universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Strategic classification: the impact of user information and non-linear classification models

verfasst von | submitted by

Blanka Visy

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 645
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Data Science
Betreut von | Supervisor:	Assoz. Prof. Dipl.-Ing. Dr.techn. Sebastian Tschiatschek BSc

Abstract

Machine learning models are often used by decision-makers to classify users with the goal of making accurate predictions. However, a fundamental assumption of these ML models is often violated — namely, that the distribution of instances after deployment matches that of the training data. When outcomes are important to users, they may attempt to manipulate their features with the goal to achieve a more favorable result i.e. a positive classification, a behavior known as gaming (e.g., students joining multiple sports clubs to improve university admission chances).

As this can affect both decision-makers and users, for example by reducing classification accuracy, various approaches address this distribution shift, such as the strategic classification framework by Hardt et al. [2016]. However, much of the existing literature focuses primarily on linear models and assumes that users have full information (i.e., exact information on model weights and structure) about the employed model. Yet, problems like the example above are often more complex and require non-linear classifiers. Also in many cases, full information about the model is not a realistic assumption - there can be different information scenarios based on what information the decision-makers choose to make avaible to the users.

In this thesis, I investigate how non-linearity and different information scenarios affect the outcomes of classificarion under gaming. Instead of limiting the analysis to extreme information cases like full transparency or no information, I also examine the effect of providing users with model explanations as information. Findings show that linear and non-linear models often result in different payoffs, especially for decision-makers. With both full transparency and full opacity (no information), the decision-makers get better accuracy with linear models, but there were also information scenarios where non-linear models provided better accuracy. Additionally, the effects of different information scenarios vary: users tend to incur higher costs in settings with less information, whereas decision-makers face disadvantages in transparent cases.

Overall, this study investigates the consequences of strategic classification for both decision-makers and users across different information scenarios and models, highlighting their often significant impact on both parties.

Abstract

Entscheidungsträger nutzen häufig Machine-Learning-Modelle, um Nutzer zu klassifizieren. Eine grundlegende Annahme dieser ML-Modelle wird jedoch oft verletzt: dass die Verteilung der Instanzen nach der Bereitstellung des Modells in der Praxis mit der der Trainingsdaten übereinstimmt. Wenn Ergebnisse für Nutzer wichtig sind, versuchen sie möglicherweise, ihre Features zu manipulieren, um ein günstigeres Ergebnis, d. h. eine positive Klassifizierung, zu erzielen. Dieses Verhalten wird als „Gaming“ bezeichnet (z. B. der Beitritt mehrerer Sportvereine zur Verbesserung der Studienplatzchancen).

Da dies sowohl Entscheidungsträger als auch Nutzer betreffen kann, beispielsweise durch eine verringerte Klassifizierungsgenauigkeit, befassen sich verschiedene Ansätze mit dieser Verteilungsverschiebung, wie zum Beispiel das Framework von "Strategic Classification" [Hardt et al., 2016]. Ein Großteil der bestehenden Literatur konzentriert sich jedoch primär auf lineare Modelle und geht davon aus, dass Nutzer über vollständige Informationen (d. h. genaue Informationen zu Modellgewichten und -struktur) über das verwendete Modell verfügen. Probleme wie das obige Beispiel sind jedoch oft komplexer und erfordern nichtlineare Klassifikatoren. Zudem ist die Annahme vollständiger Informationen über das Modell oft keine realistische Annahme – je nachdem, welche Informationen die Entscheidungsträger den Nutzern zur Verfügung stellen, können unterschiedliche Informationsszenarien entstehen.

In meiner Arbeit untersuche ich, wie sich Nichtlinearität und verschiedene Informationsszenarien auf Modelle unter der Annahme von „Gaming“ auswirken. Anstatt die Analysen auf extreme Informationsfälle wie vollständige Transparenz oder keine Informationen zu beschränken, untersuche ich auch die Auswirkungen der Bereitstellung von Modellerklärungen wie LIME als Informationen für Nutzer. Die Ergebnisse zeigen, dass lineare und nichtlineare Modelle oft zu unterschiedlichen Auszahlungen führen, insbesondere für Entscheidungsträger. Sowohl bei vollständiger Transparenz als auch bei überhaupt keine Informationen erzielen Entscheidungsträger mit linearen Modellen eine höhere Genauigkeit, es gab jedoch auch Informationsszenarien, in denen nichtlineare Modelle eine höhere Genauigkeit lieferten. Darüber hinaus variieren die Auswirkungen verschiedener Informationsszenarien: Nutzer tragen tendenziell höhere Kosten in Situationen mit weniger Informationen, während Entscheidungsträger oft in transparenten Fällen Nachteile erleiden.

Zusammengefasst untersucht diese Studie die Konsequenzen von "Strategic Classification" für Entscheidungsträger und Nutzer in verschiedenen Informationsszenarien und -modellen.

List of Figures

List of Tables

List of Algorithms

Contents

1 Introduction

Machine learning explores algorithms and techniques for automating solutions to complex problems that are challenging to address using traditional programming methods. Often, its aim is to learn a model or a set of rules from a labeled dataset (known as the training dataset) so that it can accurately predict the labels of data points not present in the dataset [Baştanlar & Özuysal, 2014]. As vast amounts of complex, high-dimensional data are generated across domains, machine learning becomes essential for extracting insights, often in almost real time. With its ability to handle non-linear relationships, it increasingly supports critical decision-making processes, and more and more crucial choices are now in the hands of machine learning algorithms. To name some examples, these algorithms dictate the information presented to us on the internet [Perlich et al., 2014], influence employment decisions such as hiring and promotions [Brynjolfsson & Mitchell, 2017], impact who qualifies for loans [Björkegren & Grissen, 2018], and even play a role in decisions regarding bail and parole [Kleinberg et al., 2018].

However, a fundamental assumption in underlying conventional machine learning is that the distribution of unseen instances is identical to that of the training dataset [Dundar et al., 2007]. But this may not hold true in many real-world scenarios. For example, perpetrators of credit card fraud adapt their activities to evade detection mechanisms employed by credit card companies. Similarly, creators of email spam tailor message templates to bypass current spam filters [Brückner & Scheffer, 2011]. Further examples of this behavior can be seen at New York's high school exit exams [Dee et al., 2019], and Chinas automatic air pollution monitoring [Greenstone et al., 2022]. On Google, there are over 32 million search results for the phrase "hack your credit score", and also numerous videos on the same cue. For example, not closing old credit cards is mentioned as a suggestion to improve the credit score [MoneyNerd, 2023; Gobler, 2022]. This behavior, where individuals manipulate their attributes to receive more favorable outcomes in scenarios where models classify them, is referred to as "gaming" in the literature [Hardt et al., 2016]. As these examples illustrate, many predictive machine learning model used in decisions affecting human lives can be affected by the concept of gaming.

In this thesis, I examine such learning scenarios with the framework of strategic classification [Brückner & Scheffer, 2011; Hardt et al., 2016]. Here, decision-makers are required to select a classifier before individuals being classified can learn about the model and potentially modify their features. To illustrate, consider university admissions. The decision-maker (the university) utilizes a classifier for grouping applicants into two categories ("accepted" and "denied"). The decision-maker aims to achieve high accuracy meaning that

truly suitable candidates are classified as accepted and others are denied. Since Contestants (applicants) may be highly motivated to secure a positive outcome, they might attempt to manipulate their features (e.g., joining multiple sports clubs without participation) to improve their chances. However, this behavior may not accurately reflect their true suitability for the university and the accuracy can decrease because of this change in behavior. In the end, the decision-maker gains a payoff in terms of accuracy, while Contestants receive the utility of the decision outcome on their new features minus the costs they incurred in altering their features as a payoff.

For decision-makers, gaming often causes a decrease in model performance because the assumption of identical distributions between training and deployment data is not met [Hardt et al., 2016]. As an example, if Contestants only change their features without their true label changing, the number of false positives increases. This outcome is undesirable for decision-makers.

1.1 State of the art

Various approaches exist for addressing distribution shifts, such as keeping decision rules confidential to prevent gaming. However, this is not always possible since there is a growing societal demand for a 'right to explanation' regarding how algorithmic decisions are made [Goodman & Flaxman, 2017]. This need is also getting formalized in law systems as in articles 13-15 of the European Union's General Data Protection Regulation. This mandates that "meaningful information about the logic" of automated systems must be made available to data subjects [2016/679, 2016-05-04]. On top of that, even if decision-makers retain the logic, information may leak, or it is often possible for an outsider to learn information from classification outcomes. Therefore, counting on keeping the decision rule private is not a reliable solution.

Another approach named performative prediction involves retraining algorithms to adapt to shifts, though this is costly and may not always guarantee convergence [Perdomo et al., 2020; Ghalme et al., 2021]. Strategic classification, a subfield of performative prediction, focuses on designing robust classifiers that remain accurate despite gaming [Hardt et al., 2016; Levanon & Rosenfeld, 2021]. This includes research on online learning [Dong et al., 2018; Chen et al., 2020; Ahmadi et al., 2021], generalization and PAC learning[Zhang & Conitzer, 2021; Sundaram et al., 2023], causality and effects of the distribution shift [Miller et al., 2020; Rosenfeld et al., 2020; Shavit et al., 2020; Bechavod et al., 2021], and social perspectives such as fairness [Hu et al., 2019; Milli et al., 2019]. Some studies distinguish between genuine improvement and faking, where Contestants may change their features for a better classification or just to game the system [Kleinberg & Raghavan, 2020; Barsotti et al., 2022].

However, most of the examined models in these studies only examine linear classifiers. This is a huge limitation as popular machine learning models as neural networks and decision trees are non-linear in general. Therefore, there is a need to understand strategic classification in cases where decision-makers use such non-linear models.

Moreover, these models typically make the following transparency assumption: the Contestants have full information about the deployed decision rule. In the context of the university example, this means that students would exactly know the classification model of the university. But this assumption is not realistic in practice. Transparency regulations aren't universally applicable. When models are kept private, there is often no unintentional information leakage, preventing Contestants from accessing information about the model. Lastly, even if decision-makers disclose the decision rule, there's no guarantee that individuals without domain knowledge of classifiers will fully understand it. There is a line of research that was looking into strategic classification where Contestants only rely on information observed by classified examples [Ghalme et al., 2021; Bechavod et al., 2022]. They demonstrate that disclosing model information does not prevent gaming but can increase disadvantages for both players. As illustrated by data stories in Ghalme et al. [2021], Contestants may make costly adjustments yet receive no change in classification or even a worse outcome (e.g. a truly suitable candidate changes features and gets classified as -1) due to their noisy model approximations. Additionally, decision-makers may encounter increased false positives and negatives. In conclusion, the drawbacks of strategic behavior become more pronounced when opacity is a factor. However, these papers only consider two cases: either Contestants have complete knowledge of the model, knowing the exact weights, or they have no available information from side of the decision-makers except for a small labeled sample of features and their classification label. Anyhow, there are other realistic scenarios that fall between these two extremes. As previously mentioned, EU regulations require decision-makers to provide "meaningful information about the logic involved" [2016/679, 2016-05-04], which indicates that using model explanations could also be possible.

With machine learning models erupting, there is also a high emphasis on finding explanations for the predictions of black box models. Algorithms are developed to make the general decision rule [Štrumbelj & Kononenko, 2014] or even separate predictions understandable [Ribeiro et al., 2016]. It is realistic that decision-makers will also rely on these explanatory tools to fulfill transparency goals. **The goal of this thesis is to deploy linear and non-linear decision rules in different information scenarios and compare the results from viewpoints of all the players.**

1.2 My contributions

To sum up, gaming can affect the decision-making processes that are frequently conducted by automated machine learning (ML) models. The existing literature examines many aspects of this trend but has two major limitations: the assumption of linearity in the classifier and the assumption that Contestants have full information. In practice, these assumptions often do not hold. The contribution of this thesis to the existing literature is the thorough analysis of non-linear models in strategic classification in the following information scenarios:

1. Full information about the model: Contestants have access to a black box function that maps any set of features to the classification outcome,
2. Partial information about the model: knowledge limited to certain explanations, such as LIME values,
3. No information about the model: no direct information provided by the decision-makers, with Contestants only having access to a labeled sample.

As (1) and (3) have already been studied for linear models, here the comparison with respect to the payoffs of players in non-linear models with previous results is a main point. The second information scenario, where decision-makers employ explanations to make the decisions interpretable for the agent, is a completely new setting as the date of the thesis.

My research questions are:

- How does the outcome of strategic classification differ for both the decision-makers and Contestants in strategic classification when non-linear models are involved?
- How do the payoffs of decision-makers and Contestants compare in the different information scenarios?
  - Does making the models completely or partially transparent result in higher accuracy for decision-makers? If not, what is the magnitude of performance loss due to transparency?
  - How do the features and predicted labels of the Contestants change in different information scenarios? In opaque scenarios, are Contestants classified as they anticipate in their optimal response?

My hypotheses, based on the results of existing literature, are as follows. I expect that in the non-linear case, when Contestants attempt to modify their models, the optimization process might pose greater chal-

lenges: i.e., Contestants need more iterations to find similar objective values, hence in the same amount if iterations linear solutions will be better for them.

Additionally, I believe that the differences between linear and non-linear models will manifest differently in each case. Regarding the decision-makers payoffs (accuracy), I expect non-linear models to be more favorable, as they typically have with higher accuracy before the feature modification of the users (i.e., when the IID assumption is uphold) and are subject to fewer user modifications due to the increased complexity of the problem that results in Contestants failing to find modifications.

Finally, I anticipate that model accuracy will decline more with increasing transparency, as Contestants would make increasingly precise adjustments that could negatively impact overall accuracy.

1.3 My methodology

In order to assess these questions, the classical framework of strategic classification proposed by Hardt et al. [2016] is used. To address non-linearity, instead of finding a closed form solution to the optimization problem of the Contestants, approximation algorithms are used. If Contestants have all the information about the model, they also have a black box function of their features and prediction outcomes that these algorithms can be applied to.

Following information scenarios were tested. (1) On one hand, Contestants can have all the information about the model. For non-linear machine learning algorithms this means that Contestants get to know a function f that maps from any attribute values to binary classes. (2) For the second case, Contestants get to know the LIME predictions of their classification [Ribeiro et al., 2016]. They change their features in the explained direction by weights defined by the feature importances in LIME. (3) In the third case, publishers do not reveal any information. Contestants take a random sample from other Contestants in the training set of the model and their predicted labels and estimate a linear model as in Ghalme et al. [2021]. On top of that, I want to compare an even more naïve approach for this no information scenario. Contestants take the same random sample, but instead of using a linear estimator, they modify their features assuming a K-Nearest Neighbors (KNN) model. This means they adjust their features as much as possible for their costs to resemble those of individuals in the sample who were classified positively.

To answer my research questions, I conducted experiments in the following settings. First, I used a synthetic clustered dataset with Gaussian clusters. Next, I use a the popular moons dataset (often used in clustering problems, see [Pedregosa et al., 2018]) to also have a better comparison for non-linear models. Finally, I compared the results with the dataset from Ghalme et al. [2021] as a benchmark. For comparison, I use vi-

sual elements. Additionally, I created data stories similar to those in Ghalme et al. [2021], where detailed descriptions of characteristic cases are provided.

1.4 My results

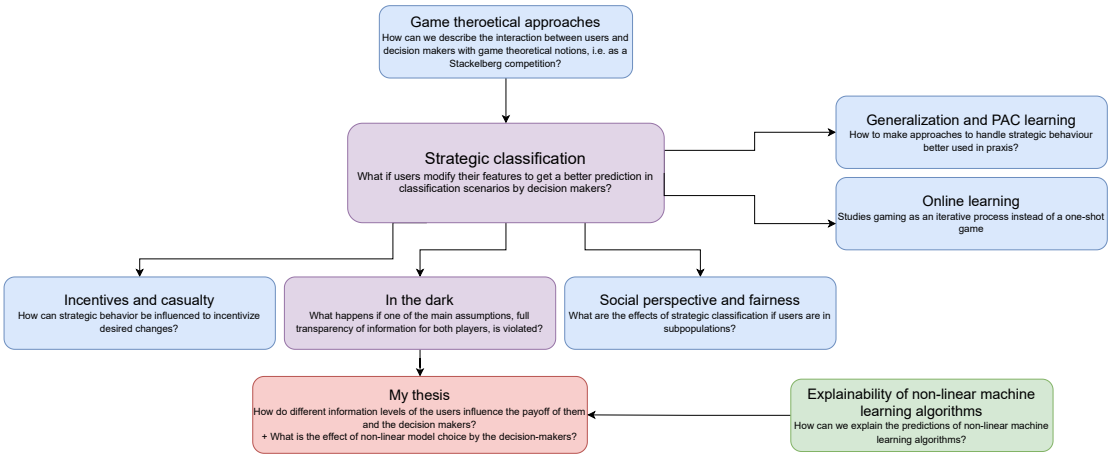
Impact of Non-Linear Models on Strategic Classification: My experiments showed notable differences in payoffs between linear and non-linear models, particularly for the Jury. These differences depended both on the information shared by the Jury and used by Contestants as well as on the dataset. Linear models performed better for the Jury in terms of accuracy under full information and in the no-information imitation setting, while non-linear models outperformed in the no-information estimation scenario. For Contestants, individual average payoffs were largely unaffected by model type, though in some cases, overall user welfare was significantly higher with linear models.

Effect of Information Disclosure on Jury Payoff Full transparency led to a 5-20% performance loss compared to the best scenarios for all model types. Surprisingly, partial information often benefited decision-makers more than complete opacity, suggesting that tools like LIME may offer a strategic advantage. The results align with Ghalme et al. [2021] without contradictions in non-linear models or varying information scenarios.

Effect of Information Disclosure on Contestant Payoff Full transparency yielded the best results for Contestants. Providing partial information was not particularly helpful, as the no-information scenario often resulted in better outcomes. The highest difference in the expected and true classified labels of the users was in the no-information case.

Characteristics of best responses in different information scenarios Contestant behavior varies across information scenarios. Full information and no-information with estimation appear similar, though the latter includes more suboptimal changes relative to the true decision boundary. In partial information, many users move to similar feature values. In opaque scenarios, there are always a significant amount of users who are not classified as they expect based on their estimations, with the no-information estimation case showing 2–3 times more such users than the partial information case.

In my thesis, I first provide a detailed literature review in Chapter 2. Next, I introduce the necessary background and notation in Chapter 3. Chapter 4 outlines the methodology and models used in this study. Chapter 5 presents the experiments I conducted, along with their setup, while Chapter 6 details the results. In Chapter 7, I interpret the findings, relate them to previous studies, and address the research questions. Finally, in Chapter 8, I summarize the results and discuss potential directions for future research.



Source: own edit.

Figure 1: A map of the related work introduced by this thesis. Purple color shows the most related work in Section 2.1. Blue color includes work that covers less directly related aspects in Section 2.2. Green indicates a different topic from strategic classification, that is still important to my work in Section 2.3.

