Erasmus School of Economics

Master Thesis

Comparison of multiple prediction models on customer conversion of financial services customers

MSc Economics and Business

Track: Data Science & Marketing Analytics

Author: Serafeim Sarafopoulos (545887)

Supervisor: Dr. N.M. (Nuno) Almeida Camacho

Abstract

This thesis goes through a comparison of machine learning models to predict conversion in financial services and uses machine learning interpretation techniques to draw information and provide insights to decision makers. The data used was provided by a financial services company in Portugal which after some data cleaning it was split into three stages: Screening Stage, Negotiation Stage and Finalization Stage. The machine learning models whose accuracy is compared are three: Random Forest, Neural network and XGBOOST. The model that shows the highest accuracy in each stage was Random Forest. After the accuracy comparison of these three models the interpretation techniques LIME, PDP and ICE are used on Random Forest on all three stages. The plots from these techniques used are explained and insights for marketing managers and business decision makers are provided.

Table of Contents

1. Introduction

The financial services sector is one of the primary drivers of a country's economy. It provides the free flow of capital and liquidity in the marketplace, supports the start and the expansion of businesses and according to the World Bank is the key to job generation and to combat poverty. The global financial services market is very great in size and it is continuing to grow. Based on www.investopedia.com, it constitutes 20-25% of the world economy. According to www.thebusinessresearchcompany.com/ and yahoo finance this market grew from $25848.74 billion in 2022 to $28115.02 billion in 2023 at a compound annual growth rate (CAGR) of 8.8% and it is expected to have a volume of US$2.50bn in 2027.

In this research the data is going to be split in three stages of the funnel which the clients have to go through until the conversion is reached. These three stages are Screening, Negotiation and Formalization. All the actions explained in the following paragraphs are going to be executed the same way in all three datasets created from the original data but each one focusing on one of the three stages mentioned. This may create more insights on conversion prediction that seem to be missing from the current literature.

There are many papers that focus on machine learning accuracy comparison but on the other hand there has been only a few research done on training machine learning techniques to predict conversion (Martínez et al., 2020), (Lee et al., 2021) and not much research on prediction of conversion in financial services. This paper is going to fill some of this gap by answering these questions:

1) Which machine learning method has the highest accuracy on predicting conversion in the financial services sector following three different stages: Screening, Negotiation and Formalization and **2) What can we learn by peeking inside the black boxes that are these machine learning techniques?**

In this paper, the first goal is going to be achieved by training machine learning methods to predict conversion and compare the accuracy of the models. The techniques that are going to be used are random forest, neural network and XGBoost. They are going to be trained on 70% of

the data and tested on the remaining 30%. Afterwards, they are going to be ranked based on the accuracy in those tests.

The second goal is going to be achieved by using black box interpretation techniques to understand what is happening inside some of the machine learning methods used as they are basically black boxes. Techniques like that are Local Surrogate (LIME) (Ribeiro et al., 2016) , Individual Conditional Expectation (ICE) and Partial Dependence Plots (PDP) (Goldstein et al., 2015). By using these techniques, managers can understand the models used to predict conversion, increase trust to the predictions made (Ribeiro et al., 2016) and explain more easily the process to the shareholders

So, after using predictive models and the current clients that are most unlikely to lead to a conversion, the marketing managers can get in touch with those customers and provide customized information and offers before making a transition to some competitor. The marketing managers could contact them to send them an customized email to inform them about a new service or a discount if they agree to stay with the company for longer (Lee, Lee, Cho, Im, & Kim, 2011). There have been studies which show that companies that used predictive models improved their revenues and profitability by hundreds of millions of US dollars (Tillmanns et al., 2017).

To conclude, this paper is going to highlight the significance of the financial sector and its role in the economy. As there is not much about predicting conversion in the financial sector, this thesis is going to fill some of the gap in the literature, show which of the trained machine learning techniques is more suitable in this case and also illustrate the value of having machine learning to predict if a person is going to become a customer and how managers could use this in their line of work. On top of all that, an explanation is going to take place on how it is possible to look into what is happening inside the used machine learning models using interpretation techniques and their results in this case.

2. Literature review

In this section a table (Table 1) is going to be shown that summarizes the existing literature on predicting conversion, machine learning methods to do so and interpretation techniques used on other cases. Afterwards, a more detailed dive in the existing literature will take place using the sources shown on the table.

***Table 1**: Existing literature*

Contributors	(Main) Methodology	Data	(Relevant) Findings
Naumzik, C., Feuerriegel, S., & Weinmann, M. (2021)	Variable-duration hidden Markov model to predict the likelihood of failure	Rating data from Yelp between January 2010 and December 2017 for restaurants listed in Phoenix, Arizona	Predicting business failures, a certain time before occurring using machine learning
Ma, L., & Sun, B. (2020)	Empirical marketing research on method, data, usage, issue, and theory	Various papers and use of existing literature	The importance of using machine learning methods in marketing research
Lemmens, A., & Gupta, S. (2020)	Profit-based loss function	1)An interactive television subscription service, provided by a firm located in continental Europe 2) subscription-based membership organization located in North America	Selecting a target size that maximizes the campaign profit leads to significantly more profitable campaign (churn focused)
Martínez et al. (2020b)	Logistic Lasso regression, Extreme learning machine, Gradient tree boosting	Transactional data provided by a large manufacturer located in central Europe from January 2009 until May 2015	The gradient tree boosting outperformed the Lasso and the extreme learning machine when computing next months purchase probabilities
Lee et al. (2021b)	1. Classification tree 2. Artificial neural network (NNET) 3. K-Nearest-Neighbor (KNN) 4. Logistic Regression (LOGIT) 5 Support vector machine with linear kernel (SVML) 6. Random forest (RF) 7. Gradient Boosting Algorithm (GBM) 8. eXtreme Gradient Boosting (XGB)	Google Merchandise Store dataset from 1 August 2016 to 15 October 2018 with 374,749 customer decision journey data and 687 explanatory variables	XGB model is the most suitable machine learning model for predicting online consumers' purchase conversion and that the possibility of explaining machine learning results in a marketing context can be enhanced by using explainable machine learning technologies
Tillmanns et al. (2017b)	A Bayesian variable selection approach	Response data gathered from a direct marketing campaign by a major German insurance. 86,741 rows and 100 variables	Direct marketers would benefit from understanding what the incremental benefits are of adding predictors to the model. Adding variables does not necessarily improve the quality of the predicted responses in profit optimal campaigns.

