Norwegian School of Economics
Bergen, Autumn 2021

Shapley values in the context of GDPR

Can Shapley Values be used as a means of interpreting black-box machine learning models while also complying with the General Data Protection Regulation?

Eirik Juelsen & Marius Andre Thoresen

Supervisor: Håkon Otneim

Master thesis, MSc in Economics and Business Administration
Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Economics and Business Administration at NHH. Please note that neither the institution nor the examiners are responsible − through the approval of this thesis − for the theories and methods used, or results and conclusions drawn in this work.

Abstract

The General Data Protection Regulation implemented in 2018 by the European Union imposes strict requirements when handling personal data regarding European citizens. This is especially true when processing said data in combination with machine learning and AI. Within these requirements lies an inherent focus on the rights of the subject and their ability to exercise these rights. Through the GDPR the subject is required to be informed about any existence of machine learning models utilising their personal data and provided meaningful information concerning the logic of the model and an explanation of the inferences made by this model.

This master thesis examines the possibility of employing Shapley values in the process of building a machine learning model and its ability to provide meaningful information to the subjects affected by this model. We review the compliance of Shapley values according to the GDPR throughout the machine learning process and highlight how the framework is affected by specific articles in the GDPR. We argue that the most applicable categories of the GDPR in relation to machine learning models explained with Shapley values are Consent, Personal Data, Processing, and the Right to be informed.

The GDPR significantly affects all aspects of a machine learning model, from data collection to prediction explanation. We argue that by utilising Shapley values as a framework throughout the process, we have trained, and are able to explain the predictions of, a binary classification model. We believe this model both complies with the strict demands set forth by the GDPR as well as provides strong predictions, indicative of the ability to utilise Shapley values within the legal framework of the GDPR.

Acknowledgements

This thesis is a part of our MSc in Economics and Business Administration at the Norwegian School of Economics (NHH). We both major in Business Analytics.

We would like to offer our sincere gratitude to Tom Robin Nilsen and Leif Erik Thorstensen with Intrum for providing the dataset and always being eager to answer any questions. We would also like to thank Morten Nestvold Løvaas at AVO consulting for his insightful comments. Special thanks to our friends and family as well, for their support during the process of writing this thesis.

Finally, we wish to thank our supervisor, Associate Professor Håkon Otneim for his valuable feedback and guidance throughout this process. His help was invaluable in improving the quality of this thesis.

Norwegian School of Economics

Bergen, December 2021

Eirik Juelsen

Marius Andre Thoresen

Contents

Table of figures

List of tables

1. Introduction

The average human spends more than half their time awake using technology (Wallace, 2020), generating 146.88GB of data each day (Bulao, 2021). Data has become an exponentially growing commodity and is expected to reach a market value of 280 billion dollars by 2025 (Allinson, 2021). Utilising data is becoming an integral part of business strategy and in the first quarter of 2019 alone, more than 28 billion dollars were allocated to machine learning research (Lazzaro, 2021).

At the same time as the companies of the world wish to obtain increasingly more data on our every move, algorithms are being integrated into the very fabric of our societies and becoming integral parts of the social safety net (Human Rights watch, 2021). As a result, there were growing concerns amongst former members of the United Nations, academics, and several civil rights movements. This concern became all the more visible when the European Union unveiled the General Data Protection Regulations (GDPR) as an effort to protect the rights of European citizens (Wolford, 2020). The European Union underlines that the cohabitation of the GDPR and AI is a potential solution to ensure that European values and rules are upheld while harnessing the full potential of AI (European Commission, 2021).

In recent years the idea of explainability in black box machine learning models has increased due to regulations such as the GDPR and a larger interest from corporations to understand and utilise data more efficiently (Stewart, 2020). Interpretable machine learning (IML) was previously a smaller field within Machine learning but has quickly risen to be a major topic in the development of AI and Machine Learning for the future (Molnar, Casalicchio, & Bischl, 2020).

There are many ways of interpreting a machine learning model. However, many popular explanation techniques make use of the Shapley value (Gopinath, 2021), developed for game theory by Lloyd Shapley more than 60 years before the introduction of the GDPR (Shapley, 1953). One such approach is the Kernel SHAP method proposed by Lee and Lundberg (2017). This method was improved upon by the Norwegian Computing Center called *shapr* (Aas, Jullum, & Løland, 2021), incorporating feature dependency to reflect the real world more accurately.

This thesis wishes to highlight how the utilisation of Shapley values interacts with the requirements set forth by the GDPR and review the legality of employing IML explained with Shapley values. In this thesis we will be applying the improved Kernel SHAP method to data used for credit scoring and debt collecting. This category of data and the model processing it is regarded by the EU as a "high risk" AI-system due to the potential effects said AI may have on the life of the subject (European Commission, 2021). This classification entails increased demands in security and the protection of the rights of the subject, which in turn increases the demands set on Shapley values and its compliance with the GDPR.

As such, the research question we wish to address is the following: *are Shapley values an appropriate framework to adequately explain predictions and provide subjects with the relevant information needed to satisfy the demands in the GDPR*? In conjunction, we will be creating a model that predicts probabilities of successful debt collection as an example of the effect GDPR has on the use of interpretable machine learning.

This thesis highlights the implications and restrictions set upon machine learning models after the introduction of the GDPR. We show how Shapley values can be utilised to isolate the most important features in a dataset. Assisting businesses in creating machine learning models that respect the right of the subject in regard to the amount of personal data collected while ensuring that models still maintain good predictive power.

We also show how Shapley values can be employed to interpret the results of a black box machine learning model, which can then be explained in layman terms to the subject and comply with the GDPR. These interpretations may be extracted from the model on both a global and individual level. This allows the subject to understand which features are important to the model, how their own feature-values are weighted by the model, and how the subject in question compares to other subjects with similar feature values.

We believe our thesis highlights how to implement interpretable machine learning in the modelling process as well as the ex-post explanation of model output. The legality of employing machine learning models is reliant on the ability of the explanation method to provide the subject with *meaningful* information, underlining the importance of utilising a robust and easy to understand explanation method.

This thesis is divided into seven different sections. Section two will introduce the relevant literature. Section three introduces the framework, which will be at the centre of our prediction explanations and model building. Section four details how we modelled our prediction model. Section five illustrates how Shapley values applied to prediction output produces meaningful explanations. Finally, section six contains our discussion regarding Shapley values and the rules and regulations set forth by the European Parliament before concluding on our research question in section seven.

2. Literature Review

2.1 GDPR

GDPR or General Data Protection Regulation is a legal framework introduced by the European Union in April 2016, where specific requirements are introduced regarding privacy and data protection for European citizens. However, the legal enactment had a two-year transition period for businesses to adapt to the new demands and was fully implemented in May 2018.

GDPR contains 99 articles detailing in large how companies handling data regarding EU citizens can process and store this data. However, the main point that affects most people is the requirement to consent to data processing, anonymising the collected data, and safely store and transfer said data.

This thesis will discuss the requirements that come into effect when dealing with Machine learning models and which restrictions this puts on the usage of these models. The critical articles mentioned in the GDPR that could come into effect are Article 5, Articles 13-15, Article 22, Article 25, and Recital (71). Beyond this, the European Parliamentary Research Service produced in June 2020 a report which analysed and discussed the impact of the GDPR on artificial intelligence and machine learning. This thesis will extract all arguments and GDPR articles brought forth in this report which is topical and can be related to the research question. There are multiple GDPR Articles included in the report but below follows a general explanation of the most important articles and recitals.

GDPR Articles 13 and 14 gives the subject the right to be informed of all aspects surrounding their data. These articles leave little room for interpretation and are very clear in how and when a company should respond if they get a request from a subject. Article 13 is specifically utilised when the information is collected from the subject itself, and Article 14 is used if the data is collected from a third party.

GDPR Article 15 gives subjects the right to access or receive a copy of all information regarding themselves that is being stored or processed. The only exceptions to Article 13 through 15 are described in Article 12(5), where the request is “manifestly unfounded or excessive” (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679,

Article 12(5)). Furthermore, Articles 13 through 15 also gives the subject the right to receive information regarding “the existence of automated decision-making, including profiling”, and “meaningful information about the logic involved” (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 13-15).

GDPR Article 22 is aptly named “Automated individual decision-making, including profiling”, where the subject is given the right to get their case reviewed by a natural person. However, this right may not be applicable in certain situations described in Article 22 Paragraph 2, where extra guidelines are nuanced in Article 22 Paragraphs 3 and 4.

GDPR Article 5 defines the “principles relating to processing of personal data”, where Article 5(1)(c) in combination with GDPR Article 25(2) create the requirement of “data minimisation”. This requirement only allows the storage and utilisation of personal data necessary for a specific purpose.

GDPR Article 25 in general defines “data protection by design and by default” and requires technical and organisational measures to ensure that only the necessary data is collected, stored and processed as well as the security and privacy of said data.

The GDPR includes multiple recitals that give further supporting context to the articles and provide further information on its usage while being legally non-binding. The most applicable recital regarding machine learning is, as mentioned previously, Recital (71), where the subject is given further rights in cases regarding profiling and decisions made solely based on automated processing. Regarding machine learning models the Recital (71)(4) provides the constraint;

> “In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision” (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Recital (71)).

Recital (71), in other words, gives subjects the right to an explanation on how the decision was reached and the logic involved, and the ability to challenge this decision or opt-out of such data usage. (Burt, 2017)

The GDPR covers 13 key issues, but not all of these are relevant regarding the utilisation of Shapley values and machine learning models. Mainly the issues reviewed in this thesis are “Consent”, “Personal Data”, “Processing” and “The right to be informed”. These four categories will serve as focus points in the discussion about the compliance of machine learning models explained with Shapley values.

2.2 Machine learning

Machine Learning (ML) is a type of Artificial Intelligence (AI) set up to become more accurate over time or make predictions with higher accuracy than if they were computed manually. Since its conception, ML has been chiefly seen as a “black box” setup where the machine can find systems or links in the data that a human would not understand. This black box setup leaves much to be desired in terms of interpretability, where the steps from input to output are somewhat unclear. (Burns, 2021)

GDPR requires, as mentioned, “meaningful information about the logic involved”, but it is somewhat unclear what “meaningful information” denotes. Depending on pre-existing knowledge, what is considered meaningful and comprehensive for some people might not be for others. However, Articles 13-15 in the GDPR does relate to the rights of the subject, and what is considered “meaningful information” should be interpreted from the point of view of the subject. It is still unclear to which degree the information needs to be meaningful, but this requirement could be decided from a functional point of view. In other words, the information should be meaningful enough for the subject to determine and exercise their rights provided by Article 22(3) GDPR. (Selbst & Powles, 2017)

Interpretable Machine Learning (IML) seeks to solve this problem by understanding what caused a specific output on either a local or global scale. Interpretability at a global scale allows the subject to understand the model and the reasoning behind each decision entirely. At a local scale, the subject does not need to be able to interpret the entire model but rather be able to trace back a single decision and understand how the model came to that conclusion. (Schmitt, 2020)

There are different degrees of interpretability that are often used interchangeably. In this thesis, we make the distinction between interpretability and explainability. Interpretability usually sets a higher standard for understanding the model, and the subject should be able to comprehend how the model reached a specific conclusion fundamentally. On the other hand, explainability is a classification on whether the subject can understand a particular node in a complex model and its effect on the output. (The Lancet Respiratory Medicine, 2018)

When working with Interpretable Machine Learning, there are two main categories of interpretation methods. These are model agnostic and model specific interpretations. When working with random forest, gradient boosting or neural networks, the analyst must employ model agnostic methods to interpret the results. There are different methods to accomplish this, such as LIME, Partial Dependencies Plots (PDPs) and Shapley Additive Explanations (SHAP). (Molnar C. , 2021)

There are five main advantages of a model agnostic method versus a model-specific approach (Molnar C. , 2021). First, as insinuated by the name, model agnostic methods are more flexible in how they work when interpreting multiple different models. In contrast, the model-specific methods are limited to specific model classes. While the model-specific methods are intrinsically interpretable such as regression weights, a model agnostic method allows the analyst to employ different explanation types. In some cases, a linear formula might be the optimal explanation system, while in others, it might be a feature importance plot. This is also called explanation flexibility.

The third main advantage of a model agnostic method is representation flexibility (Ribeiro, Singh, & Guestrin, 2016). This allows the interpretation method to be used in combination with different underlying features that in themselves might not be interpretable.

Another advantage of a model agnostic method is the lower cost of switching models during a machine learning pipeline (Ribeiro, Singh, & Guestrin, 2016). For example, if new information or a change in the dataset impairs the need for another model when using model-specific interpretation methods, the interpretation method must also be changed. This could cause a setback in the entire project or force the controllers to employ an entirely new interpretation method. When employing a model agnostic interpretation method, the underlying model can be changed without any issues or changes in the interpretation.

This ease of changing the model directly connects to the last advantage of model agnostic methods. A different output or interpretation method could complicate the comparison process if a user seeks to compare two or more models. Using a model agnostic method, the output or insight learned from the different models can be explained using the same techniques and representations (Ribeiro, Singh, & Guestrin, 2016).

As an inverse to the advantages mentioned above, a model specific method is locked to a certain model class. In other words, an interpretation method utilised on a random forest model, will not work on a linear regression model or vice versa. (Sarkar, 2018)

When reviewing a model and its usability, there is a need to evaluate how well a model recognises patterns and relationships within the data provided, also known as the training set. These relationships then lay the foundation for the model to predict a result or classification based on "unseen" data, the test set.

The model created for this thesis is a classifications model, and when performing classification predictions, there are four possible outcomes. These are true positive, false positive, true negative and false negative.

- A true positive (TP) result occurs when the model classifies an observation as belonging to a class, and the observation does belong to that class.
- A true negative (TN) result occurs when the model classifies an observation as not belonging to a class and the observations does not belong to that class.
- A false positive (FP) result occurs when the model classifies an observation as belonging to a class, and the observation does not belong to that class.
- A false negative (FN) result occurs when the model classifies an observation as not belonging to a class, and the observation does belong to that class.

The simplest method to evaluate a model is to calculate the accuracy. Accuracy can be defined as the ratio between how many times the model predicts a specific outcome and how often this outcome should occur, defined as the formula $(TP + TN)/(TP + TN + FP + FN)$. Although this formula calculates the percentage of cases where the model is correct, it is not necessarily the best way to evaluate a model.