2 Literature review

This chapter presents a comprehensive overview of key historical findings in strategic classification, discusses their relevance to the topics of in this thesis. [1] First, I introduce the most influental studies on this thesis in Section 2.1. Section 2.1.1 outlines the fundamental framework that supports the majority of recent studies in the field. This framework serves as a crucial foundation to this thesis as well, many definitions and notations are rooted in the works introduced in this section. Section 2.1.2 offers a summary of strategic classification in diverse user information scenarios, representing some of the closest related work. Then, I also present further research directions in the field of strategic classification in Section 2.2. Sections 2.2.1, 2.2.2, 2.2.3, 2.2.4, and 2.2.5 cover less directly related aspects. They discuss various perspectives, recent challenges and solutions in strategic classification, such as casuality and incentives, generalisation, PAC and online learning. Finally, Section 2.3 offers insights into explainability methods for non-linear machine learning algorithms. This section aims to delve into explainability methods as one of the key objectives of the thesis is to explore their implications in the context of strategic classification.

The structure of my literature review is summarized on Figure 1.

2.1 Closest related work in strategic classification

In this subsection, I first intorduce the related work that contains the most similarities to my thesis.

[1] For grammar checking and formatting purposes, OpenAI [2024] was used. Nevertheless, all intellectual content is the sole responsibility of the author.

2.1.1 General framework

Hardt et al. [2016] introduced the "Strategic classification framework". They formalize the manipulation of features and a framework to understand gaming. Similarly to Brückner & Scheffer [2011], they define the game between decision-makers and Contestants as a Stackelberg competition. Within this framework, the decision-maker commits to a classifier upfront, while the Contestant, in response, strategically adjusts their features in a process they refer to as gaming. Additionally, they establish the payoffs and formalize the optimal responses of both players through optimization problems, assuming both players possess full information. Subsequently, they demonstrate that solving these optimizations for non-separable cost functions is NP-hard. However, for separable cost functions, they introduce an algorithm aimed at making classifiers more robust against gaming strategies.

Their empirical evaluation, conducted on a spam filtering problem of a Brazilian social network, illustrates how their algorithm outperforms a non-modified Support Vector Machine (SVM) as gaming intensity rises. The degree of gaming is regulated by the ratio of utility to costs: a higher multiplier on utility decreases the relative weight of costs in the optimization process and hence promotes gaming.

They also consider several potential "inaccuracies" that may arise in practice: firstly, decision-makers lack access to the true classification of the entire distribution and can only observe a sample; secondly, decision-makers possess incomplete information regarding the true costs; thirdly, they lack awareness of the extent of gaming occurring.

In conclusion, the paper establishes a crucial foundation for examining gaming phenomena from a formalized perspective. However, this thesis diverges in several key aspects: Firstly, while they solely utilize a specific spam dataset, I intend to explore multiple cases. Secondly, they exclusively focus on linear SVMs, whereas my research encompasses non-linear algorithms. Lastly, their analysis is confined to separable cost functions, whereas I employ cost functions with quadratic dependencies.

2.1.2 In the dark - studies on user information

In the framework of strategic classification initiated by Hardt et al. [2016] substantial research has emerged. However, the literature typically treats the transparency/opaqueness of a classifier as a dichotomy, assuming that only transparent classifiers require a robust learning strategy. Nevertheless, in many real-life scenarios involving high-stakes classification, the situation is more nuanced.

Ghalme et al. [2021] studied cases, where users do not know the true classification model of the decision-makers f, but still aim to game and therefore try to learn from predictions and estimate a classification

model $\hat{f}$. They introduce the concept of *"Price of Opacity (POP)"*: i.e., the difference in expected error of a robust classifier f when the model is transparent to that to when it is opaque, hence in different information scenarios. To make the assumption of opacity specific, they assume that Contestants observe a random sample drawn from the distribution of Contestants iid and observe the labels of that sample made by f, the truly used model of the decision-maker. They do not only want to see the payoffs of the Jury, but also explore the so named enlargement set. This enlargement set E includes all Contestants which decision-makers classify differently under the different information scenarios. However, they show that as these points (of the Contestants) tend to concentrate around the classification boundary, this mass of this enlargement set can be large even if the $\hat{f}$ is a good approximation of f. Using this new definition of the POP they state their main result, namely a condition on having strictly positive POP. Their theorem states that if the probability mass on the enlargement set exceeds the base error rate of f, then the POP is strictly positive and decision-makers get a better payoff for the transparent strategy than for opaque one.

They also complement their theoretical results with experimental evaluation on a synthetic and a loan dataset. In both cases, they use linear SVMs and the algorithm proposed by Hardt et al. [2016] as the classification models of the decision-makers. Both models are tested in three settings: (1) non-strategic, where Contestants do not game, (2) fully informed: where Contestants perform gaming and know the classification model exactly and (3) in the dark: where Contestants perform gaming and but only know labeled examples and have to approximate the classification model. They visually compare the POP by plotting the error of the two models (linear SVM and the algorithm proposed by Hardt et al. [2016]) in the different settings. To also understand user payoffs, they project their features and the decision boundary to 2D and mark the changes they make, as well as describing some characteristic data stories from users.

This paper shares many similarities with the goal of this thesis. However, there are some key differences. In this thesis, two additional information scenarios are evaluated. One is situated between no information and "full" information, while the other assumes no information but a different best response strategy of the users with imitation. Moreover, non-linear models are also tested.

Bechavod et al. [2022] also examines strategic classification in a setting where users do not have complete information about the classifier. They propose a model for the principal-agent interaction in strategic classification, where users from different sub-populations learn information about the classifier from their peers, similar to Ghalme et al. [2021]. However, their focus diverges from Ghalme et al. [2021] by concentrating on the societal implications of opacity. Their goal is to formulate machine learning algorithms that incentivize "good" strategic behavior, i.e., true improvements, even when the model is not fully transparent

to the users.

Technically, they assume that there are multiple sub-populations of the population of users X with different distributions. In these populations, users are risk-averse, meaning they only wish to move in directions that can surely improve their outcomes. They are also fully rational and have no other information about the model except for the labels of their already classified peers. They operate with a mixed cost function comprising quadratic and linear components. The authors also assume that there exists a ground truth assessment rule, which is linear and known to the decision-maker. This differs from the previous work of Ghalme et al. [2021], where the ground truth rule is unknown, and only the labels are known to the decision-maker.

In the game setting of the paper, classification is an iterative process where users are classified one-by-one. The first few users in each sub-population have no information about the model, so they cannot estimate the model better than a random guess and as they are risk-averse, they do not engage in gaming. After enough peers in their sub-population have been classified, they conduct empirical risk minimization (ERM) on their features and labels. The goal of users is to be positively classified, as in classical strategic classification. However, the decision-makers aim to maximize *social welfare* rather than accuracy. Social welfare is defined as the average improvement of users' true labels after best-responding, summed across groups and averaged over the agents and the randomness of the labels.

After defining the game setting and their assumptions, they proceed to show the closed-form solution for the best response that maximizes users' utility and the decision-makers equilibrium classification rule that maximizes social welfare. They also state conditions for achieving equal improvement across subpopulations. Additionally, they argue that under full information and their other assumptions, the do-no-harm principle is guaranteed, meaning the number of true positives does not decline. They then evaluate their algorithms on two datasets, dividing them into subpopulations based on certain features. They show that if the costs for the groups were the same, the number of true positives strictly increased. Another key finding of the study is that people from significantly different age groups or countries can have very different understandings of the scoring rule, which can lead to disparate improvements across subpopulations.

This work is differs from my thesis in multiple aspects. Their focus lies on social improvement, while this thesis concentrates on comparing the payoffs of players in different information scenarios. They construct subgroups, a distinction not made in this thesis. Additionally, they exclusively consider linear settings.

Barsotti et al. [2022] also consider classifiers where users do not have full information about the classifier used by decision-makers. They distinguish two forms of gaming: *faking* and *improving*, whether gaming contains a difference between a user's true features and their reported features or not. The true label of

users is based on their true features, however decision-makers label them based on their reported features. When users get classified positively because they misreported their features, they call it faking, otherwise the feature manipulation is called improvement. They formalize a difference for these two forms of gaming as well, incorporating detection algorithms designed to identify and penalize faking with some success rate. Thus, in the user's best response problem, the optimization process includes a choice between true improvement (with higher costs) and faking (with lower costs but a potential high penalty). Additionally, they suggest that users may respond through *behavioral imitation*, highlighting that imitation and social learning are major drivers of behavior adaptation. Consequently, users might directly imitate the behavior of others rather than merely inferring the classification threshold.

They also explore different information scenarios, controlling the information users can infer about the decision-making model by adding noise to their estimation of the threshold. With high noise, the user's best possible estimation is a random guess of the model threshold, while with low noise, users have full knowledge of the decision-making model. Their results indicate that transparency leads to a greater increase in the number of false positives compared to scenarios with higher noise (opacity). They suggest that transparent feedback is highly desirable and investigate mechanisms to mitigate these risks associated with transparency by the decision-makers. They argue that if detection is effective and the detection probability is sufficiently high, the risk - i.e. the increase in false positives due to transparent feedback - can be minimized.

This paper shares many similarities with the goal of the thesis. It also compares different transparency scenarios and incorporates behavior imitation into users' best responses. However, there are significant differences. The main difference is that in this thesis, the level of information is controlled not by a noise parameter but by the amount and type of information that decision-makers publish. Additionally, this thesis does not explicitly distinguish between true and observed features.

2.2 Other topics in in strategic classification

Next, I would like to introduce some other, less closely related topics to this thesis in the field of strategic classification.

2.2.1 Early work on feature manipulations

Manipulating features to get a better outcome in a decision-making situation traces back through history far before the work of Hardt et al. [2016], with cautionary principles like Goodhart's Law emerging in 1975

[Goodhart, 1976]. It warns against the pitfalls of fixating a decision rule, stating: "If a measure becomes the public's goal, it ceases to be a good measure." This principle, originally in financial policy, now resonates in computer science, serving as a reminder to consider broader implications beyond mere metrics.

Early work often aimed to enhance data mining algorithms to reduce their susceptibility to feature manipulation. Dalvi et al. [2004] highlighted the risk of data being actively manipulated by users attempting to cause the classifier to produce false negatives and proposed a classifier that is optimal for decision-makers given the users' optimal strategy. However, they perceived manipulation as a zero-sum game, where one player's gain meant another's loss. In the context of this thesis, such assumptions may not always hold true. There can be cases where both parties benefit from manipulation. For example, consider a car insurance company where users opt to take additional driving courses to obtain a lower monthly insurance fee. In this scenario, both the decision-maker (the insurance company) and the Contestant or the user (customer) benefit. The insurance company faces a potentially lower risk of accidents, while users receive a better rate on their insurance. In this thesis, a zero-sum assumption is not made. Situations where both players can benefit or even where both players are adversely affected are considered.

There has been a line of research also delving into game theoretic formulations of gaming. In a study dating back to 2009, researchers investigated prediction games where the players decide simultaneously [Brückner & Scheffer, 2009]. They identified conditions under which these games exhibit a unique Nash equilibrium, exploring scenarios where the cost functions of the Contestant and decision-maker were either antagonistic or not. This differs to this thesis because it considers a simultaneous setting, i.e., the players decide at the same time and do not have the possibility to respond to the actions of the other players.

In 2012, Brückner et al. [2012] presented a revised work of the theorems outlined in the preciously described work of Brückner & Scheffer [2009]. Furthermore, the researchers introduced a kernelized variant of the methodologies of Brückner & Scheffer [2009] and empirically assessed the outcomes for Nash logistic regression and Nash support vector machine models. However, this also differs from the setting of this thesis as they again consider simultaneous decisions.

In 2011, the same authors extended their inquiry to a scenario where players do not act simultaneously [Brückner & Scheffer, 2011]. Here, decision-makers take the role of leaders by committing to decision rules initially. Then, Contestants react, similarly to the framework of strategic classification described in the introduction. The authors proceed to derive an optimization problem to determine the solution of the Stackelberg prediction game. Additionally, they conduct experiments using multiple spam labeling datasets. This contrasts with the focus of my thesis. While that approach seeks an optimal classification for the decision-

makers under the assumption of full knowledge, my thesis centers on the implications when that assumption does not hold. Additionally, a Euclidean cost function is used in their work opposed to a mixed cost function described in Section 3.1 in this thesis, and in their work experiments are exclusively conducted with linear decision rules.

Following the trajectory of game theory, a subsequent work suggests that full-information models may not always be plausible [Großhans et al., 2013]. This work delves into cases where decision-makers lack complete knowledge about the objectives of the Contestants, framing this problem as a Bayesian game. Experimentally comparing the Bayesian equilibrium strategy to the Nash equilibrium strategy, the minimax strategy, and regular linear regression, the work demonstrates that playing the identified Bayesian equilibrium yields a smaller (RMSE) in Bayesian models compared to the other mentioned models. This approach is different because it still builds and focuses on game theory. It is similar in that it does not assume full information, yet it examines the lack of information from the perspective of the decision-makers.

2.2.2 Incentives and casualty

Strategic classification assumes that users tend to modify their features to receive better predictions. On one hand, this modification can be made between the true and reported labels, viewing gaming as the difference between these two, as considered by Barsotti et al. [2022] and described in Section 2.1.2. However, there is another perspective. Miller et al. [2020] distinguish gaming based on whether the change in features causes a change in the user's true label. Thus, they formally differentiate between the two possible outcomes of users manipulating their features:

- *Faking* refers to feature changes that alter the classification outcome without changing the true label. Some literature also uses the term "gaming" for this concept. However, since traditional literature defines any manipulative behavior as gaming, this thesis uses the term "faking" to avoid confusion.
- *Improvement* refers to feature changes that modify both the classification outcome and the true label.

In many cases, genuine improvements resulting from strategic behavior are desired for decision-makers. For example, an insurance company may aim to motivate safer driving. This benefits not only the company but also the users and society as a whole, resulting in fewer accidents.

Research has as well focused on this distinction between faking and true improvement. On top of that, there are studies on how strategic behavior can be influenced to incentivize desired changes and on understanding the causal relationships within decision-making models in strategic settings.

While Bambauer & Zarsky [2018] was the first to suggest a distinction between improvement and faking from a legal perspective, their work does not directly connect to strategic classification. Nonetheless, they study the same problem setting.

In later work of Kleinberg & Raghavan [2020], the authors were also early to introduce a casual setting and approached the problem from a more economical viewpoint. Here, instead of binary classification, they deal with a continuous score that users aim to maximize. The goal of the decision-maker is not merely the accuracy of learning but rather finding mechanisms that incentivize changes in a particular direction. They demonstrate that if such a mechanism exists, a linear one is sufficient. However, their approach is quite different to this thesis due to using regression instead of binary classification and also because of different definition of payoffs due to that. On top of that, their approach is also more on how should classifiers be designed to induce the right kind of effort, while this thesis focuses more on the payoffs in different information settings.

In the work of Miller et al. [2020], the authors provide a comprehensive summary of previous literature on causality and strategic classification. They formalize definitions and unify the framework of strategic classification with the framework of structural causal models. Furthermore, they formally define the difference between improvement and faking in strategic classification, as also described in the introduction of Section 2.2.2. In their framework, they consider endogenous variables (variables that are influenced by other variables within the model) and exogenous variables (variables that are *not* influenced by other variables within the model), along with a set of structural equations that determine the values of the endogenous variables. They also assume that users only approximately best respond, meaning that the feature change of the users is not necessarily optimal, but rather close to it. Additionally, they state that incentivized design must inevitably grapple with causal analysis – it is possible to learn causal understanding, and without it, incentive design is unlikely to succeed.

Another work on causality of strategic classification by Rosenfeld et al. [2020] stresses the crucial role of decision-makers in considering the implications of their classifiers' recommendations, prioritizing models that not only improve predictive accuracy but also encourage safe behaviors. For instance, they argue that even if a model suggests a negative coefficient for alcohol consumption in cardiac risk prediction, it remains unsafe for individuals to increase their alcohol intake solely based on the model's output. They propose that strategic classification should be approached through the lens of *model selection*, where the balance between a models accuracy and safety is key. To address this, they introduce lookahead regularization, which weighs accuracy against induced changes. This method penalizes models that fail to sufficiently

improve outcomes based on user actions. Distinguishing factors from this thesis include their continuous measurement of improvement, the incorporation of certainty and uncertainty in the regularization term, as well as the utilization of differentiable optimization layers instead of utility maximization for best response. However, their perspective on model selection aligns with this thesis, as the research questions of the thesis has the aim also contribute to identifying optimal models in strategic settings.

Shavit et al. [2020] contribute to a similar line of research as the previously mentioned work of Miller et al. [2020] and Kleinberg & Raghavan [2020]. Their study introduces three different pay-off definitions for decision-makers: maximizing true poaitive labels, predictive accuracy, and estimating the true underlying model. They also note that changes often cannot be made independently between variables. They demonstrate that in the absence of hidden features, all three objectives can only be jointly satisfied with perfect approximation. Otherwise, they provide algorithms to finding optimal classifiers for all the three different cases. Moreover, they argue that decision-makers could benefit from actively informing strategic agents, rather than just passively revealing their decision rules. However, their study is limited to the linear case, focusing on linear regression without exploring other linear models. Additionally, they focus primarily on the theoretical proofs of their algorithms, whereas in this thesis, there is more emphasis on experimental evaluations.

Bechavod et al. [2021] distinguish between "meaningful" variables, which directly impact the true label when altered, and "non-meaningful" features that have no effect. They employ linear models both as the true underlying model and the decision-making classifier, re-training over time. They demonstrate that even basic actions by the decision-makers enable them to accurately identify meaningful features and motivate users to invest in them, thereby encouraging improvement. This paper differs from this thesis primarily in its emphasis on retraining over time, as opposed to a one-shot setting of my thesis, and its focus on linear models rather than non-linear ones.

Haghtalab et al. [2020] aim to design classifiers that maximize improvement across the population, i.e. social welfare. They assume that features only have a causal effect on the label and that there also exists a true underlying scoring function. They distinguish between a true high-dimensional feature vector and a visible lower-dimensional projection, such as a representation of the individual that may include vaccination records and health history (visible features) but excludes factors like exercise and dietary habits (true features). Their research explores the algorithmic aspect of identifying the welfare-maximizing decision rule within two specific settings in their model: linear classifiers and classifiers with linear thresholds. Additionally, they provide sample complexity guarantees for learning the optimal rule from samples alone. Key

differences to this thesis include the decision-maker's goal of maximizing welfare rather than accuracy, the distinction between visible and true feature spaces, as well also a theoretical focus rather than applied experiments.

2.2.3 Generalisation and PAC learning

Levanon & Rosenfeld [2021] highlight a significant gap despite the substantial recent work in strategic classification: algorithms for classification incorporating gaming factors remain underutilized, stating that work in the space of strategic classification has been predominantly theoretical. Their research prioritizes practical aspects of strategic classification. This aligns with the practical focus of this thesis. Additionally, they emphasize that strategic classification isn't necessarily a zero-sum game; the objectives of the system and its users often exhibit overlap.

In their work, they propose an approach called Strategic Empirical Risk Minimization (SERM), in which they embed users' best response mapping directly into the classifier's learning objective. However, since the best response is determined by an argmax operation and is therefore non-differentiable, they employ a differentiable proxy in its place. They also challenge the assumption of fixed costs, suggesting costs can be influenced and explore scenarios where costs are unknown. Regarding social aspects, they explore multiple regularization methods aimed at incentivizing learned models to foster positive social outcomes, emphasizing the system's responsibility in balancing accuracy and promoting favorable social outcomes. This article mainly diverges from my thesis in their assumption that users are to be fully informed. However, it's important to note that both this study and my thesis employ the proxy method and utilize the same cost function.

