universität wien

MAGISTERARBEIT

Titel der Magisterarbeit

„Statistische Analyse der Liquidität von Wertpapieren unter Verwendung von Methoden des Machine Learnings“

Verfasst von

Andreas Hackl, Bakk. rer. soc. oec.

angestrebter akademischer Grad

Magister der Sozial- und Wirtschaftswissenschaften (Mag. rer. soc. oec.)

Wien, im März 2015

Studienkennzahl lt. Studienblatt:	A 066 951
Studienrichtung lt. Studienblatt:	Magisterstudium Statistik
Betreuer / Betreuerin:	Ao. Univ.-Prof. Dr. Marcus Hudec

Eidesstaatliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht sind und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Wien, im März 2015 Unterschrift

Inhaltsverzeichnis

1) Ziele und bedeutende Aspekte der Arbeit

Diese Arbeit zielt darauf ab, Zusammenhänge zwischen der Liquidität von Wertpapieren und finanzwirtschaftlicher bzw. makroökonomischer Variablen in einem statistischen Kontext zu untersuchen. Dabei sollen allerdings nicht herkömmliche Regressionsmodelle zur Anwendung kommen, sondern es soll ein Versuch betrachtet werden, Methoden des maschinellen Lernens bei der Analyse heranzuziehen. Da eine große Vielfalt dieser Verfahren für die Anwendung infrage kommen würde, macht es Sinn, den Fokus auf ein bestimmtes Teilgebiet zu legen. Eine durchaus günstige Herangehensweise scheint die Untersuchung des oben erwähnten Zusammenhanges als ein Klassifikationsproblem zu sein, vor allem deshalb, weil sich Liquidität in einer günstigen Form als ordinal skalierte Variable definieren lässt (beispielsweise unter Verwendung von 5 Kategorien: hochliquid – liquid – mäßig liquid – gering liquid – illiquid). Bei dieser Art von Problemstellung eignen sich besonders Verfahren des sogenannten „überwachten" Lernens (engl.: „supervised learning") zur Anwendung, es kommen bei der Klassifizierung Instanzen zum Einsatz, bei denen die Klasse bereits bekannt ist. Eine detaillierte Beschreibung dieser Vorgehensweise erfolgt in Abschnitt 5.1. Die Funktionsweise der einzelnen Klassifizierungsmethoden, die zur Anwendung ausgewählt werden, wird im Abschnitt 5.2 dieser Arbeit näher beleuchtet.

Ein Aspekt sollte bei der Durchführung der Analysen, insbesondere bei der Betrachtung der zur Anwendung kommenden Variablen, die in Kapitel 2 und 3 präsentiert werden und in einem Zusammenhang mit der Liquidität stehen, und bei der Interpretation von Ergebnissen nicht vernachlässigt werden: In vielen wissenschaftlichen Arbeiten, welche die Analyse der Liquidität von Wertpapieren thematisiert haben (ein Beispiel hierfür ist J.Söderberg: „Do Macroeconomic Variables Forecast Changes In Liquidity? An Out-Of-Sample Study on the Order-Driven Stock Markets In Scandinavia" [2008]), wird die Endogenität zwischen der Liquidität und diversen mit dieser Variable in Verbindung gebrachten wirtschaftlichen Größen geschildert. Das bedeutet, dass der Zusammenhang zwischen diesen Größen in keiner einseitigen Form betrachtet werden sollte – es können nämlich wirtschaftliche Faktoren die Liquidität beeinflussen, doch umgekehrt ist dies auch möglich – Änderungen der Liquidität können auch zu einer Veränderung wirtschaftlicher Größen führen. Wird die Analyse nun als Klassifikationsproblem betrachtet (unter Verwendung bereits beobachteter Klassen bei Instanzen), so sollte man in der Lage sein, die Liquidität nicht nur als rein abhängige Variable zu sehen, sondern auch ihren möglichen eigenen Einfluss auf wirtschaftliche und ökonomische Gegebenheiten miteinzubeziehen.

Aus analytischer Sicht sollte man sich also in die Richtung einer Assoziationsanalyse bewegen, wobei hier ein explorativer Charakter im Vordergrund stehen wird (kein Testen konkreter Hypothesen, sondern die Erforschung grundlegender Zusammenhänge zwischen der Liquidität und wirtschaftlichen Faktoren steht im Mittelpunkt). Um den zuletzt erwähnten Erfordernissen entgegenzukommen, wird zusätzlich zu den überwachten Verfahren auch die Berechnung einer Clusteranalyse, die zu den nicht-überwachten Verfahren gehört und sehr

geeignet für die Untersuchung von Zusammenhängen zwischen diversen Variablen ist, durchgeführt.

Eine genauere Schilderung der Rolle der Liquiditätsbewertung von Wertpapieren im Bankenbereich (vor allem in rechtlicher Hinsicht – Basel II/III) sowie üblicher Varianten der Messung des Risikos für die bedeutendsten Wertpapierarten (im Vordergrund stehen dabei Aktien und Anleihen) erfolgt in den Abschnitten 2 und 3 dieser Arbeit.

Zur Umsetzung der Modellierung des Liquiditätsrisikos mithilfe von überwachten Klassifikationsmethoden sowie einer nicht-überwachten Variante wird ein eigens dafür zusammengestellter Datensatz verwendet, welcher Wertpapiere aus insgesamt 20 Ländern beinhaltet und sich aus 16 Variablen zusammensetzt. Die Beschreibung dieses Datensatzes geschieht in Abschnitt 4. Die Funktionsweise der Lernmethoden steht im fünften Abschnitt im Vordergrund, danach, im sechsten Abschnitt, kommt es zu einer deskriptiven Analyse des betrachteten Datensatzes, um die Eigenschaften und die Struktur des verwendeten Datensatzes besser einschätzen zu können.

Die Anwendung der präsentierten Methoden und die Darstellung der wesentlichen Resultate ist Inhalt des siebenten Kapitels. Da anfängliche Modelle, welche beim „Machine Learning“ zur Umsetzung ausgewählt wurden, in häufigen Fällen verbessert werden können, werden in der Regel Validierungsverfahren zur Beurteilung der Performance der verwendeten Modelle eingesetzt, sowie anfänglich gewählte Modellparameter durch Parametervariation optimiert. Dabei kommen auch Größen wie der „Generalization Error“ (Fehler bei der Verallgemeinerung auf neue Daten) zur Anwendung. Dies wird Thema des achten Abschnittes der Arbeit sein. Mit einer Zusammenfassung der in dieser Arbeit erzielten Ergebnisse wird die betrachtete Thematik abgeschlossen.

2) Das Liquiditätsrisiko im Bankenbereich

2.1) Grundlegende Definitionen

Der Begriff der „Liquidität“ stammt vom lateinischen Wort „liquidus“, auf Deutsch „flüssig“, und wird in der Wirtschaft als ein Maß gesehen, das widerspiegelt, unter welchen Umständen und vor allem in welcher Geschwindigkeit der Handel von wirtschaftlichen Gütern möglich ist [Amihud 2005].

Zur exakteren Beschreibung des Begriffes müssen einzelne Teilbereiche der Wirtschaft, in denen diesem Begriff eine Rolle zugespielt wird, differenziert betrachtet werden. In der Betriebswirtschaftslehre ist folgende Definition üblich:

In der Betriebswirtschaftslehre geht es um die Fähigkeit von im wirtschaftlichen Umfeld tätigen Personen bzw. Unternehmen, fällige Verbindlichkeiten zu allen Zeiten ohne Einschränkungen begleichen zu können [URL: http://www.bis.org/publ/bcbs144.htm [Zugriff am 2.11.2014]].

In der Volkswirtschaftslehre muss zwischen dem mikroökonomischen und makroökonomischen Umfeld unterschieden werden: Während in der Mikroökonomie die Qualität von Wirtschaftssubjekten betrachtet wird, Aktiva in Geld umzuwandeln [Guida 2009], geht es in der Makroökonomie um eine Größe, welche die vorhandene Geldmenge auf dem Kapitalmarkt beschreibt [Reimund 2003].

Doch besonders im finanzwirtschaftlichen Umfeld beim Handel mit Wertpapieren besitzt dieser Begriff eine überaus hochrangige Bedeutung, hier beschreibt das sogenannte Marktliquiditätsrisiko, in wie fern die Möglichkeit besteht, dass Vermögensgegenstände, im Speziellen an der Börse gehandelte Produkte wie Aktien oder Anleihen, aufgrund eines zu wenig aufnahmefähigen Marktes kaum handelbar sind bzw. in kostspieliger Form verkauft werden müssen. Dabei kann auf einem Markt letztere Situation dann auftreten, wenn entweder Zahlungsmittel in zu geringer Menge vorhanden sind, oder wenn kein für die gewünschte Transaktion geeigneter Partner gefunden werden kann, dessen Interessen im Einklang mit den eigenen Vorstellungen stehen [Pohl 2008]. Eine sehr geringe Marktliquidität zeichnet sich üblicherweise dadurch aus, dass einzelne Transaktionen große, sprunghafte Veränderungen der Marktpreise hervorrufen.

Die Beschreibung des letzten Absatzes lässt erahnen, dass das Vorhandensein von Liquidität auf Finanzmärkten abhängig ist von diversen Rahmenbedingungen, die im Umfeld der Börsen zum Handelszeitpunkt vorherrschen. Allgemeine Stimmungslagen der Märkte, deren Veränderungen über die Zeit und auch Wirtschaftszyklen können je nach gehandeltem Produkt einen großen Einfluss auf Eigenschaften von Handelsvorgängen haben. Insbesondere

seit dem Jahr 2008 erlangte die Fragestellung, wie schnell Kunden von Wertpapieren, seien es Portfoliomanager oder Privatkunden, in heiklen Situationen auf den Finanzmärkten ihre Bestände zu adäquaten Preisen verkaufen können, einen beachtlichen Aufwind, sodass das Liquiditätsrisikomanagement mittlerweile neben der Abwägung des Kreditrisikos eine der wesentlichen Aufgaben ist, um welche sich Risikomanager in Großbanken kümmern müssen. Im Kontext des Handelns von Wertpapieren bedeutet das Vorliegen hoher Liquidität also das Vorhandensein genügend vieler potentieller Handelspartner, welche ein Finanzprodukt zu einem Preis handeln (also kaufen oder verkaufen) möchten, der im Einklang zu den Wünschen ihrer Partner steht. Dies würde folglich bedeuten, dass man beim Handel von liquiden Finanzprodukten von fairen Preisen ausgehen kann – Käufer und Verkäufer können also unter dieser Voraussetzung die Geschäfte in der Regel in zufriedenstellender Form abwickeln.

Es gibt nun eine Vielzahl möglicher wirtschaftlicher Faktoren, die in der Literatur der letzten Jahre hinsichtlich ihres Zusammenhanges mit dem vorherrschenden Stand der Liquidität untersucht und diskutiert wurden:

Im Paper „Liquidity and Asset Prices“ [Amihud 2005] werden einige allgemeine Faktoren erläutert, die Illiquidität bei Wertpapieren hervorrufen können:

-) Transaktionskosten: zu diesen gehören Maklergebühren, Spesen, die bei der Durchführung von Käufen bzw. Verkäufen anfallen, oder Steuern wie die Kapitalertragsteuer

-) Druck bei der Nachfrage nach Wertpapieren – erhöhte Kosten bei Nichtauffinden passender Handelspartner zur richtigen Zeit, wenn Finanzprodukte in einem zeitlich beschränkten Rahmen verkauft werden sollen

-) Bestandsrisiko (entstehende Kosten bei Notwendigkeit des Einsetzens eines Kursmaklers beim Verkauf)

-) Privatinformation von Handelspartnern – ein Beispiel dafür ist die Entstehung erschwerter Verkaufsbedingungen, wenn negative Nachrichten, die sich auf das im Fokus stehende Unternehmen beziehen, die jeweiligen Handelspartner erreichen

-) Auffinden von geeigneten Handelspartnern, insbesondere beim Handeln von Wertpapieren, deren Horizont auf dem Markt eingeschränkt ist (ein Beispiel dafür sind exotische Optionen oder andere spezielle Derivate)

-) Neben diesen Aspekten, die im Wesentlichen abhängig von den Händlern der Wertpapiere sind, gibt es auch zahlreiche Indikatoren, die sich auf Eigenschaften der Wertpapiere im Handelsverlauf beziehen. Diese Indikatoren, zu denen unter anderem die „Geld-Brief-Spanne“ (engl.: Bid-Ask-Spread) gehört, werden in detaillierter Form im Kapitel 3 dieser Arbeit (Bedeutung der Liquiditätsrisikomessung im Bankenbereich) erläutert.

-) Liquidität ist zwar eine Größe, welche beim Handel mit Wertpapieren eher mikroökonomisch geprägt wird, doch es wurde unter anderem in der Arbeit „Does monetary policy determine stock market liquidity? New evidence from the euro zone“ [Fernandez-Amador et al. 2013] dargestellt, dass auch eine makroökonomische Größe wie die Geldpolitik einer zentralen Notenbank (also die Veränderung des Zinssatzes durch den Kauf oder den Verkauf von Anleihen) einen nicht unwesentlichen Einfluss auf die Liquidität von Wertpapieren hat. Auf diesen Aspekt wird in den nächsten Kapiteln dieser Arbeit auch noch näher eingegangen.

-) Wie bereits auf der letzten Seite dargestellt wurde, ist die Liquidität von Wertpapieren stark abhängig von allgemeinen Stimmungslagen im Bereich der Weltwirtschaft. Besonders kleinere Wertpapiere, die sich nicht in einem der großen Indizes der führenden Wirtschaftsnationen befinden, können dazu neigen, bei krisenähnlicher Stimmung in Sachen Liquidität starke Einbußen zu erleiden, während in Zeiten einer „Börsenrally“ viele Anleger dazu neigen, ein höheres Risiko bei ihren Handelsstrategien eingehen zu wollen, und somit auch zu weniger populären Papieren, denen Chancen eingeräumt werden, in deutlich häufigerer Form greifen.

2.2) Gesetzliche Rahmenbedingungen

Im letzten Kapitel stand der Begriff der Liquidität in seiner allgemeinen Form im Vordergrund. Nun soll etwas konkreter auf das Liquiditätsrisiko und auf den (von außen vorgegebenen) Umgang mit dieser Thematik eingegangen werden, welchem vor allem seit der Finanzkrise des Jahres 2008, die bereits kurz im ersten Abschnitt erwähnt wurde, erhöhte Aufmerksamkeit im Bankenbereich geschenkt wird:

Die wesentlichen Rahmenbedingungen für Banken beim Liquiditätsrisikomanagement sind Teil der Pakete der Eigenkapitalvorschriften, deren erste Version, die kurz als „Basel I“ bezeichnet wird, vom Basler Ausschuss für Bankenaufsicht im Jahr 1988 veröffentlicht wurde; eine neuere, bis vor kurzem noch gültige Fassung mit dem Namen „Basel II“ stammt aus dem Jahr 2004 (wobei nach den EU-Richtlinien 2006/48/EG und 2006/49/EG diese seit dem 1. Jänner 2007 für alle Kredit- und Finanzdienstleistungsinstitute der EU-Mitgliedsstaaten bindend war). Die damals festgelegte Regulierungsstruktur bedurfte aufgrund der im ersten Absatz erwähnten Finanzkrise einer Überarbeitung, aus dieser resultierte die seit 1. Jänner 2014 in der europäischen Union gültige Fassung der Bankenregulierung mit dem Namen „Basel III“.

Die durch die Krise notwendig gewordenen Anpassungen der Eigenkapitalvorschriften betreffen das Management des Liquiditätsrisikos in einem beträchtlichen Ausmaß. Dies liegt vor allem an der Tatsache, dass „[…] das inakkurate und ineffektive Management des Liquiditätsrisikos ein Schlüsselcharakteristikum der Finanzkrise war“, und „[…] über die Finanzkrise hinweg, die Mitte 2007 begann, viele Banken mit der Aufrechterhaltung einer adäquaten Liquidität zu kämpfen hatten“ [URL: http://www.bis.org/publ/bcbs165.pdf [Zugriff am

4.3.2014]]. In der Version von „Basel III“ („Internationale Rahmenvereinbarung über Messung, Standards und Überwachung in Bezug auf das Liquiditätsrisiko“) vom Dezember 2010, der mittlerweile in der ganzen europäischen Union gültigen Fassung dieses Paketes (einzelne Aspekte wurden in den darauf folgenden Jahren noch ergänzt), auf die auch im nächsten Absatz eingegangen wird, wird gleich zu Beginn auf Seite 1 auf die mangelnde Fähigkeit zahlreicher Bankinstitute zur Liquiditätssteuerung, unabhängig vom vorhandenen Eigenkapital, aufgrund der aufgetretenen „Missachtung elementarer Grundsätze“ hingewiesen (bezogen auf das Jahr 2007, dem Beginn der Krise) [URL: http://www.bis.org/publ/bcbs188_de.pdf [Zugriff am 6.4.2014]].

Es wird also in diesen Abschnitten der Festlegungen vom Basler Ausschuss für Bankenaufsicht verdeutlicht, dass vor allem die mangelhafte Durchführung von Tätigkeiten, die die Evaluierung des Liquiditätsrisikos betreffen, eine wesentliche Quelle für die schwerwiegendsten Probleme in der Finanzbranche seit Jahrzehnten dargestellt hat, dass dies über eine durchdachtere Mitberücksichtigung zwar nicht völlig zu verhindern war, aber zumindest negative Entwicklungen stärker eingebremst hätten werden können. Somit wurde ein rascher Handlungsbedarf hinsichtlich der Weiterentwicklung dieser Fassung herbeigeführt.

Nun steht, als Reaktion auf die in den Krisenjahren zum Vorschein gekommenen Schwachstellen, das Liquiditätsrisiko in „Basel III“ deutlich mehr im Vordergrund, als dies in der Vorgängerversion der Fall war, welche immerhin für zehn Jahre die Grundlage für viele Tätigkeiten der Großbanken dargestellt hat. Die Weiterentwicklung dieser Version bezog sich vorwiegend auf die detaillierte Beschreibung von Mindeststandards bezüglich der Ausstattung der Liquidität, an welche Banken von nun an gebunden sind, sowie der Generierung adäquater, universell einsetzbarer Überwachungsmechanismen.

Die Mindeststandards für international tätige Banken sind nun wie folgt festgelegt (der Inhalt der folgenden drei Seiten dieser Arbeit bezieht sich auf die Quelle: [URL: http://www.bis.org/publ/bcbs188_de.pdf [Zugriff am 6.4.2014]] - die gesamte Beschreibung findet sich für den ersten Standard auf den Seiten 4 bis 28 des Dokuments, für den zweiten Standard auf den Seiten 28 bis 34):

Standard 1: Festlegung einer „Mindestliquiditätsquote“ (LCR – „Liquidity Coverage Ratio“)

$$\text{LCR} = \frac{\textit{Bestand an erstklassigen liquiden Aktiva}}{\textit{Gesamter Nettoabfluss von Barmitteln in den nächsten 30 Kalendertagen}} >= 100\%$$

Dieser Quotient stellt eine unmittelbare Verbindung zwischen der Liquiditätssituation von Finanzinstituten und dem Zustand der von ihnen gehaltenen Aktiva (bzw. Wertpapiere in allgemeiner Form) dar (dies zeigt auch auf, dass eine marktdatenbasierte Umsetzung der Beurteilung von Liquiditätsrisiken, die im weiteren Verlauf der Arbeit in den Vordergrund rückt, für Banken auch durchaus Sinn macht). Er dient als Schutzmechanismus für kurz (bis zu 30 Tage) andauernde, als kritisch einzustufende Perioden und hat folgenden Hintergrund:

Ein großer Wert erstklassiger liquider Aktiva, welche den Zähler in obiger Formel darstellen, kann ein robustes Liquiditätsrisikoprofil zumindest für eine vierwöchige Periode garantieren, wodurch auch turbulente Phasen dieser Zeitdauer auf den Weltmärkten, welche Banken typischerweise schwer unter Druck setzen („Stressszenario"), gut überstanden werden können. Diese Garantie beruht darauf, dass hochliquide Bestände die Eigenschaft haben, unabhängig von den wirtschaftlichen Rahmenbedingungen jederzeit verkauft und in Barmittel umgewandelt werden zu können. Der Nenner, welcher den gesamten Nettoabfluss von Barmitteln in den nächsten 30 Kalendertagen umfasst, soll nun mit dem Bestand an erstklassigen Aktiva in einer derartigen Beziehung stehen, dass durch den Verkauf der liquiden Aktiva in einer turbulenten Phase es zu einer größeren Einnahme an Barmitteln kommt, als in dieser Periode abfließt, das heißt, der Quotient soll Werte größer gleich 1 annehmen. Die Regelung bildet damit in solchen Situationen eine Art „Schutzpolster" .

Da im weiteren Verlauf dieser Arbeit die Klassifizierung von Wertpapieren mittels verschiedener (statistischer) Verfahren, welche von diversen Eigenschaften der Finanzprodukte abhängen, im Vordergrund steht, werden die Aspekte, welche mit dem Zähler dieses ersten Kriteriums in Verbindung stehen und in „Basel III" sehr exakt und sorgfältig beschrieben werden (was dort nicht nur für den Zähler gilt), aufgrund ihrer Bedeutung für die Arbeit auch hier detaillierter behandelt:

Erstklassige liquide Aktiva besitzen in diesem Kontext unter anderem folgende wesentliche Eigenschaften, welche selbst in schwierigen Phasen, egal ob diese nur regionale oder globale Auswirkungen mit sich ziehen, eine einfache Liquiditätsbeschaffung garantieren können:

-) Notenbankfähigkeit (Idealfall – akzeptiert von der Zentralbank bei Bereitstellung von Innertagesliquidität und Overnight-Liquiditätsfazilitäten)
-) Geringes Kredit- und Marktrisiko
-) Notation an Börse mit internationalem Stellenwert
-) unproblematische Bewertung beim Handel
-) marktbezogene Merkmale (z.B. Aktivität / Bedeutsamkeit des Marktes, Präsenz engagierter Marktmacher)

Um ein klareres Bild darüber zu hinterlassen, was zu erstklassigen liquiden Aktiva gezählt werden kann, wird in weiterer Folge in „Basel III" auch eine Unterscheidung zwischen Aktiva der „Stufe 1" und Aktiva der „Stufe 2" vollzogen, wobei letztere hinsichtlich ihres Anteils am Bestand eines Wertpapierdepots innerhalb einer Stressperiode beschränkt sind (ihr Anteil darf maximal 40% betragen) . Sie unterscheiden sich vor allem durch folgende Merkmale:

-) Barmittel und Zentralbankguthaben zählen allgemein zu Aktiva der Stufe 1

-) marktgängige Wertpapiere, die an großen, tiefen und aktiven Repo- und Kassamärkten gehandelt werden und keine Verbindlichkeit eines Finanzinstituts bzw. mit ihm verbundenen Unternehmen darstellen, zählen zur Stufe 1, falls sie ein Risikogewicht von 0% unter „Basel II" aufweisen, liegt dieses Gewicht bei 20%, so fallen sie in die andere Kategorie.

-) Zur Stufe 1 zählen auch Staats-, Zentralbankschuldtitel in Landeswährung, „[…] die vom betreffenden Staat oder der Zentralbank in dem Land, in dem das Liquiditätsrisiko anfällt, oder im Herkunftsland der Bank begeben werden“ bzw. „inländische Staats- oder Zentralbankschuldtitel in Fremdwährungen, soweit das Halten solcher Wertpapiere dem Währungsbedarf für die Geschäfte der Bank in jenem Land entspricht […]“.

-) Stufe 2 umfasst auch Unternehmensanleihen bzw. gedeckte Schuldverschreibungen, welche einer anderen Quelle als einem Finanzinstitut (beim zweiten dem Institut selbst) oder einem Partnerunternehmen entstammen und welche zumindest ein „AA-“ Rating von einer Ratingagentur aufweisen können.

Typischerweise wird den Vermögenswerten der zweiten Stufe, die zum Bestand gehören, 15% oder mehr vom aktuellen Marktwert abgezogen.

Das auf Seite 6 vorgestellte Konstrukt stellt eine in „Basel III“ festgelegte, bindende Form eines Stresstests dar, im Zuge der Erläuterungen im Basler Bankenpaket wird darauf hingewiesen, dass dies als eine Mindestanforderung zu sehen ist, und es werden Institute aufgefordert, mittels Berücksichtigung des eigenen Umfeldes und ihres Tätigkeitsprofils diese durch eigene Regeln und Erweiterungen flexibler zu gestalten.

Damit ein klares Verständnis des Aufbaus der Mindestliquiditätsquote erzielt wird, ist es nun noch wichtig, den Begriff des Nettoabflusses, welcher Teil des Nenners im Quotient ist, deuten zu können: Laut Basel III handelt es sich dabei um die Differenz zwischen den Erwartungen aller vorliegender Barmittelabflüsse und aller Zuflüsse von Mitteln in einem gewissen Zeitraum (üblicherweise 30 Tage – die Dauer eines Stressszenarios). Der letzte Wert stellt dabei das Produkt zwischen aller offenen Salden vertraglicher Forderungen und der Rate, die im betrachteten Zeitraum für diese relevant ist, dar. Übersteigt dieses Produkt den 75%-Anteil der gesamten Abflüsse von Mitteln, so wird dieser Anteil anstelle der Zuflüsse den zweiten Teil der Differenz einnehmen.

Der zweite festgelegte Mindeststandard sieht nun wie folgt aus:

Standard 2: Festlegung einer „strukturellen Liquiditätsquote“ (NSFR – „Net Stable Funding Ratio“)

$$\text{NSFR} = \frac{Verfügbarer\ Betrag\ stabiler\ Refinanzierung}{Erforderlicher\ Betrag\ stabiler\ Refinanzierung} > 100\%$$

Während die „Liquidity Coverage Ratio“ einen Standard mit kurzfristiger Ausrichtung (hin auf Stressszenarien mit einer Dauer von maximal 30 Tagen) darstellt, geht es nun um die mittel- bzw. langfristige Aufrechterhaltung und Sicherstellung stabiler Geschäftsverhältnisse in einer international tätigen Bank. Mit der vorhin definierten Größe wird festgelegt, dass der verfügbare Anteil des Betrages von Mittelquellen, welcher über einen längeren Zeitraum hinweg (über ein Jahr) auch bei Auftreten schwieriger Phasen als zuverlässig gesehen werden

kann (als „stabile Refinanzierung“ bezeichnet), größer sein muss als jener Betrag, welcher für das Institut erforderlich ist, um einen stabilen Status zu gewährleisten.

Zu den Mittelquellen, die zum verfügbaren Betrag gerechnet werden können, zählen das Eigenkapital einer Bank, Vorzugsaktien mit weniger als einem Jahr Restlaufzeit, Verbindlichkeiten, die effektiv zumindest noch über diesen Zeitraum hinweg laufen, sowie Teile von Einlagen ohne Fälligkeit bzw. Termineinlagen und Mittel von Großkunden mit endender Laufzeit innerhalb eines Jahres, die auch im Falle eines kritischen Szenarios der Bank zugehörig bleiben würden. Der erforderliche Betrag hängt im Gegensatz dazu von diversen Liquiditätsmerkmalen der Aktiva einer Bank ab – hierbei wird ein Produkt zwischen der Summe der gehaltenen Aktiva und einem von der Art der Aktiva abhängigen Faktor, welcher die erforderliche stabile Refinanzierung widerspiegelt, berechnet, dazu kommt noch ein weiteres Produkt, welches sich aus dem Refinanzierungsfaktor und Geschäften, welche als ausserbilanziell gelten, zusammensetzt. Ist jener Faktor niedrig, so bedeutet dies im Allgemeinen das Vorliegen von höherer Liquidität beim betrachteten Aktivum (wodurch die Refinanzierung weniger stabil ablaufen muss).

Neben den umfangreichen Anforderungen, welche den Banken durch Basel III auferlegt werden, gibt es weitere nationale und internationale Behörden, deren Bestimmungen bezüglich des Liquiditätsrisikomanagements von den Banken zu befolgen sind:

-) European Banking Authority

.) „CEBS Guidelines on Liquidity Buffers & Survival Periods“ (2009)
.) „CEBS revised Guidelines on Stress Testing“ (2010)
.) „CEBS Guidelines on Liquidity Cost Benefit Allocation“ (2010)

Hierbei handelt es sich um Richtlinien der europäischen Finanzmarktaufsichtsbehörde zu diversen mit der Liquidität in Verbindung stehenden Themen.

-) Institute of International Finance

.) „Principles of Liquidity Risk Management“ (2007)

Dieses von einer weltweiten Verbindung von Finanzinstituten vor der Wirtschaftskrise veröffentlichte Konzept zielt auf die Vorstellung geeigneter Praktiken zur Steuerung des Liquiditätsrisikos ab und macht hierfür Vorschläge zur Messung, Überwachung und Kontrolle, auch werden dabei Stresstests und Kontingenzplanung in den Vordergrund gerückt [URL: http://www.iif.com/download.php?id=tv/yzuHHXJ0= [Zugriff am 4.3.2014]].

Neben den Richtlinien und Empfehlungen auf internationaler (europäischer) Ebene gibt es auch Bestimmungen auf nationaler Ebene, welche durch die folgenden Organe beaufsichtigt werden:

-) Österreichische Nationalbank (ÖNB)

Gemeinsam mit der Finanzmarktaufsicht (FMA) ist die österreichische Nationalbank für die Bankenaufsicht im nationalen Raum zuständig, wobei sich die Tätigkeit der ÖNB auf die Makroebene beschränkt (der Fokus liegt auf der obersten Bankenebene).

-) Finanzmarktaufischt (FMA)

Das Aufgabenfeld des neben der Nationalbank als zweites Aufsichtsorgan agierenden Institutes in Österreich umfasst die Überwachung des Bankenwesens auf der Mikroebene, also der einzelnen Akteure und Institute im Finanzsektor
[URL: http://www.fma.gv.at/de/ueber-die-fma.html [Zugriff am 16.10.2014]].

Die grundlegenden nationalen Bestimmungen sind im Bankwesengesetz abgefasst, es lautet dabei im 7. Unterabschnitt: Liquidität § 25 wie folgt [URL: http://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10004827 [Zugriff am 16.10.2014]]:

„(1) Ungeachtet der Verpflichtungen gemäß § 39 Abs. 3 und gemäß einer Verordnung der FMA gemäß § 39 Abs. 4 Z 7 haben Kreditinstitute als Mindesterfordernis flüssige Mittel ersten und zweiten Grades gemäß den Abs. 2 bis 11 zu halten. […]“

„(10) Flüssige Mittel zweiten Grades sind jeweils zum Monatsletzten zumindest im Ausmaß von 25 vH der Verpflichtungen gemäß Abs. 8 zum 15. des gleichen Kalendermonats oder des letzten vorangegangenen Geschäftstages zu halten. Die FMA kann diesen Hundertsatz innerhalb einer Bandbreite von 10 vH bis 30 vH durch Verordnung ändern, wenn dies nach den währungs- und kreditpolitischen Verhältnissen zur Aufrechterhaltung der Zahlungsbereitschaft erforderlich ist. Für Verpflichtungen gemäß Abs. 2 vermindert sich der Hundertsatz um den gemäß Abs. 5 Z 2 festgelegten Satz für flüssige Mittel ersten Grades.“

„(11) Die FMA kann die in den Abs. 4 und 8 genannten flüssigen Mittel ersten und zweiten Grades im Wege einer Verordnung durch andere Werte gleicher Flüssigkeit ergänzen. Dabei ist auf das volkswirtschaftliche Interesse an einem funktionsfähigen Bankwesen Bedacht zu nehmen.“

Auch in anderen Paragraphen finden sich Aussagen zur Steuerung der Liquidität. Paragraph 70 (1) gibt der Finanzmarktaufsicht die Befugnis zur nationalen Aufsicht auf Mikroebene, hier wird die FMA dazu verpflichtet, gemeinsam mit der ÖNB für das jeweils folgende Kalenderjahr ein Prüfungsprogramm festzulegen; die Institute, die in die Mikro-Kategorie fallen, müssen nach Paragraph 74 (2) nach Ablauf eines jeden Kalendervierteljahres Auskunft darüber erteilen, dass die Bestimmungen des Paragraph 25 von ihnen eingehalten wurden.

Bis Ende 2013 galt die im §25 des Bankwesengesetzes abgefasste, von der FMA eigens deklarierte Liquiditätsrisikomanagementverordnung (LMRV), welche Mindestanforderungen des Liquiditätsrisikomanagements beinhaltet, diese trat am 31.12.2013 aufgrund der Novelle des Bankwesengesetzes §25 Abs. 2 außer Kraft. Bis dahin musste die FMA nach bisheriger Rechtslage durch Verordnung die Mindestanforderungen für die in § 25 Abs. 1 BWG genannten Richtlinien bezüglich der Liquidität festlegen. Seit 1.1.2014 finden sich die Verordnungen der FMA hinsichtlich des Liquiditätsrisikomanagements in der gemäß § 39 Abs. 4 BWG idF BGBl. I Nr. 184/2013 von der FMA zu erlassenden Risikomanagementverordnung [URL: http://www.fma.gv.at/typo3conf/ext/dam_download/secure.php?u=0&file=11004&t=1384294085&hash=83c55aa78e0a89dba1821c176d66fd9a [Zugriff am 16.10.2014]].

In dieser Risikomanagementverordnung (kurz: KI-RMV) sind Bestimmungen zum Thema Liquidität in Paragraph 12 abgefasst. Dort ist unter anderem folgendes festgelegt [URL: http://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20008739 [Zugriff am 16.10.2014]]:

„(1) Kreditinstitute haben über geeignete Strategien, Grundsätze, Verfahren und Systeme für die Identifizierung, Messung, Steuerung und Überwachung des Liquiditätsrisikos über eine angemessene Zahl von Zeiträumen, einschließlich innerhalb eines Geschäftstages, zu verfügen, um sicherzustellen, dass sie über angemessene Liquiditätspuffer verfügen. […]“

„(4) Kreditinstitute haben unter Berücksichtigung der Art, des Umfangs und der Komplexität ihrer Geschäfte die Angemessenheit ihrer Liquiditätsrisikoprofile laufend sicherzustellen und darauf zu achten, dass das jeweilige Risikoprofil für das Funktionieren und die Solidität des Finanzsystems erforderlich ist, nicht aber darüber hinausgeht und dadurch unangemessene Systemrisiken (§ 2 Z 41 BWG) generiert.“
„(5) Kreditinstitute haben über Methoden für die Identifizierung, Messung, Steuerung und Überwachung von Refinanzierungspositionen zu verfügen. […]“

„(8) Kreditinstitute haben verschiedene Vorkehrungen zur Minderung des Liquiditätsrisikos, einschließlich Limit-Systeme und Liquiditätspuffern, zu treffen, um unterschiedlichen Stresssituationen standhalten zu können. Sie haben Vorkehrungen zur Sicherstellung einer hinreichend diversifizierten Refinanzierungsstruktur und des Zugangs zu Refinanzierungsquellen zu treffen. Diese Vorkehrungen sind regelmäßig zu überprüfen.“

„(9) Kreditinstitute haben Stresstests für Liquiditätspositionen und Risikominderungsfaktoren zu erstellen. […]“

„(12) Kreditinstitute haben wirkungsvolle Notfallkonzepte, welche die Ergebnisse der Stresstests nach Abs. 9 berücksichtigen, zu erstellen. […]“

2.3) Gängige Methoden zur Umsetzung der Liquiditätsrisikomessung in Banken

Im letzten Abschnitt wurden Konzepte und gesetzliche Vorschriften präsentiert, erstere sollen für die Banken ein Grundgerüst bilden, die Banken sind insbesondere in Basel III dazu aufgefordert, für eigene Bedürfnisse maßgeschneiderte Instrumente zu entwickeln, da das Liquiditätsrisiko stark vom Tätigkeitsprofil des jeweiligen Institutes bestimmt wird, und daher eigens abgestimmte Spezifikationen bei der Messung unumgänglich sind.

Dieser Abschnitt soll nun einen Überblick darüber geben, welche Methoden in Banken, aufbauend auf den gesetzlichen Grundlagen, zur Steuerung, Messung und Beurteilung des Liquiditätsrisikos gegenwärtig einen hohen Stellenwert besitzen:

-) Liquidity at Risk (LaR)
[URL: http://www.risiko-manager.com/index.php?id=162&tx_ttnews%5Bcat%5D=71&tx_ttnews%5Btt_news%5D=8737&tx_ttnews%5BbackPid%5D=779&cHash=5f60f846e0f4098ce12ac8345dae58ee [Zugriff am 4.5.2014]]

Dieses Maß ähnelt dem im Risikomanagement häufig betrachteten „Value at Risk“ (VaR) und hat ebenfalls einen statistischen Hintergrund. Es gibt ein vorab festgelegtes p % - Quantil der Belastung der Liquidität in einem bestimmten Zeitraum an, also diejenige Belastung, die mit einer Wahrscheinlichkeit von p% nicht übertroffen wird. Betrachtet werden hierbei die Volumina auf der Ebene der täglichen Zahlungsströme, also die von regelmäßigen Cash Flows. Eine Möglichkeit, diese Größe zu bestimmen, ist die Herannahme historischer Cash Outflows, und Bestimmung des p% - Quantils der Verteilung der Differenz zwischen den Beträgen der Cash Outflows und dem Durchschnitt aller Werte.

Dieses Maß wird für die Beurteilung des Zustandes der Liquidität in einer kurzen Frist verwendet (dispositives Liquiditätsrisiko), es kann auch als Indikator bei der Simulation von Stresssituationen verwendet werden, wobei hier Cash Flows in wirtschaftlich heiklen Phasen nachgebildet werden (hohe Flüsse in kurzer Frist), und typischerweise Quantile am Rand der Verteilung (95%, 99%) herangezogen werden.

-) Liquidity Value at Risk (LVaR)
[URL: http://www.risiko-manager.com/index.php?id=162&tx_ttnews%5Bcat%5D=71&tx_ttnews%5Btt_news%5D=8737&tx_ttnews%5BbackPid%5D=779&cHash=5f60f846e0f4098ce12ac8345dae58ee [Zugriff am 4.5.2014]]

Im Gegensatz zum LaR wird beim „Liquidity Value at Risk“ nicht das Risiko einer Zahlungsunfähigkeit bei täglichen Geschäften betrachtet, sondern das Risiko, das durch Kosten der Refinanzierung auftritt, diese Thematik spielt in mittel- und langfristiger Sicht eine tragende Rolle. Damit ist der „Liquidity Value at Risk“ eine wesentliche Kenngröße für das strukturelle Liquiditätsrisiko. Bedeutend ist in diesem Kontext die Feststellung, welche Mittel zur Refinanzierung sich für die Rentabilität als optimal herausstellen. Aufgrund der

Mitberücksichtigung des Preises bei dieser Größe hat man es hier nun nicht mit einer Volumensgröße zu tun, sondern mit einer Vermögensgröße. Daher liefert der „Liquidity Value at Risk“ eine mit einer gewissen Wahrscheinlichkeit gehaltene obere Schranke für einen Vermögensverlust aufgrund dieser Refinanzierungskosten.

Zu den wesentlichen Faktoren, die bei der Bestimmung des „Liquidity Value at Risk“ eine Rolle spielen, gehört das Rating des jeweiligen Instituts (eine Verschlechterung hat negative Folgen für das Institut, da ein schlechteres Rating dazu führt, dass der Markt vom Institut höhere Zinsen verlangen wird – in diesem Kontext spricht man von einer Erhöhung des Liquiditätsspreads) und die aktuelle Zinsstrukturkurve (Geldmarkt- bzw. Swapkurve). Verändern sich erwartete Werte bei der im Folgenden betrachteten Liquiditätslaufbilanz, so wird dies auch Auswirkungen auf diese Größe haben.