Other methods to score a machine learning model are precision, recall and F-score. Precision calculates the proportion of true positives out of all detected positives $\big(TP/(TP+FP)\big)$, while recall is the proportion of true positives out of all positives in the dataset $\big(TP/(TP+FN)\big)$. The F-score is then a mean of precision and recall, where the formula can be altered by valuing precision and recall differently to reflect the dataset better. However, the harmonic mean (precision and recall weighted equally) is calculated by $\frac{TP}{TP+\frac{1}{2}(FP+FN)}$. (Wood, 2021)

The problem with these metrics is that accuracy is sensitive to class imbalance, while precision, recall and, by extension, F-score is generally asymmetric (Shmueli, 2019). To eliminate these problems, it is possible to employ Mathews Correlation Coefficient (MCC). The MCC calculates the correlation coefficient between predicted and actual classifications. Matthews Correlation Coefficient possess beneficial properties such as easy interpretability and being perfectly symmetric, such that no class is more important than another. The Correlation Coefficient can only ever be between 1 and negative 1, where MCC is one if $FP = FN = 0$, and negative one if $TP = TN = 0$. The MCC is easy to interpret based on these ranges, where a higher MCC equals a better model and where MCC=0 means that the model is no better than flipping a coin. MCC is calculated using the formula (Shmueli, 2019):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Another method to show the performance of a classification model is using ROC curves and AUC. ROC stands for Receiver Operating Characteristic and was first invented during the Second World War in the US to improve the detection of Japanese aeroplanes (Toshniwal, 2020).

The purpose of ROC is to increase the detection of True Positives while reducing False Positives. Similarly to MCC, the ROC curve is symmetric, meaning that ROC curves will be the same no matter the composition and class distribution in the dataset. Studying the ROC curve visually makes it possible to determine the optimal prediction threshold of the model. The ROC curve shows the trade-off between the true positive rate and the false positive rate, and the threshold can be changed depending on the objective of the model. E.g., a false positive could be costly depending on the utilisation of the mode, and the threshold should then be higher.

Although visually inspecting the ROC curve can be used as a measure of model performance, another option is to evaluate the model based on Area Under Curve (AUC). Area under curve is a metric that aggregates the performance of the model at all possible thresholds. AUC can vary from zero to one, where the score is zero if every classification is wrong and one if every classification is correct. Within this range, a score of 0.5 signifies that the model is no better than flipping a coin (Mandrekar, 2015). All values above 0.5 indicates that the model has some predictive power, but what constitutes a "good" AUC differs from case to case. The AUC can then be utilised to compare multiple models, where the model with the higher AUC should be considered the more optimal model.

3. Shapley values, Kernel SHAP & dependency extension

This chapter will touch on the general theory behind the Shapley value and how it will be applied to the machine learning model to interpret it. Section 3.1 will give a general overview of the original Shapley value proposed by Lloyd Shapley in 1953 before sections 3.2 and 3.3 will go more into how the Shapley value will be applied in our case.

3.1 Shapley Values

Shapley values are a part of a field known as cooperative game theory. Chalkiadakis, Elkind and Wooldridge (2011) defines cooperative game theory as "a branch of (micro-)economics that studies the behavior of self-interested agents in strategic settings". Shapley values are for the specific games where parties cooperate to maximise their total payoff. The Shapley value (Shapley, 1953) is a "fair" way of allocating this payoff amongst the members of the game, assuming collaboration between all members.

The idea behind the Shapley value is that you let $S \subseteq \mathcal{M} = 1, \dots, \mathrm{M}$ where $|\,S\,|$ is the number of players. Each player S then has a contribution function $v(S)$, which maps subsets of players to real numbers. This mapping is the contribution of coalition S and is the total payoff the players in coalition S can expect from cooperating. Shapley values can then be used to allocate these gains amongst the members, and how much each member will be allocated can be explained by function (1) according to Aas, Jullum, and Løland (2021).

$$\phi_j(v) = \phi_j = \sum_{S \subseteq \mathcal{M}\{j\}} \frac{|S|!\,(M - |S| - 1)!}{M!}(v(S \cup \{j\}) - (v(S)), j = 1, \dots, M \tag{1}$$

They illustrate this by using an example where $\mathcal{M}$ = {1,2,3}. This makes it so that there are eight possible subsets. By then applying function (1) onto, e.g., player one, the Shapley value for player one will be calculated by the following:

$$\phi_1 = \frac{1}{3}(v(\{1,2,3\}) - v(\{2,3\})) + \frac{1}{6}(v(\{1,2\}) - v(\{2\})) + \frac{1}{6}(v(\{1,3\}) - v(\{3\})) + \frac{1}{3}(v(\{1\}) - v(\{\emptyset\})) \tag{2}$$

They also make a case for pointing out that one defines the non-distributed gain $\phi_0 = v(\emptyset)$ where there are no members of the coalition. Although this gain most often turns to 0 for usual coalitions, it will be helpful for our analysis in section 5.1.

3.1.1 Properties of Shapley values

The reasoning for using Shapley values lies in their inherently desirable properties, being the only set of values to satisfy all four properties simultaneously, this having been proven both by Lloyd Shapley himself (1953) and a later study performed by Young (1985).

Efficiency

The efficiency property of the Shapley Value ensures that all gain is distributed (Molnar C. , 2021). This property can be displayed using the following formula (3), as shown in the article by Aas, Jullum, and Løland (2021).

(3) $$\sum_{j=0}^{M} v(\mathcal{M})$$

Symmetry

The symmetric property of the Shapley value dictates that if there are two players whose contributions in all coalitions are identical, these two players will have identical Shapley Values in all coalitions in which neither contributes. This can be displayed using the following constraint:

(4) $$if: v(S \cup i) = v(S \cup j) \Longrightarrow \phi_i = \phi_j$$

(Molnar C. , 2021)

Dummy player

Shapley Values having a dummy player property means that if a player does not contribute to any coalitions, this will also lead to that player obtaining a Shapley Value of 0.

Linearity

Lastly, the Shapley value property of linearity states that "If two coalition games described by gain functions υ and w are combined, then the distributed gains correspond to the gains derived from υ and the gains derived from w:

$$\phi_i(\upsilon + w) = \phi_i(\upsilon) + \phi_i(w), \qquad \text{for every i} \tag{5}$$

" (Aas, Jullum, & Løland, 2021). This property is of high significance when using Shapley values for interpretation, as it enables the possibility of looking at individual features and their effect on the final Shapley Value.

3.1.2 Prediction explanation through Shapley values

Assume that we have a machine learning model $f(\boldsymbol{x})$ that attempts to predict y . Not only do we want to predict y, but we also want to understand how feature $\boldsymbol{x}^*$ affects this prediction. In their paper, Lee & Lundberg (2017) suggest that this be done using Shapley values. To facilitate this, one can apply the Shapley framework to a single prediction.
We substitute the pay-out for the specific prediction. Recalling back to 3.1, the features of the model replace the players. Aas, Jullum and Løland (2021) decompose it to the following equation:

$$f(\boldsymbol{x}^*) = \phi_0 + \sum_{j=1}^{M} \phi_j^*, where\ \phi_0 = E[f(x)]\ and\ \phi_j^* is\ the\ \phi_j\ for\ the\ prediction\ \boldsymbol{x} = \boldsymbol{x}^*. \tag{6}$$

Adding up the Shapley values for all features in x describes the difference between the specific prediction,

$$y^* = f(\boldsymbol{x}^*), \tag{7}$$

and what to expect on a general basis (global mean prediction). A model that behaves in this manner is known as an additive feature attribution method (Aas, Jullum, & Løland, 2021) and is the only model of its kind that satisfies all four of the properties described earlier by Lee and Lundberg (2017). This aspect of Shapley values makes it more suited for explaining than other additive feature attribution methods, e.g., LIME (Ribeiro, Singh, & Guestrin, 2016),

as no other method possesses all four properties, making them more susceptible to inconsistencies.

To be able to compute the Shapley values for prediction explanation, we define the contribution function $v(S)$ for the subset S, and this should resemble $f(\boldsymbol{x}^*)$, given that only the values of subset S are known for these features (Aas, Jullum, & Løland, 2021). We then recreate the work of Lee & Lundberg (2017) in (8), using the expected outcome of the predictive model with feature values $\boldsymbol{x}_S = \boldsymbol{x}_S^*$, summarised in equation (2).

$$v(S) = E[f(x) \mid \boldsymbol{x}_S = \boldsymbol{x}_S^*] \tag{8}$$

3.1.3 Advantages and disadvantages of Shapley Values

As described earlier, Shapley Values is the only additive feature attribution method that can satisfy all four properties described in 3.1. By possessing all four attributes, Shapley values as a method can fairly and evenly distribute amongst all the features while at the same time being an excellent tool for individual predictions (Aas, Jullum, & Løland, 2021).

Another critical aspect of the Shapley value relates to the GDPR. In section 2.1, Articles 13-15 refers to *meaningful information*. Since Shapley Values can display percentage changes in the prediction in an intuitive and informative manner, data scientists and laypeople alike can interpret and understand the output of the model.

One downside when utilising Shapley values to explain predictions is the added computational cost of including the new feature *m*. This is due to the exponential growth in the number of possible subsets. There are 2^M different possible subsets, where M is the number of features (Aas, Jullum, & Løland, 2021). Another caveat of the Shapley value is that it requires an approximation for all $\boldsymbol{x}_S$ in equation (2).

3.2 Kernel Shap

There are several ways of going about the prediction explanation shown in section 3.1.2, and one way of doing this is the Kernel SHAP method, proposed by Lundberg and Lee (2017). This method can be divided into two distinct parts:

i) Computing an approximation of the Shapley values in an easy to replicate way

There exist many different and equally correct ways to formulate the Shapley value. Lee and Lundberg (2017), as well as Charnes et al. (1988) before them, define Shapley values as a weighted least square (WLS) problem and state that calculating the Shapley values can be reduced to solving the following minimization problem:

(9) $$min \sum_{S \subseteq \mathcal{M}} (v(S) - (\phi_0 + \sum_{j \epsilon S} \phi_j))^2 k(M,S),$$

where k(M,S) are the kernel weights .This equation can again be simplified by making some assumptions. We let Z represent all possible combinations with the inclusion and exclusion of the M features in a $2^M \times (M+1)$ binary matrix. Let v denote the vector that contains $v(S)$, and $\boldsymbol{W}$ denote a diagonal matrix for $k(M, |S|)$ (Aas, Jullum, & Løland, 2021). This equation can be written as such:

(10) $$(v - Z\phi)^T W (v - Z\phi)$$

This can, in turn, be solved by:

(11) $$\phi = (Z^T W Z)^{-1} Z^T W v$$

This is a computationally heavy exercise, and when M increases, it can impose problems due to an exponential increase in subsets. That is why the weighted least square formulation in (9) is used to approximate. Due to the differing sizes of the Shapley Kernel weights, many of the rows in Z contribute insignificant amounts. One can then approximate utilising a subset D of $\mathcal{M}$, and then only use the corresponding rows in Z. This yielded Aas, Jullum and & Løland (2021) the approximation:

(12) $$\phi = [(\boldsymbol{Z}_D^T \boldsymbol{W}_D \boldsymbol{Z}_D)^{-1} \boldsymbol{Z}_D^T \boldsymbol{W}_D] \boldsymbol{v}_D = \boldsymbol{R}_D \boldsymbol{v}_D$$

ii) A method of estimating $v(S)$

As we remember from **i)** v contains all the $v(S)$ values, and $v(S) = E[f(x)|x_S = x_S^*$. $\bar{S}$ will be the complement of S, and $x_{\bar{S}}$ is all the x that is not part of x_S. The expected value can then be calculated by the following formula:

$$E[f(x)|x_S = x_S^*] = E[f(x_{\bar{S}}, x_S|x_S = x_S^*] = \int f(x_{\bar{S}}, x_S^*)p(x_{\bar{S}}|x_S = x_S^*)dx_{\bar{S}} \quad (13)$$

We observe that knowing $p(x_{\bar{S}}|x_S = x_S^*)$ is integral for calculating $v(S)$, but knowing this distribution is seldom a fact. The kernel SHAP method therefor assumes feature independence (Aas, Jullum, & Løland, 2021). We can then replace $p(x_{\bar{S}}|x_S = x_S^*)$ with just $p(x_{\bar{S}})$. With this assumption, we can now approximate the integral to:

$$^{v}KerSHAP(S) = \frac{1}{K}\sum_{k=1}^{K} f\left(x_{\bar{S}}^{k}, x_S^*\right), \quad (14)$$

where $x_{\bar{S}}^{k}$ and k =1,.., K are sampled from the training data.

3.3 Kernel Shap dependancy extension

As described in equation (8) in section 3.2, the original kernel SHAP method proposed by Lee and Lundberg (2017) assumes that all features are feature independent. However, this assumption does not naturally lend itself to an intuitive understanding of the world, where many aspects often have some degree of interconnection. Therefore, Aas, Jullum & Løland (2021) propose an extension to the original kernel SHAP to incorporate feature dependency rather than independence.

3.3.1 Relaxation

Aas, Jullum & Løland (2021) propose a relaxation of the independence assumption by approximating $p(\boldsymbol{x}_{\bar{S}}|x_S = \boldsymbol{x}_S^*)$ and generating samples from this estimation instead of the original solution of independent generation. This does however come with the caveat that the proper distribution for the estimation is found, and the authors propose four different distributions:

- Multivariate Gaussian
- Gaussian Copula
- Empirical conditional
- A combined approach

The mathematics behind these distributions are, however, beyond the scope of this thesis. For those interested, we highly recommend reading the work of Aas, Jullum & Løland (2021), as it provides a great insight into the topic of Shapley values and the kernel SHAP and their extension to incorporate dependant variables. We will discuss which distribution is most applicable to our data in section 4.3.3.

4. Modelling

Before we could start applying the Shapley framework to explain predictions; we needed to decide whether to use a pre-existing machine learning model or train a new one. The GDPR affects all aspects of the machine learning process. A new model will be trained to ensure full compliance and avoid uncertainty surrounding older models in relation to GDPR requirements. This section will present our data, highlight the feature selection process, provide descriptive statistics, and highlight the implications the GDPR has on the process of building a machine learning model.

4.1 Introduction of the data

Our data was obtained through the debt collector Intrum, formerly known as Lindorff. The complete dataset contains information regarding debt collection activities. It includes information about the subject in question, their monetary situation, prior debt collection history, and the outcome of the debt collection activity in question. All data that could be used to identify a specific subject has been either removed or anonymised, and all data were collected simultaneously from the database of the data supplier. Due to the removal of several features in the original dataset as a part of the data treatment process, section 4.3.3 will further explain the remaining data.