Zhang & Conitzer [2021] follow a similar path by integrating the modified distribution of users into Empirical Risk Minimization (ERM), much like SERM. However, their focus lies in examining the problem through the lens of PAC (Probably Approximately Correct) learnability. They introduce the notion of incentive compatibility, where users are disincentivized from gaming the system. However, their work primarily consists of theoretical analysis and guarantees as opposed to this thesis.

Sundaram et al. [2023] present a novel approach that also integrates PAC-learning with strategic classification considerations. They introduce a richer setting where points can be assigned mixed labels based on their preference for +1 or -1, while assuming that points retain their true labels despite gaming. They assert that any strategic classification problem can be agnostically PAC-learned through the empirical risk minimization paradigm with given sample size. Additionally, they define the strategic VC-dimension $SVC(H, R, c)$, which encapsulates the learnability of any hypothesis class H when test data points' strategic behaviors are

influenced by the cost function c and preference values from any set R. Their findings indicate that the more general instance-wise costs pose significantly greater challenges in terms of both statistical learnability and computational tractability.

2.2.4 Online learning

There is a line of research that studies gaming as an iterative process instead of a one-shot game.

The work of Dong et al. [2018] addresses strategic classificaton as a repetitious online learning procedure. In their setting decision-maker iteratively publishes a classifier, one users best responds, the decision-maker learns from this response, and then republishes. The study assumes that neither the original set of features nor the cost functions are known, so the decision-maker can only learn from the modified feature of the users. The focus of the article is on linear classifiers, with user payoff based on the output functions value (encoding the probability of a positive label) rather than the label. One of the significant contributions of this study is the introduction of the concept of Stackelberg regret. Moreover, they propose a regret minimization algorithm for convex optimization, which is similar to the bandit approach in general online learning as discussed by Flaxman et al. [2004]. Then, they identify the appropriate conditions under which the best-response function of a user, written as a function of the decision-maker's classifier, is concave. This concavity ensures that the decision-maker's optimization problem is convex. They develop an algorithm for minimizing Stackelberg regret in this convex setting.

Chen et al. [2020] explore strategic classification but differs from the work of Dong et al. [2018] by not relying on convexity, instead assuming less smooth utility functions for decision-makers and less smooth loss functions for users. However, both studies share the same setting where users are strategic, and decision-makers lack knowledge of the users' costs and unmanipulated features. They compare two types of regret: Stackelberg regret, as discussed by Dong et al. [2018], and external regret. They argue that achieving sublinear external and Stackelberg regret simultaneously is generally impossible, indicating a strong incompatibility between the two. On top of that, they propose the GRINDER algorithm, an adaptive discretization algorithm, and provide simulations to demonstrate its effectiveness.

Ahmadi et al. [2021] combine a fixed utility model with non-smooth utility for users. Their main contribution is an online learning algorithm that is robust to manipulation, capable of finding a linear classifier with a bounded number of mistakes when costs are known. They further generalize the approach to scenarios without knowledge of costs and with heterogeneous agents.

However, these above mentioned studies differ from this thesis in a very important point: they look at

strategic classification as an iterative game, while in my thesis, strategic classification has a one-shot setting similarly to [Hardt et al., 2016; Ghalme et al., 2021].

Perdomo et al. [2020] study iterative retraining from a broader perspective. They categorize all predictions where the prediction influences the target as performative, with strategic classification being a specific instance of such problems. Unlike other frameworks, they do not assume a specific cost function for the distribution shift but rather focus on the sensitivity of the data-generating distribution to changes in model parameters, making their approach more general. They introduce *Repeated Risk Minimization* (RRM) that they define as a process of retraining the model repeatedly on shifted distributions. They also define the concept of performative stability, where a model has minimal loss on the distribution it generates when RRM converges to an objective value. Their key findings include the necessary and sufficient conditions for the retraining process to reach a performatively stable point with nearly minimal loss, one of which is that the classifier's loss function ℓ should be smooth and strongly convex. This framework is much more general, but could also be studied in different information scenarios. However, the focus of this thesis is on the specific case where users respond based on specific utility and cost functions.

2.2.5 Social perspectives and fairness

Hu et al. [2019] are the first to argue that in many social contexts where machine learning tools are used, agents face different costs when altering the attributes evaluated by the classifier. They formalize a setting where societal groups incur varying costs for modifying their features, with each group potentially having different distributions over unmanipulated features and distinct true labeling functions. This introduces higher uncertainty for the decision-maker due to heterogeneous groups, although the possible decision rule has to remain non-discriminatory.

They define the payoff for the decision-maker not based on accuracy but rather as penalties for false positives (FP) and false negatives (FN), weighted by specific costs. Their findings indicate that when the learner is aware of the costs different groups face, the equilibrium classifier tends to reinforce existing inequalities. Providing subsidies to alleviate manipulation burdens for disadvantaged groups can enhance the decision-maker's classification performance and reduce inequality-reinforcing errors. However, this intervention can also improve performance solely for the decision-maker, potentially lowering the welfare of both groups compared to a scenario without subsidies.

In contrast, this current thesis assumes that the population is homogeneous with respect to their cost functions.

Milli et al. [2019] investigate another social risk of strategic classification: users with a true positive label face a higher bar to be classified correctly. The concept of social burden is introduced to measure the expected cost that an individual with a truly +1 label needs to incur to be correctly classified. The study demonstrates that because of this cost there is the trade-off between the payoff of decision-makers and users in strategic classification.

Moreover, the study highlights the fairness impact of strategic classification. When decision-makers make their classification algorithms robust against gaming, the incurred costs can fall unevenly on different subpopulations, leading to disparities between these groups. They also conduct a case study on FICO scores, which are widely used in the United States to predict creditworthiness, focusing on two subpopulations determined by race ("black" and "white"). This case study validates their modeling assumptions and illustrates the trade-offs and fairness concerns associated with strategic classification in a real-world context.

This work differs from the work of Hu et al. [2019] by focusing more on the trade-off between decision-makers and users rather than differences between users. It also differs from my thesis, which uses non-linear models and assumes a homogeneous population, rather than focusing on linear models, subpopulations, and social burden.

2.3 Explainability of non-linear machine learning algorithms

In strategic classification, decision-makers use models in order to classify users. Hence, their model chocie is also a crucial part of the game. As stated in the introduction, neural networks and other non-linear machine learning methods are gaining popularity due to their often superior accuracy compared to linear classifiers. However, a major drawback of these methods is their lack of interpretability. This poses a problem because decision-makers are less likely to trust and use models they cannot interpret. Additionally, data regulations require algorithms used in automated decision-making to provide at least “meaningful information about the logic involved” [2016/679, 2016-05-04]. Therefore, the ability to explain both the model itself and individual predictions is crucial in practical applications of machine learning models.

Doshi-Velez & Kim [2017] summarize the importance and definitions of interpretability in machine learning. They state that interpretability is crucial if there are significant consequences for unacceptable results. They summarize the necessity of interpretability stemming from incompleteness and argue that interpretability is important for scientific understanding, safety, ethics, among other factors. Molnar [2020] also argues that interpretability is important for a model to gain social acceptance.

Interpretability is also a foundation for gaming in strategic classification. Users need to understand a

model to know, or at least have an estimate on how to manipulate their features to get better predictions. This is why many studies in strategic classification focus on linear models: they argue that linear classifiers are the most common case of classification where users game due to their easy interpretability. However, as machine learning models become more popular, there has been a growing focus on making them interpretable as well. Thus, it is plausible to assume that gaming also occurs when non-linear machine learning models with some explanations are deployed.

There are two main directions of model explanations: global and local interpreting mechanisms. Global methods describe the average behavior of a machine learning model and are often expressed as expected values based on the distribution of the data. For example, the partial dependence plot, a type of feature effect plot, shows the expected prediction when all other features are marginalized out [Friedman, 2001]. Since global interpretation methods describe average behavior, they are particularly useful when the modeler wants to understand the general mechanisms in the data or debug a model. Further examples for global explanations include Accumulated Local Effects (ALE) Plots [Apley & Zhu, 2020], feature interactions [Friedman & Popescu, 2008], functional ANOVA [Hooker, 2004], permutation feature importance [Breiman, 2001; Fisher et al., 2019] and global surrogates [Molnar, 2020].

However, for users to understand their predictions, local interpretations can be even more relevant. Local interpretations namely explain individual predictions.

An example for local interpretations are *Shapley values*. The Shapley values, introduced by Shapley [Shapley et al., 1953], are a method in game theory for distributing payouts to players based on their individual contributions to the total payout. In this context, players form a coalition and share the resulting profit from their collaboration. In the realm of machine learning, the terms "players," "game," and "payout" take on specific meanings related to prediction and interpretability. Here, the "game" refers to the task of making a prediction for a single instance within a dataset. The "gain" represents the actual prediction for this instance minus the average prediction for all instances. The "players" are the feature values of the instance that work together to achieve the gain (i.e., to predict a specific value). The objective is to explain the difference between the actual prediction and the average prediction.

Other local interpretations include local surrogate models. The concept of Local Interpretable Model-agnostic Explanations (LIME) proposed by Ribeiro et al. [2016] involves training this kind of local surrogate models to approximate the predictions of the underlying black box model for specific instances rather than globally.

LIME operates by generating variations of the input data and observing the resulting predictions from the

black box model. It then creates a new dataset of these perturbed samples and their corresponding predictions. An interpretable model, such as Lasso or a decision tree, is trained on this new dataset, with weights based on the proximity of samples to the instance of interest. This local surrogate model approximates the black box model's predictions locally. The resulting explanations exhibit several desirable qualities:

- they are **easy to interpret**, as they employ an easily interpretable model as a surrogate,
- they are **locally faithful**, meaning the explanations accurately reflect the model's behavior near the instance being predicted,
- they are **model-agnostic**, as perturbed samples and predictions can be generated for any model.

This thesis evaluates the effect of using model explanations for black-box models in strategic situations, focusing specifically on LIME. The motivation behing choosing LIME are its popularity, that it is model-agnostic methods and also relatively easy for laypeople to understand due to its visual outputs. However, it would be interesting for future research to investigate and compare how different explanation methods might influence gaming behavior and outcomes.

3 Background

In this chapter, I intorduce the used notation and definitions adopted from previous works.

3.1 Notation and definitions

I mostly base my notations on the frameworks of Hardt et al. [2016]; Ghalme et al. [2021]. The game in strategic classification consists of two players:

- *Jury* or decision-maker (e.g. University)
- *Contestant* or user (e.g. University applicants)

Briefly, the Jury has to classify the Contestants into two binary classes with the goal of getting high accuracy. The Contestants goal is to get a positive classification.

Formally, let $x \in \mathcal{X} \subseteq \mathbb{R}^d$ represent a d-dimensional feature vector describing one user. In the example of the introducton with the university admissions, features could consist of the grades of an applicant in different subjects. Let their population $\mathcal{X}$ have distribution D, with the assumption that the users form a homogenous population and therefore can be described with one distribution. The users are by assumption fully rational.

There is by assumption a ground truth mapper, i.e. a true underlying classifier $h : \mathcal{X} \to \mathcal{Y}$, that assings binary labels $y \in \mathcal{Y} = \{-1, 1\}$ to the users. The Jury has access to a sample set of n Contestants, that is $T = \{(x_i, y_i)\}_{i=1}^n$ with $x_i \overset{iid}{\sim} D$ and $y_i = h(x_i)$ and wants to find a classifier $f : \mathcal{X} \to \mathcal{Y}$ from class $\mathcal{H}$ that predicts well on D. After committing to the model f, they make some information I_f describing the model avaible to the users.

Users operate based on their payoff utility and cost functions. Their goal is to maximize their overall utility (for ease of reference simpley utility) $u : \mathcal{X} \to \mathbb{R} = u_{\text{payoff}} - c$. The payoff utility function $u_{\text{payoff}} : \mathcal{X} \to \mathbb{R}$, is assumed to be linear, i.e. $u_{\text{payoff}}(x) = t \cdot f(x)$. As the binary labels allow only for $f(x) \in \{-1, 1\}$, $u_{\text{payoff}}(x) \in \{-t, t\}$ and the maximal gain of utility is $2 \cdot t$. Hence, t encodes the intensity of gaming: the higher t, the more users are wanting to reach a label +1 and the more they are willing to manipulate their features. The cost function $c(x, x')$ is a publicly known, non-negative function quantifying the cost of feature manipulation. It depends on a cost vector α that represents the varying difficulties associated with changing different features. Given that $\|\alpha\|_2^2 = 1$, it reflects the relative cost differences among the features. By

assumption, a mix of linear and quadratic costs is used :

$$c(x, x') = (1 - \varepsilon)\langle \alpha, x' - x \rangle_+ + \varepsilon \|x - x'\|_2^2 \tag{1}$$

where ε defines the weight of the quadratic element and satisfies $\varepsilon \in [0, 1]$.

From the definition it also shows that there is no cost of not manipulating features: $c(x, x) = 0$.

Best response. In strategic classification, users modify their features to gain a positive prediction while incuring the least possible costs. How they do it, depends on the information I_f as well as their utility and cost functions. Let Δ_{I_f} be their utility maximizing feature vector:

$$\Delta_{I_f} = \arg\max_{x' \in \mathcal{X}} u(x') = \arg\max_{x' \in \mathcal{X}} u_{\text{payoff}}(x') - c(x, x') \tag{2}$$

$\Delta_{I_f} : \mathcal{X} \to \mathcal{X}$ is a response mapping based on I_f, hence representing the step where users change their features based on their knowledge. I_f is the information that users possess, comprehend, and utilize in their best response on the model f. In the university example, I_f could be the information on the webpage of the university on application requirements. It is worth to remark that in neither the utility and therefore neither the best response of the users take the true labeling function h in consideration. Users are not interested in true improvement: their goal is only a positive classification outcome. From the specific utility and cost functions defined above, the set of feasible and worthwhile new feature vectors can be derived:

$$C_f(x) = \{x' | c(x, x') < 2t, f(x') = 1\} \tag{3}$$

Hence, users only change their features if:

1. *feasibility*: the costs are less than the utility they gain. As the maximal gain in utility is $2t$, users only game if the costs are lower than that.
2. *worthiness*: they believe that the classification outcome of their new features is +1. This means that users only invest in changing if the believe that they will get a +1 prediction.

Therefore, they best response mapping of the users can be rewritten as:

$$\Delta_f(x) := \begin{cases} \arg\min_{x' \in C_f(x)} c(x, x') & f(x) = -1 \wedge C_f(x) \neq \emptyset; \\ x & \text{otherwise.} \end{cases} \tag{4}$$

In words, users change their features only if they are classified negatively and there are feasible and worthwhile changes possible. Among the possible changes in $C_f(x)$, they choose the lowest-cost point.

Payoffs of the players The objective in strategic classification for the Jury is to find a function f that minimizes the *strategic error*:

$$\text{err}(f) := \mathbb{P}_{x\sim D}\{h(x) \neq f(\Delta_f(x))\},$$

This represents the probability of classification error by f when each user x strategically modifies their features to $\Delta_f(x)$ in response to f.

The users optimize their best response in order to gain individual payoffs based on their utility function. Their final payoff is: $u(\Delta_{I_f}) = u_{\text{payoff}}(\Delta_{I_f}) - c(x, \Delta_{I_f})$.

3.2 Estimation errors and related definitions

When users lack complete information about the classifier f, the label they get may differ from their estimates. Hence it is important to identify and define the discrepancies that may arise as a result. In this section I introduce the relevant definitions also used by Ghalme et al. [2021].

Definition 3.1 (Strategic error)**.** *The strategic error when the Jury plays f and Contestant responds to I_f is given by:*

$$err(f, I_f) = \mathbb{P}_{x\sim D}\{h(x) \neq f(\Delta_{I_f}(x))\}. \tag{5}$$

Note that $\text{err}(f, f) = \text{err}(f)$.

Definition 3.2 (Price of Opacity [Ghalme et al., 2021])**.** *When Jury plays f, however Contestants do not best respond based of f but other information, the Price of Opacity equals:*

$$err(f, I_f) - err(f, f) \tag{6}$$

Thus, intuitively, the Price of Opacity or *POP* the difference in the error of a model in a fully transparent and another information setting. POP depends on the model f, the information provided by the decision-maker I_f and implicitly also on the ground-truth labeling function h.

The difference in errors is caused by the different best responses of the users. These discrepancies can be labeled as follows [Ghalme et al., 2021].

Definition 3.3 (Disagreement set)**.** *The disagreement set S is the set of the users who are classified differently than they expected:*

$$S := \{x | f(x) \neq I_f(x)\}$$

This means, users in the disagreement set have a different estimation on their label based on the information on the model f to what they get truly classified with. For example, if a user only sees a small sample of labels from the model, they might have a different estimation $\hat{f}$ on the model f resulting in $\hat{f}(x) \neq f(x)$.

Definition 3.4 (Enlargement set)**.** *The enlargement set includes all users that the decision-maker classifies differently in different information scenarios:*

$$S := \{x | f(\Delta_f(x)) \neq f(\Delta_{I_f}(x)))\}$$

Partition of E. The points in E can be partitioned further to:

$$E^+ = \{x \mid h(x) = f(\Delta_f(x)) \wedge h(x) \neq f(\Delta_{I_f}(x))\};$$
$$E^- = \{x \mid h(x) \neq f(\Delta_f(x)) \wedge h(x) = f(\Delta_{I_f}(x))\}.$$

In words, E^+ are the users who would have been classified correctly in transparent scenario, i.e. where all possible information I_f is known to the users, but cause an erroneous prediction in opaque scenarios. On the other hand, E^- are the users who would have been classified correctly in an opaque scenario, but cause an erroneous prediction in transparent scenario.

Ghalme et al. [2021] relate this POP, E and h directly by defining:

$$\text{POP}^+ := \mathbb{P}_{x\sim D}\{x \in E^+\}, \quad \text{POP}^- := \mathbb{P}_{x\sim D}\{x \in E^-\}.$$

From this, they deduct that:

$$\text{POP} = \text{POP}^+ - \text{POP}-$$

In words, the difference in errors is equal to the increase in errors due to transparency minus the decrease in errors for the same reason.

Ghalme et al. [2021] suggest sufficient conditions on when the price of opacity is positive. They argue that if these conditions are fulfilled, transparency from the side of a Jury has better payoff to them as opaqueness.

Theorem 3.1 (Sufficient conditions on strictly positive POP [Ghalme et al., 2021])**.** *The price of opacity is*

strictly positive, i.e., POP > 0*, if the following condition holds:*

$$P_{x\sim D}\{x \in E\} > 2\,\mathrm{err}(f^\star, f^\star) + 2\epsilon_1$$

Where $\mathrm{err}(f^\star, f^\star) = \mathrm{err}(f, f) - \epsilon_1$ *and* $f^\star = \mathrm{argmin}_{f\in\mathcal{H}}\,\mathrm{err}(f, f)$*.*

In words, $f^\star$ *is the optimal strategy-aware classifier in the transparent case and* ϵ_1 *is the chosen to be the difference between the errors of the optimal and actual models.*

Remark 3.2. *From Theorem 3.1 and the definition of* $\mathrm{err}(f^\star, f^\star)$ *we can restate the condition as follows:*

$$P_{x\sim D}\{x \in E\} > 2\,\mathrm{err}(f, f)$$

3.3 Linearity vs non-linearity

In my thesis, I aim to compare a strategic setup using both linear and non-linear model choices made by the Jury. Therefore I want to give a small overview about the models that I am using. To achieve this, I have selected the following model types:

Linear Model **Linear Support Vector Classifier (Linear SVC)** – A Support Vector Machine (SVM) [Cortes & Vapnik, 1995] is a widely used classification model that finds the optimal hyperplane to separate data points. In its linear form, it is particularly effective for problems where the data is linearly separable. I chose the Linear SVC for its strong theoretical foundation and efficient performance on high-dimensional datasets as well as its application in previous studies in the field of Strategic classification (for example [Sundaram et al., 2023]).