-) Liquiditätsablaufbilanz: [Fink 2012]

Wie der „Liquidity Value at Risk“ liegt die Bedeutung dieser Größe im Bereich des strukturellen Liquiditätsrisikos. Sie wird zur Darstellung der Geschäfte im aktivseitigen und passivseitigen Umfeld und im Bereich außerhalb der Bilanz verwendet (Anzeige eines strukturellen Bedarfs bzw. eines strukturellen Überschusses an Liquidität), und dient somit als Indikator, durch welchen maßgebliche Informationen über den aktuellen Zustand der Liquidität und der mit ihr unmittelbar in Verbindung stehenden Größen gewonnen werden können. Sie kann sowohl als Warnsystem in der kurzen Frist als auch als Planungsinstrument eingesetzt werden, bei letzterem muss allerdings auf lange Sicht die Aussagekraft von dieser und mit ihr in Verbindung stehender Größen aufgrund nicht völlig vorhersehbarer Entwicklungen und Trends auf den Märkten (auch in engerem Rahmen – unvorhersehbare Zahlungsströme sind häufig möglich) mit Vorsicht betrachtet werden.

Die Erstellung dieser Bilanz erfolgt unter Betrachtung verschiedener Szenarien, auch Stresssituationen werden hierbei mitberücksichtigt.

2.4) Beispiele für Kenngrößen und spezieller Verfahren zur Bestimmung der Liquidität verschiedener Bestandsarten (im Kontext des stammdatenbasierten Liquiditätsrisikos)

Es sollen in weiterer Folge nun speziellere Varianten bzw. Größen vorgestellt werden, die im Bankenbereich zur Bestimmung der Liquidität von verschiedenen Bestandsarten eingesetzt werden. Diese Bestimmung steht in einem engen Zusammenhang mit der allgemeinen Liquiditätssituation des Institutes, dem diese Bestände zuzuordnen sind, da die Liquidität der Bestände bestimmt, in wie fern bzw. wie schnell das Institut bei Bedarf die Bestände in

Kapital umwandeln kann. Wie vor allem an der ersten Variante erkannt werden kann, spielen spezielle Ansätze aus dem Bereich der Statistik dabei eine nicht zu unterschätzende Rolle:

BVI-Methode

[URL: http://www.risiko-manager.com/index.php?id=162&tx_ttnews%5Bcat%5D=185&tx_ttnews%5Btt_news%5D=17501&tx_ttnews%5BbackPid%5D=773&cHash=fab343bffe67888d2b8a908cf2133d54 [Zugriff am 8.5.2014]]

Im Bereich des Fondsmanagements existiert ein vom Bundesverband Investment und Asset Management e. V. (BVI) getragenes Konzept, welches auf die Überwachung der mit der Liquidität zusammenhängenden Prozesse abgestimmt ist. Ein Herzstück dieser Methodik bildet dabei die Quantifizierung der für die Abschätzung des Liquiditätsrisikos bedeutenden Nettomittelabflüsse (im letzten Abschnitt als Cash Outflows bezeichnet), wobei der Fokus auf die Abflüsse unter Stresssituationen gelegt wird. Aus methodischer Sicht bedient man sich dabei Verfahren aus dem Bereich der Extremwertstatistik – dies kann vor allem aus jenem Grund von Nutzen sein, dass man sich bei der Prognose nicht nur auf historische Simulationen stützen muss (auftretende Ereignisse sind in der Regel nur zu wenigen Zeitpunkten als extrem einzustufen – Extremwertverfahren erlauben auch Schluss auf historisch noch nie verifizierte Stresssituationen) .

Aus statistisch-methodischer Sicht ist eine der Extremwertverteilungen, die zum Einsatz für die Schätzung der Abflüsse kommt, die folgende Extremwertverteilung in generalisierter Form:

$$H_{\xi,\sigma,\mu}(x) = \begin{cases} \exp\left\{-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right\}, & \text{falls } \xi \neq 0 \\ \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}, & falls\ \xi = 0 \end{cases} \quad (1)$$

Auch die generalisierte Paretoverteilung wird für diese Aufgabe in Betracht gezogen:

$$H_{\beta,\gamma}(x) = \begin{cases} 1-\left(1+\frac{\gamma x}{\beta}\right)^{-\frac{1}{\gamma}} & für\ \gamma \neq 0 \\ 1-\exp\left(-\frac{x}{\beta}\right) & für\ \gamma = 0 \end{cases} \quad (2)$$

Hier gilt: $\beta > 0$, $x \geq 0$, $\gamma \geq 0$, $0 \leq x \leq -\frac{\beta}{\gamma}$, falls $\gamma < 0$

Eine wichtige Rolle ist die adäquate Schätzung der Parameter für die hier gegebene Ausgangslage. Dabei können typische statistische Verfahren der Punktschätzung, wie die

Maximum-Likelihood-Methode, zum Einsatz kommen. In diesem Kontext setzt man, aufgrund der hohen Komplexität der Punktschätzung von Parametern der Extremwertverteilungen, oft auf die Parametrisierung von Verteilungsfunktionen (Anwendung genetischer Algorithmen zur Parameterschätzung) und daraus folgender Ermittlung von Quantilen über die Umkehrfunktion, wo dann auch grafische Darstellungen (QQ-Plots) bzw. quadratische Fehlerwerte (RMSE – Wurzel des mittleren quadratischen Fehlers) eine Rolle einnehmen.

Bloomberg BVAL Scores

[URL: http://cdn.gotraffic.net/enterprise/pdf/DS_bval_score_primer_final_9.21.10.pdf [Zugriff am 11.5.2014]]

Der Finanzdienstleister Bloomberg zählt zu den wichtigsten Datenlieferanten für institutionelle Kunden sowie Investmentbanken. In den letzten Jahren wurde von diesem der sogenannte „Bloomberg Valuation Service“ (BVAL) ins Leben gerufen. Der Hauptgedanke besteht darin, Kunden im Finanzumfeld ein hochwertiges Service im Bereich der Datenqualität anzubieten. Dabei sollen die großen Datenmengen für die im Bloomberg-System implementierten Wertpapiere, welche einer sonst selten erreichten Vielzahl an Quellen entstammen, genutzt werden, um bei besonders unregelmäßig gehandelten und daher schwierig zu bepreisenden Papieren eine Metrik bereitzustellen, welche ein objektives und gut vergleichbares Resultat liefert. In diesem Kontext stellt diese Metrik eine Art „Konfidenzgrad“ zur Beurteilung des Preises von Wertpapieren dar.

Laut Bloomberg wird dabei ein Service bereitgestellt, „[…] das durch seine Breite an Beobachtungen, zusammen mit innovativen Analysewerkzeugen, Ausdrücken und Datenqualität, hochwertige und objektive Preisbewertungen im Sinne eines Drittanbieters für „Fixed-Income“ – Wertpapiere (wichtigste Formen von Anleihen, unter anderem Staats- und Unternehmensanleihen) und Derivate erstellt“
[URL: http://www.bloomberg.com/enterprise/data/pricing-services/ [Zugriff am 16.10.2014]].

Ein wesentlicher weiterführender Schritt ist die Beurteilung der Qualität und Zuverlässigkeit der Daten, die für die vorher genannte Preisbestimmung verwendet werden. Die für die Ermittlung des BVAL-Preises verwendeten Daten werden im Kontext der Ermittlung der sogenannten „BVAL Score“ – Metrik hinsichtlich ihrer Qualität und Quantität beurteilt. Dies geschieht wie folgt: Zum einen wird die Anzahl von „qualitativen Beobachtungen“ herangezogen, wobei die Qualität durch Faktoren wie Pünktlichkeit, Losumfang und Notierungstyp bestimmt wird. Ein gewichtetes Mittel aus Einzelnotierungen bestimmt dann die sogenannte „effektive Zahl“. Zum anderen wird die Standardabweichung der für das Wertpapier erhältlichen Marktdaten ermittelt.

Ein großer Vorteil der von Bloomberg festgelegten Methodik besteht darin, dass auch die Evaluation von Wertpapieren möglich wird, für welche Marktdaten nicht bzw. nur unregelmäßig verfügbar sind (indirekte Methodologie). Dabei werden für eine Menge von vergleichbaren Wertpapieren die zuvor beschriebenen Größen erhoben (Gewichtung dieser Größen bezüglich der Korrelation der vorhandenen Daten) und folglich über Extrapolation die

Maße für das betrachtete Wertpapier geschätzt. Sollten diese Daten nicht zielführend angewendet werden können, können auch gewisse regelbasierte Techniken zur Anwendung kommen.

Als Ergebnis liefert die Metrik einen Indexwert zwischen 1 (sehr ungenaue Preisfestlegung) und 10 (Preis mit hohem Sicherheitsniveau).

3) Relevante Einflussfaktoren aus der Wirtschaft und dem Finanzbereich

Die Liquidität von Wertpapieren ist, wie viele Faktoren im Börsenumfeld, eine Größe, welche mit diversen (finanz-)wirtschaftlichen und ökonomischen Merkmalen in einem relativ komplexen Zusammenhang steht. Aus diesem Grund gibt es verschiedene Arten von Ansätzen, mit denen versucht werden kann, Information über die Liquidität aus derem Zustand und deren Eigenschaften zu gewinnen. Typischerweise werden diese dann in einer spezifischen Form bei Banken zur Bestimmung der Liquidität von Beständen verwendet. Die Einflussfaktoren werden auch ab dem fünften Kapitel bei der Anwendung statistischer Methoden (aus dem Bereich des Machine Learning) eine große Rolle spielen, hier sollen zunächst allgemeine Ansätze beim Versuch der Modellierung von Zusammenhängen zur Gewinnung der oben erwähnten Information diskutiert werden:

Bereits bei der Beschreibung der beiden Standards aus „Basel III“ im zweiten Kapitel dieser Arbeit sind einige der wesentlichen Merkmale in Erscheinung getreten. Ein Beispiel hierfür ist die Notation des Wertpapiers an einer oder mehreren Börsen mit einem internationalen Stellenwert. Insbesondere garantiert (bei der Betrachtung von Aktien) in diesem Kontext die Zugehörigkeit zu einem der größten Indizes der Welt eine konstant hohe Liquidität, da in Indizes wie dem „Dow Jones Industrial Average“ oder dem „DAX 30“ per Definition die meistgehandelten Aktien des Landes, aus dem der Index stammt, zusammengefasst sind. Ein alternativer, relativ ähnlicher Indikator kann gewonnen werden, wenn anstelle der Überprüfung, ob Wertpapiere eines Unternehmens an einem der größten Finanzplätze der Welt gehandelt werden, betrachtet wird, an wie vielen Börsen diese notieren, und welche internationale Bedeutung diese im Mittel aufweisen. Da viele Unternehmen, deren Wertpapiere typischerweise liquid sind, nicht nur im Ursprungsland, sondern auch in jenen Ländern auf der Börse präsent sind, auf die sich ihre Interessenssphären besonders stark ausgeweitet haben (Beispiele hierfür sind Gazprom oder Rosneft, die als russische Unternehmen auch an der Londoner Börse gehandelt werden) bzw. wenn man sie auch auf Börsen mit speziellem Hintergrund findet (Börsen für Optionsscheine, Zertifikate, Anleihen, Futures), würde sich für viele Konzerne in Sachen Liquidität eine ähnliche Situation bei letzterer Betrachtung wie bei der Erfüllung des ersten Kriteriums ergeben.

Daneben kam auch die Rolle des Ratings zur Geltung, das im Falle von „Basel III“ zwar nur in Verbindung mit Anleihen bzw. Schuldverschreibungen gebracht wurde, jedoch auch bei anderen Wertpapiergattungen nicht außer Acht gelassen werden sollte. Gerade das Rating von Unternehmen kann große Effekte bei der Entwicklung der Attraktivität im internationalen Umfeld bewirken, und somit die Handelsumfänge im mittelfristigen Zeitraum stark beeinflussen.

Neben diesen beiden eher allgemeinen Merkmalen gibt es nun weitere Einflussfaktoren, über welche die Möglichkeit besteht, die Beziehung zwischen der Liquidität und diverser wirtschaftlicher Variablen in einem formaleren Kontext zu behandeln. Folgende Indikatoren

werden in der Arbeit „Does monetary policy determine stock market liquidity? New evidence from the euro zone“ [Fernandez-Amador et al. 2013] präsentiert (die Wertpapiere sind im Folgenden Teil eines Portfolios, welches sich aus einer bestimmten Menge an Wertpapieren zusammensetzt):

-) Der relative „Bid-Ask-Spread“ (deutsch: „Geld-Brief-Spanne“) gibt an, in wie fern sich jene Preise, zu welchen Käufer Wertpapiere eines Unternehmens erwerben möchten (Ask-Price), von den Preisen relativ unterscheiden, welche Verkäufer für den Verkauf derselben Wertpapiere in Erwägung ziehen (Bid-Price),:

$$S\_REL_{iyd} = (PA_{yid}\text{-}PB_{yid}) / ((PA_{yid} + PB_{yid})/2) \quad (3)$$

S_REL_{iyd} steht dabei für den relativen Bid-Ask-Spread des i-ten Wertpapieres am Tag d des Jahres y, PA_{yid} für den Ask-Price und PB_{yid} für den Bid-Price (jeweils mit äquivalenten Indizes wie beim ersten Ausdruck).
Ist diese Spanne sehr hoch, so kann davon ausgegangen werden, dass das Zustandekommen von Handelsvorgängen durch unterschiedliche Vorstellungen der Interessensparteien erschwert wird, worunter die Liquidität des Wertpapiers leiden könnte.

-) Die Monatssumme der gehandelten Wertpapiere des Unternehmens (Summe über alle Handelstage) bzw. der mittlere gehandelte Wert dieser Wertpapiere an den Handelstagen des betrachteten Monats (die Zahl täglich gehandelter Mengen des i-ten Wertpapiers im Jahr y, Monat m und Tag d wird als VO_{iymd} bezeichnet). Ein hoher mittlerer Wert spricht für hohe Liquidität.

-) Die „Turnover-Rate“ („Umsatzrate“) setzt sich aus der eben betrachteten Monatssumme gehandelter Wertpapiere (auf Tagesbasis), welche durch die Anzahl ausstehender Papiere dividiert wird („Number of outstanding shares“, bei diesen noch kein abgeschlossener Handel), zusammen:

$$TO_{iym} = \frac{\sum_{d=1}^{D_{iym}} VO_{iymd}}{NOSH_{iym}} \quad (4)$$

TO_{iym} entspricht hier der Turnover-Rate für das i-te Wertpapier im Monat m des Jahres y. Der Nenner $NOSH_{iym}$ ist die im letzten Absatz erwähnte Anzahl ausstehender Papiere unter Betrachtung derselben Indizes wie zuvor.

-) Wird die den Zähler der letzten Formel bestimmende Zahl gehandelter Wertpapiere mit den entsprechenden Preisen (P_{iymd} – Preis des i-ten Wertpapiers an Tag d in Monat m des Jahres y) multipliziert und das resultierende Produkt anschließend logarithmisiert, gelangt man zum gehandelten Volumen („traded volume“, abgekürzte Form ist TV_{iym} unter Verwendung derselben Indizes wie zuvor) des i-ten Wertpapiers im Monat m des Jahres y in der entsprechenden Währung. Zudem kann aus der „Turnover-Rate“ der „Turnover Price Impact“

(TPI) bestimmt werden, indem man die absolute tägliche Rendite des i-ten Wertpapiers ($|R_{iyd}|$) durch diese Rate dividiert:

$$TV_{iym} = \ln(\sum_{d=1}^{D_{iym}}(VO_{iymd} * P_{iymd})) \quad (5)$$

$$TPI_{iyd} = \frac{|R_{iyd}|}{TO_{iyd}} \quad (6)$$

-) Das gehandelte Volumen eines Wertpapiers in der entsprechenden Währung spielt wiederum eine Rolle im „Illiquiditätsquotienten“ (abgekürzte Form: ILLIQ), welchen der Finanzwissenschaftler Yakov Amihud im Jahr 2002 in der Arbeit „Illiquidity and Stock Returns: Cross-Section and Time-Series Effects“ [Amihud 2002] präsentierte: Dabei wird der Grad der Illiquidität durch das Verhältnis zwischen der absoluten täglichen Rendite des i-ten Wertpapiers (entspricht dem Betrag der Tagesveränderung des Wertes des Papiers) und dem zuvor betrachteten gehandelten Volumen bestimmt:

$$ILLIQ_{iyd} = \frac{|R_{iyd}|}{TV_{iyd}} \quad (7)$$

Amihuds Indikator besagt also, dass eine hohe absolute Rendite (gleichbedeutend mit stärkeren Kursschwankungen) und ein geringes monatlich gehandeltes Volumen mit einer stärkeren Tendenz zur Illiquidität in Zusammenhang stehen.

-) Ein weiterer, ebenfalls auf einer wissenschaftlichen Arbeit basierender Indikator ist der sogenannte „Roll Impact“. Im Gegensatz zu den meisten vorangegangenen Variablen weist die methodische Grundlage dieser Variable auch einen stärkeren Bezug zur Statistik auf. Die Basis bildet die Arbeit „A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market“ [Roll 1984], in welchem über die Modellierung des Preises eines Wertpapiers als Random Walk, wo nicht nur die Störterme unabhängig und identisch verteilt sind (Mittelwert 0 und Standardabweichung σ), sondern dies auch für die Wahrscheinlichkeit eines Kauf- und Verkaufsauftrages (jeweils 0,5) gilt, folgende Beziehungen zwischen Preisveränderungen (ΔP) und dem Bid-Ask Spread (S) gewonnen werden:

$$\mathrm{Cov}\big(\Delta P(t), \Delta P(t-1)\big) = -\frac{1}{4}S^2 \quad \rightarrow \quad \mathrm{S} = 2*\sqrt{-Cov(\Delta P(t), \Delta P(t-1)} \quad (8)$$

Daraus gewinnt man den „Roll-Schätzer“ wie folgt:

$$\mathrm{ROLL}_{iyd} = \begin{cases} 2 * \sqrt{-Cov(\Delta P(t), \Delta P(t-1)}, & \text{falls } \mathrm{Cov}\big(\Delta P(t), \Delta P(t-1)\big) < 0 \\ 0 & sonst \end{cases} \quad (9)$$

Zum „Roll-Impact“ (R_IMP) gelangt man schließlich über die Division des Roll-Schätzers durch das gehandelte Volumen am Tag d des Jahres y in der entsprechenden Währung:

$$R\_IMP_{iyd} = \frac{ROLL_{iyd}}{TV_{iyd}} \quad (10)$$

-) Der vorhin betrachtete Roll-Schätzer kann auch in einer relativen Form angegeben werden – dafür muss er durch den Preis des i-ten Wertpapiers dividiert werden.

-) Ein weiteres Maß, das zur Messung der Liquidität verwendet werden kann, ist die sogenannte „Amivest Liquidity Ratio“ (AR) [Goyenko et al. 2009]. Sie ergibt sich aus dem Quotienten des täglichen gehandelten Volumens und der absoluten täglichen Rendite. Es wird insbesondere die Sensitivität des Preises durch das gehandelte Volumen beim betrachteten Wertpapier berücksichtigt– welcher Geldwert gehandelt wird, wenn eine Preisänderung von einem Prozent vorausgesetzt wird:

$$\mathrm{AR_{iy}} = \frac{\sum_{d=1}^{D_{iy}} TV_{iyd}}{\sum_{d=1}^{D_{iy}} |R_{iyd}|} \quad (11)$$

Zur allgemeinen Bestimmung der Liquidität des i-ten Wertpapiers wird bei Fernandez-Amador et. al. nun die Berechnung eines Mittelwertes aller betrachteter relevanter (finanz-)wirtschaftlicher Parameter in Erwägung gezogen. Auf Gewichtungen wird hierbei verzichtet, alle Variablen liefern bei dieser Berechnung den gleichen Beitrag. Einige Variablen, die zuvor vorgestellt wurden, werden ab dem Kapitel 5 dieser Arbeit Teil der statistischen Analyse sein, in Form von Prädiktorvariablen wird dabei ihr Beitrag zur Erklärung von Liquiditätsgraden evaluiert (keine exakte Liquiditätsberechnung, sondern eine Klassifizierung der Liquidität in verschiedene Grade wird dabei anvisiert) – stellt sich dabei heraus, dass manche der zuvor erwähnten Faktoren in einer stärkeren Form mit der Liquidität assoziiert sind als andere, so könnte man auch die Bildung eines gewichteten Mittelwertes anstelle eines gewöhnlichen Durchschnittes präferieren.

In der zuletzt betrachteten Arbeit, vor allem aber auch im Paper „Do Macroeconomic Variables Forecast Changes in Liquidity? An Out-of-Sample Study on the Order-Driven Stock Markets in Scandinavia“ von Jonas Söderberg werden die finanzwirtschaftlichen Variablen von vorhin mit diversen makroökonomischen Größen ergänzt, deren Einfluss auf die Liquidität von Wertpapieren ein wesentlicher Untersuchungsgegenstand der beiden Werke ist. Besonders hier ist auf die bereits auf Seite 1 erwähnte Tatsache zu achten, dass der Zusammenhang zwischen den Variablen, die hier vorgestellt werden, und der Liquidität zumeist in einer endogenen Form vorliegt – Beziehungen liegen also in einer wechselseitigen Form vor, die Änderung der einen Variable kann eine Auswirkung auf den Zustand der anderen Variable haben, aber es ist auch umgekehrt möglich. Das heißt, man sollte eher von Assoziationen als von einseitigen Abhängigkeiten sprechen.

Die bedeutendsten makroökonomischen Größen sind die folgenden [Söderberg 2008]:

-) Variablen, die mit Geldpolitik in Verbindung stehen:

.) Wachstum des „breiten“ Geldes (Prozentveränderung innerhalb eines Monats bzw. eines Jahres)
.) Veränderung des zwischenbanklichen Zinssatzes (monatliche Veränderung)
.) monatliche Differenz des Leitzinssatzes

In der Arbeit von J. Söderberg wird dabei festgehalten, dass Senkungen des Leitzinssatzes, welche von Zentralbanken in den letzten Jahren in einer häufigeren Form durchgeführt wurden (insbesondere in der Phase der Bankenkrise), einen stimulierenden Effekt auf die Liquidität haben, weil es beispielsweise für Banken dadurch weniger kostspielig wird, Geldmengen (für Finanzierungszwecke) bei der Zentralbank zu beschaffen. Ähnlich ist der Effekt bei einem niedrigen zwischenbanklichen Zinssatz, wenn sich Banken untereinander Geld zu günstigen Konditionen leihen können.

-) Variablen, die von wirtschaftlichen Zyklen beeinflusst werden:

.) monatliche Inflationsrate
.) monatliche Industrieproduktion

Es liegt nahe, dass Handelsvorgänge an den Börsen eine sehr starke Abhängigkeit von allgemeinen Stimmungen haben, die an den Weltmärkten vorherrschen: liegt eine trübere Stimmung aufgrund (drohender) schlechter marktwirtschaftlicher Zahlen vor oder sind ungünstige Vorzeichen zu erkennen, so neigen viele Händler zur Vorsicht und sind in vielen Situationen nicht so leicht als Handelspartner zu gewinnen. Auch in umgekehrter Sicht sind laut Söderberg Zusammenhänge erwähnenswert: Eine höhere Produktivität kann Marktteilnehmer dazu neigen lassen, riskantere Investitionen zu tätigen, um (vorübergehende) Chancen auf höhere Gewinne zu wahren. Dies kann klarerweise die Liquidität von Wertpapieren, die nicht von Unternehmen von großer internationaler Bedeutung stammen, in einem stärkeren Ausmaß beeinträchtigen.

Die Verwendung der monatlichen Inflationsrate und der monatlichen Industrieproduktion in diesem Kontext liegt in der starken Korrelation zwischen der Entwicklung dieser beiden Indikatoren und dem allgemeinen Zustand der Weltwirtschaft: Eine sehr hohe oder stark schwankende Inflationsrate und eine zur Deflation neigende Volkswirtschaft haben in der Regel genauso eine Verbindung zu einer in einer Depression liegenden oder dazu neigenden Konjunktur wie schwache Zahlen der monatlichen Industrieproduktion. Im weiteren Sinne bedeutet dies also, dass über das Bindeglied des aktuellen Konjunkturstandes in der wirtschaftlichen Welt ein Zusammenhang zwischen Größen wie Inflationsrate bzw. Industrieproduktion und der Liquidität, der einfachen Handelbarkeit von Wertpapieren auf dem Weltmarkt, besteht.

-) Variablen, die mit Investorflüssen zusammenhängen:

.) monatlicher Nettofluss von Investmentfonds

.) monatlicher Nettofluss von fremden / ausländischen Investoren
.) logarithmierte Differenz des real – effektiven Wechselkurses

Hintergrund von diesen Faktoren und ihrem Einfluss auf die Liquidität von Wertpapieren ist ein belastender Effekt auf Bestände von Händlern mit wesentlichem Einfluss auf den Handelsmärkten, wenn große Geldmengen von Fonds abfließen. Dabei soll der Grad des Informationsvorsprunges gegenüber kleineren Marktteilnehmern eine Rolle spielen, welcher in diese Handelsvorgänge einwirkt – je größer dieser ist, desto schlechtere Auswirkungen kann dies auf Faktoren wie die Liquidität haben.

Wie zu erkennen ist, kann die Liquidität mit diversen Faktoren aus unterschiedlichen Bereichen der Wirtschaft, insbesondere dem Finanzbereich und der Makroökonomie, in Verbindung gesetzt werden. Eines der wesentlichen Ziele, welches ab dem vierten Kapitel dieser Arbeit verfolgt wird, ist die Bearbeitung der Fragestellung, ob und wie diese Beziehung zwischen der Liquidität und den Einflussfaktoren in einer konkreteren Form dargestellt werden kann: dabei können unter anderem die Fragen betrachtet werden, ob manche Faktoren eine stärkere Verbindung zur Zielvariable aufweisen als andere Faktoren (Auffinden der wesentlichsten Einflüsse - Möglichkeit des Aufbaus einer hierarchischen Struktur der Einflussfaktoren, gereiht nach ihrer Bedeutung für die Erklärung von Variabilität) beziehungsweise ob es eine Teilmenge der zahlreichen in diesem Kapitel behandelten Faktoren gibt, die bereits eine annähernd optimale Erklärung ermöglicht (dies würde dann bedeuten, dass hohe Korrelationen zwischen den im Modell verbliebenen und den eliminierten Variablen bestehen und durch Modellselektion das Problem der Multikollinearität vermieden werden kann), oder auch ob die endogene Beziehungsstruktur zwischen der Liquidität und den wirtschaftlichen Faktoren, welche bereits am Beginn der Arbeit erwähnt wurde, bei allen Variablen erkennbar ist, oder ob auch einseitige Abhängigkeiten zum Vorschein kommen (wo die Liquidität die Variable beeinflusst, nicht aber umgekehrt, oder wo die Variable nur auf die Liquidität wirkt).

Dies wird mithilfe des Einsatzes von Verfahren des Machine Learnings unter Verwendung der zuletzt vorgestellten Variablen in Angriff genommen, dabei steht ein Datensatz, welcher in Kapitel 4 vorgestellt wird, im Mittelpunkt.

4) Beschreibung der für die Analyse verwendeten Daten

Nachdem in den letzten beiden Kapiteln die Bedeutung der Liquidität im Bankenbereich sowie diverse Größen präsentiert worden sind, die eine Rolle bei der Erklärung der Liquidität und ihrer Entwicklung spielen können, wird in den folgenden Abschnitten dieser Arbeit das Augenmerk auf eine statistische Analyse gelegt, bei der mithilfe von Methoden des maschinellen Lernens Aussagen über den Zusammenhang der Wertpapierliquidität mit jenen Größen getroffen werden sollen, welche die letzten beiden Kapitel bestimmt haben. Damit es zu einer geeigneten Durchführung dieser Analyse kommen kann, ist es wichtig, die Auswahl der verwendeten Daten und die Zusammenstellung des Datensatzes in einer sorgfältigen Form zu bewerkstelligen. Der in weiterer Folge im Fokus stehende Datensatz besitzt nun folgenden Aufbau:

Betrachtet werden 902 Aktien, welche aus 20 Ländern stammen, die aufgrund ihrer Größe und ihres Wohlstandes eine unterschiedliche Rolle bzw. Bedeutung in der weltweiten Wirtschaft einnehmen. Neun der zwanzig Länder können im Bereich der hochentwickelten Länder eingeordnet werden (Österreich, Deutschland, Großbritannien, Norwegen, Spanien, Italien, USA, Japan, Irland), weitere neun sind mäßig entwickelte Länder, deren wirtschaftlicher Entwicklungs- und Wachstumsprozess jedoch starke Züge aufweist und deren Abstand zu führenden Nationen sich bereits reduziert hat (Tschechische Republik, Ungarn, Bulgarien, Türkei, Russland, Brasilien, Venezuela, China, Südafrika). Zur dritten Gruppe gehören zwei weitere Länder, deren Entwicklungsstufe noch einen Grad unter jener der zweiten Gruppe liegt (Indonesien, Ägypten).

Einer der wesentlichen Faktoren, der zur Wahl dieser Länder beigetragen hat, ist das Ziel der Ermittlung des Einflusses des Entwicklungsgrades eines Landes auf die Liquidität von Wertpapieren, deren Unternehmen aus dem jeweiligen Land stammen. Doch ein weiterer Punkt von hoher Bedeutung ist hierbei, dass neben einer bereits geschehenen Etablierung im Kreis der hochentwickelten Länder Faktoren wie der politische Einfluss des Landes sowie das aktuell vorliegende Wirtschaftswachstum eine genauso große Rolle bei der Auswahl von Ländern in diesem Kontext spielen sollten. Länder wie China, Russland, Südafrika oder Brasilien können aufgrund diverser wirtschaftlicher Indikatoren (beispielsweise das Bruttoinlandsprodukt pro Kopf) nicht zu den bestentwickeltsten Ländern gerechnet werden, der Einfluss dieser Länder bzw. ihrer Unternehmen (beispielsweise aufgrund ihrer Zugehörigkeit zu den sogenannten „BRICS“ – Staaten) ist aber insbesondere durch diese beiden letzten Faktoren genauso groß wie der der Länder, die zur ersten Gruppe gehören.

Der dritte motivierende Aspekt, welcher die Wahl der Länder mitbestimmt hat, ist die höchst unterschiedliche Auswirkung, welche die Wirtschafts- und Finanzkrise in den jeweiligen Staaten hatte. Während einige der 20 Länder die entstandenen Folgen in einem eher geringen Ausmaß gespürt haben (wie China), führte sie in manchen europäischen Staaten, zu denen auch Irland, Spanien und Italien zählen, zu kritischen Szenarien, durch welche bei den beiden ersten Staaten die Aufnahme in einen internationalen Rettungsschirm notwendig wurde. Die

Auswirkungen sind dabei noch heute an den Finanzplätzen dieser Länder in einem beträchtlichen Ausmaß sichtbar.

In einzelnen Ländern, wie Ägypten und Russland, kann auch untersucht werden, ob politische Ereignisse außerordentlichen Ausmaßes, wie Unruhen oder Kriegszustände, in welchen diese Länder verwickelt sind, auf die Handelbarkeit diverser Papiere auf dem Finanzmarkt einwirken.

Neben der Auswahl geeigneter Länder für die Analyse spielt die Auswahl passender börsennotierter Unternehmen, deren Zahl pro Land möglichst gleich groß sein soll (50 für jedes Land), eine entscheidende Rolle. Die Auswahl richtet sich, wie bei den Ländern, auch hier nach diversen Kriterien: Primär geschieht diese nun nach der durchschnittlichen Anzahl gehandelter täglicher Wertpapiere pro Tag (betrachtet über mehrere Monate). Dieser Indikator gibt wesentliche Auskunft über die Bedeutung des jeweiligen Papieres im internationalen Handel. Es scheint klar, dass eine Größe wie die Liquidität davon maßgeblich bestimmt wird – schließlich wird über diese Größe auch die „Illiquidity Ratio“ von Amihud, wie sie in Kapitel 3 beschrieben wurde, berechnet. Aus diesem Grund wird in der Analyse diese Zahl nicht als Prädiktorvariable verwendet, sondern sie wird als Teil der „Illiquidity Ratio“ im Kontext des betrachteten „überwachten Lernens“ zur Bestimmung der Liquiditätsklassen des Trainingssets, also bei der Klassifizierung der 902 Aktien in Form einer abhängigen Variable, eingesetzt. Die wesentliche Aussage dieser abhängigen Variable ist dabei, in wie fern das täglich gehandelte Volumen sich auf den Wert der Aktie niederschlägt – je größer der betragsmäßige Return (im Zähler des Quotienten) im Verhältnis zum Volumen ist, desto stärker beeinflussen einzelne Handelsvorgänge den Preis des Wertpapieres, was in der Regel für eine geringere Liquidität spricht.

Aufgrund der Tatsache, dass das sich im Nenner der Illiquidity Ratio befindliche gehandelte Volumen deutlich größere Werte als der betragsmäßige Return im Zähler annimmt, und deshalb die abhängige Variable in der herkömmlichen Form eine etwas unanschauliche Größe bildet, macht es Sinn, sie über die Festlegung von Intervallen auf eine ordinale Größe zu bringen. Deshalb werden folgende Werte von nun an als Intervallgrenzen festgelegt:

Intervall	Liquiditätsgrad	Bezeichnung
ILL-Ratio < 1*10^-6	1	hochliquid
1*10^-6 < ILL-Ratio < 5*10^-5	2	liquid
5*10^-5 < ILL-Ratio < 0,001	3	mäßig liquid
0,001 < ILL-Ratio < 0,05	4	gering liquid
0,05 < ILL-Ratio	5	illiquid

Tabelle 4.1.: Kategorisierung der Liquidität anhand der Amihud Illiquidity Ratio

Wie viele Aktien pro Tag im Schnitt von einem börsennotierten Unternehmen gehandelt werden, kann zumeist damit in Verbindung gebracht werden, ob das Papier einem größeren

oder kleineren Aktienindex zugehörig ist: Typischerweise ist die Zahl der gehandelten Papiere in einem gewissen Zeitraum ausschlaggebend dafür, schließlich gehören den Indizes die bedeutendsten Aktien des jeweiligen Landes an.

Deshalb soll bei der Auswahl der etwa 50 Wertpapiere pro Land, damit sowohl Unternehmen mit stärkerem als auch jene mit schwächerem Handelsvolumen in die Analyse einfließen, die folgende Aufteilung anvisiert werden: 15 Wertpapiere werden dem größten Aktienindex des Landes angehören, 20 Aktien einem weiteren Aktienindex des Landes (falls es diesen im jeweiligen Land gibt), die restlichen 15 Aktien sollen aufgrund ihrer geringeren Bedeutung keinem (bedeutenden) Aktienindex angehören. Dies ist selbstverständlich nur in jenen Ländern möglich, auf deren Finanzplätzen eine mehrstufige Aufteilung der Wertpapiere nach ihrem Marktvolumen bzw. anderen Kriterien gehandhabt wird (dies scheint beispielsweise in Irland, Venezuela oder der tschechischen Republik nicht der Fall zu sein).

Die durch dieses Auswahlschema berücksichtigten Wertpapiere sollen, damit der resultierende Datensatz eine für die Analyse konforme Variabilität aufweist, einige weitere Bedingungen erfüllen: Ähnlich wie bei der Länderauswahl macht es auch hier Sinn, Wertpapiere zu nehmen, wo das Rating des dazugehörenden Unternehmens von Papier zu Papier unterschiedlich ist (hierbei soll eine Mischung aus Kreditwürdigkeit (dabei kommen Ratings der Agentur „Moody's" zum Einsatz) und Analystenratings herangezogen werden). Eine fix festgelegte Aufteilung wird bei der Auswahl von Wertpapieren nach Ratingkategorie zwar nicht erwogen, jedoch soll eine gewisse Menge von Unternehmen aller möglichen Ratingkategorien (A-C, je nach Verfügbarkeit auch D) in die Analyse einbezogen werden.

Eine durchaus bedeutende Fragestellung bei der Planung des Datensatzes ist, in welcher Art und Weise makroökonomische Daten in die Analyse inkludiert werden sollen. Faktoren aus diesem Bereich stehen mit vielen unternehmensspezifischen Merkmalen in Verbindung und können zusätzlich zu finanzwirtschaftlichen Merkmalen einen Beitrag zur Erklärung der abhängigen Variable leisten. Klar ist, dass diese Daten sich über die Zeit ändern, üblicherweise unabhängig von den auf internationalen Finanzplätzen gehandelten Wertpapieren. Jedoch kann umgekehrt die Liquidität, also der Wertpapierhandel, von makroökonomischen Faktoren durchaus in einer stärkeren Form beeinflusst werden. Am naheliegensten ist die Strategie, die finanzwirtschaftlichen Variablen bei jedem Wertpapier zu unterschiedlichen Zeitpunkten zu beobachten, damit eine ausreichende Variabilität bei allen Variablen gewährleistet ist. Dafür muss eine Zeitspanne gewählt werden, in dessen Rahmen die (zufällige) Wahl von Zeitpunkten liegen kann. Optimal wäre die Betrachtungsmöglichkeit verschiedener Positionen im mittel- oder langfristigen Wirtschaftszyklus (sodass sowohl Boomphasen als auch Rezessionsphasen und Krisensituationen miteinfließen), jedoch darf ein derartiger Zeitraum auch nicht zu sehr in die Länge gezogen werden, da Aussagen über die Liquidität im Hier und Jetzt getroffen werden sollen – nachdem in langfristiger Hinsicht Veränderungen von Zusammenhängen zwischen der Liquidität und diversen Prädiktorvariablen durchaus möglich sind, sollte er nur eine Länge haben, bei welcher eine signifikante Änderung dieses Zusammenhanges als eher unwahrscheinlich angesehen werden kann.

Aus diesen Gründen wird der Zeitraum für die Durchführung der Analyse auf 10 Jahre festgesetzt – die Daten der jeweiligen Unternehmen stammen somit von einem zufällig gewählten Zeitpunkt zwischen 2004 und 2013 (üblicherweise wird ein Monat gewählt). Dieser Zeitraum sollte sowohl die Variabilität von finanzwirtschaftlichen Variablen, als auch von makroökonomischen Variablen in einer passenden Form garantieren.

In weiterer Folge werden nun jene Prädiktorvariablen im Detail in jener Form präsentiert, wie sie zur Durchführung der Analyse im Datensatz betrachtet werden:

-) Die Geld-Brief-Spanne (engl.: Bid-Ask-Ratio) wird beim Datensatz als der Mittelwert des Quotienten zwischen Geldpreis (Bid-Wert) und Briefpreis (Ask-Wert) innerhalb eines Monates (30 Tage) betrachtet

-) Die Volatilität der Preisentwicklung des Wertpapiers – diese soll in Form der Standardabweichung des täglichen Returns innerhalb eines Monates (30 Tage) betrachtet werden (in %). Der Return eines Wertpapieres ist in diesem Kontext definiert als der Quotient zwischen dem letzten Preis des Handelstages t und dem letzten Preis des Handelstages t-1

-) Die Variable des innerhalb eines Monats gehandelten Volumens wird, wie in diesem Kapitel bereits festgelegt wurde, zur Klassifizierung der Trainingsdaten aufgrund ihrer Aussagekraft verwendet, jedoch kann die Veränderung derartiger monatlicher Volumina über einen gewissen Zeitraum hinweg als Variable verwendet werden – hier wird diese definiert als relative Veränderung des mittleren täglichen Handelsvolumens im Monat eines Jahres zum Handelsvolumen im selben Monat des vorangegangenen Jahres (in %)

-) Als nicht-numerische Variable wird neben dem bereits erwähnten Rating des Wertpapieres (kategorielle Prädiktorvariable) und dem Land auch der Sektor des betrachteten Unternehmens miteinbezogen

Dabei kommen folgende Sektoren infrage: Basic Materials, Communication Services, Consumer Cyclical, Consumer Defensive, Energy, Financial Services, Healthcare, Industrials, Real Estate, Technology, Utilities

Unter „Consumer Cyclical“ versteht man dabei Konsumgüter mit einer engen Beziehung zu wirtschaftlichen Zyklen und Bedingungen (dazu gehören unter anderem die Automobil- und Unterhaltungsindustrie). Hingegen gehören zur Kategorie „Consumer Defensive“ nicht-zyklische Industriezweige wie die Nahrungsmittel- oder Verpackungsproduktion [URL: http://www.morningstar.com/InvGlossary/Default.aspx?letter=C [Zugriff am 15.9.2014]].