Coussement (2014)	Logistic regression	1. Information media provider dataset with 134,084 rows. 2. Another Information media provider dataset with 143,198 rows 3. Food from supermarket dataset with 100,000 rows 4. Financial institution dataset with 117,808 rows 5.Do-it-yourself company dataset with 32,371 rows 6. Financial institution dataset with 102,279 rows	Suggests managerial recommendations for optimizing the decision-making process in a customer churn prediction context, considering the cost ratio and churn incidence
Ganesh et al. (2000)	Data collection and data analysis	12-minute phone call interviews. Sample was randomly drawn from the residential section of the current then local telephone directory of a major metropolitan area in the southeast region of the United States.	Switching companies as a client is largely initiated by the customer. switching costs are relatively high and/or consumer involvement with the product/service is high; (3) customer contact and personal relationship issues take precedence over other aspects of the service in terms of selection, satisfaction, and switching;
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016)	LIME, a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. SVMs, DT, LR, RF. Google's pre-trained Inception neural network.	1. Unigrams to differentiate "Christianity" from "Atheism 2. (books and DVDs, 2000 instances each)	Suggested LIME, a modular and extensible approach to faithfully explain the predictions of any model in an interpretable manner
Goldstein et al. (2015)	Using ICE and PDP plots to look into machine learning techniques	1.Dataset containing 156 subjects from a depression clinical trial (DeRubeis et al. 2014). with Hamilton Depression Rating Scale as the response variable 2. Dataset with 5000 white wines produced in the vinto verde region of Portugal obtained from the UCI repository (Bache and Lichman 2013). 3. Dataset that consists of 332 Pima Indians (Smith and Everhart 1988)	Using real data and visualizing how a given black box learning algorithm makes use of covariates to generate predictions with PDP plots and how ICE plots provide a tool for visualizing an arbitrary fitted model's map between predictors and predicted values.
Kumar and Reinartz (2016b)	Review of academic studies	Integrate and synthesize existing findings	Firms need to align the perceived value customers receive with the resources spent on them through the adoption of forward looking metrics, to create net value for the firm. The customer contribution to firm profitability occurs directly, through customer purchases, and also indirectly, through actions that might include referrals or influencing others

			via social networks, customer reviews, and feedback to the firm.
Kumar et al. (2008b)	1.field experiment with a multinational firm that provides three product categories in the information technology industry to business customers 2.field experiment in the telecommunications industry	1. 6350 observations that belonged to the 566 customers 2. B2B companies with annual revenues of more than $50 million and have between 100 and 999 employees	Customers who were exposed to the customer-focused sales campaign believed that the firm understood their needs better

The high importance of the ability of a company in the financial sector to be able to understand if a client is going to lead to a conversion or not has already been mentioned above. There has been a lot of research on churn prediction but less on conversion prediction in this sector even though it is a key for the road of a country's prosperity. In table 1 there is a non-exhaustively summary of these research papers.

Both for the survival and increasing profitability of a business, it is crucial to be able to predict issues but also know the probability of success of the business. Many businesses fail in the first years of operations and to be able to predict what is coming in the short future, it has a great beneficial and practical impact on the business (Naumzik et al., 2022). If business managers have the knowledge about which customers are more likely to lead to conversion, they can shift the business plans accordingly based on the levels of the profits that they are expecting in the near future.

It is really important for a company in the financial services sector to understand which client is likely going to lead to a conversion so it can take action to give an incentive to the client to do so. Furthermore, the business might try to persuade a client that would still stay with the company either way and not use the same resources to persuade a potential churner, so it misses clients and loses possible profits (Gordini, N., & Veglio, V. (2017). By focusing on the right clients, the firm can understand each customer's needs and create customer-based promotion campaigns. In addition, with the benefits for the company, the customer-based campaigns protect the customers themselves from multiple marketing communication and sales calls which they are not interested in or they are not tailored to their interests and needs. Additionally, the company ends up reducing the marketing spending without having a negative impact on conversions (Kumar et al., 2008).

Unfortunately, every business has limited resources, and it cannot pursue every available possible customer lead. Some customers are different from others and the same can be said about their needs and desires. Additionally, marketing managers have to create marketing campaigns focusing on a certain size of possible clients that maximizes the profits of the business because this is what leads to a successful campaign (Lemmens, A., & Gupta, S. 2020). So, those managers could really take advantage of knowing the probability of conversion of certain customers and act accordingly when planning each marketing campaign.

In recent years, companies have become more focused on the clients than the product. They are developing strategies with the customer in the center rather than the product. They are also taking advantage of the increased amount of data on customers which are available nowadays (Coussement 2014). For the company to be able to focus on the right customer it must develop or acquire the right tool to help them on their task.

Customers that are exposed to campaigns that have them as their main focus are feeling that the company understood their needs better (Kumar et al. 2008b) . As their needs are understood better, the customer feels that their personal relationship with the contact person is of higher importance than other elements of the service or product. Additionally, the customers felt that the companies would provide more value in those cases, and they were more likely to suggest the company to their friends and family so the business would also benefit from Word Of Mouth. Furthermore, in the cases where customers switch companies, which is most likely initiated by themselves, is not a very easy task as it requires high involvement with the product or service and in general high costs for the customer like time, effort and money in the process of evaluating information the next possible companies before switching to them(Ganesh et al. 2000).

Businesses are now using machine learning techniques to predict if a customer is going to lead to a conversion or if a current customer is going to churn. Machine learning is gaining ground as a crucial tool for all kinds of businesses. Machine learning techniques are able to make use of large amounts of unstructured data. Additionally, these techniques are capable of handling the available large datasets a lot better when compared to economic models. Other advantages of

machine learning methods compared to economic models are their flexibility and focus on the accuracy of predictions compared to economic models and their focus on causal identification and interpretation (Ma, L., & Sun, B. 2020).

When companies reach the stage of becoming familiar with utilizing the tools that help them predict the probabilities of customer conversion, they can take it a step further and have marketing managers work together with other departments so they can evaluate the cost of resources spent on a customer and the profit which the customer brings to the company (Kumar and Reinartz 2016b) . The profit that the customer brings to the company is the direct profit from the purchase of service or product but also, as mentioned before, the indirect profit such as referrals or the customer influencing others through their social networks, digital and non-digital, customer reviews and feedback to the firm.

