ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Master Thesis
MSc Economics & Business

The acceptance of algorithms for automated profiling with a focus on Black-box models

J.C. Rovekamp
Student ID number: 434713
15-12-2020

Supervisor: prof.dr. D. Fok
Second assessor: dr. C.S. Bellet

Abstract

The number of crucial decisions made based on Artificial Intelligence is increasing rapidly. The discussion to what extent these decisions have to be explainable is still ongoing. This study contributes to this discussion for people living in the Netherlands. We find that people are willing to lose explainability if this means a higher accuracy and fairness. However, in many cases people do not trust the decisions made by algorithms more than decisions made by humans. Especially because they believe humans can give a more interpretable explanation. When the consequences of the decision become less serious more people tend to be more comfortable with an algorithm making the decision.

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Acknowledgements

This study is written under supervision of Capgemini Invent. Capgemini Invent is part of the Capgemini Group. The history of the Capgemini group starts back in 1967 when Serge Kampf found Sogeti. Since then the company has grown a lot and is now one of the top consultancy firms worldwide. The practice of Capgemini is briefly described by them as: *“Capgemini is a global leader in consulting, digital transformation, technology and engineering services. The Group is at the forefront of innovation to address the entire breadth of clients’ opportunities in the evolving world of cloud, digital and platforms.”* A more specific description of the practice of Capgemini Invent: *“Capgemini Invent combines strategy, technology, data science and creative design to solve the most complex business and technology challenges.”*

During this research I had the privilege to be part of the Data Science & Data strategy team which is part of the capability Insight Driven Enterprise. This team deals with a large variety of clients with a focus on the financial sector. The team has provided me with two supervisors, which helped me a lot in a supportive role. Beside the support Capgemini offered for this study, it provided a very educational environment. Therefore, I would like to thank Capgemini Invent for their support. In addition, I would like to thank two colleagues in particular. Tijmen Wintjes and Henk Vermeulen, thank you for being my supervisors at Capgemini Invent and for the feedback you have given.

Lastly, I would like to thank professor Fok and the Erasmus University. Professor Fok for his guidance and feedback during the thesis and Erasmus University for all the great years I have spent there.

List of Figures and Tables

List of Figures

List of Tables

Contents

1 Introduction

Lately more and more crucial decisions are based on Artificial Intelligence (AI) (Adadi & Berrada, 2018). However, even though there is an increase in the use of AI there is still a lot of skepticism around the application of AI in multiple areas (Hengstler, Enkel, & Duelli, 2016). Reasons to justify the scepticism regarding the use of AI in critical decision making are some well-known examples where there were biases in the data. Especially when AI is applied to human beings for automated profiling, the consequences of mistakes can be distressing. An example from Amazon within Human Resources (HR), which is among the sectors that is affected by the development of AI (Bodie, Cherry, McCormick, & Tang, 2017; Gulliford & Dixon, 2019), is the use of an algorithm to review resumés. This project started in 2014, however Amazon decided to stop using it in 2018 when they found out the algorithm favoured men (Reuters, 2018). Another example comes from Unilever, which claims to save millions of euros by using AI to asses job interviews (Wilson & Daugherty, 2018). Furthermore, an area where AI is used for automated profiling is the criminal justice system. The system that is used in the United States is called COMPAS. The algorithm behind COMPAS does not use race as feature, however it still turned out to be biased against blacks (Angwin, Larson, Kirchner, & Mattu, 2016). This can happen when other features in the model are correlated with a certain race.

Situations, in which algorithms turn out to be biased, can cause companies or institutions a great deal of problems. Transparency about why algorithms make certain decisions is therefore deemed important by a large group of people. However, especially when so called Black-box models are used for automated profiling, transparency is very low. Low transparency means that the reasons for a certain outcome are unknown. This puts organisations in a difficult position since a simple: “the algorithm told me to do so”, will not suffice.

With the entrance of the GDPR rules in May 2018, new rules were set regarding the use and storage of data. Two articles that have an impact on the use of machine learning (ML) and AI when it comes to automated profiling are article 15 and 22. If automated decision making is used, the data subjects have the right to obtain meaningful information about the consequences of such processing and the logic and significance of how the decision was made (article 15.1.h). And, data subjects have the right that a decision is not solely based on automated processing, when the effects of the automated processing produces legal effects or similar effects which have a significant impact (article 22.1). What is meant by legal effects or similarly significant and the right to meaningful information are somewhat vague. However, recital 71 says people have a right to an explanation about how a decision was reached. Examples for the context that are given are an online credit card application or e-recruiting without human intervention.

However, there is also another opinion regarding the use of Black-box models, one could argue that the best solution should always be used. Therefore, if a Black-box model provides a solution that is best it should be preferred over other solutions (Holm, 2019). This raises the question what, from an ethical point of view, is the best option? In other words, what would be the ethical thing to do if one did not have to take GDPR rules into consideration? Ethics deal mostly with what is deemed good for humans (Alder & Gilbert, 2006). Furthermore, making ethical decisions has been proven to be beneficial for companies (Key & Popkin, 1998). As Mantelero (2018) stated, the GDPR rules are mainly focused on data protection but fail to tackle societal and ethical issues.

This study will focus on the opinion of the public concerning the use of algorithms, with a focus on the use of Black-box models, for automated profiling. This study challenges the call of the GDPR to only allow profiling which is not solely based on automated decision making, and the right to an explanation. An important question is what would be lost if the use of Black-box models will be completely ruled out? And it is important to recognise the alternatives, Black-box solutions might not be optimal, but do we, as a society, currently have a better alternative? Van den Heuvel and Bondarouk (2017, p. 29) describe the situation within HR analytics nicely: “Conducting analytics on employee data can probably go as far, and develop as fast, as employees approve. Of course, there is the need for organisations to remain compliant to data privacy legislations, but they may be most dependent on the trust they have from these employees to use ‘their’ data for the greater good of the business.”. This quote suggests that the trust in the models that are being used is of high importance. This study will find out how much trust people have in different kind of models, when clearly explained. This is, to our knowledge, the first study to do so. Since not everyone is aware of what the difference is between black-box models and explainable AI, the public is not asked directly if they approve of black-box models, but indirectly by selecting the model they prefer for a certain task.

Research Questions:

Research question 1: "To what extent does the public in the Netherlands approve the use of algorithms, with a focus on Black-box models, for profiling?"

Research question 2: "What are the differences in the amount of trust people in the Netherlands tend to have in AI-based decision-making models based on their experience-level of working with data and the given context?"

To answer the research questions the performance of transparent and Black-box models will be evaluated on three datasets. These three datasets contain the following data: a dataset to predict loan defaults, a dataset for bank marketing and a dataset to predict student performance. On these three datasets five different algorithms will be trained and their results will be compared. These results will be used to map the difference in performance for the different models. Reason to include multiple datasets is to check for stability among the results. Therefore, each model is trained on each dataset to get a fair comparison of the performance of the different models. The difference between the outcome of the transparent and Black-box models once trained on these datasets will be clearly presented in a survey. Meaning that the accuracy, interpretability and a check for fairness will be presented. Based on this the survey can determine which attribute is deemed most important and perhaps there are differences based on whether someone has experience on working with data. This survey will give insight into what is deemed right by the public, concerning the use of AI-based decision-making models in profiling. Answering the question if the public requires an explanation, implicitly answers the approval of Black-box models for this study. The following hypotheses will be tested:

Hypothesis 1 (H1): *The public in the Netherlands approves the use of Black-box modelling for profiling, if the model is more accurate then the alternative and has proven to be fair.*

Hypothesis 2 (H2): *People will be more hostile towards AI-based decision-making if the consequences are more serious.*

Hypothesis 3 (H3): *People who are familiar with data have more trust in AI-based decision-making.*

First several concepts and theories concerning this study will be evaluated in the literature review. In the methodology the data collection, technical methods and the survey are described. In the data section the datasets that are being used and the transformations that are made are described. The result section is split up in technical results, containing the results of the different models and in a section containing the results of the survey. Finally, in the conclusion the most important results are summarised, and the research questions are answered.

2 Literature Review

Due to the recent increase of big data there is a lot of interest in AI and recently more companies are starting to act on it. However, AI as a field of study originates from the 1950's. There are several ways to classify "different" kinds of AI. Kaplan and Haenlein (2019) provide two ways of classifying AI. One way of classifying AI is to split into analytical, human-inspired, and humanised AI. This way of classifying is based on the type of intelligence the AI exhibits. Analytical AI exhibits cognitive intelligence, such as pattern recognition. Human-inspired AI possess the ability of cognitive intelligence and emotional intelligence. Lastly Humanised AI, exhibits cognitive, emotional and social intelligence. Another way of classifying AI is into Artificial Narrow, General and Super intelligence depending on the evolutionary stage of AI. Narrow AI can be described as AI that is designed to do one specific task as good or better than humans, like driving a car. This study focuses on Artificial Narrow, because the system only has to deal with one task. Artificial General Intelligence is an AI that performs as well or better than humans in multiple areas. Lastly Artificial Super intelligence is an AI that outperforms humans in all areas. What tends to happen to AI, however, is that when it reaches a state where it becomes mainstream, it is no longer considered AI. An example of this are search engines like the one from Google. Nowadays AI systems are capable of tasks which were deemed impossible a few decades ago. For example, an AI beating the world champion in chess or GO. Due to this development and this process where AI might become a part of everyone's life one might ask if there is a need for regulation of the use of AI? Perhaps not the use of the final AI itself needs regulation since AI is inherently unbiased, but a set of rules might be needed for the training of these models (Haenlein & Kaplan, 2019). In the literature review the following items will be discussed: Ethics, Profiling and the GDPR, Human decision making, The difference between transparent and black-box models, Transparency of AI, and Fairness of AI.

2.1 Ethics

Ethics is the study of what is the "right" thing to do.Within Ethics there are several movements with their own opinions regarding ethics. Two large contradicting movements are deontology and consequentialism. Deontology has a focus on the intention of peoples action. Therefore, the results of such an action are not of importance to determine wether an action is good or bad. This means that when the results of an action are good, but the intentions are not, the action should not be undertaken (Alexander & Moore, 2016). Consequentialism as described by Sinnott-Armstrong (2019) is focussed on the actions themselves. In contrast to Deontology the intentions of an action are not important. Which movement is used to determine what is right can have a significant effect on what is deemed right regarding the use of black-box models. The use of Black-box models with the intention to get the most accurate results without bias might be a good idea from a Deontology point of view. However, when some of the consequences are taken into account, e.g. not being able to explain your actions, the result might be different. This study will look at ethics from a Consequentialism point of view. Therefore, not the intentions of the AI matter, but the actions, the results or the effects it causes, are of importance. Reason for this is that in this study it is assumed that the intentions are good when the AI models are trained.

2.2 Profiling and the GDPR

For this study the GDPR definition of profiling is used. Profiling is defined as follows by the GDPR (article 4.4): " 'profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements." The definitions of personal data and processing can also be found in appendix A.

According to Wachter (2017) the rules regarding profiling stated by the GDPR in articles 15 and 22 contain a lot of ambiguity. They state that in the current form of the GDPR one does not have a right to an explanation, which is often thought, but rather a right to be informed. Wachter (2017) divide the 'right to an explanation' into two groups: system functionality and specific decisions. Whereas the system functionality refers to e.g. the significance and other details about the model, and specific decisions are about reasons or decision rules. The GDPR only enforces the right to be informed about system functionality. A right to an explanation is only written in the recitals which are not legally binding. For this study it is interesting to find out whether people actually demand detailed information. This study aims to answer this question with a survey. If people demand an explanation one could argue this should be enforced within the GDPR, if not perhaps the focus should be elsewhere.

2.3 Human Decision Making

In order to discuss whether algorithms and especially Black-box models should be used for profiling, the alternative of human decision making also needs to be evaluated. In this section the differences will be outlined, and the disadvantages of human decision-making will be touched upon.

Shrestha et al. (2019) clearly state the differences in the decision-making process of AI versus a human being, Table 1 is based on their insights. Based on Table 1, AI-based decision making outperforms Human decision making on most of the features. An AI performs better on the size of the alternative set, the decision-making speed and the replicability of the outcomes. A human being is better when it comes to the specificity of the decision search space and the interpretability of the decision-making process and outcome. However, the explanation given by humans is sensitive to retrospective sense-making. Therefore, based on the Table 1, the AI-based decision making is generally preferred over human decision making. However, this does not mean the role of humans should be completely ignored. Shrestha et al. (2019) also propose organisational decision making where both the AI and the human are used. Kahneman also addresses the difference in decision making between humans and algorithms. Kahneman has a focus on the difference in accuracy between clinical predictions and statistical predictions. He states that statistical predictions have proven to be better. However, humans seem to have a prejudge against algorithms when the decisions are consequential. Meehl, writer of *Clinical vs. Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* and others argue that it is highly unethical to rely on intuitive judgement over algorithms when the algorithms have proven to make less mistakes. Furthermore, he points out that the fact that people care about the cause of a mistake, is one of the reasons humans dislike the algorithm alternative.The idea of not getting a job because some computer made a mistake is

Table 2.1: Differences between AI and Human based decision-making.

Decision-Making Conditions	**AI-Based Decision Making**	**Human Decision Making**
Specificity of the decision search space	Requires a well-specified decision search space with specific objective functions.	Accommodates a loosely defined decision search space.
Interpretability of the decision-making process and outcome	Complexity of the functional forms can make it difficult to interpret the decision process and outcomes.	Decisions are explainable and interpretable, though vulnerable to retrospective sense-making.
Size of the alternative set	Accommodates large alternative sets.	Limited capacity to uniformly evaluate a large alternative set.
Decision-making speed	Comparatively fast. Limited trade- off between speed and accuracy.	Comparatively slow. High trade-off between speed and accuracy.
Replicability of outcomes	Decision-making process and outcomes are highly replicable due to standard computational procedure.	Replicability is vulnerable to inter- and intra-individual factors such as differences in experience, attention, context, and emotional state of the decision maker.

Adapted source: Shrestha, Ben-Menahem, and Von Krogh (2019) more distressing than a human error (Kahneman, 2011). The interpretability of an AI and the "feelings" towards AI making decisions seem to be the biggest issue surrounding the use of AI for (high stake) decision-making. The interpretability will be further addressed in the transparency of AI section.

For fair comparison it is of importance to highlight some of the disadvantages of human decision making. Humans are prone to all kinds of biases like availability, representativeness and anchoring (Tversky & Kahneman, 1974). Furthermore, even though discrimination based on for example gender or age is forbidden, this still happens when humans make decisions. Several examples can be named in hiring where this does happen (Kelly & Roedder, 2008; Rooth, 2010; Weichselbaumer, 2003). Possible reasons are that a large number of managers still use their intuition in the hiring process (Miles & Sadler-Smith, 2014) and intergroup bias, which causes people to favour people who are like them (Hewstone, Rubin, & Willis, 2002). Similar effects of discrimination have been found for loan denial rates. A study performed by Blanchflower, Levine, and Zimmerman (2003) showed evidence that black-owned firms had higher loan denial rates even when they corrected for creditworthiness. In conclusion, the advantage of human decision making is that it is understood by other human beings. However, there is a risk of retrospective sense making which causes a disruption between the given and the real explanation. Furthermore, research has shown that human decision making is flawed in many cases, humans are prone to many biases which influence the decision-making process. These flaws in the human decision-making process should be taken into account when it is being compared with AI-based decision making.

2.4 Transparency of AI

As explained in the introduction, this study will compare the results of interpretable algorithms and Black-box models. The model framework of questions from Gunning (2017) in Table 2.2, helps determining whether something is a Black-box or not, but also in understanding the difference between the two. The main difference is in whether the internal workings of the model and the reason for specific decisions are understandable for non-expert humans which leads to being able to know the shortcomings of the model. Techniques that try to explain these Black-box models are often refeGDPRrred to as ”explainable AI (XAI)”. Other models are inherently explainable. For models that are inherently explainable the same conditions apply as for the XAI shown in Table 2.2.

Table 2.2: The difference between a Black-box and explainable AI

Black-Box	**XAI**
Why did you do that?	I understand why
Why not something else?	I understand why not
When do you succeed?	I know when you'll succeed
When do you fail?	I know when you'll fail
When can I trust you?	I know when to trust you
How do I correct an error?	I know when you erred

Adapted source: Gunning (2017).