Non-Linear Models **Neural Networks (NN)** – Neural networks form the foundation of many AI models and have demonstrated high performance with strong generalization capabilities, making them widely used. Unlike linear models, they can learn complex patterns and decision boundaries. For this study, I chose a simple fully connected neural network with ReLU activations [Agarap, 2019]. The network is optimized using the ADAM optimizer [Kingma & Ba, 2017]. I decided to include neural networks due to their popularity.

K-Nearest Neighbors (KNN) – KNN [Cover & Hart, 1967] is a simple yet effective non-linear model that classifies data points based on the majority class of their nearest neighbors. It is particularly useful for non-parametric learning, as it does not make strong assumptions about the data distribution. I included this

model due to its simplicity and the distinct optimization approach it introduces to the users compared to a linear model.

Random Forest Classifier (RNF) – Random Forest is an ensemble learning method that builds on multiple decision trees and combines their predictions to improve accuracy and reduce overfitting [Ho, 1995]. I chose random forests because they represent a fundamentally different non-linear optimization problem for the user's best response.

For ease of reference, I refer to these three non-linear models using the short terms provided in brackets.

4 Methodology

Next, I introduce my methodology for addressing the research questions of this thesis. First, I define the different information scenarios that I am comparing, each characterized by the algorithms describing the game in corresponding information scenario. Following that, I mention the linear and non-linear models that are used. Finally, I give a general outline of the overall process of my experiments for answering the research questions.

4.1 Information scenarios

In this thesis, three different information scenarios are compared. These scenarios vary in the information I that users possess, comprehend and utilize for determining their best responses.

[FULL_INFO] I. Full information. In this case, the decision-maker makes the classifier f public. This means, users can access and evaluate $f(x')$ for any $x' \in \mathcal{X}$ before being classified themselves, without any additional costs.

[PART_INFO] II. Partial information The case of partial information is also evaluated. In this scenario, users have access to LIME explanations $L(x)$ [Ribeiro et al., 2016] of the model for their x feature vector before being classified themselves, without any additional costs. LIME values consist of "feature conditions" Z and a "feature importance score" w for each feature $1, ..., d$ as well as a prediction of the LIME on the instance $l(x)$. In binary classification, this prediction reflects the probability of the instance belonging to the classes.

An example:

Feature conditions Z	**Feature importance score** w
maths $\leq$ 90	-0.3
82 < literature $\leq$ 11.86	0.12
overall gpa > 3.5	0.05
Prediction on the instance	0: 0.22, 1: 0.78

Table 1: An example of $L(x)$ within the recurring university example. Each row shows a feature condition Z and its associated importance score w, reflecting how much that condition contributes to the model's prediction. Positive scores push the prediction toward +1, while negative scores push it toward 0. For instance, this particular student's math score was below 90, which contributed most strongly to a negative prediction.

The best response of users who have access to the LIME values on their features is designed as in Al-

gorithm 1. The motivation behind the greedy approach lies in the logic of making the most effective improvement with the fewest possible changes. Users aim to modify their features so that negative conditions—those that decrease the probability of a positive classification—no longer apply to them. To achieve this, they adjust their features to the closest possible value that causes the condition Z to no longer hold, specifically moving just past the threshold Z_i where the condition becomes inactive.

Algorithm 1 Feature Adjustment Algorithm Based on LIME Output

Inputs:
feature vector x of user
LIME output $L = [Z, w, l(x)]$, where list of Z feature conditions and w weights, both sorted descending by the absolute values of w. $l(x) = \{l_0(x), l_{+1}(x)\}$ contains the predicted probabilities of the instance belonging to the classes 0 and +1.
Cost function in the form: $c_{true}(x, x') = (1 - \varepsilon)\langle \alpha, x' - x \rangle_+ + \varepsilon \|x - x'\|_2^2$
Initialize $\Delta_L \leftarrow x$, $c_{cumulative} = 0$
for each i in changeable features in descending order by the absolute values of w **do**
  Evaluate the cost of changing the feature: $c_i = (1 - \varepsilon)\alpha_i(x_i - Z_i) + \varepsilon(x_i - Z_i)^2$
  $c_{\text{cumulative}} \mathrel{+}= c_i$
  if $c_{\text{cumulative}} \geq u(f([Z_i, \Delta_{L_{\bar{i}}}]))$ **where** $\Delta_{L_{\bar{i}}}$ contains the other feature values not i **then**
    $\Delta_{L_i} \leftarrow x_i$ ▷ The cost of change is too high, keep original feature value
  else
    $\Delta_{L_i} \leftarrow Z_i$ ▷ Change the feature to the condition specified by F
  end if
  if $f(\Delta_L) = +1$ **then**
    break ▷ Stop if the desired prediction is achieved
  end if
end for
Output: Δ_L

III. No information In this case, decision-makers do not share any information with the users. Here, the users have to gather information and estimate the model themselves. For that, this thesis uses a similar approach to Ghalme et al. [2021]. Users observe the features and predictions of other users, drawing a sample T_C of size m, $T_C = \{(x_i, f(x_i)\}_{i=1}^m$. If there are relations in the data, this sample is drawn as all the neighbors of a users that are at most two edges away. If the dataset does not contain social relations, a distance based approach is used. As in many cases, users with similar features are more connected in life. In this case, the following weighted sampling method is used: $\text{WS}(X, x, \sigma)$ denotes a weighted sampling from the population $\mathcal{X}$ with a Gaussian kernel centered at x and a bandwidth parameter σ. The weights are computed as: $w_i = \exp\left(-\frac{d(x,x_i)^2}{2\sigma^2}\right)$, where $d(x, x_i)$ is the Euclidian distance between x and x_i, and the sampling probabilities are normalized such that $p_i = \frac{w_i}{\sum_{j=1}^N w_j}$. The bandwidth parameter σ controls the importance of distance. However, for being able to use ERM, the users need to have access to a sample set that has both labels. Therefore, both sampling methods are corrected to have both labels in the sample. In the

first case, the closest +1 labeled neighbor is sampled additionally. In the second case, rejection sampling is used.

On this sample set T_C users best response in two different ways established below: either (1) they perform utility maximization as described in Section 3.1 or they (2) directly imitate the behaviour of their positively classified peers.

For (1) **[NO_INFO_EST]** they use ERM to estimate a linear SVM model on the sample set T_C, denoted as $\hat{f}$. Then, they proceed to best respond to the model the exact same way, however, as they only have an estimate, the final label might be different to their expectations.

For (2) **[NO_INFO_IMIT]** they take the average behavior of positively classified individuals in T_C. They take the subset of users in their sample that have been classified in the desired way $T_C^+ = \{(x_i, f(x_i) | f(x_i) = +1\}_{i=1}^p$, consisting of p users, and create the average point of their features $\overline{x}^P = \frac{1}{p}\sum_{(x,f(x))\in T_C^+} x$. The best response of the users in this case will be:

$$\Delta_{\text{imitation}} = \beta \cdot \overline{x}^P + (1-\beta)\left(\frac{1}{p}\sum_{(x,f(x))\in T_C^+} x\right) \text{ such that } \beta \in [0,1] \text{ and } c(x, \Delta_{\text{imitation}}) = 2t \qquad (7)$$

Hence, the users try to change their features to be as closed to the positive average, as possible, meaning that they spend all their budget on it. Since they do not estimate their own label or prediction score, they have no clear stopping point where they can be sure that they prediction is positive—so their goal becomes getting as close as possible to $\overline{x}^P$. The exact value of β to find the highest feasible changes has a closed form solution, which is summarized in Lemma 4.3.

Remark 4.1. *In the no information - imitation scenario, the change vector of the best response $x' - x$ can be rewritten as follows:*

$$x' - x = \Delta_{imitation} - x = (\beta - 1)(x - \overline{x}^P)$$

where $\overline{x}^P = \frac{1}{p}\sum_{(x,f(x))\in T_C^+} x$ and $\Delta_{I_f} = \beta\underline{x} + (1-\beta)\left(\frac{1}{p}\sum_{(x,f(x))\in T_C^+} x\right)$

Lemma 4.2. *Contestants with given mixed cost function (see Equation 1) with $\varepsilon \neq 0$ and a point they want to imitate $\overline{x}^P = \frac{1}{p}\sum_{(x,f(x))\in T_C^+} x$, choose β in their best response algorithm (see Algorithm 1) in the following way:*

$$\beta^* = \max(0, \frac{-(1-\varepsilon)\langle\alpha, x-\overline{x}^P\rangle_+ - \sqrt{(1-\varepsilon)^2\langle\alpha, x-\overline{x}^P\rangle_+^2 + 4\cdot\varepsilon\|x-\overline{x}^P\|_2^2\cdot 2t}}{2\varepsilon\|x-\overline{x}^P\|_2^2})$$

Lemma 4.3. *Contestants with given mixed cost function (see Equation 1) with $\varepsilon \neq 0$ and a point they want to*

imitate $\overline{x}^P = \frac{1}{p}\sum_{(x,f(x))\in T_C^+} x$*, choose* β *in their best response algorithm (see Algorithm 1) in the following way:*

$$\beta^* = \max(0, \frac{-(1-\varepsilon)\langle\alpha, x-\overline{x}^P\rangle_+ - \sqrt{(1-\varepsilon)^2\langle\alpha, x-\overline{x}^P\rangle_+^2 + 4\cdot\varepsilon\|x-\overline{x}^P\|_2^2\cdot 2t}}{2\varepsilon\|x-\overline{x}^P\|_2^2})$$

Proof. By definition, Contestants choose the smallest feasible β indicating the largest change. The feasibility constraints on beta are: $\beta \in [0,1]$ and $c(x, \Delta_{\text{imitation}}) \leq 2t$. We use the equation on the mixed cost function from Equation 1:

$$c(x, \Delta_{\text{imitation}}) = (1-\varepsilon)\langle\alpha, \Delta_{\text{imitation}} - x\rangle_+ + \varepsilon\|\Delta_{\text{imitation}} - x\|_2^2 \leq 2t$$

Now we use Remark 4.1

$$(1-\varepsilon)\langle\alpha, (\beta-1)(x-\overline{x}^P)\rangle_+ + \varepsilon\|(\beta-1)(x-\overline{x}^P)\|_2^2 \leq 2t$$

This is a quadratic equiation for β:

$$(\beta-1)^2\varepsilon\|(x-\overline{x}^P)\|_2^2 + (\beta-1)(1-\varepsilon)\langle\alpha, (x-\overline{x}^P)\rangle_+ - 2t \leq 0$$

As $\varepsilon \in [0,1]$ and $t \geq 0$ the discriminant $D = (1-\varepsilon)^2\langle\alpha, x-\overline{x}^P\rangle_+^2 + 4\cdot\varepsilon\|x-\overline{x}^P\|_2^2\cdot 2t$ is always non-negative hence there exists at least one exact solution $\beta^* \in \mathbb{R}$. If we define $\beta_1 \leq \beta_2$, the solutions for equality are:

$$\beta_1 = \frac{-(1-\varepsilon)\langle\alpha, x-\overline{x}^P\rangle_+ - \sqrt{D}}{2\varepsilon\|x-\overline{x}^P\|_2^2} + 1 \text{ and } \beta_2 = \frac{-(1-\varepsilon)\langle\alpha, x-\overline{x}^P\rangle_+ + \sqrt{D}}{2\varepsilon\|x-\overline{x}^P\|_2^2} + 1$$

For all other $\beta \in [\beta_1, \beta_2]$ the costs are strictly less than $2t$. However, Contestants want the largest change in their budget therefore they prefer $c(x, \Delta_{\text{imitation1}}) = 2t$ over $c(x, \Delta_{\text{imitation2}}) < 2t$. From the property of the mixed cost function $c(x,x) = 0$ which happens if $\beta = 1$ we know that $1 \in [\beta_1, \beta_2]$. Hence, their choice depends on the sign of β_1. If $\beta_1 \leq 0$ they will choose $\beta^* = 0$ and move completely to the average point $\Delta_{I_f} = \frac{1}{p}\sum_{(x,f(x))\in T_C^+} x$. Else $\beta_1 > 0$ they choose $\beta^* = \beta_1$. These two cases can be summarized as following:

$$\beta^* = \max(0, \beta_1) = \max(0, \frac{-(1-\varepsilon)\langle\alpha, x-\overline{x}^P\rangle_+ - \sqrt{(1-\varepsilon)^2\langle\alpha, x-\overline{x}^P\rangle_+^2 + 4\cdot\varepsilon\|x-\overline{x}^P\|_2^2\cdot 2t}}{2\varepsilon\|x-\overline{x}^P\|_2^2} + 1)$$

□

For the ease of reference, I used a slightly modified wording from the literature to define the FULL_INFO case as a ***transparent*** scenario, and the other cases (PART_INFO, NO_INFO_EST and NO_INFO_IMIT) as ***opaque*** scenarios.

4.2 The algorithms

The game setup of the different information scenarios can be summarized as:

- Full information: Algorithm 2
- Partial information: Algorithm 3
- No information: Algorithm 4

Algorithm 2 The interaction protocol between Jury and Contestants with *full information* [FULL_INFO]

1: Nature selects ground truth scoring function h
2: Jury commits to classifier f
3: Jury publishes the decision rule f
4: Contestants observe f and predicted labels for their features $(x_i, f(x_i))$
5: Based on classifier f, their utility function u and cost function c Contestants best respond:
6: $\Delta_{I_f} = \Delta_f = \arg\max_{x' \in X} u(f(x')) - c(x, x')$
7: Contestants gain $u(f(\Delta_f)) - c(x, \Delta_f)$
8: Jury gains $\mathrm{acc}(f) = \mathbb{P}_{x \sim D}\{h(x) = f(\Delta_f(x))\}$

Algorithm 3 The interaction protocol between Jury and Contestants with *partial information* [PART_INFO]

1: Nature selects ground truth scoring function h
2: Jury commits to classifier f
3: Contestants observe predicted labels for their features $(x_i, f(x_i))$ and LIME values L
4: L is a list of feature conditions F and respective weigths w
5: Based on L, their utility function u and cost function c, Contestants best respond:
6: $\Delta_{I_f} = \Delta_L$ as described in Algorithm 1
7: Contestants gain $u(\Delta_L) - c(x, \Delta_L)$
8: Jury gains $\mathrm{acc}(f) = \mathbb{P}_{x \sim D}\{h(x) = f(\Delta_L(x))\}$

It is important to note that in the NO_INFO_IMIT case, users do not receive information about their own estimated label—only the labels within their sample. As a result, even users with initially positive predictions may choose to alter their features, without being able to verify whether their decision is worthwhile. Consequently, a Contestant's payoff can even be negative.

By assmuption, the cases PART_INFO, NO_INFO_EST and NO_INFO_IMIT are defined as opaque scenarios. It is assumed that the NO_INFO cases are more opaque than the PART_INFO case, as less information is published by the decision-maker.

Algorithm 4 The interaction protocol between Jury and Contestants with *no information*

1: Nature selects ground truth scoring function h
2: Jury commits to classifier f
3: Contestants sample a set of instances and their labels T_C of size m, $T = \{(x_i, f(x_i))\}_{i=1}^{m}$ where $x \sim \text{WS}(X, x, \sigma)$ (see definition of weighted sampler in Section 4.1)
4: **if** Contestants perform utility maximization [NO_INFO_EST] **then**
5: Contestants estimate the classifier based on T_C resulting in the estimated model $\hat{f}$
6: Based on the estimated classifier $\hat{f}$, their utility function u, and cost function c, Contestants best respond:

$$\Delta_{I_f} = \Delta_{\hat{f}} = \arg\max_{\Delta_{\hat{f}} \in X} u(\Delta_{\hat{f}}) - c(x, \Delta_{\hat{f}})$$

7: Contestants gain $u(\Delta_{\hat{f}}) - c(x, \Delta_{\hat{f}})$
8: **else if** Contestants imitate behavior [NO_INFO_IMIT] **then**
9: Contestants calculate $\overline{x}^P = \frac{1}{p} \sum_{(x, f(x)) \in T_C^+} x$
10: Contestants best respond as defined in Lemma 4.3:

$$\Delta_{I_f} = \Delta_{\text{imitation}} = \beta \underline{x} + (1 - \beta) \left(\frac{1}{p} \sum_{(x, f(x)) \in T_C^+} x \right)$$

such that $\beta \in [0, 1]$ and $c(x, \Delta_{\text{imitation}}) = 2t$
11: Contestants gain $u(\Delta_{\text{imitation}}) - c(x, \Delta_{\text{imitation}})$
12: **end if**
13: Jury gains $\text{acc}(f) = \mathbb{P}_{x \sim D}\{h(x) = f(\Delta_{\text{imitation}}(x))\}$

4.3 Overall process

In this section, I provide a summary of the methodology employed to address the research questions, specifically detailing how the previously introduced algorithms are applied to answer these questions.

As my goal is to compare the payoffs of different players using linear and non-linear models across various information scenarios, I apply the same steps in multiple settings and two datasets. These steps are:

1. **Model training**: The original model is trained using the training dataset $X_{train} = \{x_1, ...x_k\}$ resulting in a trained classifier f.

2. **Simulating best responses**: The best responses of the users in the test set $X_{test} = \{x_1, ...x_n\}$ with the information from the trained model I_f is simulated using the different algortihms defined in the section above. The result is an updated test dataset with the new features of the users $X'_{test} = \{\Delta_{I_f\,1}, ...\Delta_{I_f\,n}\}$

3. **Evaluation**: The updated model and the differences between the original and updated test datasets with the metrics defined below.

To compare non-linearity versus linearity, I modify the model choice f during the model training step,

keeping all other factors constant. To compare different information scenarios, I adjust the information I_f during the step of simulating best responses ceteris paribus.

The experiments are conducted on three datasets: a synthetic dataset with two Gaussian clusters, a synthetic dataset with two moon-shaped clusters and the dataset from Ghalme et al. [2021]. The synthetic datasets offer flexibility in exploring various dimensions and edge cases, as well as promising easy interpretability in low-dimensional cases. On the other hand, the latter dataset serves as a benchmark, allowing for a comparison of the results from the new information scenarios with those from previous studies.

As the goal of this thesis is to compare the players' payoffs in different settings, the following aspects are evaluated both before and after the best responses of the Contestants:

- **Jury's payoff**: Assessed by model accuracy. Formally: $\mathrm{acc}(f) = \mathbb{P}_{x\sim D}\{h(x) = f(\Delta_{I_f}(x))\}$. Accuracy is used because it is the primary metric employed in other papers on strategic classification and because it is the most commonly used model selection criterion, which decision-makers aim to make robust. In praxis, it is evaluated as $\frac{\sum_{i=1}^{n} \mathbb{1}[f(\Delta_{I_f\, i})\neq h(x_i)]}{n}$.