The European Commission released a statement concerning high-risk AI systems, where AI systems regarding essential services, such as debt collection, are specifically mentioned. These systems will be subject to a higher standard and *must* abide by all requirements before being deployed. Failure to comply will result in the AI system being banned, and any use of said AI system would be regarded as illegal. If Shapley values can abide by the strict requirements set upon high-risk systems, it can be argued that, by extension, Shapley values can be applied to systems defined as “limited & minimal risk” (European Commission, 2021) as well.

4.2 Data treatment

4.2.1 Result variable

The model trained for the purpose of this thesis aims to predict the probability of the feature *Full payment*. This is a binary feature that takes the value one if, and only if, the instance of debt collection was paid in full. The feature would take the value of zero if the subject paid the debt collection partly or not at all. As the data contained more than one possible result feature, several features had to be removed from the dataset due to a direct correlation with the chosen response variable.

4.2.2 Missing values

Although the data provided to us from Intrum was, for the most part, complete, there were some problems related to the migration of the data from the data providers internal network to our possession. This resulted in missing values in more than 74.000 different entries. Intrum confirmed that a large part of the missing values was a by-product of differences between their internal language and R. These missing values were intended to be zero and could safely be replaced.

GDPR Article 4(4) defines profiling as: “processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person”. By this definition, our model and the intended use must be regarded as profiling, which entails specific requirements. GDPR Recital (71)(6) states that:

> “the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised” (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Recital 71).

One way of dealing with the remaining missing values is to delete these entries from our data. However, this method results in the loss of information (Kumar, 2020) and, if applied when values are missing for the wrong reasons, may introduce bias into the model (Swalin, 2018). A common approach to avoid deleting rows containing missing data is to perform feature imputation, replacing missing values with substituted ones. We argue that there is a non-zero possibility that introducing substituted values for individuals with missing values will introduce, rather than correct, inaccuracies. The substituted values may also falsely represent the subject and could lead to conclusions derived from inaccurate information. Feature imputation is therefore considered non-compliant with GDPR in our opinion. When observing the non-distributed gain before and after removing the remaining rows containing missing values, there is a change of 0.17%. Therefore, we argue that removing rows containing missing values has not introduced a significant amount of bias and is an acceptable way of treating the missing values while complying with GDPR Recital (71).

4.2.3 GDPR demands

In correlation with the dataset provided, the GDPR provides a requirement of data minimisation. This is done through Articles 5(1)c, which states data should be:

> "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 5)

and is further expanded upon in Article 25(2):

> "The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility." (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 25)

Before building a model, what constitutes relevant or necessary data can be challenging to determine. Fraser, Ohno-Machado & Ohrn (1998) found that removing redundant data had no significant effect on classification performance. Therefore, we argue that redundant data must be considered irrelevant or unnecessary for our model. Multicollinearity exists when two or more assumed independent variables correlate (Khanna, 2020), indicating that they portray the same information. It is possible to argue that correlated variables provide diminishing amounts of new information, deeming all but one of the collinear variables irrelevant or unnecessary according to the GDPR. One way to identify collinearity within datasets is by utilising the variance inflation factor. The variance inflation factor assigns each variable a collinearity score, where scores above five are deemed to be above acceptable levels (Bhandari, 2020). We applied the variable inflation factor on a linear regression that utilised all available features in our dataset, identifying and removing the feature with the highest inflation factor. This process was repeated until no features displayed an inflation factor above acceptable levels.

4.3 Machine learning model

4.3.1 Choice of model

The model type trained for this thesis was an XGBoost model (Chen & Guestrin, XGBoost: A Scalable Tree Boosting System, 2016). The reasons behind this were twofold, both rooted in demands put forth in Recital (71). As specified in subsection (6), the controller is responsible for ensuring that the risk of error is minimised. As highlighted by Vishal Morde (2019), current data science leader at Apple, in Figure 1 below, XGBoost performs better than other common machine learning models. Due to the low training times of the model, we are also able to process more data in less time. This property of XGBoost facilitates the possibility of reducing variance and enhancing the predictive performance of our model (Chawla, 2020) and consequently reducing error.

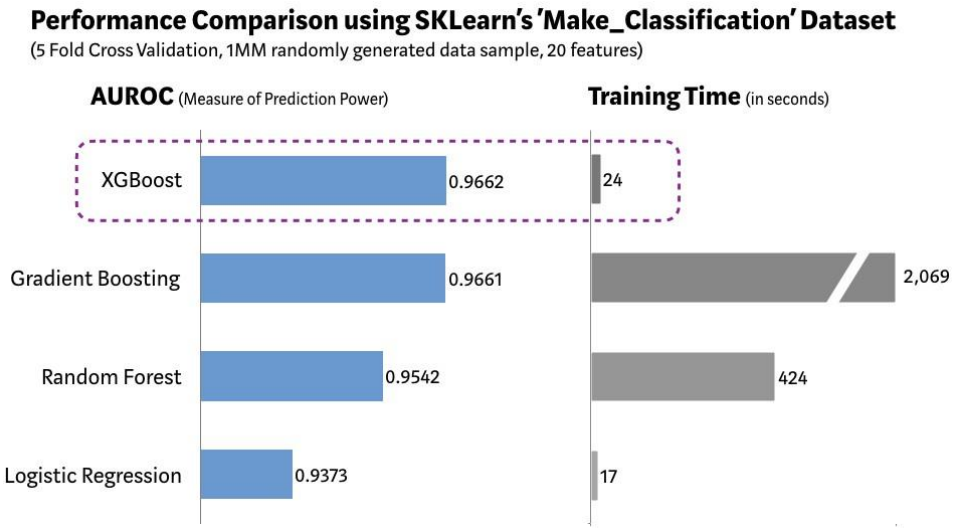

Figure 1 - XGBoost classification results

However, larger pools of data also facilitate more significant chances of overfitting the model. Therefore, several hyperparameters were introduced and tuned with this in mind. The parameters presented in Table 1 were chosen after testing different hyperparameter values during the cross-validation process. The explanations below were sourced from Chen et al. (2021).

Table 1 - List of XGBoost Hyperparameters

Parameter	***Chosen value***	***Explanation***
ETA	0.001	Controls the learner rate of the model. Is on a range from 0 to 1, lower values makes the boosting process more conservative, but more robust again overfitting
Evaluation metric	AUC	Evaluation metric for the XGBoost model. The model uses AUC as the criterion for optimisation.
Early stopping rounds	20	Will stop the model if the evaluation metric does not improve for 20 rounds.
Objective	binary:logistic	Specifies the learning task. Binary:logistic implies logistic regression for binary classification and outputs probability.
Number of rounds	10000	Maximium number of boosting iterations
Number of folds	5	Number of cross-validation folds

A low learner rate was chosen to utilise the speed of the model and the large dataset provided by Intrum. The evaluation metric was set to AUC for its properties in distinguishing between classes (Bhandari, 2020). The hyperparameter early stopping rounds is specified to avoid overfitting. If the model does not improve for 20 consecutive iterations, XGBoost reverts to the iteration with the best AUC. As probabilities of a binary outcome were the desired outcome of this model, the objective binary: logistic was chosen. In tandem with the low ETA and the fast nature of XGBoost, the model utilises 10.000 iterations.

4.3.2 Initial model

By utilising the hyperparameters in Table 1, the trained model obtained a training set AUC of 93.3% and a testing set AUC of 90.6%. The result of the cross-validation process can be seen in Figure 2 below:

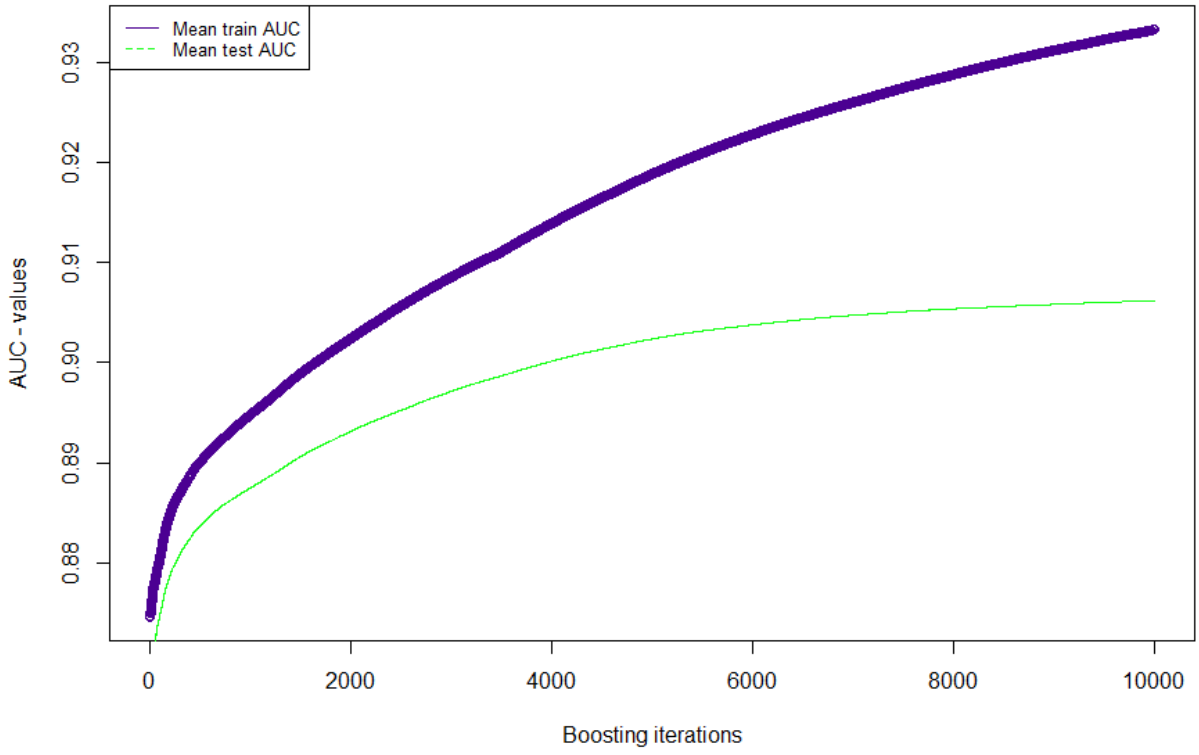

Figure 2 - Initial model AUC performance[1]

[1] We would like to comment on the choice of colour in our thesis. Wherever distinguishing between colours is relevant, we have opted for a colour palette that is distinguishable for all peoples, also those affected by varying degrees of colour blindness.

When inspecting Figure 2, iterations above 6.000 offer diminishing increases to the AUC, but the early stopping rounds argument did not come into effect. This indicates the possibility of performing additional iterations before overfitting. Still, due to the flat nature of the AUC, we argue that the increased computational time added by additional iterations does not warrant a minor increase in AUC performance.

As presented in the literature review, the model will also be evaluated using the Matthews Correlation Coefficient and accuracy. The Matthews Correlation Coefficient will be normalised ($\frac{MCC+1}{2}$) to fit a scale of zero to one, to make the interpretation more intuitive. Values of 0.5 will then indicate the same predictive power as filliping a coin, one indicating a perfect model and zero the opposite. Due to the model producing probabilities as output, there was a need to define a threshold for binary classification. To comply with GDPR Recital (71)(6) and minimise the risk of error, the optimal threshold was calculated to be 52% by iterating through all possible integer thresholds between one and one hundred. The optimal threshold produced a normalised Matthews Correlation Coefficient of 0.82 and an accuracy of 82.83%. The GDPR does not explicitly define requirements for the overall performance of machine learning predictions. Still, we argue that the results in AUC, normalised Matthews Correlation Coefficient and accuracy produced by the model abides by GDPR Recital (71)(6).

As explained in section 3.2 regarding Kernel Shap, some features have smaller weights than others due to Shapley values being a weighted least square problem. GDPR Articles 13-15, presented in the literature review, defined the right of the subject to be presented with *meaningful information*. Presenting the subject with 69 different features can exceed the amount of information a subject can reasonably be expected to comprehend. We, therefore, argue that a feature importance analysis is required to discern which features constitute *meaningful information*. Such an analysis causes implications with GDPR Articles 5(1)c and 25(2). Features that do not contribute *meaningful information* are challenging to describe as relevant or necessary and should be removed.

To investigate the feature importance within the model, we employed absolute Shapley values calculated using Kernel SHAP (Lee & Lundberg, 2017). The absolute Shapley value represents the importance of a feature, where the most important features obtain the highest values (Molnar C. , 2021). To obtain the global importance, we took the average absolute values for all features and Table 2 below displays the nine features most important to the model. All features with an importance score of less than 0.15 were deemed too insignificant to provide *meaningful information,* no longer warranting their inclusion. As discussed above, these features cannot be seen as relevant or necessary and are removed from the dataset. This decision is further supported by evaluating the nine-feature model in section 4.3.4.

Table 2 - Variable importance scores

Variable	***Importance score***
2 Years from extraction	1.08
Principal balance	0.31
Percentage without legal process	0.27
New Fee	0.25
Multiple cases same creidtor	0.24
Income	0.22
Payment remarks	0.17
Outstanding balance monitored	0.17
Deposit last 6 months	0.15

It is essential to distinguish between Kernel SHAP (Lee & Lundberg, 2017)and *shapr* (Aas, Jullum, & Løland, 2021). While *shapr* is working with transformed probabilities, Kernel SHAP uses margins. Lundberg (2018) detailed this problem in his GitHub repository, where he states that all models of the original Kernel SHAP will utilise log-odds during calculations. Therefore, the output of importance is not applicable to understanding feature effects. The intuition is still easily understood where higher values indicate increased importance. We argue that diverging from the Shapley value approach to introduce other means of calculating feature importance would cause more confusion for the average subject and that Shapley values alone allow the subject to understand "*the logic involved*" as stated in GDPR Articles 13-15.

4.3.3 Descriptive statistics

As a result of thorough data cleansing, we have obtained a GDPR compliant dataset. This section will further examine this data to learn more about its features. We will start by explaining the nine features present in the model, seen in Table 3. All monetary variables have been normalised to represent the number of standard deviations they diverge from the mean value to minimise the risk of differing scales creating bias in the data (Lakshmanan, 2019).