Guidotti et al. (2018) define a Black-box model as a machine learning (ML) or data mining model of which the internal working is unknown or uninterpretable. Tree based models such as decision trees are often referred to as interpretable models. These models can give useful insights as long as the model is up for the task (Ribeiro, Singh, & Guestrin, 2016a). The interpretability and usefulness of the model also depends on the targeted dataset. Sometimes classification rules, which are very similar to decision trees, can be more comprehensive than decision trees for example (Freitas, 2014). A lack of transparency is not always a result of the complexity of the model. It can also be a result of maintaining a competitive advantage for businesses. In this case companies want their algorithms to remain secret, because competitive firms might steal it from them. This is sometimes referred to as a Black-box too, but this is not what is meant by a Black-box in this study. Reason for this is that in this case it is solely a matter of not wanting to give insights in their algorithm instead of not being able to give insights. Tutt (2017), argues a lack of transparency due to maintaining a competitive advantage could be solved by involving a third party.

Black-box models have become increasingly popular due to their high accuracy. However, it is important that applied algorithms do not only provide better accuracy but also offer a better understanding of why certain decisions are made and are more fair than the alternative of human decision making. The new GDPR rules also push companies and computer scientist to not only make their algorithms efficient, but also fair and transparent (Goodman & Flaxman, 2017). However, the meaning of transparency could be different for e.g. a developer compared to the user (Weller, 2017). Whereas the developer needs transparency in order to know if and where the model makes a mistake, the user needs it in order to trust the system. Doshi-Velez et al. (2017) focus on the accountability in AI and more precisely on the ability to explain Black-box models. Coglianese and Lehr (2016) argue that ML can be transparent and accountable even when some components of the algorithm are not available for the public. The main problem is that it is way less intuitive how a ML comes from input to output, compared to traditional data analysis methods. However, analysts are able to understand the inner workings of algorithms. Another important distinction to make is whether the internal workings need to be explained, or the reason why a certain decision was made (Guidotti et al., 2018). This distinction is referred to as global and local interpretability. Whereas in global interpretability the reasoning to get to different outcomes is completely understandable and the whole logic of the model can be explained, for local interpretability only a single decision can be made interpretable. Several solutions have been offered by researchers to increase transparency of Black-box models. Some examples are building in status updates in the model (Zhou, Khawaja, Li, Wang, & Chen, 2016), the importance of visualisations (Zhou & Chen, 2015) and LIME. LIME stands for Local Interpretable Model-Agnostics Explanations, which will help gain trust in the outcome of these models (Ribeiro, Singh, & Guestrin, 2016b). However, local interpretability still fails at providing global interpretability. Furthermore, Rudin argues that the explanation provided by ad-hoc methods such as LIME must be wrong, because if it would perfectly mimic the Black-box model, the Black-box model would not be needed. When there is uncertainty about the explanation, one cannot fully trust the explanation and therefore not fully trust the original model according to Rudin (2019). The use of post-hoc methods to explain the original model can lead to flawed conclusions and one could argue that the focus should be on using models that are inherently explainable.

Loyola-Gonzalez (2019) performed a study on the advantages and weaknesses from a practical point of view for both Black-box and White-box models. The term White-box models refers to what in this study is called interpretable models. One of his findings is that if the output is in the same form as the input data, the outcome of the Black-box model is very easy to understand. An example of this is when a mammography image is used where malignant cells are highlighted. The model will still be very hard to understand but the output data is straightforward. Furthermore, the study of Loyola-Gonzalez argues that for experts in a field it is not needed to understand the mathematics behind the models but only the output. The experts in machine learning however should understand the inner workings. Which model is deemed to be better depends on the input data. Therefore, the use of Black-box models should not be banned, but carefully selected according to Loyola-Gonzalez.

For this study the level of transparency/interpretability of the different models is of great importance. Therefore, the meaning of interpretability has to be clearly explained in the survey which will be conducted to find out how important people deem interpretability. The survey will focus on the global interpretability of the model. Reason for this is that the local interpretability cannot be fully trusted yet and therefore one cannot fully rely on the outcome of methods such as LIME.

2.5 Fairness of AI

The field of fair machine learning has become increasingly popular. The main goal of fair machine learning is to make sure that decisions based on AI or ML are equitable. A model itself is not racist or sexist, but it can inherit bias through the training data. Wilson et al. (2019) for example have proven that standard models trained on a standard data set for object detection, score better on light skin tones than on darker skin tones. Another study, that is mentioned in the introduction, shows that black people who were labeled high risk less often reoffended compared to white people rated as high risk (Angwin et al., 2016). The causes for these biases differ. One of the causes for a bias in the output of ML or AI is that it inherits the bias that is present in the training data. Another cause is that the data will fit to majority groups in the data (Chouldechova & Roth, 2018).

Corbett-Davies and Goel (2018) mention three different formal definitions that have formed concerning fairness in ML. The first one, anti-classification, is to make sure that attributes like race, gender and attributes related to this are not used to make the decision. The second definition, classification parity, is about whether the performance measures that are commonly used are equal across all groups. For example, the false positive rate should be equal for different skin colours or different genders. The final definition, calibration in the context of machine learning, means that for example defendants with a certain risk score, should have the same chance of reoffending when they are released. However, these three definitions seem to have significant statistical limitations. They can even harm the people it is supposed to protect. An example is when the protected attribute sex is removed. When female defendants have a lower risk of reoffending compared to men with a similar criminal history in a gender-neutral model there would be taste-based discrimination against women. Thus, the removal of the attribute sex, causes that women get a higher risk score than they deserve. However, the question what is deemed fair from a policy point of view is beyond the scope of this study and will therefore not be discussed in further detail. Hardt, Price, and Srebro (2016), consider non-discrimination within ML as predicting the true outcome from a set of features, while making sure that the prediction is non-discriminatory with respect to a protected feature such as gender. They speak of equalised odds when the predicted outcome and the protected attribute are independent conditional on the true outcome. This means, for example, that the percentage of correct predictions for women is equal to the percentage of correct predictions for men. This way equal bias and equal accuracy are enforced. They also propose equalised opportunities where the only the "advantaged" group, e.g. the group that gets the job, is taken into account. The formula is then as follows:

$$Pr\{\hat{Y} = 1|A = 0, Y = 1\} = Pr\{\hat{Y} = 1|A = 1, Y = 1\} \tag{2.1}$$

- Y is the observed class
- $\hat{Y}$ is the predicted class
- A is the protected attribute

In this study the models are checked for equalised opportunities. This will serve as a measure of fairness for this study. The formula compares the probability of a correct prediction for a certain class for two different groups within the protected attribute. For example, if the protected attribute is gender, the probability of a correct prediction in the advantaged group for females is compared to the probability of a correct prediction for males in the advantaged group.

3 Methodology

3.1 Data Collection

In the ideal situation, several large real data sets would be available regarding the topic 'profiling'. Especially because Black-box models are known to perform well with large amounts of data. Furthermore, the use of real data gives a good indication of how the models would perform in real world problems. In addition, it is of importance that the size of the dataset resembles the size of the dataset that would be used in real world situations. For this study Capgemini data and open source data are taken into consideration. In the following sections both the advantages and disadvantages of these options are touched upon.

Open Source Data

The first option that is reviewed for this study is open source data. A tool that can be used to find open source data is dataset search from Google. Results will often refer to popular platforms such as Kaggle. Usually the original data source is also mentioned which can then be used to check the quality of the data. The benefit of open source data is that it is available for everyone and free to use. The downside, however, is that sometimes the quality of the data is poor, and it can be hard to trace the origin of the data. Furthermore, the nature of this study requires data that is usually personal and not available to the public. This issue can be solved with the use of artificial data, but then it is uncertain how well this reflects the real world.

Capgemini Data

The second option to gather data for this study is to use data from Capgemini. Since this study is performed in collaboration with Capgemini, the possibility of using Capgemini data can be explored. However, since the nature of the data needed for this study is personal, it is challenging to get access to this data. In order to get access to for example recruitment data the HR or Staffing can be contacted. The benefit of using and collecting data from Capgemini is that the source of the data is clear and can be reached out to when things are unclear. Therefore, the quality of data is known and can be influenced. The downside however is that the nature of the data is sensitive.

This complicates the process of getting access to the data. Therefore the use of open source data is more suitable for this study and will be used.

3.2 The importance of accuracy, interpretability and fairness by the public

For the survey it is of great importance that the difference in performance of the models is understandable to its readers. The emphasis is on the accuracy, interpretability and fairness of each model. The technical details of each model are of less importance. The idea is to find out which of the attributes is deemed most important. The outcome might differ by age, gender, or their experience with data. Furthermore the context for which the model is used could be of importance. Since this study is interested in the acceptance of algorithms for automated profiling in the Netherlands, only Dutch people or people who live and/or work in the Netherlands are allowed to participate. The context of the question is focussed on problems which adults face, therefore there is an age restriction from eighteen to sixty-five year olds.

The questionnaire consists of some general questions and the main questions. The general questions are to determine the respondents age, gender and the respondents knowledge of the subject. The main questions are about the respondents preferences regarding the algorithms. Instead of asking the respondents directly whether they approve of Black-box models, several trade-offs are given. Each respondent has to decide between two algorithms, which differ on accuracy, interpretability and fairness. For each context three trade-offs are given. When a respondent chooses one of the options there are some follow-up questions to determine the decisive factor. Lastly, at the end of each context respondents are asked if they would rather have a human being performing the task.

To determine whether results are significant, for example if the results for experienced respondents are different from the non-experienced respondents, the Chi square test will be used. However, if the sample size is not big enough for the Chi square test to be reliable, the Fishers' exact test will be used. Both of these test determine whether there is a relation between categorical variables. More on the Chi square test and Fishers' exact test can be found in Agresti (2018).

3.3 Technical methods

In this section five different types of models will be discussed. Testing these five models is needed to get an accurate idea of the difference in their performance. The five different models ordered based on their interpretability are: Multiple Linear Regression (MLR), Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM) and Artificial Neural Networks (ANN). This section consists of technical and intuitive descriptions and diagnostics of the models. These methods are chosen based on their differences in complexity, and their availability of useful packages.

Diagnostics

A technique that has been used frequently for tuning the parameters of the different models is K-fold Cross Validation. In this study K-fold Cross Validation is used for every model except the MLR. For K-fold Cross Validation the training data is split up in K folds. Each run one of these folds is held out as a validation set. A model will then be trained on all folds except the held out validation set. This process is repeated K times and then the values for the parameters are set to the values that performed best. For most of the tuning purposes a 10-fold Cross Validation will be used. However, for computationally expensive tasks a 5-fold Cross Validation will be performed. This process of tuning the models is used to chose the best value for various parameters. Tuning will reduce the error on the final test set.

Classification models vs Regression models

The DT, RF and SVM are well-known classification models. The MLR and ANN are known as regression models. However, the problems these models are faced with are all classification tasks. This can be solved by introducing a decision rule. For example if there are two class labels 0 and 1, 0.5 can be used as a decision rule. Therefore, when an observation for a continuous regression model is predicted to be above 0.5 it will be classified as 1.

Multiple Linear Regression

As mentioned before, the MLR is a method that is relatively easy to interpret. The model is based on linear combinations between the response variable and the explanatory variables. The technical description and the diagnostics of the Linear Regression are based on James, Witten, Hastie, and Tibshirani (2013).

The Formula is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{3.1}$$

- Y = the response variable
- β_0 = the intercept of the linear regression
- β_p = the weight of the feature p
- X_p = explanatory of variable p
- ϵ = the error term

The goal of the Linear Regression is to find the (linear) relationship between the response variable and the features. A way of achieving this is by minimising the residual sum of squares (RSS). The residual is the difference between the prediction $\hat{y}_i$ for the ith observation and the real value of the ith observation y_i. The residual sum of squares can be calculated as follows:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.2}$$

The values of β_0, ..., β_p that minimise the RSS are used as coefficients for the Multiple Linear Regression. The interpretation of the Linear Regression then becomes relatively easy. The sign of the coefficient tells whether the effect is positive or negative and the size of the coefficient determines the magnitude of the effect. Interaction effects will not be included to maintain the linear relationship which keeps the interpretation simple.

To tune the Linear Regression, backwards selection is used. Backward selection starts with including all the variables and then removes the variable with the highest p-value, meaning that it is the least significant variable. This process continues until some stopping rule is reached. For this study the optimal model was chosen based on the highest adjusted R-squared, meaning that the model that explains the largest part of the variance. Backward selection was preferred over Forward Selection because Forward selection is a greedy approach. Furthermore, Backward selection was preferred over Mixed-selection, which is a combination of forwards and backwards selection, because this is computationally very expensive when the number of variables becomes large.

Decision Trees

The second method that is used is the DT. The reason that a DT is very intuitive, is that they consist of a set of ”if then” statements. An example is if person A has a Data Science profile and has more than 4 years of experience in consultancy then he would be a good fit. Whereas if the answer to one of those questions had been no, it would not have been a good fit (Song & Ying, 2015). Another reason to include DT is that they are widely used. The C4.5 and CART for example, which are a widely used DT methods, were voted to be among the top ten most influential algorithms (Wu et al., 2008). A DT is called ”greedy” because of the top-down approach it uses. The package that will be used for the DT is called rpart (Therneau & Atkinson, 2019).

A DT consists of a root node, internal nodes and leaf nodes. To form these nodes a DT uses recursive partitioning of the data based on independent variables. This means that the data is split in smaller subsets by the use of feature values. The root node consists of the entire data set and keeps splitting into internal nodes until it arrives at a leaf node where no further splits are made. The splits of the nodes can be decided upon using various methods. Two of the best-known methods are Gini impurity of entropy information gain. In this study Gini impurity will be used, because the focus is on classification models. If the DT becomes very large it is likely to overfit. Therefore, pruning should be applied. For this study cost complexity pruning is used, which is a parameter chosen by cross validation. The value of the cost complexity parameter is chosen to minimise the average error.

Gini Index can be calculated as follows:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{3.3}$$

- k stands for the class
- $\hat{p}_{mk}$ stands for the proportion of training observations in the mth region from class k

The Gini Index is a measure of node purity, Gini takes on a small value if all $\hat{p}_{mk}$'s are close to either one or zero. A small value means that the node is pure, which indicates the node contains mostly observations from a certain class. Therefore, this method is used to determine the best possible split. The best split is the split that creates the most pure nodes.

Random Forest

A RF builds on the idea of the DT. Whereas the DT is still relatively easy to interpret this becomes harder for the RF, especially when the number of trees becomes large. The idea of the RF is that a large number of decision trees are generated and each of those trees gets a vote in the final decision. Each tree is generated on a bootstrap sample from the original data. From the original data random observations are selected to create a new dataset. The sampling for the new dataset is performed with replacement, meaning that observations can be included multiple times in the bootstrap sample. The RF is generated as follows (Liaw & Wiener, 2002a):

1. Draw bootstrap samples from the original dataset equal to the amount of n_{tree}
2. For each bootstrap sample build a tree without pruning it. However, now only a random subset of the features (m_{try}) is used for classification
3. perform the classification based on majority of votes across all generated trees

To determine the optimal numbers of n_{tree} and m_{try} K-fold cross validation can be performed. This process should be performed multiple times because both values cannot be optimised simultaneously. For the RF the package ”randomForest” will be used (Liaw & Wiener, 2002b).

When the n_{tree} becomes large it is difficult to determine which features caused a certain decision. A way to determine the feature with the highest impact is variable importance. Variable importance is based on the mean decrease in accuracy. The mean decrease in accuracy is computed by recording the Out-of-Bag error. The Out-of-Bag error is computed on a sample from the data that was not used for constructing the tree. The Out-of-Bag error is then compared with the error after permuting each predictor variable. The results are averaged over all trees and normalised by dividing it with the standard deviation. The size of the difference in accuracy determines the importance of the variable. However, it is important to mention that variable importance does not offer the same information as models that are inherently interpretable. For example, it is unknown whether the effect is negative or positive. Furthermore, the variable importance can be biased due to the different scales of variables.

Support Vector Machines

The fourth model used for this study is SVM. SVM can be seen as a black box due to the high dimensionality of the model. The SVM is explained based on Cortes and Vapnik (1995); De Brabanter, De Moor, Suykens, Van Gestel, and Vandewalle (2002); James et al. (2013). SVM builds on the relatively simple and interpretable maximal margin classifier. The package that will be used for SVM is called ”e1071” (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2019a). The maximal Margin classifier uses a hyperplane to split between two classes. The idea is to have a margin as big as possible between the two classes. A hyperplane is a flat affine subspace of dimension $p-1$ in a p-dimensional space. This means that in a two dimensional space the hyperplane is a one-dimensional subspace (a line). In a three dimensional space it becomes a plane. In more than three dimensions, visualisation of hyperplanes becomes hard however. The Formula for a hyperplane is:

$$w^T x + b = 0 \tag{3.4}$$

This means that any combination of $(x_1, x_2, ...x_p)$ that satisfies the condition of (4.4) lies on the hyperplane. Every combination that is smaller or bigger than 0 lies above or below the hyperplane.