- **Contestants' payoff**: Evaluated as the gained payoff utility minus the cost of changing their features. Formally: $u(\Delta_{I_f}) = u_{\text{payoff}}(\Delta_{I_f}) - c(x, \Delta_{I_f})$. This metric is used because it mirrors the players' motivations and expectations, which are the most important factors driving their manipulations. Average Contestant payoff is calculated as $\frac{\sum_{i=1}^{n}(t\cdot y_i - c(x_i, \Delta_{I_f})}{n}$, where n is the number of users.

- **Change indicators**: A user is changing their features if $x \neq \Delta_{I_f}$. Let the number of changes be l. The percentage of user changes is evaluated as l/n. The average cost of change per change is calculated as $\frac{\sum_{i=1}^{l} c(x_l, \Delta_{I_f})}{l}$.

- Disagreement set: the number of users who get classified differently by f than they estimated based on $\hat{f}$. As they only have estimated labels in the PART_INFO and the NO_INFO_EST case, this measure is only evaluated on those two. By definition, it would be always 0 for the FULL_INFO case. The proportion of the disagreement set is evaluated as: $\frac{\sum_{i=1}^{n} \mathbb{1}[f(x_i)\neq I_f(x_i)]}{n}$

- **User welfare**: Measured as the number of users whose classified label is $+1$. Formally: $\frac{\sum_{i=1}^{n} \mathbb{1}[f(\Delta_{I_f})=1]}{n}$.

The results are visualized for varying levels of opaqueness and user utilities. Additionally, a small subset of users from the disagreement set is sampled and analyzed to provide deeper insights into the observed changes.

5 Experimental setup

In this section technical details of the experiments are described. The code for the thesis can be found at https://github.com/blank-v-248/BV-master-thesis-project.

5.1 Dataset description

Synthetic dataset There are two synthetic datasets used, as motivated above. One is a dataset with Gaussian clusters, while contained two moon-shaped clusters. For the Gaussion clusters, one cluster was centered a (12,180) as the other was centered around (55,30) both with standard deviation of (20,90). The moons dataset was created with the `make_moons` package from `sklearn.datasets` and a noise of 0.3 was used Pedregosa et al. [2018].

Loan dataset Additionally, experiments are conducted using the dataset provided by Ghalme et al. [2021]. This loan dataset is created by cross-referencing two distinct sources related to the Prosper.com peer-to-peer lending platform. The authors detail their dataset creation process: they identified the intersection of publicly available loan data[2] and the network data from the work of Kumar & Zinger [2014].

The resulting dataset contains $20,222$ loan requests, each characterized by a set of informative features. Additionally, the corresponding users are linked to others within the associated social network. This dataset is publicly available and can be accessed through the authors' repository [3].

5.2 General settings

Across experiments, as many as possible parameters were kept the same in order to facilitate comparison. The weight of the quadratic element in the cost function was $\varepsilon = 0.2$. The budget for changes in the users best response was $t = 0.2$ as it was also in the work of Ghalme et al. [2021]. In the synthetic cases, the number of datapoints n was choosen to be $n = 1138$ with $x \in \mathbb{R}^2$ and $y \in \{0; 1\}$, with a 10% test and 90% training split. For the loan dataset, I used the dataset described above with the same settings and the same subset of features as in the work of Ghalme et al. [2021], and also applied their train-test split. However, for computational reasons, I ended up using a subset of size 500 of the training set for evaluation. This is because the in order of the optimizer to work, many rounds of iterations where needed and sometimes even multiple initializations - this made computational costs very high. Even though some of the comparing power with

[2] `https://www.kaggle.com/datasets/yousuf28/prosper-loan`
[3] `https://github.com/staretgicclfdark/strategic_rep/tree/main/data`. Downloaded on: 03.01.2025

the baseline gets lost with using the subset, comparisons within information scenarios and model types are not affected.

For α, the cost vector representing the varying difficulties associated with changing different features, I choose $\alpha_{\text{synthetic}} = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}$ and $\alpha_{\text{loan}} = \begin{pmatrix} 0.5 & 0.5 & 1.5 & -2.5 & -0.5 & 0.5 \end{pmatrix}$. For the classifiers, the following hyperparameters were selected. Since this study does not focus on optimizing the models f, I did not conduct an extensive hyperparameter search. Instead, I used either standard parameters or values comparable to those in Ghalme et al. [2021]. The Linear SVC was configured with a regularization parameter of $C = 0.01$ and an L2 penalty. The neural network consisted of four layers, each with 10 units, using ReLU activations and the Adam solver, with a maximum of 1000 iterations. In the KNN model, the number of neighbors was set to n_neighbors $= 11$. For the RNF model, I chose 10 estimators while keeping all other parameters at their default values.

To ensure reproducibility, all experiments were conducted with a constant random seed.

Optimizer For the optimizer used in computing the Contestants' best response, I employed the `"trust-constr"` method from `scipy.optimize`, as it is a general-purpose optimizer suitable for non-convex constrained optimization problems. It was essential to choose an optimizer that could be applied consistently across all models and that performed reasonably well -meaning it successfully found solutions in the majority of cases.

The `cvxpy` package was not suitable, as it assumes convex-constrained optimization, which does not necessarily hold for models such as NNs, RNFs and KNNs. I also experimented with other methods from `scipy.optimize`, such as `"COBYLA"` and `"SLSQP"`. While these often yielded better objective values (i.e., lower minimization results compared to `"trust-constr"`), they frequently failed to converge on solutions for many instances, particularly with KNN and RNF models. This is likely due to the non-smooth nature of their decision surfaces. Since the inability to find a solution in numerous cases introduces more distortion than slightly suboptimal solutions, I opted to use `"trust-constr"` consistently in my experiments.

For each instance, I defined an initial feasible starting point as a positively predicted sample, since the optimization constraint required a positive prediction. In the FULL_INFO setting, I selected positively predicted samples from the training set, computed their Euclidean distances to the target instance, and chose the closest feasible point as the initial starting point. The intuition behind this was similar to the NO_INFO_IMIT case, meaning that users in real world scenarios would try to learn from positive examples amongst their peers. If the initial attempt failed to yield a solution, I proceeded iteratively by selecting increasingly distant points—

specifically, every 10^{th} closest point—up to a maximum of 11 attempts. If all attempts failed, I assumed that the instance would retain its original features.

In the NO_INFO_EST scenario, where users only had access to a subset of the training data, it sometimes happened that no positively predicted samples by $\hat{f}$ were present in their subset—even when the model's true prediction f was positive. In such cases, I generated synthetic instances by sampling from a multivariate normal distribution over $\mathcal{X}$, then applied the same distance-based selection strategy used in the FULL_INFO case.

This approach had some limitations. First, the resulting solutions were often suboptimal, with relatively high objective values compared to closed-form solutions (where available, such as with linear models) or those obtained via alternative optimizers. Second, the results were quite sensitive to the choice of the initial starting point. Lastly, in some cases, the method still failed to find a solution. Nevertheless, within the scope of this thesis, this method was the most effective overall, given its relatively high success rate in identifying feasible solutions across different model types.

5.3 Table of experiments

In total, I tested 4 model choices on three datasets and all 4 information scenarios in the following ways:

Model and experiments	**Linear**		**NN**		**KNN**		**RNF**	
Dataset	**Decision boundary**	**POP**	**Decision boundary**	**POP**	**Decision boundary**	**POP**	**Decision boundary**	**POP**
Synthetic: Gaussian clusters	✓	✓						
Synthetic: moons	✓	✓	✓	✓	✓	✓	✓	✓
Loans dataset	✓		✓		✓		✓	

Table 2: Conducted experiments

For all scenarios, the metrics defined in Section 4.3 are evaluated. To address the Jury's payoff in linear vs. non-linear and transparent vs. non-transparent cases, accuracy is used. To evaluate the Contestants' payoff, the previously defined overall utility, change indicators, and user welfare are assessed. These calculations allow me to quantitatively answer my research questions.

Additionally, in the **decision boundary** experiments, the decision boundary, dataset, and users' best responses are visualized to better understand the changes and their effects. These experiments also address my second research question, focusing on how Contestants' features and predicted labels change across different information scenarios. Since these plots are in feature space and labels also are represented with the decision boundary, they allow for a clearer interpretation of user behavior patterns. In higher-dimensional cases, however, I rely solely on metric-based evaluation rather than visualizations, because using dimensionality reduction techniques would potentially distort the patterns.

In the **POP** experiments present plots showing accuracy across various information scenarios. These visualizations provide insight into the Jury's payoff, offering further support in answering my second research question.

In the decision boundary experiments, the sample size m was chosen depending on the total number of users in the training and test set $k+n$ respectively, the number of total users in the test set: $m = (k+n)/75$.

6 Results

In this chapter, I present the results of my experiments. I begin with analyses on synthetic data in Section 6.1, where I first illustrate various scenarios using visualizations of feature changes in a linear model in Section 6.1.1. I then compare different model choices in Section 6.1.2, followed by an evaluation of various information scenarios through the POP metric and corresponding plots in Section 6.1.3. This section concludes with a discussion of selected data stories in Section 6.1.4.

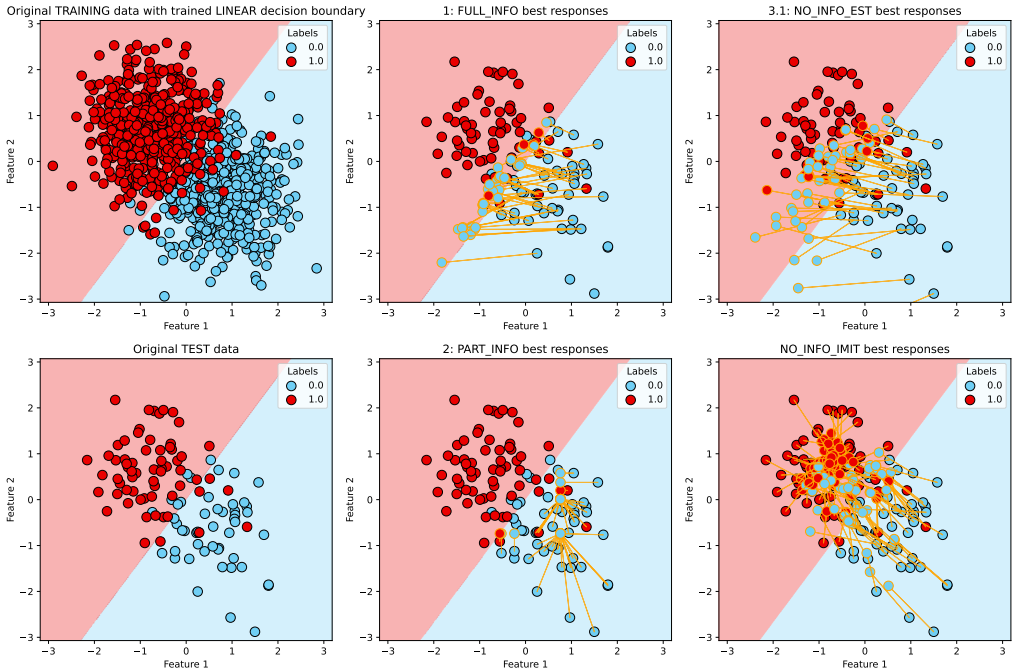
The second half of the chapter in Section Section 6.2 focuses on loan data. Here, I once again compare linear and non-linear model choices in 6.2.1, followed by an assessment of different information scenarios using the POP metric in 6.2.2. Finally, in Section 6.3, I examine the impact of additional parameters, such as the weight of the quadratic element in the cost function.

6.1 Synthetic data

6.1.1 Understanding the feature changes

First, I simulated users' best responses for the information scenarios defined in the previous chapter using a linear classifier on synthetic data. It is assumed that individuals' ground truth labels $h(x)$ remain unchanged, hence changing the features of the users can only have impact on their predicted labels $f(x)$ but not on the ground truth labels $h(x)$. This approach enables to visualize the different algorithms of information scenarios described in Chapter 4 and allows for their interpretation in a basic scenario. The results can be see in Figure 2. The quantitative results of the metrics defined in the section above can be found in Appendix A.1.

We can see two clusters and that on the original data the linear model is classifying quite well with 88.6% accuracy on the test set. In the FULL_INFO scenario, most users initially classified as 0 change their labels, and they incur the least possible cost to do so as the best responses are along the decision boundary. A key limitation of the thesis is evident in the plot: the optimization process does not consistently find the exact optimal solutions. Even when provided with full information, not all shifted features align with the decision boundary—despite this being a more cost-effective outcome. However, since a central objective of this thesis is to compare linear and non-linear approaches, it is crucial to apply the same optimization method for determining the user's best response across all models. As this particular optimizer performed best for non-linear models, it was used uniformly throughout the thesis. Even with suboptimal solutions, the cost of change per change is very low with a value of 0.01.



Source: own edit.

Figure 2: Strategic classification on synthetic data. Left: Synthetic dataset with the fitted LinearSCV. Middle: Best responses from users with full and partial information. Right: Best responses from users with no information and sampling. The labels show the true labels of the users. A yellow outline indicates that a users best response, i.e. their new, changed feature values, the yellow arrows connect their original and new feature values.

The PART_INFO scenario produces significantly different results: users do not move along the shortest path to the decision boundary (i.e., adjusting both features). Instead, they typically alter only one feature within their budget, which is in almost all cases insufficient to change their classification as user welfare only changes by 0.8%point. We can also see on the plots that they tend to move to the same points: this indicates that LIME often gives similar explanations to multiple users.

In the NO_INFO_EST scenario we see a similar shift to the FULL_INFO case from negative to positive classifications, though often at a higher cost with an average cost of change per change of 0.04. Also, there are multiple Contestants, who change their features but their classification by f does not change - this is mirrored in the moderate average Contestant payoff of 0.16.

Finally, in the NO_INFO_IMIT scenario all users adjust their features.

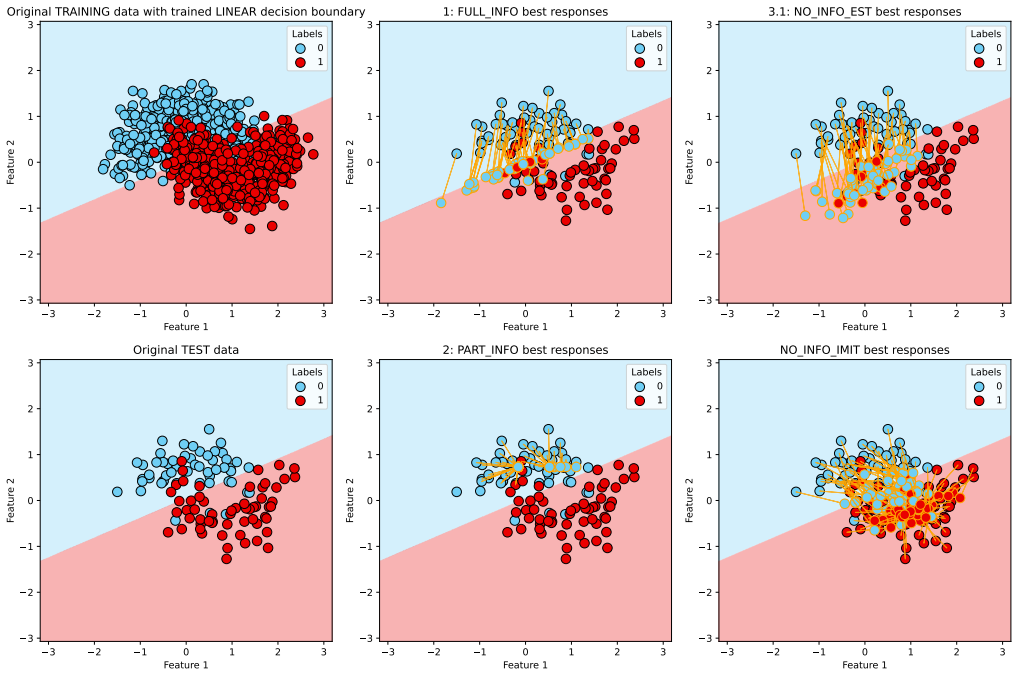
6.1.2 Linearity VS non-linearity

Subsequently, I repeated the experiments altering f, the model selection employed by the Jury. For this analysis, I opted for a synthetic scenario that necessitates non-linearity to evaluate the performance of lin-

ear versus non-linear models effectively - as with the gaussian clusters even a linear model had almost perfect performance making differences in accuracies harder to compare. To this end, I transitioned from two circular Gaussian clusters to the widely-used "moons" dataset.

The results of the non-linear models introduced in Section 3.3 can be found on following figures:

- Linear model: Figure 3
- NN: Figure 4
- KNN: Figure 5
- RNF: Figure 6

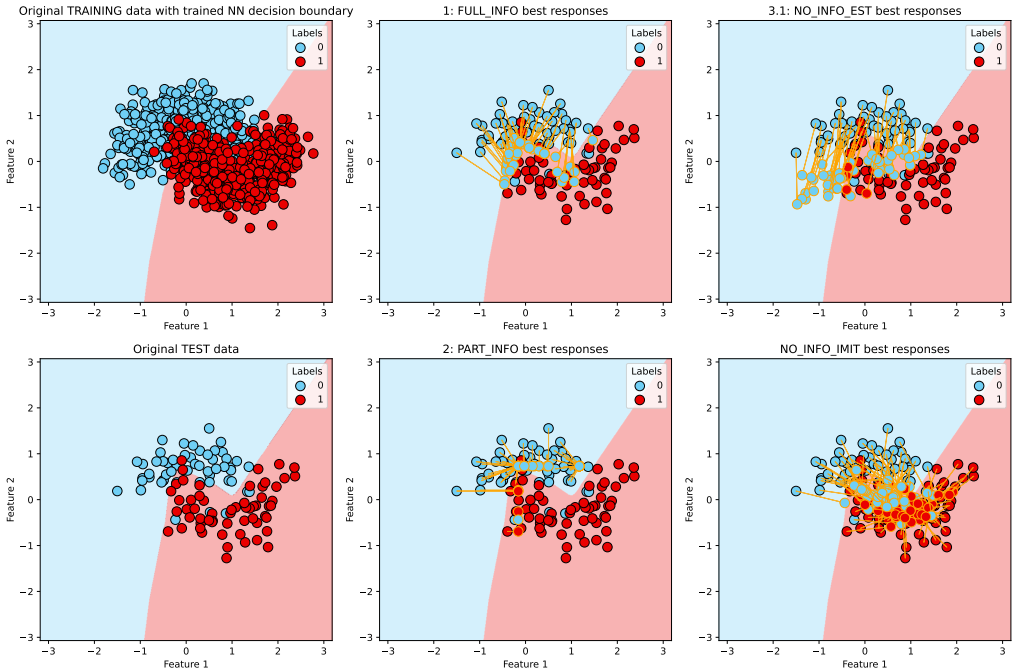


Source: own edit.

Figure 3: Strategic classification on synthetic data, f: LinearSVC

The exact numerical values for the different payoffs and results can be found in the Appendix A.1.