-) Weitere numerische Größen aus dem finanzwirtschaftlichen Bereich, die zur Analyse herangezogen werden, können wie folgt beschrieben werden:

.) der mittlere Monatsumsatz (Turnover), der durch den Wertpapierhandel generiert wird – in US-Dollar

.) die Anzahl der Börsen, an denen ein Handel mit dem betrachteten Wertpapier möglich ist (als Informationsbasis dienen hierfür Online-Handelsplattformen, über welche Handelsvorgänge von internationalen Wertpapieren getätigt werden können)

-) Die aus dem makroökonomischen Umfeld ausgewählten Größen können weiters wie folgt beschrieben werden (betrachtet wird jeweils die monatliche Veränderung in Prozent):

.) Consumer Price Index (Verbraucherpreisindex – Preisveränderung von Waren und Dienstleistungen in Bezug auf den Privatbedarf)

.) Industrieproduktion

.) Veränderung des Money Supply (verfügbare Geldmenge) auf Niveau M3

Zu dieser Form der verfügbaren Geldmenge zählen Münzen und Banknoten, die sich außerhalb des Bankensystems im Umlauf befinden, Sichteinlagen der Nichtbanken sowie der gesamte Bargeldumlauf, der von Kreditinstituten verwaltete Zentralbankgeldbestand, Einlagen mit maximaler vereinbarter Laufzeit von zwei Jahren und Einlagen mit maximaler gesetzlicher Kündigungsfrist von drei Monaten sowie Geldmarktfondsanteile, Repoverbindlichkeiten, Geldmarktpapiere und Bankschuldverschreibungen mit einer maximal zweijährigen Laufzeit [URL: http://www.ecb.europa.eu/stats/money/aggregates/aggr/html/hist.en.html [Zugriff am 26.11.2014]].

.) Veränderung des (zwischenbanklichen) Zinssatzes in der kurzen Frist

.) Veränderung des langfristigen Zinssatzes (10-Jahres-Staatsanleihen)

.) Veränderung des real-effektiven Wechselkurses

Vergleich der Veränderung von Kosten und Preisen in verschiedenen Ländern in deflationierter Form (inflationsbereinigt bezüglich des Landes, welches im Fokus steht)

.) Veränderung des Economic Sentiment Indicators

Dieser Indikator wird seit einigen Jahren von der europäischen Statistikbehörde Eurostat veröffentlicht. Dabei handelt es sich um einen zusammengesetzten Index, der sich aus dem Industrial Confidence Indicator, Services Confidence Indicator, Consumer Confidence Indicator, Construction Confidence Indicator sowie dem Retail Trade Indicator

berechnet. Allgemein werden derartige „Vertrauensindikatoren“ als Mittelwerte saisonbereinigter Bilanzen berechnet, die sich aus Fragestellungen ergeben, welche sich auf Beurteilungen der Stimmungslage in den jeweiligen Bereichen beziehen und in stichprobenartiger Form regelmäßig an Unternehmen, Konsumenten etc. gerichtet werden [URL: http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&language=en&pcode=teibs010 [Zugriff am 26.11.2014]].

Die Quellen für die makroökonomischen Variablen stellen die über „Thomson Reuters Datastream“ verfügbaren Daten des internationalen Währungsfonds (World Economic Outlook) sowie der OECD (Main Economic Indicators) dar, für die finanzwirtschaftlichen Variablen dient großteils der Finanzdienstleister Bloomberg als Quelle, in einem geringen Ausmaß stammen die Daten auch von Finanzportalen aus dem Internet (finanzen.net, onvista.de).

5) Modellierung des Liquiditätsrisikos durch dessen Betrachtung als Klassifikationsproblem – Anwendung verschiedener Methoden des Machine Learnings

5.1) Grundlegende Aspekte von Klassifikationsproblemen im Machine Learnings

Die bedeutendsten Methoden des „Machine Learning“ wurden in den letzten Jahrzehnten des 20. Jahrhunderts entwickelt (Leo Breiman, Tom Mitchell), seitdem erlebt dieser Teilbereich der Computerwissenschaften, genauer gesagt der künstlichen Intelligenz, einen enormen Zuwachs an Bedeutung. Ein Grund dafür ist die stetige Zunahme der Leistungsfähigkeit von Rechnern und Softwarepaketen, wodurch selbst die Durchführung von komplizierten und umfangreichen Modellberechnungen bzw. Algorithmen kein großes Problem mehr darstellt. Eine andere Ursache für diesen Trend liegt darin, dass die vernetzte Gesellschaft von heute enorme Mengen an Daten produziert, welche für Felder wie die Wirtschaft oder Wissenschaft von großer Bedeutung sind.

Jedoch haben die Daten in der rohen Gestalt, wie man sie in den unzähligen Datensätzen finden kann, keinen direkten Wert für Forschungszwecke. Was von Bedeutung ist, ist die Information, die sie beinhalten:

„[…] Wenn die Daten als aufgezeichnete Fakten charakterisiert werden, dann ist die Information die Zusammensetzung von Mustern, welche den Daten zugrunde liegt“ [Witten & Frank 2005].

Eine wesentliche Zielsetzung bei der Anwendung von Methoden des Machine Learning besteht nun darin, bislang unbekannte, jedoch nützliche Information aus den Rohdaten zu gewinnen, um daraus praxisrelevante Schlussfolgerungen ziehen zu können. Das bedeutet, dass mithilfe der gewonnenen Information die Möglichkeit bestehen sollte, sie zu verallgemeinern und somit auch hochwertige Prognosen bei neuen, vom ursprünglichen Datensatz unabhängigen Instanzen erzielen zu können [Witten & Frank 2005].

Um diese Zielsetzung in einer brauchbaren Form umsetzen zu können, bedarf es am Beginn der Analyse einer Wahl geeigneter Methoden aus der Vielzahl an Verfahren, die das maschinelle Lernen mittlerweile umfasst. Dabei hängt es insbesondere von der Struktur und der Eigenschaft des betrachteten Datensatzes (z.B. Anzahl der Prädiktorvariablen, Größe des Datensatzes) ab, welche Methoden für die Analyse als brauchbar gesehen werden können.

In dieser Arbeit steht die Betrachtung der Analyse des Liquiditätsrisikos als Klassifikationsproblem im Vordergrund: Bei diesem Konzept geht es darum, einer diskreten (das bedeutet, einer nominal oder ordinal skalierten) abhängigen Variable, welche

üblicherweise als Klasse bezeichnet wird, anhand einer gewissen Zahl von Variablen, die als Prädiktoren dienen und deren Skalentyp unterschiedlicher Art sein kann, einen von mehreren möglichen Werten zuzuordnen. Man zählt dieses Konzept zu den sogenannten „überwachten" Lernmethoden (engl.: „supervised learning"), es kommen somit bei der Anwendung Trainingsdaten zum Einsatz, bei denen eine Klassifizierung bereits durchgeführt wurde, und aus denen nun das Lernen mithilfe von Klassifikationsverfahren (Classification Trees, Random Forests, Support Vector Machines, etc.) erfolgt [Witten & Frank 2005].

Ob die Anwendung einer solchen Methode zu einem brauchbaren Erfolg geführt hat, wird in der Praxis häufig mithilfe vom Trainingsset unabhängiger Validierungs- bzw. Testdatensätze evaluiert – für diese Daten sind zwar die Klassen der dazugehörenden Instanzen bekannt, sie haben aber keinen Einfluss auf das Ergebnis, welches das gelernte Klassifikationsmodell liefert, da sie während des Trainings nicht zur Verfügung stehen. Ist so ein Testdatensatz in genügend großer Form vorhanden, so kann, vor allem bei Nutzung mehrerer Durchläufe und der Generierung multipler Klassifikatoren, damit eine hochwertige Abschätzung der Performance von Klassifikationsverfahren erzielt werden [Alpaydin 2008].

Bezüglich Performancemaß ist es im Falle von Klassifikationsproblemen bei der Validierung eines Klassifikationsregelwerkes dann üblich, die Fehlerrate zur Beurteilung der Trefferrate (engl.: „success rate" bzw. „accuracy") heranzuziehen, also die Berechnung des relativen Anteils der richtig prognostizierten Instanzen in einem Testset an der Gesamtzahl der darin vorhandenen Instanzen [Witten & Frank 2005].

$$Trefferrate = \frac{Zahl\ der\ Erfolge}{Zahl\ der\ Erfolge + Zahl\ der\ Fehler} \tag{12}$$

Eine häufig eingesetzte Darstellungsform der Trefferrate von Klassifikationsverfahren bzw. deren Fehlerraten (1 - Trefferrate) ist die Verwendung einer sogenannten „Konfusionsmatrix", die sich für den Vergleich der Häufigkeiten korrekter und falscher Klassifizierungen eignet: In ihr finden sich in den Zeilen die wahren Klassen, in den Spalten die vom berechneten Regelwerk vorhergesagten Klassen (in manchen Fällen tritt dies in umgekehrter Form auf). Durch die Häufigkeitswerte, die sich ergeben, wenn diese wahren und vorhergesagten Klassen bei den jeweiligen Instanzen bezüglich ihrer Übereinstimmung kombiniert werden, ergeben sich folglich die Werte in den Zellen dieser Matrix [Witten & Frank 2005].

Geht es beispielsweise darum, eine Klasse mit den möglichen Werten „positiv" und „negativ" vorherzusagen, dann hat eine Konfusionsmatrix die folgende Gestalt:

	Vorhergesagte Klassen	
Wahre Klassen	Richtig positiv	Falsch Negativ
	Falsch positiv	Richtig Negativ

Tabelle 5.1: Gestalt einer typischen Konfusionsmatrix

Wird ein Datensatz in Trainings- und Testset unterteilt, so ist es von großer Wichtigkeit, dass die jeweiligen Teilmengen der Stichprobe so wie die gesamte Stichprobe repräsentativ für eine Grundgesamtheit sind, bei welcher gewonnene Regelwerke einsetzbar sein sollen. Das bedeutet, es sollte gelingen, Verzerrungen bei der Gewinnung von Regeln und den daraus resultierenden Prognosen (Bias) möglichst zu vermeiden. Eine hierfür insbesondere bei kleineren Datensätzen brauchbare Variante ist die iterative Durchführung der Erzeugung von Trainings- und Testsets, dies kann durch den Einsatz eines Verfahrens, welches sich Kreuzvalidierung nennt, bewerkstelligt werden: Dabei wird ein Datensatz in eine gewisse, vorab festgelegte Zahl an gleich großen Teilen aufgeteilt (üblich sind 10 Teile), wobei von diesen alle Teile bis auf einen die Rolle von Trainingssets übernehmen, mithilfe der darin sich befindenden Instanzen wird die Berechnung eines Modells / Regelwerkes durchgeführt, und die Qualität der damit erhältlichen Prognosen wird mithilfe des übrig gebliebenen Teils, welcher als Testset dient, überprüft. Dieses Vorgehen wird nun so oft in iterativer Form wiederholt, bis jeder Teil einmal die Rolle des Testsets übernommen hat [Alpaydin 2008].

Die Anwendung dieses Verfahrens führt in vielen Fällen zu einer brauchbaren Evaluation eines Regelwerks, insbesondere aufgrund der günstigen Situation, dass die Zugehörigkeit der Instanzen zu Trainingssets bzw. dem Testset durchgehend wechselt, und somit alle Instanzen eine gleichberechtigte Rolle für die Erstellung und für die Validierung erhalten (speziell unter Betrachtung der Tatsache, dass das Testset keine Rolle bei der Erstellung des Modells spielt).

In vielen Anwendungen von Methoden des Machine Learnings ist es üblich, bei Vorliegen genügend vieler Instanzen die Performance verschiedener Methoden anhand Fehlerwerten, welche sich im Mittel bei mehrmaliger Anwendung einer 10-fachen Kreuzvalidierung herausstellen, zu vergleichen [Grossmann 2011].

Eine ebenfalls häufig verwendete, oft noch vielversprechendere Vorgehensweise bei der Evaluation derartiger Verfahren ist die Dreiteilung des vorliegenden Datensatzes in Trainingsset, Testset und Validierungsset, wobei nun ersteres zur Erstellung eines Basis-Modells verwendet wird, und mit letzterem die Parameter in Folge optimiert werden und der Prediction Error geschätzt werden soll. Das daraus erzielte Resultat wird dann wieder mithilfe des Testsets beurteilt. Die übliche Aufteilung, welche in den meisten Anwendungen festgelegt wird, ist die Verwendung von 50% der Instanzen als Trainingsset, 25% als Validierungsset und weitere 25% als Testset [Grossmann 2011].

Neben der bereits erwähnten Trefferrate als brauchbaren Indikator für die Prognosequalität eines Modells / Regelwerkes gibt es noch weitere ähnliche Maße, welche in der Praxis zur Anwendung kommen. Beispiele, welche bei Fragen der Evaluation häufig eine Rolle spielen, sind die folgenden Maße:

- Sensitivität & Spezifität

Insbesondere in der Medizin, wenn bei Tests die Klassen „positiv“ und „negativ“ betrachtet werden, spielen die Begriffe „Sensitivität“ und „Spezifität“ von Tests eine überaus große Rolle in vielen Anwendungen. Definiert sind diese Begriffe wie folgt [Kroschel et al. 2011]:

$$Sensitivität = \frac{Korrekt\ positiv\ vorhergesagte\ Fälle}{Korrekt\ positiv\ vorhergesagte\ Fälle + Falsch\ negativ\ vorhergesagte\ Fälle} \quad (13)$$

$$Spezifität = \frac{Korrekt\ negativ\ vorhergesagte\ Fälle}{Korrekt\ negativ\ vorhergesagte\ Fälle + Falsch\ positiv\ vorhergesagte\ Fälle} \quad (14)$$

Die Sensitivität berücksichtigt also nur die Menge aller Einheiten, die in der Realität eine positive Merkmalsausprägung hat, und gibt den Anteil an, der davon korrekt erkannt wurde. Mit dem Rest, also jene Einheiten mit negativer Merkmalsausprägung, beschäftigt sich das Maß der Spezifität, wiederum wird dabei der Anteil korrekter Prognosen widergespiegelt.

- Recall & Precision (Trefferquote & Genauigkeit)

Neben den beiden zuvor beschriebenen Gütekriterien spielen in vielen Anwendungen auch die Kenngrößen „Recall“ und „Precision“ eine Rolle, wobei das erste der beiden Maße analog zur Sensitivität aufgebaut ist (statt dem Nenner in (13) wird hier im Nenner häufig die Zahl aller positiven Fälle angegeben). Das zweite Maß berücksichtigt sowohl im Zähler als auch im Nenner nur die positiv vorhergesagten Einheiten, und gibt an, welcher Anteil daraus auch in der Realität positiv ist und damit ein korrektes Vorhersageresultat erhält [Fawcett 2006]:

$$Recall = \frac{Korrekt\ positiv\ vorhergesagte\ Fälle}{Gesamtzahl\ aller\ positiven\ Fälle} \quad (15)$$

$$Precision = \frac{Korrekt\ positiv\ vorhergesagte\ Fälle}{Korrekt\ positiv\ vorhergesagte\ Fälle + Falsch\ positiv\ vorhergesagte\ Fälle} \quad (16)$$

Oft macht es Sinn, den Nutzen von Vorhersagen bei Klassifikationsproblemen und mögliche Kosten, die durch falsche Einstufungen anfallen, direkt gegenüberzustellen. Eine Möglichkeit der Veranschaulichung der Maße Sensitivität bzw. Spezifität bei binären abhängigen Variablen (die zwei Klassen aufweisen, zum Beispiel positiv und negativ) bieten sogenannte ROC-Kurven („Reciever Operating Characteristic“), wo die Rate richtig positiver Vorhersagen (Sensitivität, diese wird typischerweise auf der Y-Achse dargestellt) in Abhängigkeit von der Falsch-Positiv-Rate (1-Spezifität, typischerweise auf der X-Achse) grafisch dargestellt wird. Bei der Generierung einer derartigen Kurve geht man nach folgendem Muster vor: Zuerst werden die Wahrscheinlichkeiten berechnet, mit denen eine Instanz als positiv klassifiziert wird. Dann werden die Beobachtungen absteigend nach den resultierenden Wahrscheinlichkeiten geordnet. Bei einer endlichen Zahl an Instanzen hat die

ROC-Kurve dann, ähnlich wie die empirische Verteilungsfunktion, die Form einer monoton wachsenden Treppenfunktion: Beginnend mit der höchsten Wahrscheinlichkeit durchläuft man nun die Instanzen – ist die echte Klasse der prognostizierten Instanz positiv, so wandert die ROC-Kurve um den Wert 1/n, mit n als Anzahl der positiven Instanzen in der Stichprobe, senkrecht nach oben. Ist die echte Klasse jedoch negativ, so läuft die Kurve horizontal um 1/n weiter. [Fawcett 2006]

Eine besondere Bedeutung besitzt bei diesen Kurven die Diagonale (x=y), weil eine Kurve, die entlang dieser verläuft, einen Klassifikator widerspiegelt, bei dem die Klassen von Instanzen nach dem Zufallsprinzip vorhergesagt werden, hier wären dann Treffer genauso wahrscheinlich wie fehlerhafte Beurteilungen. Mithilfe dieser Tatsache lässt sich ein Gütekriterium für ROC-Kurven finden, die sogenannte „Area under the curve“ (AUC): Diese stellt den flächenmäßigen Anteil der quadratischen Diagrammfläche dar, der sich unterhalb der resultierenden ROC-Kurve befindet. Deshalb haben Klassifikatoren, die Klassen nach dem Zufallsprinzip auswählen, typischerweise einen AUC-Wert von 0,5, während ein perfekter Klassifikator, der alle positiven Instanzen richtig vorhersagt und auch den nicht positiven Rest korrekt erkennt, das gesamte Quadrat umrahmt, was einem AUC-Wert von 1 entspricht [Fawcett 2006].

Die Anwendbarkeit von ROC-Kurven ist nicht nur auf binäre Klassifikationsprobleme beschränkt, sie lässt sich auch auf das mehrkategorielle Umfeld erweitern. Hier bietet es sich an, jeweils den Fokus auf einzelne Klassen zu legen und dann die Kurven entweder über einen paarweisen Vergleich zweier Klassen („one versus one“) oder über eine Gegenüberstellung einer Klasse mit allen übrigen Klassen („one versus the rest“) zu generieren. Werden beispielsweise fünf Klassen bei der abhängigen Variable betrachtet, so würde der erste Ansatz, wenn alle Klassenkombinationen berücksichtigt werden sollen, zu einer Zahl von $\binom{10}{2}$ = 10 ROC-Kurven führen, im zweiten Fall sind es fünf Kurven. Hier kann folglich der AUC-Wert einzeln für jede Klasse bewertet werden, aber auch insgesamt durch Mittelung der Werte für die jeweiligen Kurven. [Hudec & Grossmann 2013].

Folgende Tabelle (8 Instanzen) bzw. Grafik zeigt ein Beispiel für eine ROC-Kurve:

Instanz	Klasse	Wahrscheinlichkeit
1	Positiv	0,95
2	Positiv	0,95
3	Negativ	0,9
4	Positiv	0,8
5	Negativ	0,75
6	Negativ	0,35
7	Positiv	0,3
8	Negativ	0,25

Tabelle 5.2.: Tabelle für Darstellung eines Beispiels für ROC-Kurve

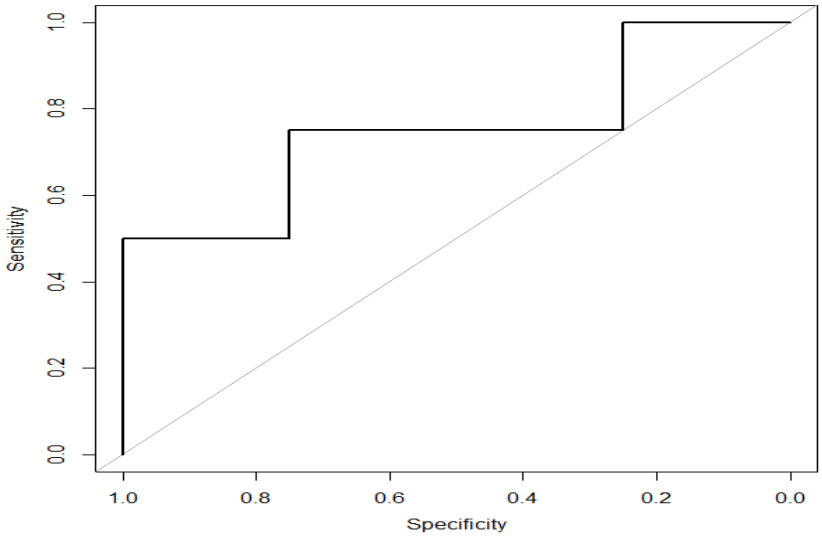


Abbildung 5.1.: Beispiel einer ROC-Kurve. Die Fläche zwischen der grauen Linie und der Kurve stellt den AUC-Bereich dar (hier: 0,75 – spricht für mäßige Qualität)

Neben den ROC-Kurven werden auch „Recall & Precision Curves“ häufig als grafische Darstellungsform von vorhin beschriebenen Maßen eingesetzt.

5.2) Vorstellung in Betracht gezogener Methoden

5.2.1) Classification Trees

Zu den am häufigsten in der Praxis verwendeten Klassifikationsverfahren des Machine Learnings zählt die Methode der Classification Trees (deutsch: Klassifikationsbäume). Unter einem Klassifikationsbaum versteht man eine in hierarchischer Form angeordnete Menge von Knoten, die durch Kanten miteinander in Verbindung stehen. Der Aufbau dieser hierarchischen Struktur geschieht durch das Lösen eines Klassifizierungsproblems, welches in Form des in Kapitel 5.1 beschriebenen „überwachten“ Lernens geschieht: Das bedeutet, auf Basis von vorliegenden Trainingsdaten wird ein Regelsatz generiert, schließlich erlaubt dieser die Zuordnung von Klassen zu allgemeinen Instanzen, deren Attributswerte bekannt sind [Petersohn 2005].

Generiert werden Classification Trees durch die rekursive Durchführung der Auswahl eines Attributs bei jedem Knoten, mit der die Aufspaltung des Trainingssets in disjunkte Teilmengen (die folglich die Töchterknoten festlegen) in einer derartigen Form gelingen soll, dass jede Teilmenge am Ende nur noch Werte beinhaltet, die zu ein- und derselben Klasse gehören [Witten & Frank 2005].

Begonnen wird dabei bei der Baumwurzel, mit dem vollständigen Trainingsdatensatz. Nun ist die Verwendung einer geeigneten Methodik für die Entwicklung des Baumes durch Aufspalten der Knoten von großer Bedeutung: Das typische Maß, welches zur Auswahl der geeignetsten Attribute verwendet wird, ist ein Maß, welches sich der Entropie bedient, in diesem Fall der sogenannte Informationsgehalt. Mit diesem lässt sich der Grad der „Reinheit" beurteilen, den ein betrachteter Knoten aufweist. Die benötigte Information, um ein Ereignis vorherzusagen, wird dabei in Form von „Bits" gemessen und berechnet sich über die bereits erwähnte Entropie. Für diese bestimmt man den negativen dualen Logarithmus der Wahrscheinlichkeit dieses Ereignisses [Petersohn 2005]:

$$Entropy\ (p_1,p_2,\ldots,p_n) = -\sum_{i=1}^{n} p_i * log_2(p_i) \qquad (17)$$

Die Attributsauswahl wird dann wie folgt bewerkstelligt: Man bestimmt relative Häufigkeiten der Klassen, welche bei den Instanzen der Töchterknoten auftreten, wenn die Aufteilung des Elternknotens mit diesem Attribut durchgeführt wird. Durch Berechnung der Entropie unter Verwendung dieser relativen Häufigkeiten als Wahrscheinlichkeiten lässt sich die Reinheit beurteilen, welche durch diese Aufteilung generiert wird. Nun wird sich jenes Attribut als am günstigsten für die Verwendung an einer gewissen Position im Baum herausstellen, durch welches der Informationszuwachs am höchsten ausfällt – dies ist die Differenz zwischen dem Informationsgehalt im Elternknoten und dem Gehalt im betrachteten Tochterknoten. Das Ziel ist es nun, durch Aufspaltungen zu Knoten im Baum zu kommen, bei denen der Informationsgehalt möglichst klein wird. Befinden sich in einem Knoten nur mehr Instanzen mit derselben Klasse, so wird die Entropie den Wert 0 annehmen, was maximale Reinheit kennzeichnet. Zu einem Endknoten wird man somit dadurch gelangen, dass sich durch eine neue Aufteilung nur noch ein geringer Informationsgewinn ergibt [Petersohn 2005].

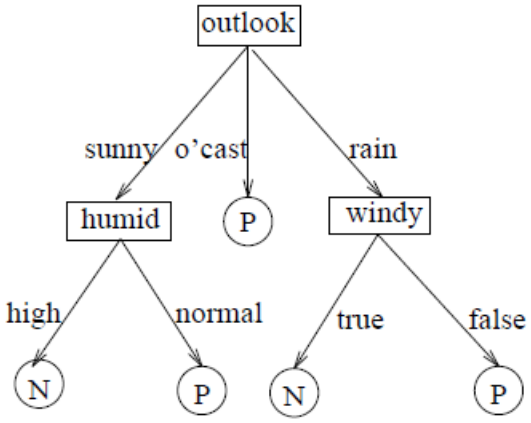


Abbildung 5.2.: typischer, einfach gehaltener Klassifikationsbaum mit zwei Klassen (Quelle: Breslow & Aha 1997)

Jedoch ist es bei der Generierung eines Klassifikationsbaumes von großer Bedeutung, dass mit diesem auch Aussagen bei Anwendung der generierten Regelsätze auf neue Daten getroffen werden können. Wird die Größe eines Baumes hinsichtlich der Knotenzahl, welche die Komplexität eines Baumes widerspiegelt, stark reduziert, so ergibt sich dadurch, wie in verschiedenen Studien gezeigt wurde [Holte 1993], in den meisten Fällen nur eine

überschaubare Reduktion der Vorhersagequalität eines Baumes. Ebenfalls wurde nachgewiesen, dass ein gut durchdachtes, an passenden Stellen durchgeführtes Zurechtschneiden sogar zu signifikanten Verbesserungen der Nutzbarkeit des Baumes führen kann [Elomaa 1994]. Der erzeugte Baum sollte daher am Ende eine eher kompakte Struktur aufweisen und nicht zu einer starken Komplexität aufgrund zu vieler Knoten und Kanten neigen.
Das bedeutet, der Baum sollte am Ende nur aus jenen Knoten bestehen, durch welche die Regelsätze auch tatsächlich aufgewertet werden. Die sogenannten „Pruning"-Verfahren liefern die Möglichkeit, einen zu komplex geratenen Baum auf eine im statistischen Kontext optimale Größe zu komprimieren.

Deren Anwendung ist zu verschiedenen Zeitpunkten möglich: Eine Variante ist die des „Pre-Prunings", dies geschieht durch das Vermeiden von Überspezialisierung mittels eines geeigneten Stoppkriteriums vor Beginn der Berechnungen. Typischerweise ist dies ein Homogenitäts-Stoppkriterium, mit dem der erwartete Performancezuwachs durch Weiterbildung des Baumes bestimmt wird und die Grenze dadurch erreicht ist, wenn dieser eine kritische Schranke unterschreitet. Eine Alternative ist das inkrementelle Anpassen der Größe des Baumes durch Anwenden von Algorithmen, die schrittweise den Baum weiterbilden und, nach Beurteilung gewisser Kriterien, wieder zurechtschneiden.
Die in der Praxis am häufigsten verwendete Variante ist die des „Post-Prunings", also das Zurechtschneiden nach Fertigstellung des Modells. Hierbei kommt es zum Entfernen von Verzweigungen in jenen Bereichen des Baumes, wo bereits die Überprüfung vieler Regeln und dadurch häufiges Splitten der Knoten notwendig war, um in diese Bereiche vorzudringen [Breslow & Aha 1997].

Auch beim Post-Pruning wird nun zwischen verschiedenen Varianten unterschieden: Die bekannteste Variante ist die des „Cost-Complexity-Prunings", die sich nach dem Begründer, Leo Breiman, wie folgt beschreiben lässt [Breiman et al. 1984]: Die Zielsetzung ist ein akzeptabler „Tradeoff" (Kompromiss) zwischen der Vermeidung von Overfitting und Optimalität der geschätzten wahren Rate, mit der Instanzen einer falschen Klasse zugeordnet werden. Die sogenannte „cost-complexity" ist für einen Baum durch die Summe der Resubstitutionsfehlerraten (Fehler, die bei Anwendung des Resultats an den Trainingsinstanzen auftreten, diese stehen für „cost" und sind bei den Formeln mit D(T) gekennzeichnet), dem Produkt aus der Summe der Endknoten (im Englischen als „leaves" – „Blätter" - bezeichnet) und einem Tradeoff-Parameter definiert:

$$D_k(T) = D(T)+k*size(T), \tag{18}$$

wobei $$D(T) = \sum_{i \in R_m} I(c_m \neq \hat{c}_m)$$

mit I=1, wenn die Bedingung in der Klammer erfüllt ist (Fehlerrate), und $\hat{c}_m$ betrachtet wird als geschätzte Klasse des Modells.

size(T) = *Summe der Knoten im vorliegenden Baum* (19)

Das Element k ermöglicht eine Abstimmung zwischen Kosten und Komplexität, welche vom Benutzer selbst bestimmbar ist. Man geht bei der Anwendung wie folgt vor: Zunächst minimiert man die „cost-complexity“ durch iteratives Abschneiden der untersten Verzweigungen, um eine Reihe immer kleinerer Bäume zu erzeugen, dann wird der hochwertigste Baum bezüglich seiner Größe (relativ zur Ausgangssituation) und einem benutzerspezifischem Beurteilungsmaß aus dieser Reihe gewählt (hier verwendet man beim „Minimal Cost-Complexity Pruning“ entweder die Vorhersagegüte des Pruningsets oder, wie in häufigeren Fällen, die kreuzvalidierte Vorhersagegüte). Schließlich resultiert der Baum mit den wenigsten Splits, dessen Fehler beim Pruningset den kleinsten Fehler unter allen Bäumen nicht um mehr als eine Standardabweichung übertrifft [Breslow & Aha 1997].

Alternativen zum „Cost-Complexity Pruning“ sind beispielsweise das „Reduced Error Pruning“, das „Minimum Error Pruning“ oder das „Pessimistic Error Pruning“.

5.2.2) K-Nearest Neighbours

Ein weiteres Standard-Klassifikationsverfahren, welches sich durch seine unkomplizierte Anwendbarkeit auszeichnet, ist das der „K-Nearest-Neighbours“. Dem Namen entsprechend zielt diese Methode darauf ab, dass man, unter der Verwendung eines geeigneten Abstandsmaßes, aus einem Trainingsset jene K Instanzen herausfiltert, welche bezüglich ihrer Attributswerte den zu klassifizierenden Instanzen am ähnlichsten sind. Schließlich gewinnt man eine Prognose für diese Instanzen durch ein Voting, welches man unter den K ausgewählten Instanzen durchführt (das bedeutet, die unter den gewählten Instanzen am häufigsten auftretende Klasse kommt zur Anwendung) [Duda et al. 2001].

Die beiden Fragen, die vor der Anwendung noch zu klären sind, sind die Wahl eines günstigen Abstandsmaßes zur Bestimmung der „Entfernung“ der Instanzen des Trainingssets von den zu klassifizierenden Objekten, und die Wahl eines passenden Wertes für den Parameter K.

Im Falle nominal oder ordinal skalierter Attribute erweist sich als Abstandsmaß die sogenannte „Hamming-Distanzfunktion“ als günstig. Diese ist definiert „als die Anzahl an unterschiedlichen Stellen, welche in zwei Blöcken unterschiedlicher Länge auftreten“. [Hamming 1950]. Formal lässt sie sich wie folgt darstellen:

$$\Delta(x, y) = |\{i \in \{1, \dots, p\} \mid x_i \neq y_i\}| \quad (20)$$

Im vorliegendem Fall betrachtet man die Anzahl unterschiedlicher Attributsausprägungen bei zwei Instanzen. Bezüglich oben erwähnter Skalierungen hat also jene bereits klassifizierte

Instanz den geringsten Abstand zu einer zu klassifizierenden Instanz, wenn sie die geringste Anzahl unterschiedlicher Attributsausprägungen hat.

Diese Distanz kann in dieser Form für metrische Attributswerte nicht analog angewendet werden. Hier erweist sich vor allem die euklidische Distanzfunktion als eine brauchbare Alternative [URL: http://de.wikipedia.org/wiki/Euklidischer_Abstand [Zugriff am 10.05.2014]]:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \tag{21}$$

Nun muss man eine Variante finden, wie bei unterschiedlich skalierten Attributen (wenn nominal, ordinal und metrisch skalierte Variablen allesamt auftreten) die Abstimmung auf ein universelles Maß erfolgen soll, sodass die Abstände aller Attribute untereinander vergleichbar werden. Nachdem bei der Modellierung der Liquidität von Wertpapieren eher Prädiktorvariablen in metrischer Form im Vordergrund stehen werden, ist es nicht günstig, eine Diskretisierung der metrischen Attribute zu favorisieren. Durchaus vielversprechend scheint eine Normierung dieser Attribute zu sein – das bedeutet, die Bildung eines Quotienten, bei dem im Zähler die Differenz zwischen Attributswert und Mittelwert aller Instanzen bezüglich dieses Attributs aufscheint, und im Nenner die Standardabweichung des Attributs, die sich über alle Instanzen im Trainingsset ergibt. Das Maß bei nominal oder ordinal skalierten Attributen könnte dann so bleiben, wie es in (20) formuliert ist [Ghosh 2006].

$$Z = \frac{x_i - \mu}{s} \text{ mit } x_i = \text{i-te Instanz, } \mu = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ und } s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu) \tag{22}$$

Eine häufig untersuchte Fragestellung bei der Anwendung der K-Nearest Neighbors-Methode ist die der optimalen Wahl des Parameters K, also der Anzahl der Nachbarn, die am Ende über die Klassifikation der Instanzen entscheidet. Zu den klassischen Methoden bei der Bestimmung dieses Parameters zählt der Vergleich der Schätzungen von Misklassifikationsraten, welche durch Kreuzvaliderung bei unterschiedlichen Werten dieses Parameters bestimmt wurden - schließlich wird am Ende derjenige Parameterwert gewählt, bei dem diese Rate den minimalen Wert geliefert hat [Lachenbruch & Mickey 1968].

A.K.Ghosh wies in seiner Arbeit „On optimum choice of k in nearest neighbour classification“ auf die Problematik dieses Verfahrens hin: „[…] diese Kreuzvalidierungs-Techniken nutzen naive empirische Anteile zur Schätzung der Misklassifizierungswahrscheinlichkeiten. Als Konsequenz bekommen wir häufig zwei oder mehr Werte von K als Minimierer der geschätzten Misklassifizierungsrate, von der man schwer den optimalen Wert aussuchen kann […]“ [Ghosh 2006]. Als Alternative schlägt er einen Ansatz aus der bayesianischen Statistik vor, dessen Komplexität den Rahmen dieser Arbeit übertrifft und daher auf eine genauere Beschreibung verzichtet wird.

5.2.3) Random Forests

Die Basis für die Methodik der „Random Forests“ stellen, wie der Name bereits vermuten lässt, Baumstrukturverfahren wie jenes von Kapitel 5.2.2 dar. Jedoch geht es hier nicht um die Gewinnung von Prognosen aus einem einzelnen Lern-Algorithmus (z.B. einem Klassifikationsbaum, aus dem ein Regelwerk zur Gänze resultiert). Dieses Lernverfahren ist das am häufigsten verwendete und bekannteste Beispiel sogenannter „Ensemble Learner“, welche in detaillierter Form in Abschnitt 5.2.6 dargestellt werden sollen. Aufgrund der hohen Relevanz in der Praxis werden aber bereits hier die grundlegenden Aspekte der „Random Forests“ erläutert:

Leo Breiman, einer der Begründer des hier erläuterten Verfahrens, definiert in seinem Werk „Random Forests“ [Breiman 2001] diese wie folgt:

„Ein Random Forest ist ein Klassifikator, bestehend aus einer Sammlung von in Baumform strukturierten Klassifikatoren $\{h(x,\theta_k), k=1\ldots\}$, wobei die $\{\theta_k\}$ i.i.d. Zufallsvektoren sind und jeder Baum ein Teilvotum für die populärste Klasse beim Input x abgibt“.

Diese Definition soll nun genauer beleuchtet werden. Wie hier erwähnt wird, kommt es bei diesem Verfahren zu einer Kombination der Ergebnisse von einer Vielzahl an gelernten Bäumen. Die Durchführung geschieht wie folgt: Man generiert zunächst aus einem ursprünglichen Trainingsset eine Vielzahl an neuen Trainingssets durch Anwendung der Bootstrap-Methode (das bedeutet, durch Ziehen mit Zurücklegen aus den vorhandenen Trainingsdaten). Dann erfolgt einzeln für jeden Baumknoten eine zufällige Auswahl einer vorab festgelegten Zahl k an Attributen, welche für den Split in Betracht gezogen werden können (anders als bei den Klassifikationsbäumen, wo die Auswahl jedes Attributes an jeder Position möglich ist). Allgemein gilt dabei, dass k kleiner als die Gesamtzahl p der Attribute ist, als passende Wahl stellt sich häufig $k = \sqrt{p}$ heraus. Die gewählten Elemente bilden dann den in der Definition beschriebenen Zufallsvektor θ_k . Nach der Attributswahl wird der Baum ohne Betrachtung einer Pruning-Methode vollständig ausgebildet. Schließlich wird ermittelt, welche Klassifizierung einer neuen Instanz bezüglich dessen Attributsausprägungen die einzelnen Bäume des Random Forests implizieren, und diejenige Klasse gewählt, die bei den meisten Bäumen resultierte [URL: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm [Zugriff am 15.5.2014]].

Diese Methode der multiplen Ausbildung von Klassifikationsmodellen und Nutzung der dadurch gewonnenen Information baut auf der „Bagging“ – Methode auf (Kurzform für „Bootstrap Aggregating“), welche ebenfalls von Leo Breiman im Jahr 1996 entwickelt wurde: Dabei kommt es auch zu einer Kombination von Prognosen, die verschiedenen Klassifikations- bzw. Regressionsmodellen entstammen, im Form einer (gewichteten) Mittelung der erhaltenen Resultate. Dies wird als „Voting“ bezeichnet und lässt sich formal wie folgt anführen:

$$y_i = \sum_{j=1}^{L} w(j) * d(ji) \quad (23)$$

Dabei steht d_{ji} für die Stimme des j-ten Lerners für die i-te Klasse und w_j für das Gewicht, das er erhält. Auch hier werden für die Erstellung der Modelle zufällig Instanzen aus dem Trainingsset mithilfe der Bootstrap-Methode gezogen, jedoch gibt es hier keine Zufallsauswahl der Attribute bei der Baumgenerierung [Breiman 2001; Alpaydin 2008].

Die Methode der „Random Forests" stellt also eine Weiterentwicklung des herkömmlichen „Bagging"-Verfahrens dar – vor allem soll sich die Qualität der Modelle bei der Anwendung auf neue Daten durch den Einsatz der Attribute in zufälliger Form (über die Bildung unabhängiger, identisch verteilter Zufallsvektoren) im Vergleich zur älteren Methode erhöhen.

Übrige Vorteile, die sich ergeben, entsprechen im Wesentlichen denen des „Baggings": Liefern Berechnungen einzelner Klassifikations- oder Regressionsbäume Ergebnisse mit einer verhältnismäßig hohen Streuung (das bedeutet, falls erneute Modellberechnungen zu Resultaten führen können, die größere Unterschiede zur ersten Auswertung aufweisen), dann kann eine Kombination von Resultaten mehrerer Modelle in Form von Mittelung bzw. einer Mehrheitsentscheidung dieses Problem erheblich einschränken und somit stellt sich auch dadurch in der Regel eine deutliche Verbesserung der Vorhersagegüte ein [Breiman 2001].