When the two classes are linear separable, the hyperplane can function as a very natural classifier. The margin M between the two classes should be maximised.

However, when the two classes are not linear separable, one can fix this problem by allowing for some misclassifications, known as a soft margin. This has several benefits, such as greater robustness to individual observations.

The position of the hyperplane is based on a maximal margin. Therefore, given the training data the distance between the hyperplane and the closest observation of each class has to be maximised. The positive class will be labeled +1 and the negative class labeled -1. Therefore if the class is +1:

$$w^T x + b \geq 1 \tag{3.5}$$

- w is a normal vector perpendicular to the hyperplane
- x is the input vector
- b is the bias term

if the class is -1:

$$w^T x + b \leq -1 \tag{3.6}$$

Mathematically maximising the distance between the two classes, subject to equation 3.4, 3.5 and 3.6, means maximising $2/||w||$ which is the same as:

$$min \frac{1}{2} ||w||^2 \tag{3.7}$$

One can expand on this by allowing misclassifications:

$$min \frac{1}{2} ||w||^2 + C \sum_{i=1}^{N} \xi_i \tag{3.8}$$

- ξ_i allows for misclassification, but penalises them
- C is a trade-off between the width of the margin and misclassifications

C Functions as a trade-off. When C is equal to 0, one does not allow for misclassifications. When C increases the tolerance towards misclassifications increases. These support vector classifiers work well when the data is linear separable. However, when the data is not linear separable like non linear separable case in Figure 3.1 the support vector classifier will classify very poorly. The SVM will be able to deal with this much better.

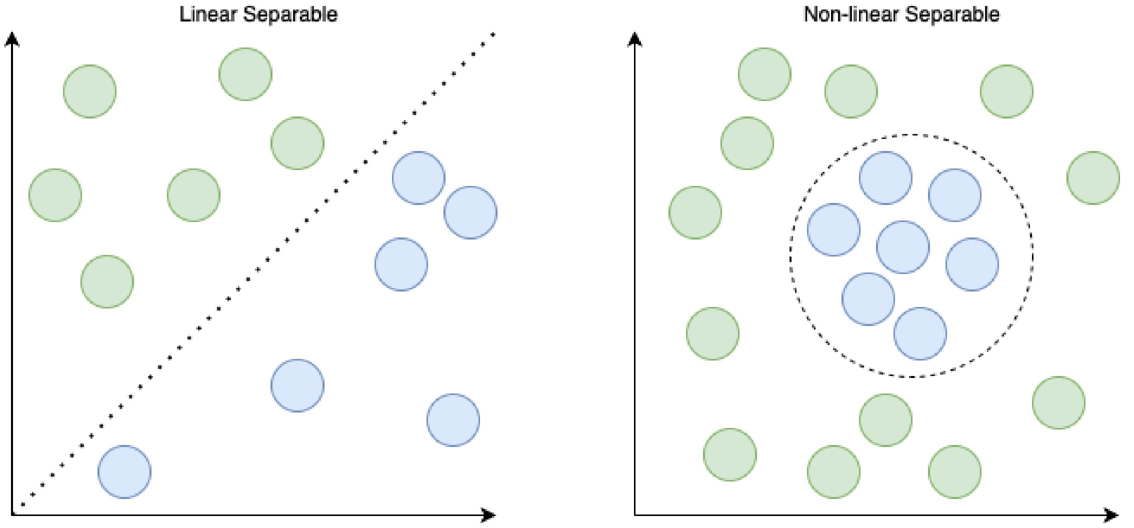

Figure 3.1: The Linear Separable Case vs Non-linear separable case

SVM can use non-linear boundaries to classify observations. SVM does this by enlarging the feature space and kernels. The kernels create an enlarged feature space in which the decision boundary is actually linear, however in the original data the data is not linear separable. Misclassifications can still happen, however the SVM might perform better after applying one of the kernels. Therefore, this is computationally an efficient way of dealing with this problem. To determine which kernel should be applied, cross-validation can be used.

The kernels use the inner product of observations represented by:

$$K(x_i, x_{i'}) \tag{3.9}$$

Where K is a function, referred to as kernel, which can have multiple forms for SVM. The kernel quantifies the similarity of two observations. Only the support vectors determine the class of an observation the other vectors are not taking into account.

The linear kernel, which is the support vector classifier:

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij}x_{i'j} \tag{3.10}$$

If a non-linear kernel is used SVM is introduced, three commonly used non-linear kernels are the polynomial kernel, the radial kernel and the sigmoid kernel which take the following forms (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2019b):

Polynomial kernel:

$$K(x_i, x_{i'}) = (\gamma \sum_{j=1}^{p} x_{ij}x_{i'j} + \alpha)^d \tag{3.11}$$

Radial kernel:

$$K(x_i, x_{i'}) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2) \tag{3.12}$$

sigmoid:

$$tanh(\gamma \sum_{j=1}^{p} x_{ij}x_{i'j} + \alpha) \tag{3.13}$$

The kernels have different parameters which should be tuned to get the optimal result. The tuning of these parameters will be performed by cross validation. The parameter C is not part of a specific kernel but should always be tuned to determine the tolerance towards misclassifications. All kernels except the linear kernel use γ. A large γ reduces the variance and increases the bias and a small gamma leads to high variance and small bias. The degree d determines de degree of the polynomial kernel. Furthermore, the polynomial and sigmoid kernel use the parameter α, which is an independent term.

Due to the high dimensionality of the model SVM it cannot be visualised in a human interpretable way, especially the non-linear kernels.

Artificial Neural Networks

An ANN is inspired by the way biological neural systems deal with processing data (Jain, Mao, & Mohiuddin, 1996). The ANN is considered to be a Black-box, due to the fact that the inner process is very hard for humans to follow. The initial input is usually transformed multiple times, making the interpretations for humans very hard. For this study ANN are used as a supervised learning method, meaning that the model is trained on a labeled dataset. The package that will be used is the ”neuralnet” package (Fritsch, Guenther, & Wright, 2019). An ANN learns by iteratively adapting its parameters to fit the labeled data as well as possible. Neural networks consist of an input layer, an output layer and hidden layers in between. Those layers consist of neurons and these neurons are connected with synapses (Günther & Fritsch, 2010). The formula for an ANN with one hidden layer is as follows:

$$o(x) = f(w_0 + \sum_{j=1}^{J} w_j * f(w_{0j} + \boldsymbol{w_j}^T \boldsymbol{x})) \tag{3.14}$$

Where:

- $o(x)$ is the calculated output for the given inputs and current weights
- f is the calculated output of the output neuron and hidden neurons
- w_0 is the intercept of the output neuron
- w_{0j} is the intercept of the jth hidden neuron
- w_j is the weight of the synapse starting at the jth hidden neuron going to the output neuron
- $\boldsymbol{w_j}$ is the vector of all synapses leading to the hidden neurons
- $\boldsymbol{x}$ is the vector of all the features

If the neural network performs well, the predicted output $o(x)$ will be close to the observed output y.

Furthermore, ANN's use an activation function which determines which neurons should be activated and which should not be activated. The neuralnet package in R, which is used for this study, uses the same activation function for all neurons. For this study the logistic function is used since this works well for binary output. The formula for a logistic output function is as follows:

$$f(u) = \frac{1}{1 + e^{-u}} \tag{3.15}$$

The amount of layers and nodes will be chosen based on K-fold cross validation. The neuralnet package in R allows for a maximum of three layers to be cross validated.

4 Data

In this section the datasets that are used to answer the research question are discussed. For this study three different data sets related to profiling will be used to answer the research question. Reason to analyse multiple datasets is to check for stability among the results. Since the exact details of each dataset are less important for this study, there will only be a brief discussion of each dataset in this section. For more information on the datasets there is a link to the source of each dataset in Appendix B.

Student Performance

The Student Performance dataset is about the performance of secondary school students. The data was originally collected to predict the performance of students using data mining by Cortez and Silva (2008). The explanatory variables are mostly categorical (27) and a small part is numerical (3). The explanatory variables have information about the school, the students' living situation and personal information about the student. For this study the independent variable is whether the student passes the class. The students who got an insufficient grade (below 10) fail the class and the students who passed got a 10 or higher. The complete dataset has 395 students of which 164 failed the class and 231 passed.

Credit Card Clients

The Credit Card dataset can be used to predict which clients will default on their loan. The dataset was originally collected by Yeh and Lien (2009). The dataset consists of 30.000 observations from clients in Taiwan from April 2005 to September 2005. The data consists of the independent variable wether someone defaulted on their loan and 21 explanatory variables. The explanatory variables are a mix of fourteen numeric variables and nine categorical variables. The data has personal information about the client, and the loan of the client. Out of the 30.000 observations, 23.364 did not default on their loan and 6.636 did.

Bank Marketing

This dataset is focused on predicting the success of telemarketing for a bank. The data was originally collected by Moro, Cortez, and Rita (2014). Our approach slightly differs because a different performance measure is used. In this study accuracy is used to determine the performance of a model. For this study a subset of 4119 randomly selected observations was used that was provided to test more computationally expensive methods. The data consists of twenty explanatory variables and one dependent variable, whether the client has subscribed to a term deposit. The independent variables are a mix of ten categorical variables and ten numeric variables. The data has information about the client, the current campaign, previous campaigns and the economic context. Out of the 4119 clients 3668 did not subscribe a term deposit and 451 clients did subscribe to a term deposit.

5 Results

In the results section the results of the different models are discussed and explained. These results were needed to get accurate information on these models' performances for the survey. On each data set five different algorithms were trained for classification purposes. The models differ on interpretability and accuracy and have also been checked on discrimination.The results and interpretation of each model are discussed separately for each dataset. Furthermore, an overview is presented which summarises the results of the different models. Additional information on the tuning of the models can be found in Appendix C. The results of the survey are presented after the results of the models.

5.1 Student Performance

The first data set that is reviewed is the student dataset regarding the mathematic grades. The different algorithms were used to predict whether the final grade of the students would be sufficient (10 or higher on a scale of 20) or not. The students who failed the class were labeled as 0 and the students who pass the class are labeled as 1. For the fairness metric the group that is labeled as 0, meaning that they will fail the math class, was studied. The goal is to find out whether the model works better for a certain gender in predicting whether someone will fail the math class.

Multiple Linear Regression

The final model for MLR achieved an accuracy of 65,55% using only 18 predictor variables, which is shown in Table 5.1.

Table 5.1: Confusion Matrix of the Multiple Linear Regression on the Student Performance dataset.

	Reference 0	**Reference 1**
Prediction 0	22	18
Prediction 1	23	56

The two most important variables seem to be failures and schoolsup, based on significance and size. Failures stands for the amount of failures on a continuous numerical scale from 0 to 3 and then 4 includes 4 failures and everything above. The feature schoolsup represents whether someone received extra support, yes or no. The interpretation of the features failures and schoolsup is as follows. For failures the coefficient is -0.2146 meaning that when failures goes up by one for a student ceteris paribus, the prediction goes down with 0.2146 making the student less likely to pass. The interpretation of schoolsup is slightly different because it is a categorical variable. The coefficient for schoolsup yes is -0.3417, meaning that someone with extra school support is less likely to get a sufficient grade compared to the base level. The base level in this case is someone who does not get extra school support, once again taking ceteris paribus into account. Intuitively these results make sense. The full list of coefficients can be seen in Appendix C.

The algorithm had also been checked on fairness based on gender. The MLR worked better for men compared to women. The proportion of men that was predicted correctly was 1.5 times bigger.

Decision Tree

The second algorithm that has been used to classify the students is the DT. The DT achieved an accuracy of 63,87% after pruning, which can be seen in Table 5.2. The DT can be seen in Figure 5.1.

Table 5.2: Confusion Matrix of the Decision Tree on the Student Performance dataset.

	Reference 0	**Reference 1**
Prediction 0	25	23
Prediction 1	20	51

The results of a DT are very intuitive and therefore easy to interpret for humans. Furthermore, the importance of certain features can also be obtained relatively easy since the splits on the top have a bigger impact than the splits further down. Therefore, the same two variables seem to be of great importance for the Decision Tree as for the Linear Regression.

The DT is slightly more fair based on the correct classifications for men and women. Whereas for the Linear Regression the proportion of men predicted correctly was 1.5 times bigger for Decision Tree the correctly predicted proportion of men was only 1.15 times bigger.

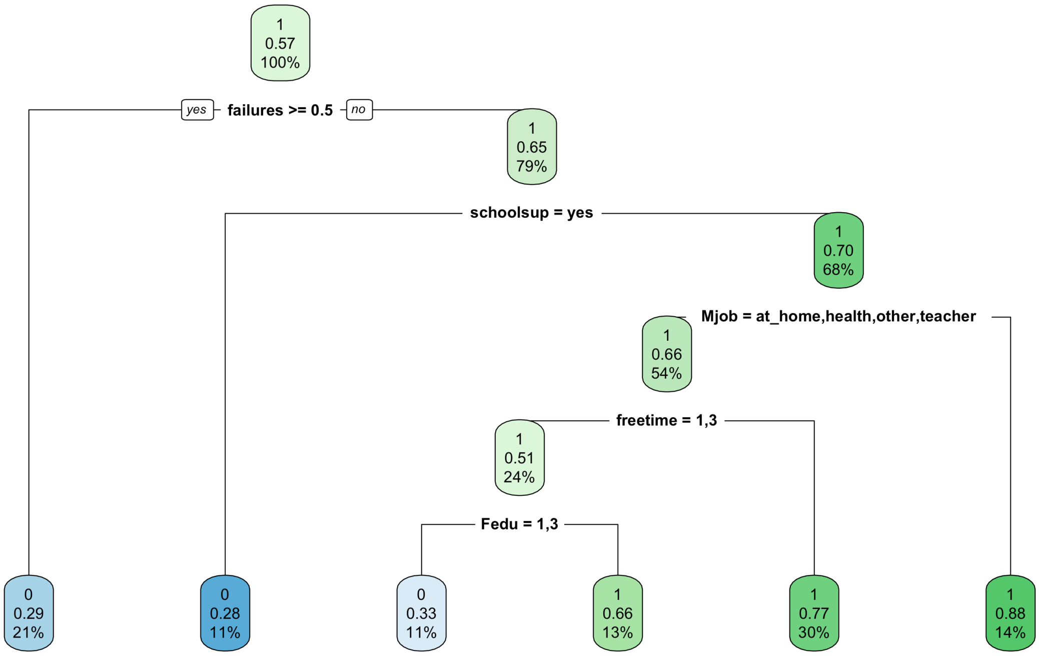

Figure 5.1: Decision Tree Student Performance

Random Forest

After running a 10-fold cross validation for RF the tuning parameters were set at 650 for n_{tree} and 15 for m_{try}. These values led to an accuracy of 71,43% which can be seen in Table 5.3.

Table 5.3: Confusion Matrix of the Random Forest on the Student Performance dataset.

	Reference 0	**Reference 1**
Prediction 0	25	14
Prediction 1	20	60

The interpretation of the RF is a lot harder than the interpretation of the DT. Whereas the DT clearly shows the path it takes to make a certain decision this is more complicated for the RF. This is because of the large number (650) of trees it uses to make a decision. However, as said in the methodology the variable importance can be assessed by computing the decrease in accuracy when a variable is permuted. The results of the top 10 most important features are shown in Figure 5.2. Based on this information one can say that once again the features failures and schoolsup are most important. One could also inspect the individual trees grown by the RF to see how the model makes decisions. Therefore, the model is still somewhat interpretable.