There have been recent papers that tested machine learning methods to predict conversion. In one paper, when trying to predict the purchases of products in the next month, gradient tree boosting outperformed the Lasso and the extreme learning machine (Martínez et al. 2020b).
In another paper there were multiple techniques tested and compared to predict online custom purchases. Those techniques were Classification Tree, Artificial Neural Network, KNN (K-Nearest Neighbor algorithm), logistic regression analysis, SVM (Support Vector Machine), random forest, GBM (Gradient Boosting Machine and XGB (eXtreme Gradient Boosting) with the last one to be most suitable one for this kind of predictions (Lee et al. 2021b).

An issue with the machine learning techniques is that they seem like black boxes as it is not clear how they come up with their predictions. There have now been some methods to take a look inside those black boxes. In a relatively recent paper, a method called LIME was used to explain machine learning methods locally (Ribeiro, M. T., Singh, S., & Guestrin, C. 2016). In that paper, Lime was used to understand the predictions of two models: a SVP and a neural network. In the second case by using Lime the authors saw that the machine learning method worked in a reasonable manner but in the first case even though the model had accuracy of 94%, the predictions were based on arbitrary reasons, and they were mainly based on variables that did not really have impact on the target variable.

Another way to look into what is happening inside the machine learning techniques is ICE and PDP plots (Friedman 2001). For example, in Goldstein et al. (2015), the authors used 3 datasets: 1) Dataset containing 156 subjects from a depression clinical trial with Hamilton Depression Rating Scale as the response variable. 2) Dataset with 5000 white wines produced in the vinto verde region of Portugal obtained from the UCI repository. 3) Dataset that consists of 332 Pima Indians and they used BART, NN and RF respectively. In the first case they used ICE plots of the BART model to check the effect of the treatment on depression score after 15 weeks. In the second case an ICE plots of the NN model to check wine ratings versus pH of white wine was used and in the third case they used ICE plots of the RF model for estimated centered logit of the probability of contracting diabetes versus skin colored by subject age. The same way in this paper ICE and PDP plots are going to be used to check interactions and understand more what is happening inside the machine learning methods used to predict conversion in the financial services sector.

3. Data

3.1 General dataset

The data that is going to be used in this thesis come from a financial consultant company in Algés, Portugal.The company is a mid-sized European financial intermediary which is focused on technology and has been in the TOP 30 of the Portugal Fintech Report for four years. A similar company in the Netherlands could be Independer.nl.

As the data provided is in Portuguese, it is going to be translated to English, so it is easier to grasp the process and the results. The data is cleaned from NAs and some of the columns that mostly consist of NAs were completely removed. Furthermore, some columns include some data that do not seem to be right, so they are removed, like the Age column shows clients age lower than 18 and higher than 100. Also, a new column is going to be created that contains the day of the week based on the Date column.

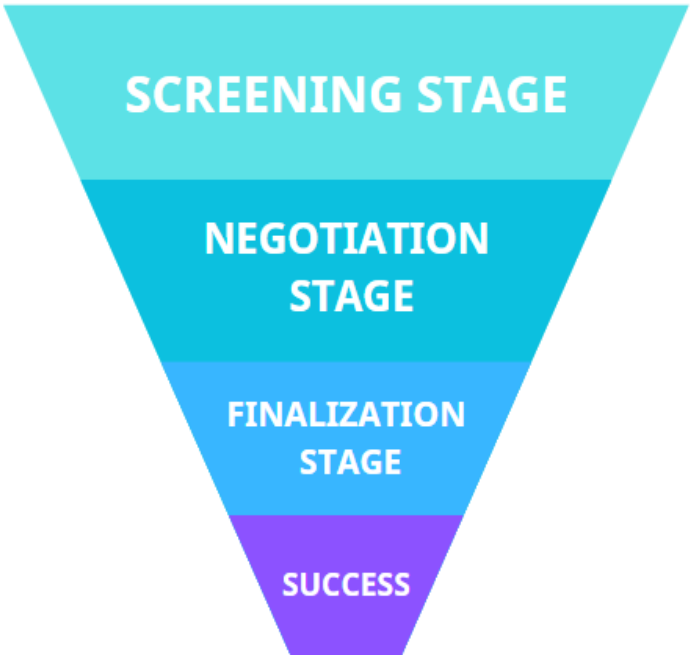


Image 1: *Funnel and Stages*

The data is then going to be split into 3 datasets which the first contains data from the screening stage, negotiation stage and the formalization stage which are parts of the funnel that lead to customer conversion as shown in Image 1. All 3 of them contain the data of the clients that eventually finished the whole process successfully, and each of the datasets also include the data of the clients that decided that they are not interested in continuing with the process. Additionally, the Status column that shows where the client is in the process funnel is replaced with the binary Success column where 0 means that client did not finish the process and 1 where the client successfully finished the process. Table 2 shows all the columns that are included in the datasets.

Table 2: *Columns and examples of values*

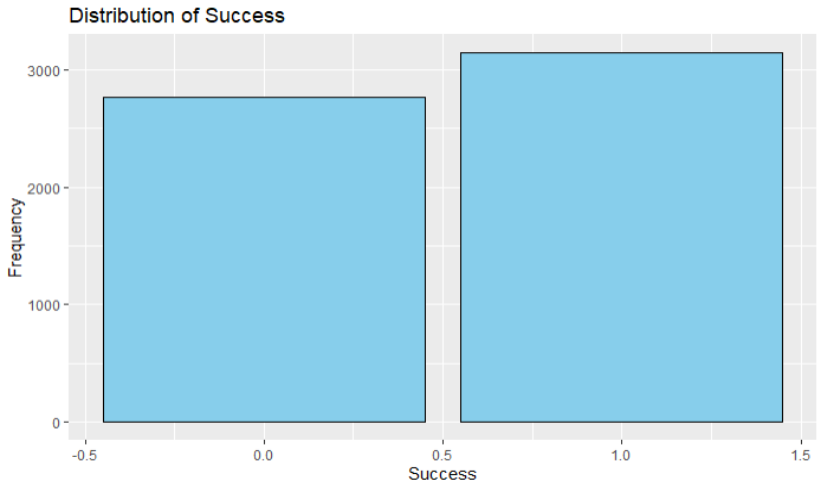
COLUMNS	EXAMPLE #1	EXAMPLE #2
Analise_Date	2021-01-01 11:34:04	2021-01-01 14:06:45
Finance_amount	180000	200000
Lead_Main_Source	Facebook ads	Google ads
Proposed_Spread	1	1.05
Age	30	35
Marital_status	Single	Divorced
Dependents	0	2
Net_amount_available	1444.86	1362.01
Gender	Male	Female
Effort_rate	24	30
Rating	A^-	A^+
Loan_to_value	80	69
Day_of_Week	Friday	Monday
Success	**0**	**1**

According to the business "Effort rate is the ratio between the sum of all the monthly loan payments of a certain person or family and the net monthly income" and Loan_to_value is the amount requested to be loaned divided by house value. The rest of the variables are mostly self-explanatory.