In Bezug auf die Vorhersagegüte der jeweiligen Bäume sowie der Korrelation zwischen Bäumen im Random Forest besitzt der bereits vorhin beschriebene Parameter k, der die Anzahl der relevanten, zufällig gewählten Attribute bei jedem Split darstellt, eine wesentliche Rolle in Form eines Tuning-Parameters: Eine Verkleinerung dieser Größe verursacht sowohl einen Rückgang bei der Vorhersagegüte als auch bei der Korrelation in Bezug auf die Bäume des Random Forests; wird sie angehoben, gibt es in beiden Fällen einen Anstieg. Da allerdings im Gegensatz zum positiven Effekt der Vorhersagegüte die Baumkorrelation einen negativen Effekt auf die Fehlerrate des Random Forests hat (höhere Korrelation bedeutet größere Fehlerrate, bei der Vorhersagegüte ist es umgekehrt), ist weder eine sehr kleine, noch eine sehr große Wahl von k zweckgemäß, das Optimum liegt typischerweise in einem mittleren Größenbereich [URL: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm [Zugriff am 15.5.2014]].

Zur Beurteilung des Fehlers bei der Verallgemeinerung (wenn das Modell auf neue Daten angewendet wird) des kombinierten Ensembles von Bäumen und zur Optimierung des Parameters k wird der sogenannte „Out of Bag" - Fehler betrachtet. Dieser basiert auf der Tatsache, dass ein Bootstrap-Sample wegen der Möglichkeit des mehrmaligen Ziehens ein- und derselben Instanz etwa 35% der Trainingsinstanzen gar nicht beinhaltet. Es kann bei diesen nun die Klasse (die Zielvariable) unter Verwendung jener Bäume vorhergesagt werden, wo durch das Bootstrapping die Beobachtungen nicht berücksichtigt wurden, sie also „out of bag" waren. Durch Mittelung oder Mehrheitsentscheidung über alle Vorhersagen derjenigen Bäume, für die dies zutrifft, kann man diesen „Out of Bag" Fehler bestimmen [James et al. 2013].

Neben der Varianzreduktion ist die durch das starke Gesetz der großen Zahlen gegebene fast sichere Konvergenz des Fehlers der Bäume bei der Verallgemeinerung ein bedeutender Vorteil der hier beschriebenen Methode. Dadurch ist Overfitting in der Form, wie es bei den Klassifikationsbäumen durch Pruning-Maßnahmen vermieden werden soll, von Beginn an kein Thema, welchem man entgegensteuern müsste [Breiman 2001].
Jedoch muss man in der Regel auch einen größeren Aufwand bei der Berechnung (Beanspruchung von Zeit und Rechenleistung) sowie im Vergleich zu Klassifikationsbäumen schlechter interpretierbare Ergebnisse hinnehmen.

5.2.4) Support Vector Machines

Die Methode der „Support Vector Machines“ zählt ebenso wie die der „Random Forests“ zu den eher komplexeren Verfahren des Machine Learnings, sie gehört aber auch zu den qualitativ hochwertigsten von diesen. Die theoretische Grundlage findet sich hier in der Geometrie, konkret in der Konstruktion von Ebenen (ist die Dimension größer als 2, spricht man von Hyperebenen) zur Trennung von (meist konvexen) Objekten im Raum. Im p-dimensionalen Raum kann dabei eine Hyperebene wie folgt definiert werden [James et al. 2013]:

$$ß_0 + ß_1X_1 + ß_2X_2 + \ldots + ß_pX_p = 0 \qquad (24)$$

Dabei ist $(X_1,\ldots,X_p)$ ein p-dimensionaler Vektor und $ß_0, ß_1,\ldots,ß_p$ seien reellwertige Parameter. Auf der Ebene liegen damit jene Punkte, welche die obige Gleichung erfüllen. Nun gibt es neben diesen Punkten noch jene, die eine der beiden übrigen Möglichkeiten erfüllen:

$$ß_0 + ß_1X_1 + ß_2X_2 + \ldots + ß_pX_p < 0 \quad \text{bzw.} \quad ß_0 + ß_1X_1 + ß_2X_2 + \ldots + ß_pX_p > 0 \qquad (25)$$

Das Vorzeichen bestimmt dabei, ob sich der Punkt links von der Hyperebene oder auf der gegenüberliegenden Seite befindet. Das bedeutet, dass mit diesem Setup die Möglichkeit besteht, Punkte im Raum durch (Hyper-)Ebenen zu trennen - je nachdem auf welcher Seite sich diese befinden, ist das Resultat der obigen Linearkombination positiv oder negativ (oder gleich 0 für Punkte auf der Ebene). Nun kann diese geometrische Überlegung zur Generierung einer Herangehensweise für die Lösung von Klassifikationsproblemen verwendet werden: Dabei verwendet man die Trainingsinstanzen als jene Punkte im Raum bei der vorherigen Überlegung. Die Schilderung der Anwendung lässt sich gut durchführen, wenn p Instanzen herangezogen werden, deren Klasse nur 2 mögliche Ausprägungen hat: Man kann jeder Instanz den Wert einer Indikatorfunktion y zuordnen, entweder sei $y_i = -1$, falls die i-te Beobachtung der ersten Klasse angehört, oder es sei $y_i = 1$, falls es sich um die zweite Klasse handelt [James et al. 2013].

Werden die Gleichungen (24) und (25) in diesem Kontext betrachtet, sind folgende Beziehungen für die i-te Beobachtung möglich [James et al. 2013]:

$$ß_0 + ß_1X_{i1} + ß_2X_{i2} + \ldots + ß_pX_{ip} < 0, \text{ falls } x_i = -1 \quad \text{bzw.}$$
$$ß_0 + ß_1X_{i1} + ß_2X_{i2} + \ldots + ß_pX_{ip} > 0, \text{ falls } x_i = 1 \tag{26}$$

Damit ergibt sich die Möglichkeit einer Klassifikation von Beobachtungen durch die Ermittlung des Vorzeichens, welches sich bei Bildung solcher Linearkombinationen für die jeweilige Instanz herausstellt. In diesem Kontext kann auch durch den Betrag der Linearkombination eine nützliche Information über die Eindeutigkeit der Klassifikation (durch die Entfernung von der Ebene) gewonnen werden [James et al. 2013].

Dieses Konzept stellt das Grundgerüst für die Methode der „Support Vector Machines“ dar. Es gilt als nächstes zu klären, wie man aus der Menge aller möglichen Ebenen, die die vorliegenden Klassen des Datensatzes trennen können, die geeignetsten Objekte herausfiltert: Im Falle separierbarer Daten wird die „Maximal Margin Hyperplane“ (Ebene mit der größten Spannweite) als günstigste Wahl gesehen. Diese wird wie folgt bestimmt [Burges 1998]: Man generiert, wie zuvor in diesem Abschnitt geschildert wurde, eine Ebene, welche den Raum, in dem sich die Instanzen befinden, in zwei Teile trennt, folglich bestimmt man die kürzeste Distanz der Ebene zur nächsten Instanz auf der einen sowie auf der anderen Seite (bezeichnet mit d_+ bzw. d_-) und deren Spannweite („margin“) durch Addition der Werte ($d_+ + d_-$). Nun wird nach einer separierenden Ebene mit maximaler Spannweite gesucht: Dafür ist es günstig, die Gleichung in (24) so zu formulieren, dass anstelle von < 0 und > 0 (25) die Beziehung <= -1 bzw. >= 1 in den jeweiligen Fällen gilt (sodass ein Zwischenbereich zwischen Instanzen der beiden Seiten und der Ebene vorliegt), daraus kann dann folgendes Set von Ungleichungen gewonnen werden:

$$y_i(ß_0 + ß_1X_{i1} + ß_2X_{i2} + \ldots + ß_pX_{ip}) - 1 >= 0 \; \forall \; i \tag{27}$$

Werden nun Punkte betrachtet, die am Rande des Zwischenbereichs liegen, also die Beziehung

$$ß_0 + ß_1X_{i1} + ß_2X_{i2} + \ldots + ß_pX_{ip} = 1 \text{ oder } ß_0 + ß_1X_{i1} + ß_2X_{i2} + \ldots + ß_pX_{ip} = -1 \tag{28}$$

erfüllen und bei denen der Vektor ß normal zur Ebene liegt, so ergibt sich durch Betrachtung des Abstandes vom Ursprung, der im ersten Fall $|1-ß_0|/||ß||$ (der Ausdruck im Zähler stellt die euklidische Norm von ß dar) und im zweiten Fall $|-1-ß_0|/||ß||$ beträgt, für die Spannweite der Wert $2/||ß||$. Die Maximierung der Spannweite wird in diesem Falle also durch die Minimierung der euklidischen Norm von ß erlangt.

Nach „An Introduction To Statistical Learning with Applications in R“ [James et al. 2013] kommt man zu einem Klassifikator mit maximaler Spannweite durch Maximierung des vorhin betrachteten Zwischenbereiches zwischen den der Ebene am nächsten liegenden

Beobachtungen beider Seiten unter Einhaltung der Gleichheits-Nebenbedingung $\sum_{j=1}^{p} ß(j)^2 = 1$ (euklidische Norm hier damit auf 1 festgesetzt).

Dies stellt den einfachsten Fall der Klassifizierung mithilfe separierender (Hyper-)Ebenen dar. Im Falle der Nicht-Existenz einer (Hyper-)Ebene, die die zuvor beschriebenen Konzepte erfüllt, also falls sich keine direkte Möglichkeit einer Trennung der Klassen durch geometrische Objekte ergibt, so kann mithilfe sogenannter „Slack-Variablen" ξ_i eine Erweiterung der beschriebenen Methodik angewendet werden. In diesem Fall erzielt man die Erzeugung von Ebenen, die die Separierung der Klassen „so gut wie möglich" bewerkstelligen (die aber auch Fehler erlauben): Bei diesen sogenannten „Support Vector Classifiers" spricht man von einer „Soft Margin", also einer „weichen Spannweite" [James et al. 2013].

Laut „An Introduction To Statistical Learning with Applications in R" [James et al. 2013] definiert sich das Grundprinzip des „Support Vector Classifiers" wie folgt: „Größere Robustheit bezüglich individueller Beobachtungen, und bessere Klassifikation der meisten Trainingsinstanzen. Das bedeutet, dass es durchaus sinnvoll ist, einige Trainingsinstanzen falsch zu klassifizieren, wenn dafür ein besserer Job bei der Klassifikation der übrigen Beobachtungen gemacht wird […]".

In formaler Hinsicht schreiben sich die Gleichungen von vorhin unter der Verwendung der Slack-Variablen ξ_i, die so gewählt werden müssen, dass eine Zuordnung von Instanzen zur falschen Klasse unter minimalen Kosten möglich ist, wie folgt [Burges 1998]:

$$ß_0 + ß_1X_{i1} + ß_2X_{i2} + \ldots + ß_pX_{ip} <= -1 + \xi_i \text{ , falls } x_i = -1 \quad \text{bzw.}$$

$$ß_0 + ß_1X_{i1} + ß_2X_{i2} + \ldots + ß_pX_{ip} >= 1 - \xi_i \text{ , falls } x_i = 1 \qquad (29)$$

$$\xi_i >= 0 \; \forall \; i$$

Man erkennt leicht, dass ein Fehler bei der i-ten Instanz eintritt, falls $\xi_i > 1$ gilt, zudem ist die Summe aller ξ_i als obere Schranke für die Zahl der Fehler, die beim Trainingsset begangen werden, verwendbar. Nun ist bei Betrachtung des Ziels, wieder eine maximale Spannweite unter den hier vorliegenden Konditionen zu erlangen, eine Ergänzung der euklidischen Norm von ß um das Produkt der Summe der Slack-Variablen ξ_i sowie einer Kostenvariable C notwendig, die umso höher gewählt wird, je stärker Fehler bestraft werden sollen.

Auch in „An Introduction To Statistical Learning with Applications in R" [James et al. 2013] wird der Fall der Nichtseparierbarkeit mithilfe einer (Hyper-)Ebene diskutiert, dabei wird die Spannweite in einer Nebenbedingung, welche die minimale Entfernung der Instanzen verschiedener Klassen festlegt, um den Faktor $(1- \xi_i)$ gestaucht (da auch hier nun Instanzen falsch klassifiziert werden dürfen), und die Summe der Slackvariablen wird über einen Tuning-Parameter (analog zu den Kosten von vorhin) nach oben beschränkt.

Nun fehlt noch eine grundlegende Erweiterung des Konzeptes der „Support Vector Classifier“, um zur anvisierten Methodik der „Support Vector Machines“ zu kommen: Bis jetzt wurde als Funktion, welche die Entscheidungsfindung bei der Klassifizierung bestimmt, eine lineare Funktion der Daten betrachtet. Dies trifft in den Anwendungen allerdings oft nicht zu. Eine Verallgemeinerung vorangegangener Erkenntnisse auf nichtlineare Entscheidungsgrenzen ist in diesem Fall zielführend:

Laut „An Introduction To Statistical Learning with Applications in R“ [James et al. 2013] liegt die grundlegende Idee der „Support Vector Machines“ nun darin, „[…] mithilfe von Kernfunktionen den Raum, in welchem sich die Instanzen befinden, in einer solchen Form zu erweitern, sodass auch nichtlineare Begrenzungen zwischen den Klassen untergebracht werden können […]“. Unter der folgenden Erkenntnis, dass die Bestimmung einer maximalen Spannweite bei den „Support Vector Classifiers“ alleinig anhand der Betrachtung innerer Produkte der Instanzen möglich ist, lässt sich der lineare „Support Vector Classifier“ in folgender Form darstellen:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle \tag{30}$$

Die Koeffizienten β_0 bzw. α_i , die zu den Beobachtungen gehören, ermittelt man über die Bestimmung der inneren Produkte zwischen allen vorhandenen Trainingsinstanzen. Nun können die oben bereits erwähnten (typischerweise nichtlinearen) Kernfunktionen anstelle der inneren Produkte zwischen den Trainingsinstanzen eingesetzt werden, um eine Verallgemeinerung des Konzeptes der „Support Vector Classifier“ zu erzielen. Formal lässt sich diese Verallgemeinerung wie folgt darstellen [James et al. 2013]:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i) \tag{31}$$

Typische Beispiele für Kernfunktionen sind der polynomiale Kern oder der radiale Kern, diese sind wie folgt definiert [James et al. 2013]:

$$K(x_i, x_{i'}) = (1+\sum_{j=1}^{p} x_{ij}\, x_{i'j})^{d} \text{ (polynomial)} \tag{32}$$

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2) \text{ (radial)} \tag{33}$$

Der Ausdruck der letzten Formel wird einen umso kleineren Wert annehmen, je größer die Summe der quadrierten Differenzen ausfallen wird, wobei dies gerade dann eintreten wird, wenn die Differenz zwischen x_{ij} und $x_{i'j}$ groß ist. Daraus resultiert, dass die Instanz x_i nach Formel (33) hier eine untergeordnete Bedeutung in der Bestimmung von f(x) besitzt.

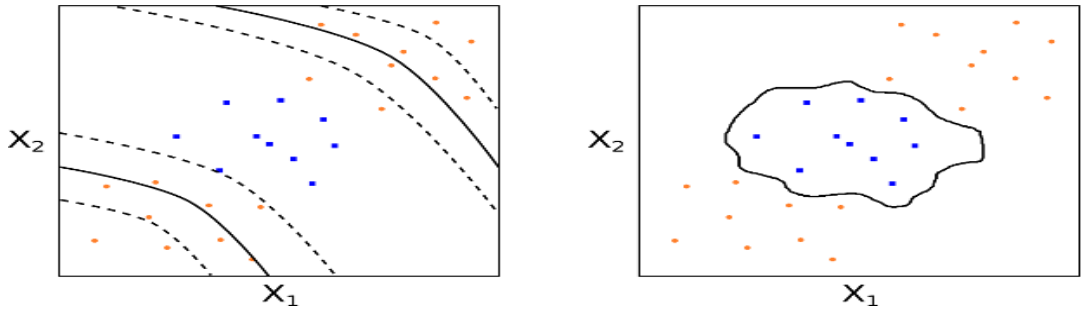


Abbildung 5.3: Darstellung der Anwendung der Support Vector Machine-Methode mittels einem polynomialen Kern 3.Ordnung (links) bzw. einem radialen Kern (rechts)
[Quelle: URL: http://www.quantstart.com/articles/Support-Vector-Machines-A-Guide-for-Beginners [Zugriff am 15.12.2014]]

5.2.5) Unsupervised Learning - Clusteranalyse

Die Clusteranalyse zählt ebenso wie die bisher besprochenen Techniken zu den bedeutendsten Verfahren des maschinellen Lernens, jedoch liegt zwischen den ersten vier Verfahren und diesem ein essentieller Unterschied vor: Während zuvor das sogenannte „überwachte" Lernen (supervised learning) im Vordergrund stand, ist dies hier das einzige berücksichtigte Verfahren, welches zu den „nicht-überwachten" Verfahren (unsupervised learning) zählt. Das bedeutet, dass hier bei der Durchführung einer Klassifikation oder Regression kein Bedarf an Trainingsinstanzen besteht, bei denen der Wert der abhängigen Variable bereits bekannt ist, wobei es an den Eigenschaften und am Aufbau der Verfahren liegt, dass solche Instanzen nicht benötigt werden. Lediglich die Werte der Prädiktorvariablen sind für die jeweiligen Instanzen bekannt [Witten & Frank 2005].

Hier findet vor allem die sogenannte Clusteranalyse in vielerlei statistischen Problemstellungen ihre Anwendung. Die Zielsetzung, welche angepeilt wird, ist weniger die Untersuchung, wie stark einzelne Prädiktoren die Variabilität der abhängigen Variable erklären können. Vielmehr geht es hier um Beziehungen zwischen Prädiktorvariablen, welche sich über eine möglichst große Anzahl von Instanzen erstrecken sollten. Im Vordergrund steht die Ermittlung von disjunkten Gruppen, deren Elemente möglichst homogen sind, und nach außen hin (andere Cluster) eine möglichst große Heterogenität aufweisen [James et al. 2013].

Eine der grundlegenden Fragestellungen, welche maßgeblich für die Anwendung dieses Analyseverfahrens ist, ist die Wahl der passenden (mehrdimensionalen) Distanzmetrik, über welche Cluster und deren Grenzen definiert werden. Dabei kommt es besonders auf die Struktur und Skalierung der vorliegenden Daten an. Einige typische Distanzmaße für die jeweiligen Skalenniveaus sind dabei [URL: http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_distance_sect016.htm [Zugriff am 17.10.2014]]:

-) metrisch / interallskalierte Daten: Euklidische Distanz: $d(a,b) = \sqrt{\sum_{i=1}^{p}(a_i - b_i)^2}$ (34)

Manhattan-Distanz: $d(a,b) = \sum_{i=1}^{p} |a_i - b_i|$ (35)

Gower-Maß: $d(a,b) = \sum_{k=1}^{p} \frac{|x_{ak} - x_{bk}|}{r_k}$ (36)

(mit x_{ak} als Wert der a-ten Instanz bei Variable k,
r_k als Spannweite der Variable k)

-) ordinal/nominal skalierte Daten: Hammingdistanz: d(a,b) = 1, wenn $a \neq b$ (37)
d(a,b) = 0, wenn a = b

Die Hammingdistanz kann für ordinal skalierte Variablen in einer angepassten Form angewendet werden, um Abstände zwischen verschiedenen Kategorien zu berücksichtigen. Ebenso kann das Gower-Maß bei passender Transformation metrisch skalierter und intervallskalierter Variablen bei allen Skalenniveaus verwendet werden. Dafür sollte eine Transformation dieser Variablen auf deren Rangstatistiken angewendet werden [URL: http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_distance_sect016.htm [Zugriff am 17.10.2014]].

Das populärste Verfahren zur Durchführung einer Clusteranalyse ist das sogenannte „K-means-Clustering“, welches für metrisch skalierte Beobachtungen zugeschnitten ist (zur Anwendung im Kontext anderer Skalierungen sind geeignete Transformationen notwendig). Der Fokus liegt dabei auf den Distanzen zwischen Beobachtungen, welche sich innerhalb eines Clusters befinden – deren Summe soll minimiert werden. Es geht somit um die Minimierung einer Zielfunktion mit folgender Gestalt [James et al. 2013]:

$$\min_{C_1,\ldots,C_K}\{\sum_{k=1}^{K}\frac{1}{|C_K|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2\} \quad (38)$$

Die Lösung des Minimierungsproblems wird über ein iteratives Verfahren aufgebaut. Die Variante der zufälligen Initialisierung sieht dabei wie folgt aus [James et al. 2013]:

-) Erste Clusterzuordnung - Teile den Instanzen eine zufällig gewählte Zahl zwischen 1 und K zu
-) Iteration bis zur Erfüllung eines Stoppkriteriums:
  .) Berechnung des Cluster-Zentroides (Mittelpunkt, Mittelwertsvektor) für jeden der K Cluster
  .) Zuordnung jeder Instanz zu Cluster, dessen Zentroid am nächsten zur Beobachtung liegt (kleinste euklidische Distanz)

Typischerweise wird als Stoppkriterium eine sich nicht mehr verändernde Zuteilung der Instanzen zu den Clustern verwendet. Die Durchführung dieses Verfahrens liefert eine kontinuierliche Verbesserung (Verkleinerung) der Zielfunktion hin zu einem lokalen Minimum.

Um zu besseren Resultaten als jenen zu kommen, welche durch einzelne lokale Minima geliefert werden, macht es Sinn, den gesamten K-Means-Prozess (mit unterschiedlich gewählten Zufallszahlen in Schritt 1) in einer iterativen Form anzuwenden, und verschiedene erlangte lokale Minima miteinander zu vergleichen.

Die schwierigste Aufgabe bei der Anwendung dieses Clusterverfahrens ist die zielführende Wahl des Parameters K. Auch hier ist es angebracht, die Iteration des gesamten K-Means-Verfahrens unter Verwendung verschiedener Werte von K anzusetzen (Betrachtung kreuzvalidierter Distanzen von Zentroiden, Anwendung von Stoppkriterien). Falls dies nicht zu einem gewünschten Ergebnis führt oder von Beginn weg mit einem zu hohen Rechenaufwand verbunden ist, gibt es noch Alternativen zur K-Means-Methode: Erwähnenswert ist vor allem das „hierarchische" Clustern, bei welchem es darum geht, Cluster mit minimaler (interner) Varianz schrittweise solange miteinander zu vereinen, bis ein vorab gewähltes Unähnlichkeitsmaß (beispielsweise 1 – Gowermaß, dieses wurde in diesem Abschnitt bereits beschrieben) eine kritische Schranke unterschreitet. Hinsichtlich der Vereinigung der Cluster gibt es diverse Verfahren: Zu den hochwertigsten Verfahren zählt die Clustermethode nach Ward, wo der Anstieg der Distanz-Quadratsummen von Instanzen zu den jeweiligen Zentroiden durch die Vereinigung beurteilt wird (die Minimierung dieses Anstiegs ist die Zielsetzung) [Aggarwal & Reddy 2013].

5.2.6) Ensemble Learning – Kombination diverser Verfahren

Das Grundprinzip des Ensemble Learnings - die kombinierte Verwendung von Verfahren des Machine Learnings - wurde bereits in Punkt 5.2.3, bei der Erläuterung der Random Forests, in Betracht gezogen. Dort wurde die gleiche Prozedur, nämlich die Generierung von Klassifikationsbäumen mithilfe zufällig gewählter Attribute und Bootstrap-Samples, in multipler Form durchgeführt, und am Ende ein Voting betrachtet. Dies garantiert, wie damals bereits erwähnt wurde, einerseits eine Reduktion der Varianz im Verhältnis zu einzelnen Modellberechnungen, andererseits die Konvergenz des Fehlers bei der Verallgemeinerung. Nun ist es auch möglich, in einer ähnlichen Form eine Kombination verschiedener Lern-Algorithmen zur Verbesserung der Qualität von Prognosen zu bewerkstelligen. Die grundlegende Motivation dahinter, welche den Nutzen des Ensemble Learnings nahelegt, steckt im sogenannten „No Free Lunch Theorem": Laut diesem vorwiegend aus dem ökonomischen Bereich bekannten Theorem, welches auch im Machine Learning Anwendung findet, gibt es bei der Betrachtung aller möglichen statistischen Probleme, welche mithilfe des Machine Learnings gelöst werden können, kein universell bestes Verfahren, welches angewendet werden kann [Wolpert 1996].

Das heißt, dass es besonders von der Art der Problemstellung und vom Aufbau des vorliegenden Datensatzes abhängig ist, ob ein Verfahren zu genauen Prognosen führt oder nicht. Es bedeutet vor allem auch, dass Anwender von Machine Learning – Verfahren am

Beginn der Analysen, basierend auf den Eigenschaften des Datensatzes, sich Gedanken über die Thematik der Performance machen und eine sorgfältige Selektion der Menge der für die gegebene Problemstellung anwendbaren Verfahren durchführen sollten, um bereits hier Stärken und Schwächen von diesen bei der vorliegenden Aufgabe einschätzen zu können. Insbesondere in der Situation, wenn sich das Auffinden des günstigsten Verfahrens als schwierig herausstellt, kann die Variante der Kombination verschiedener Methoden eine sehr effektive Alternative darstellen. Dies kann sich besonders dann als nützlich erweisen, wenn Basislerner, die unterschiedlichen Ansätzen bei der Generierung der Entscheidungen folgen, einander ergänzen, also wenn eine Lernmethode die Schwäche einer anderen Methode ausgleichen kann [Alpaydin 2008].

Nun gibt es verschiedene Varianten, mit denen die Kombinationen von einander ergänzenden Lernverfahren vorangetrieben werden können: Im Werk „Maschinelles Lernen“ [Alpaydin 2008] werden folgende Möglichkeiten vorgestellt:

<u>Möglichkeiten für die Generierung</u>:

-) Training unterschiedlicher Basislerner durch Nutzen verschiedener Lernalgorithmen, deren Annahmen an den Daten sich unterscheiden

-) Generierung unterschiedlicher Varianten einer Lernmethode mithilfe von Parametervariation (z.B. bei K-Nearest-Neighbors)

-) Nutzung unterschiedlicher Repräsentationen vom selben Eingabeobjekt bzw. Eingabeereignis bei unterschiedlichen Basislernern

-) Verwendung unterschiedlicher Trainingssets zur Generierung der Basislerner (Bagging, bereits in Abschnitt 5.2.3 beschrieben)

<u>Möglichkeiten für die Kombination:</u>

-) Kombination von Expertensystemen über ein separates Kombinationsverfahren, das Entscheidungen der Systeme heranzieht und sie miteinander verknüpft

Die einfachste Variante einer Umsetzung stellt das Voting dar (welches bereits in Abschnitt 5.2.3 beschrieben wurde).

-) Mehrstufige Kombinationsmethoden: Schrittweises Aufwerten der Performance der verwendeten Basislerner: Die i+1 – te Lernmethode soll dabei jene Instanzen möglichst gut vorhersagen können, bei denen die i-te Methode Schwächen gezeigt hat

Ein bekanntes Verfahren von dieser Form, welches in mehrstufiger Form abläuft, ist das sogenannte „Boosting“ [Freund & Schapire 1996].

Wie bei der Methode der K-Nearest Neighbors können auch beim Voting Ansätze aus der Bayesianischen Statistik bei der Implementierung betrachtet werden: Bei der Bayesschen Modellkombination, wie sie im Buch „Maschinelles Lernen“ [Alpaydin 2008] geschildert wird, werden Annäherungen der a-priori – Wahrscheinlichkeiten eines Modells durch Gewichte und von modellbezogenen Likelihood-Werten durch Modellentscheidungen betrachtet:

$$P(C_i|x) = \sum_{alle\ Modelle\ M_j} P\left(C_i\middle|x, M_j\right) * P(M_j) \qquad (39)$$

Der erste Ausdruck im Produkt der rechten Seite (bedingte Wahrscheinlichkeit für Klasse gegeben Modell und Daten) spiegelt dabei die Stimmabgabe des Lerners j für die Klasse C_i wider. Der zweite Ausdruck umfasst das Gewicht der Stimme, hier im Kontext der Wahrscheinlichkeit des Modells M_j.
In diesem Kontext kann man das ungewichtete Voting mit einer einheitlichen a-priori-Wahrscheinlichkeit deuten.

6) Vorbereitung des Datensatzes für die Analyse

6.1) Deskriptive Analyse

Wie im vorletzten Kapitel beschrieben wurde, besteht der Rohdatensatz aus 902 Wertpapieren, der sowohl finanzwirtschaftliche als auch makroökonomische Variablen beinhaltet, sie ergeben zusammen insgesamt 15 Prädiktorvariablen, die in weiterer Folge im Vordergrund stehen. Ebenfalls beschrieben wurde die Form der abhängigen Variable, dabei handelt es sich um die „Amihud Illiquidity Ratio“, welche als Quotient zwischen dem absoluten Return und dem gehandelten Volumen am betrachteten Tag beschreibt, wie stark sich Preisentwicklungen durch einzelne, kleinere Handelsvorgänge beeinflussen lassen. Im Falle liquider Wertpapiere ergibt sich praktisch kein Einfluss, während bei gering liquiden oder illiquiden Beständen das Zustandekommen einzelner Handelsvorgänge im Extremfall Schwankungen im Bereich von 10 Prozent auslösen kann.

Am Beginn dieses Kapitels sollen grundlegende Eigenschaften des vorliegenden Datensatzes untersucht werden. Dabei sollte auf eine differenzierte Behandlung von Variablen mit unterschiedlicher Skalierung geachtet werden, um einen besseren Überblick generieren zu können (bei nominal und ordinal skalierten Variablen werden dabei Häufigkeitsauswertungen, bei intervall- und metrisch skalierten Variablen Verteilungsanalysen und Auswertungen von Lage- und Streuungsmaßen anvisiert).

Es ergeben sich für die jeweiligen Variablen nun folgende Werte:

Häufigkeiten (Skalenniveau nominal oder ordinal):

Rating

RATING	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C	NA
Anzahl	8	45	91	153	164	61	11	1	0	182

Tabelle 6.1.: Rating-Häufigkeiten gesamt, für Instanzen mit nicht-fehlender Liquidität

RATING	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C	NA
Anzahl (LIQ = 1)	3	25	40	48	40	3	1	1	0	19
Anzahl (LIQ = 2)	2	12	38	64	62	23	2	0	0	55
Anzahl (LIQ = 3)	3	6	9	27	30	21	6	0	0	56
Anzahl (LIQ = 4)	0	2	4	12	25	11	2	0	0	39
Anzahl (LIQ = 5)	0	0	0	2	7	3	0	0	0	13

Tabelle 6.2.: Rating-Häufigkeiten, gruppiert nach Liquidität

Listungen auf Handelsplätzen:

NUM_LIST	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	NA
Anzahl	155	115	63	50	58	57	36	31	37	51	33	12	8	3	1	1	1	4

Tabelle 6.3.: Anzahl der Listungen auf Handelsplätzen, gesamt

Anzahl Listungen auf Handelsplätzen (nach Liquidität):

NUM_LIST	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	NA
Anzahl (LIQ = 1)	28	27	2	4	8	6	9	10	19	21	25	8	7	3	1	1	1	0
Anzahl (LIQ = 2)	47	29	23	20	19	30	19	16	16	24	7	4	1	0	0	0	0	3
Anzahl (LIQ = 3)	34	38	21	15	25	11	3	4	2	6	0	0	0	0	0	0	0	0
Anzahl (LIQ = 4)	35	14	15	10	5	9	5	0	0	0	0	0	0	0	0	0	0	1
Anzahl (LIQ = 5)	11	7	2	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0

Tabelle 6.4.: Anzahl der Listungen auf Handelsplätzen, gruppiert nach Liquidität

Wertpapier-Sektor (Gesamt)

SECTOR	Basic Materials	Communication Services	Consumer Cyclical	Consumer Defensive	Energy
Anzahl	71	17	102	62	44

SECTOR	Financial Services	Healthcare	Industrials	Real Estate	Technology	Utilities
Anzahl	94	34	117	34	65	26

Tabelle 6.5.: Anzahl Wertpapiere jeweiliger Sektoren, gesamt

Land

COUNTRY	AT	DE	GB	IT	ES	NO	US	JP	IR
Anzahl	42	44	47	46	41	44	48	49	41

COUNTRY	TR	CZ	HU	BG	RU	CHN	BR	VZ	SA
Anzahl	34	11	31	26	21	48	33	12	33

COUNTRY	ID	EG
Anzahl	34	31

Tabelle 6.6.: Anzahl Wertpapiere aus jeweiligen Ländern, gesamt

COUNTRY	AT	DE	GB	IT	ES	NO	US	JP	IR
Anzahl (LIQ=1)	1	9	24	10	10	10	17	14	4
Anzahl (LIQ=2)	11	20	16	17	22	14	16	27	17
Anzahl (LIQ=3)	12	8	6	15	8	13	11	8	13
Anzahl (LIQ=4)	14	6	1	4	1	6	4	0	7
Anzahl (LIQ=5)	4	1	0	0	0	1	0	0	0

COUNTRY	TR	CZ	HU	BG	RU	CHN	BR	VZ	SA	ID	EG
Anzahl (LIQ=1)	9	0	0	0	4	41	9	0	5	12	1
Anzahl (LIQ=2)	22	6	9	1	7	6	12	0	10	9	16
Anzahl (LIQ=3)	1	3	8	4	5	1	7	7	10	7	12
Anzahl (LIQ=4)	1	2	6	16	3	0	3	4	8	6	2
Anzahl(LIQ=5)	1	0	8	5	2	0	2	1	0	0	0

Tabelle 6.7.: Anzahl Wertpapiere aus jeweiligen Ländern, gruppiert nach Liquidität

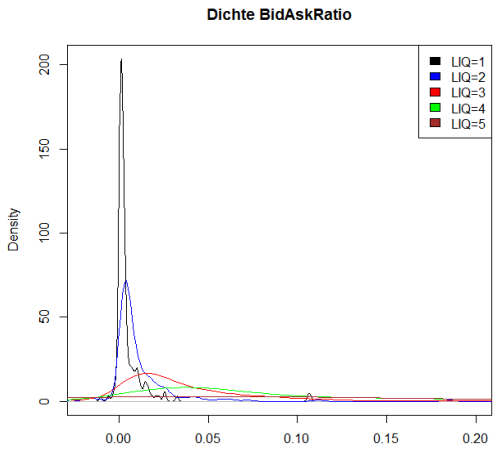
Lage-/Streumaße

Variable / Maßzahl	MIN	1.QTL	MW	MED	3.QTL	MAX	STD
BidAskRat	-0,0911	0,0025	0,0518	0,008	0,0284	10,293	0,4214
Vol-Vrd.	-0,7583	0,0385	0,3986	0,1264	0,3442	23,0206	1,4151
Turnover	6,000	47970,00	267300000	946400	12810000	2,805*10^8	1,612*10^-9
STD Retrn	0,00089	0,01468	0,02922	0,02123	0,03312	0,24425	0,0271
CPI	-2,200	0,0003	0,00455	0,00212	0,00487	2,16667	0,1291
Ind.Prod.	-0,1784	-0,0089	0,00084	0,00277	0,0107	0,1513	0,0310
REER	-0,1212	-0,0066	0,00019	-0,0001	0,0068	0,1480	0,0175
Verf. GM	-0,0703	0,00194	0,00834	0,00747	0,01503	0,1258	0,0157
Kurzfr. ZS	-0,8279	-0,0294	0,00098	0	0,02597	1,1705	0,1456
Langfr. ZS	-0,3067	-0,0351	-0,00126	-0,0065	0,0336	0,3785	0,0682
EconSent	-23,000	-0,00811	-0,0881	0,00029	0,00784	10,000	1,3195

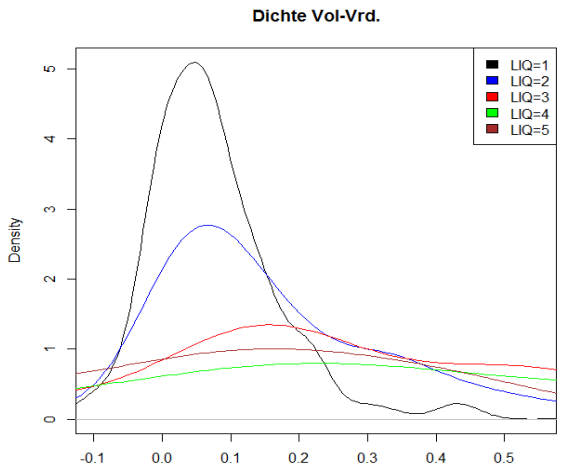
Tabelle 6.8.: Lage- und Streumaße für intervall- und metrisch skalierte Variablen, gesamt

Dichteplots für intervall- und metrisch skalierte Variablen (für Liquiditätskategorien):