In the **FULL_INFO** scenario, every model—except the KNN model—allows all users to achieve a positive outcome by changing their features, as indicated by a user welfare of 100%. For the KNN model, this metric is similarly high at 99.12%. The percentage of users who made changes was also comparable across all models, ranging from approximately 43% to 48%. The average cost per change was lowest for the linear model at 0.01, while the non-linear models had slightly higher costs, between 0.02 and 0.03. Similarly, the average Contestant payoff was highest in the linear case at 0.2, compared to around 0.18 for the non-linear



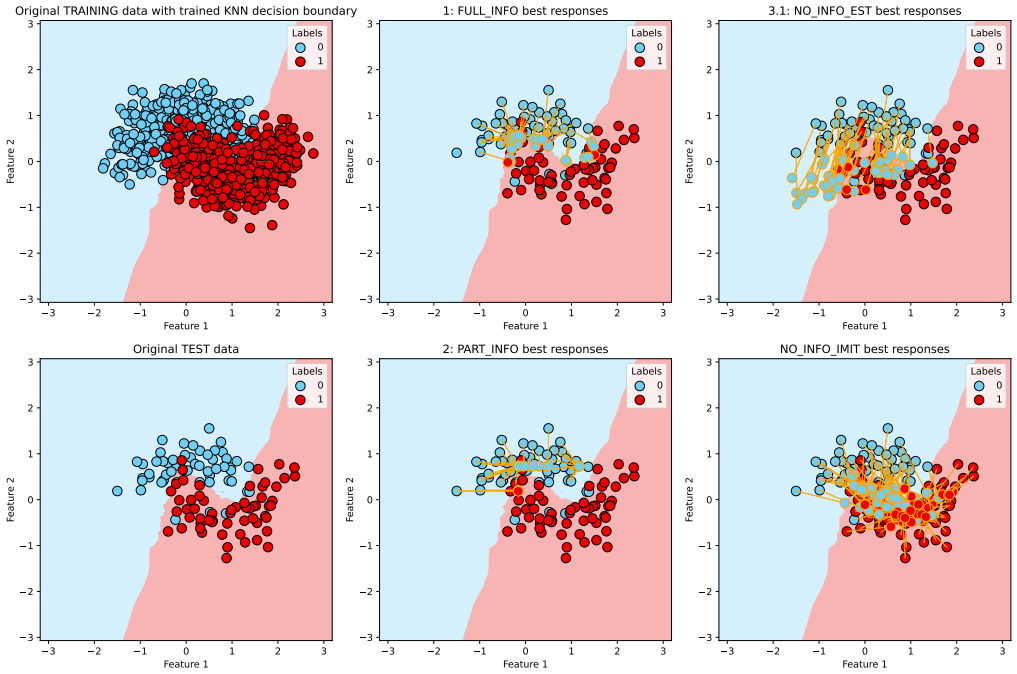
Source: own edit.

Figure 4: Strategic classification on synthetic data, f: Neural network with 4 layers, each containing 10 neurons

models. In summary, while model accuracy from the Jury's perspective was similar across all models, a slight distinction exists between linear and non-linear models: Contestants tend to receive better payoffs when the Jury used the linear model.

In the **PART_INFO** case, the opposite is the case: the payoff of the Jury is higher for non-linear models as for the linear model choice, as with former models had a 91.23% accuracy and latter only 87.72%. There is also a cut as much less users changed with the linear model with only 32.46% changing, while for all non-linear models at least 41% of the users changed their features. Average cost per change and also average Contestant payoff are similar. In summary, while individual payoffs of the Contestants where similar across models, in the PART_INFO case the Jury faced better payoff with non-linear models. Visually, the feature changes seem to be similar, indicating that the LIME explanations of different models are used similarly by the users.

In the **NO_INFO_EST** scenario, users approximate the model f with a linear substitute $\hat{f}$. Users modify their inputs the most with the linear model, showing a change rate of 48.25%, while this metric is around 47% for the other models. However, model accuracy is the lowest also for the linear case, achieving only 52.63% accuracy after the changes, compared to approximately 72-74% for the other models f. When examining the Contestants' payoffs, the models provide a similar average, with the NN and linear models having slightly



Source: own edit.

Figure 5: Strategic classification on synthetic data, f: KNN with $k = 11$

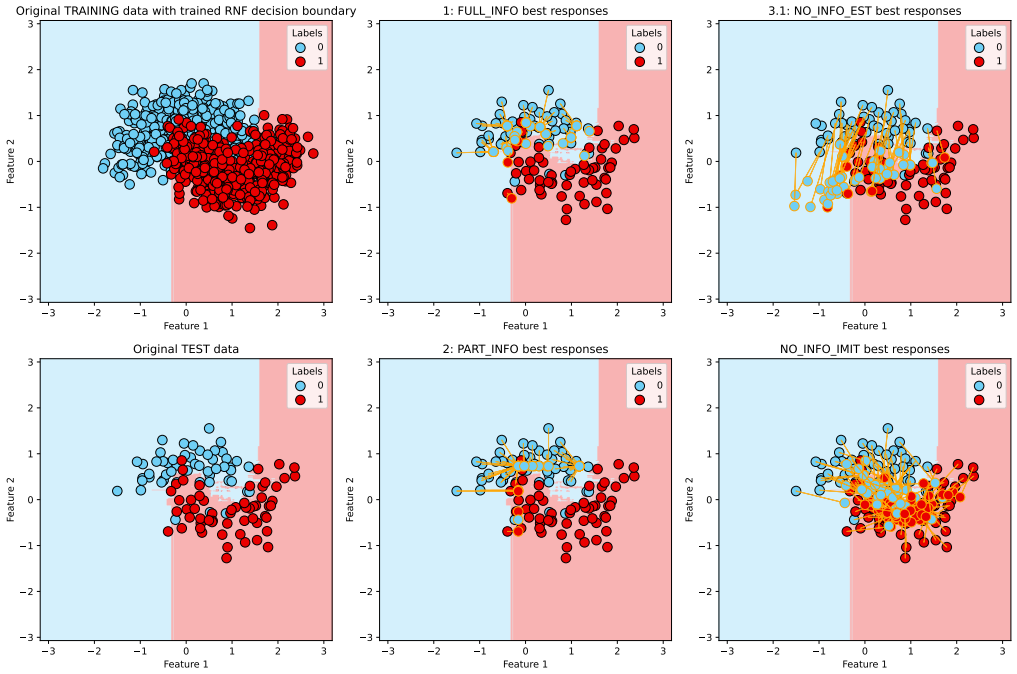
the highest value of 0.15. The average cost per change is fairly consistent for the NN, KNN and RNF models at 0.02, but it rises to 0.09 for the linear model. Regarding the size of disagreement set, the linear model has the lowest percentage with 2.63%, while the other models have at least 23% of users with differing labels to what they expected. In conclusion, in this case a subtle difference between linear and non-linear models is evident in the payoffs: the linear model offers lower accuracy, has a higher cost of change per change but also a lower disagreement set proportion than the other models.

In the **NO_INFO_IMIT** case, there are also some differences between linear and non-linear models. Model accuracy is the best for the linear model with 61.4%, while being around 50-54% for non-linear models. All users modify their features, with an average cost per change of around 0.15 for the linear and 0.1 for the non-linear models. This is reflected in the average Contestant payoff, which is 0.03 for the linear model, while for the other models, it is around 0.08. Hence, here a conflict of interest is showing: linear models provide better payoff for the Jury but lower payoff for the Contestants compared to non-linear models.

6.1.3 Price of Opacity

Next, I also evaluated the price of opacity as defined in Definition 3.2. The Price of Opacity in different information scenarios is depicted on Figure 7.

On the plot's x-axis, the sample size m is varied. As the sample size m does not influence cases where



Source: own edit.

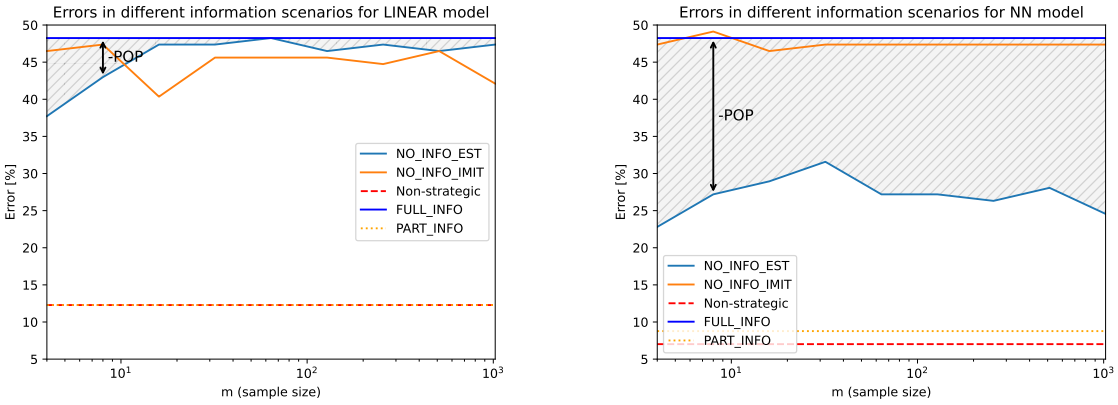
Figure 6: Strategic classification on synthetic data, f: Random Forest Classifier with 10 estimators.

no sampling is happening, the errors of the non-strategic, FULL_INFO and PART_INFO case are horizontal lines. The y-axis shows the error on the test set in percentage, $\text{err}_{\%}(f) := |h(x) \neq f(\Delta_{I_f}(x))|/n \cdot 100$ with n being the number of instances in the test set. The grey shading shows the price of opacity between the FULL_INFO and NO_INFO_EST case, however it is important to note that also areas between FULL_INFO and PART_INFO as well as between FULL_INFO and NO_INFO_IMIT correspond to POP values, just for other information scenarios.

For the linear model the price of opacity is negative for all opaque information scenarios. However, for larger sample sizes $m > 64$, the difference of errors between the FULL_INFO and the NO_INFO_EST cases is diminishing. The error in the PART_INFO case equals the error of the non-strategic model, i.e. the original error of the test data with model f.

The non-linear models on the other hand show a distinct pattern from this, being however similar to each other. The POP remains negative for almost all sample sizes - there is one case where the POP is positive: for sample size $m = 8$, the error of NO_INFO_IMIT is higher than the full information case with the NN model. This can however also be a statistical fluctuation, as the difference is quite small.

In order to compare my findings to Theorem 3.1 by Ghalme et al. [2021], I also estimated the probabilities of the enlargement set and the errors by calculating their proportions in the sample. I used $|E|/n$ as an estimate for $P_{x\sim D}\{x \in E\}$ and $|h(x) \neq f(\Delta_f(x))|/n$ as an estimate for $\text{err}(f, f)$. I calculated the size of the

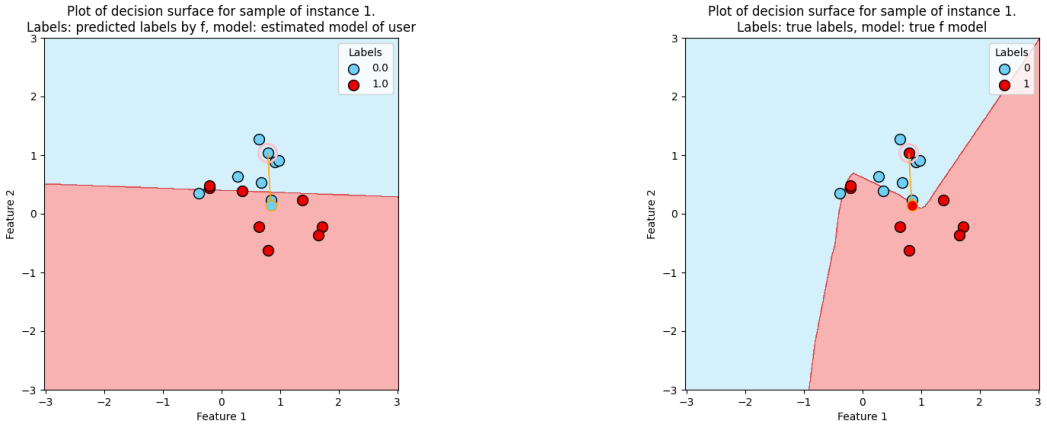


(a) Linear (b) NN

(c) KNN (d) RNF

Source: own edit.

Figure 7: Price of Opacity on the moons dataset depending on sample size m with different model choices f.



(a) The predicted labels in the sample $f(x)$ and the estimate of the model $\hat{f}$ (b) The true labels y and the model f used by the Jury

Source: own edit.

Figure 8: The sample T_C of user 1 (circled with pink) and their best response on the synthetic moons dataset with model choice f being NN.

enlargement set pairwise between the FULL_INFO case and the opaque cases (PART_INFO, NO_INFO_EST, NO_INFO_IMIT). The results can be seen in Table 3.

	PART_INFO	NO_INFO_EST	NO_INFO_IMIT	$2 \cdot \mathrm{err}(f, f)$
Linear	48.246	2.632	11.404	96.491
NN	42.982	23.684	5.263	96.491
KNN	47.368	28.070	5.263	94.737
RNF	44.737	29.825	9.649	96.491

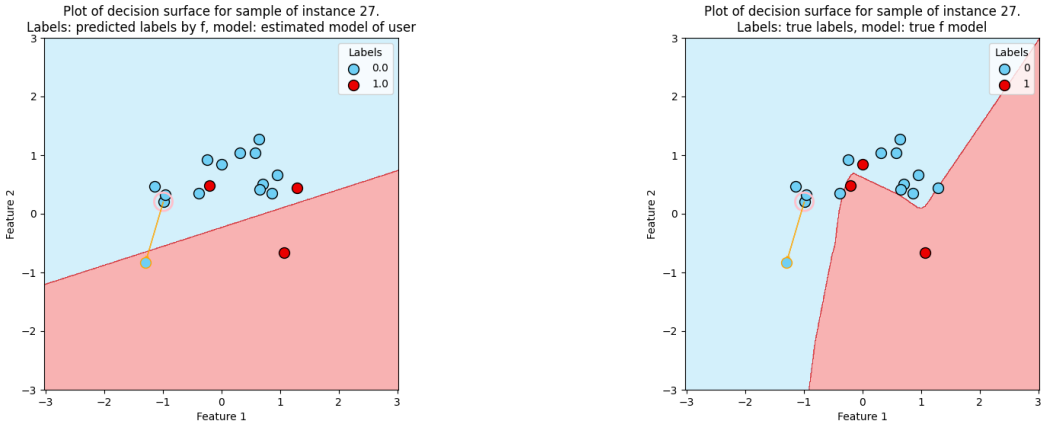
Table 3: Evaluation of the sufficient condition for the POP to be negative on the synthetic moon dataset. Rows: model choice for f. Columns: the enlargement set between FULL_INFO and given information scenario. All values are percentages.

Based on Theorem 3.1, if a value in one of the middle three columns is larger than the value in the rightmost column, the POP of that scenario should be strictly positive. This condition do not hold for any of the cases, and also, there are no positive POP values accordingly.

6.1.4 Data stories

In order to better understand the estimations of users, I also plotted some cases for the NO_INFO scenarios. The experiment is the same as before with the NN dataset, however the plots focus on the samples of a individuals that they base their best response on.

Figure 8 shows a user form the moons dataset in the NO_INFO_EST case that is classified by the Jury with a NN model. The user's true label $y = +1$, however by the model f of the Jury they get initially classified as 0. As this is also the case in their linear estimation $\hat{f}$ based on their sample T_C, they best respond to shift



(a) The predicted labels in the sample $f(x)$ and the estimate of the model $\hat{f}$ (b) The true labels y and the model f used by the Jury

Source: own edit.

Figure 9: The sample of user 27 (circled with pink) and their best response on the synthetic moons dataset with model choice f being NN.

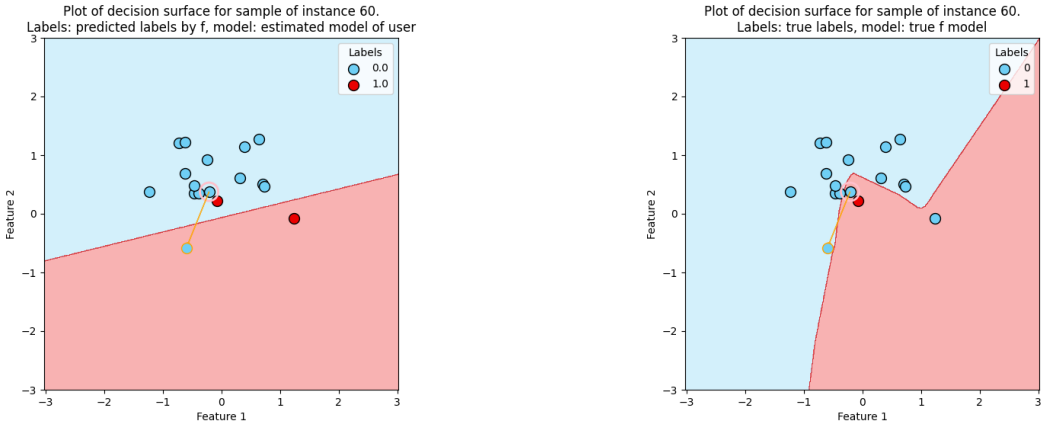
their features to the decision boundary of $\hat{f}$. In this case, they to also succeed to get a positive classification with the model f that the Jury actually employs, as $f(\Delta_{\hat{f}}) = 1$. This case is a win-win: as their true label $h(x)$ remains +1, from an initial false negative they get correctly classified as a +1, so also the Jury earns higher accuracy while the Contestant gets a positive prediction as well.

But this is not always the case. In Figure 9, we see another user who is classified negatively by both the true model of the Jury f, as well as the linear estimation of it $\hat{f}$ that the user calculates on their sample T_C. They find a feasible and worthwhile change as a best response to their estimated model, $\hat{f}$, however it does not change their outcome by f as it remains $f(x) = 0$. This way, the Jury's payoff from this user is unchanged, however the user changes their features and incurs costs to do so, however they do not get classified positively.

Another example, also noted by Ghalme et al. [2021], involves users who would have initially been classified positively, but based on their estimated model $\hat{f}$, mistakenly believe they will be classified negatively and thus alter their features. This occurs in the case of user 60 in the synthetic moons dataset, shown in Figure 10. The user incorrectly estimates that their original classification would be 0 and consequently modifies their features. However, the true model would have classified them as +1 before the change—after the modification, they end up receiving a negative classification.

6.2 Loan data

Next, I conducted the same experiments on the loan dataset introduced in Chapter 5.



(a) The predicted labels in the sample $f(x)$ and the estimate of the model $\hat{f}$ (b) The true labels y and the model f used by the Jury

Source: own edit.

Figure 10: The sample of user 60 (circled with pink) and their best response on the synthetic moons dataset with model choice f being NN.

6.2.1 Linearity VS non-linearity

To evaluate the differences in payoffs for linear and non-linear models, I ran the experiments for all 4 information scenarios for 4 different model choices f also on the loan dataset. The results for the different payoffs and results can be found in Tables 4, 5, 6, and 7. In these tables, each column represents an information scenario, with the "Original" column showing performance before the feature changes, i.e. non-strategic performance.