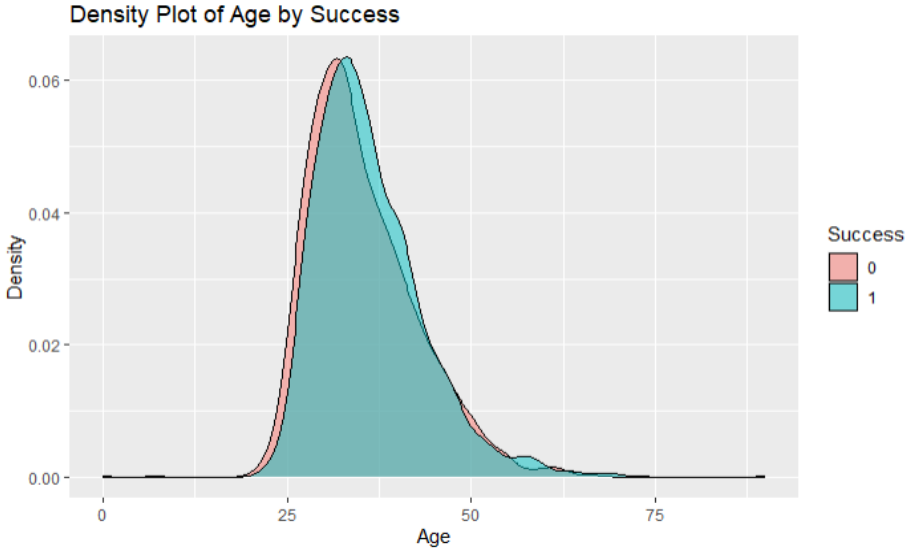
Next, a couple plots are going to be shown to better understand the data that is split into the three stages. The stages are: Screening stage, Negotiation stage and Finalization stage.

3.2 Screening data

The screening process data consists of 5914 observations and 14 variables. There are two plots below. These plots show the distribution of success and the density age in this dataset.



***Plot 1**: Distribution of Success for Screening dataset*

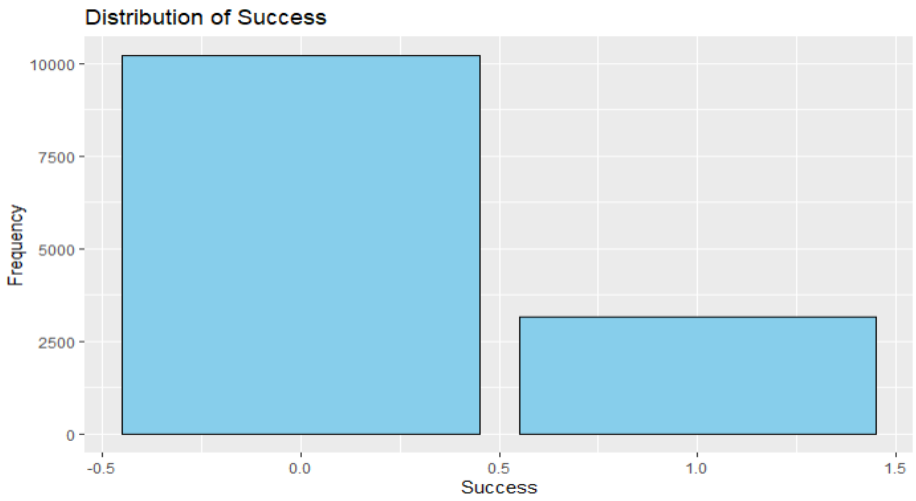


***Plot 2**: Distribution plot of Age by Success in Screening dataset*

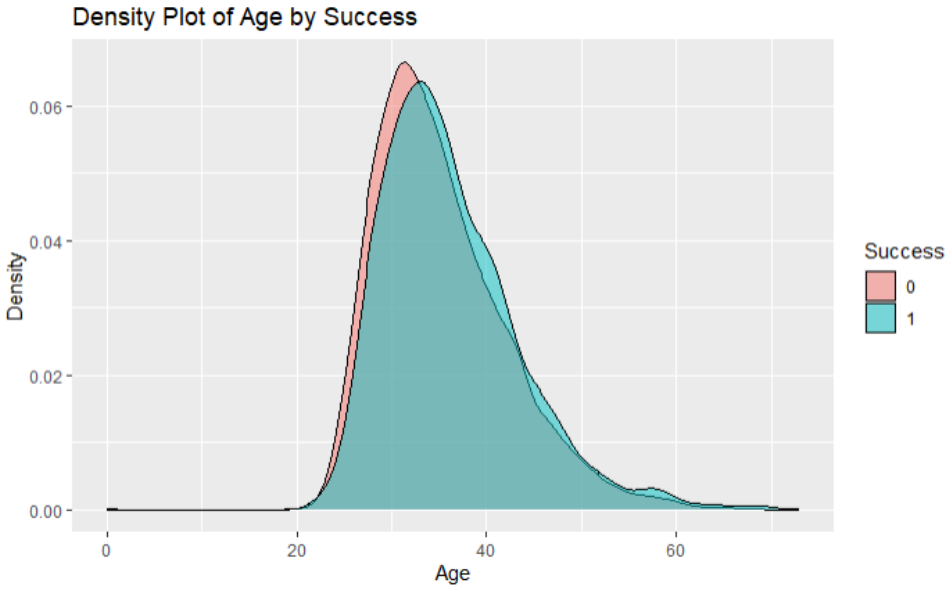
Plot 1 shows that the dataset looks balanced in terms of success and failure to become a customer. The age distribution shown in Plot 2 above seems almost the same for both failure and success to become a customer.

3.3 Negotiation data

The negotiation process data consists of 13353 observations and 14 variables. As previously, two plots are going to be shown below, one shows the distribution of success and the other shows the density of age in this dataset. The first one looks different compared to the same plot for the other two stages.