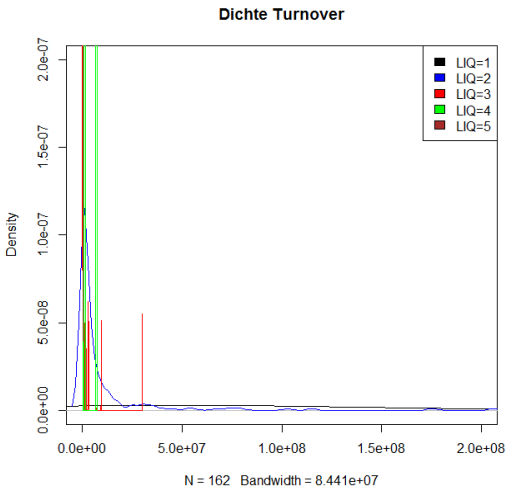
BidAskRatio



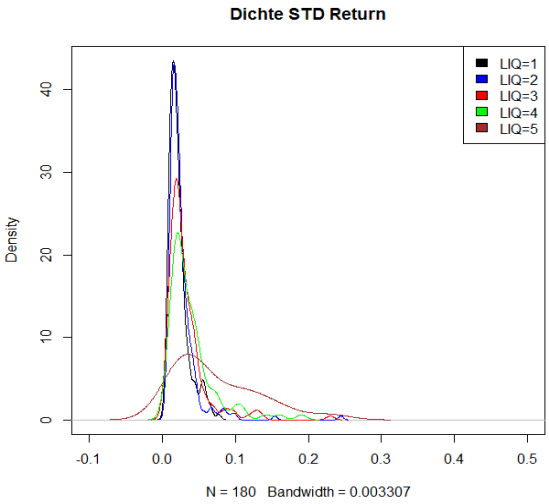
Volumsveränderung



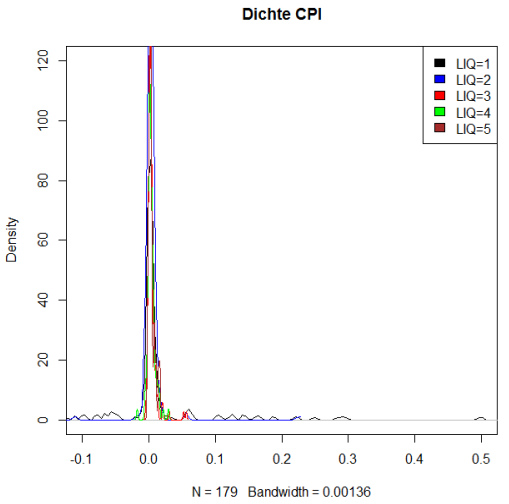
Turnover



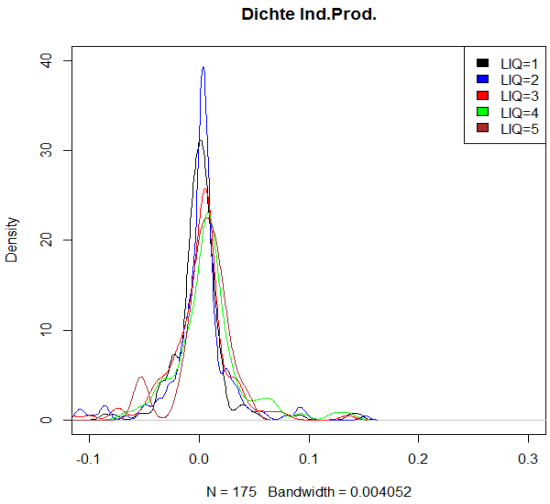
STD Return



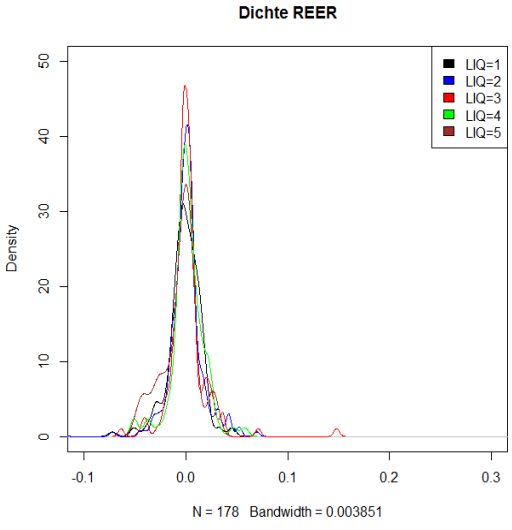
Consumer Price Index



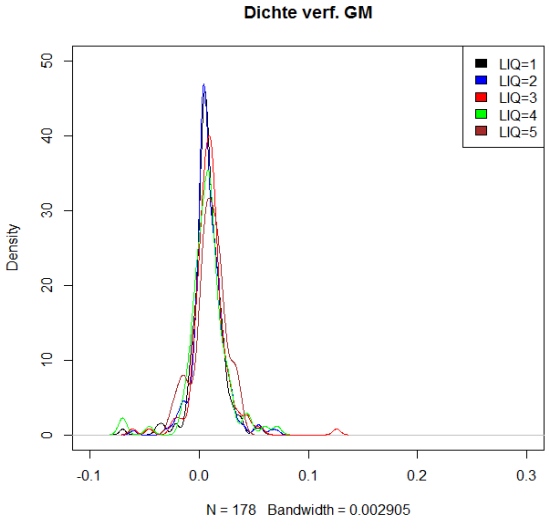
Industrieproduktion



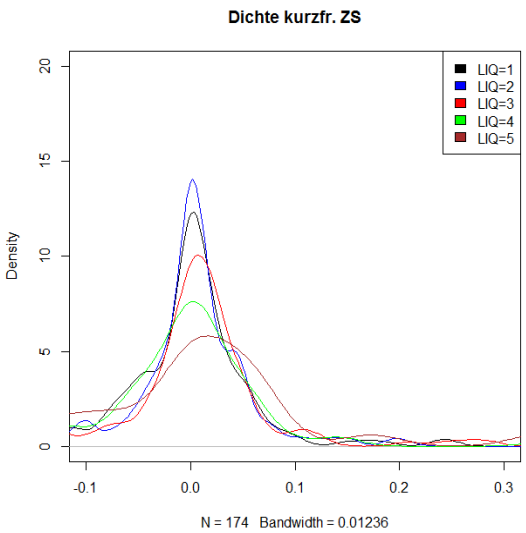
REER



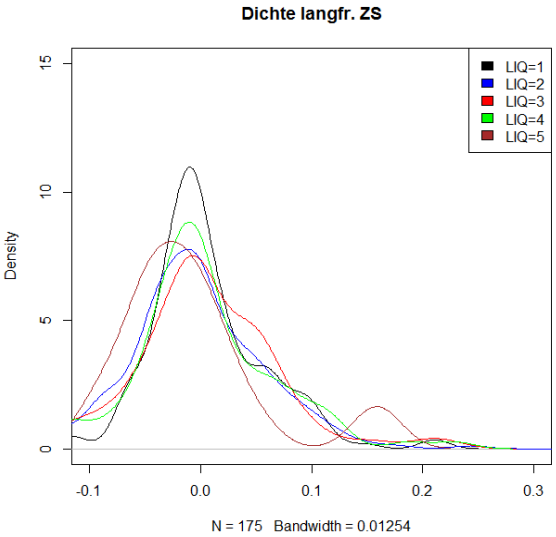
Verfügbare Geldmenge



Kurzfristiger ZS



Langfristiger Zinssatz



Economic Sentiment

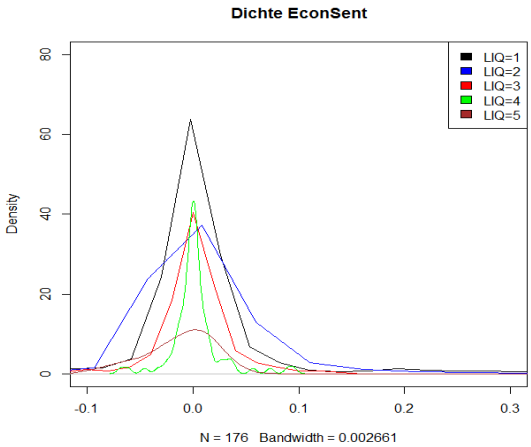


Abbildung 6.1.: Dichteplots für metrische Prädiktorvariablen, differenziert nach Liquiditätsgrad

Boxplots (zur Identifikation von Ausreissern)

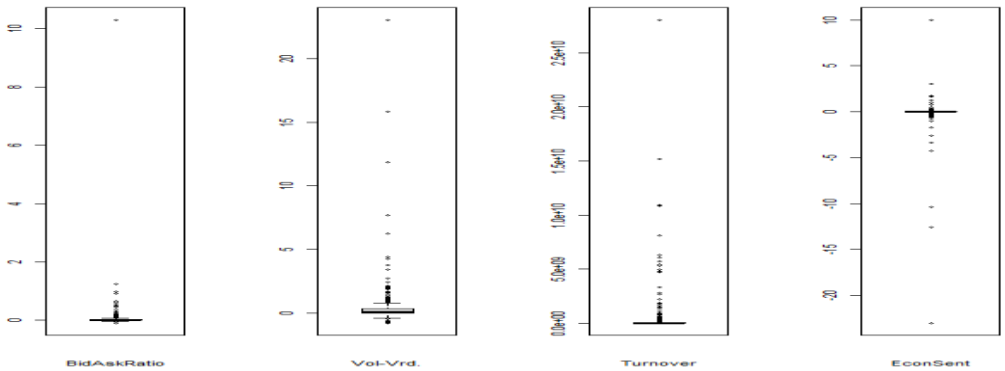


Abbildung 6.2.: Boxplots zur Identifikation von Ausreissern bestimmter Variablen

Im vorliegenden Fall können aus den deskriptiven Auswertungen bereits Schlüsse gezogen werden, welche für die Fortführung der Analyse von großer Hilfe sein können: So liegen bereits Indizien vor, welche Variablen eine größere Aussagekraft hinsichtlich der Erklärung der Liquidität zu haben scheinen. Von besonders großer Wichtigkeit ist die deskriptive Analyse am Beginn einer statistischen Untersuchung meist deshalb, weil ein Überblick über die vorliegende Verteilung gewonnen und vor allem die Identifikation von Ausreissern (sogenannte „high leverage points“) erzielt werden kann, diese können auch bei Anwendung spezieller Klassifikationsmethoden einen ungünstigen Einfluss haben.

Es stechen nun bei näherer Betrachtung der Auswertungen der nominal und ordinal skalierten Variablen das Rating des Wertpapiers und die Zahl der Listungen auf Börsenplätzen hinsichtlich auftretender Unterschiede zwischen den verschiedenen Liquiditätsklassen hervor, bei den intervallskalierten und metrisch skalierten Variablen sind dies die Bid-Ask-Ratio, die Volumsveränderung, der kurzfristige Zinssatz und der Economic-Sentiment-Indikator, welche bei unterschiedlichen Liquiditätsklassen ein unterscheidbares Bild aufweisen. Bei einzelnen Indikatoren gibt es auch Unterschiede, die nur eine der fünf Klassen betrifft, wie beispielsweise bei der Standardabweichung des Returns, wo die Dichte bei der illiquiden Klasse einen deutlich flacheren Verlauf zeigt als die der anderen Klassen.

Im nächsten Abschnitt soll der Fokus nun besonders auf die Anpassung und Elimination von fehlenden und problembehafteten Werten gerichtet werden.

6.2) Bereinigung des Datensatzes (fehlende / nicht akkurate Werte)

Da die Daten einerseits aus Ländern stammen, wo die Verfügbarkeit von Daten, insbesondere in historischer Hinsicht, keine gleichmäßigen Züge vorweist, andererseits die Daten innerhalb einer Zeitspanne betrachtet werden, in welcher einige Unternehmen noch nicht denjenigen Status erreicht haben, der sie aktuell auszeichnet (und daher Finanzdaten für das entsprechende Wertpapier zu diesem Zeitpunkt nur mangelhaft oder gar nicht zur Verfügung stehen könnten), musste damit gerechnet werden, dass im verwendeten Datensatz viele fehlende Werte zum Vorschein kommen. Das bedeutet, dass bei einem Teil der fehlenden Werte von rein zufälligem Fehlen („missing at random“) ausgegangen werden kann (wo durch Zufall ein zu früher Zeitpunkt ausgewählt wurde), beim Rest jedoch eine gewisse Systematik denkbar ist (fehlende Marktdaten aufgrund reiner Nichtverfügbarkeit – Erhebung der benötigten Daten hat beispielsweise bei kleineren Unternehmen nicht stattgefunden).

Es ergeben sich nun für die 15 Prädiktorvariablen sowie für die abhängige Variable folgende Zahlen fehlender Werte:

Variable	Fehlende Werte	Fehlende Werte nach Reduktion der Instanzen
BidAskRatio	250	87
Veränderung Volumen	204	50
Turnover	290	110
Rating	314	182
Anzahl Listungen Börsen	7	4
Standardabw. Return	161	4
Sektor	97	50
Land	0	0
CPI	40	31
Industrieproduktion	82	48
Real-effektiver WK	68	51
Verfügbare Geldmenge	42	32
Kurzfristiger Zinssatz	73	55
Langfristiger Zinssatz	105	62
Economic Sentiment	117	73
Liquidität	186	0

Tabelle 6.9.: Zahl fehlender Werte für unabhängige und abhängige Variablen, vor und nach Elimination fehlender Werte bei abhängiger Variable

Es kann nun überlegt bzw. überprüft werden, ob über eine geeignete Imputationsmethode das Auffüllen der fehlenden Werte gelingen kann. Jedenfalls scheint es wenig Sinn zu machen, Werte der kategoriellen abhängigen Variable durch Näherungswerte zu ersetzen, da es gerade hier keine günstigen Auswirkungen hätte, wenn die Variabilität anderer Beobachtungen (die nicht zwingend derselben Systematik wie bei fehlenden Elementen folgen muss) zur Ersetzung eingesetzt werden würde, oder durch verwendete Zufallsmechanismen bei der Imputation gewisse Verzerrungen bei der abhängigen Variable und folglich auch bei der Abhängigkeitsstruktur entstehen würden. Daher sollte man auf diese 186 Beobachtungen verzichten und die übrigen 716 Beobachtungen in den Datensatz aufnehmen, mit welchem die Verfahren zur Analyse der Zusammenhänge zwischen abhängiger und erklärender Variablen durchgeführt werden.

Um eine klare Aussage darüber treffen zu können, ob eine Imputation fehlender Werte, deren Zahl bei den Prädiktorvariablen sich durch Einschränkung des Datensatzes bezüglich der abhängigen Variable zumeist deutlich reduziert hat (30%-75% weniger fehlende Werte als zuvor, lediglich beim Rating sind noch 25% der Werte fehlend, sonst liegt deren Zahl zwischen 0% und 15%), einen gewinnbringenden Effekt hat, sollte ein Überblick über die wesentlichen deskriptiven Eigenschaften vorliegen. Dies wurde bereits in Abschnitt 6.1 verwirklicht, wodurch die Evaluierung der Sinnhaftigkeit einer Imputation vorangetrieben werden kann.

Abbildung 6.1 sowie die im letzten Abschnitt erzeugten Auswertungen zeigen recht deutlich, dass sich die Situation von Variable zu Variable unterscheidet, und dass bei der Beurteilung,

ob und wo Anpassungen wie Transformationen oder Elimination von Ausreissern nötig sind, alle Variablen differenziert betrachtet werden sollten.

Eine sehr günstige Situation ergibt sich bei der Standardabweichung des Returns, beim CPI, bei der Industrieproduktion, beim real-effektiven Wechselkurs, bei der verfügbaren Geldmenge sowie beim langfristigen Zinssatz – hier deuten eine sehr kleine Standardabweichung bzw. kleine Extremwerte und eine geringe Schiefe (welche fast überall zu erkennen ist) keine Notwendigkeit von Korrekturmaßnahmen hinsichtlich der Verteilungseigenschaften an.

Etwas anders ist die Ausgangslage bei Variablen wie der Bid-Ask-Ratio, der Volumsveränderung, dem Turnover oder dem Economic-Sentiment-Indikator, wo sich teilweise starke Ausreisser, welche typische Werte um ein Vielfaches über- bzw. untertreffen, bemerkbar machen (siehe Tabelle 6.3). Bei jenen Variablen, deren Werte nicht kleiner als 0 sein können, könnte eine Logarithmierung der Werte nun als zielführender Weg angesehen werden. Für die übrigen problembehafteten Variablen müsste ein anderer Weg gesucht werden. Da jedoch die Zahl der massiven Ausreisser eher gering ist und Verfahren des maschinellen Lernens nicht allzu sensitiv bezüglich Verteilungsannahmen sind, wird lediglich eine obere Schranke (sowie eine untere Schranke beim Economic-Sentiment – Indikator) für diese vier Variablen festgelegt, und auftretende Extremwerte auf diese Schranke reduziert.

Bezüglich fehlender Werte wird eine zweiseitige Herangehensweise anvisiert: bei den nominal und ordinal skalierten Variablen macht die Betrachtung der Nichtverfügbarkeit eines Wertes als eigenständige Kategorie durchaus Sinn (vor allem beim Rating, wo Nichtverfügbarkeit ein eher schlechtes Bild von der Instanz generiert), während bei den übrigen Skalenniveaus eine Imputation im Kontext der erfolgenden Standardisierung günstig erscheint, selbst wenn bei einem Teil der fehlenden Werte ein systematisches Fehlen vorliegt. Jedoch kann dies hier nur über das Auffinden von Ähnlichkeiten zwischen Instanzen, bei denen keine fehlenden Werte vorhanden sind, und jenen, bei denen Daten fehlen, in einer brauchbaren Weise geschehen. Als brauchbare Variante zur Imputation wird deshalb die Mittelung der Werte der 10 nächsten Nachbarn betrachtet, wobei dies nach dem in Kapitel 5.2 beschriebenen Verfahren durchgeführt wird. Diese Vorgehensweise verursacht in den meisten Fällen (vor allem bei geringen Varianzen und einer geringen Zahl extremer Werte) auch keine großen Veränderungen bei der Dichte der Variablen, sodass in der Regel die Qualität des Datensatzes erhalten bleibt.

6.3) Zusammenhänge zwischen den Variablen

Im Vorfeld jeder statistischen Analyse sollte überprüft werden, ob und in welcher Größenordnung Korrelationen zwischen den Prädiktorvariablen vorliegen. Während eine hohe Korrelation zwischen Prädiktorvariablen und der abhängigen Variable vorteilhaft ist, sind sehr große (lineare) Zusammenhänge zwischen den Prädiktorvariablen besonders bei

parametrischen Modellen unerwünscht, da Multikollinearität bei diesen Variablen dazu führt, dass deren Vorhersagekraft schlecht evaluiert werden kann, und Freiheitsgrade in den Modellen verschwendet werden könnten.

Bei Methoden des maschinellen Lernens, wo Voraussetzungen an die Daten typischerweise schwächer sind als im vorhin erwähnten Fall, sollte man auch auf diese Art von Variableneigenschaften achten, da mit einer größeren Menge wenig korrelierter Prädiktorvariablen deutlich vielversprechendere Resultate zu erhalten sind als in anderen Fällen (es ist mehr erklärte Varianz durch das Modell zu erwarten).

Als Darstellung von Korrelationen zwischen Variablen eignen sich sowohl Korrelationstabellen als auch sogenannte Scatterplot-Darstellungen, in der jede Variable gegen jede andere Variable geplottet wird. Stärkere Zusammenhänge zeigen sich hier, wenn sich die Werte im Plot entlang einer Linie aufhalten, im Optimalfall sind sie ohne auffälliges Muster gleichmäßig im Raum verteilt.

Es zeigt sich für die metrischen Prädiktorvariablen nun folgendes Resultat:

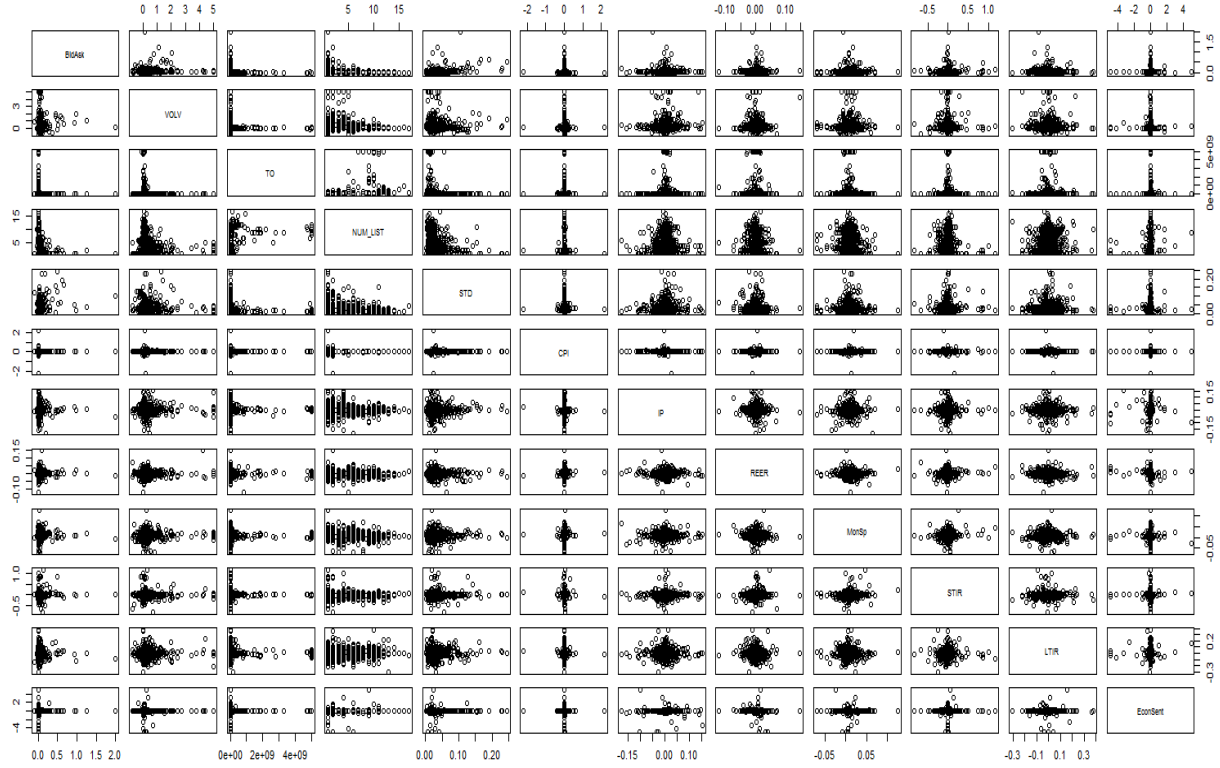


Abbildung 6.3.: Scatterplot-Matrix zur Darstellung der Zusammenhänge zwischen metrischen Prädiktorvariablen

Es gibt in nahezu allen Fällen kein Indiz auf einen stärkeren Zusammenhang zwischen den verschiedenen Prädiktorvariablen. Als Ergänzung wurden auch Korrelationstabellen mit allen Kombinationen der Prädiktorvariablen ausgewertet, dabei zeigt sich in keinem Fall ein Wert von über 0,4 bei der Korrelation. Leichte Korrelationen gibt es beispielsweise zwischen der Bid-Ask-Ratio und der Standardabweichung des Returns (0,37), der Anzahl der Listungen an

Börsen und dem Turnover (0,28) und zwischen der Anzahl der Listungen an Börsenplätzen und der Standardabweichung des Returns (-0,26 – leicht negativ). Etwas überraschend sind die außerordentlich geringen Korrelationen zwischen den makroökonomischen Variablen, welche in nahezu allen Fällen betragsmäßig kleiner als 0,1 sind. Zwischen diesen scheint es also so gut wie keinen linearen Zusammenhang zu geben.

6.4) Vorgangsweise bei der Klassifizierung und Anwendung der beschriebenen Verfahren

Bevor die in Kapitel 5 beschriebenen Verfahren zur Anwendung kommen können, müssen noch Vorkehrungen zur Vorbereitung des Klassifizierungsprozesses getroffen werden. Insbesondere ist die Überlegung notwendig, wie eine Aufteilung in Trainingsset, Testset bzw. Validierungsset (zur Optimierung von Parameterwerten) erfolgen soll, um die Qualität von Klassifizierungsverfahren zu evaluieren. Typischerweise wird bei der Anwendung von Verfahren des „supervised learnings“ eine Aufteilung des Datensatzes in einer derartigen Form empfohlen, sodass 50% der Instanzen in ein Trainingsset und jeweils 25% in ein Validierungsset und ein Testset fallen (siehe Kapitel 5.1). Aufgrund des nach der Bereinigung relativ klein gewordenen Datensatzes sollte man sich hier jedoch das Vorgehen bei einer derartigen Aufteilung etwas genauer überlegen, da es sein kann, dass zu kleine Sets resultieren, deren Variabilität und Eigenschaften die der gesamten Datenmenge bzw. der Grundgesamtheit nicht in einer passenden Form repräsentieren.

Als günstige Variante, um das Ziel dieser Evaluierung zu erreichen, kann der Einsatz eines Kreuzvalidierungsverfahrens gesehen werden, wie es in Abschnitt 5.1 bereits beschrieben worden ist. Bei der Anwendung der 10-fach-Kreuzvalidierung zur Beurteilung einer Klassifizierungsmethode würde das bedeuten, dass bei jedem Iterationsschritt ein Zehntel des Datensatzes (im vorliegenden Fall wären dies etwa 72 Instanzen) als Testset herangezogen wird, auf welches die Resultate angewendet werden, welche sich aus der Berechnung der betrachteten Klassifizierungsmethode bei einem Trainingsset mit allen übrigen Instanzen ergeben haben.

Der Einsatz eines derartigen Verfahrens in einer mehrfachen Hinsicht (zum Beispiel eine fünfmalige Anwendung einer 10-fach-Kreuzvalidierung bei jeder Methode) sollte in der vorliegenden Arbeit genügen, um einen brauchbaren Vergleich der Performance verschiedener Methoden zu erreichen. Diesem wird im Anschluss an die nun folgende Darstellung der allgemeinen Resultate, die für alle in Kapitel 5 beschriebenen Verfahren berechnet werden, ein eigener Abschnitt gewidmet. Daneben wird auch über Konfusionsmatrizen versucht, ein Bild über Performancemaße wie Trefferrate, Sensitivität oder Spezifität zu erhalten.

7) Klassifikation des Liquiditätsrisikos – Anwendung der präsentierten Verfahren und Darstellung der Ergebnisse

7.1) Classification Trees

Nach der entsprechenden Vorbereitung werden in weiterer Folge die vorgestellten Klassifikationsverfahren auf den im Fokus stehenden Datensatz angewendet, die Ergebnisse präsentiert und erläutert sowie ein erster Vergleich der Routinen durchgeführt.

Die Berechnung der angewendeten Modelle sowie die Ermittlung von Resultaten und grafischen Darstellungen wird mithilfe der Statistik-Software R unter Verwendung diverser Packages für Data-Mining-Anwendungen durchgeführt.

Zu Beginn kommen die im Kapitel 5 zuerst vorgestellten Klassifikationsbäume zur Anwendung. Von großer Bedeutung ist dabei eine adäquate Bestimmung des im Kapitel 5.2.1 bereits erwähnten Cost-Complexity – Parameters (kurz CP), dessen Wert eine Quantifizierung der Komplexität des berechneten Baumes ermöglicht (je kleiner der Wert, desto komplexer das Baummodell – dieses wird dann mehr Knoten umfassen).

Bei der Anwendung erweist es sich als günstig, zunächst einen recht komplexen Baum (durch Setzen des CP-Parameters auf einen niedrigen Wert, hier wird 0,001 verwendet) zu generieren, und folglich über die ebenfalls bereits beschriebene Variante des „Post-Prunings“ die Komplexität durch das Abschneiden der untersten Reihen des Baumes wieder zu reduzieren. Um dies möglichst effizient zu gestalten, kann das Augenmerk auf den sogenannten „kreuzvalidierten“ Fehler gerichtet werden, welcher sich nicht mit zunehmender Baumgröße immer weiter reduziert, sondern ein lokales Optimum erreicht (über Kreuzvalidierung wird eine Pönalisierung einer zu umfangreichen Baumstruktur ermöglicht). Typischerweise wird zusätzlich die „1-Standardabweichungs-Regel“ berücksichtigt: dabei wird nicht die Baumkomplexität für das Prunen herangezogen, bei welcher sich der minimale kreuzvalidierte Fehler herausstellt, sondern diejenige, welche sich nach zusätzlicher Addition der resultierenden Standardabweichung ergibt.

Die Anwendung dieser Methodik auf den Datensatz liefert nun folgendes Bild für die Entwicklung des kreuzvalidierten Fehlers, abhängig von der Baumkomplexität:

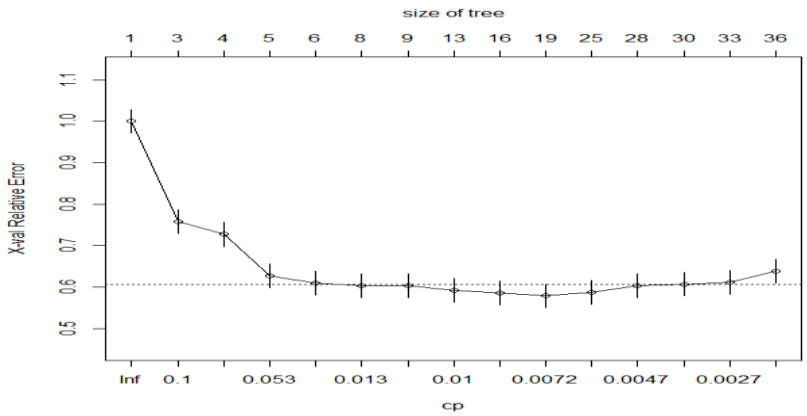


Abbildung 7.1: Entwicklung des kreuzvalidierten Fehlers bei Veränderung des Cost-Complexity-Parameters

Es gibt also einen relativ großen Unterschied zwischen den CP-Werten, welche sich je nach Entscheidung für oder gegen die 1-Standardabweichungs-Regel ergeben. Dies zeigt sich sehr deutlich bei der grafischen Darstellung der resultierenden Klassifikationsbäume:

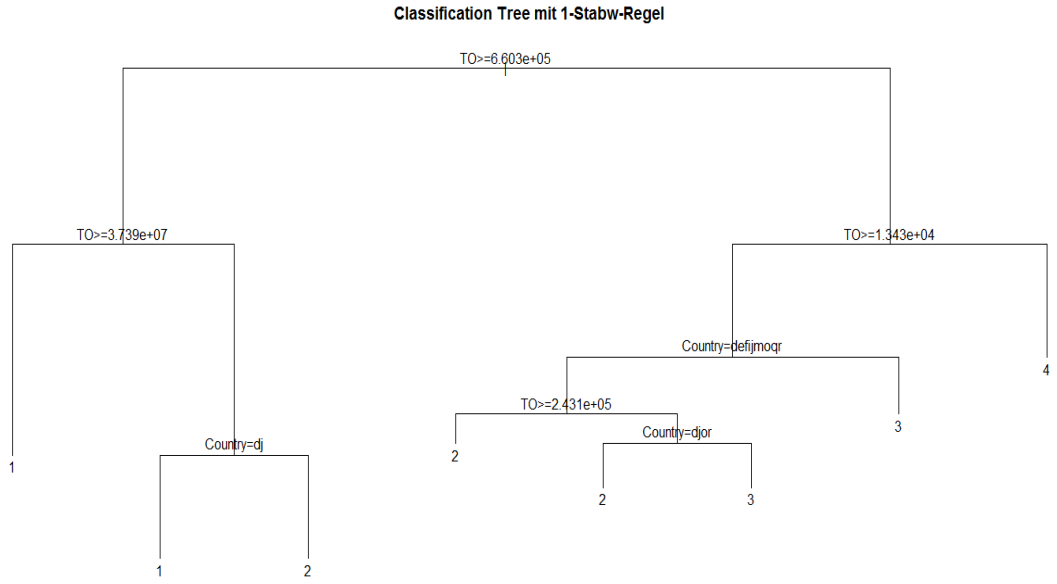


Abbildung 7.2: Darstellung des resultierenden Klassifikationsbaumes unter Miteinbeziehung der 1-Standardabweichungs-Regel

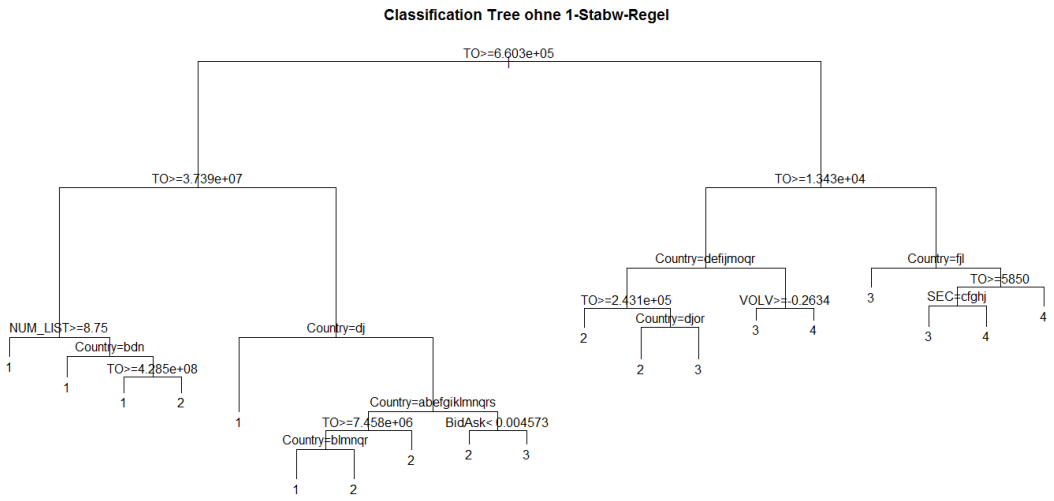


Abbildung 7.3: Darstellung des resultierenden Klassifikationsbaumes ohne Miteinbeziehung der 1-Standardabweichungs-Regel

Während bei Berücksichtigung dieser Regel sich ein Baum mit gerade einmal 6 Knoten herausstellt und nur 2 Variablen für die Verzweigungen (Turnover, Land) verwendet werden, sind es bei der komplexeren Variante ohne Standardabweichungsregel 18 Knoten unter Nutzung von 6 Prädiktorvariablen. Doch obwohl der zweite Baum schon in eine recht komplexe Richtung tendiert, hat dieses Ergebnis auch seine Vorzüge gegenüber dem ersten Baum, nachdem hier die Regeln für die verschiedenen Liquiditätskategorien Werte mehrerer Prädiktorvariablen einschließen.

Man erkennt bei Betrachtung des Outputs, dass eine Darstellung der Ländernamen bei der entsprechenden Variable innerhalb des Baumes kaum möglich ist. Die dabei verwendeten Abkürzungen (Buchstaben, welche der Reihe nach den Ländern zugeordnet werden) gehören zu folgenden Ländern:

a	Österreich	f	Ägypten	k	Irland	p	Südafrika
b	Brasilien	g	Deutschland	l	Italien	q	Spanien
c	Bulgarien	h	Großbritannien	m	Japan	r	Türkei
d	China	i	Ungarn	n	Norwegen	s	USA
e	Tschechien	j	Indonesien	o	Russland	t	Venezuela

Tabelle 7.1: Darstellung der Länderabkürzungen in den Baummodellen

Als 5 wichtigste Variablen nach der Beurteilung über ein Entropiemaß, wie es in Kapitel 5.1 vorgestellt wurde, stellen sich der Turnover, das Land, die Bid-Ask-Ratio, die Anzahl der Listungen an Börsenplätzen und die Volumsveränderung heraus.

Nun soll für beide Fälle über Konfusionsmatrizen und für die Variante mit der besseren Trefferrate auch über ROC-Kurven eine erste Evaluation der Qualität der Prognosen

durchgeführt werden, welche über die beiden Baummodelle gewonnen werden können. Vor allem auf die Trefferrate soll ein besonderes Augenmerk gerichtet werden.
Es ergeben sich nun folgende Tabellen:

Mit 1-Standardabweichungs-Regel:

Echter Wert/Prognosewert	1	2	3	4	5	Trefferrate
1	129	50	0	1	0	0,717
2	27	182	42	7	0	0,705
3	1	22	112	24	0	0,704
4	0	6	28	60	0	0,638
5	0	3	7	15	0	0,000

Tabelle 7.2: Konfusionsmatrix unter Einbeziehung der Standardabweichungs-Regel

Gesamt-Trefferrate bei Trainingsinstanzen: 67,5%

Ohne 1-Standardabweichungs-Regel:

Echter Wert/Prognosewert	1	2	3	4	5	Trefferrate
1	152	27	1	0	0	0,844
2	21	184	53	0	0	0,713
3	1	12	142	4	0	0,893
4	0	2	36	56	0	0,596
5	0	3	10	12	0	0,000

Tabelle 7.3: Konfusionsmatrix ohne Einbeziehung der Standardabweichungs-Regel

Gesamt-Trefferrate bei Trainingsinstanzen: 74,6%

ROC-Kurven als grafische Darstellung von Sensitivität und Spezifität:

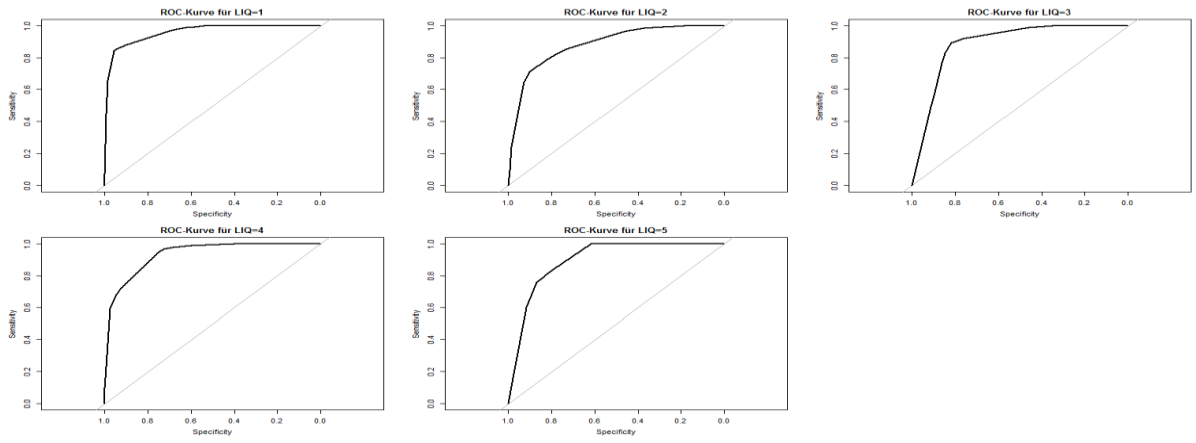


Abbildung 7.4.: Darstellung von ROC-Kurven für jede Liquiditätskategorie

Area under the curve	
LIQ=1	0,9600
LIQ=2	0,8839
LIQ=3	0,8885
LIQ=4	0,9338
LIQ=5	0,8979

Tabelle 7.4.: „Area under the Curve“ – Maß für jede Liquiditätskategorie

Die Ergebnisse dieser Matrizen samt Trefferraten dürfen an dieser Stelle nicht überbewertet werden, da zu berücksichtigen ist, dass bei der Berechnung der Modelle alle verfügbaren Instanzen mitgewirkt haben, und kein vom Trainingsdatensatz unabhängiges Testset gebildet worden ist. Die Gesamttrefferrate mag zwar beim Modell ohne Standardabweichungsregel höher sein, diese Beurteilung basiert allerdings ausschließlich auf denjenigen Daten, die bei der Berechnung des Modells verwendet wurden. Bessere Aussagen würden über Testdaten erzielbar sein, die keinerlei Beitrag zur Erstellung des Baummodells geliefert haben. In diesem Zusammenhang ist eine validere Beurteilung der Prognosequalität hinsichtlich „neuer“ Instanzen zum Beispiel über die Betrachtung einer Kreuzvalidierung (dies findet im Kapitel 8 statt) möglich.

7.2) Random Forests

Wie bereits bei der Vorstellung der Methoden erwähnt wurde, gibt es zwischen Klassifikationsbäumen und Random Forests einen starken Zusammenhang. Auch bei den Random Forests steht das iterative Generieren von Verzweigungen, ausgehend von einem Wurzelknoten, im Vordergrund, doch hier wird insgesamt eine Vielzahl von Bäumen aus einer Reihe von Trainingssets, welche mittels Bootstrap erzeugt werden, herangezogen.

Die Parameter, deren Wahl für die hier durchgeführten Modellberechnungen bedeutend ist, sind die Anzahl der Bäume, aus denen der Random Forest schließlich bestehen soll, sowie die Anzahl der Variablen, die bei jedem Split innerhalb der einzelnen Bäume zufällig ausgewählt wird, diese Variablen stehen folglich als Kandidaten für die Teilung des Baumes an dieser Stelle zur Verfügung (ist beispielsweise die Anzahl gleich 3, stellen 3 zufällig gewählte Variablen Kandidaten dar). Welche dieser gewählten Variablen schließlich zum Zug kommt, wird über ein Importance-Maß entschieden.

Da die Prognosequalität sich typischerweise mit ansteigender Baumzahl erhöht, soll an dieser Stelle die Baumzahl fix bei einer ansprechenden Größe (500 Bäume) gehalten werden. Im Gegensatz dazu soll die Variablenanzahl m, die bei jedem Split zur Kandidatenfindung dient, unterschiedlich betrachtet werden: zunächst wird m=3 betrachtet, dann m=5, und schließlich

m=10. Aufgrund der schlechten Darstellbarkeit bzw. schwer möglichen direkten Interpretierbarkeit von Random Forests wird das Hauptaugenmerk hier auf Ergebnissen in Tabellenform gerichtet (Konfusionsmatrizen, Variablenimportance).