Table 3 - Variable descriptions

Variable	***Description***
2 years from extraction	A binary varaible that indicates if the case was more than two years from the extraction date
Principal balance	The total principal balance the subject owes across all current debt collection cases
Percentage without legal process	The fraction of previous cases related to the subject that were concluded without the use of legal process
New Fee	Fees handed to the subject in other cases prior to the extraction date
Multiple cases same creditor	The number debt collection cases to the same creditor of the same type (i.e. credit card debt)
Income	The last available income registered to the subject on the extraction date
Payment remarks	Number of payment remarks registered to the subject
Outstanding balance monitoring	Outsanding balance on insolvent demands at time of extraction
Deposit last 6 months	A binary variable that indicates if the subject has made a deposit on the debt collection in the last six months
Full Payment	A binary feature that takes the value one if, and only if, the instance of debt collection was paid in full

As mentioned in section 3.3, the Kernel SHAP extension is only required if dependencies exist within the initially assumed independent variables. To explore this, we plotted a correlation matrix of the features present in the nine-feature dataset in Figure 3.

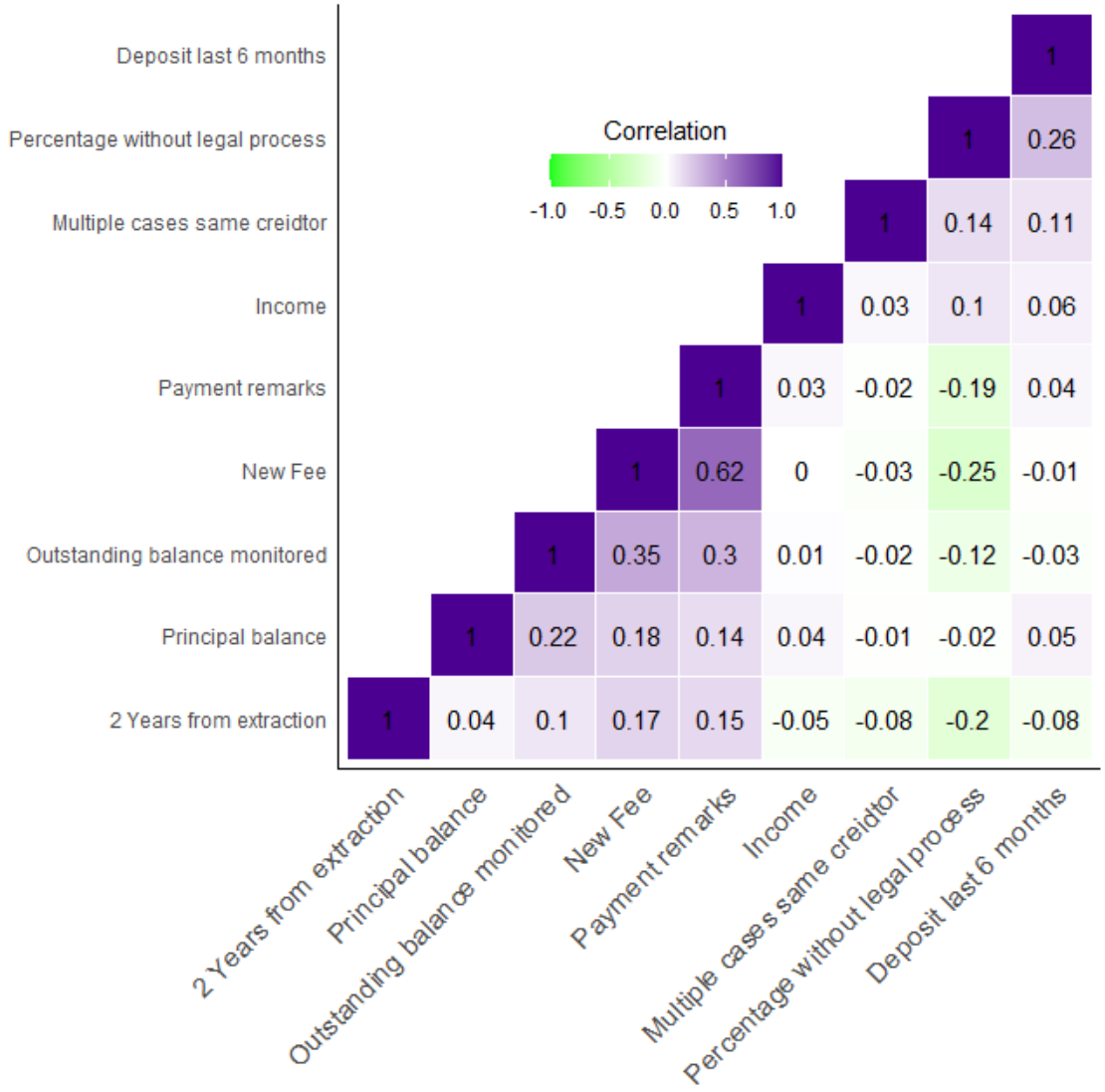

Figure 3 - Nine-feature correlation matrix

By analysing the figure above, the highest cases of correlation are observed between the features *New Fee* & *Payment remarks* and *New Fee* & *Outstanding balance monitored*. The only instance of total feature independence is observed between *New Fee* and *Income*. The consensus regarding correlation in the data aligns with the opinion of Aas, Jullum & Løland (2021) that the assumption of feature independence taken by Lee and Lundberg (2017) does not accurately portray the real world. The extension of Kernel SHAP is therefore warranted when working with this data. To understand which feature distribution is most fitting for our data, we must inspect the histograms of all nine features. As explained earlier, the Kernel SHAP extension can choose from four distributions: Gaussian, Gaussian Copula, empirical and a combined approach. Figure 4 shows the nine feature distributions. The feature distributions show clear signs of not resembling a standard Gaussian distribution. The existence of two binary features further supports this. Both the Gaussian and Gaussian copula methods are therefore not applicable to our data, and to ensure the best results, the model will use the empirical approach.

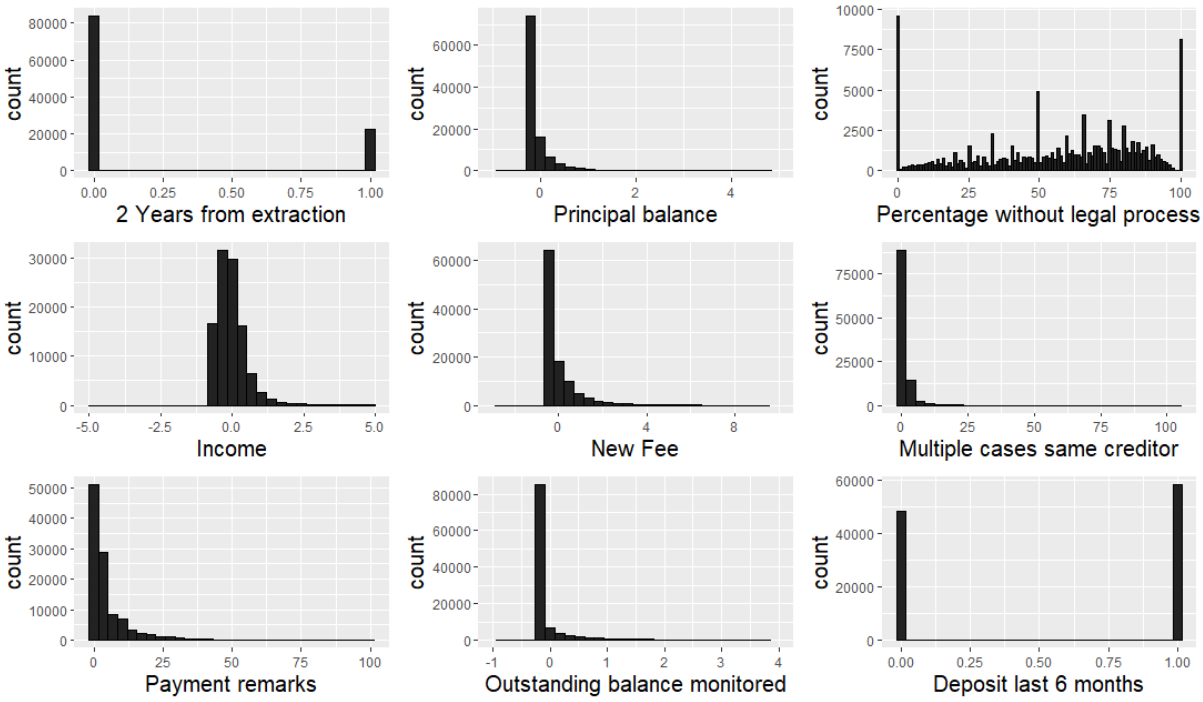

Figure 4 - Feature distribution[2]

[2] In figure 4, the most extreme outliers were omitted to produce more legible plots.

Another key aspect of descriptive statistics is the non-distributed gain within the dataset. As explained in the section regarding Shapley values, the non-distributed gain is the predicted outcome before any feature values are considered. By observing the mean of the feature *Full Payment*, we find a 54.7% chance that any given subject will pay their debt in full.

Table 4 - Summary statistics

Variable	***Mean***	***Trimmed mean***	***SD***	***Min***	***1. QT (25%)***	***3. QT (75%)***	***MAX***
2 Years from extraction	0.213	0.140	0.409	0	0	0	1
Principal balance	0	-0.144	1	-0.241	-0.233	-0.053	185.128
Percentage without legel process	55.603	57.129	30.214	0	33	80	100
New Fee	0	-0.227	1	-0.486	-0.486	0.054	23.746
Multiple cases same creditor	0.919	0.368	3	0	0	1	104
Income	0.00	-0.093	1	-0.703	-0.398	0.205	125.727
Payment remarks	4.615	2.935	7.415	0	0	6	100
Outstanding balance monitored	0	-0.152	1	-0.200	-0.192	-0.155	150.732
Deposit last 6 months	0.546	0.556	0.498	0	0	1	1
Full payment	0.547	0.558	0.498	0	0	1	1

We would also like to draw attention to the max values within the dataset. As GDPR Recital (71)(6) states, it is the job of the controller to ensure minimal amounts of error in the data. Outliers are a part of the real world and removing said outliers could introduce error by training a model under the assumption that no subject has extreme feature values. Due to high maximum values, we suspect that outliers may interfere with the summary statistics. We, therefore, compute the trimmed mean of features, trimming ten per cent from each end. The average subject has a noticeable reduction in all features excluding *Deposit last 6 months* and *Full Payment*. This indicates that the ten per cent trimmed from the high end of features values had a more significant impact on mean values, highlighting that outliers stem from having substantially higher, not lower, feature values.

4.3.4 Nine-feature model

Due to the requirements set forth by GDPR Article 5(1)c and 25(2) in combination with the requirements specified in GDPR Recital (71), there is a trade-off between the amount of data employed and the predictive power of the model. By reducing the original dataset to nine features, we argue that all features are critical to the model and contain meaningful information. The final step of our modelling process is to secure that the predictive power still satisfies the "minimisation of error" requirement described in GDPR Recital (71). This model will be trained using the same hyperparameters as shown in Table 1.

When reviewing the cross-validated AUC in Figure 5, there is an expected drop, where the new model produces a test AUC of 88.8%. By removing 60 features, the test-AUC is reduced by less than two per cent, further strengthening our assumption that the initial 69 feature model contained non-relevant and unnecessary data. The test-AUC appears to flatten out at approximately 4.000 iterations. Still, there is no reason to believe overfitting has occurred, as the early stopping rounds argument does not stop the model early. The new optimal threshold for classification is at 51% probability, which produces a normalised Matthews Correlation Coefficient of 0.80, with an accuracy of 80.67%. Although the results in all three categories are slightly weaker than the model trained on all 69 features, we argue that the demands to minimise the risk of error, set forth by GDPR Recital (71), are still met.

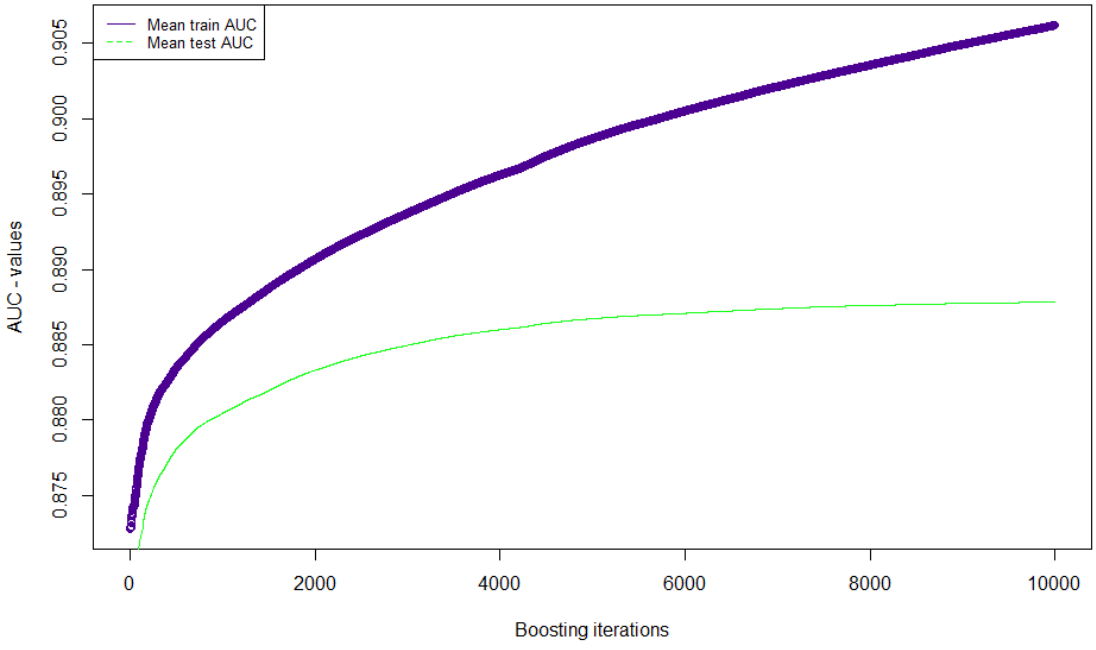

Figure 5 - Nine-feature AUC performance

5. Results

As proven by the evaluation above, the nine-feature model performs to a satisfactory level. Based on the model evaluation we can confidently trust the estimated Shapley values and their effect on the prediction outcome. We will be using the R-package *shapr* (Jullum, et al., 2021) to compute Shapley values for local interpretations and the package SHAPforxgboost (Liu, 2019) in tandem with *shapr* when computing the global interpretations.