***Plot 3**: Distribution of Success for Negotiation dataset*

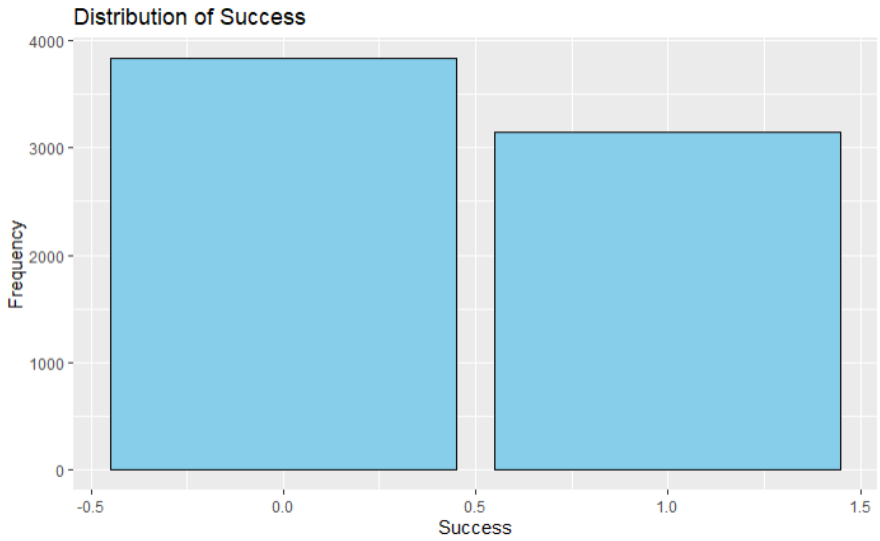


***Plot 4**: Density plot of Age by Success in Negotiation dataset*

Based on Plot 3, the Negotiation dataset does not look balanced in terms of success and failure to become a customer. This issue is going to be solved by oversampling the successful cases. For this task the Rose R package is going to be used. Again, the age density, as shown in Plot 4, seems almost the same for both failure and success to become a customer.

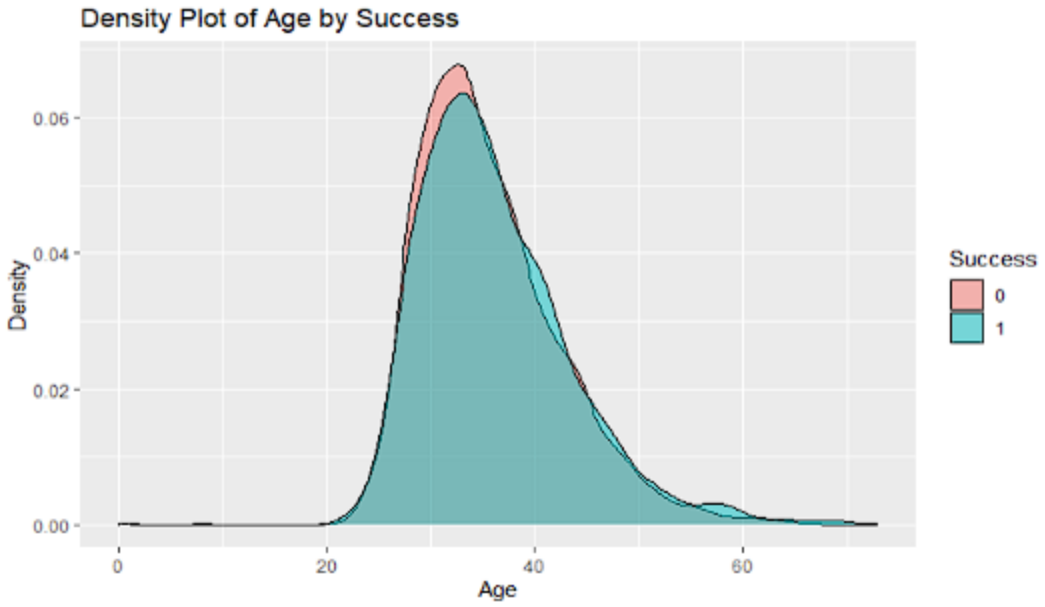
3.4 Formalization data

The formalization process data consists of about 7000 observations and 14 variables. As in the previous two subsections, two plots are going to be shown: the distribution of successful cases which looks like a reversed case of Plot 1 and the density of age in the Formalization data.



***Plot 5**: Distribution of Success for Finalization dataset*

The dataset in finalization stage (Plot 5) looks balanced in terms of success and failure to become a customer. The same way, the age density on success in the finalization stage as shown in Plot 6, seems almost the same for both failure and success to become a customer.



Plot 6: *Distribution plot of Age by Success in Finalization dataset*

4. Methodology

This part of the thesis is going to illustrate the techniques which are going to be used. As this paper is a comparative study of the accuracy of Random Forest, XGBOOST and Neural Network on predicting conversion in the financial services, an explanation of these techniques is going to take place in this section. In addition, with the accuracy results of the models in the Results section, the metrics Precision, Recall and F1 score are going to be included for each model in each stage. Precision shows the accuracy on true positive predictions, Recall shows accuracy on identifying positive cases and F1 score shows the balance between the previous two. The following table (Table 3) shows the formulas for each metric. Since this research is also going to take a peek at what is happening inside the methods used, the black box interpretation methods are going to be explained afterwards.

Table 3: *Metrics Formulas.*

Metric	Formula
Accuracy	$\frac{True\ Positive\ +\ True\ Negative}{True\ Positive\ +\ False\ Positive\ +\ False\ Positive\ +\ False\ Neg}$
Precision	$\frac{True\ Positive}{True\ Positive\ +\ False\ Positive}$
Recall	$\frac{True\ Positive}{True\ Positive\ +\ False\ Negatives}$
F1 score	$2 \times \frac{Precision\ \times\ Recall}{Precision\ +\ Recall}$

4.1 Machine Learning Methods

4.1.1 Random Forest

The first method used is Random Forest. In this research, this method is used to predict if a person is going to become a customer so it is used for classification, but it could also be used for regression tasks. It was introduced by Breiman in 2001.

Random Forest is a supervised machine learning method that uses a combination of bootstrap aggregating and random feature selection. It builds multiple random decision trees using training data and only a subset of variables available which lessens the risk of overfitting. A decision tree is a tree made of nodes and branches where the root splits into branches that lead to first internal nodes that are mutually exclusive subgroups. Every branch represents the variable value for the node. Each of those nodes are then split and continue to do so until the final classification is met. The final classification is this thesis is the conversion prediction is made taking into account the majority of the Decision Trees made.

There are multiple hyperparameters that can be tuned when using Random Forests. The number of trees to be produced, the number of variables per split and the node size. Random forests can handle big datasets and they are robust to outliers, but the risk of overfitting gets higher when the number of trees produced is also getting high. *to be continued base of the parameters used here

4.1.2 Artificial Neural Network

The second method used is Artificial Neural Network (ANN). This machine learning method was made to mimic the structure and function of the human brain. It creates a network made of nodes (Image 2) which analyze the data and are linked by weighted connections.