Für die jeweiligen Werte von m zeigen sich nun folgende Resultate:

m = 3:

OOB-Fehlerrate: 36.59%
Gesamt-Trefferrate: 63,4%

Konfusionsmatrix:

Echter Wert/Prognosewert	1	2	3	4	5	Trefferrate
1	132	46	2	0	0	0,733
2	24	193	34	7	0	0,748
3	1	52	85	21	0	0,535
4	0	6	45	42	1	0,447
5	0	2	10	11	2	0,080

Tabelle 7.5.: Konfusionsmatrix für m=3

m=5:

OOB-Fehlerrate: 35.34%
Gesamt-Trefferrate: 64,7%

Konfusionsmatrix:

Echter Wert/Prognosewert	1	2	3	4	5	Trefferrate
1	134	46	0	0	0	0,744
2	25	191	37	5	0	0,740
3	1	47	89	22	0	0,560
4	0	5	41	48	0	0,511
5	0	1	9	14	1	0,040

Tabelle 7.6.: Konfusionsmatrix für m=5

m=10:

OOB-Fehlerrate: 35.20%
Gesamt-Trefferrate: 64,8%

Konfusionsmatrix:

Echter Wert/Prognosewert	1	2	3	4	5	Trefferrate
1	139	40	1	0	0	0,772
2	25	186	41	6	0	0,721
3	1	45	90	23	0	0,566
4	0	4	39	48	3	0,511
5	0	1	8	15	1	0,040

Tabelle 7.7.: Konfusionsmatrix für m=10

ROC-Kurven als grafische Darstellung von Sensitivität und Spezifität:

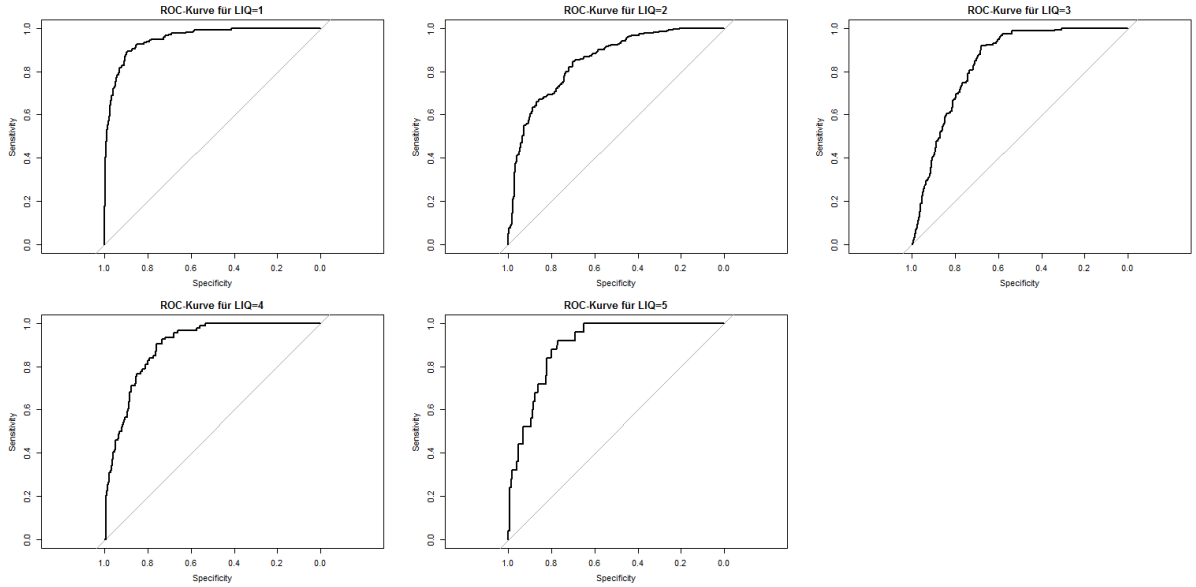


Abbildung 7.5.: Darstellung von ROC-Kurven für jede Liquiditätskategorie

Area under the curve	
LIQ=1	0,9542
LIQ=2	0,8499
LIQ=3	0,8401
LIQ=4	0,8963
LIQ=5	0,8988

Tabelle 7.8.: „Area under the Curve“ – Maß für jede Liquiditätskategorie

Variablen-Importance (für m=3/5/10):

Variable	Mittl. Verbesserung Trefferrate			Mittlere Verbesserung Gini-Index		
	m = 3	m=5	m=10	m=3	m=5	m=10
Bid-Ask-Ratio	30,82	32,34	29,62	62,18	64,52	55,31
VOLV	15,45	14,57	14,54	36,26	32,34	26,00
Turnover	50,93	61,66	89,45	101,25	119,06	148,50
Rating	5,602	5,919	4,913	15,22	12,72	10,13
Anzahl Listungen	20,12	20,05	19,22	27,81	23,81	18,73
STD Return	3,264	3,366	2,139	26,09	22,49	20,38
Sektor	2,428	3,022	4,708	33,33	34,02	34,88
Land	29,71	33,86	37,83	60,62	67,18	75,75
CPI	12,43	13,01	12,95	29,17	27,27	24,92
Industrieproduktion	3,240	5,097	4,967	22,87	20,40	17,01
REER	5,061	4,461	3,475	23,60	21,06	19,43
Money Supply	4,638	3,990	5,651	23,33	21,96	20,82
Kurzfr. Zinssatz	2,364	3,175	4,880	21,35	18,89	16,80
Langfr. Zinssatz	2,130	2,486	2,871	22,45	20,62	19,39
Economic Sentiment	4,255	2,315	6,762	23,21	22,12	20,65

Tabelle 7.9: Variablen-Importance für n=500, m=3

Die Ergebnisse der Random Forests unterscheiden sich geringfügig von denen der Klassifikationsbäume. Die im Vordergrund stehenden Variablen sind auch hier der Turnover, das Land des Unternehmens sowie die Bid-Ask-Ratio. Von den makroökonomischen Variablen spielt als einzige der Consumer Price Index eine vordergründige Rolle.

Die Trefferraten liegen in etwa bei 64%, das bedeutet in Summe etwas weniger vom Modell korrekt prognostizierte Instanzen. Ein größer gewählter Wert von m führt dabei zu etwas besseren Ergebnissen. Die ROC-Kurven und das Gütemaß, das als „Area under the curve“ bezeichnet wird, weisen ebenso wie die Konfusionsmatrizen auf eine bessere Prognosegenauigkeit des Modells bei den höheren Liquiditätskategorien hin (insbesondere für die erste Kategorie).
Aber wie zuvor muss auch hier darauf hingewiesen werden, dass die Prognose am Trainingsset durchgeführt wurde, doch lediglich die Vorhersage bei Werten, die bei der Modellerstellung nicht beteiligt waren, kann eine echte Aussage über die Qualität eines derartigen Verfahrens liefern.

7.3) Support Vector Machines

Das Verfahren der Support Vector Machines zeichnet sich unter anderem dadurch aus, dass es hier zwei verschiedene Parameter sind, welche bei der Erstellung von Prognosen einen großen Einfluss haben. Dazu zählt einerseits eine Kostenvariable, welche den Grad der Pönalisierung angeben kann, wenn Datenpunkte durch die Trennfunktion nicht im richtigen Bereich (zu Instanzen derselben oder einer ähnlichen Klasse) eingeordnet wurden. Von großer Bedeutung ist aber auch die Wahlmöglichkeit einer passenden Kernfunktion für die Klassifizierung. Je nachdem welche Eigenschaften ein Datensatz besitzt, kann sich ein linearer, polynomialer oder radialer Kern (deren Definitionen wurden in Kapitel 5.4 betrachtet) als günstig für die weitere Durchführung der Analyse herausstellen. In diesem Abschnitt wird der Fokus auf den radialen Kern sowie einer Kostenfunktion von 1 (geringe Bestrafung bei ungünstig liegenden Datenpunkten) gelegt.

Für diese Kombination ergeben sich bei der Klassifikation der Liquidität folgende Resultate:

Anzahl von Support-Vektoren: 683

Konfusionsmatrix:

Echter Wert/Prognosewert	1	2	3	4	5	Trefferrate
1	82	98	0	0	0	0,456
2	5	245	8	0	0	0,950
3	1	120	37	1	0	0,233
4	0	56	28	10	0	0,106
5	0	10	12	3	0	0,000

Tabelle 7.10.: Konfusionsmatrix für die Support-Vector-Machine-Methode

Gesamt-Trefferrate: 51,5%

ROC-Kurven als grafische Darstellung von Sensitivität und Spezifität:

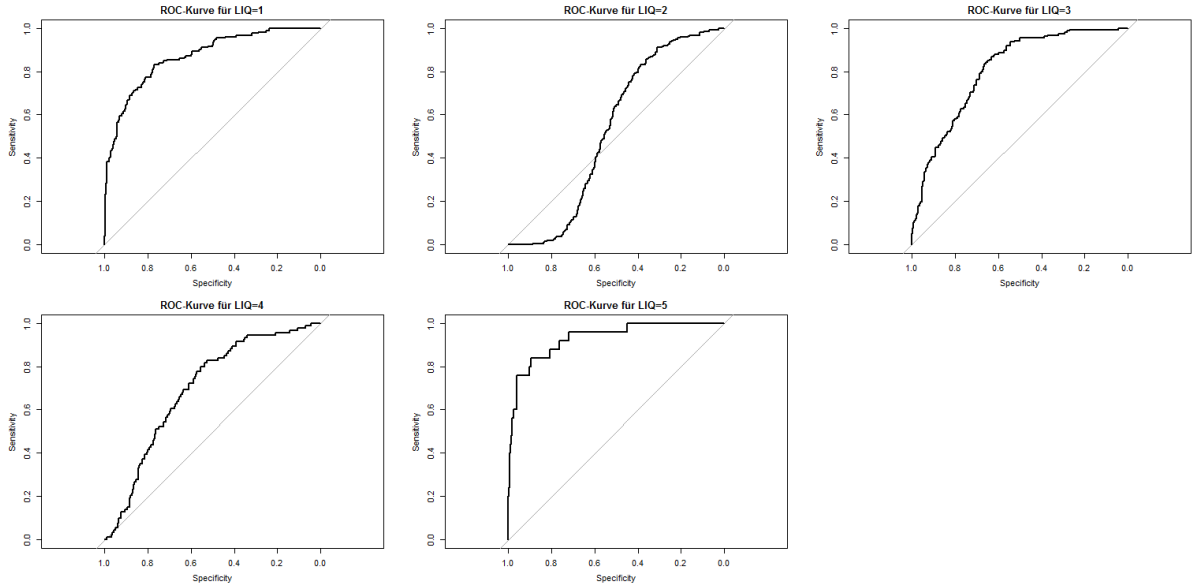


Abbildung 7.6.: Darstellung von ROC-Kurven für jede Liquiditätskategorie

Area under the curve	
LIQ=1	0,8709
LIQ=2	0,5292
LIQ=3	0,8064
LIQ=4	0,6971
LIQ=5	0,9314

Tabelle 7.11.: „Area under the Curve“ – Maß für jede Liquiditätskategorie

Die Support Vector Machines liefern ein schwächeres Ergebnis als die ersten beiden Methoden, wenn wiederum nur eine Evaluation anhand der Trainingsdaten herangezogen wird. Auch hier soll die Betrachtung einer Kreuzvalidierung bessere Schlussfolgerungen ermöglichen. Auffallend ist bei obiger Konfusionsmatrix vor allem die deutlich geringere Trefferrate bei den hochliquiden Wertpapieren im Vergleich zu den Baumstruktur-Verfahren.

Da Support Vector Machines ihren Ursprung in der Geometrie haben, wäre eine anschauliche Darstellung der Supportvektoren und der Klassentrennung anhand deren Funktionsweise von großer Bedeutung. Es bietet sich die Möglichkeit, über eine multidimensionale Skalierung die Distanzen zwischen den Beobachtungen bezüglich der betrachteten Prädiktorvariablen zu aggregieren und sie in Form einer zweidimensionalen Punktwolke darzustellen, und zu dieser Darstellung Punkte hinzuzufügen, welche die Supportvektoren repräsentieren. In dieser Analyse ergibt sich folgende Darstellung:

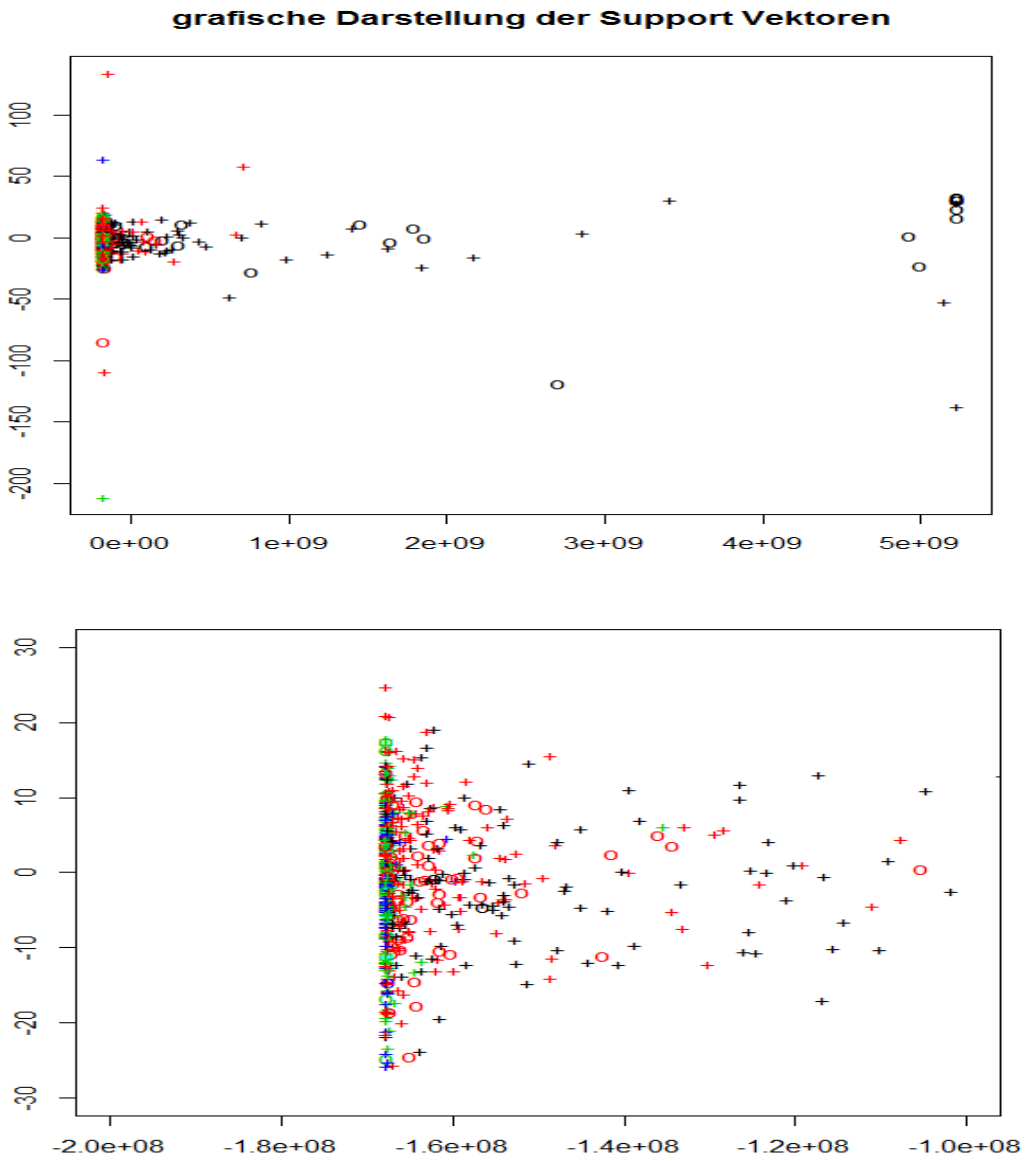


Abbildung 7.7: grafische Darstellung der Support Vector Machines. Die Supportvektoren werden dabei durch Kreuze, die Klassen der Liquidität durch Farben dargestellt. Die zweite Grafik stellt einen vergrößerten Abschnitt des linken zentralen Bereiches der ersten Grafik dar, auf den sich die Daten konzentrieren

7.4) Clusteranalyse als Beispiel für nicht-überwachtes Lernen

Im Gegensatz zu den bisher angewendeten Verfahren liegt der Fokus bei der Anwendung einer Clusteranalyse nun nicht direkt bei der Klassifikation der Liquiditätskategorie, sondern bei der Auffindung von Gruppen, die aus Instanzen (Wertpapieren) bestehen, welche eine homogene Struktur aufweisen. Die Zielsetzung, welche dabei an vorderster Stelle stehen sollte, ist die Untersuchung und Beschreibung von Beziehungen zwischen den Instanzen,

welche sich hinsichtlich ihrer Variablenwerte und deren Eigenschaften ergeben. Welche Variablen weisen bei den Instanzen in den jeweiligen Clustern Ähnlichkeiten auf? Und welche Variablen liefern einen wesentlichen Beitrag für die Generierung der Cluster?
In Kapitel 5 wurden grundlegende Konzepte, die bei der Clusteranalyse zum Tragen kommen, bereits beschrieben. Basierend auf einer generalisierten Form des Gower-Maßes wird in weiterer Folge die Berechnung einer Distanzmatrix durchgeführt. Dieses Maß wird deshalb gewählt, weil es eine simultane Verwendung von unterschiedlich skalierten Variablen erlaubt. Es lässt sich in folgender Form darstellen [URL: https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/daisy.html [Zugriff am 20.10.2014]]:

$$d_{ij} = d(i,j) = \frac{\sum_{k=1}^{p} \omega_k \delta_{ij}^{(k)} d_{ij}^{(k)}}{\sum_{k=1}^{p} \omega_k \delta_{ij}^{(k)}} \quad (40)$$

Dabei ist d_{ij} das gewichtete Mittel von $d_{ij}^{(k)}$, wobei $d_{ij}^{(k)}$ die Gower-Distanz für die k-te Variable in ihrer gewöhnlichen Form darstellt (siehe Formel 36). Insbesondere ist hier für nominal und binär skalierte Variablen der Wert von $d_{ij}^{(k)}$ gleich 0, wenn die Variablenwerte bei der i-ten und j-ten Instanz gleich sind, ansonsten ist $d_{ij}^{(k)}$ gleich 1. Einen weiteren günstigen Vorteil besitzt diese Methode hinsichtlich der Berücksichtigung fehlender Werte – hier kann das Gewicht $\delta_{ij}^{(k)}$ im Falle fehlender Werte einfach auf 0 festgesetzt werden.

Als Clustermethode wird der k-Medoids-Algorithmus herangezogen („Partitioning around medoids“ – kurz „PAM“), welcher eine robuste Variante des k-Means-Algorithmus darstellt. Zwar wird hier, genauso wie bei letzterer Methode, die Minimierung der Abstände zwischen Datenpunkten, die schließlich in einen Cluster fallen, anvisiert, doch hier wird das Clusterzentrum allgemein durch einzelne Datenpunkte gebildet (die als „medoids“ bezeichnet werden) und nicht durch Berechnung eines Cluster-Schwerpunkts aus Beobachtungen, die diesem Cluster zugeordnet werden. Dabei kann eine größere Anzahl verschiedener Distanzmatrizen verwendet werden [Kaufman & Rousseeuw 1990].

Eine häufig angewendete Form dieses Algorithmus sieht wie folgt aus [Theodoridis & Koutrumbas 2006]:

Schritt 1: Wähle zufällig k der n Datenpunkte ohne Zurücklegen als „medoids“
Schritt 2: Ordne jeden Datenpunkt dem nächsten „medoid“ zu (bezüglich einer gewählten Distanzmatrix)
Schritt 3: Für jedes „medoid“ m sowie für alle „non-medoids“ o: Tausche m und o aus und berechne Kosten für jeweilige Auswahl (Kosten hängen dabei typischerweise mit Distanzen zusammen)
Schritt 4: Wähle die Kombination mit den niedrigsten Kosten
Schritt 5: Wiederholung der Schritte 2-4, bis Stoppkriterium erfüllt ist

Es stellt sich als sinnvoll heraus, die Variable „Sektor“ aufgrund ihrer Eigenschaften (nominelle Variable, viele Instanzen mit gleichen Kategorien und daher keine Distanz

hinsichtlich dieser Variable) zunächst nur in einer deskriptiven Form in diese Analyse miteinzubeziehen, also nur eine Auswertung von Häufigkeiten bei den erstellten Clustern voranzutreiben, jedoch keine Einbeziehung bei der eigentlichen Segmentierung zu betrachten. Da für die Variable „Land“ ähnliches gilt, wird auch sie, trotz der größeren Rolle bei den vorherigen Auswertungen, bei den ersten Berechnungen zunächst nur als deskriptives Merkmal betrachtet. Auch die Liquidität, als abhängige Variable, wird vorerst nur in dieser Form einbezogen.

Im Zuge der Anwendung der oben beschriebenen Mehodik wird in einem ersten Schritt erwogen, eine fixe, vorab gewählte Zahl an Clustern zu generieren, um allgemeine Erkenntnisse über Anwendbarkeit und Güte des Verfahrens zu gewinnen. Als Bedingung wird dabei jedoch betrachtet, dass sich die Instanzen relativ gleichmäßig auf die Cluster aufteilen. Wird nun versucht, eine Aufteilung der Daten in 5 Gruppen zu erzielen, so zeigen sich folgende Resultate:

Clustereigenschaften

Cluster	Größe	Rel. Größe	Anzahl LIQ = 1	Anzahl LIQ = 2	Anzahl LIQ = 3	Anzahl LIQ = 4	Anzahl LIQ = 5	Mittlere Liquidität	Mittlere Anzahl Börsenlistungen
1	179	19,84%	85	66	13	6	2	1,69	9,66
2	232	25,72%	29	94	59	26	1	2,41	5,53
3	166	18,40%	51	30	32	8	6	2,12	2,14
4	184	20,40%	15	54	32	22	4	2,57	1,59
5	141	15,63%	0	14	23	32	12	3,52	1,54

Tabelle 7.12: Grundlegende Eigenschaften der resultierenden Cluster, Teil 1

Cluster	Instanzen mit A-Ratg	Instanzen mit B-Ratg	Instanzen mit C-Ratg	Instanzen ohne Rtg	Häufigste Sektoren	Häufigste Länder
1	69	96	0	14	Fin. Services(27), Industrials(24)	GER (46), GBR (28), USA (19)
2	33	130	9	60	Consumer Cyclical(48), Industrials(33)	ITA(50), IRL(23), SPA(21)
3	27	73	1	65	Industrials(38) Consumer Cyclical(29),	CHN(50), VEN(14), JPN(12)
4	12	80	2	90	Industrials(40), Consumer Cyclical(36)	TUR(51), EGY(27), SAF(19)
5	6	47	3	85	Industrials(27), Financial Services(26)	BUL(49), VEN(15), IND(14)

Tabelle 7.13: Grundlegende Eigenschaften der resultierenden Cluster, Teil 2

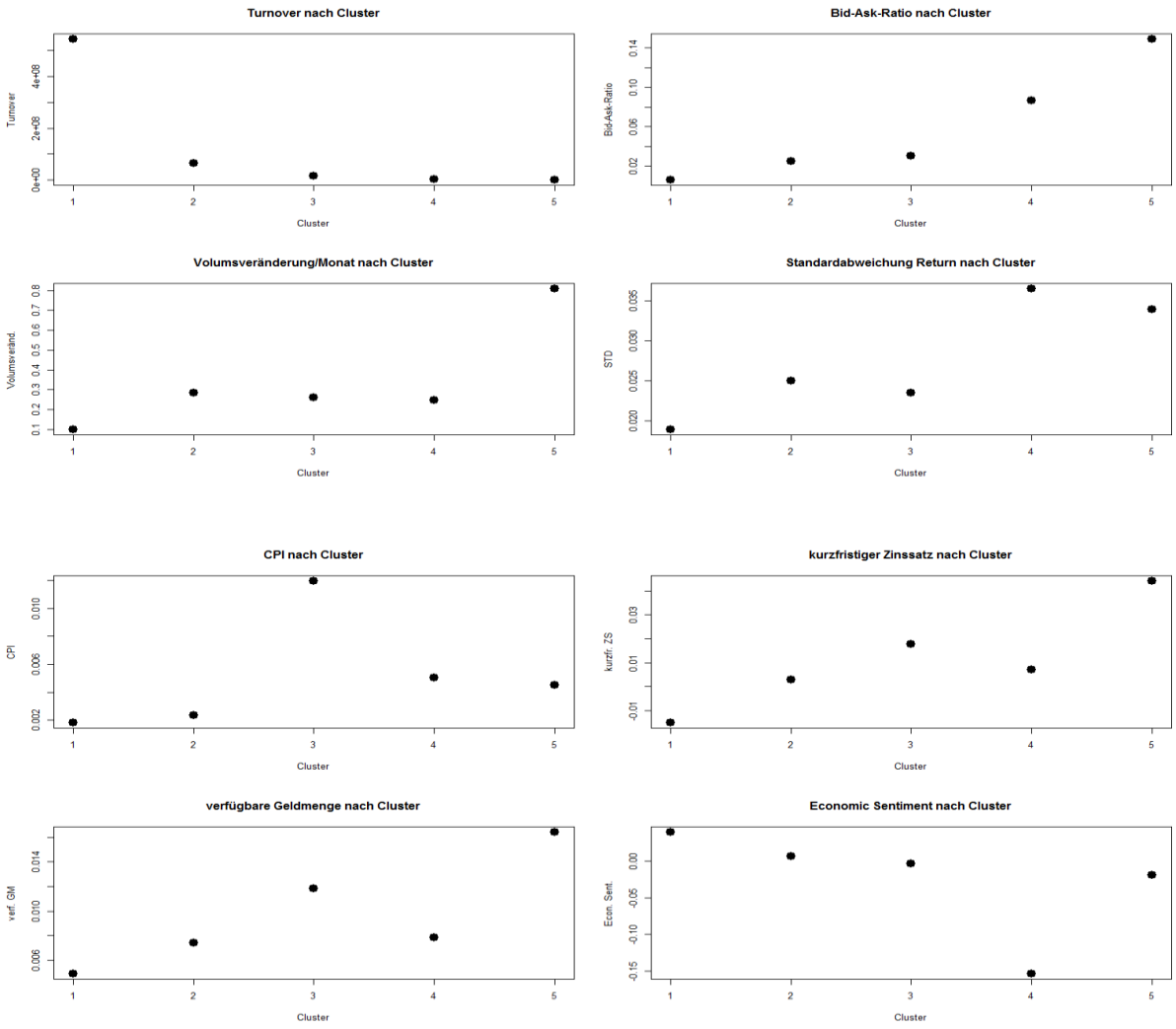


Abbildung 7.8: Darstellung der Veränderung wesentlicher Variablenmerkmale der 5 Cluster

Als Resultat der Clusteranalyse zeigen sich also 5 Cluster mit teilweise ähnlichen, teilweise aber doch sehr unterschiedlichen Merkmalen. Alle von ihnen weisen, wie für den Anfang gewünscht, eine recht ähnliche Größe auf (zwischen 141 und 232 Beobachtungen).

Es sind mehrere Variablen, deren Werte den Aufbau der jeweiligen Clustern mitbestimmt haben dürften: So zeichnet sich der erste Cluster durch eine recht hohe Liquidität, sehr viele Börsennotationen, durch Ratingnoten im A- oder B-Bereich und einem sehr hohen Turnover aus (diese Merkmale haben eine deutlich stärkere Ausprägung als bei den übrigen Gruppen). Die Cluster 2 und 3 liegen hinsichtlich Liquidität noch in einem überdurchschnittlich guten Bereich, bei Cluster 2 kann man dabei die höhere Anzahl an Börsennotationen (im Schnitt um 4 weniger als bei Cluster 1) und gute Ratings (vorwiegend B-Ratings, aber auch einige mit fehlenden Angaben) hervorheben. Bei Cluster 3 fallen besonders die höheren Werte der makroökonomischen Variablen (CPI, verfügbare Geldmenge) im Vergleich zu den ersten beiden Clustern auf. Die vierte Gruppe zeigt bereits eine etwas schwächere Liquidität, bei den Prädiktorvariablen zeigen sich deutlich höhere Werte bei der Standardabweichung des Returns (knapp 0,04 im Vergleich zu 0,02-0,025 in den ersten drei Gruppen) und bei der Bid-

Ask-Ratio. Aber die auffallendste Gruppe ist die letzte, bei der eine deutlich niedrigere Liquidität (im Schnitt 1-2 Kategorien tiefer als sonst) und stark überdurchschnittliche Werte bei der Bid-Ask-Ratio, bei der Veränderung des gehandelten Volumens pro Monat, der verfügbaren Geldmenge sowie des kurzfristigen Zinssatzes erkennbar sind. Es scheint also durchaus möglich, dass ein Zusammenhang zwischen den erwähnten Prädiktorvariablen und der in den Clustern auftretende Liquidität besteht, über welchen sich vor allem hochliquide und gering liquide Wertpapiere voneinander trennen lassen.

Eine Frage, die man sich stellen sollte, ist jene, ob die hier gewählte Clusterzahl als gute Wahl angesehen werden kann, und wie dies zu überprüfen ist. In Kapitel 8.3 wird versucht, über spezielle Verfahren, mit denen die Qualität der Clusterauswahl evaluiert werden kann, eine optimale Anzahl von Clustern für die Analyse zu finden.

8) Validierung und Optimierung der angewendeten Verfahren

8.1) Validierung der überwachten Lernverfahren im Kontext einer Parametervariation

Im Mittelpunkt des letzten Abschnittes stand die Anwendung der in Kapitel 5 vorgestellten Verfahren des maschinellen Lernens unter Verwendung einer bestimmten, meist als Standardvariante betrachteten Kombination von Parametern. Doch dabei kann keineswegs garantiert werden, dass diese die bestmögliche Auswahl für die Problemstellung ist, welche man bearbeitet. Um nun herausfinden zu können, ob sich eine alternative Wahl als günstiger erweist, kann auf verschiedene Evaluationsmethoden zurückgegriffen werden, wie sie unter anderem in Kapitel 5.1 erwähnt wurden.

Für die Verfahren des maschinellen Lernens, die in die Kategorie des „supervised learnings“ fallen, bietet die sogenannte k-fache Kreuzvalidierung eine Möglichkeit, in einer hochwertigen Form die Brauchbarkeit diverser Parametervarianten abzutesten. Da das Verfahren in Kapitel 5.1 bereits näher beschrieben wurde, liegt das Augenmerk in diesem Abschnitt auf der Präsentation von Ergebnissen, die unter den jeweiligen Abstimmungen resultieren. In diesem Abschnitt wird eine 10-fache Kreuzvalidierung für alle betrachteten Verfahren fünf Mal berechnet.

Als Parametervarianten werden nun für die drei Verfahren des maschinellen Lernens (Klassifikationsbäume, Random Forests, Support Vector Machines) folgende Varianten zur Validierung herangezogen:

-) Klassifikationsbäume: Berechnung ohne bzw. mit Berücksichtigung der 1-Standardabweichungs-Regel (in Abbildung 8.1.: cv.rpart.v1 bzw. cv.rpart.v2)

-) Random Forests: 50/500/800 Bäume (bei fixer Zahl von 10 Variablen bei Split, cv.rf.v1-cv.rf.v3 in Abbildung 8.1.), 3/5/10 Variablen als Kandidaten bei Split (bei fixer Zahl von 500 Bäumen, cv.rf.v1-cv.rf.v3 in Abbildung 8.2.)

-) Support Vector Machines: radialer/polynomialer (Grad 3)/sigmoider Kern (bei fixiertem Kostenparameter auf 1, cv.svm.v1-cv.svm.v3 in Abbildung 8.1.); Kostensetzung auf 1/10/50 (bei fixer Verwendung von radialem Kern, cv.svm.v1-cv.svm.v3 in Abbildung 8.2.)

Insgesamt werden somit 14 verschiedene Varianten der Verfahren getestet. Zusätzlich wird die Validierung auch auf ein Vergleichsmodell angewendet, dies ist in diesem Fall eine gewöhnliche logistische Regression (unter Verwendung von Dummy-Variablen für die nominal bzw. ordinal skalierten Variablen, in Abbildung 8.2. ist sie als cv.glm.v1 gekennzeichnet). Als Beurteilungskriterium wird die bereits in den vorangegangenen Kapiteln

verwendete Trefferrate (Accuracy) verwendet, die den Anteil an korrekt klassifizierten Instanzen angibt.

Für die auf der letzten Seite erwähnten Modellvarianten ergibt sich nun folgende grafische Darstellung der Werte für die Accuracy:

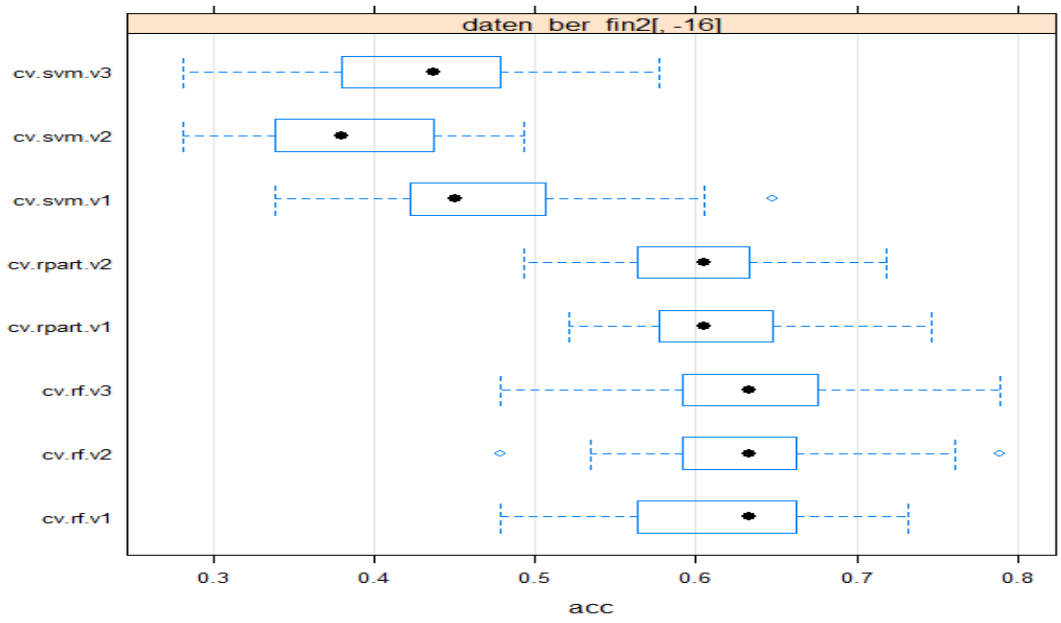


Abbildung 8.1.: grafische Übersicht der Ergebnisse der Kreuzvalidierung, Teil 1

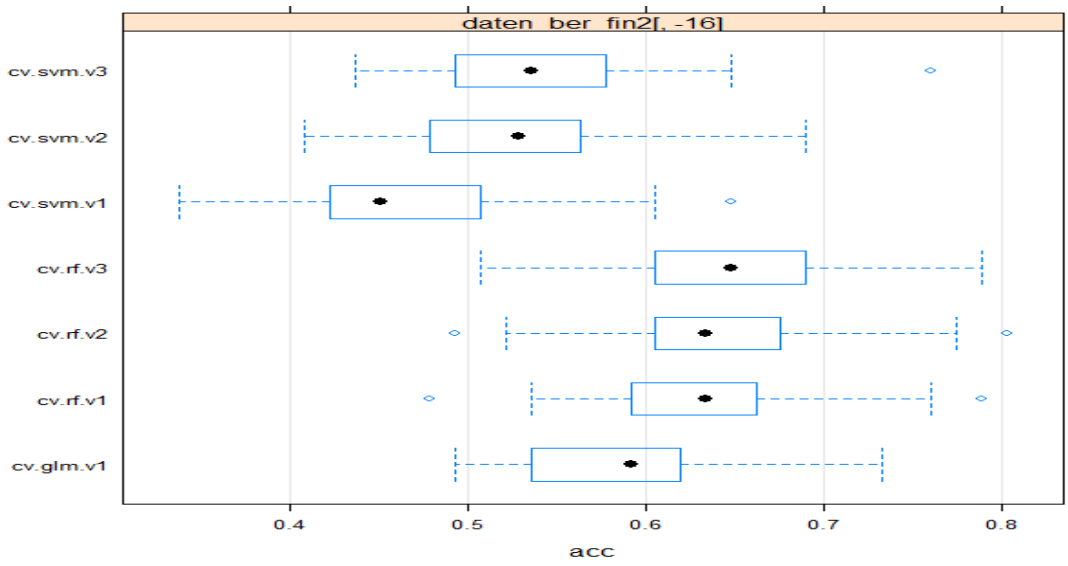


Abbildung 8.2.: grafische Übersicht der Ergebnisse der Kreuzvalidierung, Teil 2

Es zeigen sich also bei einigen der gewählten Modellvarianten sehr ordentliche Ergebnisse. Am besten schneiden die Varianten der Random Forests ab, bei denen sich die Trefferrate im Mittel bei etwa 0,63 bewegt. Zwar treten stärkere Schwankungen bei den jeweiligen Auswertungen der Modellvarianten auf (bei einer Mittelung der 6 Varianten, die für die Random Forests betrachtet wurden, liegt das 5%-Quantil (bezogen auf alle 50 Iterationen der

Validierung) bei etwa 0,51, das 95%-Quantil bei 0,77), doch auch in diesem Kontext liegen die Random Forests im Spitzenfeld. Nicht weit hinter diesem Modell folgen die Varianten der herkömmlichen Klassifikationsbäume, diese weisen eine mittlere Trefferrate von 0,6 auf, wobei sich hier die Spannweite als etwas niedriger als bei den Random Forests herausstellt (das 5%-Quantil ist im Mittel mit 0,52 knapp über dem des ersten Modells). Als allgemein etwas schwächer erweist sich das logistische Vergleichsmodell mit einer mittleren Trefferrate von 0,59, das klare Schlusslicht sind hier die Support Vector Machines, wo die Trefferrate im Schnitt unter 0,5 bleibt. Die Spannweite liegt bei den letzten beiden Varianten im selben Niveaubereich wie bei den Baumstrukturmethoden.
Insgesamt zeigt sich aber eine passable Prognosetauglichkeit der jeweiligen verwendeten Modelle, vor allem Baumstrukturverfahren dürften bei einer Problemstellung wie in der gegebenen Form von großem Nutzen sein.

Als eine sehr nützliche grafische Darstellung erweist sich, ähnlich wie in den Kapiteln 7.1-7.3, die Verwendung von ROC-Kurven zur Visualisierung von Sensitivität und Spezifität. Hier kann diesen Darstellungen eine höhere Bedeutung zugewiesen werden als im letzten Kapitel, da im Gegensatz zu vorhin hier die Prognosen bei Instanzen berechnet werden, die bei der Erstellung des Modells keine Rolle gespielt haben.

Es wird nun für die drei Verfahren jeweils für jeden Durchlauf im Zuge der Validierung die ROC-Kurve ermittelt und diese wird gemeinsam mit jenen der anderen Durchläufe grafisch dargestellt. Die rote Kurve zeigt das Ergebnis an, wenn alle grafischen Ergebnisse gemittelt werden (dies entspricht einer Zusammenlegung aller Prognosen bei der Validierung der Lernmethode und darauffolgender Berechnung).