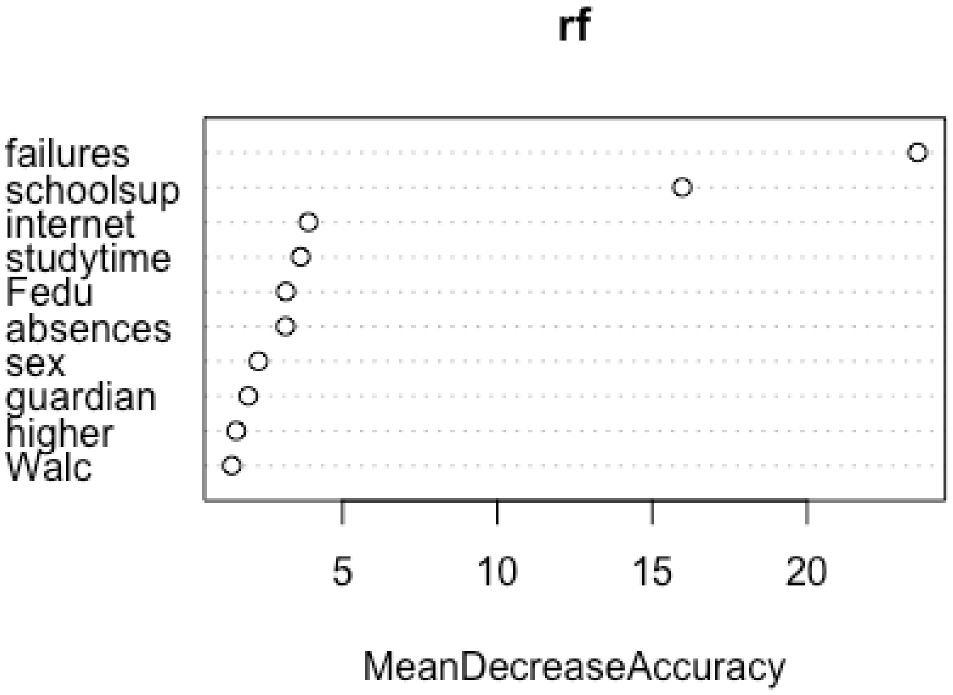

Figure 5.2: Variable importance of the Random Forest on Student Performance dataset

On the fairness criteria the RF performs slightly worse than the DT. The proportion of male students that is correctly predicted to fail is 1.25 times bigger than the proportion of women.

Support Vector Machines

For SVM the sigmoid kernel turned out to be the best for this binary classification problem based on cross validation. To determine the values for γ, α and C a 10-fold cross-validation was used. The best value for γ was 0.076, -1 for α and 10 for C. After training the SVM on the training set, the SVM achieved an accuracy of 68,91$\%$ on the test set which can be seen in Table 5.4.

Since the sigmoid kernel has been used, the transparency of the model is very low making it hard to interpret. The SVM has a relatively high accuracy, however the

Table 5.4: Confusion Matrix of the Support Vector Machines on the Student Performance dataset.

	Reference 0	**Reference 1**
Prediction 0	18	10
Prediction 1	27	64

model works much better for men than for women. For the group that is predicted to fail, the proportion of male students classified correctly is 1.96 times bigger than the proportion of women.

Artificial Neural Networks

The final model for the student dataset is the ANN. A 5-fold cross validation was performed to estimate the best number of hidden nodes in the three layers. The results of the cross-validation can be seen in Figure 5.3. The horizontal axis represents the number of nodes in the first layer, the colour represents the number of nodes in the second layer and the four different graphs represent the four different options for the number of nodes in the third layer.

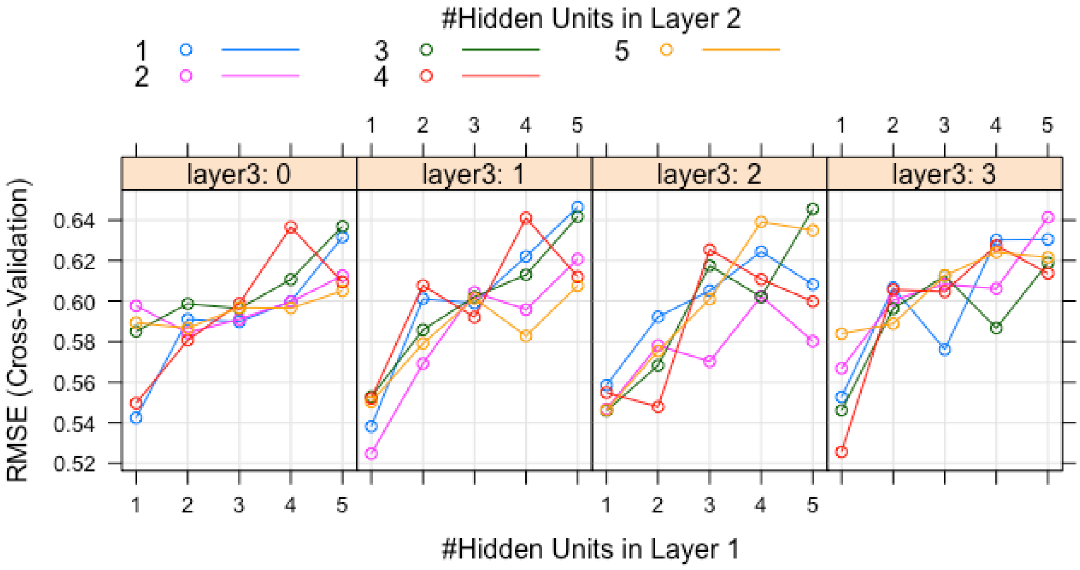

Figure 5.3: Cross validation for the Artificial Neural Network on the Student Performance dataset

The final ANN achieved an accuracy of 70,59% on the test set ,which can bee seen in Table 5.5. This is very similar to the RF which has the highest accuracy (71,43%) for this dataset.

Table 5.5: Confusion Matrix of the Artificial Neural Network for the Student Performance dataset

	Reference 0	**Reference 1**
Prediction 0	20	10
Prediction 1	25	64

Like the SVM the ANN functions as a black box, making interpretation very hard. However, in contrast to the previously mentioned model the ANN actually works relatively well for women. The proportion of women predicted correctly was 1.49 times higher than for men. This is remarkable since all other models performed better for male students. Therefore, the ANN had likely used other features to reach a decision.

5.2 Credit Card Default

The Second dataset was concerned with predicting whether people will default on their loan. Within the dataset people who defaulted on their loan were labeled with a 1 and the people who did not default on their loan were labeled with a 0. For the fairness check people who were predicted to default on their loan were investigated. The goal was, again, to check whether a model works better for a certain gender.

Multiple Linear Regression

The MLR has an accuracy of 79,68% after performing backwards selection. The final model used only sixteen of the available variables. The confusion matrix can be seen in Table 5.6. For the Linear model coefficient with a negative sign mean that it reduces the chance of defaulting. Coefficients with a positive sign increase the chance of someone defaulting on their loan. Two variables that seem to have a relatively big impact are PAY 0 with a coefficient of 0.09 and PAY 2 with a coefficient of 0.02. Both PAY 0 and PAY 2 are continuous features and the features tells you how many months this specific payment is delayed. Therefore, an increase in the delay of these payments increases the chances of a default on someones loan. The full list of coefficients can be seen in Appendix C.

Table 5.6: Confusion Matrix of the Multiple Linear Regression on the Credit Card Default dataset.

	Reference 0	**Reference 1**
Prediction 0	6892	1727
Prediction 1	102	279

The fairness check has shown that the proportion of male students that was correctly predicted to default was 1.16 times bigger than the proportion of women. Therefore women were more often classified to default incorrectly.

Decision Tree

The DT did better than the Linear Regression with an accuracy of 81,77%. The confusion matrix can be seen in Table 5.7

Table 5.7: Confusion Matrix of the Decision Tree on the Credit Card Default dataset.

	Reference 0	**Reference 1**
Prediction 0	6669	1312
Prediction 1	325	694

What might also be of importance for the comparison of these models is that even though an increase of only two percent points might not seem like much the amount of correctly predicted defaults more than doubled. As can be seen in Table 5.7 the group that was correctly classified to default was much bigger compared to the linear regression.

Due to the intuitive structure of the DT it can be seen that the three most important variables seemed to be Pay 0, Pay 2 and Bill AMT1 as can be seen in 5.4. The DT also performed much better on fairness with the proportion of men being classified correctly to default only 1.04 times bigger than the proportion of women.

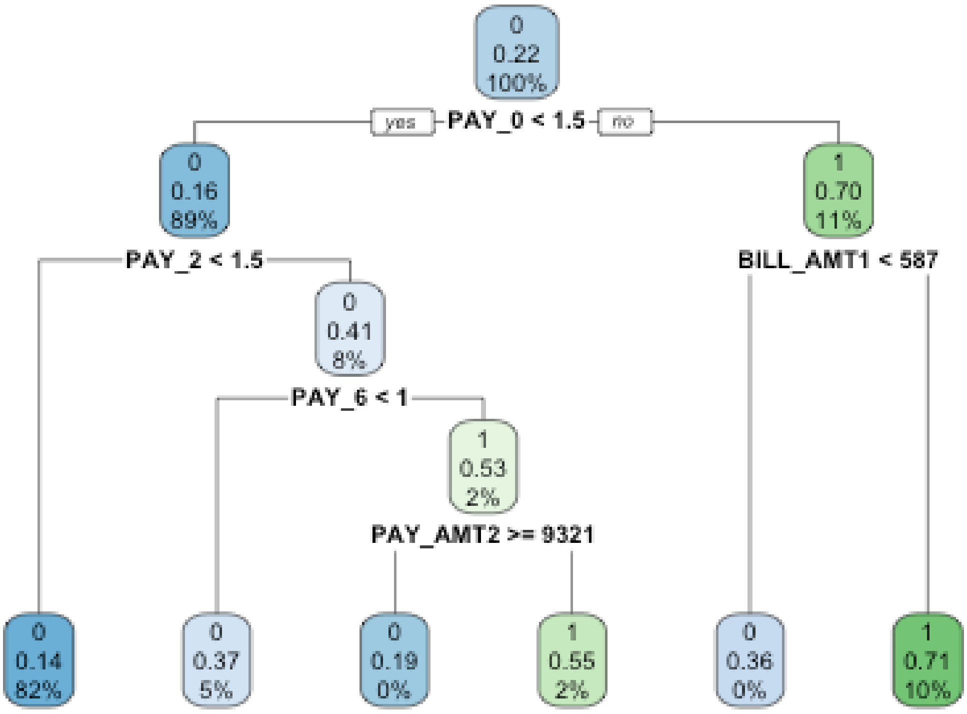

Figure 5.4: Decision Tree for the Credit Card Defaults

Random Forest

The RF outperformed the other models based on accuracy for the student dataset. For the Credit Card default dataset the differences in percent points are a lot smaller. After performing a 5-fold Cross Validation m_{try} was set at 4 and n_{tree} at 450. The RF achieved an accuracy of 81,40%, which can be seen in Table 5.8.

Table 5.8: Confusion Matrix of the Random Forest on the Credit Card Default dataset.

	Reference 0	**Reference 1**
Prediction 0	6608	1288
Prediction 1	386	718

Once again the number of correctly predicted defaults is much higher than the amount for MLR. The results of the top 10 most important features are shown in

Figure 5.5. The three most important variables based on the importance measure are PAY 0, PAY 2 and BILL AMT3.

The performance on fairness is very similar to the performance of the DT. The proportion of men classified correctly is 1.06 times bigger than the proportion of women. That the performance is similar to the performance of the DT makes sense. The RF is a large collection of DT and for both models the same features seem to be important.

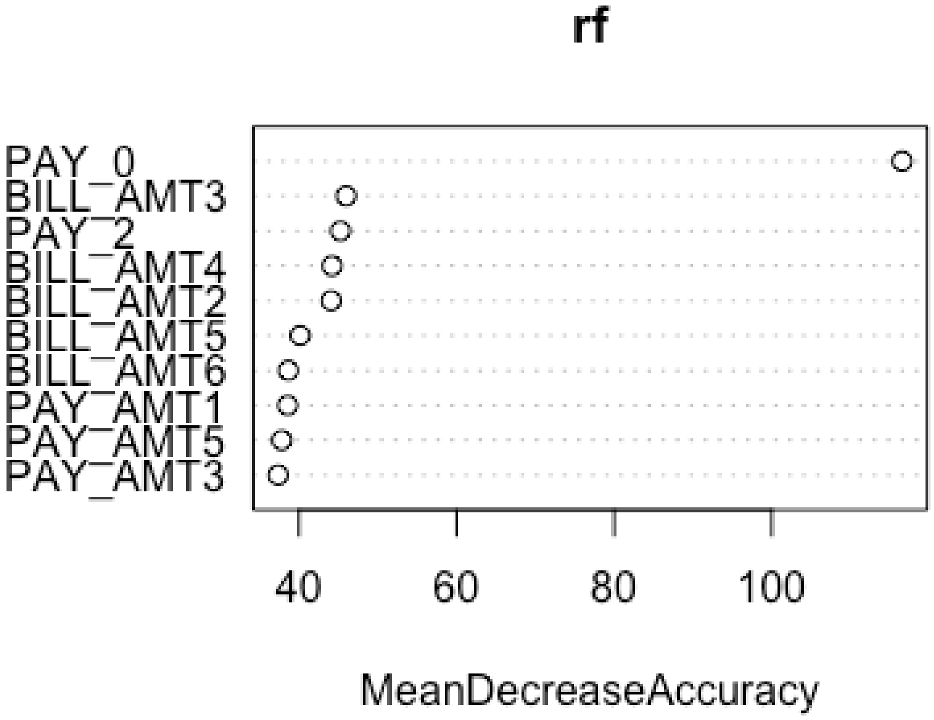

Figure 5.5: Variable importance of the Random Forest on the Credit Card Default dataset

Support Vector Machines

The SVM is the model with the second-highest performance for this dataset. With the radial kernel being the best choice for this classification problem, the SVM achieved an accuracy of 81.47% as can be seen in Table 5.9. Based on a five-fold cross validation the best value for γ was 0.04 and the best value for C was 10.

Table 5.9: Confusion Matrix of the Support Vector Machines on the Credit Card Default dataset.

	Reference 0	**Reference 1**
Prediction 0	6646	1320
Prediction 1	348	686

On the fairness criterium the SVM scores relatively low. With the proportion of men being classified correctly 1.12 times bigger than the proportion of women, only the linear regression performs worse.

Neural Network

The ANN has been trained based on trial and error instead of K-fold cross validation due to the high number of observations. The ANN with the lowest error, on the training set had three hidden layers with six, three and one hidden nodes. On the test set this ANN achieved an accuracy of 80,87% of which the results can be seen in Table 5.10

Table 5.10: Confusion Matrix of the Artificial Neural Network on the Credit Card Default dataset.

	Reference 0	**Reference 1**
Prediction 0	6505	1288
Prediction 1	489	773

Even though the ANN was outperformed by every model except the MLR based on accuracy, the ANN had the highest number of correctly classified defaults. In practice this can be an incentive to still pick the ANN over the other classification models. On the fairness criterium the ANN scored slightly worse then the tree based models but better than the SVM and the Linear Regression. The proportion of men classified correctly is 1.09 times bigger than the proportion of women.

5.3 Bank Marketing

The final dataset consisted of data of an E-commerce campaign from a bank. The goal was to predict whether customers will subscribe to a term deposit yes (1) or no

(0). For the fairness check gender was not available therefore marital status was used. The group that was checked for fairness was the group that had been classified to be a success.

Multiple Linear Regression

The MLR achieved an accuracy of 89,21%. This result was achieved after selecting the variables through backwards selection, resulting in a models with eleven variables. The Confusion matrix can be seen in Table 5.11.

Table 5.11: Confusion Matrix of the Multiple Linear Regression on the Bank Marketing dataset.

	Reference 0	**Reference 1**
Prediction 0	806	81
Prediction 1	19	21

For the MLR the features about the economic context seem to be of great importance. The consumer price index has a positive coefficient of 0.27 and the quarterly average of the total number of employed citizens has a negative coefficient of -0.17 both variables are continuous. A categorical variable with a relatively big impact is the type of contact. If people are contacted on their landline phone the coefficient is -0.13 compared to the base level, cellular phone. The fairness check showed that the model worked better for singles than married people. The proportion of singles classified correctly was 1.25 times bigger than the proportion of married people. The full list of coefficients can be seen in Appendix C.

Decision Tree

The DT did slightly worse than the Linear Regression with an accuracy of 88,67% as can be seen in Table 5.12.

Table 5.12: Confusion Matrix of the Decision Tree on the Bank Marketing dataset.

	Reference 0	**Reference 1**
Prediction 0	794	74
Prediction 1	31	28

The DT can be seen in Figure 5.6. The most important features for the

DT were the quarterly average of the total number of employed citizens, the number of days that passed since the previous contact and the type of contact. Thus, the important features overlap with the MLR a lot.