5.1 Local interpretation

Before computing local interpretations by means of *shapr*, some additional actions need to be performed. As discussed in the literature review, interpretable machine learning can be performed using model-specific or model agnostic methods. Shapley values for prediction explanation were created as a uniform way of explaining model output and is thus a model agnostic explanation method. Therefore, we need to create an explainer object, normalising the explanation output and enabling the utilisation of model agnostic methods.

The next step is to define the non-distributed gain in our model. As explained in section 3.1 regarding Shapley Values, the non-distributed gain was defined as the prediction when no features are considered and equals the global mean prediction. For our nine-feature prediction model, the global mean equals 0.547 rounded to three decimals as shown in section 4.3.3. Finally, the distribution parameter needs to be specified. Referring to the feature distributions highlighted in Figure 4, we deemed it necessary to use the empirical approach due to the nature and distribution of our features. Below is the local explanation of a randomly picked subject within the testing set.

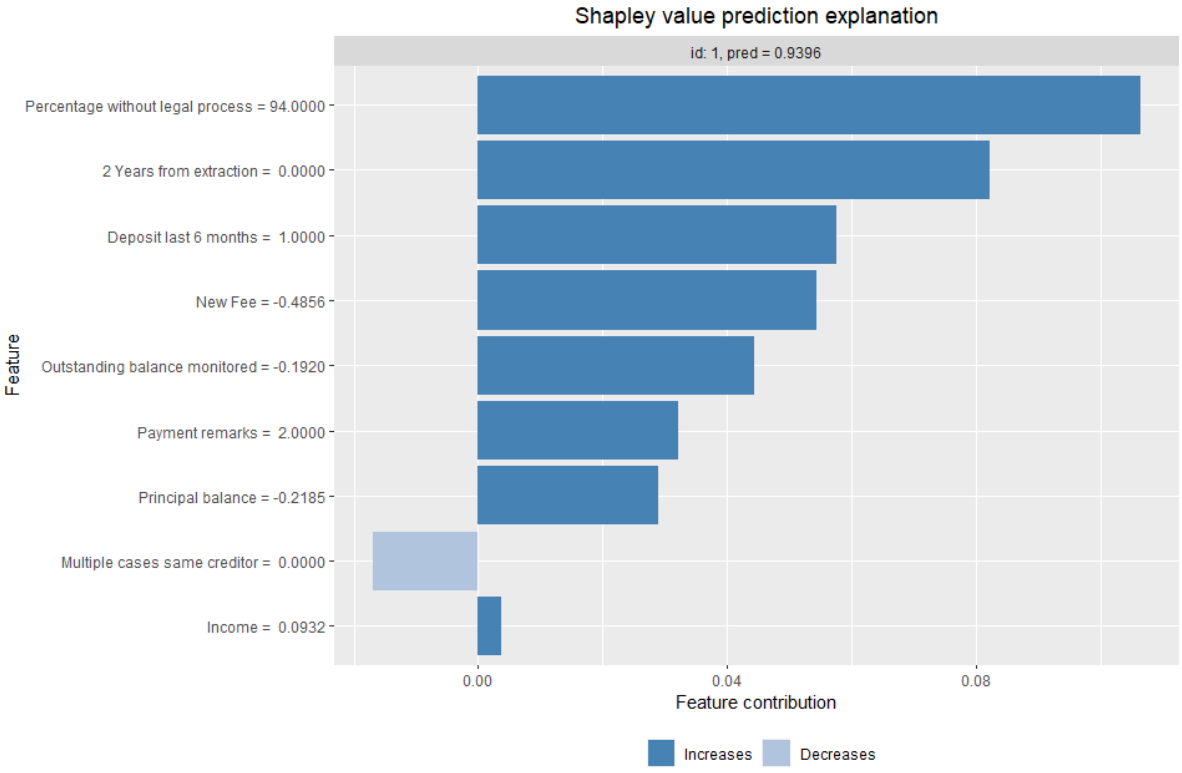

Figure 6 - Local explanation example

As displayed, the subject is given a probability of paying the debt collector of 93.96%. The subject is predicted to pay their debt most likely. All feature values indicate a high willingness to pay, except *multiple cases same creditor*. The prior willingness of the debtor to pay without Intrum needing to use legal process and the age of the case contribute the most to this prediction. Without understanding the finer details of black-box machine learning models, the subject can visually inspect how their feature values impact the prediction outcome.

5.2 Global interpretation

In the plot above, the subject can visually inspect their prediction probability, and according to Michaels Correlation Coefficient and the model accuracy, the prediction should be trusted. However, is this enough to constitute the *meaningful information* referred to in GDPR Articles 13-15? Intuition tells us that subjects would want to compare their prediction up against other similar predictions.

We then return to the absolute mean SHAP values used for feature importance to better gauge how the nine-feature model weights the different features when predicting the outcome on a global level.

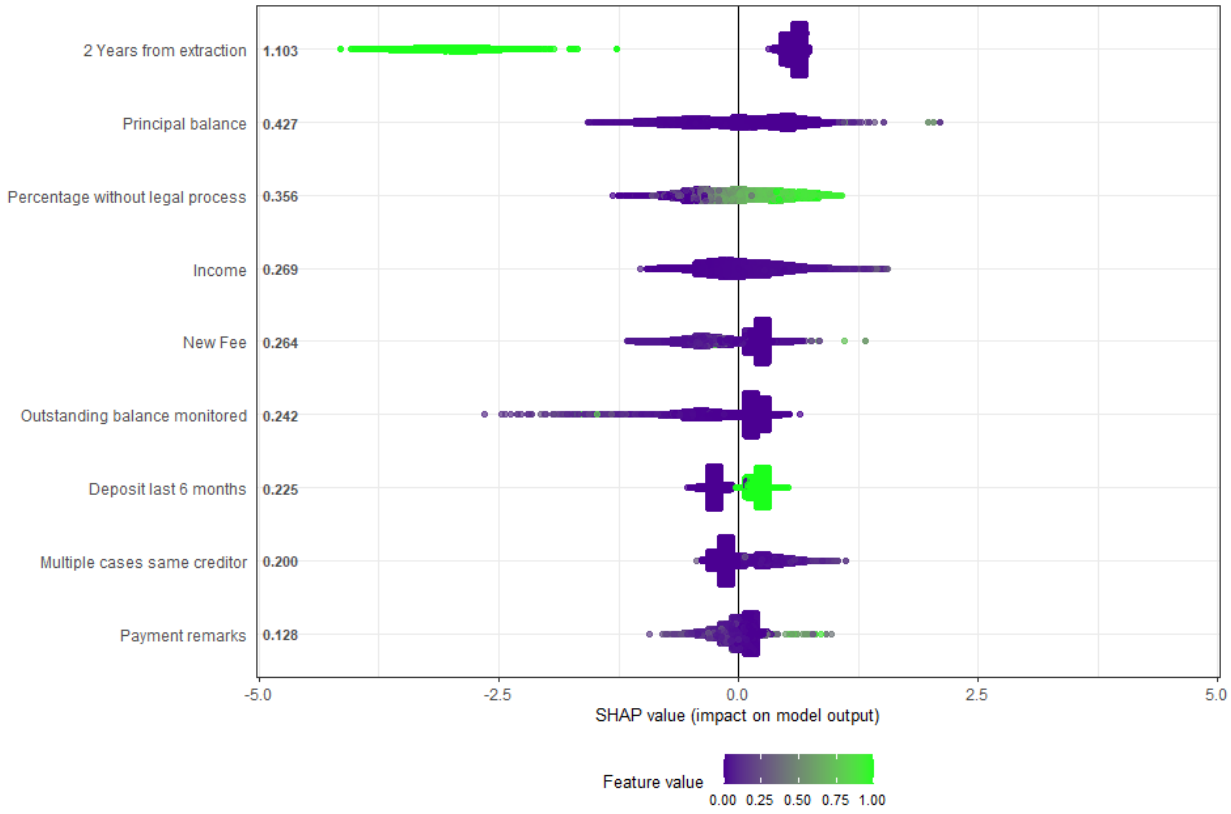

Figure 7 - Global feature importance

The plot presented above in Figure 7 displays the feature importance of the nine-feature model. This time we have also plotted the SHAP value impact each feature has on a global scale. As in section 4.3.2, the variable importance was calculated using the Kernel SHAP (Lee & Lundberg, 2017) method and used log-odds instead of probabilities. Due to this, we cannot conclude how much each feature, on average, affects the model, but we can still gain a high-level understanding of how feature values affect the prediction outcome. The plot shows different positive and negative value feature-values for the binary features *2 Years from extraction* and *Deposit last 6 months*. Whereas having paid in the last six months indicates a positive SHAP value impact (meaning that the probability of the debt collector being paid increases), negative feature values show a distinct negative trend. Some features also display no clear sign as to what values constitute positive and negative impacts. Using purely overall global feature importance might not be a solution without exceptions to understand which

features are important to predictions. We suspect that the features, and their importance, are different for high and low probability predictions.

To investigate our suspicion presented above, we will be sampling 25 random subjects from the predictions deemed not to pay and the predictions deemed to pay according to the optimal threshold found in 4.3.4 and analysing whether there are categorical differences in how the model predicts these two groups. We will be starting with the low probability predictions.

5.2.1 Low probability predictions

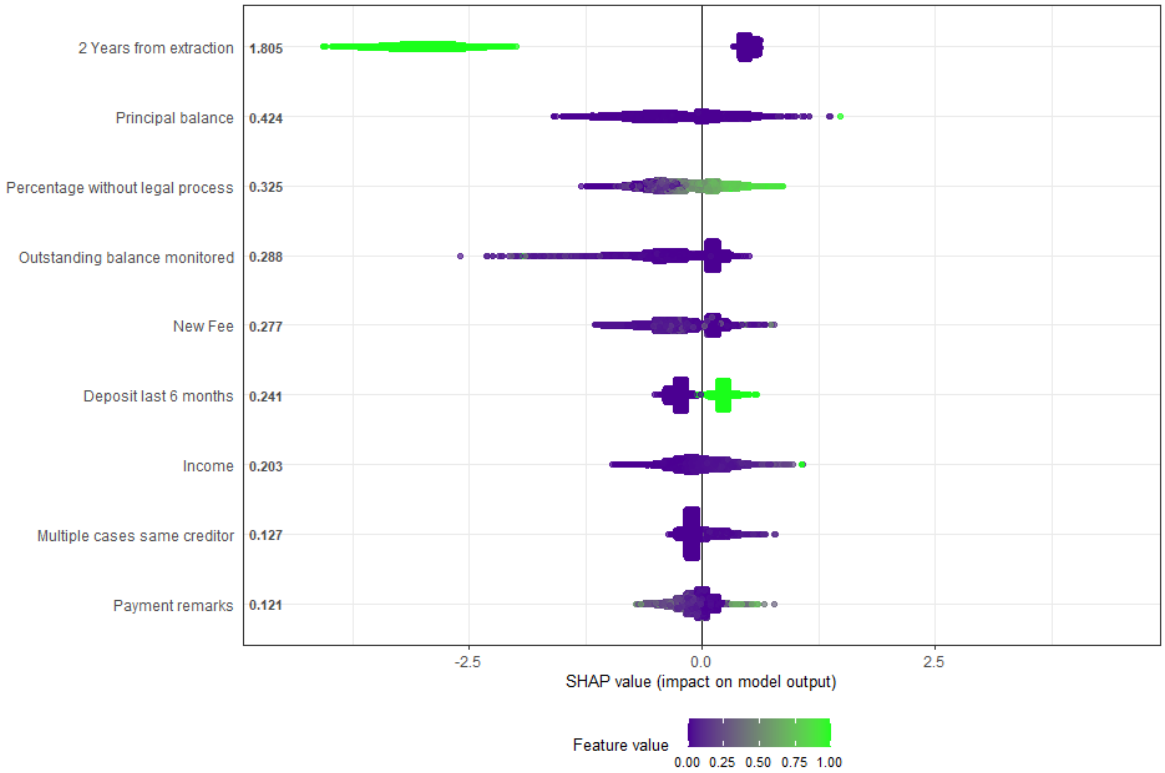

Figure 8 - Low probability feature importance

Figure 8 - Low probability feature importanceillustrates how feature values along the y-axis have changed, both in value and in order of importance. The importance of the feature *2 Years from extraction* has changed from 1.103 to 1.805, highlighting how the model emphasises this feature to a higher degree when predicting subjects that will not pay. The importance of *Income* has also shifted, now being weighed less than *Outstanding balance monitored*, *New Fee* and *Deposit last 6 months* in contrast to Figure 7.

As mentioned in section 3.1.2, computing Shapley values are a computationally expensive procedure and calculating Shapley values for tens of thousands of subjects to gauge how each feature on average affects the prediction is not feasible. However, it is possible to do this without needing too much computational power on the randomly selected sample of 25 subjects. The result of mean Shapley values can be seen in the plot below. We observe that *2 Years from extraction*, on average, reduces the probability of a subject paying with over 17%, with *Principal balance* and *New Fee* reducing that probability further by roughly 12% combined. An interesting note to take away is that *Income*, despite being viewed almost twice as much as *Payment remarks*, contributes comparatively equal amounts to the average prediction.