Die Ergebnisse sehen bei Verwendung der „one vs. all“ – Variante, die bereits in Kapitel 7 benützt wurde, für die einzelnen Liquiditätskategorien wie folgt aus:

Classification Trees:

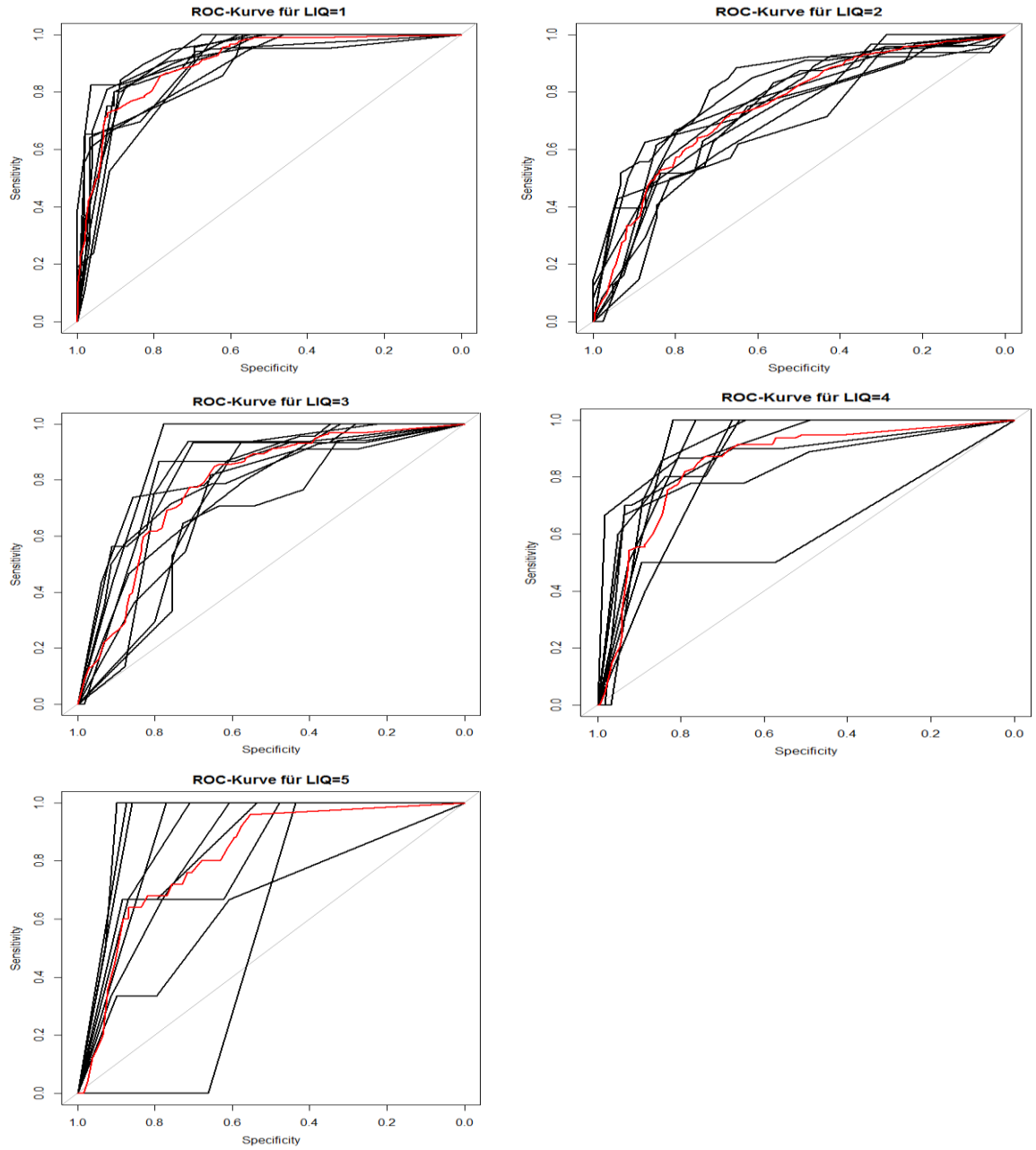


Abbildung 8.3.: ROC-Kurven für alle Läufe der Kreuzvalidierung und gruppiert nach Liquiditätskategorie (rot: Ensemble-Mittel)

Klasse	Area under the curve
LIQ = 1	0,9044
LIQ = 2	0,7485
LIQ = 3	0,7891
LIQ = 4	0,8531
LIQ = 5	0,8274
Mittelwert	**0,8245**

Tabelle 8.1.: „Area under the Curve“ – Maß für jeweilige Liquiditätskategorie

Random Forests:

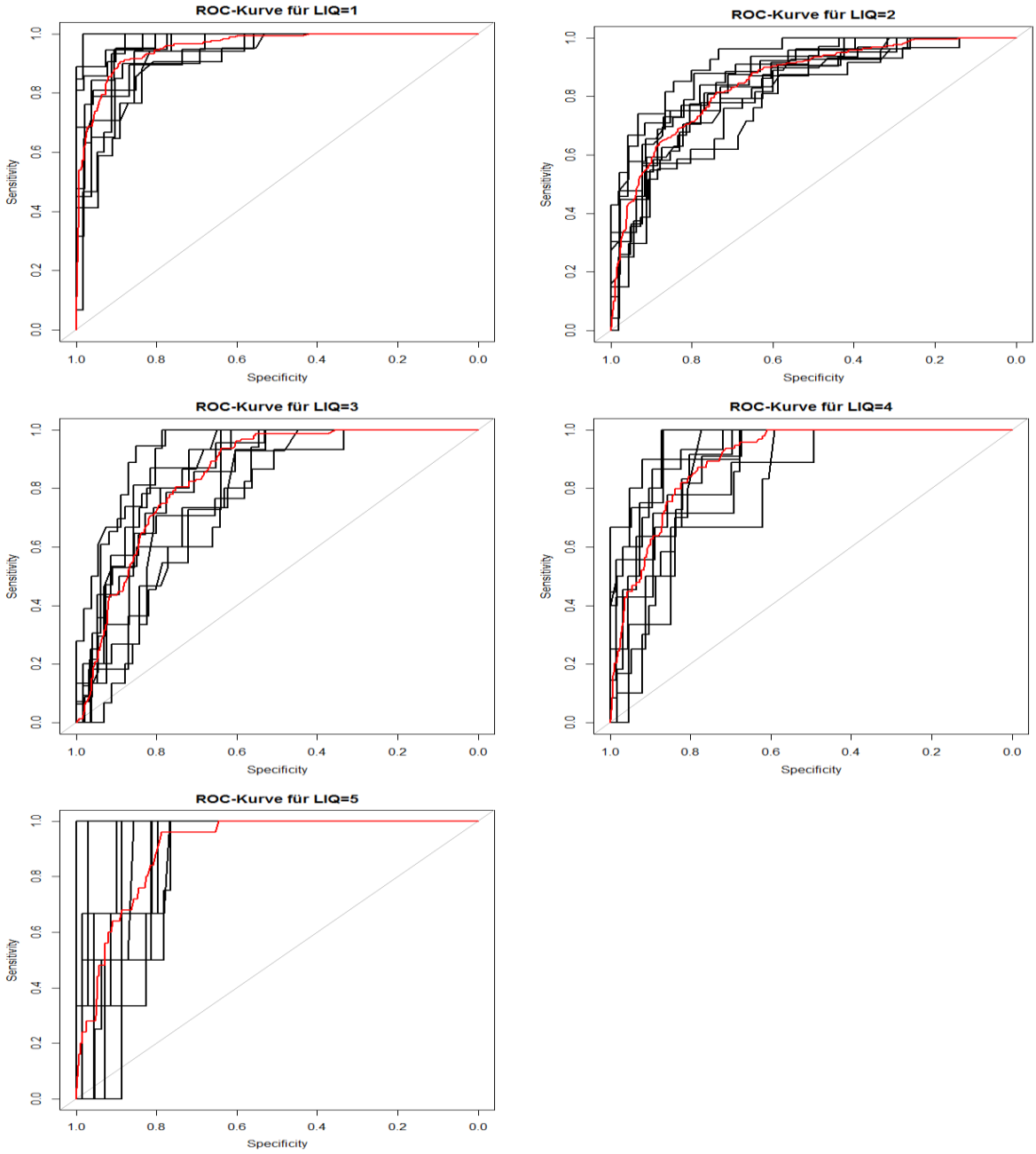


Abbildung 8.4.: ROC-Kurven für alle Läufe der Kreuzvalidierung und gruppiert nach Liquiditätskategorie (rot: Ensemble-Mittel)

Klasse	Area under the curve
LIQ = 1	0,9564
LIQ = 2	0,8522
LIQ = 3	0,8400
LIQ = 4	0,8985
LIQ = 5	0,9135
Mittelwert	**0,8921**

Tabelle 8.2.: „Area under the Curve“ – Maß für jeweilige Liquiditätskategorie

Das Bild, welches sich durch den Vergleich der ROC-Kurven ergibt, ähnelt jenem, welches die Boxplots zuvor geliefert haben.

Support Vector Machines:

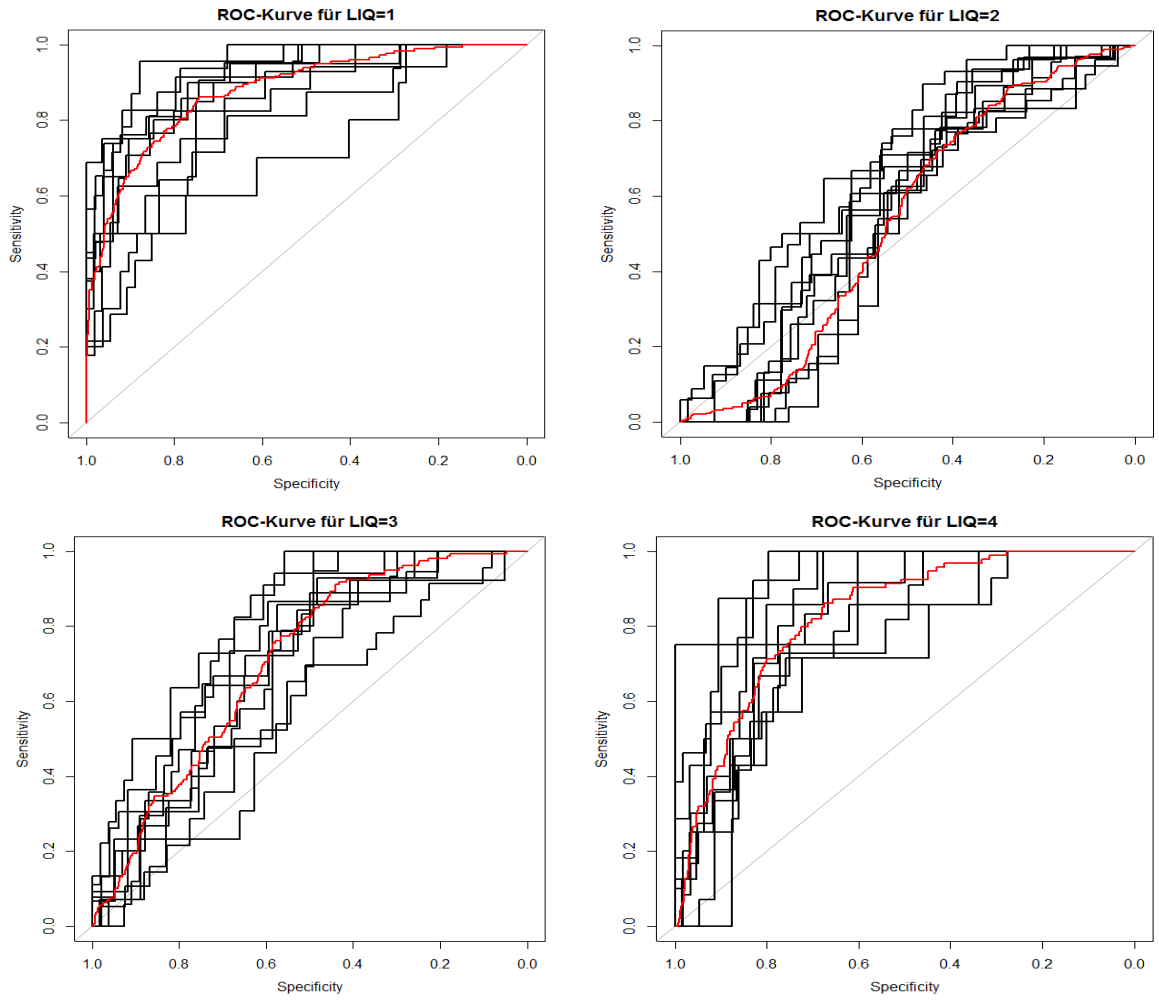


Abbildung 8.5.: ROC-Kurven für alle Läufe der Kreuzvalidierung und gruppiert nach Liquiditätskategorie (rot: Ensemble-Mittel)

Für den fünften Liquiditätsgrad konnte keine Grafik ermittelt werden, da in zumindest einer Iteration keine Beobachtung zur Verfügung stand, welche als „Fall“ interpretiert werden konnte.

Klasse	Area under the curve
LIQ = 1	0,8696
LIQ = 2	0,5326
LIQ = 3	0.7033
LIQ = 4	0.8424
LIQ = 5	-
Mittelwert	**0,7370**

Tabelle 8.3.: „Area under the Curve“ – Maß für jeweilige Liquiditätskategorie

Die Support-Vector-Machines weisen also bei Beurteilung der Ergebnisse dieses Kapitels weniger hochwertige Ergebnisse auf als die Baumstruktur-Methoden. Wie bei den Boxplot-Darstellungen stellen davon die Random Forests die beste Methode dar, hier sogar mit einem größeren Abstand zu den übrigen Verfahren.

8.2) Kombination überwachter Lernverfahren in Form eines Majority Votings

Nachdem im vorangegangenen Abschnitt die Performance der überwachten Lernmethoden in einer detaillierten Form abgeprüft wurde, soll hier eine Variante des Ensemble Learnings getestet werden, welche eine Kombination dieser angewendeten Methoden darstellt. Dies soll wie folgt bewerkstelligt werden:

Nach Berechnung der drei Verfahren (in unabhängiger Form unter Anwendung vorab festgelegter Parameter) wird ein Voting durchgeführt, bei welchem die von den jeweiligen Verfahren prognostizierten Klassen bezüglich jeder Instanz miteinander verglichen werden. Schließlich wird jene Klasse am Ende als Prognosewert für die Instanz gewählt, welche bei der Mehrheit der Modelle resultierte („Majority Voting“). Hier bedeutet dies, dass, wenn sich bei zumindest zwei Verfahren die gleiche Klasse bei der Prognose ergibt, diese Klasse folglich den Prognosewert darstellt. Sollte sich bei allen drei Verfahren eine andere Klasse ergeben, so erhält die Klasse jenes Verfahrens den Vorzug, welches sich im Kapitel 8.1 als das hochwertigste Verfahren herausgestellt hat. In diesem Fall ist dies das Verfahren der „Random Forests“.

Ziel dieses Abschnittes ist es, herauszufinden, ob mithilfe einer Kombination der in separater Form bereits validierten Lernverfahren eine weitere Steigerung der Prognosequalität erzielbar ist, und besonders, ob Schwachstellen der Verfahren, die sich bislang gezeigt haben, bei gemeinsamer Anwendung behoben oder zumindest abgemindert werden können.

Die gewählten Parameter für die Verfahren stellen hier jene Varianten dar, die sich bei der Validierung in Kapitel 8.1 als vielversprechend herausgestellt haben, dies sind:

-) Classification Trees: keine Berücksichtigung der 1-Standardabweichungs-Regel

-) Random Forests: 500 Bäume, Betrachtung von 10 Variablen bei jedem Split

-) Support Vector Machines: radialer Kern, Kosten auf 50 gesetzt

Das Hauptaugenmerk wird nun auf den Vergleich der (mittleren) Trefferraten einer 10-fachen Kreuzvalidierung bei den einzelnen Verfahren (die bereits in Abschnitt 8.1 ermittelt wurden) und beim Ensemble Learning-Verfahren gerichtet. Dabei zeigt sich folgendes Ergebnis:

Verfahren	Mittlere Trefferrate	5% - Quantil	95% - Quantil
Random Forests (500 Bäume, 10 Split-Var)	0,647	0,505	0,790
Classification Trees (ohne Std.Abw.-Regel)	0,605	0,520	0,745
Support Vector Machines (radialer Kern, Kosten= 50)	0,535	0,440	0,650
Majority Voting	0,639	0,542	0,750

Tabelle 8.4.: Vergleich der Ergebnisse aus Abschnitt 8.1. mit denen des Majority Votings

Das Ergebnis für das Ensemble Learning-Verfahren entspricht annähernd jenem der Random Forests, wobei sich hier die Spannweite zwischen dem 5%-Quantil und dem 95%-Quantil als kürzer herausstellt (eine etwas geringere Varianz der Resultate ist erkennbar). Eine echte Steigerung der Qualität kann aber mit dem hier betrachteten Ensemble Learning-Verfahren offenbar nicht erzielt werden.

Weiters wird auch versucht, Ergebnisse von ROC-Kurven im Kontext des Ensemble Learners zu generieren. Da diese allerdings nicht auf prognostizierte Klassen selbst, sondern auf Wahrscheinlichkeiten für diese Klassen beruhen, muss am Beginn eine geeignete Vorgehensweise überlegt werden. Zwei mögliche Varianten sehen wie folgt aus:

1) Mittelung der Wahrscheinlichkeiten, die jedes Verfahren liefert
2) Mittelung der Wahrscheinlichkeiten jener Verfahren, die dieselbe Klasse prognostizieren

Die zweite Variante besitzt eine stärkere Assoziation zur grundlegenden Idee des Majority Votings, deshalb wird diese nun näher beleuchtet. In Tabellenform zeigen sich im Vergleich zu den Ergebnissen aus Abschnitt 8.1 folgende Resultate:

Verfahren	Mittlerer AUC-Wert
Random Forests (500 Bäume, 10 Split-Var)	0,8921
Classification Trees (ohne Std.Abw.-Regel)	0,8245
Support Vector Machines (radialer Kern, Kosten= 50)	0,6602
Majority Voting	0,8300

Tabelle 8.5.: Vergleich der Ergebnisse aus Abschnitt 8.1. mit denen des Majority Votings

Auch hier kann das Ergebnis der Random Forests durch ein Ensemble-Verfahren nicht verbessert werden. Allgemein ist dies ein Anzeichen dafür, dass weder das Verfahren der Classification Trees noch jenes der Support Vector Machines wesentliche Information aus den Daten zur Erklärung der abhängigen Variable nützen kann, die nicht schon in den Random Forests verarbeitet wird.

8.3) Auffinden einer optimalen Clusteranzahl für die angewendete Clustermethode

Auch bei nicht-überwachten Lernmethoden spielt die Optimierung verwendeter Parameter eine große Rolle, vor allem dann, wenn Gewissheit herrschen soll, dass beim angewendeten Modell die Möglichkeiten, die geboten werden, vollständig ausgenutzt werden. Um dies zu garantieren, soll hier getestet werden, wie viele Cluster bei der in Kapitel 7.4 angewendeten Variante einer Clusteranalyse einen optimalen Wert bezüglich eines vorab festgelegten Beurteilungsmaßes liefern. In diesem Fall wird als Beurteilungskriterium der sogenannte Silhouettenkoeffizient herangezogen, welcher für Beobachtungen aussagt, wie gut eine Zuordnung zu jenen Clustern ist, die ihnen am nächsten liegen. Für Beobachtungen o ϵ A ist die Silhouette daher wie folgt definiert [Rousseeuw 1987]:

$$S(o) = \begin{cases} 0, & wenn\ d(A,o) = 0 \\ \frac{d(B,o)-d(A,o)}{\max\{d(A,o),d(B,o)\}}, & sonst \end{cases} \quad (41)$$

Dabei ist $d(A,o) = \frac{1}{n_A}\sum_{a \in A} d(a,o)$ mit n_A als die Anzahl aller Objekte im Cluster a und d(b,o) als der Abstand des Objektes zum nächstgelegenen Cluster b. Die Differenz der Distanzen im Zähler wird mit der maximalen Distanz für Objekte in den jeweiligen Clustern gewichtet, wodurch die Maßzahl zwischen -1 und 1 liegt. Als günstig stellen sich Werte nahe 1 heraus, da daraus resultiert, dass das im Fokus liegende Objekt in einem Cluster liegt.

Eine grafische Darstellung des Silhouettenplots unter Angabe der Resultate für die jeweiligen 5 Cluster zeigt folgendes Bild:

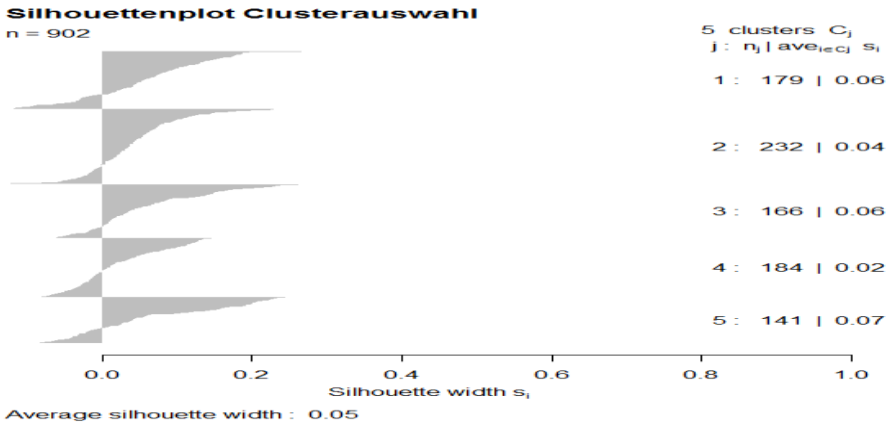


Abbildung 8.6.: Silhouettenplot für die in Kapitel 7.4. berechnete Clusterung

Die Darstellung zeigt vor allem anhand der darin mitgelieferten Silhouetten-Werte, dass eine Aufteilung in die 5 Cluster, welche im Kapitel 7.4 ermittelt wurden, ein Ergebnis liefert, welches relativ weit vom Optimum entfernt zu sein scheint. Die mittlere Silhouettenweite von 0,06 und die Tatsache, dass bei allen Clustern der Wert in einem Bereich zwischen 0,02 und

0,07 liegt, unterstreichen, dass die Struktur des Datensatzes nicht dafür ausgerichtet zu sein scheint, eine passable Aufteilung in 5 Cluster zu erreichen. Dies heißt aber keineswegs, dass eine Clusteranalyse beim vorliegenden Datensatz nicht funktioniert. Es legt vielmehr nahe, zu überprüfen, ob nicht eine andere Wahl der Clusteranzahl zu verbesserten Ergebnissen führen kann. Dies wird im nächsten Schritt anvisiert.
Überprüft wird im folgenden Schritt nun, wie sich die mittlere Silhouettenweite bei einer Clusteranzahl zwischen 2 und 20 ergibt. Dabei kann folgendes Resultat erlangt werden:

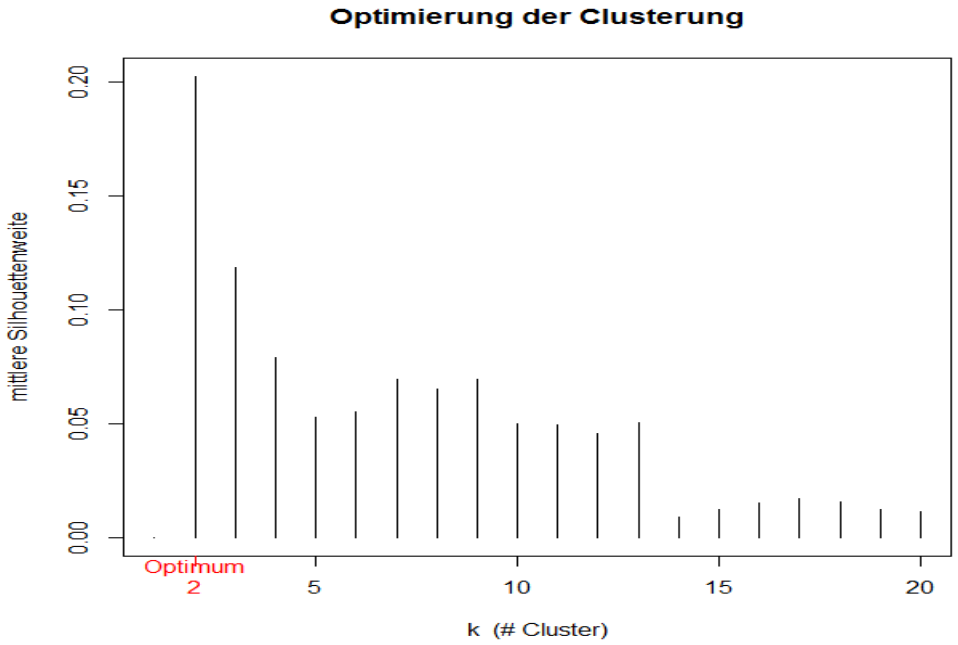


Abbildung 8.7.: Darstellung der mittleren Silhouettenweiten für verschiedene Anzahl berechneter Cluster

Die Berechnung liefert bereits bei 2 Clustern den besten Wert für die Silhouettenweite, dieser liegt mit einem Wert von 0,21 immer noch in einem eher niedrigen Niveau, hier kann bestenfalls von einer schwachen Clusterstruktur gesprochen werden. Unter der Voraussetzung, dass die Variablen „Sektor“ und „Land“ aus dem Datensatz entfernt wurden, kann mit einer Clusteranalyse kein hochwertiges Ergebnis erzielt werden.

Wird diese Auswertung unter Miteinbeziehung der beiden kategoriellen Variablen ein weiteres Mal durchgeführt, so zeigt sich folgendes Resultat:

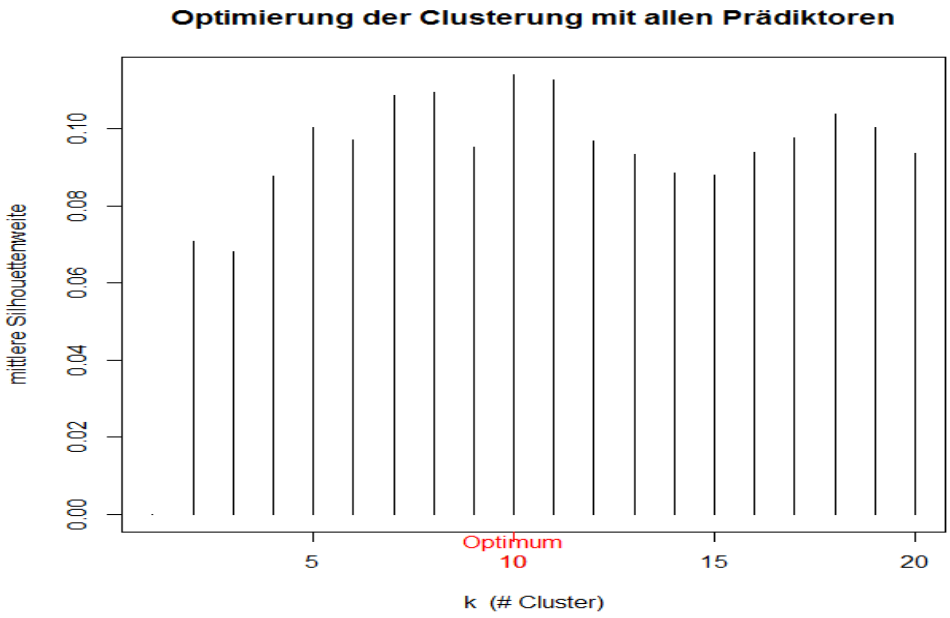


Abbildung 8.7.: mittlere Silhouettenweiten unter Verwendung aller Prädiktorvariablen

Hier zeigt sich ein anderes, jedoch kaum besseres Bild als zuvor. Die optimale Zahl an Clustern ist von 2 auf 10 angestiegen, jedoch zeigt die dabei auftretende mittlere Silhouettenweite von 0,12 auf, dass auch unter Einbeziehung der beiden kategoriellen Variablen ein hochwertiges Resultat hier offenbar nicht zu erzielen ist.

Die Hoffnung, im Kontext eines nicht-beobachteten Lernverfahrens eine Clusteraufteilung unter Verwendung der Prädiktorvariablen zu finden, sodass sich im besten Fall die Beobachtungen mit unterschiedlichen Liquiditätsklassen in die verschiedenen Cluster aufteilen, hat sich somit kaum erfüllt. Für hochliquide Instanzen zeigt die Clusterung zwar gute Ansätze, insgesamt dürfte aber die Variabilität der Prädiktorvariablen in Bezug auf die Liquidität nicht ausreichen, um eine Clusterung in höherwertiger Form erreichen zu können.

9) Zusammenfassung

In dieser Arbeit wird versucht, mithilfe spezieller statistischer Verfahren, die im Bereich des „maschinellen Lernens“ einzuordnen sind, eine Klassifizierung von Wertpapieren in Liquiditätskategorien anhand verschiedener Eigenschaften durchzuführen, welche mit der Liquidität in Verbindung gebracht werden können. Im Vordergrund steht dabei die Fragestellung, in wie fern sich jene Liquiditätseinstufungen, welche sich mithilfe eines traditionell verwendeten Maßes zur Kategorisierung, der sogenannten „Amihud Illiquidity Ratio“, ergeben, über die Berechnung von Klassifikationsmodellen mit finanzwirtschaftlichen und makroökonomischen Prädiktoren nachbilden lassen (das bedeutet, wie gut die tatsächliche Liquiditätskategorie mithilfe derartiger Prädiktoren vorhergesagt werden kann).

Diese Fragestellung wird in Form einer Analyse mit einem explorativen Charakter behandelt: Explorativ bedeutet dabei, dass bei der Verfassung dieser Arbeit nicht das Ziel verfolgt wird, explizit nachzuweisen, dass gewisse wirtschaftliche Faktoren einen hochsignifikanten Beitrag zur Erklärung der Liquidität liefern, dafür ist die verwendete Datenmenge auch deutlich zu gering. Es geht hier eher um das Auffinden grundlegender Zusammenhänge zwischen derartigen Faktoren und der Liquidität von Wertpapieren, und der Erforschung, in welcher Qualität eine derartige Aufgabe mithilfe von Methoden des maschinellen Lernens gelöst werden kann.

Der für die Untersuchung erhobene Datensatz besteht aus 902 Instanzen, allesamt Aktien aus 20 verschiedenen Ländern, und 15 verschiedenen Prädiktorvariablen, von denen 7 Variablen einen finanzwirtschaftlichen Ursprung und 7 Variablen einen makroökonomischen Ursprung haben, daneben stellt auch das Land, aus welchem das Unternehmen stammt, dessen Wertpapiere herangezogen werden, eine zusätzliche erklärende Variable dar. Weiters wird auch eine zeitliche Komponente in die Analyse eingebettet, damit eine ausreichende Variabilität der makroökonomischen Variablen gewährleistet ist. Dies wird durch die zufällige Wahl eines einmonatigen Zeitraums zwischen 2004 und 2013 für jede Instanz realisiert, wobei für jede Instanz die Werte der Prädiktorvariablen aus dem gewählten Zeitraum stammen bzw. diese über Daten aus diesem Monat berechnet werden. Nach der Bereinigung des Datensatzes reduzierte sich dessen Größe auf 716 Instanzen, mit denen schließlich die Analyse durchgeführt wird.

Als Verfahren kommen in weiterer Folge drei sogenannte „überwachte“ Lernverfahren (klassische Klassifikationsbäume, Random Forests sowie Support Vector Machines) sowie eine Methode aus dem nicht-überwachten Bereich (eine verallgemeinerte Variante des k-Means Clustering) zum Einsatz. Überwacht bedeutet dabei, dass Trainingsinstanzen für das Lernen benützt werden, bei welchen der Wert der abhängigen Variable bereits über eine andere Quelle bekannt ist (in diesem Fall waren dies Instanzen mit bekanntem Wert bei der Amihud Illiquidity Ratio) und diesen Wert man über die Prädiktoren möglichst genau treffen möchte. Im Gegensatz dazu kommt bei den nicht-überwachten Methoden ein Trainingsset nicht zur Anwendung, hier steht in vielen Fällen auch nicht die Erklärung einer speziellen abhängigen Variable im Mittelpunkt, sondern zumeist die Erforschung der Beziehung

zwischen Variablen bzw. Instanzen, wie es unter anderem bei der Clusteranalyse der Fall ist. Hier ist das Ziel, Information über die vorliegenden Instanzen zu gewinnen, indem man sie in möglichst homogene Gruppen unterteilt, die disjunkt sind und sich voneinander so stark wie möglich unterscheiden sollten.

Die anfängliche Anwendung der überwachten Verfahren auf das Trainingsset (noch ohne Testinstanzen) liefert eine durchschnittliche Trefferrate von etwa 65%, wobei Baumstrukturverfahren die beste Performance aufzeigen. Darauf aufbauend wird als hochwertiges Validierungsverfahren die 10-fache Kreuzvalidierung gewählt, welche bei allen verschiedenen Verfahren unter diversen Parameterkonstellationen durchgeführt wird.
Dabei werden die im ersten Schritt ermittelten Ergebnisse im Wesentlichen bestätigt, auch hier können dieselben Modellvarianten wie zuvor am besten abschneiden. Bemerkenswert ist dabei, dass diese Varianten auch ein hochwertigeres Ergebnis als das zusätzlich miteinbezogene Vergleichsmodell liefern, welches die klassische logistische Regression darstellt. Auch die Ergebnisse eines Ensemble-Learning Verfahrens (Majority Voting) liegen aus qualitativer Sicht nicht über den Baumstrukturverfahren. Das Resultat der nicht-überwachten Lernmethode (Clusteranalyse) kann die Erwartungen nicht in vollem Ausmaß erfüllen, hierfür scheint eine zu schwache Struktur beim verwendeten Datensatz vorzuliegen.

Werden nun neben den Ergebnissen der jeweiligen Methoden auch jene aus inhaltlicher Sicht beurteilt, so kann ausgesagt werden, dass die verwendeten makroökonomischen und finanzwirtschaftlichen Variablen recht stark mit der Liquidität (unter deren hierarchischer Einteilung in verschiedene Kategorien) assoziiert sind. Dabei stechen vor allem einzelne Prädiktorvariablen hinsichtlich ihrer Bedeutung hervor: Bei der Beurteilung der Variablen unter Verwendung von Importance-Maßen (Gini-Index, Mean Decrease Accuracy) zeigen der Turnover und das Land des Wertpapiers (in dem dessen Unternehmen seinen Sitz hat) die mit Abstand besten Resultate. Die Bid-Ask-Ratio erreicht dabei auch einen verhältnismäßig hohen Wert. Allgemein können die finanzwirtschaftlichen Variablen eine deutlich höhere Importance erzielen als die makroökonomischen Variablen, von denen am ehesten noch der Consumer Price Index (CPI) eine Rolle spielt.

Um noch aufschlussreichere Resultate hinsichtlich der Fähigkeit der verwendeten Variablen, die Liquiditätskategorien zu erklären, erhalten zu können, wäre vor allem der Einsatz eines größeren Datensatzes (5- bis 10-fache Menge der hier verwendeten Anzahl an Instanzen) nötig, um die Fähigkeiten der eingesetzten Methoden noch stärker zur Geltung kommen zu lassen. In der hier durchgeführten Analyse erkennt man besonders an Stelle der Klassifikationsbäume, dass die Anzahl an Trainingsinstanzen mit einer illiquiden Liquiditätskategorie so niedrig ausgefallen ist, dass selbst bei Bäumen mit einer mittleren Splitanzahl diese Kategorie am Ende in einigen Fällen nicht aufgetreten ist, und daher keine Regeln für diese Kategorie zum Schluss verfügbar waren. Dies ist keineswegs optimal, und wäre zu verhindern, wenn die Zahl an Instanzen und damit auch die Zahl an illiquiden Wertpapieren, die darin enthalten sind, höher wäre, als es hier der Fall ist. Dennoch kann von einem brauchbaren Ergebnis gesprochen werden, welches einen ordentlichen Einblick in die Beziehungen zwischen Liquidität und diversen wirtschaftlichen Faktoren geben kann.