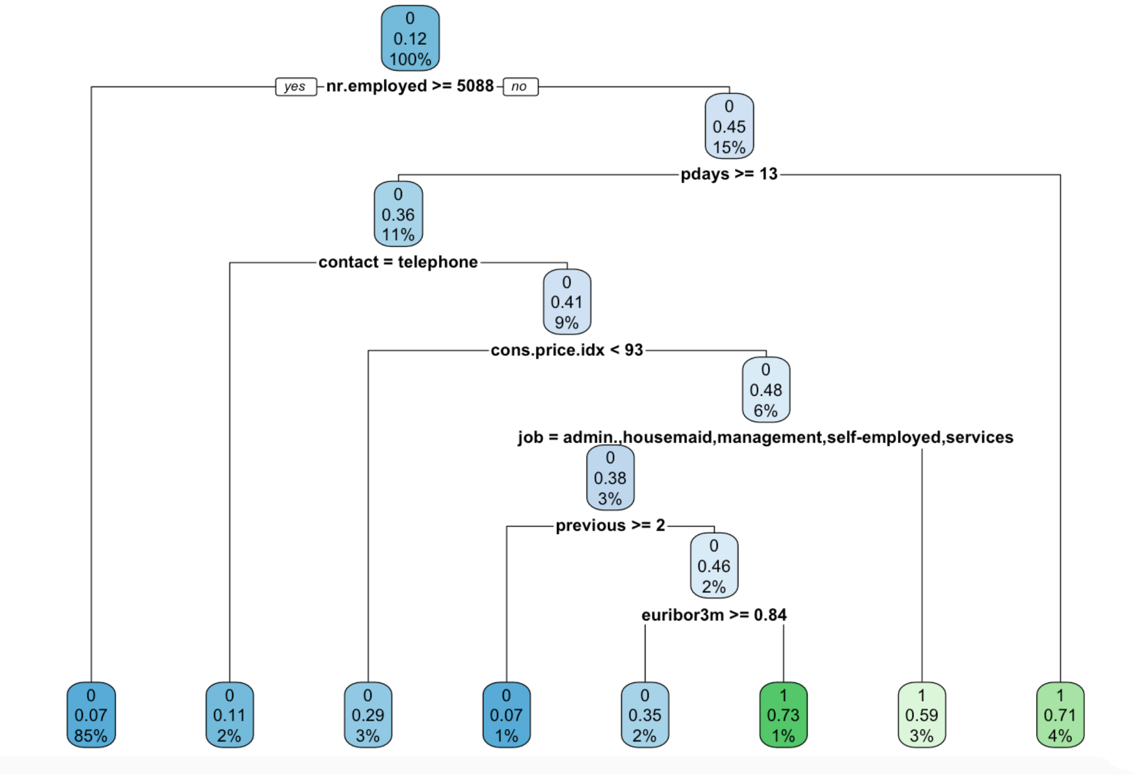

Figure 5.6: Decision Tree for Bank Marketing

The DT was slightly more fair than the MLR. The model worked slightly better for singles than for married people, with the proportion of correctly classified singles being 1.05 times bigger.

Random Forest

The RF was tuned by using a 5-fold Cross Validation, setting the value for m_{try} at 2 and the n_{tree} at 200. The performance of the RF was slightly higher than the performance of the DT with 89,21%. This increase in accuracy is thanks to a better performance for the group that is predicted to be unsuccessful as can be seen in Table 5.13.

The difference in performance compared to the DT can be due to the features that are used. Based on Figure 5.7, the Euribor three month rate seems to be of great importance, which is less important for the DT. The fairness check showed

Table 5.13: Confusion Matrix of the Random Forest on the Bank Marketing dataset.

	Reference 0	**Reference 1**
Prediction 0	805	80
Prediction 1	20	22

comparable results for the RF as the MLR. The proportion of correctly classified singles was 1.25 times bigger than the proportion of correctly classified married people.

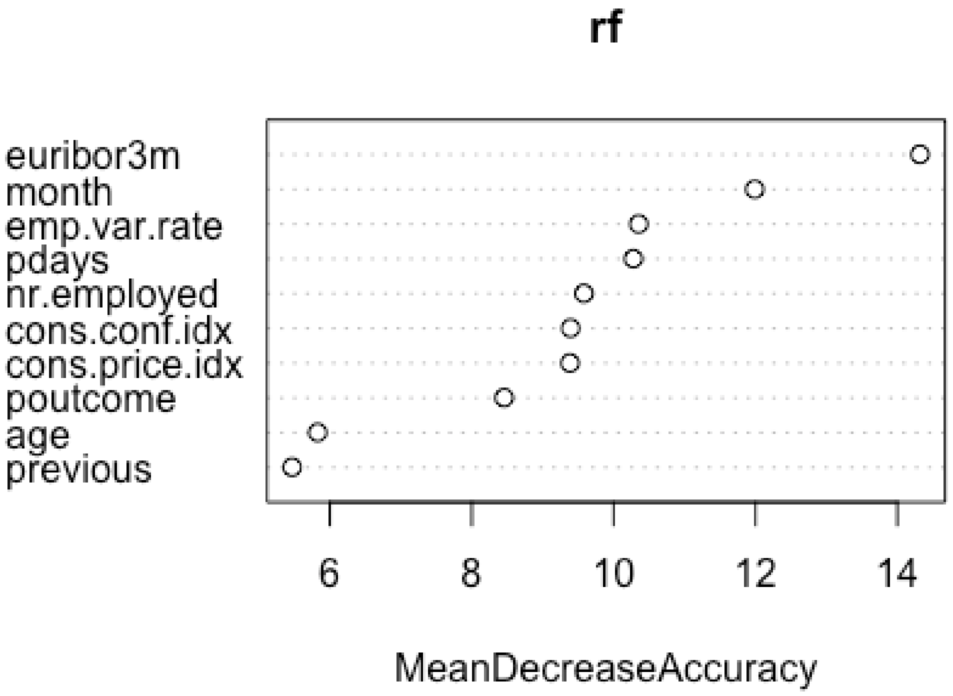

Figure 5.7: Variable importance of the Random Forest on Bank Marketing dataset

Support Vector Machines

The SVM achieved the highest accuracy for the Bank marketing dataset with an accuracy of 89,97%. The polynomial kernel was the best choice for this dataset and a 10-fold Cross Validation was used to determine the best value for the hyper parameters. The best values were 0.015 for γ, 1 for C , 0 for α and 3 for d. The Confusion Matrix for the SVM can be seen in Table 5.14.

However, as can be seen in the Confusion Matrix the SVM performs espe-

Table 5.14: Confusion Matrix of the Support Vector Machines on the Bank Marketing dataset.

	Reference 0	**Reference 1**
Prediction 0	816	84
Prediction 1	9	18

cially well on the group that is not targeted. Due to the use of the polynomial kernel, the importance of certain features remains unknown. The SVM scored similar to the DT based on fairness. The proportion of correctly classified singles was 1.07 times bigger than the correctly classified proportion of married people.

Neural Network

The ANN achieved the lowest accuracy for the Bank marketing dataset with an accuracy of 86,41% as can be seen in Table 5.15. A 5-fold Cross Validation was performed to determine the optimal amount of hidden nodes within the different layers. Based on the Cross Validation which can be seen in Figure 5.8 two hidden layers with each one node was chosen.

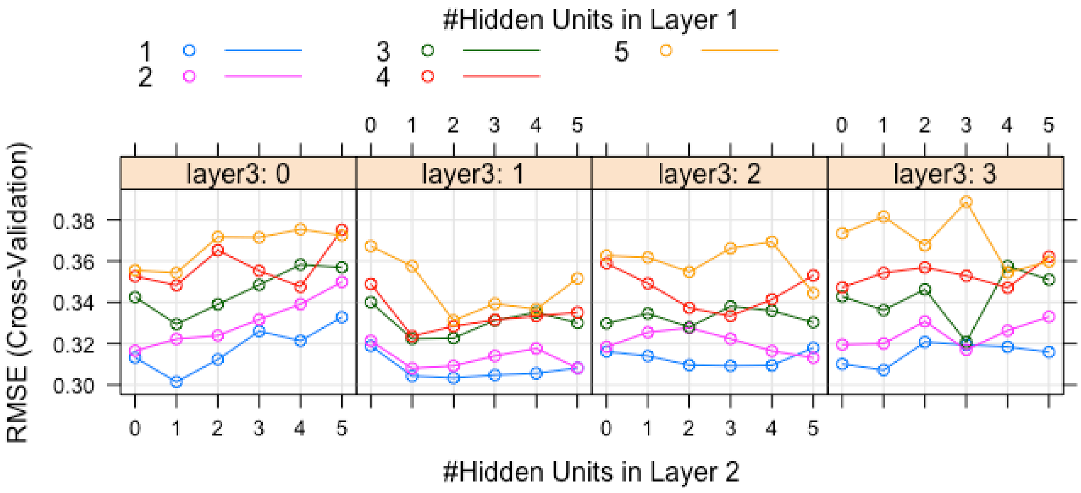

Figure 5.8: Cross Validation of the Artificial Neural Network for the Bank Marketing dataset

However, as can be seen in Table 5.15 the relatively low accuracy is due to wrong predictions in the unsuccessful group. The correct predictions of the group that should be targeted are as high as the DT as for the ANN. Therefore, the ANN might still be valuable. Furthermore, the fairness check showed that the proportion of correctly classified married people was only 1.02 times bigger than the proportion of correctly classified singles. Therefore the ANN scores best on fairness.

Table 5.15: Confusion Matrix of the Artificial Neural Networks on the Bank Marketing dataset.

	Reference 0	**Reference 1**
Prediction 0	790	74
Prediction 1	35	28

5.4 Model performance overview

To summarise the results of the various models on each dataset, each aspect (accuracy, interpretability and fairness) will be reviewed.

Accuracy

As can be seen in Table 5.16, there was not one model that was always the worst or best model in terms of accuracy. If the best model has to be picked based on accuracy it would be SVM. The SVM had the best performance on the Bank marketing dataset, the second best model for the credit-card dataset and the third best model for the students dataset. However, it would be best to take multiple models into consideration for each classification problem based on the type of data.

Table 5.16: Summary of the accuracy results

Dataset	**Linear Regression**	**Decision Tree**	**Random Forest**	**SVM**	**ANN**
Students	65,55%	63,87%	**71,43%**	68,91%	70,59%
Creditcard	79,68%	**81,77%**	81,40%	81,47%	80,87%
Bank marketing	89,21%	88,67%	89,21%	**89,97**	88,24%

Interpretability

The interpretability of each model is determined based on how difficult is is to extract information out of the model on the classification process. For the classification problems discussed in this study the Decision Tree proved to be very interpretable. The tree structure is very intuitive and is interpretable even for people without knowledge about machine learning algorithms. The Linear Regression is also a very interpretable solution to the classification problems. The interpretation needs little explanation to make it comprehensive. Due the fact that each coefficient has a sign and an amplitude it is easy to get insights into the decision-making process. The Random Forest becomes a little more complicated. The Random Forest has the same tree based structure as the Decision Tree. However, due to the large number of trees the interpretation becomes a lot harder. The SVM and ANN were the two models with the lowest interpretability. For these two models it is impossible to extract global interpretability of the model. There are ways to get information on local interpretability off the model. However, as explained in the transparency section, models such as LIME cannot be fully trusted.

Fairness

Table 5.17: Summary of the fairness results

Dataset	**Linear Regression**	**Decision Tree**	**Random Forest**	**SVM**	**ANN**
Students	Low	**Very high**	High	Very Low	Average
Creditcard	Very Low	**Very high**	High	Low	Average
Bank marketing	Low	Low	High	Average	**Very high**

The results for the fairness check are summarised in Table 5.17. The fairness check that was conducted on gender for the Students and Credit-card dataset and on marital status for the Bank marketing dataset should be interpreted accordingly. Firstly, the results are for this specific dataset, parameters, feature selection etc. The results do not tell anything about how fair, for example, a DT is in general. However, it can still be a useful measure to check if a model is biased towards certain characteristics.

Especially when the so called Black-box models are used it can be a very useful tool to check whether the model works better for a certain gender or marital status for example. That the results differ for the different datasets can be due to the fact that for each dataset the model is trained and it picks features to use for its prediction.

Verdict

Based on these results there is not one model that always dominated the others. Therefore, it is important to keep in mind that the performance of a model is dependent on the type of data. In some cases a Linear Regression might work very well if the relationship between the dependent and explanatory variables is linear. However, in some cases it might be better to use more complicated algorithms that are better at dealing with non-linear relationships. Therefore, from an overall performance point of view, each of these models should be considered. Based on the problem and the type of data and the importance of interpretability a decision can then be made. Instead of enforcing a right to an explanation perhaps a set of rules on how to deal with data and how to choose a model would be better. In some cases the benefits of a Black-box model might outweigh the costs and in some cases it might not. The Black-box models sometimes had the highest accuracy or performed best for the specific group that was studied. Not giving Black-box models a chance would be a loss.The ethically correct answer on which model should be used will be derived from the publics' opinion.

5.5 The publics' opinion

To determine whether the public approves algorithms and especially Black-box models for profiling, a questionnaire was conducted. Instead of directly asking the respondents if they approve of the use of Black-box models for profiling, they were asked to choose the model they prefer. The Questionnaire got a total of 87 responses.

Firstly, some descriptive statistics of the data will be presented. Based on the descriptive statistics shown in Table 5.18, the decision was made to merge the age groups of 26-35 and 36-45 and to merge the age groups 46-55 and 56-65.

Table 5.18: Descriptives of the Questionnaire

Age	**Numbers**			**Row percentages**		
	Total	Male	Female	Total	Male	Female
Total	87	47	40	100	54.02	45.98
18-25	51	25	26	100	49.02	50.98
26-35	14	7	7	100	50.00	50.00
36-45	9	6	3	100	66.67	33.33
46-55	7	4	3	100	57.14	42.86
56-65	6	5	1	100	83.33	16.67

Algorithms to predict performance

The first context that was presented in the survey is when algorithms are used to predict performance, e.g. whether a student will fail or pass, or if employees will perform well at their job.

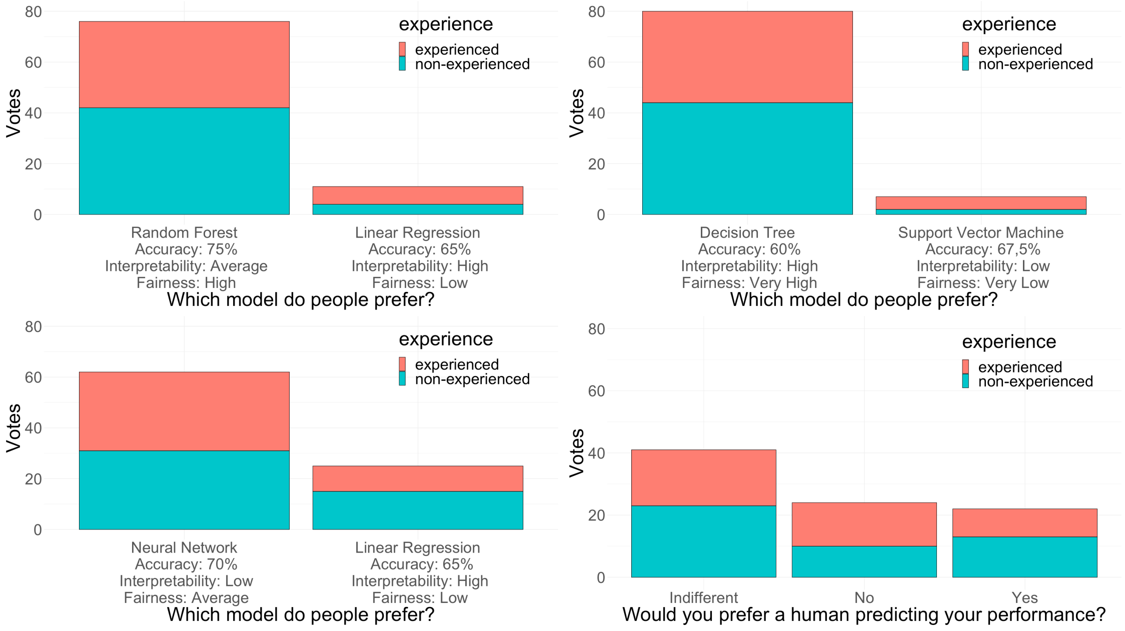

Figure 5.9: Algorithms to predict performance

The most important results for the performance context are shown in Figure 5.9. What can be seen is that every time the model that outperforms the other model on two features won. Furthermore fairness is always higher for the model that has won. However, if the respondents were asked why they chose for the model that performed best on two features, a majority said it was due to the combination of the two features this can be seen in Appendix D. Therefore most respondents did not just focus on one of the features, but the overall performance of the model seems to be decisive. This is also confirmed by the fact that the majority of people that chose the RF over the MLR, would change their vote if the performance of the linear regression would increase. However, it is important to mention that the increase needed to change their vote differs across respondents. The respondents who chose the DT over SVM were asked if their response would change if the difference in accuracy was bigger, is also interesting. The results show that a very small majority would indeed change their response, however, for most of them the difference in accuracy would need to be at least 20 or 30% this can be seen in Appendix D. Therefore, if one of the models really is far more accurate than the other, only then are people more willing to accept a lower fairness and interpretability level. Finally, the respondents were asked if they prefer a human being predicting their performance. The results show "indifferent" (47.13%) is the most popular option and the remainder of people are almost equally distributed over "yes" (25.29%) and(27,59%) "no". From respondents who would rather have a human being predict their performance, the majority (54.55%) said this was due to the fact that they believe humans can give a more interpretable explanation.