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	82.40	63.60	82.20	80.00	72.80
User welfare	41.20	75.60	42.60	36.00	63.60
Average cost of change per change		0.17	0.18	0.01	0.31
Average contestant payoff		0.09	0.02	0.07	-0.18

Table 4: The results on loan data, linear model

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	83.60	55.80	80.40	80.00	70.80
User welfare	42.80	85.40	48.40	35.60	68.00
Average cost of change per change		0.14	0.14	0.00	0.30
Average contestant payoff		0.11	0.03	0.07	-0.16

Table 5: The results on loan data, NN model

The accuracies on the test dataset were between 79%-84% for all the models before the change ("Original" column), with the neural network performing best. Also the user welfare on the data before the change

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	79.40	79.60	78.20	77.60	70.00
User welfare	39.40	54.80	43.80	41.60	59.20
Average cost of change per change		0.00	0.16	0.00	0.30
Average contestant payoff		0.11	0.00	0.08	-0.18

Table 6: The results on loan data, KNN model

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	81.80	52.60	81.20	79.20	69.20
User welfare	35.00	89.00	37.60	30.40	61.60
Average cost of change per change		0.08	0.15	0.00	0.28
Average contestant payoff		0.13	0.00	0.06	-0.16

Table 7: The results on loan data, RNF model

is the highest for the NN model with around 42.8% of the users being classified positively.

In the **FULL_INFO** scenario with non-linear models, there were some instances where the solver, related to the Contestant's best response problem, failed to produce a solution. If the optimization failed with the closest positive sample from the training set, a further example was chosen as the starting point for the optimization - choosing new iteration points until a solution was found up to 11 tries in total. If the 11th optimization also failed, it was assumed that the users did not modify their features. This decision was motivated both by computational limitations and by a realistic setup, where users also have limited computational resources. However, the issue remained for KNN and RNF models. There were more failed optimizations cases with less-smooth decision boundaries, such as those involving KNN with 45.2% failed optimizations or RNF with 9.4% failed optimizations - as opposed to linear and NN models with smooth predictions and 0 failed optimizations. Changing parameters such as the maximum iteration did not solve the problem either. The extremely high number of failed optimizations made the results on KNN models less comparable with other models, hence for this dataset I focused my comparison of linear and non-linear models to linear, NN and RNF models.

In the FULL_INFO scenario, the accuracies of all models dropped significantly compared to the accuracy before the feature changes. However, the accuracy of the linear model remained slightly higher at 63.6%, while the NN and RNF models dropped to around 55%. There were also notable differences in the proportion of users who changed their responses. The highest number of changes occurred with the RNF model, where 54% of users altered their decisions, followed by the NN model at approximately 42% and then the linear model with 34.4%. In the KNN model, only 15.4% of the users changed their features - this low percentage mirrors the high number of cases where the optimization process failed to find a solution and by my

assumption they remained at their original features. The average cost of change was highest for the linear model, followed by the NN model, while the RNF model had the lowest average of 0.08. The average Contestant payoff was highest for the RNF models with 0.13, while it was slightly lower—around 0.1—for the linear and NN models. Here the cut between linear and non-linear models was only distinguished in the payoff of the Jury. For the Contestants, the payoff varied more between non-linear models, showing no clear cut between the linear and non-linear cases.

In the **PART_INFO** case, the accuracies did not differ that much for the linear and non-linear cases. There were some differences in the percentage of user changes: the linear model has a lower rate of about 40% while with the non-linear models always around half of the users change. Also, the average Contestant payoff was extremely low for every model, with the highest value of 0.03 for the NN model, 0.02 for the linear and 0 for the RNF model. The disagreement set was the lowest for the NN model with 17%, while the linear and RNF models had values of around 30%.

In the **NO_INFO_EST** case, there was no clear difference in the accuracies based on model choice as it was around 80% for all model types. Also a similar amount of users changed with all models, around 60% of the test set. In this case, the cost of change was quite low with all models compared to other information scenarios. The average Contestant payoff was similar for all different model choices, between 0.07 and 0.08. The size of the disagreement set is much larger, than in the PART_INFO case: in this case around 63% of the users are classified differently than they "expect".

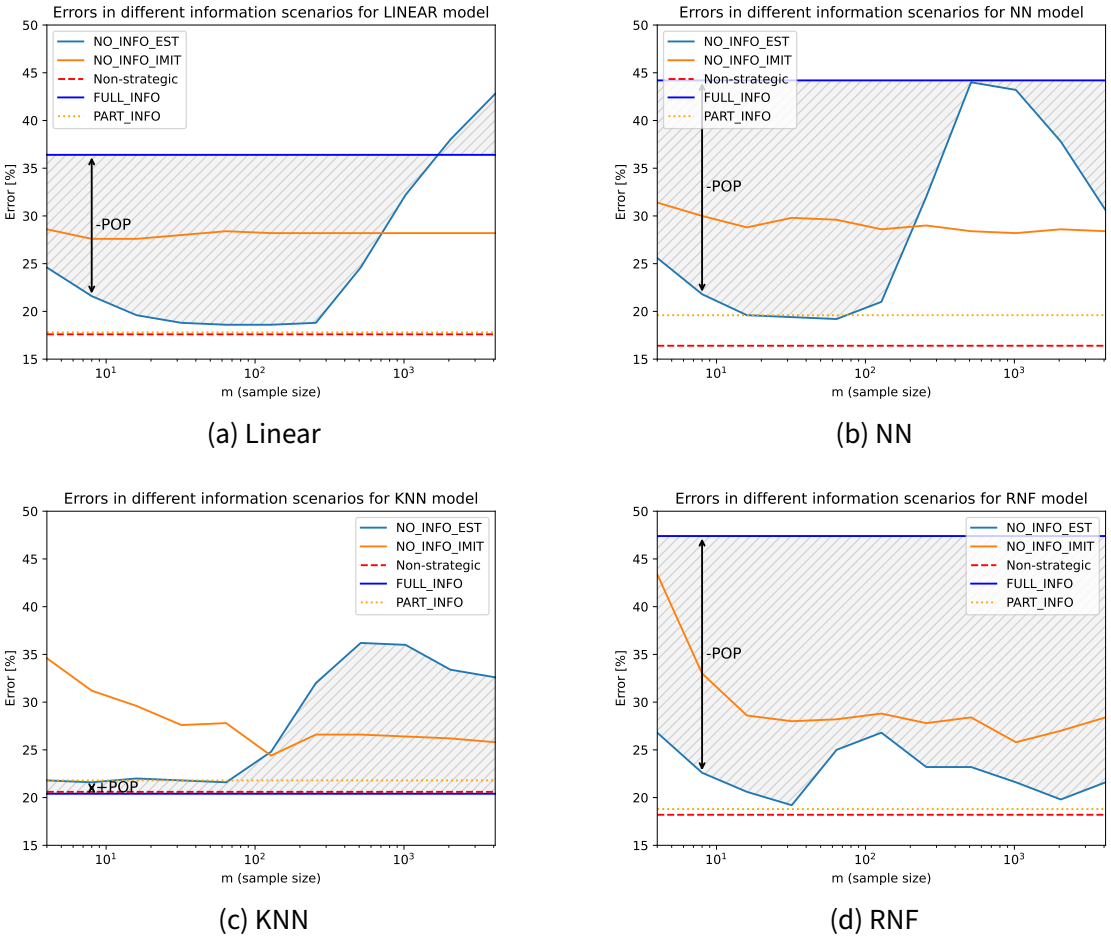
In the **NO_INFO_IMIT** case, the accuracies drop similarly in the linear and non-linear cases by around 10 percentage points compared to the "original" accuracy. The average cost of change is also very similar for all the models at around 0.3, while also the average Contestant payoff is similar at -0.18. The only differences show in user welfare, as for the NN it is around 68%, while for the other models it is slightly lower. To sum up, there is no disticts trend between linear and non-linear models.

6.2.2 Price of Opacity

The price of opacity is evaluated in Figure 11 for the loans dataset.

The top-left figure in Figure 11 corresponds to Figure 3 of the work of Ghalme et al. [2021]. The non-strategic and FULL_INFO lines have identical values in both cases. Similarly, the NO_INFO_EST case, which aligns with their "SVM (in the dark)" scenario, exhibits very similar behavior. Minor differences arise due to variations in our sampling processes as well as the optimisation process.

Despite these differences, the main trend remains: the POP is negative with the linear model, indicating



(a) Linear (b) NN

(c) KNN (d) RNF

Source: own edit.

Figure 11: Price of Opacity on the loans dataset depending on sample size m with different model choices f.

that the NO_INFO_EST case results in lower error compared to the FULL_INFO scenario. Additionally, all my other opaque information scenarios also yield negative POP values, reinforcing this pattern. Only when the sample size becomes very large does the NO_INFO_EST case show a higher error than the FULL_INFO scenario.

For the NN model, we observe similar patterns. The PART_INFO case has the lowest error, indicating a negative POP in this scenario. The error in the NO_INFO_IMIT case falls in the middle range and remains relatively stable across different sample sizes. In contrast, the error of the NO_INFO_EST scenario fluctuates significantly, even resulting in an almost zero POP for sample sizes $m = 512$ and $m = 1024$.

The KNN models results are again influenced by the high number of failed optimizations, hence they are not evaluated in detail.

The RNF model also exhibits negative POP across all sample sizes and information scenarios. The error in the NO_INFO_IMI case is quite high for small sample sizes, then more or less stabilizes at a lower level for larger samples. In contrast, the NO_INFO_EST case shows a more consistent relationship with sample size compared to other models.

Finally, I also evaluated the sufficient conditions of Ghalme et al. [2021], described in Theorem 3.1 for this dataset as well in Table 8. Table 8 demonstrates that the estimated sufficient conditions are not met in any

	PART_INFO	NO_INFO_EST	NO_INFO_IMIT	$2 \cdot \mathrm{err}(f, f)$
Linear	34.2	39.6	12.0	72.8
NN	37.4	49.8	17.4	88.4
RNF	51.8	59.0	32.2	94.8
KNN	15.0	16.0	19.6	40.8

Table 8: Evaluation of the sufficient condition for the POP to be negative on the loan dataset.. Rows: model choice for f. Columns: the enlargement set between FULL_INFO and given information scenario.

case—neither for any model choice f nor in any information scenario I_f. There are again no contradictions with my results.

6.3 Variations on cost function and budget

To address the generality of my results, I also conducted experiments in which I varied some parameters that were otherwise kept constant. I evaluated these variations on the NN model using both datasets.

6.3.1 Weight of Quadratic Element in Cost Function

On one hand, I varied the weight of the quadratic element, ε, in the mixed cost function described in Equation 1. By default, ε was set to 0.2 in all experiments. However, I also tested two additional values: a balanced mixture of linear and quadratic costs with $\varepsilon = 0.5$ and a more quadratic-dominant variant with $\varepsilon = 0.8$. Detailed results can be found in Section A.1.3 of the Appendix.

The effect of different quadratic weights varies significantly between the two datasets. On the synthetic moons dataset, higher ε generally leads to lower error across almost all information scenarios. However, this pattern does not hold for the loan dataset. Here, some information scenarios exhibit similar behavior (e.g., PART_INFO), some remain largely unaffected by changes of ε (e.g., NO_INFO_EST), and others show the opposite trend, with increasing error for higher ε values. Regarding the payoffs of the Contestants, in the moons dataset the average Contestant payoff decreases as ε increases for all information scenarios. On the loans dataset, the average Contestant payoff remains often similar for different ε values or does not show a linear trend.

6.3.2 Budget for changes

Next, I also varied the budget for changes, denoted as t. According to Equation 3, users modify their features only if the change is at most $2t$. By default, the budget was set to $t = 0.2$ to align with the experimental setup of Ghalme et al. [2021]. However, I also tested a lower budget of $t = 0.01$ and a higher budget of $t = 10$. Detailed results can be found in Section A.1.4 of the Appendix.

From the Jury's perspective, a higher cost budget often led to lower accuracy, particularly in the FULL_INFO and NO_INFO_IMIT cases for both datasets. However, there appears to be a saturation point beyond which users do not make more costly changes. This effect is evident in the synthetic dataset, where the differences between $t = 0.2$ and $t = 10$ are minimal.

From the Contestants' perspective, directly comparing the average payoff is not straightforward, as changes in t also affect utility. A more insightful approach is to examine the number of changes and the average cost per change. Here, the FULL_INFO and NO_INFO_IMIT cases are again the most sensitive to t: with a higher budget, a greater percentage of users make changes, and the average cost per change also increases. The saturation effect is observed here as well.

To summarize, variations in different parameters lead to even greater variability in the results other than variations in information scenarios and models.

7 Discussion

In this sections, I summarize my results and their implications for my research questions and compare them to existing literature. I structured this chapter according to my research questions: first, I compare linear and no-linear models in Section 7.1 and then I adress the payoffs of the two players in different information scenarios in 7.2.

7.1 Linearity VS non-linearity

One of my research questions examined how the payoffs of decision-makers and Contestants compare when non-linear models are involved. A key challenge in the non-linear case was determining the best response of users, as the optimization problem became non-convex at times depending on the model choice. Using non-convex constrained optimizers from `scipy.optimize` proved effective for linear models and certain non-linear cases, such as neural networks (NN), where solutions were always found—even if they exceeded the cost budget. However, for models like KNN and RNF there were cases, where the optimization failed, even after multiple initializations. This was especially the case for the real-world dataset with higher-dimensions. This finding shows, that even though gaming can be applied to certain non-linear models, black box optimization with `scipy.optimize` is not always applicable.

Regarding the payoffs of the players, the payoffs with linear and non-linear models remained highly dependent on the **information scenario**.

- **FULL_INFO:** In this setting, the Jury achieved at least as high accuracy with the linear model as with non-linear models—and in some cases, even higher. Contestant payoffs were similar across both model types. This finding is important, as it suggests that *linear models, when provided with full information, can lead to better overall outcomes*: the Jury benefits from higher accuracy, while individual users receive comparable payoffs on average.

- **PART_INFO:** Here, no consistent pattern emerges in terms of Jury accuracy across datasets. However, in both datasets examined, Contestant payoffs did not differ significantly between linear and non-linear models. This implies that under partial information, the Jury may experience varying accuracies depending on the dataset, but Contestant payoffs remain robust across model types.

- **NO_INFO_EST:** This scenario showed the *largest discrepancies* in Jury accuracy between linear and non-linear models. In one dataset, the linear model yielded accuracy that was 20 percentage points

lower than that of the non-linear model. Although Contestant payoffs were again similar, the overall user welfare was significantly higher under the linear model.

- **NO_INFO_IMIT:** In contrast, under this setting—where the Jury only observes user behavior and imitates choices—linear models performed better in terms of accuracy across both datasets. Again, Contestant payoffs remained similar. The last two information scenarios present an interesting asymmetry: when the Jury provides no information, the gaming choice of the users (estimating vs. imitating) can result in opposite outcomes in terms of Jury accuracy.

These findings emphasize that the relative performance of linear and non-linear models is *contingent on the information policy* and sometimes the dataset. The key takeaways can be summed up as:

- Under full transparency, linear models are more favorable for the Jury.
- Under no information with estimation, non-linear models achieve higher accuracy.
- Contestant payoffs tend to be similar regardless of the model type, but
- Overall user welfare can differ significantly between linear and non-linear models—even when individual payoffs do not. Non-linear models often provided better overall user welfare than linear models.

7.2 Information scenarios

My second research question was how the payoffs of decision makers and Contestants compare to each other in the different information scenarios. In this section, I would like to interpret the results from the previous chapter from the distinct viewpoints of the two players.

7.2.1 Payoffs of the decision-maker

From the perspective of the decision-maker, the results show that in all cases, accuracy did not improve in the strategic cases compared to the non-strategic, initial accuracy. Thus, any form of gaming harmed the Jury's payoff, regardless of the information scenario or the model choice f. This highlights the significance of gaming, demonstrating that it has a consistently negative effect for the Jury. To facilitate comparison across information scenarios, averages over the different model choices were also calculated, resulting in a single measure for each information scenario.

In the **FULL_INFO** case, the decrease in accuracy was highly dependent on the model choice f, but similar across both datasets. For linear models, the results align with previous studies [Hardt et al., 2016; Ghalme

et al., 2021], showing a substantial drop in accuracy as many users adapt their behavior — although this effect varied with different gaming budgets. Also for non-linear model choices, the effect is similar: many users successfully game the system, leading to a significant increase in error compared to the non-strategic performance before the feature changes.

In average over the model choices, FULL_INFO has a lower accuracy than the other information scenarios, meaning that using a transparent policy is worse for decision-makers than sharing less information. This is consistent with my hypothesis and the results from previous works [Hardt et al., 2016; Ghalme et al., 2021]. There were only a few cases where it resulted in better accuracy, for example the NO_INFO_EST scenario for the KNN model on the moons dataset.

The **PART_INFO** information scenario has not yet been evaluated in the field of strategic classification, so there are no prior results available for comparison. The findings of this thesis suggest that when the Jury publishes LIME-based explanations and users base their best responses on them, the decrease in accuracy is moderate or almost negligible. Notably, this is not due to a low number of users attempting to change their features — a significant percentage of users did try to adapt. However, many failed to achieve a positive prediction, as also illustrated in Figure 3. The figure shows that often many users move to the same level in at least one of the features, probably indicating the the boundaries in the LIME explanations are similar for users. Although users reached these boundaries of the LIME explanations, they were often unable to change the model's output. This suggests that the boundaries provided by LIME's local explanations may not accurately reflect the true global decision boundaries of the model.

Comparing this case with FULL_INFO case, the drop of accuracy in this case was smaller or comparable to the former, resulting in a negative (or near-zero) price of opacity. Thus, using LIME explanations as a means of model transparency appears to be a favorable choice for the Jury, as it preserved accuracy in most model choices across both datasets. Although in some cases, accuracy under PART_INFO was slightly lower than in the FULL_INFO scenario, the difference was minimal.

The **NO_INFO_EST** scenario corresponds to a setting similar to what Ghalme et al. [2021] proposed in their work. The results of this thesis align with their findings in the linear case on the loan dataset, as expected: the non-strategic and full-information accuracies are nearly the same, and there are only slight differences for the NO_INFO_EST case due to sampling and the different optimization processes. This alignment makes the other results even more valuable for comparison.

However, beyond this, the accuracy of the model in my results strongly depended on the dataset. In the synthetic case, the NO_INFO_EST scenario led to a significant drop in accuracy compared to the non-

strategic case, whereas in the loan dataset, the difference was minimal. A key similarity across both datasets is the strong influence of the user sample size on accuracy. Also, for both datasets, the models linearity had a high impact on the Price of Opacity (POP).

These results indicate that this scenario is highly sensitive to factors such as the data distribution, user sample size m, and model choice f. The significance of these findings highlights a critical downside of this opaque scenario compared to the transparent case — namely, its many dependencies, which makes the predictions on results in practical applications much harder. Thus, not only might the decision-maker achieve a worse payoff than in a transparent case, but their outcomes may also vary significantly from expectations if they misestimate m or the dataset distribution is different to what they expect (i.e. because of wrong sampling for the training process).

The **NO_INFO_IMIT** case was inspired by the work of Barsotti et al. [2022], which measures results based on the difference in false positives between strategic and non-strategic classification. While their methodology is not identical to mine, some results are comparable. Notably, their findings show that imitation leads to greater error compared to the non-strategic case, similar to what I observe.

Additionally, in their work, the number of imitators—comparable to the sample size *m* in my study—had a decreasing effect on error: as the number of imitators increased, the error increase diminished. In my experiments, this pattern was less consistent, with trends often remaining stable or fluctuating. However, for certain models, such as KNN and RNF in Figure 11, a similar trend was observed.