10) Abbildungsverzeichnis

11) Tabellenverzeichnis

12) Literaturverzeichnis

Bücher:

- *Aggarwal, C.C. und C.K.Reddy*: Data Clustering: Algorithms and Applications, S100-106, CRC Press, London 2013

- *Alpaydin*, E.: Maschinelles Lernen, S347-349, 351-352, 375-382, Oldenbourg Verlag, München 2008

- *Breiman L., J. Friedman, C.J.Stone und R.A.Olshen*: Classification and Regression Trees, Taylor & Francis Group, London 1984

- *Duda, R.O., P.E. Hart und D.G. Stork*: Pattern Classification, S182-183,Wiley, New York 2001

- *Guida, J.J.*: Mikroökonomie und Management: die Grundlagen, S252, Kohlhammer, Stuttgart 2009

- *James, G., D. Witten, T. Hastie und R. Tibshirani*: An Introduction To Statistical Learning with Applications in R, S317-318, 338-352, 373, 385, 387-388, Springer Science & Business Media, New York 2013

- *Kaufman, L. und P.J. Rousseeuw*: Finding Groups in Data – An Introduction to Cluster Analysis, Wiley, S68-71, Hoboken/New Jersey 1990

- *Kroschel, K., G. Rigoll und B.Schuller*: Statistische Informationstechnik, 5.Auflage, S215-217, Springer Verlag, Berlin/Heidelberg 2011

- *Petersohn, H.*: Data Mining – Verfahren, Prozesse, Anwendungsstruktur, S136-141, Oldenbourg Verlag, München 2005

- *Pohl, M.*: Das Liquiditätsrisiko in Banken – Ansätze zur Messung und ertragsorientierten Steuerung, S8, GRIN Verlag, München 2008

- *Theodoridis, S. und K. Koutrumbas*: Pattern Recognition, 3. Ausgabe, S635, Academic Press, Waltham/Massachusetts 2006

- *Witten, I.H. und E. Frank*: Data Mining - Practical Machine Learning Tools and Techniques, Elsevier, xxiii (Preface), S42-43, 97, 136-137, 163, 171-172, San Francisco 2005

Artikel/Zeitschriften:

- *Amihud, Y.:* Illiquidity and Stock Returns: Cross-Section and Time-Series Effects, Journal of Financial Markets 5 (1), S31-56, Elsevier, London/Amsterdam 2002

- *Amihud, Y., H.Mendelson und L.H.Pedersen:* Liquidity and Asset Prices, Foundations and Trend in Finance 1 (4), S270, Now Publishers, Boston 2005

- *Breiman, L.:* Random Forests, Machine Learning 45 (1), S5-32, Springer Science & Business Media, New York 2001

- *Breslow, L.A. und D.W. Aha*: Simplifying Decision Trees: A Survey, S8-11, 14, 16-17, Navy Center for Applied Research in Artificial Intelligence, Washington 1997

- *Burges, C.*: A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery 2 (2), S128-129, 135, Boston 1998

- *Elomaa, T.*: In Defense of C4.5: Notes on learning one-level decision trees, Proceedings of the eleventh international conference on Machine Learning, S62-69, New Brunswick/New Jersey 1994

- *Fawcett, T.*: An introduction to ROC analysis, Pattern Recognition Letters 27, S862-864, 868, North Holland Publishing, London/Amsterdam 2006

- *Fernandez-Amador, O., M. Gächter, M. Larch und G. Peter*: Does monetary policy determine stock market liquidity? New evidence from the euro zone, Journal of Empirical Finance 21, S54-68, Elsevier, London/Amsterdam 2013

- *Freund, Y. und R.E. Schapire*: Experiments with a New Boosting Algorithm, Machine Learning: Proceedings of the Thirteenth International Conference, 1996

- *Ghosh, A.K.*: On optimum choice of k in nearest neighbor classification, Computational Statistics & Data Analysis 50 (11), S3113-3123, Elsevier, London/Amsterdam 2006

- *Goyenko, R.Y., C.W. Holden und C.A. Trzcinka*: Do liquidity measures measure liquidity?, Journal of Financial Economics 92 (2), S161, Elsevier, London/Amsterdam 2009

- *Hamming, R.W.*: Error-detecting and error-correcting codes, Bell System Technical Journal XXIX (2), S154-155, New York 1950

- *Holte, R.C.*: Very simple classification rules perform well on most commonly used datasets, Machine Learning 11 (1), S63-91, Springer Science & Business Media, New York 1993

- *Lachenbruch, P.A. und M. Mickey*: Estimation Of Error Rates in Discriminant Analysis, Technometrics 10 (1), S1-5, Taylor & Francis Group, London 1968

- *Roll, R.*: A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market, The Journal of Finance 39 (4), Wiley-Blackwell, Hoboken/New Jersey 1984

- *Rousseeuw, P.J.*: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20, S53-65, 1987

- *Söderberg, J.*: Do Macroeconomic Variables Forecast Changes in Liquidity? An Out-of-Sample Study on the Order-Driven Stock Markets in Scandinavia, CAFO Working Paper, Växjö University, 2008

- *Wolpert, D.*: The Lack of A Priori Distinctions between Learning Algorithms, Neural Computation 8 (7), S1352-1365, MIT Press, Cambridge/Massachusetts 1996

Diplomarbeiten, Dissertationen:

- *Fink, K.*: Liquiditätsrisiko – aufsichtsrechtliche und betriebswirtschaftliche Anforderungen für Banken, Diplomarbeit, S22-23, Wien 2012

- *Reimund, C.*: Liquiditätshaltung und Unternehmenswert. Dissertation.: Erklärungsansätze, Modelle und empirische Untersuchung deutscher börsennotierter Unternehmen, S8, 2003

Lehrveranstaltungen/Seminare:

- *Grossmann, W.*: Kapitel „Statistische Modellierung“ aus Vorlesung „Angewandte Statistik“, Universität Wien, 2011

- *Hudec, M. und W. Grossmann*: Kapitel „Klassifikation“ aus Vorlesung „Machine Learning“, Universität Wien, 2013

Onlinequellen:

- URL: http://www.bis.org/publ/bcbs144.htm [Zugriff am 2.11.2014]

- URL: http://www.bis.org/publ/bcbs165.htm [Zugriff am 4.3.2014]

- URL: http://www.bis.org/publ/bcbs188_de.pdf [Zugriff am 6.4.2014]

- URL: http://www.iif.com/download.php?id=tv/yzuHHXJ0= [Zugriff am 4.3.2014]

- URL: http://www.fma.gv.at/de/ueber-die-fma.html [Zugriff am 16.10.2014]

- URL: http://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen &Gesetzesnummer=10004827 [Zugriff am 16.10.2014]

- URL: http://www.fma.gv.at/typo3conf/ext/dam_download/secure.php? u=0&file=11004&t=1384294085&hash=83c55aa78e0a89dba1821c176d66fd9a [Zugriff am 16.10.2014]

- URL: http://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen& Gesetzesnummer=20008739 [Zugriff am 16.10.2014]

- URL: http://www.risiko-manager.com/index.php?id= 162&tx_ttnews%5Bcat%5D=71&tx_ttnews%5Btt_news%5D=8737&tx_ttnews%5Bb

ackPid%5D=779&cHash=5f60f846e0f4098ce12ac8345dae58ee [Zugriff am 4.5.2014]

- URL: http://www.risiko-manager.com/index.php?id=162&
  tx_ttnews%5Bcat%5D=185&tx_ttnews%5Btt_news%5D=17501&tx_ttnews%5BbackPid%5D=773&cHash=fab343bffe67888d2b8a908cf2133d54 [Zugriff am 8.5.2014]

- URL: http://cdn.gotraffic.net/enterprise/pdf/DS_bval_score_primer_final_9.21.10.pdf [Zugriff am 11.5.2014]

- URL: http://www.bloomberg.com/enterprise/data/pricing-services/ [Zugriff am 16.10.2014]

- URL: http://www.morningstar.com/InvGlossary/Default.aspx?letter=C [Zugriff am 15.9.2014]

- URL: http://www.ecb.europa.eu/stats/money/aggregates/aggr/html/hist.en.html [Zugriff am 26.11.2014])

- URL: http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&language=en&
  pcode=teibs010 [Zugriff am 26.11.2014])

- URL: http://de.wikipedia.org/wiki/Euklidischer_Abstand [Zugriff am 10.05.2014]

- URL: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm [Zugriff am 15.5.2014]

- URL: http://support.sas.com/documentation/cdl/en/statug/63962/HTML/
  default/viewer.htm#statug_distance_sect016.htm [Zugriff am 17.10.2014]

- URL: https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/daisy.html [Zugriff am 20.10.2014]]

- URL: http://www.quantstart.com/articles/Support-Vector-Machines-A-Guide-for-Beginners [Zugriff am 15.12.2014])

13) R-Code zur Berechnung der Resultate

```
library(Amelia)
library(DMwR)
library(rpart)
library(randomForest)
library(e1071)
library(cluster)
library(pROC)
library(nnet)

#Herstellung der Datenverbindung, Anzahl fehlender Werte

setwd("C:/Users/andreashackl/Andreas/Studium/Mag Statistik/Magisterarbeit")
daten <- read.csv("Datensatz_MGA_1.csv",sep=";",dec=".",header=TRUE)
nmiss <- NULL
nmiss_ber <- NULL
for (i in 1:19)
nmiss[i] <- length(which(is.na(daten[,i])))

#Bereinigung des Datensatzes - Schritt 1

daten_ber <- daten[which(!is.na(daten[,21] == TRUE)),c(2:9,12:18,21)]

daten_ber_cls <- daten[,c(2:9,12:18,21)]

for (i in 1:16)
nmiss_ber[i] <- length(which(is.na(daten_ber[,i])))

#deskriptive Auswertungen und Grafiken

table(daten_ber[,4])
table(daten_ber[,5])
table(daten_ber[,7])
table(daten_ber[,8])
table(daten_ber[which(daten_ber[,16] == 1),4])
table(daten_ber[which(daten_ber[,16] == 2),4])
table(daten_ber[which(daten_ber[,16] == 3),4])
table(daten_ber[which(daten_ber[,16] == 4),4])
table(daten_ber[which(daten_ber[,16] == 5),4])
table(daten_ber[which(daten_ber[,16] == 1),5])
table(daten_ber[which(daten_ber[,16] == 2),5])
table(daten_ber[which(daten_ber[,16] == 3),5])
table(daten_ber[which(daten_ber[,16] == 4),5])
table(daten_ber[which(daten_ber[,16] == 5),5])
table(daten_ber[which(daten_ber[,16] == 1),8])
table(daten_ber[which(daten_ber[,16] == 2),8])
table(daten_ber[which(daten_ber[,16] == 3),8])
table(daten_ber[which(daten_ber[,16] == 4),8])
table(daten_ber[which(daten_ber[,16] == 5),8])

summary(daten_ber[,c(1,2,3,6,9:15)])
sd(daten_ber[,c(1,2,3,6,9:15)])

daten_ber[which(daten_ber[,3] > 3e+09),2] <- NA
plot(density(daten_ber_fin2[(!is.na(daten_ber_fin2[,1]) &
daten_ber_fin2[,16] == 1),1]), main="Dichte BidAskRatio", xlim=c(-
0.05,0.2), ylim=c(0,200))
lines(density(daten_ber_fin2[(!is.na(daten_ber_fin2[,1]) &
daten_ber_fin2[,16] == 2),1]),col="blue")
```

```
lines(density(daten_ber_fin2[(!is.na(daten_ber_fin2[,1]) &
daten_ber_fin2[,16] == 3),1]),col="red")
lines(density(daten_ber_fin2[(!is.na(daten_ber_fin2[,1]) &
daten_ber_fin2[,16] == 4),1]),col="green")
lines(density(daten_ber_fin2[(!is.na(daten_ber_fin2[,1]) &
daten_ber_fin2[,16] == 5),1]),col="brown")
legend("topright", c("LIQ=1", "LIQ=2", "LIQ=3", "LIQ=4", "LIQ=5"),
fill=c("black", "blue", "red", "green", "brown"))

par(mfrow=c(1,4))
boxplot(daten_ber[,1])
boxplot(daten_ber[,2])
boxplot(daten_ber[,3])
boxplot(daten_ber[,15])
par(mfrow=c(1,1))

#Festlegung einer oberen bzw. unteren Schranke für Ausreisser

daten_ber[which(daten_ber[,1] > 5),1] <- 2
daten_ber[which(daten_ber[,2] > 5),2] <- 5
daten_ber[which(daten_ber[,3] > 5.0e+09),3] <- 5.0e+09
daten_ber[which(daten_ber[,15] < - 5),15] <- -5
daten_ber[which(daten_ber[,15] > 5),15] <- 5

#Imputation fehlender Werte (Bereinigung Schritt 2)

daten_ber2 <- amelia(daten_ber, m=5, noms=c("SEC", "Country"), ords="RTG")
daten_ber_fin <- daten_ber2$imputations$imp3
daten_ber_fin[,4] <- daten_ber[,4]

daten_ber_fin2 <- knnImputation(daten_ber, k=10, meth="median")
daten_ber_fin2[,4] <- daten_ber[,4]

#Korrelationstabellen, Scatterplot-Diagramme

cor(daten_ber_fin2[,c(1:3,5,6,9:15)])
pairs(daten_ber_fin2[,c(1:3,5,6,9:15)])

attach(daten_ber_fin2)

#Classificaton Trees – erster Versuch

liq.rpart <- rpart(formula = ILL ~ BidAsk + VOLV + TO + RTG + NUM_LIST +
STD + SEC + Country + CPI + IP + REER + MonSp + STIR + LTIR + EconSent,
method="class", data=daten_ber_fin2)
plot(liq.rpart)
text(liq.rpart)

plotcp(liq.rpart)
printcp(liq.rpart)

#komplexere Variante, Zurechtschneiden auf adäquate Größe

liq.rpart_comp <- rpart(formula = ILL ~ BidAsk + VOLV + TO + RTG + NUM_LIST
+ STD + SEC + Country + CPI + IP + REER + MonSp + STIR + LTIR + EconSent,
method="class", control = rpart.control(cp=.001), data=daten_ber_fin2)
plotcp(liq.rpart_comp)
printcp(liq.rpart_comp)

liq.rpart_comp1 <- prune(liq.rpart_comp, cp=0.0065503)
plot(liq.rpart_comp1, main = "Classification Tree ohne 1-Stabw-Regel")
text(liq.rpart_comp1)
```

```
liq.rpart_comp2 <- prune(liq.rpart_comp, cp=0.015)
plot(liq.rpart_comp2, main = "Classification Tree mit 1-Stabw-Regel")
text(liq.rpart_comp2)

#Prognose, Generierung von ROC-Kurven

pred_res <- predict(liq.rpart_comp1,daten_ber_fin2,"class")
pred_res_pr <- predict(liq.rpart_comp1,daten_ber_fin2)
rbind(daten_ber_fin2$ILL, pred_res)
table(daten_ber_fin2$ILL, pred_res)

pred_res2 <- predict(liq.rpart_comp2,daten_ber_fin2,"class")
pred_res_pr2 <- predict(liq.rpart_comp1,daten_ber_fin2)
rbind(daten_ber_fin2$ILL, pred_res2)
table(daten_ber_fin2$ILL, pred_res2)

pred1 <- pred_res_pr[,1]
pred2 <- pred_res_pr[,2]
pred3 <- pred_res_pr[,3]
pred4 <- pred_res_pr[,4]
pred5 <- pred_res_pr[,5]

roc1 <- roc(as.numeric(daten_ber_fin2$ILL==1),pred1)
roc2 <- roc(as.numeric(daten_ber_fin2$ILL==2),pred2)
roc3 <- roc(as.numeric(daten_ber_fin2$ILL==3),pred3)
roc4 <- roc(as.numeric(daten_ber_fin2$ILL==4),pred4)
roc5 <- roc(as.numeric(daten_ber_fin2$ILL==5),pred5)

par(mfrow=c(2,3))

plot.roc(roc1, main="ROC-Kurve für LIQ=1")
plot.roc(roc2, main="ROC-Kurve für LIQ=2")
plot.roc(roc3, main="ROC-Kurve für LIQ=3")
plot.roc(roc4, main="ROC-Kurve für LIQ=4")
plot.roc(roc5, main="ROC-Kurve für LIQ=5")

#Random Forests – erster Versuch

daten_ber_fin2$ILL <- as.factor(daten_ber_fin2$ILL)
liq.rf1 <- randomForest(ILL ~ BidAsk + VOLV + TO + RTG + NUM_LIST + STD +
SEC + Country + CPI + IP + REER + MonSp + STIR + LTIR + EconSent,
data=daten_ber_fin2, ntree=500, mtry=3, importance=TRUE)
print(liq.rf1)
tuneRF(x=daten_ber_fin2[, -16], y=daten_ber_fin2$ILL, trace=FALSE)
importance(liq.rf2)

#Prognose für ersten Versuch

summary(liq.rf1)
pred_res1 <- round(as.numeric(liq.rf1$predicted),0) #Prognosewerte
rbind(daten_ber_fin2$ILL,pred_res1)

#Festlegung auf 500 Bäume und Splitvariablenzahl = 10

liq.rf2 <- randomForest(ILL ~ BidAsk + VOLV + TO + RTG + NUM_LIST + STD +
SEC + Country + CPI + IP + REER + MonSp + STIR + LTIR + EconSent,
data=daten_ber_fin2, ntree=500, mtry=5, importance=TRUE)
liq.rf3 <- randomForest(ILL ~ BidAsk + VOLV + TO + RTG + NUM_LIST + STD +
SEC + Country + CPI + IP + REER + MonSp + STIR + LTIR + EconSent,
data=daten_ber_fin2, ntree=500, mtry=10, importance=TRUE)
```

```
#Prognose, ROC-Kurven

pred_res3a <- as.numeric(liq.rf3$predicted),0)
pred_res3b <- predict(liq.rf3, type='prob')
pred1 <- pred_res3b[,1]
pred2 <- pred_res3b[,2]
pred3 <- pred_res3b[,3]
pred4 <- pred_res3b[,4]
pred5 <- pred_res3b[,5]

roc1 <- roc(as.numeric(daten_ber_fin2$ILL==1),pred1)
roc2 <- roc(as.numeric(daten_ber_fin2$ILL==2),pred2)
roc3 <- roc(as.numeric(daten_ber_fin2$ILL==3),pred3)
roc4 <- roc(as.numeric(daten_ber_fin2$ILL==4),pred4)
roc5 <- roc(as.numeric(daten_ber_fin2$ILL==5),pred5)

par(mfrow=c(2,3))

plot.roc(roc1, main="ROC-Kurve für LIQ=1")
plot.roc(roc2, main="ROC-Kurve für LIQ=2")
plot.roc(roc3, main="ROC-Kurve für LIQ=3")
plot.roc(roc4, main="ROC-Kurve für LIQ=4")
plot.roc(roc5, main="ROC-Kurve für LIQ=5")

#Support Vector Machines

liq.svm <- svm(ILL ~ ., probability=TRUE, data=daten_ber_fin2)

print(liq.svm)
summary(liq.svm)

#Prognose, ROC-Kurven

daten_pr <- subset(daten_ber_fin2, select= -ILL)
pred_res3 <- predict(liq.svm, daten_ber_fin2, decision.values=TRUE)
pred_res3 <- round(summary(liq.svm,daten_ber_fin2)$fitted) #alternativ
pred_res3_pr <- predict(liq.svm,daten_ber_fin2,probabilities=TRUE)
table(pred_res3, daten_ber_fin2$ILL)

pred1 <- pred_res3_pr[,1]
pred2 <- pred_res3_pr[,2]
pred3 <- pred_res3_pr[,3]
pred4 <- pred_res3_pr[,4]
pred5 <- pred_res3_pr[,5]

roc1 <- roc(as.numeric(daten_ber_fin2$ILL==1),pred1)
roc2 <- roc(as.numeric(daten_ber_fin2$ILL==2),pred2)
roc3 <- roc(as.numeric(daten_ber_fin2$ILL==3),pred3)
roc4 <- roc(as.numeric(daten_ber_fin2$ILL==4),pred4)
roc5 <- roc(as.numeric(daten_ber_fin2$ILL==5),pred5)

par(mfrow=c(2,3))

plot.roc(roc1, main="ROC-Kurve für LIQ=1")
plot.roc(roc2, main="ROC-Kurve für LIQ=2")
plot.roc(roc3, main="ROC-Kurve für LIQ=3")
plot.roc(roc4, main="ROC-Kurve für LIQ=4")
plot.roc(roc5, main="ROC-Kurve für LIQ=5")

#grafische Darstellung der Support-Vektoren

plot(cmdscale(dist(daten_ber_fin2[,c(-4,-16)])),
```

```
     col = as.integer(daten_ber_fin2[,16]),
     pch = c("o","+")[1:716 %in% liq.svm$index + 1],xlim=c(-2e+08,-
1e+08),ylim=c(-30,30),main="grafische Darstellung der Support Vektoren")

#Clusteranalyse

daten_ber_cls[which(daten_ber_cls[,1] > 5),1] <- 2

# Ausreisser ist unbrauchbarer Wert - lege obere Schranke für jeweilige
# Variablen fest

daten_ber_cls[which(daten_ber_cls[,2] > 5),2] <- 5
daten_ber_cls[which(daten_ber_cls[,3] > 5.0e+09),3] <- 5.0e+09
daten_ber_cls[which(daten_ber_cls[,15] < - 5),15] <- -5
daten_ber_cls[which(daten_ber_cls[,15] > 5),15] <- 5

#Imputation für Clusteranalyse

daten_ber_cls_fin <- knnImputation(daten_ber_cls[,-16], k=10,
meth="median")
daten_ber_cls_fin[,4] <- daten_ber_cls[,4]

#Berechnung der Distanzmatrix, Erstellung der Cluster

var_dist1 <- daisy(daten_ber_cls_fin[,c(-7,-16)], metric="gower",
stand=FALSE)

var_clus <- pam(var_dist1, 5, diss=TRUE, stand=FALSE)

var_clus1 <- which(var_clus$clustering == 1)
clus1 <- daten_ber_cls_fin[var_clus1,]
var_clus2 <- which(var_clus$clustering == 2)
clus2 <- daten_ber_cls_fin[var_clus2,]
var_clus3 <- which(var_clus$clustering == 3)
clus3 <- daten_ber_cls_fin[var_clus3,]
var_clus4 <- which(var_clus$clustering == 4)
clus4 <- daten_ber_cls_fin[var_clus4,]
var_clus5 <- which(var_clus$clustering == 5)
clus5 <- daten_ber_cls_fin[var_clus5,]
var_clus6 <- which(var_clus$clustering == 6)
clus6 <- daten_ber_fin2[var_clus6,]
var_clus7 <- which(var_clus$clustering == 7)
clus7 <- daten_ber_fin2[var_clus7,]
var_clus8 <- which(var_clus$clustering == 8)
clus8 <- daten_ber_fin2[var_clus8,]

#Eigenschaften der Cluster

summary(clus1) # analog für andere Cluster

mean(daten_ber_cls[var_clus1,16][which(is.na(daten_ber_cls[var_clus1,16])==
FALSE)]) #analog für andere Cluster

daten_pl <- NULL
daten_pl[1] <- mean(daten_ber_cls_fin[var_clus1,15])
daten_pl[2] <- mean(daten_ber_cls_fin[var_clus2,15])
daten_pl[3] <- mean(daten_ber_cls_fin[var_clus3,15])
daten_pl[4] <- mean(daten_ber_cls_fin[var_clus4,15])
daten_pl[5] <- mean(daten_ber_cls_fin[var_clus5,15])
```

```
#grafische Darstellung

par(mfrow=c(2,2))
plot(1:5,daten_pl,main="Turnover nach Cluster",
xlab="Cluster",ylab="Turnover", pch=1,lwd=7)
plot(1:5,daten_pl,main="Bid-Ask-Ratio nach Cluster",
xlab="Cluster",ylab="Bid-Ask-Ratio", pch=1,lwd=7)
plot(1:5,daten_pl,main="Volumsveränderung/Monat nach Cluster",
xlab="Cluster",ylab="Volumsveränd.", pch=1,lwd=7)
plot(1:5,daten_pl,main="Standardabweichung Return nach Cluster",
xlab="Cluster",ylab="STD", pch=1,lwd=7)
plot(1:5,daten_pl,main="CPI nach Cluster", xlab="Cluster",ylab="CPI",
pch=1,lwd=7)
plot(1:5,daten_pl,main="kurzfristiger Zinssatz nach Cluster",
xlab="Cluster",ylab="kurzfr. ZS", pch=1,lwd=7)
plot(1:5,daten_pl,main="verfügbare Geldmenge nach Cluster",
xlab="Cluster",ylab="verf. GM", pch=1,lwd=7)
plot(1:5,daten_pl,main="Economic Sentiment nach Cluster",
xlab="Cluster",ylab="Econ. Sent.", pch=1,lwd=7)


plot(var_clus, main="Silhouettenplot Clusterauswahl") # Silhouettenplot,
wird in Kapitel 8 dann wichtiger

#Vorbereitung Kreuzvalidierung

cv.rpart <- function(form,train,test,...) {
  m <- rpartXse(form,train,...)
  p <- predict(m,test)
  mse <- mean((p-resp(form,test))^2)
  c(nmse=mse/mean((mean(resp(form,train))-resp(form,test))^2))
}

cv.rf <- function(form,train,test,...) {
  m <- randomForest(form,train,...)
  p <- predict(m,test)
  mse <- mean((p-resp(form,test))^2)
  c(nmse=mse/mean((mean(resp(form,train))-resp(form,test))^2))
}

cv.svm <- function(form,train,test,...) {
  require(e1071)
  t <- svm(form,train,...)
  p <- predict(t,test)
  mse <- mean((p-resp(form,test))^2)
  c(nmse=mse/mean((mean(resp(form,train))-resp(form,test))^2))
}

cv.lm <- function(form,train,test,...) {
  t <- lm(form,train,family="binomial",...)
  p <- predict(t,test)
  mse <- mean((p-resp(form,test))^2)
  c(nmse=mse/mean((mean(resp(form,train))-resp(form,test))^2))
}
daten_ber_fin2[which(daten_ber_fin2[,4] == "Ca"),4] <- "Caa"

res <- experimentalComparison(
                  c(dataset(ILL ~., daten_ber_fin2[,-16])),
                  c(variants('cv.rpart',se=c(0,1)),
                    variants('cv.rf',ntree=c(50,500,800)),
                    variants('cv.svm', kernel=c("radial", "polynomial",
"sigmoid")),
```

```
                  variants('cv.lm')),
                cvSettings(5,10,1234))

res2 <-  experimentalComparison(
                c(dataset(ILL ~., daten_ber_fin2[,-16])),
c(variants('cv.rf',ntree=500, mtry=c(3,5,10)),
                  variants('cv.svm', kernel="radial", cost=c(1,10,50))),
                cvSettings(5,10,1234))

cv.rpart <- function(form, train, test, ...) {
require(rpart)
model <- rpartXse(form, train, method='class',...)
preds <- predict(model, test, type='class',...)
class.eval(resp(form, test), preds,
stats=c('acc'))
}

cv.rf <- function(form, train, test, ...) {
model <- randomForest(form, train, importance=TRUE, ...)
preds <- predict(model, test, type='class',...)
class.eval(resp(form, test), preds,
stats=c('acc'))
}

cv.svm <- function(form, train, test, ...) {
require(e1071)
model <- svm(form, train,...)
preds <- predict(model, test, type='class',...)
class.eval(resp(form, test), preds,
stats=c('acc'))
}

cv.glm <- function(form, train, test, ...) {
model <- multinom(form, train,...)
preds <- predict(model, test,...)
class.eval(resp(form, test), preds,
stats=c('acc'))
}

#Kreuzvalidierung (5x10)

ILL <- as.factor(ILL)
eval.res <- experimentalComparison(
                c(dataset(ILL ~., daten_ber_fin2[,-16])),
                c(variants('cv.rpart',se=c(0,1)),
                  variants('cv.rf',ntree=c(50,500,800)),
                  variants('cv.svm', kernel=c("radial", "polynomial",
"sigmoid"))),
                cvSettings(5,10,1234))

eval.res2 <-  experimentalComparison(
                c(dataset(ILL ~., daten_ber_fin2[,-16])),
                c(variants('cv.glm'),
                  variants('cv.rf',ntree=500, mtry=c(3,5,10)),
                  variants('cv.svm', kernel="radial", cost=c(1,10,50))),
                cvSettings(5,10,1234))

#Optimierung der Clusteranalyse

asw <- numeric(20)
for (k in 2:20)
  asw[k] <- pam(var_dist1, k, diss=TRUE, stand=FALSE) $ silinfo $ avg.width
```

```
k.best <- which.max(asw)

plot(1:20, asw, type= "h", main = "Optimierung der Clusterung",
     xlab= "k  (# Cluster)", ylab = "mittlere Silhouettenweite")
axis(1, k.best, paste("Optimum",k.best,sep="\n"), col = "red", col.axis =
"red")

var_dist2 <- daisy(daten_ber_fin2[,-16], metric="gower", stand=FALSE)

asw2 <- numeric(20)
for (k in 2:20)
  asw2[k] <- pam(var_dist2, k, diss=TRUE, stand=FALSE) $ silinfo $
avg.width
k.best2 <- which.max(asw2)

plot(1:20, asw2, type= "h", main = "Optimierung der Clusterung mit allen
Prädiktoren",
     xlab= "k  (# Cluster)", ylab = "mittlere Silhouettenweite")
axis(1, k.best2, paste("Optimum",k.best2,sep="\n"), col = "red", col.axis =
"red")


#Berechnung von ROC-Kurven für Schritte der Kreuzvalidierung
#hier für Classification Trees - bei anderen Methoden analog durchgeführt

daten_cv <- sample(1:716,716,replace=FALSE)
daten_cv1 <- list(NULL)
daten_tr <- list(NULL)
daten_cv1[[1]] <- daten_cv[1:72]
daten_cv1[[2]] <- daten_cv[73:144]
daten_cv1[[3]] <- daten_cv[145:216]
daten_cv1[[4]] <- daten_cv[217:288]
daten_cv1[[5]] <- daten_cv[289:360]
daten_cv1[[6]] <- daten_cv[361:432]
daten_cv1[[7]] <- daten_cv[433:504]
daten_cv1[[8]] <- daten_cv[505:576]
daten_cv1[[9]] <- daten_cv[577:648]
daten_cv1[[10]] <- daten_cv[649:716]

daten_tr[[1]] <- daten_cv[73:716]
daten_tr[[2]] <- daten_cv[c(1:72,145:716)]
daten_tr[[3]] <- daten_cv[c(1:144,217:716)]
daten_tr[[4]] <- daten_cv[c(1:216,289:716)]
daten_tr[[5]] <- daten_cv[c(1:288,361:716)]
daten_tr[[6]] <- daten_cv[c(1:360,433:716)]
daten_tr[[7]] <- daten_cv[c(1:432,505:716)]
daten_tr[[8]] <- daten_cv[c(1:504,577:716)]
daten_tr[[9]] <- daten_cv[c(1:576,649:716)]
daten_tr[[10]] <- daten_cv[1:648]

daten_tr1 <- list(NULL)
daten_test1 <- list(NULL)
pred_res_cv <- list(NULL)
pred1 <- list(NULL)
roc1 <- list(NULL)
pred2 <- list(NULL)
roc2 <- list(NULL)
pred3 <- list(NULL)
roc3 <- list(NULL)
pred4 <- list(NULL)
roc4 <- list(NULL)
pred5 <- list(NULL)
```

```
roc5 <- list(NULL)

for(i in 1:10) {
daten_tr1[[i]] <- daten_ber_fin2[daten_tr[[i]],]
daten_test1[[i]] <- daten_ber_fin2[daten_cv1[[i]],]
liq.rpart_comp <- rpart(formula = as.factor(ILL) ~ BidAsk + VOLV + TO + RTG
+ NUM_LIST + STD + SEC + Country + CPI + IP + REER + MonSp + STIR + LTIR +
EconSent, method="class", data=daten_tr1[[i]])
pred_res_cv[[i]] <- predict(liq.rpart_comp,daten_test1[[i]],type="prob")

pred1[[i]] <- pred_res_cv[[i]][,1]
pred2[[i]] <- pred_res_cv[[i]][,2]
pred3[[i]] <- pred_res_cv[[i]][,3]
pred4[[i]] <- pred_res_cv[[i]][,4]
pred5[[i]] <- pred_res_cv[[i]][,5]

roc1[[i]] <- roc(as.numeric(daten_test1[[i]]$ILL==1),pred1[[i]])
pred_1 <- unlist(pred1)
daten_test2a <-
c(daten_test1[[1]][,16],daten_test1[[2]][,16],daten_test1[[3]][,16],daten_t
est1[[4]][,16],daten_test1[[5]][,16],daten_test1[[6]][,16],daten_test1[[7]]
[,16],daten_test1[[8]][,16],daten_test1[[9]][,16],daten_test1[[10]][,16])
roc1a <- roc(as.numeric(daten_test2a==1),pred_1)
roc2[[i]] <- roc(as.numeric(daten_test1[[i]]$ILL==2),pred2[[i]])
pred_2 <- unlist(pred2)
daten_test2b <-
c(daten_test1[[1]][,16],daten_test1[[2]][,16],daten_test1[[3]][,16],daten_t
est1[[4]][,16],daten_test1[[5]][,16],daten_test1[[6]][,16],daten_test1[[7]]
[,16],daten_test1[[8]][,16],daten_test1[[9]][,16],daten_test1[[10]][,16])
roc2a <- roc(as.numeric(daten_test2b==2),pred_2)
roc3[[i]] <- roc(as.numeric(daten_test1[[i]]$ILL==3),pred3[[i]])
pred_3 <- unlist(pred3)
daten_test2c <-
c(daten_test1[[1]][,16],daten_test1[[2]][,16],daten_test1[[3]][,16],daten_t
est1[[4]][,16],daten_test1[[5]][,16],daten_test1[[6]][,16],daten_test1[[7]]
[,16],daten_test1[[8]][,16],daten_test1[[9]][,16],daten_test1[[10]][,16])
roc3a <- roc(as.numeric(daten_test2c==3),pred_3)
roc4[[i]] <- roc(as.numeric(daten_test1[[i]]$ILL==4),pred4[[i]])
pred_4 <- unlist(pred4)
daten_test2d <-
c(daten_test1[[1]][,16],daten_test1[[2]][,16],daten_test1[[3]][,16],daten_t
est1[[4]][,16],daten_test1[[5]][,16],daten_test1[[6]][,16],daten_test1[[7]]
[,16],daten_test1[[8]][,16],daten_test1[[9]][,16],daten_test1[[10]][,16])
roc4a <- roc(as.numeric(daten_test2d==4),pred_4)
roc5[[i]] <- roc(as.numeric(daten_test1[[i]]$ILL==5),pred5[[i]])
}
pred_5 <- unlist(pred5)
daten_test2e <-
c(daten_test1[[1]][,16],daten_test1[[2]][,16],daten_test1[[3]][,16],daten_t
est1[[4]][,16],daten_test1[[5]][,16],daten_test1[[6]][,16],daten_test1[[7]]
[,16],daten_test1[[8]][,16],daten_test1[[9]][,16],daten_test1[[10]][,16])
roc5a <- roc(as.numeric(daten_test2e==5),pred_5)
plot.roc(roc1[[1]], main="ROC-Kurve für LIQ=1")
for(j in 2:10) {
lines.roc(roc1[[j]])
}
lines.roc(roc1a, col="red")
plot.roc(roc2[[1]], main="ROC-Kurve für LIQ=2")
for(j in 2:10) {
lines.roc(roc2[[j]])
}
lines.roc(roc2a, col="red")
```

```
plot.roc(roc3[[1]], main="ROC-Kurve für LIQ=3")
for(j in 2:10) {
lines.roc(roc3[[j]])
}
lines.roc(roc3a, col="red")
plot.roc(roc4[[1]], main="ROC-Kurve für LIQ=4")
for(j in 2:10) {
lines.roc(roc4[[j]])
}
lines.roc(roc4a, col="red")
plot.roc(roc5[[1]], main="ROC-Kurve für LIQ=5")
for(j in 2:10) {
lines.roc(roc5[[j]])
}
lines.roc(roc5a, col="red")


#Majority Voting
#Trainings- und Testsetgenerierung äquivalent zu oben - daten_cv1 und
#daten_tr von oben erzeugt

daten_tr1 <- list(NULL)
daten_test1 <- list(NULL)
pred_res_cv1 <- list(NULL)
pred_res_cv2 <- list(NULL)
pred_res_cv3 <- list(NULL)
pred_res_cv1_pr <- list(NULL)
pred_res_cv2_pr <- list(NULL)
pred_res_cv3_pr <- list(NULL)
#pred5 <- list(NULL)
#roc5 <- list(NULL)

#Berechnung der Modelle & Prognosen, Durchführung des Votings

for(i in 1:10) {
daten_tr1[[i]] <- daten_ber_fin2[daten_tr[[i]],]
daten_test1[[i]] <- daten_ber_fin2[daten_cv1[[i]],]
liq.rpart_mj <- rpart(formula = as.factor(ILL) ~ BidAsk + VOLV + TO + RTG +
NUM_LIST + STD + SEC + Country + CPI + IP + REER + MonSp + STIR + LTIR +
EconSent, method="class", data=daten_tr1[[i]])
liq.rf_mj <- randomForest(as.factor(ILL) ~ BidAsk + VOLV + TO + RTG +
NUM_LIST + STD + SEC + Country + CPI + IP + REER + MonSp + STIR + LTIR +
EconSent, data=daten_tr1[[i]], ntree=500, mtry=10, importance=TRUE)
liq.svm_mj <- svm(formula = as.factor(ILL) ~ BidAsk + VOLV + TO + RTG +
NUM_LIST + STD + SEC + Country + CPI + IP + REER + MonSp + STIR + LTIR +
EconSent, probability=TRUE, data=daten_tr1[[i]])
pred_res_cv1[[i]] <- predict(liq.rpart_mj,daten_test1[[i]],"class")
pred_res_cv2[[i]] <- predict(liq.rf_mj,daten_test1[[i]])
pred_res_cv3[[i]] <- predict(liq.svm_mj,daten_test1[[i]])
pred_res_cv1_pr[[i]] <- predict(liq.rpart_mj,daten_test1[[i]])
pred_res_cv2_pr[[i]] <- predict(liq.rf_mj,daten_test1[[i]],type='prob')
pred_res_cv3_pr[[i]] <- predict(liq.svm_mj,daten_test1[[i]],prob=T)
}
pred_res_mj <- list(NULL)
for (k in 1:9) {
pred_res_mj[[k]] <- numeric(72)
}
pred_res_mj[[10]] <- numeric(68)
for(i in 10:10) {
for(j in 1:68) {
if ((pred_res_cv1[[i]][j] == pred_res_cv2[[i]][j]) | (pred_res_cv1[[i]][j]
== pred_res_cv3[[i]][j])) pred_res_mj[[i]][j] = pred_res_cv1[[i]][j] else {
```

```
if (pred_res_cv2[[i]][j] == pred_res_cv3[[i]][j]) pred_res_mj[[i]][j] =
pred_res_cv2[[i]][j]
if ((pred_res_cv1[[i]][j] != pred_res_cv2[[i]][j]) & (pred_res_cv2[[i]][j]
!= pred_res_cv3[[i]][j])) pred_res_mj[[i]][j] = pred_res_cv1[[i]][j] }
}
}
val_res <- numeric(10)
for(m in 1:10) {
val_res[m] <- length(which(pred_res_mj[[m]] == daten_test1[[m]][16]) ==
TRUE) / (length(pred_res_mj[[m]])) }

#ROC-Kurven - Versuch Mittelwert

roc_val <- NULL
for(i in 1:72)
roc_val[i] <- mean(as.numeric(pred_res_cv1_pr[[3]][i]),
as.numeric(pred_res_cv2_pr[[3]][i]),
attr(pred_res_cv3_pr[[3]],"probabilities")[i,3])

#keine günstigen Resultate

#ROC-Kurven - Versuch über Toleranzbereich - Probabilities für Modelle, die
#in diesen Bereich fallen, werden gemittelt
#Versuch Toleranzbereich = 0.2

pred_diff1_2 <- abs(pred_res_cv1_pr[[10]]-pred_res_cv2_pr[[10]])
pred_diff1_3 <- abs(pred_res_cv1_pr[[10]]-
attr(pred_res_cv3_pr[[10]],"probabilities")[,c(1,2,3,5,4)])
pred_diff2_3 <- abs(pred_res_cv2_pr[[10]]-
attr(pred_res_cv3_pr[[10]],"probabilities")[,c(1,2,3,5,4)])
pred_diff1_2b <- which(pred_diff1_2 <= 0.2)
pred_diff1_3b <- which(pred_diff1_3 <= 0.2)
pred_diff2_3b <- which(pred_diff2_3 <= 0.2)
attr(pred_res_cv3_pr[[10]],"probabilities") <-
attr(pred_res_cv3_pr[[10]],"probabilities")[,c(1,2,3,5,4)]
prob_res10 <- matrix(numeric(0),72,5)
for(i in 0:67) {
for(j in 1:5) {
if((length(which(pred_diff1_2b == i*5+j) == TRUE) == 1) &
(length(which(pred_diff1_3b == i*5+j) == TRUE) == 1) &
(length(which(pred_diff2_3b == i*5+j) == TRUE) == 1)) prob_res10[i+1,j] =
mean(pred_res_cv1_pr[[10]][i+1,j],pred_res_cv2_pr[[10]][i+1,j],attr(pred_re
s_cv3_pr[[10]],"probabilities")[i+1,j]) else {
if((length(which(pred_diff1_2b == i*5+j) == TRUE) == 1)) prob_res10[i+1,j]
= mean(pred_res_cv1_pr[[10]][i+1,j],pred_res_cv2_pr[[10]][i+1,j])
if((length(which(pred_diff1_3b == i*5+j) == TRUE) == 1)) prob_res10[i+1,j]
=
mean(pred_res_cv1_pr[[10]][i+1,j],attr(pred_res_cv3_pr[[10]],"probabilities
")[i+1,j])
if((length(which(pred_diff2_3b == i*5+j) == TRUE) == 1)) prob_res10[i+1,j]
=
mean(pred_res_cv2_pr[[10]][i+1,j],attr(pred_res_cv3_pr[[10]],"probabilities
")[i+1,j]) }
}
}
for(i in 1:68) {
for(j in 1:5) {
if(is.na(prob_res10[i,j])) prob_res10[i,j] = 1-
sum(prob_res10[i,which(is.na(prob_res10[i,]) == FALSE)])
}
}
```

```
prob_res_ges5 <- c(prob_res[,5], prob_res2[,5], prob_res3[,5],
prob_res4[,5], prob_res5[,5], prob_res6[,5], prob_res7[,5], prob_res8[,5])
daten_test2 <-
c(daten_test1[[1]][,16],daten_test1[[2]][,16],daten_test1[[3]][,16],daten_t
est1[[4]][,16],daten_test1[[5]][,16],daten_test1[[6]][,16],daten_test1[[7]]
[,16],daten_test1[[8]][,16],daten_test1[[9]][,16],daten_test1[[10]][,16])
roc_ges5 <- roc(as.numeric(daten_test2==5),prob_res_ges5[1:716])
```

14) Lebenslauf

PERSÖNLICHE DATEN:

Name: Andreas Hackl, Bakk.

Geburtsdatum : 29. September 1989
Geburtsort: Zwettl-NÖ

SCHULBILDUNG:

1996-2000	Volksschule in Groß Gerungs
2000-2008	Gymnasium in Zwettl/NÖ

STUDIUM:

2009-2012	Bakkalalaureatsstudium Statistik an der Universität Wien Kernfächer: Grundzüge der Wahrscheinlichkeitsrechnung und Inferenzstatistik, Mathematik und Optimierung, Lineare Modelle, Angewandte Statistik, Ökonometrie und Zeitreihenanalyse, Finanz- und Versicherungsmathematik
2012-2015	Magisterstudium Statistik an der Universität Wien Kernfächer: Wahrscheinlichkeitstheorie und Asymptotische Statistik Stochastik, Ökonometrie, Vertiefung Statistik, Quantitative Finance & Decision Support, Biometrie

EDV-KENNTNISSE	MS Office, VBA, R, SAS, SPSS, Matlab, SQL, Bloomberg, Reuters Datastream, Tableau
SPRACHKENNTNISSE	fließendes Englisch Grundkenntnisse Französisch
WEITERE AUSBILDUNG	SAS-Kurs (2014): Programmierung 1

PRAKTIKA

September 2011	Erste Group Bank AG Abteilung: Haftungsverbundrevision
Juli-August 2012	Erste Sparinvest GmbH Abteilung: Global Strategies & Research
Juli-September 2013	Erste Asset Management GmbH Abteilung: IT (Bereich: Account Management)

August-September 2014	Erste Group Bank AG Abteilung: Datamining & Business Analysis

ARBEITSSTELLEN

April 2013-Jänner 2014	Studienassistent an der Universität Wien (rechtswissenschaftliche Fakultät, Abteilung für Rechtsvergleichung)
Seit Oktober 2013	Erste Asset Management GmbH Projektmitarbeiter in der IT-Abteilung