Algorithms to predict loan default risk

The second context that was presented was what if algorithms are used to predict whether someone gets a loan or not. This decision could be based on whether the bank expects you to default on your loan. The differences in performances are smaller in percent points compared to the performance example. However, people were reminded that when the numbers of observations is high a 2% change can mean an increase over a hundred correct predictions.

The most important results for the results of the loan acceptance context are shown in Figure 5.10. Once again in the two graphs at the top the model that outperforms the other on two features get most of the votes. Both winning cases have in common that fairness is higher compared to the losing algorithm. However, just as in the performance context, the majority of respondents says that the combination of two features being higher was decisive this can be seen in Appendix D. Furthermore, in the trade-off between the MLR and the SVM people indicated that their

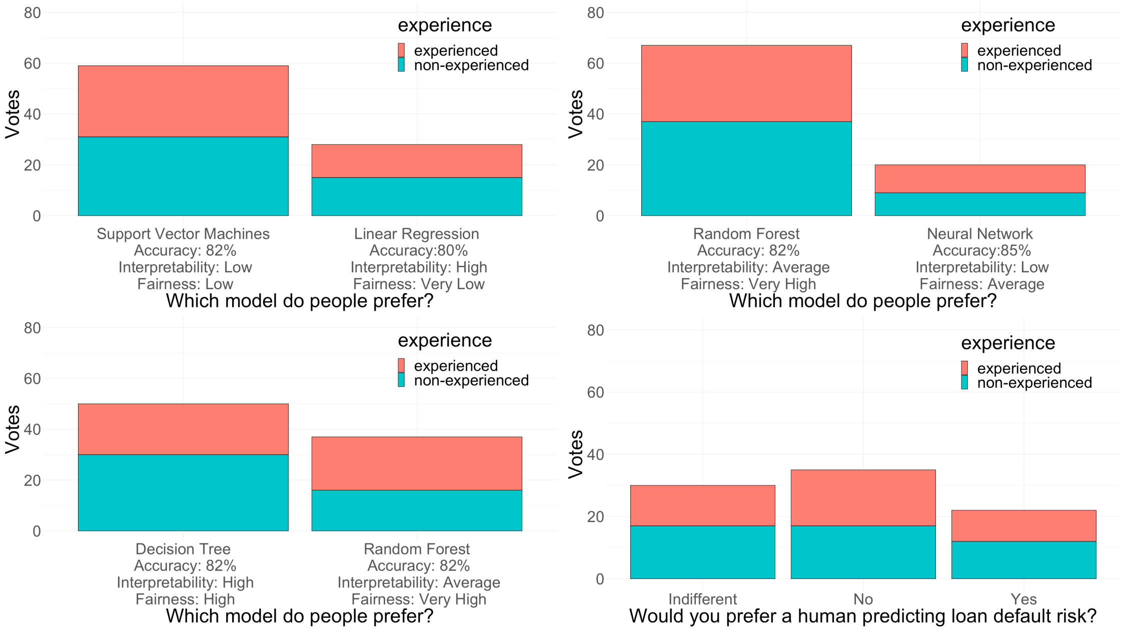

Figure 5.10: Algorithms to predict loan default risk

choice would change if the accuracy of MLR was higher. The respondents that chose RF over the ANN were asked if their answer would change if the difference in accuracy was bigger. Once again the majority of respondents would change their answer, however the amplitude of the change needed differs among the respondents this can be seen in Appendix D. At the bottom left of Figure 5.10 there is an example where the accuracy of both models is equal. This creates a trade-off solely between interpretability and fairness and in this example experienced people seem to favour the higher fairness and non-experienced people favour a higher interpretability. Furthermore, the respondents were asked if they would rather have a human being predicting their loan default risk. The results now show that the ”no” becomes the most popular answer (40.02%) , ”yes” gets the same number of votes as in the performance example (25.29%) and the number of votes for ”indifferent” shrunk (33.48%). Like in the previous context, a majority (59.09%) of the people that would rather have a human being predict whether they get a loan or not, says this is due to the fact that they believe humans give a more interpretable explanation.

Algorithms to predict who to target

The final context that will be discussed is when algorithms are used to predict who to target for certain products. Targeting them is based on their expectancy to buy the product. The most important results are shown in Figure 5.11.

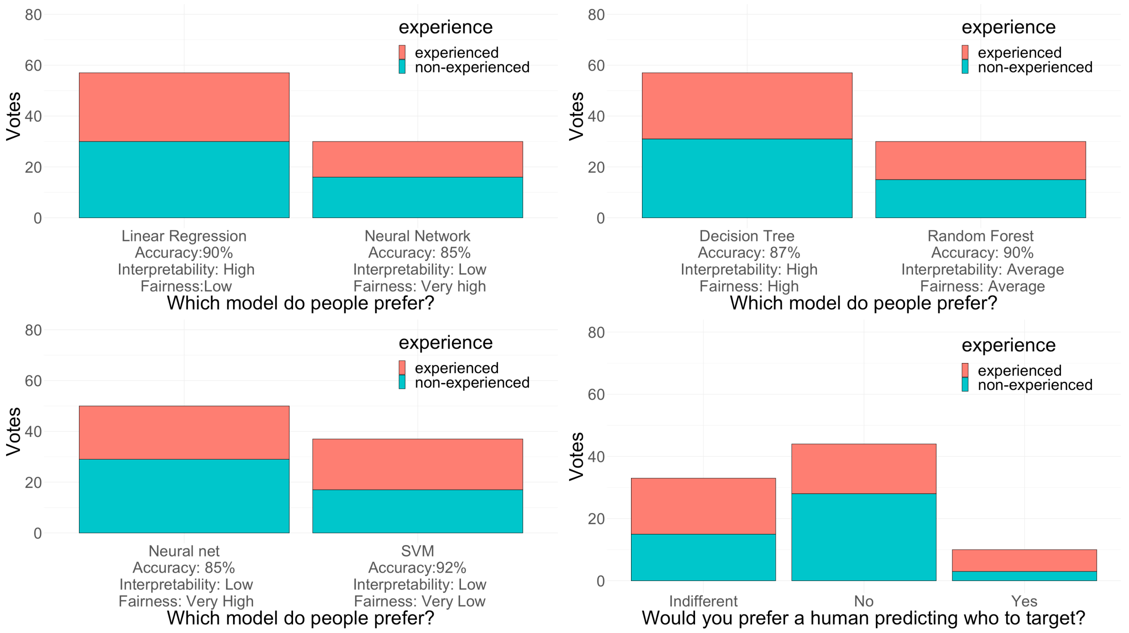

Figure 5.11: Algorithms to predict who to target

In the two graphs shown at the top of Figure 5.11 once again the algorithm that outperformed the other algorithm on two features was chosen. This is also the case for the graph shown in the bottom left, however this is mostly due to the non-experienced group. The results for the experienced group were very close with 51.22% choosing the ANN and 48.78% choosing the SVM. A distinction with the two previous contexts however is that in the trade-off between the MLR and the ANN most people said this was due to the higher accuracy this can be seen in Appendix D. In the two preceding contexts most people chose for a model because two features were higher. Furthermore, respondents indicated that their choice between RF and DT would change if the difference in accuracy was bigger this can be seen in Appendix D. In this example the majority (50.57%) chose ”no” when they were asked if they would rather have a human being determining who to target. From the remaining respondents 37.93% chose ”indifferent” and only 11.49% chose ”yes”.

The publics opinion

For each context the overall performance of the model seems to be decisive. Respondents always chose the model that outperformed the other models on two out of the three features. For all the choices between two models there were no significant differences between what the experienced respondents chose versus the non-experienced respondents. Furthermore, the difference between experienced and non-experienced respondents in whether people would rather having human decision-making was also not significant. The context does have a significant effect on a 5% significance level on whether people prefer human decision-making.

6 Conclusion

This study aimed to determine to what extent people in the Netherlands approve the use of algorithms and especially of Black-box models for profiling. Furthermore this study questions if the answer to this question differs based on experience and context. To answer this question, research has been conducted to determine the differences in the results of Black-box models and more interpretable models. To determine these differences three different classification problems were solved using the same five models for each problem. The models were assessed on their accuracy, interpretability and fairness on a protected attribute. The empirical results for these models indicated that it really depends on the data which model performs the best. Therefore, the suggestion is to always try multiple models and use the one that performs best. Naturally, a model with a high level of accuracy, interpretability and fairness would be the ideal option. Therefore, scientist should still aim at making Black-box models interpretable or making interpretable models perform well on accuracy. However, the aim of this study is to determine what should be done if there is not one model outperforming the other models on all three of these features. The results of the empirical study on the performance of the different models has been used as a base for the questionnaire.

The results of the questionnaire answer the research question: *"To what extent does the public in the Netherlands approve the use of Black-box models for profiling?"*. The questionnaire has shown that people do not always choose for the algorithm with the highest accuracy, interpretability or fairness. In fact, the majority of the respondents seemed to asses the overall performance of the algorithms and then made a decision. This meant that in every example people preferred the option that scored better than the alternative on two out of the three features. Therefore, people seem to accept a lower interpretability if this means a higher accuracy and fairness. Thus, there is strong evidence to support that Hypothesis 1 is true. Furthermore, if the difference in accuracy becomes very large, respondents even accepted a lower fairness and interpretability. The second research question: *"What are the differences in the amount of trust people tend to have in Black-box models based on experience and context?"* can be answered. At a 5% significance differences were detected in the context. Whereas in the performance context only 27.59% of the respondents would rather have an algorithm making the decision than a human being, this became 40.04% for the loan acceptance and 50.57% for the targeting of certain products. This is in line with what is described in Kahneman (2011), that there is a prejudge against algorithms when the decision is consequential. Therefore, there is strong evidence that Hypothesis 2 is also true. Reason to prefer human decision making, was usually due to the belief that humans can give a more interpretable explanation. Lastly, there were no significant differences in the choices experienced respondents made compared to the non-experienced respondents. Therefore, there seems to be no evidence for the third hypothesis to be true.

Based on these results one could suggest to look critically at the right to an explanation as formulated in the GDPR. Instead, a set of rules can be put in place to assure an ethical process regarding the use of algorithms and especially Black-boxes for profiling as suggested by Haenlein and Kaplan (2019). For example, like Holm (2019) suggests, if a Black-box model provides the best solution it should be used. However, what is deemed to be the best solution should be clearly stated. The questionnaire has shown that people are willing to accept a lower degree of one of the features if this is compensated by the two other features. Therefore, if a Black-box model outperforms an interpretable model on accuracy and has shown to be more fair towards the protected attributes that play a role, it should be used. The context of the decision-making process should also still be taken into account. The extent to which the models needs to be explainable depends on the how serious the consequences of the decision are. Furthermore, as shown by Loyola-Gonzalez (2019), if the output is in the same format as the input Black-box models are very easy to interpret. More precisely, the outcome of the model is easy to interpret while the mathematics are not. However, this is not of importance to the people who have to interpret the outcome. In conclusion, instead of always having the right to an explanation when automated decision-making is used for profiling, one could look more into specific cases.

Discussion

The results and the resulting conclusion are based on a relatively small sample which consisted predominantly of people between the age of 18 and 25. For further studies a bigger sample size would be suggested to have a better look at the results and to confirm whether they represent the publics' opinion. In addition, it would be interesting to compare these results to other cultures since this may yield different answers. Furthermore, it would be interesting to look into the suggested set of rules that could be implemented in the GDPR. Possible questions to study what this set of rules should look like, and how to enforce them. Finally, the search of algorithms with high accuracy and interpretability should continue. The ideal situation remains an algorithm that is inherently transparent and provides the desired accuracy. If these algorithms then outperform humans on accuracy and fairness, while still giving interpretable results, people might become less hostile towards the use of algorithms for high stake decision making.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Acces*, *6*, 52138–52160.

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

Alder, G. S., & Gilbert, J. (2006). Achieving ethics and fairness in hiring: Going beyond the law. *Journal of Business Ethics*, *68*(4), 449–464.

Alexander, L., & Moore, M. (2016). Deontological ethics. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/.

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016). *Machine bias.* Retrieved 2020-05-04, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Blanchflower, D. G., Levine, P. B., & Zimmerman, D. J. (2003). Discrimination in the small-business credit market. *Review of Economics and Statistics*, *85*(4), 930–943.

Bodie, M., Cherry, M., McCormick, M., & Tang, J. (2017). The law and policy of people analytics. *U. Colo. L. Rev*, *88*, 961.

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *preprint arXiv:1810.08810*.

Coglianese, C., & Lehr, D. (2016). Regulating by robot: Administrative decision making in the machine-learning era. *Geo. LJ*, *105*, 1147.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023.*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

De Brabanter, J., De Moor, B., Suykens, J. A., Van Gestel, T., & Vandewalle, J. P. (2002). *Least squares support vector machines*. World scientific.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., … Wood, A. (2017). Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134.*.

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM*

SIGKDD explorations newsletter, *15*(1), 1–10.

Fritsch, S., Guenther, F., & Wright, M. N. (2019). neuralnet: Training of neural networks [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=neuralnet` (R package version 1.44.2)

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

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1–42.

Gulliford, F., & Dixon, A. P. (2019). Ai: the hr revolution. *Strategic HR Review*.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2.

Günther, F., & Fritsch, S. (2010). neuralnet: Training of neural networks. *The R journal*, *2*(1), 30–38.

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4), 5–14.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 3315–3323.

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, 105–120.

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual review of psychology*, *53*(1), 575–604.

Holm, E. A. (2019). In defense of the black box. *Science*, *364*(6435), 26–27.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer*, *29*(3), 31–44.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kaplan, A., & Haenlein, M. (2019). Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25.

Kelly, D., & Roedder, E. (2008). Racial cognition and the ethics of implicit bias.

Philosophy Compass, *3*(3), 522–540.

Key, S., & Popkin, S. (1998). Integrating ethics into the strategic management process: doing well by doing good. *Management Decision*.

Liaw, A., & Wiener, M. (2002a). Classification and regression by randomforest. *R news*, *2*(3), 18–22.

Liaw, A., & Wiener, M. (2002b). Classification and regression by randomforest. *R News*, *2*(3), 18-22. Retrieved from `https://CRAN.R-project.org/doc/Rnews/`

Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, *7*, 154096–154113.

Mantelero, A. (2018). Ai and big data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*, *34*(4), 754–772.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019a). e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=e1071` (R package version 1.7-3)

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019b). e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=e1071` (R package version 1.7-3)

Miles, A., & Sadler-Smith, E. (2014). "with recruitment i always feel i need to listen to my gut": The role of intuition in employee selection. *Personnel Review*.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

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

Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, *17*(3), 523–534.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Shrestha, Y. R., Ben-Menahem, S. M., & Von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, *61*(4), 66–83.

Sinnott-Armstrong, W. (2019). Consequentialism. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2019 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/sum2019/entries/consequentialism/`.

Song, Y.-Y., & Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.