Regarding the role of model choice in this case, no clear trend emerges from the experiments. One consistent observation across both datasets is that RNF experiences the most significant accuracy drop in the NO_INFO_IMIT case compared to the strategic setting. These findings suggest that the effects of imitation are not strictly dependent on model linearity, though certain model choices can cause higher variation in results.

7.2.2 Payoffs of the Contestants

For the Contestants' payoffs, I evaluated three key aspects: the percentage of users who changed their features, the average cost per change, and the Contestants average payoff. For the PART_INFO and the NO_INFO_EST cases, I also evaluated the percentage of test users who had different estimations for their labels to what they got classified in the end, hence the proportion of the disagreement set to the test size. This section evaluates these results across different information scenarios from the Contestants' perspective.

In the **FULL_INFO** case, users altered their features in a similar percentage and had similar average pay-

offs for different models. Compared to other information scenarios, the FULL_INFO setting consistently yielded the highest average payoff in both datasets and for all model types, supporting the hypothesis that greater information leads to better user payoffs. Not only was the individual average consistently higher, but also overall user welfare exceeded the values of other information scenarios.

In the **PART_INFO** case, the percentage of users who modified their features was in the middle range of all the information scenarios, but this percentage was different for linear and non-linear models for both datasets. The average Contestant payoff followed a stable pattern across models, though it remained consistently low compared to other information scenarios. In some instances, the payoff was even zero, likely because LIME explanations, that the users rely on, do not perfectly predict the model's true outcome. The average cost per change was also relatively similar across models. In summary, while the PART_INFO case produced similar results across different models, the average payoff remained low in all cases compared to other information scenarios.

In the **NO_INFO_EST** case, users almost always exhibit the second highest change rate across all information scenarios, regardless of the dataset or model type. The average cost per change varies significantly between datasets, ranging from near zero to moderate (0.1), and is also higher for linear models than non-linear ones. On the other hand, the average Contestant payoff shows little variation between models, though it differs slightly between datasets. It is consistently the second highest among all information scenarios. The ratio of the disagreement set is also almost always higher compared to the PART_INFO case: for the loans dataset sometimes even being 2-3x larger than with LIME-explanations.

Ghalme et al. [2021] point out that in their opaque scenarios several users with initially positive prediction also change their features. My results confirm those cases as Section 6.1.4 shows. These are the cases that partly cause the average Contestant payoff to be less than in the FULL_INFO case.

Overall, the average Contestant payoff is largely unaffected by the decision-maker's model choice, which benefits Contestants since they have no control over that. Additionally, Contestants receive relatively high payoffs compared to other information scenarios, second only to the FULL_INFO scenario, which has better payoffs harmonizing with the results of Ghalme et al. [2021].

In the **NO_INFO_IMIT** case, all users modified their features across all models. This is because the assumption that they do not infer anything about the model itself, relying solely on the labels provided by it. Since Barsotti et al. [2022] does not focus on participant payoffs, there is no direct basis for comparison. However, as even users with initially positive predictions altered their features, this scenario resulted in the lowest average payoff among all information scenarios. The average cost per change was often close to the

budget limit, suggesting that most users exhausted their available budget.

Overall, user payoffs were highest, on average, in the **FULL_INFO** case, aligning with the initial hypotheses. The **NO_INFO_EST** scenario typically yielded the second-best payoffs, while **PART_INFO** and **NO_INFO_IMIT** resulted in the lowest payoffs. Interestingly, this result is counterintuitive—users achieved better outcomes when provided with no information compared to when they had access to model explanations and used those. However, there were also the most users who expected a different label based on their estimations to what the model assgned to them in the end. Notably, imitation—the simplest form of adaptation—produced the poorest payoffs.

8 Summary

In my thesis, I explored strategic classification, a framework that analyzes decision-making scenarios where decision-makers choose classifiers before individuals being classified can learn about the model and potentially adjust their features. The goal of the decision-makers is to achieve high classification accuracy, while users aim to get positive predictions. Existing literature examines this setting from various perspectives; however, previous studies lack a comparison of different amounts of information disclosed by decision-makers to users, as well as an analysis of non-linear model choices. My research addresses these gaps.

I defined four distinct information scenarios: full transparency, partial information with LIME explanations, and two cases with no information, where users base their responses on a sample. I tested these scenarios across multiple models, including linear SVMs, neural networks, random forests, and KNNs, to investigate differences between linear and non-linear classifiers. The experiments were conducted on two datasets: a synthetic dataset and a loan dataset.

8.1 Key research findings

1. Impact of Non-Linear Models on Strategic Classification Outcomes My findings indicate that the there can be differences in the payoffs between linear and non-linear models, especially for the Jury. However, the differences depended both on the information that the Jury published and the Contestants used in their best response, as on the data as well. Linear models provided better accuracy with full information and no information and imitation, while non-linear models had higher accuracy in the no information and estimation scenario. For the Contestants, the individual average payoffs were not affected by non-linearity, however typically overall user welfare was higher for non-linear models than linear ones.

2. Effect of Information Disclosure on Jury payoff: In all cases, full transparency led to a performance loss in accuracy compared to less transparent scenarios, reducing accuracy by 5-20% relative to the best possible outcome among the scenarios. However, the ranking of the other information scenarios was not as pronounced. Contrary to my hypothesis, partial information disclosure was often more beneficial to decision-makers than complete opacity. This suggests that providing LIME explanations could be a strategic advantage for decision-makers.

3. Effect of Information Disclosure on Contestant payoff: From the users' perspective, full transparency yielded the best results. Surprisingly, partial information was not particularly helpful, as the no-information scenario often resulted in better outcomes. The largest discrepancy between anticipated and

actual classifications occurred in the no information case.

3. Characteristics of best responses in different information scenarios: Based on the figures, the best responses for different information scenarios take different forms: not only is the set of users often different, the same users often react differently for different information scenarios. The most similar responses show visually between the full transparent and the no-information estimation cases, however the best responses in the latter are often suboptimal by either being more costly or not resulting in a positive prediction. In the partial information case, users often move to the same feature values.

Key takeaways: My results underscore a recurring theme: what benefits decision-makers often disadvantages Contestants, both in model choice and information scenario. However, there were counter-examples for both aspects. For model choice, all the non-linear models proved to be benefitial for both players compared to a linear model choice on one of the datasets. For information scenario on the other hand, the no information with imitation case was suboptimal for both players. This highlights an important insight: if decision-makers anticipate imitation, they may be better off providing partial information or assisting users in estimating the model's behavior.

Overall, my research demonstrates the diverse payoffs that arise from different model types, datasets, parameters, and information strategies. It shows how an machine learning models can not only affect decision-makers, but also the individuals being classified - opening many further areas for further exploration.

8.2 Limitations and further research

This thesis compares linear and non-linear models across different information scenarios, but several important aspects remain unexplored.

One of the main limitations of this thesis lies in the use of imperfect optimization algorithms. Since the focus was on black-box optimization to enable a broad comparison across different models, this came at the cost of the optimizer not always finding the best possible solutions. Additionally, the optimizer used in this work is highly sensitive to the initialization point. Further experiments analyzing the statistical distribution of results based on different initialization strategies would be a valuable extension of this thesis.

Moreover, especially in the complex full-information setting, the computed best response may not accurately reflect how humans would act in real-world scenarios. While some information setups, such as partial or no-information cases, attempt to imitate realistic conditions, there is no guarantee that they fully capture human behavior.

Future research could examine alternative optimization strategies used by individuals, particularly for

random forests and KNN models, where challenges persist. Expanding the study to additional datasets would help assess the generalizability of these findings. Additionally, further analysis of parameters that remained fixed in this study—such as budget constraints, cost functions, or threshold settings in LIME explanations—could provide deeper insights.

A key limitation of this thesis is its reliance on LIME for model explanations, without considering alternative methods. Also, with the LIME explanation the algorithm designed in this thesis allows room for improvement, as often it leads to changes in features that do not change the Contestants labels.

Also the inherent randomness and variability of non-linear models also introduce uncertainty, which was not extensively analyzed, leaving room for evaluating the statistical significance of my findings. Furthermore, robustness algorithms designed to counteract gaming were excluded, as they are either tailored for linear classifiers or incompatible with non-linear models.

Previous research differentiates between "faking" and "gaming" in strategic classification. Exploring these behaviors under different information scenarios, particularly in non-linear settings, would be a valuable extension. Additionally, real-world case studies on transparency policies could refine the assumptions underlying these scenarios, making them more applicable to practical decision-making contexts. The same applies to users: it would be interesting to study how users in real-world scenarios modify their features.

9 References

2016/679, R. E. (2016-05-04). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *OJ*, *L-119*, 40-43.

Agarap, A. F. (2019). *Deep learning using rectified linear units (relu).* Retrieved from https://arxiv.org/abs/1803.08375

Ahmadi, S., Beyhaghi, H., Blum, A., & Naggita, K. (2021). The strategic perceptron. In *Proceedings of the 22nd acm conference on economics and computation* (pp. 6–25).

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(4), 1059–1086.

Bambauer, J., & Zarsky, T. (2018). The algorithm game. *Notre Dame L. Rev.*, *94*, 1.

Barsotti, F., Koçer, R. G., & Santos, F. P. (2022). Transparency, detection and imitation in strategic classification. In *31st international joint conference on artificial intelligence, ijcai 2022* (pp. 67–73).

Baştanlar, Y., & Özuysal, M. (2014). Introduction to machine learning. *miRNomics: MicroRNA biology and computational analysis*, 105–128.

Bechavod, Y., Ligett, K., Wu, S., & Ziani, J. (2021). Gaming helps! learning from strategic interactions in natural dynamics. In *International conference on artificial intelligence and statistics* (pp. 1234–1242).

Bechavod, Y., Podimata, C., Wu, S., & Ziani, J. (2022). Information discrepancy in strategic learning. In *International conference on machine learning* (pp. 1691–1715).

Björkegren, D., & Grissen, D. (2018). The potential of digital credit to bank the poor. In *Aea papers and proceedings* (Vol. 108, pp. 68–71).

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Brückner, M., Kanzow, C., & Scheffer, T. (2012). Static prediction games for adversarial learning problems. *The Journal of Machine Learning Research*, *13*(1), 2617–2654.

Brückner, M., & Scheffer, T. (2009). Nash equilibria of static prediction games. *Advances in neural information processing systems*, *22*.

Brückner, M., & Scheffer, T. (2011). Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th acm sigkdd international conference on knowledge discovery and data mining* (pp. 547–555).

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? workforce implications. *Science*, *358*(6370), 1530–1534.

Chen, Y., Liu, Y., & Podimata, C. (2020). Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems*, *33*, 15265–15276.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273–297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21–27.

Dalvi, N., Domingos, P., Mausam, Sanghai, S., & Verma, D. (2004). Adversarial classification. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 99–108).

Dee, T. S., Dobbie, W., Jacob, B. A., & Rockoff, J. (2019). The causes and consequences of test score manipulation: Evidence from the new york regents examinations. *American Economic Journal: Applied Economics*, *11*(3), 382–423.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., & Wu, Z. S. (2018). Strategic classification from revealed preferences. In *Proceedings of the 2018 acm conference on economics and computation* (pp. 55–70).

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dundar, M., Krishnapuram, B., Bi, J., & Rao, R. B. (2007). Learning classifiers when the training data is not iid. In *Ijcai* (Vol. 2007, pp. 756–61).

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

Flaxman, A. D., Kalai, A. T., & McMahan, H. B. (2004). Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles.

Ghalme, G., Nair, V., Eilat, I., Talgam-Cohen, I., & Rosenfeld, N. (2021). Strategic classification in the dark. In *International conference on machine learning* (pp. 3672–3681).

Gobler, E. (2022, Dec). *8 hacks to boost your credit score quickly.* erin gobler. (URL: https://eringobler.com/boost-your-credit-score/, Accessed: 2024-04-25)

Goodhart, C. (1976). *Monetary relationships: a view from threadneedle street*. University of Warwick.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50–57.

Greenstone, M., He, G., Jia, R., & Liu, T. (2022). Can technology solve the principal-agent problem? evidence from china's war on air pollution. *American Economic Review: Insights*, *4*(1), 54–70.

Großhans, M., Sawade, C., Brückner, M., & Scheffer, T. (2013). Bayesian games for adversarial regression problems. In *International conference on machine learning* (pp. 55–63).

Haghtalab, N., Immorlica, N., Lucier, B., & Wang, J. Z. (2020). Maximizing welfare with incentive-aware evaluation mechanisms. *arXiv preprint arXiv:2011.01956*.

Hardt, M., Megiddo, N., Papadimitriou, C., & Wootters, M. (2016). Strategic classification. In *Proceedings of the 2016 acm conference on innovations in theoretical computer science* (pp. 111–122).

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282).

Hooker, G. (2004). Discovering additive structure in black box functions. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 575–580).

Hu, L., Immorlica, N., & Vaughan, J. W. (2019). The disparate effects of strategic manipulation. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 259–268).

Kingma, D. P., & Ba, J. (2017). *Adam: A method for stochastic optimization.* Retrieved from https://arxiv.org/abs/1412.6980

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, *133*(1), 237–293.

Kleinberg, J., & Raghavan, M. (2020). How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, *8*(4), 1–23.

Kumar, A., & Zinger, M. (2014). *Network analysis of peer-to-peer lending networks.* (CS224W Project Final Paper)

Levanon, S., & Rosenfeld, N. (2021). Strategic classification made practical. In *International conference on machine learning* (pp. 6243–6253).

Miller, J., Milli, S., & Hardt, M. (2020). Strategic classification is causal modeling in disguise. In *International conference on machine learning* (pp. 6917–6926).

Milli, S., Miller, J., Dragan, A. D., & Hardt, M. (2019). The social cost of strategic classification. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 230–239).

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

MoneyNerd. (2023, Apr). *3 hacks to boost your credit score in under a minute!* YouTube. YouTube. (URL: https://www.youtube.com/watch?v=yMDv2p3pVbk, Accessed: 2024-04-25)

OpenAI. (2024). *Chatgpt (june 14 version).* `https://chat.openai.com/`. (Large language model, Retrieved from https://chat.openai.com/)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Édouard Duchesnay (2018). *Scikit-learn: Machine learning in python.* Retrieved from `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html` (Accessed: 2025-01-30)

Perdomo, J., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. In *International conference on machine learning* (pp. 7599–7609).

Perlich, C., Dalessandro, B., Raeder, T., Stitelman, O., & Provost, F. (2014). Machine learning for targeted display advertising: Transfer learning in action. *Machine learning*, *95*(1), 103–127.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Rosenfeld, N., Hilgard, A., Ravindranath, S. S., & Parkes, D. C. (2020). From predictions to decisions: Using lookahead regularization. *Advances in Neural Information Processing Systems*, *33*, 4115–4126.

Shapley, L. S., et al. (1953). A value for n-person games.

Shavit, Y., Edelman, B., & Axelrod, B. (2020). Causal strategic linear regression. In *International conference on machine learning* (pp. 8676–8686).

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, *41*, 647–665.

Sundaram, R., Vullikanti, A., Xu, H., & Yao, F. (2023). Pac-learning for strategic classification. *Journal of Machine Learning Research*, *24*(192), 1–38.

Zhang, H., & Conitzer, V. (2021). Incentive-aware pac learning. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 5797–5804).

A Appendix

A.1 Complete results of experiments

A.1.1 Synthetic Gaussian-clustered dataset

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	88.60	59.65	89.47	63.16	67.54
User welfare	57.02	96.49	57.89	91.23	85.09
Average cost of change per change		0.01	0.13	0.04	0.22
Average contestant payoff		0.19	0.07	0.16	-0.05

Table 9: The results on synthetic Gaussian-clustered data, linear model

A.1.2 Synthetic moons dataset

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	87.72	51.75	87.72	52.63	61.40
User welfare	51.75	100.00	51.75	97.37	88.60
Average cost of change per change		0.01	0.05	0.09	0.15
Average contestant payoff		0.20	0.09	0.15	0.03

Table 10: The results on synthetic moons data, linear model

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	92.98	51.75	91.23	71.93	53.51
User welfare	57.02	100.00	57.02	76.32	94.74
Average cost of change per change		0.03	0.04	0.01	0.11
Average contestant payoff		0.19	0.10	0.15	0.08

Table 11: The results on synthetic moons data, NN model

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	92.11	52.63	91.23	73.68	50.88
User welfare	52.63	99.12	53.51	71.05	93.86
Average cost of change per change		0.03	0.04	0.01	0.10
Average contestant payoff		0.18	0.09	0.14	0.09

Table 12: The results on synthetic moons data, KNN model

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	91.23	51.75	91.23	72.81	54.39
User welfare	53.51	100.00	55.26	70.18	90.35
Average cost of change per change		0.02	0.04	0.03	0.11
Average contestant payoff		0.19	0.09	0.13	0.08

Table 13: The results on synthetic moons data, RNF model

A.1.3 Weigh of quadratic element in cost function

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	92.98	53.51	90.35	80.70	56.14
User welfare	57.02	98.25	59.65	71.05	92.11
Average cost of change per change		0.08	0.09	0.13	0.21
Average contestant payoff		0.16	0.08	0.09	-0.02

Table 14: The results on synthetic moons data, NN model, epsilon = 0.5

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	92.98	57.02	86.84	83.33	58.77
User welfare	57.02	94.74	63.16	66.67	89.47
Average cost of change per change		0.12	0.16	0.16	0.26
Average contestant payoff		0.14	0.05	0.09	-0.09

Table 15: The results on synthetic moons data, NN model, epsilon = 0.8

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	83.60	51.80	80.80	80.20	72.80
User welfare	42.80	89.80	46.80	35.80	66.00
Average cost of change per change		0.15	0.46	0.06	0.35
Average contestant payoff		0.11	-0.14	0.04	-0.22

Table 16: The results on loan data, NN model, epsilon = 0.5

A.1.4 Budget for changes

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	83.60	48.00	82.80	81.40	74.00
User welfare	42.80	94.40	44.40	37.80	64.40
Average cost of change per change		0.11	0.24	0.13	0.38
Average contestant payoff		0.13	-0.03	0.04	-0.25

Table 17: The results on loan data, NN model, epsilon = 0.8

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	92.98	60.53	91.23	75.44	67.54
User welfare	57.02	91.23	57.02	72.81	82.46
Average cost of change per change		0.00	0.04	0.00	0.01
Average contestant payoff		0.01	-0.01	0.01	-0.00

Table 18: The results on synthetic moons data, NN model, t = 0.01

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	92.98	51.75	91.23	71.93	50.88
User welfare	57.02	100.00	57.02	76.32	97.37
Average cost of change per change		0.03	0.04	0.01	0.13
Average contestant payoff		9.99	5.69	7.63	9.60

Table 19: The results on synthetic moons data, NN model, t = 10

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	83.60	81.20	80.60	80.00	83.20
User welfare	42.80	52.80	48.20	35.60	44.40
Average cost of change per change		0.00	0.14	0.00	0.02
Average contestant payoff		0.01	-0.07	0.00	-0.01

Table 20: The results on loan data, NN model, t = 0.01

	Original	FULL_INFO	PART_INFO	NO_INFO_EST	NO_INFO_IMIT
Accuracy	83.60	42.40	80.40	80.00	42.80
User welfare	42.80	100.00	48.40	35.60	99.60
Average cost of change per change		0.27	0.14	0.02	0.67
Average contestant payoff		9.85	4.77	3.55	9.29

Table 21: The results on loan data, NN model, t = 10