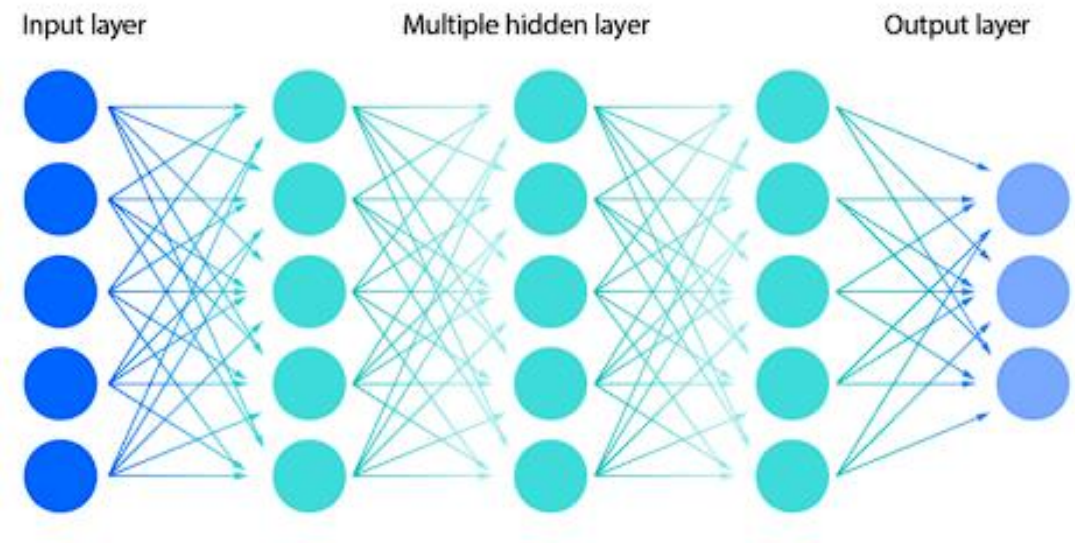


Image 2: *Picture of an ANN, ibm.com*

During the training of an ANN the weights of the connections are adjusted through optimization algorithms such as gradient descent with the goal to reduce the difference between the prediction output and the actual values.

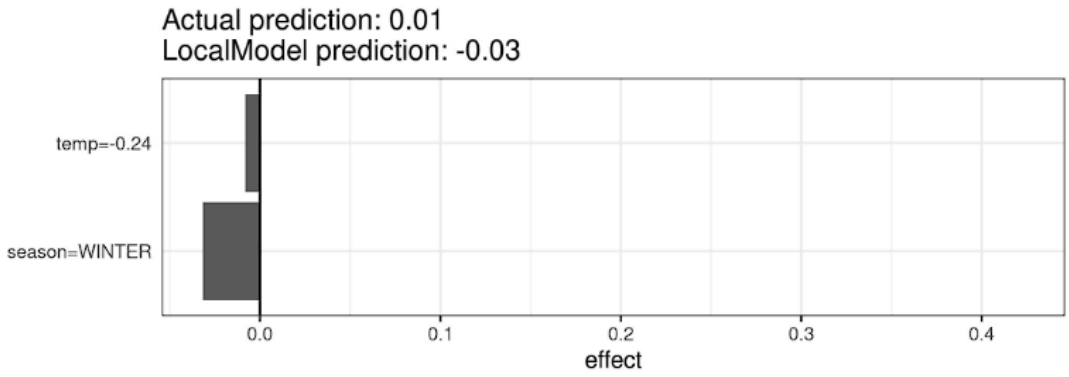
4.1.3 XGBOOST

The third machine learning algorithm that is going to be used is Extreme Gradient Boosting (XGBOOST). It is another supervised machine learning algorithm that uses the predictions of simple prediction methods, such as decision trees, and with ensemble learning and gradient boosting it produces a new robust model. In the scenario of this research is going to be used for classification but XGBOOST can be used for both classification, regression and ranking problems. It is faster and has low computational cost on large data compared to other algorithms that also includes regularization to prevent overfitting.

4.2 Interpretation Techniques

4.2.1 LIME

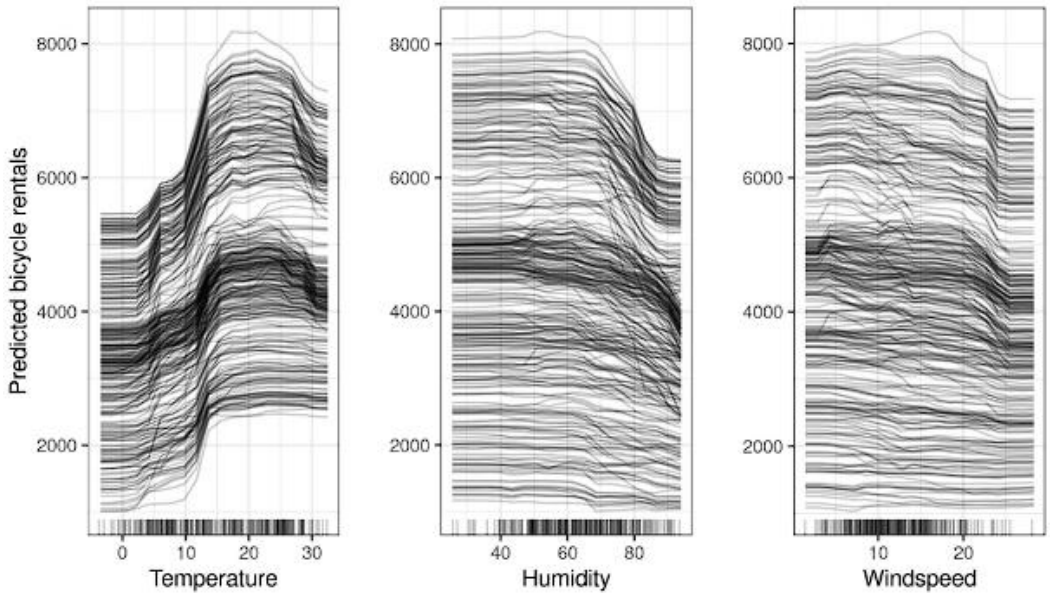
Local Interpretable Model-agnostic Explanations or LIME is a technique for explaining what is happening inside machine learning models locally for individual predictions compared to other techniques that provide explanations for the whole machine learning algorithm. The goal of this model is to give a simple interpretation of a very complex algorithm using a simple model such as linear regression on a specific data point so a person can understand how the complex model works in that specific point. Lime is a very valuable tool as it works on most machine learning algorithms. For example, Plot 7 is a LIME plot which shows the negative effect of low temperature and season(winter) on prediction for bike rentals



***Plot 7**: Example of a LIME plot. (Christoph Molnar 2023)*

4.2.2 ICE

Individual Conditional Expectation or ICE is a technique which graphically shows the impact of a feature used by a machine learning algorithm for each data point. Those graphs are helping with understanding how changes of the selected feature influence the predictions made by the machine learning algorithm when the rest of the features stay the same. For illustration, the ICE plot in Plot 7 shows a model predicting bike rentals and how the prediction changes by the change of a feature.