Therneau, T., & Atkinson, B. (2019). rpart: Recursive partitioning and regression trees [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=rpart` (R package version 4.1-15)

Tutt, A. (2017). An fda for algorithms. *Admin. L. Rev.*, *69*(1), 83–124.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Van den Heuvel, S., & Bondarouk, T. (2017). The rise (and fall?) of hr analytics. *Journal of Organizational Effectiveness: People and Performance*.

Wachter, M. B. . F. L., S. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99.

Weichselbaumer, D. (2003). Sexual orientation discrimination in hiring. *Labour Economics*, *10*(6), 629–642.

Weller, A. (2017). Challenges for transparency. *arXiv preprint arXiv:1708.01870*.

Wilson, H. J., & Daugherty, P. R. (2018). Collaborative intelligence: humans and ai are joining forces. *Harvard Business Review*, *96*(4), 114–123.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1–37.

Yeh, I.-C., & Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473–2480.

Zhou, J., & Chen, F. (2015). Making machine learning useable. international journal of intelligent systems technologies and applications. *International Journal of Intelligent Systems Technologies and Applications*, *14*(2), 91–109.

Zhou, J., Khawaja, M., Li, S. J., Z., Wang, Y., & Chen, F. (2016). 'making machine learning useable by revealing internal states update – a transparent approach'. *Int. J. Computational Science and Engineering*, *13*(4), 378–389.

A Definitions

Definitions from the GDPR, article 4:

'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;

'processing' means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;

B Dataset Links

Links to Datasets:

`https://archive.ics.uci.edu/ml/datasets/Student+Performance`

`https://archive.ics.uci.edu/ml/datasets/Bank+Marketing`

`https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`

C Model Results

Table C.1: MLR output for the Student Performance dataset

	Coefficients	
(Intercept)	0.89^{*}	(0.44)
sexM	0.11	(0.07)
famsizeLE3	0.10	(0.06)
Fjobhealth	0.22	(0.18)
Fjobother	0.07	(0.13)
Fjobservices	0.08	(0.14)
Fjobteacher	0.26	(0.18)
Fedu1	−0.37	(0.33)
Fedu2	−0.27	(0.33)
Fedu3	−0.32	(0.33)
Fedu4	−0.21	(0.34)
Mjobhealth	−0.02	(0.13)
Mjobother	−0.08	(0.09)
Mjobservices	0.05	(0.10)
Mjobteacher	-0.34^{**}	(0.12)
reasonhome	0.06	(0.07)
reasonother	−0.08	(0.10)
reasonreputation	0.05	(0.08)
traveltime2	−0.06	(0.06)
traveltime3	0.15	(0.13)
traveltime4	−0.14	(0.20)
studytime2	0.08	(0.07)
studytime3	0.24^{*}	(0.10)
studytime4	0.27^{*}	(0.13)
failures	-0.21^{***}	(0.04)
schoolsupyes	-0.34^{***}	(0.08)
famsupyes	-0.17^{**}	(0.06)
internetyes	0.09	(0.08)
romanticyes	−0.08	(0.06)
famrel2	0.04	(0.24)
famrel3	−0.09	(0.20)
famrel4	0.02	(0.20)
famrel5	−0.08	(0.20)
freetime2	0.24	(0.14)
freetime3	0.10	(0.13)
freetime4	0.24	(0.14)
freetime5	0.35^{*}	(0.16)
goout2	0.00	(0.12)
goout3	−0.07	(0.13)
goout4	−0.19	(0.13)
goout5	−0.18	(0.14)
Dalc2	0.04	(0.08)
Dalc3	−0.06	(0.12)
Dalc4	−0.23	(0.20)
Dalc5	0.22	(0.26)
health2	−0.21	(0.11)
health3	−0.16	(0.10)
health4	−0.18	(0.11)
health5	−0.16	(0.09)
R^2	0.37	
Adj. R^2	0.24	
Num. obs.	276	

Note: Standard errors are in parentheses; $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table C.2: MLR output for the Credit Card Default dataset

	Coefficients	
(Intercept)	0.13^{***}	(0.03)
LIMIT_BAL	-0.00^{**}	(0.00)
SEXFemale	-0.01^{*}	(0.01)
EDUCATIONGraduate	0.12^{***}	(0.02)
EDUCATIONUniversity	0.11^{***}	(0.02)
EDUCATIONHigh school	0.10^{***}	(0.02)
MARRIAGEMarried	0.02	(0.02)
MARRIAGESingle	−0.01	(0.02)
AGE	0.00^{**}	(0.00)
PAY_0	0.09^{***}	(0.00)
PAY_2	0.02^{***}	(0.00)
PAY_3	0.01^{*}	(0.00)
PAY_4	0.01	(0.00)
PAY_5	0.01^{*}	(0.00)
BILL_AMT1	-0.00^{***}	(0.00)
BILL_AMT2	0.00	(0.00)
PAY_AMT1	-0.00^{***}	(0.00)
PAY_AMT2	−0.00	(0.00)
PAY_AMT4	−0.00	(0.00)
PAY_AMT5	−0.00	(0.00)
R^2	0.13	
Adj. R^2	0.12	
Num. obs.	21000	

Note: Standard errors are in parentheses; $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table C.3: MLR output for the Bank Marketing dataset

	Coefficients	
(Intercept)	-24.83^{***}	(3.67)
jobblue-collar	-0.02	(0.02)
jobentrepreneur	-0.05	(0.04)
jobhousemaid	0.01	(0.04)
jobmanagement	-0.02	(0.02)
jobretired	0.03	(0.04)
jobself-employed	-0.03	(0.03)
jobservices	-0.03	(0.03)
jobstudent	-0.02	(0.05)
jobtechnician	0.01	(0.02)
jobunemployed	0.02	(0.04)
educationbasic.6y	0.05	(0.04)
educationbasic.9y	0.02	(0.03)
educationhigh.school	0.02	(0.03)
educationilliterate	-0.28	(0.29)
educationprofessional.course	0.01	(0.03)
educationuniversity.degree	0.02	(0.03)
contacttelephone	-0.13^{***}	(0.02)
monthaug	0.03	(0.05)
monthdec	0.18^{*}	(0.09)
monthjul	-0.04	(0.04)
monthjun	-0.05	(0.04)
monthmar	0.34^{***}	(0.06)
monthmay	-0.04	(0.03)
monthnov	-0.06	(0.04)
monthoct	-0.01	(0.06)
monthsep	-0.05	(0.06)
campaign	-0.00	(0.00)
pdays	-0.00	(0.00)
poutcomenonexistent	0.05^{*}	(0.02)
poutcomesuccess	0.21^{*}	(0.09)
emp.var.rate	-0.17^{***}	(0.04)
cons.conf.idx	0.01^{**}	(0.00)
cons.price.idx	0.27^{***}	(0.04)
euribor3m	0.05	(0.03)
R^2	0.26	
Adj. R^2	0.24	
Num. obs.	2163	

Note: Standard errors are in parentheses; $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

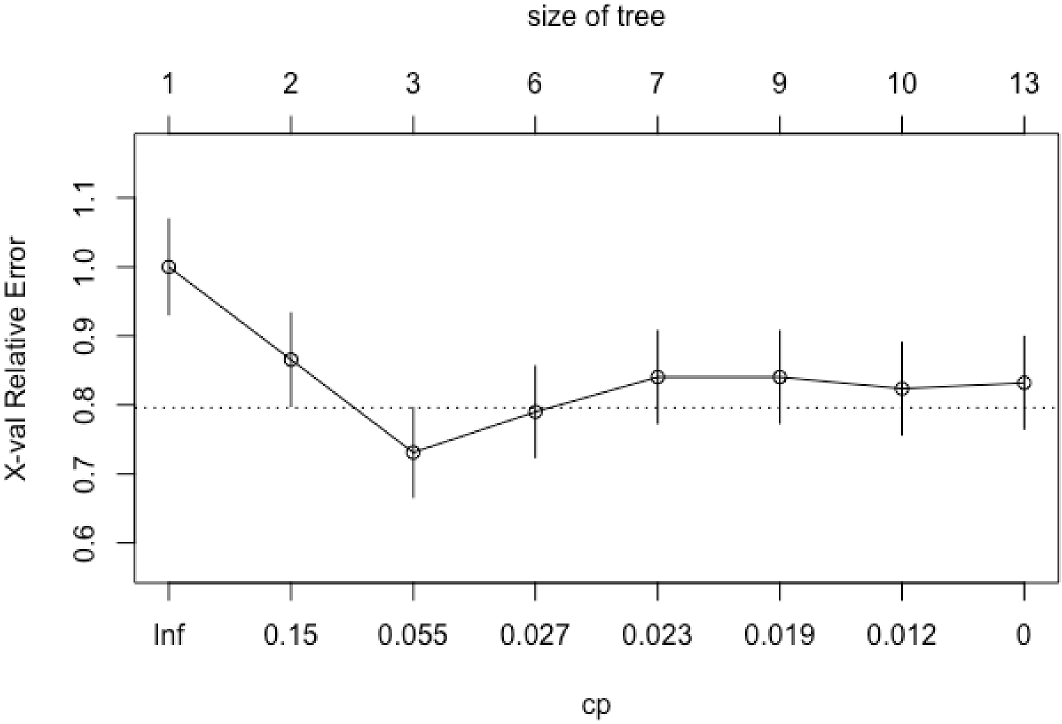

Figure C.1: Tuning of the DT size for the Student Performance dataset

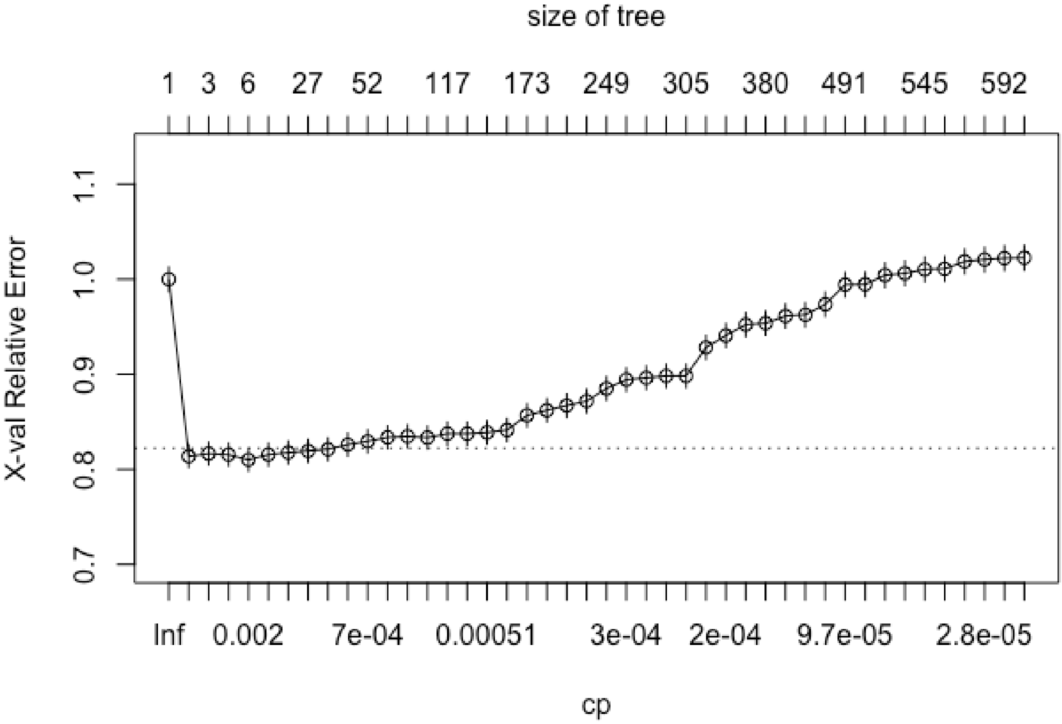

Figure C.2: Tuning of the DT size for the Credit Card Default dataset.