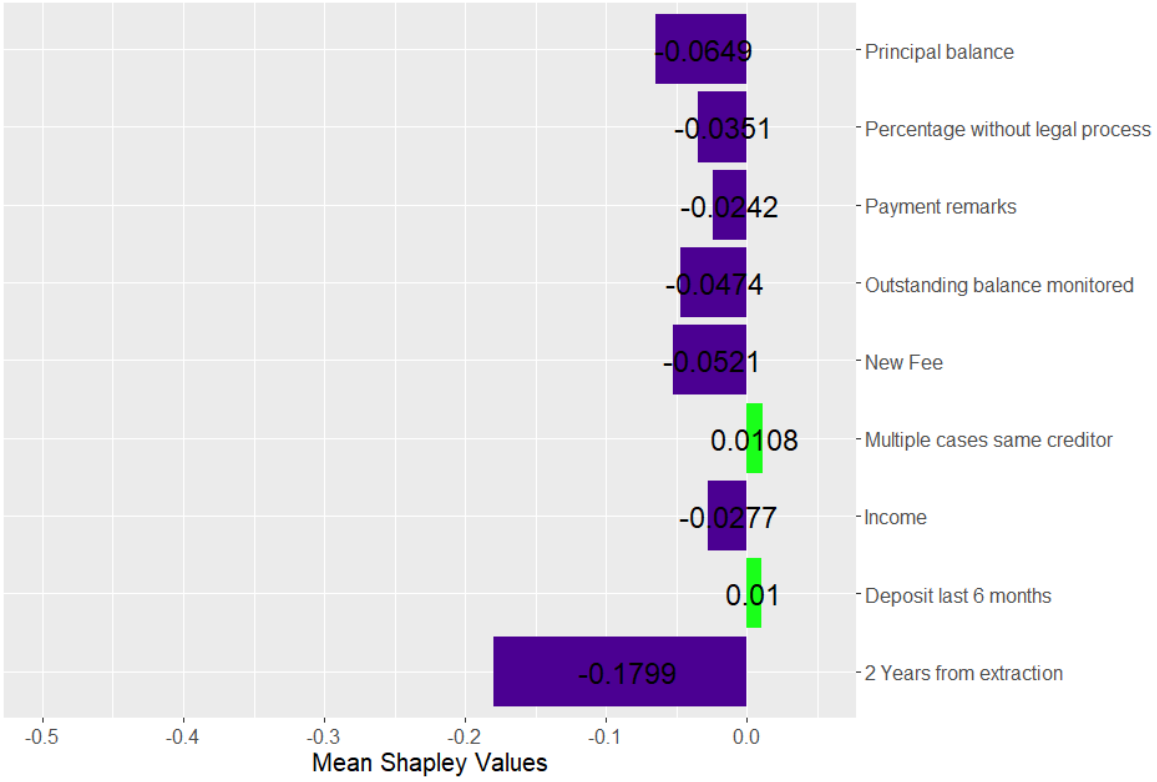

Figure 9 - Average Shapley values low probability predictions

5.2.2 High probability predictions

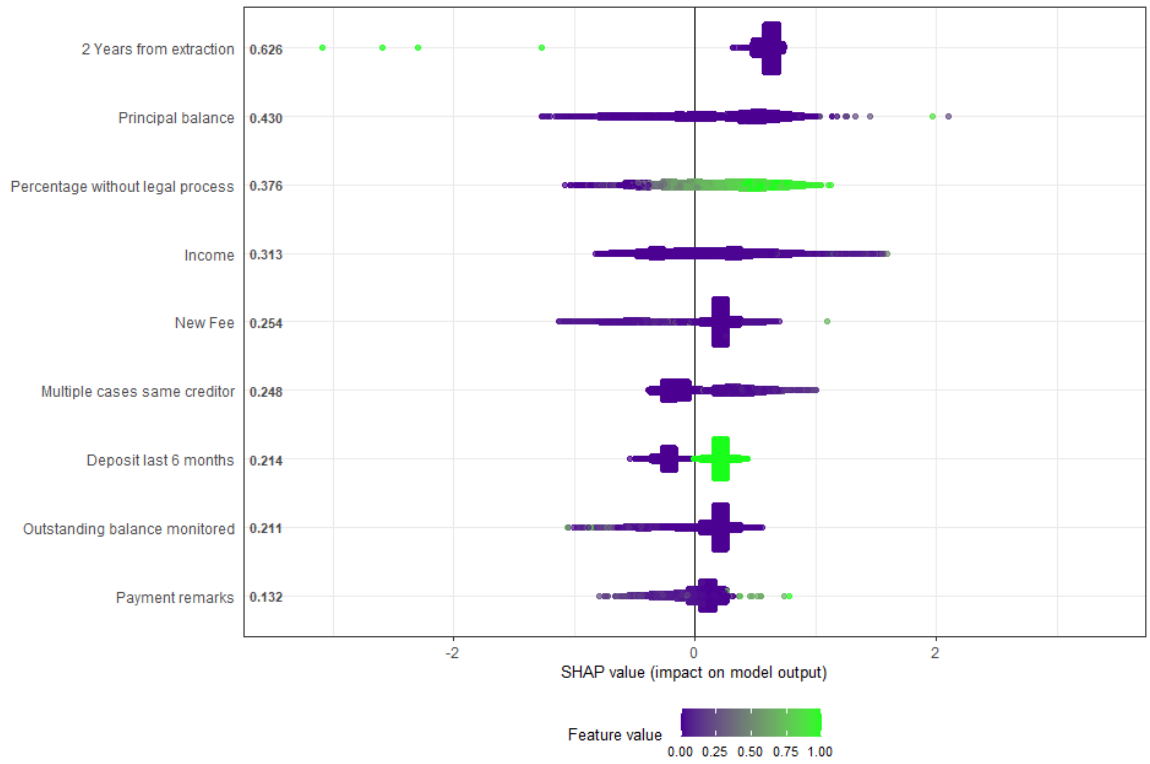

Figure 10 - High probability feature importance

Compared to the subjects predicted not to pay, the feature values of positive predictions appear mostly equal, with a few exceptions. The most obvious difference is the change in the importance of the feature *2 Years from extraction* and the extremely one-sided distribution for positive predictions. From having feature importance of 1.003 in global importance, 1.805 for low probability predictions, and 0.626 for high probability predictions, this feature interacts very differently for positive and negative predictions. *Income* is for high probability predictions weighted the same as the global feature importance. This is distinctly different from the low probability predictions, with a feature importance value more than doubled from 0.203 for low probability to 0.313 for high probability predictions.

When computing the Shapley values and taking the mean for the 25 randomly selected subjects with positive predictions, we can observe distinct differences in Shapley values for several features. By looking only at *2 Years from extraction*, the average Shapley value has had a net change on the predicted probability of almost 30%. Whether the subject has made a deposit in the last six months and the fraction of cases not requiring legal process round out the three most influential features.

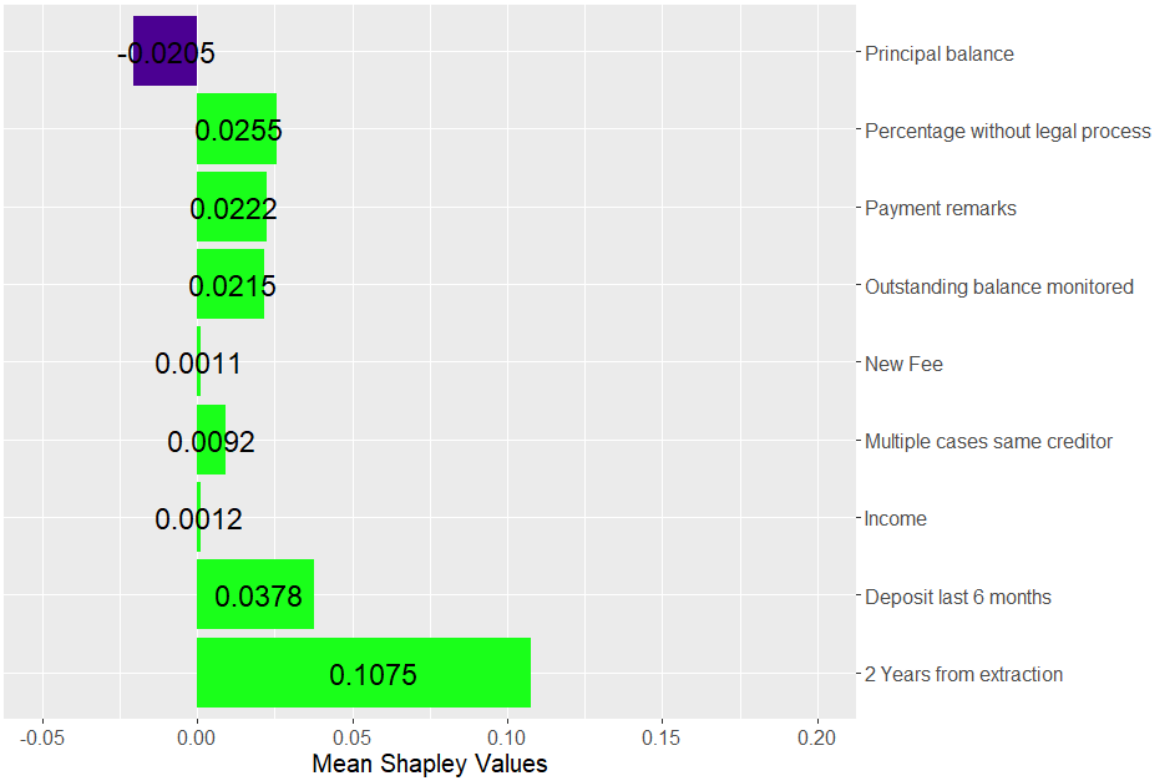

Figure 11 - Average Shapley values high probability predictions

It has become clear from comparing the explanations of positive and negative predictions regarding feature importance and feature contribution that there are differences in how the model predicts low probability and high probability outcomes. This supports our suspicion and highlights the need for producing two distinct global explanations to enable subjects to compare and understand their own predicted probability in comparison to others.

6. Discussion

Neither machine learning nor AI is explicitly mentioned in the GDPR. Still, some articles can be interpreted with broad or ambiguous wording, and multiple articles must be addressed for the model and the explanation to be compliant. The implication of the GDPR on AI and machine learning has been discussed and interpreted by the European Parliament Research Service (EPRS). They delve deeper into how AI under the GDPR can still be permitted. Their interpretation will create the basis for which we analyse the compliance of our model and Shapley values in regards to the appropriate GDPR articles.

One major aspect of the GDPR is the safety of the subject and their data. This is true for all stages of the machine learning process, especially when storing or collecting data. This thesis and discussion take for granted that the data provided to us is collected and stored correctly to comply with the GDPR.

6.1.1 Consent

While it is assumed that all data provided for this thesis was collected and stored with proper consent, it is important to specify that according to the GDPR Article 4(11):

> “'consent' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her;” (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 4(11))

Recital (32) further explains the scope of consent:

> "Consent should cover all processing activities carried out for the same purpose or purposes. When the processing has multiple purposes, consent should be given for all of them” (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Recital (32)).

In regard to our thesis, this could be interpreted in such a way that the debt collector needs explicit and separate consent from the subject to collect, store and process their data.

There are multiple criticisms regarding the definitions of consent, but there are three main issues that must be discussed regarding machine learning models and AI. Issue number one is the specificity of the consent. By the former definition, there needs to be a separate consent for each specific data usage. This issue is, however, partly solved through Recital (33), which states:

- "It is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research. Data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Recital (33)).

Recital (33) does, however, specify that this consent within an area is only regarding "scientific research purposes". Recital (159) further explains that:

> "scientific research purposes should be interpreted in a broad manner including for example technological development and demonstration, fundamental research, applied research and privately funded research. In addition, it should take into account the Union's objective under Article 179(1) TFEU of achieving a European Research Area" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Recital (159)).

Based on this interpretation, it is hard to argue that building a machine learning model in the field of debt collection and including the data of the subject within a training set could be considered "scientific research". This is also true for employing said model on personal data regarding a subject, which would require separate, explicit consent for both actions.

Issue number two regards the granularity of the consent. This is explained in GDPR Recital (43), which states:

> "Consent is presumed not to be freely given if it does not allow separate consent to be given to different personal data processing operations despite it being appropriate in the individual case" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Recital (43)).

Regarding AI and machine learning, this can be interpreted in such a way that consent to one form of data processing does not necessarily enable the data to be used for other purposes. For this thesis, such an interpretation would mean that if a subject gives consent to be reviewed by our model, it would not automatically be acceptable to utilise the data in another application. It could be argued that this also entails the data not being included in the training set.

The final issue with consent is evaluating whether the consent was freely given. GDPR Recital (42) specifies that:

> "Consent should not be regarded as freely given if the data subject has no genuine or free choice or is unable to refuse or withdraw consent without detriment" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Recital (42).

Furthermore, Recital (43) expands on this criterion with:

> "In order to ensure that consent is freely given, consent should not provide a valid legal ground for the processing of personal data in a specific case where there is a clear imbalance between the data subject and the controller, in particular where the controller is a public authority" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Recital (43)).

Based on this, the debt collector would be unable to utilise force or nudging in any way to acquire consent. It is possible to argue that there exists a certain "imbalance between the data subject and the controller". This could, in turn, entail that the standards for acceptable consent should be raised in order to further protect the subject.

Alternatives to consent

Still, consent is not the only legal base available to allow personal data processing. GDPR Article 6(1) clarifies these options in more detail, and there are three main legal bases that can be called upon. GDPR Article 6(1)(a) is the aforementioned consent, GDPR Article 6(1)(b-e) describes different situations in which processing is a necessity, and GDPR Article 6(1)(f) details the legitimate interests pursued by the controller or by a third party.

Article 6(1)(f) could be utilised when including data pertaining to a subject within a training set to build a model if the requirement "except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 6(1)(f)) is met. This would not be the case when utilising personal data as input since such an analysis might not be in the best interest of the subject. In these cases, GDPR Article 6(1)(a) must be applied, where consent and an option to opt-out would be needed (EPRS, 2020).

6.1.2 Personal data

Personal data is defined as: "information relating to an identified or identifiable natural person ("data subject")" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 4(1)). This definition is further extrapolated in Recital (26), where the problem of identifiability is addressed, where it states: "Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person". (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Recital (26)).

This definition was then commented on by the Court of Justice of the European Union and the Advocate General in Joint Cases C-141 and 372/12. (Y.S v Minister voor Immigratie, 2014) By their ruling, only the input data or the data concerning the subject as well as the conclusion of the analysis should be regarded as personal data. By this definition, it can be argued that Shapley values on individual subjects are not counted as "personal data". However, this was later contradicted in Case C-434/16 (Peter Nowak v Data Protection Commissioner, 2017), where comments or the reasoning leading to the conclusion were regarded as personal data.

The Article 29 Working Party later corroborated the latter definition in both Opinion 4/2017 with their broad definition of personal data and Opinion 216/679 where it was decided that in cases involving profiling, the subject has the right to access both the input data and the final or intermediate conclusions inferred from the input data (Working Party, 2018). Based on this, Shapley Values are to be regarded as personal data and should be handled appropriately.

Profiling

GDPR Article 4(2) defines profiling as:

> "'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" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 4(2)).

This definition can further be seen in the light of Article 29 Working Party, which dealt with issues related to the protection of privacy and personal data pre GDPR and their definition of profiling.

"Broadly speaking, profiling means gathering information about an individual (or group of individuals) and evaluating their characteristics or behaviour patterns in order to place them into a certain category or group, in particular, to analyse and/or make predictions about, for example, their:

- ability to perform a task;
- interests;
- or likely behaviour." (Working Party, 2018, p. 7)

This definition of profiling is quite broad, but in connection with our model, it could be argued that it classifies people on the expected ability to pay their debt or the likely behaviour of paying or not paying their debt. Based on this definition, our model and its usage should be counted as profiling, and hence all the ramifications that this entails should be identified and taken into consideration.