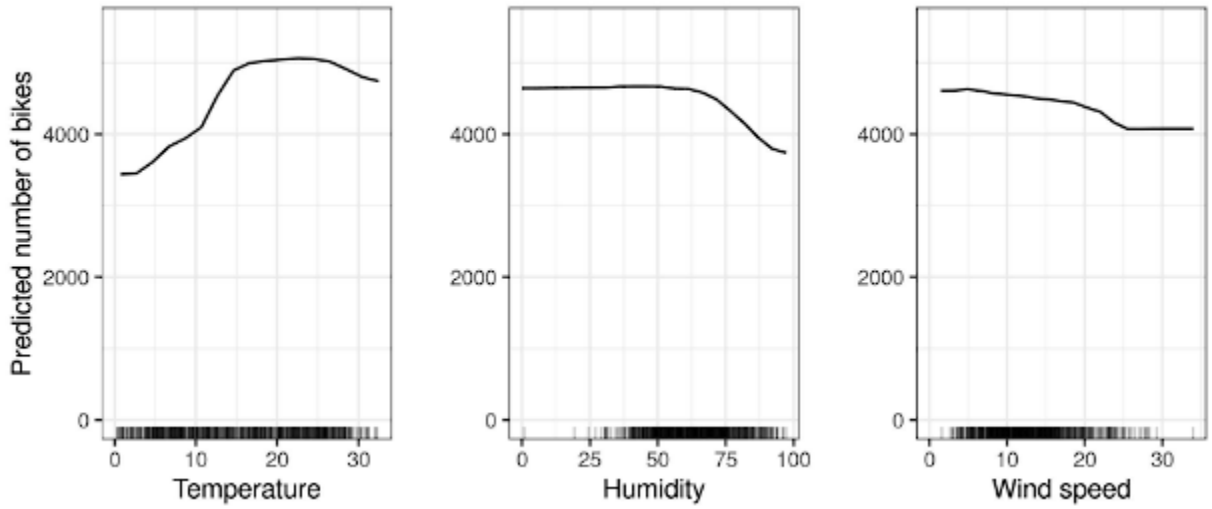


***Plot 8**: Example of ICE plots. (Christoph Molnar 2023)*

4.2.3 PDP

Partial Dependence Plot or PDP technique creates plots that show the marginal effect of one or two variables on the prediction made by a machine learning algorithm (J. H. Friedman 2001). For classification predictions like in this research, the PDP shows the probability of an outcome by changing one or two variables but marginalizing the rest of the features used by the model. Contrary to LIME, PDP is a global technique, meaning that it shows the global relationship of the variable and the prediction made by the model. Using the model mentioned on the previous

plot, Plot 9 shows an example of a PDP plot of a model predicting bike rentals and how the prediction changes by the change of Temperature, Humidity and Wind speed.



Plot 9: *Example of PD plots. (Christoph Molnar 2023)*

5. Results

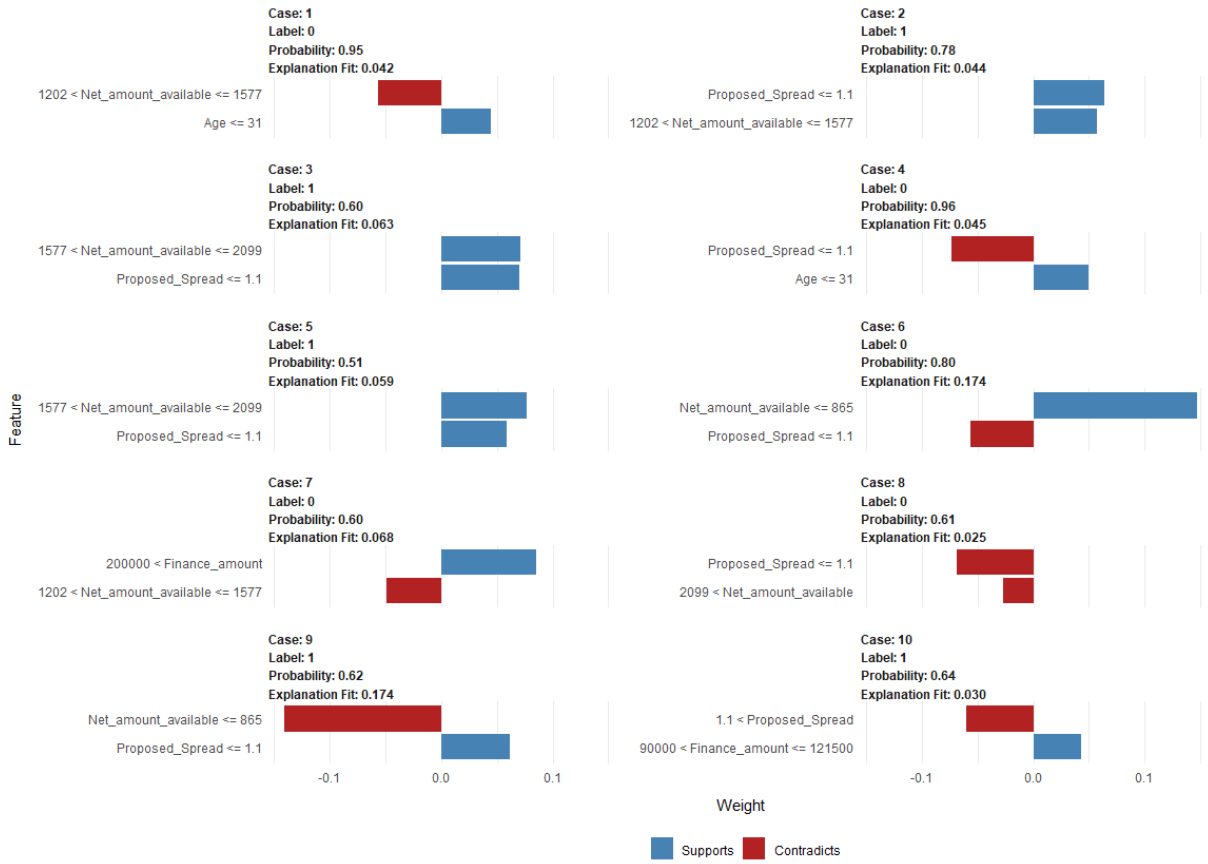
In this section, the Accuracy, Precision, Recall and F1 score for each model for each stage are going to be shown. The models are trained in 70% of the dataset of each stage and tested in the rest 30%. The Random Forest model was the one with the highest accuracy in all three stages. It also had the highest scores overall in the rest of the metrics, so it seems that is the model that is suitable in most cases in all stages. For these reasons, the interpretation techniques that were mentioned in the previous section are going to be used on Random Forest and analyzed in all three stages. In the end, a comparison of the results from all the three stages is going to take place and insights for marketing managers and company's decision makers are going to be given based on the information extracted from the metrics results and the the interpretation techniques plots.