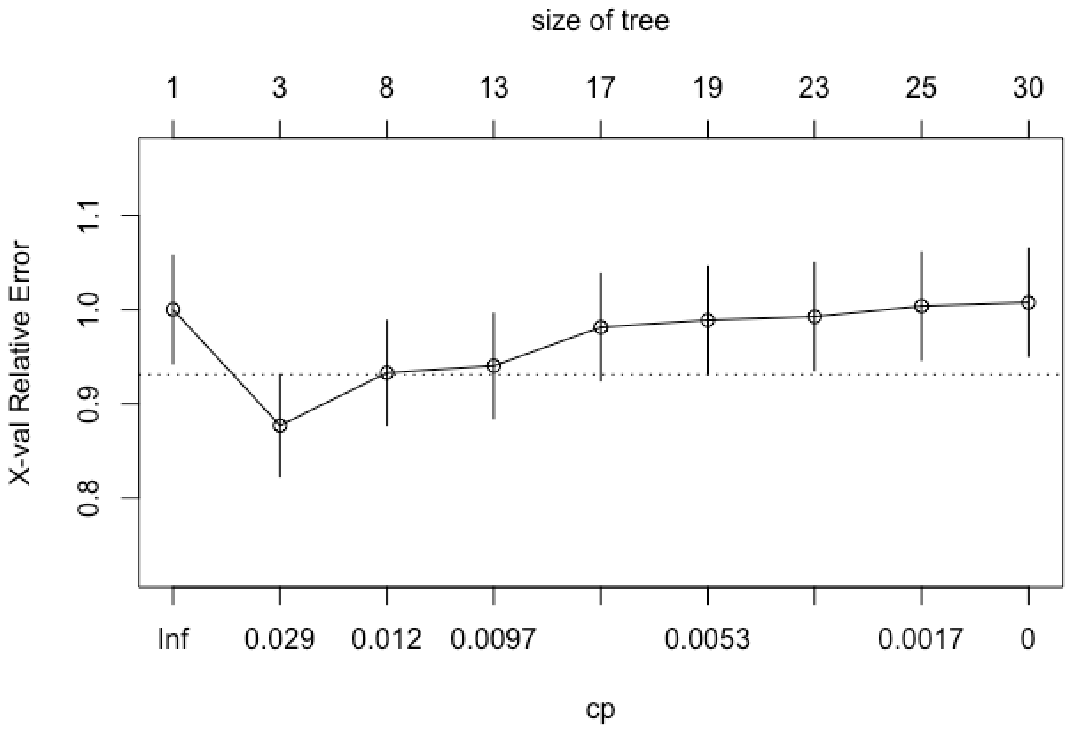

Figure C.3: Tuning of the DT size for the Bank Marketing dataset

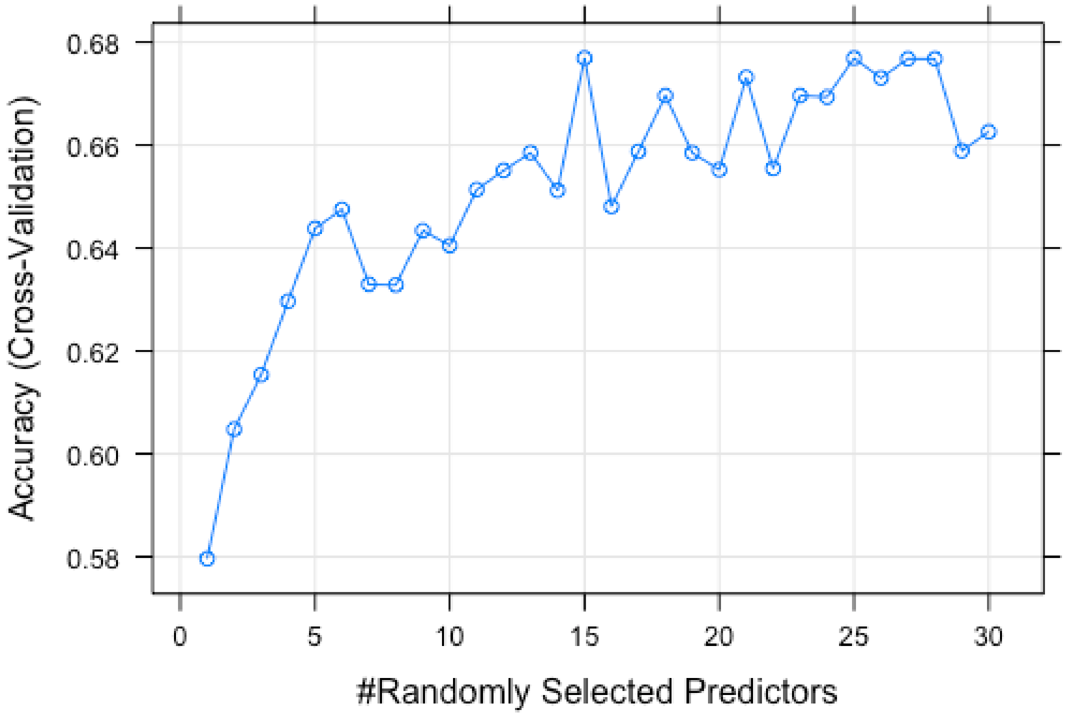

Figure C.4: Tuning of m_{try} for the Student Performance dataset

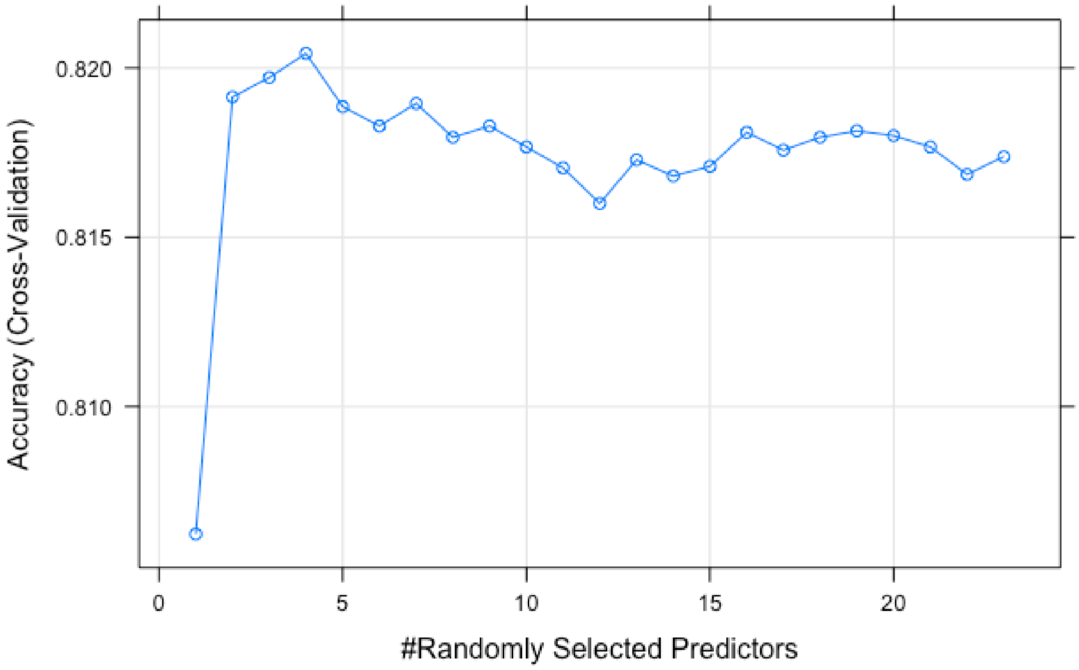

Figure C.5: Tuning of m_{try} for the Credit Card Default dataset

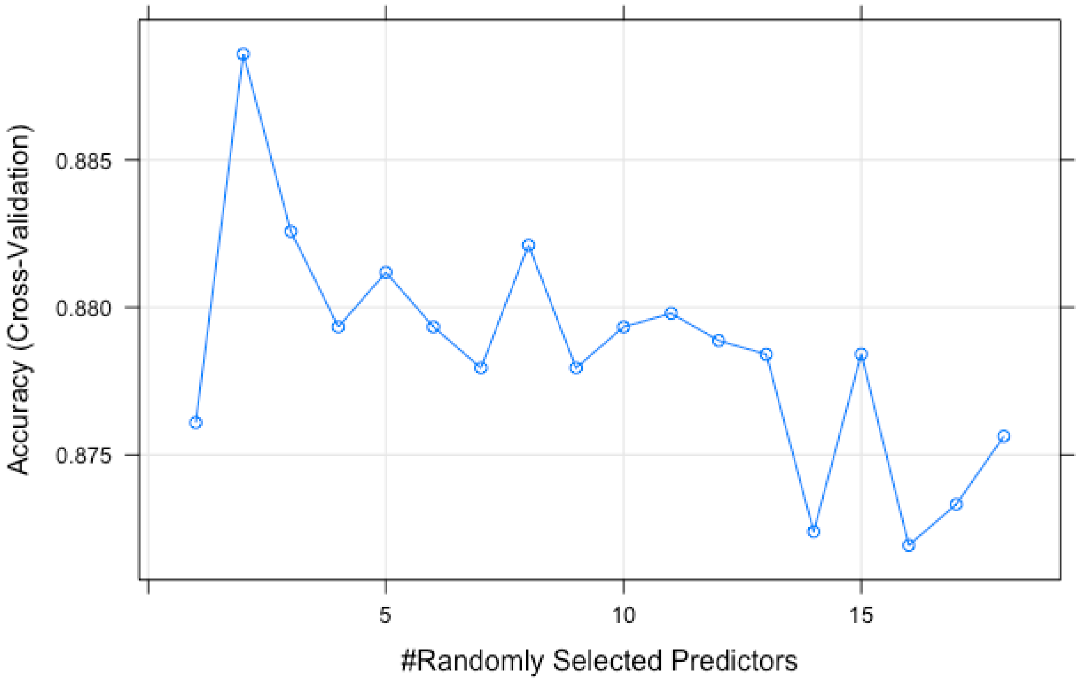

Figure C.6: Tuning of m_{try} for the Bank Marketing dataset

D Survey Results

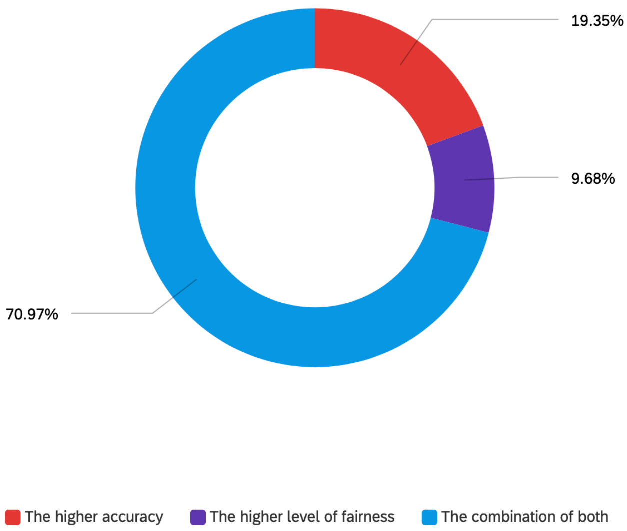

Figure D.1: Performance prediction context: Why did you choose a certain model?

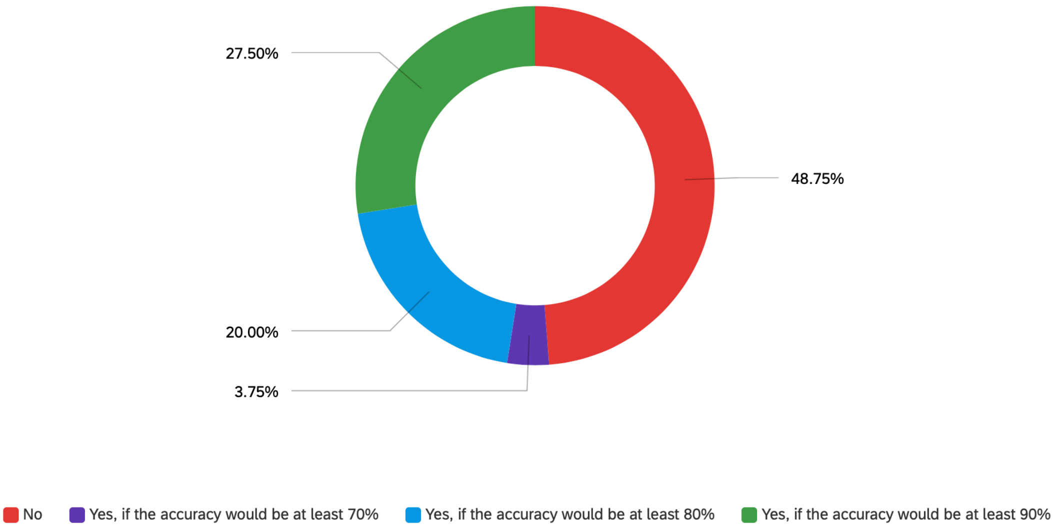

Figure D.2: Performance prediction context: Would your answer change if the difference in accuracy between DT and the SVM was even bigger?

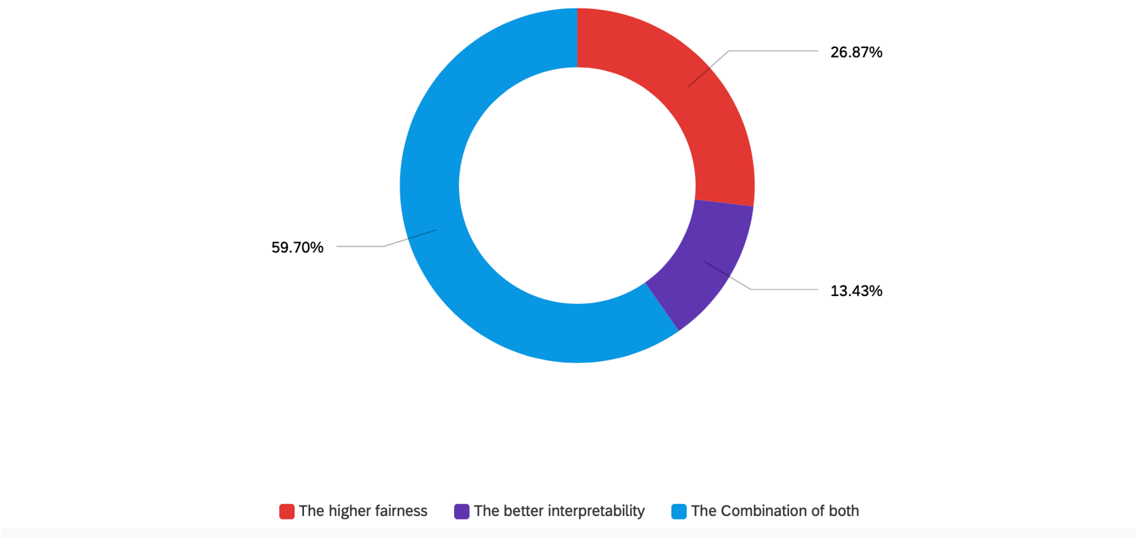

Figure D.3: Risk prediction context: Why did you choose a certain model?

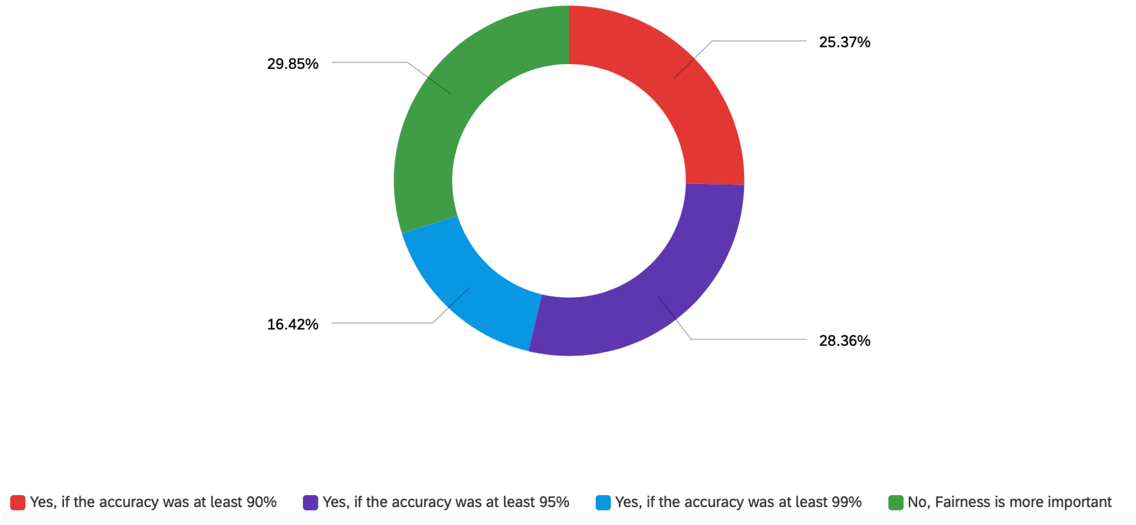

Figure D.4: Risk prediction context: Would your answer change if the difference in accuracy between RF and the ANN was even bigger?

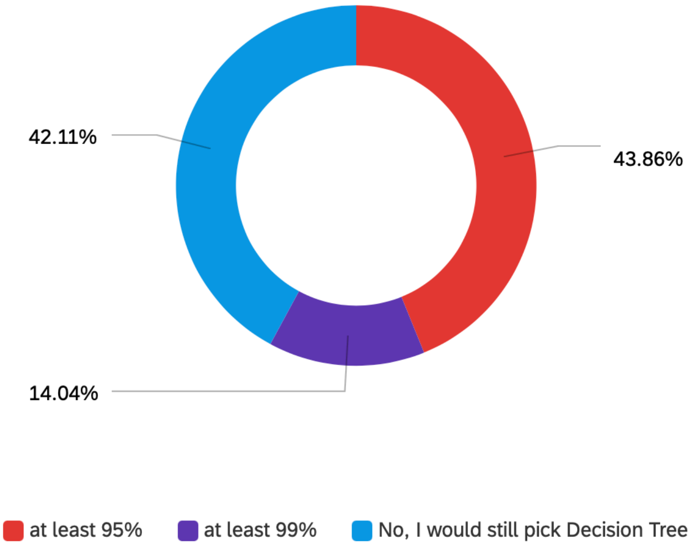

Figure D.5: Who to target context: Why did you choose a certain model?

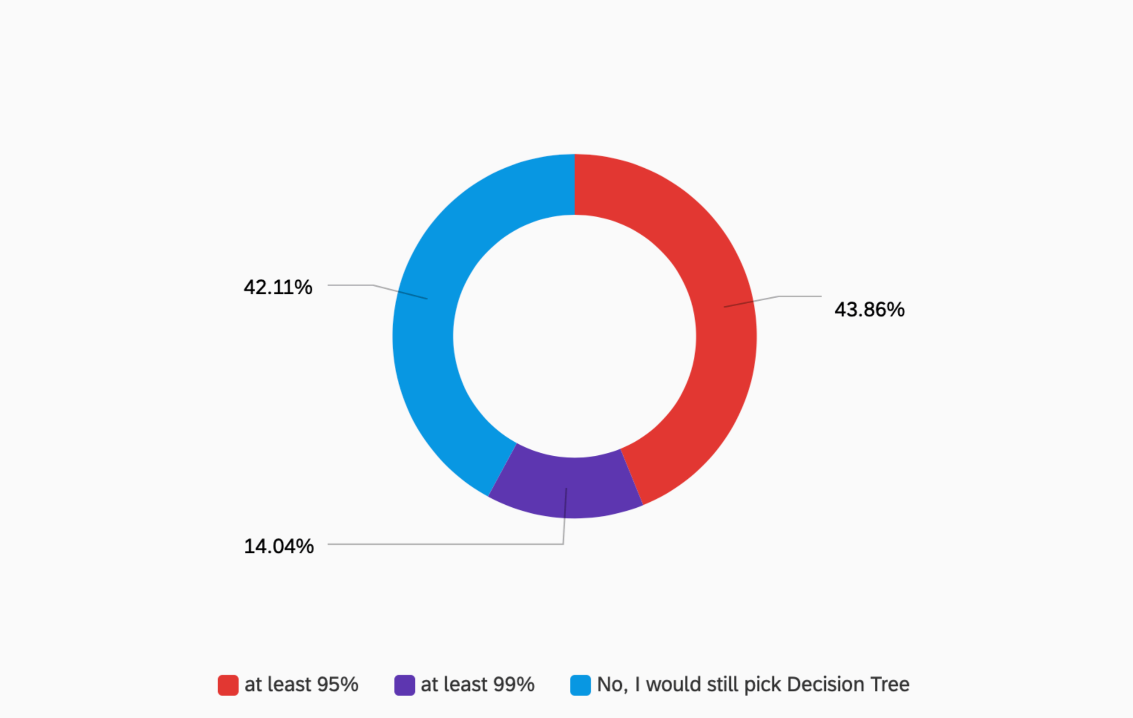

Figure D.6: Who to target context: Would your answer change if the difference in accuracy between RF and the DT was even bigger?