In automated profiling based on a machine learning model, the model predicts based on a training set consisting of personal data from a large sample of people. On the other hand, the trained algorithmic model by itself does not contain personal data. Earlier, we concluded that Shapley values are to be regarded as personal data, but "The correlations embedded in the algorithmic model are not personal data, since they apply to all individuals sharing similar characteristics" (European Commision, 2018). This means that any global explainer such as Shapley values are not to be seen as personal data, while Shapley values on individual subjects are.

When utilising an automated decision model within profiling the subject should be able to contest not only the decision based on this inference but also the inference itself. In our case, this would mean the debtor would not only be able to contest the decisions based on the result of the machine learning model but also the inference taken by the model. The one utilising the model or making decisions based on it should be able to demonstrate that the inference is reasonable (EPRS, 2020). For the inference to be regarded as reasonable, it should be upheld to three main standards:"

- Acceptability: the input data (the predictors) for the inference should be normatively acceptable as a basis for inferences concerning individuals (e.g., to the exclusion of prohibited features, such as sexual orientation);

- Relevance: the inferred information (the target) should be relevant to the purpose of the decision and normatively acceptable in that connection (e.g., ethnicity should not be inferred for the purpose of giving a loan).

- Reliability: both input data, including the training set, and the methods to process them should be accurate and statistically reliable." (EPRS, 2020, p. 41)

There seem to be no issues when comparing these criteria to our model. In the training data there are no variables that could be regarded as "prohibited data". In our opinion, all variables seem relevant for the model, further supported by the origin of our data. In the section detailing our modelling process, the choice of model and all subsequent steps were evaluated to ensure that the model is as reliable as possible while still being within reason for laypeople to understand.

6.1.3 Processing

All processing of personal data is subject to GDPR Article 5(1)(a), which states:

> "processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency');" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 5)

The criterion of transparency is further explained in GDPR Recital (58) which states that:

> "The principle of transparency requires that any information addressed to the public or to the data subject be concise, easily accessible and easy to understand, and that clear and plain language and, additionally, where appropriate, visualisation be used" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Recital (58)).

We have in this thesis focused on transparency in every choice when building the model. The transparency principle included in Article 5(1)(a) ensures that the subject should be completely aware of the processing of personal data and is further explained in GDPR Recital (39). The recital states that the subject should be made aware of all possible risks and conditions, including how to exercise their rights. To put it in context, the debt collector should, by this requirement, be able to deploy our model if they provide the subject with adequate information beforehand.

Within the "fairly" principle stated in GDPR Article 5(1)(a), the European Parliament Research Service has distinguished between two separate notions of fairness. Firstly, information fairness is linked with transparency to secure the subject from being misinformed or misled about their own data. GDPR Recital (60) further explains that the controller should provide the subject with correct and enough information regarding the usage of profiling, to allow the subject to exercise their rights, and ensure GDPR compliance.

The second notion of fairness is "Substantive fairness" and can, in theory, hinder the utilisation of the model and Shapley values within this thesis. Substantive fairness "concerns the fairness of the content of an automated inference or decision" (EPRS, 2020, p.45). Within GDPR Recital (71), it states:

> "In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Recital (71)).

In order to ensure fair procressing in our model we need to confirm that "factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised".

Estimation tolerance

Earlier, we concluded that Shapley values on the individual subject are to be regarded as personal data. Because Shapley values, by nature, are estimated, it raises an issue regarding the accuracy and error in said estimation. A possible counterargument is found in the Kernel SHAP results laid forward by Lee and Lundberg (2017), where they show that their estimations of Shapley values are robust. The GDPR was created to ensure the rights of the individual subject, and it is possible to view the word error in connection with the subject being misinformed. When the Shapley values are estimated, there is a possibility of error, and if the subject bases their decision on whether to exercise their rights on wrongful information, it can be regarded as "unfair".

There is clearly a discussion to be had on the acceptable usage of Shapley values depending on the interpretation of the word error and the expected accuracy of such explainers. It is possible to interpret Recital (71) in such a way that only the model itself is affected. "The controller should use appropriate mathematical or statistical procedures for the profiling" (GDPR Recital (71)), can be interpreted to where the profiling itself is the main part that needs to be secured from errors. An XGBoost model is, as shown in Figure 1, one of the best machine learning models for classification, and the model itself is not affected by the problem of estimation. By optimising this model, it could be argued that the factors that result in inaccuracies in personal data, which in our case are the individual Shapley values, are corrected, and the risk of errors is minimised. Depending on the interpretation and the weight laid on Shapley values being estimated, it can hinder its usage according to GDPR guidelines.

6.1.4 The right to be informed

In situations where data on a subject will be processed, GDPR Article 13 and GDPR Article 14 secure the right to information for the subject when the data is either collected from the subject themself or a third party, respectively. GDPR Article 13(2)(b-c) and GDPR Article 14(2)(c-d) in regards to our model, secure the rights of the subject to be informed about and object to processing. This is also true in cases where consent has been provided, as the subject is allowed to withdraw consent at any time as stated in GDPR Article 7(3). This must also be seen in connection with GDPR Article 15, which secures "the right to access" during the analysis process.

Automated decision-making

In connection with machine learning and automated decisions, the most applicable paragraphs are GPDR Article 13(2)(f), GDPR Article 14(2)(g) and Article 15(1)(h), which state that the subject should be provided with the following information:

> "the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject." (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679)

What constitutes *meaningful information* must be seen from the perspective of the subject, as explained in the literature review. This is also true for the explainability of a model, as explainability most often helps secure the rights of the subject. The European Parliamentary Research Service brings forth the idea of ex-ante and ex-post explainability. These are not explicitly stated in the GDPR, but they are brought forth as the "core of current research on explainability" (EPRS, 2020, p.54).

Different approaches to explainability

One approach to explainability is the social aspect, where explainability is focused and tailored to the subject. This also supports the notion that the GDPR and *meaningful information* should be interpreted in relation to the subject and their perspective. Social scientists focus on making the explanations understandable for laypeople and introduce four ideas they believe to be critical. These include:

> "**Contrastive explanation**: specifying what input values made a difference, determining the adoption of a certain decision (e.g., refusing a loan) rather than possible alternatives (granting the loan);
>
> **Selective explanation**: focusing on those factors that are most relevant according to human judgement;
>
> **Causal explanation**: focusing on causes, rather than on merely statistical correlations (e.g., a refusal of a loan can be causally explained by the financial situation of the applicant, not by the kind of Facebook activity that is common for unreliable borrowers);
>
> **Social explanation**: adopting an interactive and conversational approach in which information is tailored according to the recipient's beliefs and comprehension capacities" (EPRS, 2020, p. 55).

In theory, the implementation of Shapley values accompanied with text explanation should fulfil all these ideas in an ex-post explanation. Shapley values by themselves or a plot such as in Figure 6 might not be interpretable to a layperson, but the inclusion of an explanation regarding the basic idea should make the explanation sufficient.

On the other hand, computer scientists bring forth three other aspects with a more technical and ex-ante focus. These are (1) *model explanation*, which focus on the explanation of the model and the logic of the opaque system involved. (2) *Model inspection* focuses on the ability to inspect the properties of the opaque model where the patterns or sensitivity involved is represented. (3) *Outcome explanation*, where the outcome of the model is explained in terms of the choices that led to the exact outcome, which could include the specification on which input value fell below a threshold or similar (EPRS, 2020, p. 54-55).

Depending on the interpretation, these explanations could be fulfilled with the correct utilisation of Shapley values. Still, as the European Parliamentary Research Service states, these explanations are "intended for technological experts and assume ample access to the system being explained" (EPRS, 2020, p. 55). This would contradict the idea that the explanations should be understandable and aimed towards the subject themselves, and hence they are not directly analysed in this thesis.

All the ideas for explainable AI and machine learning stated above are, as already mentioned, not included in the GDPR itself. Still, their inclusion in the report by the European Parliamentary Research Service indicates that they might be important for the interpretation of the GDPR or serve as guidelines for future additions made to the GDPR.

Decision-based modelling

The most significant piece in determining if our model explained with Shapley values can be regarded as legal and in compliance with the GDPR is GDPR Article 22. GDPR Article 22(1) states:

> "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 22).

This paragraph should not be interpreted as a right for the subject but rather as a constraint on the processing part. Within this paragraph, there are four conditions that are all required for the application of Article 22.

1. A decision has to have been taken
2. The decision is based solely on automated processing
3. The automated processing must include profiling
4. The decision must have legal or significant effect (EPRS, 2020)

The first condition is met by utilising the model to decide which subjects are pursued, and which subjects show no indication of being profitable. The third condition is also met through the definition of profiling, as explained earlier.

The second and fourth conditions are not necessarily as straightforward. The second requirement depends entirely on how the debt collector utilises our model. If the model and its results are only used as a guiding tool for the debt collector and the final decision is made by a person, this condition is not met. This human involvement and its degree are further specified in Working Party 251/2017, and it states:

> "To qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision" (Working Party, 2018).

In connection with our model, it is entirely possible for the debt collector to employ the model in such a way that meets this requirement. Condition four can also be seen as met, based on the fact that a debt collector has multiple means to collect on the money that the subject owes. These means, such as disbursements in either housing or wages, or forced sale of a house or vehicle, can for many be seen as a decision with "significant effect". In GDPR Recital (71), there are explicit examples mentioned:

> "such as automatic refusal of an online credit application or e-recruiting practices without any human intervention" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Recital (71)).

These examples fall in line with the means available to the debt collector. However, we believe it might be possible to combine conditions two and four to optimise the utilisation of the model depending on the interpretation of the article. For example, if the model is utilised and the resulting prediction states that the chance of full payment is negligible, the effect of applying the aforementioned means are equally negligible and advised against. If none of the means are deployed, it can hardly be seen as a "significant effect" on the subject, and for these subjects, the decision can be automated since the fourth condition is not met. On the other hand, for the subjects that require possible additional actions from the part of the debt collector, the model can be used as guidance. The debt collector would then review the individual case in order for the decision to not be solely based on an automated process, thereby not satisfying condition number two. This approach can partly automate the debt collection process and cut down on total processing time. We would however like to emphasize that such an approach would need

to be further investigated by the proper authorities to ensure that the rights of the subject are being protected.

The constraint stated in GDPR Article 22(1) does not necessarily stop all usage of automated decision models and can be waivered if any of the requirements stated in GDPR Article 22(2) is present.

> "a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
>
> b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
>
> c) is based on the data subject's explicit consent." (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 22(2))

Depending on the existence of a contract between the debt collector and the debtor, it could be possible to employ point a), but this would also require the use to be "necessary". If there is a large number of cases to be reviewed, it could be categorised as necessary to reduce time, cost or the existence of bias introduced by a human reviewer. In most cases, point c) and consent would be necessary if others were not specified.

The rights of the subject

GDPR Article 22(3) next explicitly mentions Article 22(2) a) and c), and in these cases, there is a significant requirement for the security of the rights and freedoms of the subject. It is also specified:

> "at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 22(3))

This would mean that if the debt collector employed the model, they would have to secure the ability of the subject to understand the results and contest the inference of the model, including the decisions based on these inferences as explained earlier.

GDPR Article 22(4) sets another criterion that the model shall not include, or the decision be based on "special data categories of personal data". These categories are specified in GDPR Article 9(1):

> "Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 9).

When reviewing the explainer variables included in our model, there is no need to utilise GDPR Article 22(4) since none of the variables seems to be connected to the data categories mentioned above. However, it can be argued that the inclusion of *Income* in the dataset can lead to a discriminatory or biased training set. This is due to some areas having a higher average earning and that immigrants or women earn less in the general population (Statistisk Sentralbyrå, 2020b) (Statistisk Sentralbyrå, 2020a). If the dataset were to be proven biased, it is possible to doubt the safeguarding of "the data subject's rights and freedoms and legitimate interests" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 22(3)). If this were proven to be accurate it is possible to argue against the model being employed.

Requirments of explainability

To secure the legitimate interest of the subject and the ability to contest any decision based on *meaningful information* in the eyes of the subject, as specified in GDPR Article 13-15, the European Parliamentary Research Service brings forth the question of explainability. Their analysis combines GDPR Articles 13, 14 and 15 with GDPR Article 22 and GDPR Recital (71). By combining these references, they interpret the requirements of

> (1) "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 13-15),

(2) “the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision” (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679 Article 22) and

(3) the right “to obtain an explanation of the decision reached after such assessment and to challenge the decision” (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Recital (71)).

These requirements are further reviewed into a list of nine possible requirements for explainability:

“1. information on the existence of profiling, i.e., on the fact that the data subject will be profiled or is already being profiled;

2. general information on the purposes of the profiling and decision-making;

3. general information on the kind of approach and technology that is adopted;

4. general information on what inputs factors (predictors) and outcomes (targets/predictions), of what categories are being considered;

5. general information on the relative importance of such input factors in determining the outcomes;

6. specific information on what data have been collected about the data subject and used for profiling him or her;

7. specific information on what values for the features of the data subject determined the outcome concerning him or her;

8. specific information on what data have been inferred about the data subject;

9. specific information on the inference process through which certain values for the features of the data subject have determined a certain outcome concerning him or her” (EPRS, 2020, p. 64-65)

They clarify that due to the different ways automated decision models can be implemented, it is hard to specify exact “fit-all” requirements. However, we will try to review our model within these requirements based on their interpretation of the GDPR.

Requirements 1-5 are regarded as explanations demanded before or ex-ante of the processing of data concerning the subject. (1) The debt collector must ensure that the subject understands that they are being profiled based on the data provided. (2) The debt collector must also explain the purpose of said profiling and the possible effects such processing can have on the decision.

Requirement (3) is more challenging to interpret, but it is possible to understand it in such a way that a layman explanation of XGBoost and machine learning, in general, is required. Again, The European Parliamentary Research Service brings up the example of explaining that the model "may be inappropriate or lead to errors and biases" (EPRS, 2020, p. 65). This could mean that the debt collector would have to explain possible biases or errors in the dataset and possible overfitting and other problems with such models.

Requirement (4) would demand that the debt collector provide a general explanation of the variables and categories considered for the model and possible target or predictor variables.

By utilising absolute mean Shapley values to discern global feature importance, it can be argued that requirement (5) is also met. This is because the absolute mean values provide the "relative importance". How much each variable affects the prediction can be understood by examining feature importance shown in Figure 7.

As a continuation of requirement (4), requirement (6) demands more specific information on the features included in the model, described in Table 3. Requirement 6-9 is classified as ex-post explanations, and this would entail that a more detailed explanation would only be required after the processing has been performed.

Individual Shapley values could be regarded as sufficient to meet requirement (7). By interpreting and explaining the individual Shapley values, the subject would gain insight into how each value in the data provided affects the prediction produced by the model. As an extension to this, requirement (8) would require the debt collector to give specific information on the target variable *Full payment* and what the model has inferred from the data provided.

Requirement (9) does not necessarily affect our model, depending on the interpretation. The European Parliamentary Research Service mentions that "information about (9) might also be provided, though information on (7) and (8) should generally be sufficient to provide adequate individualised explanations" (EPRS, 2020, p. 65). In other words, informing the subject about

the information specified in requirements (7) and (8) should suffice to comply with the ex-post demands.

All of the articles in the GDPR mentioned within this thesis and in connection with AI and machine learning are supported by GDPR Article 24. Article 24 specifies the responsibilities of the controller or, in our case, the debt collector and states:

> "the controller shall implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation. Those measures shall be reviewed and updated where necessary". (EU General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, Article 24)

In layman terms, this requires the debt collector to review their model and the whole machine learning process whenever "necessary" and update the model if any aspect is found to not comply with the GDPR. In terms of machine learning in general, and our model specifically, this would include:" the adequacy and completeness of training sets, over reasonableness of the inferences, over the existence of causes of bias and unfairness" (EPRS, 2020, p. 67)

7. Conclusion

This thesis has attempted to answer the question: *are Shapley values an appropriate framework to adequately explain predictions and provide subjects with the relevant information needed to satisfy the demands in the GDPR*? Before a prediction model could be trained, there was a need to ensure that all data presented to the model was compliant with the GDPR. We reviewed the dataset and removed irrelevant and unnecessary elements according to Articles 5(1)c and 25(2) concerning data minimisation. The dataset was further reduced by utilising absolute mean Shapley values to discern the features that provided *meaningful information* in reference to GDPR Articles 13-15. This highlighted how the application of Shapley values in the collection and processing stages of a machine learning model enables it to comply with the GDPR.

We argue that machine learning models are primarily impacted by four different aspects of the GDPR: *consent*, *personal data*, *processing*, and the *right to be informed*. Although this thesis assumed that proper consent was acquired, it is a highly relevant aspect concerning machine learning models and we highlighted that Shapley values require consent from the subject to be utilised. We also concluded that Shapley values calculated for individuals should be regarded as personal data. The ramifications of this classification are increased demands when deploying Shapley values in regard to profiling. We still argue that models built with and explained using Shapley values are compliant with the GDPR.

The estimated nature of Shapley values could theoretically be a concern due to the possibility of inaccuracy and error in the estimation. Still, we argue that the leading classification performance of XGBoost in combination with the robust results produced by *shapr* proves the reliability of the method and that the risk of error is sufficiently minimised, complying with GDPR Recital (71). We also argue that Shapley values provide good enough explanations to comply with the nine requirements extrapolated from the GDPR Articles 13-15, 22 and Recital (71) by the European Parliamentary Research Service.

We have throughout this thesis proven how Shapley values abide by the requirements set forth by the relevant aspects of the GDPR. Based on this we believe Shapley values adequately explain predictions and provide subjects with the relevant information needed to satisfy the demands in the GDPR.

Further work

Lastly, we wish to highlight the potential future work that can be done with our model and Shapley values. At its current stage the model only predicts whether a subject will pay in full, not through which means this payment is procured. Creating an even more complex model that also seeks to understand through which means the subject will pay is an interesting way to further test the capabilities of Shapley values and its correlation to the GDPR. The more complex the model becomes, the more important it becomes for the explanation method to be robust and easy to understand. It would be interesting to investigate if there are current limitations with Shapley values that only presents itself on more complex models and understanding this could be vital in pushing interpretable machine learning even further.

References

Aas, K., Jullum, M., & Løland, A. (2021, September). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence Volume 298*.

Allinson, M. (2021, July 22). *How has Data Become the World's Most Valuable Commodity?* Retrieved from Robotics&Automation news: https://roboticsandautomationnews.com/2021/07/22/how-has-data-become-the-worlds-most-valuable-commodity/44267/

Bhandari, A. (2020, June 16). *AUC-ROC Curve in Machine Learning Clearly Explained*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Bhandari, A. (2020, March 20). *What is Multicollinearity? Here's Everything You Need to Know*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/

Bryan, J., Hester, J., & Wickham, H. (2021). *readr: Read Rectangualr Text Data*. Retrieved from https://readr.tidyverse.org

Bulao, J. (2021, December 7). *How Much Data Is Created Every Day in 2021?* Retrieved from techjury: https://techjury.net/blog/how-much-data-is-created-every-day/#gref

Burns, E. (2021). *machine learning*. Retrieved from searchenterpriseai: https://searchenterpriseai.techtarget.com/definition/machine-learning-ML

Burt, A. (2017, June 1). *Is there a 'right to explanation' for machine learning in the GDPR?* Retrieved from iapp: https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/

Chalkiadakis, G., Elkind, E., & Wooldridge, M. (2011). *Computational Aspects of Cooperative Game Theory*.

Charnes, A., Golany, B., Keane, M., & Rousseau, J. (1988). Extremal Principle Solutions of Games in Characteristic Function Form: Core, Chebychev and Shapley Value Generalizations. *Econometrics of Planning and Efficiency*, pp. 123-133.

Chawla, V. (2020, September 16). *Is More Data Always Better For Building Analytics Models?* Retrieved from Analytics India Magazine: https://analyticsindiamag.com/is-more-data-always-better-for-building-analytics-models/

Chen, et al. (2021, November 21). *xgboost: Extreme Gradient Boosting*. Retrieved from https://CRAN.R-project.org/package=xgboost

Chen, T., & Guestrin, C. (2016, June 10). XGBoost: A Scalable Tree Boosting System. DOI: 10.1145/2939672.2939785.

EPRS. (2020, 6). *The Impact of General Data Protection Regulation (GDPR) on Artifical Intelligence.* Retrieved from European Parliment: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf

EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ 2016 L 119/1 (European Parliament April 27, 2016).

European Commision. (2018, 8 22). *ec.europa.* Retrieved from Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679: https://ec.europa.eu/newsroom/article29/items/612053

European Commission. (2021, April 21). Europe fit for the Digital Age: Commission proposes new rules and actions. Brussels, Belgium.

François, R., Henry, L., Müller, K., & Wickham, H. (2021). *dplyr*. Retrieved from https://dplyr.tidyverse.org/

Gopinath, D. (2021, October 26). *The Shapley Value for ML Models*. Retrieved from towardsdatascience: https://towardsdatascience.com/the-shapley-value-for-ml-models-f1100bff78d1

Grolemund, G., & WIckham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software, 40(3)*, pp. 1-25 .

Harrell Jr, F. E., & Dupont, C. (2021, October 7). *Hmisc: Harrell Miscellaneous.* Retrieved from https://cran.r-project.org/package=Hmisc

Human Rights watch. (2021, November 10). *How the EU's Flawed Artificial Intelligence Regulation Endangers the Social Safety Net: Questions and Answers*. Retrieved from Human Rights Watch: https://www.hrw.org/news/2021/11/10/how-eus-flawed-artificial-intelligence-regulation-endangers-social-safety-net

Jullum, M., Sellereite, N., Lingjærde, C., Løland, A., Redelmeier, A., Regnesentral, N., & Wahl, J. C. (2021, January 28). *shapr: An R-package for explaining machine learning.* Retrieved from https://cran.r-project.org/package=shapr

Kassambara, A. (2019, May 19). *ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'.* Retrieved from https://cran.r-project.org/package=ggcorrplot

Kassambara, A. (2020, June 27). *ggpubr: 'ggplot2' Based Publication Ready Plots.* Retrieved from https://cran.r-project.org/web/packages/ggpubr/index.html

Khanna, C. (2020, November 29). *Multicollinearity — Why is it bad?* Retrieved from Towards data science: https://towardsdatascience.com/multicollinearity-why-is-it-bad-5335030651bf

Kumar, S. (2020, July 24). *7 Ways to Handle Missing Values in Machine Learning*. Retrieved from Towards Data Science: https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e

Lakshmanan, S. (2019, May 16). *How, When, and Why Should You Normalize/ Standardize / Rescale Your Data?* Retrieved from Towards AI: https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

Lazzaro, S. (2021, June 21). *Machine learning's rise, applications, and challanges*. Retrieved from Venturebeat: https://venturebeat.com/2021/06/21/machine-learnings-rise-applications-and-challenges

Lee, S.-I., & Lundberg, S. M. (2017, November 25). A Unified Approach to Interpreting Model. *Conference on Neural Information Processing Systems*.

Liu, Y. (2019, July 18). *SHAP for XGBoost in R: SHAPforxgboost.* Retrieved from Yang's Research Blog: https://liuyanguu.github.io/post/2019/07/18/visualization-of-shap-for-xgboost/

Lüdecke, D. (2021, October 1). *performance: Assessment of Regression Models Performance.* Retrieved from https://cran.r-project.org/package=performance

Lundberg, S. (2018, February 1). *Output value in binary classification task is outside [0, 1] range #29*. Retrieved from Github.com: https://github.com/slundberg/shap/issues/29

Mandrekar, J. N. (2015, 11 20). *Receiver Operating Characteristic Curve in Diagnostic Test Assessment*. Retrieved from Science Direct: https://www.sciencedirect.com/science/article/pii/S1556086415306043

Miller, E., Smith, D., & Wickham, H. (2021). *haven.* Retrieved from https://haven.tidyverse.org/

Molnar, C. (2021, 11 11). *Interpretable Machine Learning*. Retrieved from Github: https://christophm.github.io/interpretable-ml-book/agnostic.html

Molnar, C., Casalicchio, G., & Bischl, B. (2020). *Interpretable Machine Learning - A Brief History, State-of-the-Art and Challenges.*

Morde, V. (2019, April 8). *XGBoost Algorithm: Long May She Reign!* Retrieved from Towards Data Science: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Ohno-Machado, L., H.S, F., Ohrn, & A. (1998). Improving machine learning performance by removing redundant cases in medical data sets. *Proc AMIA Sym*.

Peter Nowak v Data Protection Commissioner, C-434/16 (CJEU (Second Chamber) 12 20, 2017).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, 6 16). *Model-Agnostic Interpretability of Machine Learning.* Retrieved from Arxiv.org: https://arxiv.org/pdf/1606.05386.pdf

Sarkar, D. (2018, May 24). *The Importance of Human Interpretable Machine Learning*. Retrieved from towardsdatascience: https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476

Schmitt, M. (2020). *Interpretable Machine Learning*. Retrieved from datarevenue: https://www.datarevenue.com/en-blog/interpretable-machine-learning

Selbst, A. D., & Powles, J. (2017, November). Meaningful information and the right to explanation . *International Data Privacy Law*, pp. 233-242.

Shapley, L. S. (1953). A value for n-person games. In L. S. Shapley, *Contributions to the Theory of Games (AM-28), Volume II* (pp. 307-318). Princeton University Press.

Shmueli, B. (2019, 11 22). *Matthews Correlation Coefficient is The Best Classification Metric You've Never Heard Of.* Retrieved from Towards Data Science: https://towardsdatascience.com/the-best-classification-metric-youve-never-heard-of-the-matthews-correlation-coefficient-3bf50a2f3e9a

Silge, J., Chow, F., & Wickham, H. (2021, November 8). *rsample: General Resampling Infrastructure.* Retrieved from https://cran.r-project.org/package=rsample

Statistisk Sentralbyrå. (2020a, 3 3). *Flyktninger i og utenfor arbeidsmarkedet 2018.* Retrieved from SSB Arbeid og Lønn: https://www.ssb.no/arbeid-og-lonn/artikler-og-publikasjoner/flyktninger-i-og-utenfor-arbeidsmarkedet-2018

Statistisk Sentralbyrå. (2020b, 11 16). *Alder og Yrke Påvirker lønnsgapet.* Retrieved from SSB arbeid og lønn: https://www.ssb.no/arbeid-og-lonn/artikler-og-publikasjoner/alder-og-yrke-pavirker-lonnsgapet

Stewart, M. (2020, March 19). *Guide to Interpretable Machine Learning*. Retrieved from towards data science: https://towardsdatascience.com/guide-to-interpretable-machine-learning-d40e8a64b6cf

Swalin, A. (2018, January 31). *How to Handle Missing Data*. Retrieved from Towards Data Science: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

The Lancet Respiratory Medicine. (2018, October 17). Opening the black box of machine learning. *The Lancet Respiratory Medicine*. doi:https://doi.org/10.1016/S2213-2600(18)30425-9

Toshniwal, R. (2020, 1 15). *Demystifying Roc Curves*. Retrieved from Towards Data Science: https://towardsdatascience.com/demystifying-roc-curves-df809474529a

Ushey, K. (2018, June 5). *RcppRoll*. Retrieved from https://cran.r-project.org/package=RcppRoll

Wallace, A. (2020, March 17). *The average person spends more than 8 hours a day using technology, study says.* Retrieved from DesertNews: https://www.deseret.com/u-s-world/2020/3/17/21184252/technology-screen-time-bed-sleep-8-hours-pbs-cnn-fox-news

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.

Wickham, et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4(43)*, p. 1686 https://doi.org/10.21105/joss.01686.

Wolford, B. (2020). *What is GDPR, the EU's new data protection law?* Retrieved from GDPR: https://gdpr.eu/what-is-gdpr/

Wood, T. (2021). *F-Score*. Retrieved from DeepAI: https://deepai.org/machine-learning-glossary-and-terms/f-score

Working Party. (2018, 2 6). *Guidelines on Automated individual decision-making and Profiling.* Retrieved from ARTICLE 29 DATA PROTECTION WORKING PARTY: https://ec.europa.eu/newsroom/article29/redirection/document/49826

Y.S v Minister voor Immigratie, C-141/12 and C-372/12 (Third Chamber of the Court of Justice of the European Union 7 17, 2014).

Young, H. (1985). Monotonic solutions of cooperative games. *International Journal of Game Theory 14*, pp. 65-72.