5.1 Screening process

Table 4 shows that Random Forest had the highest accuracy in this stage (76.1%). The cases that the model predicted as successes, about 72.3% of them were actual successes and about 70.6% of the actual successes were predicted correctly by the model.

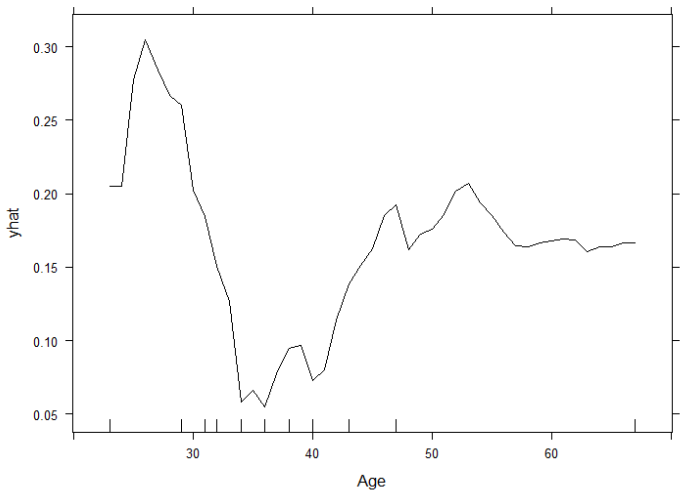
Table 4: *Screening stage metrics.*

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.7614213	0.7236181	0.7058824	0.7146402
Neural Network	0.6463621	0.6250000	0.3676471	0.4629630
XGBoost	0.7394247	0.7093596	0.7058824	0.7076167



Plot 10: *Lime plot of Random Forest in Screening stage.*

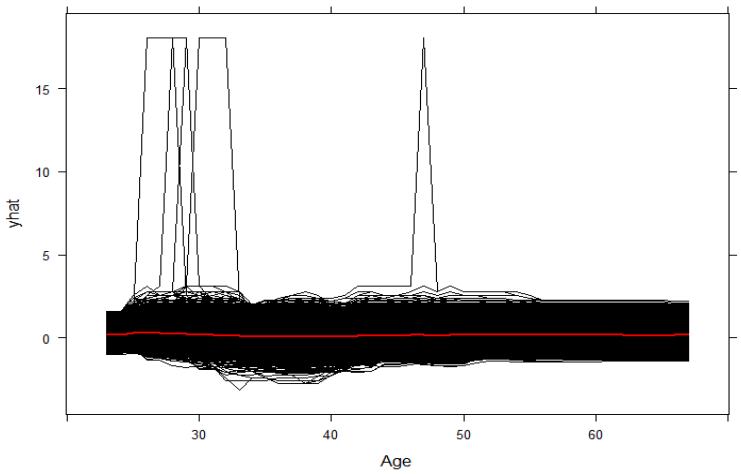
The LIME plot above (Plot 10) shows the effect of the variables Net_amount_available, Proposed_Spread, Finance_amount and Age on 10 predictions made by the random forest model in the screening stage with Label as 1 for successfully becoming a customer. In Case 1 where the Net_amount_available is between 1202 and 1577 and Age being less than 31, the potential customer has a high possibility 0.95 of not actually becoming a customer. On the other hand, for the person in case 2 there is a possibility of becoming a customer of 0.78 with proposed spread less than 1.1 and Net amount available higher than 1577.



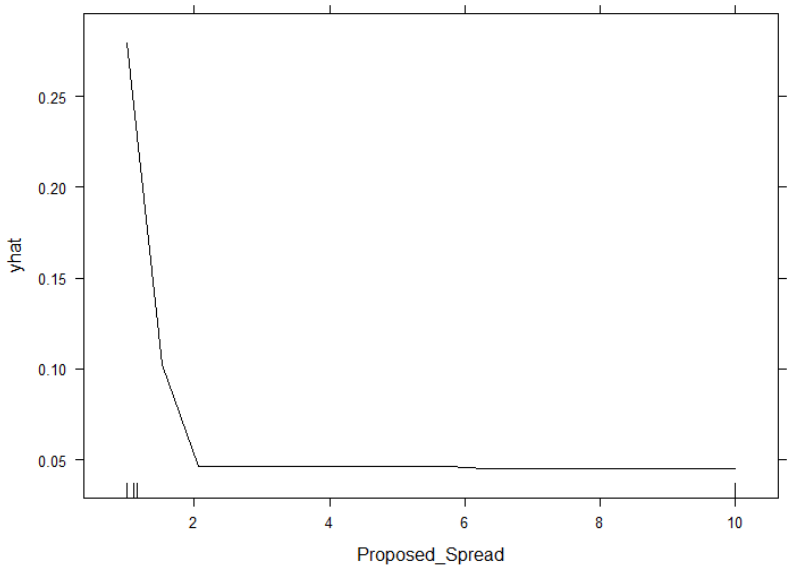
Plot 11: *PDP plot of Age of Random Forest in Screening stage.*

Based on Plot 11, the plot PDP for Age, the people that are 30 or younger have a higher probability of becoming customers. That probability decreases on higher ages but there is still some variation in the probability.

Based on Plot 12, ICE plot for Age, there are some spikes in the probability of becoming a customer in people of 30 years of Age or less. The spikes show that the model is sensitive in that Age bracket. After that mark the predictions become more stable with much fewer spikes. Even though there is a variability in the predictions, it seems that it is more likely that young people are more likely to eventually become customers.

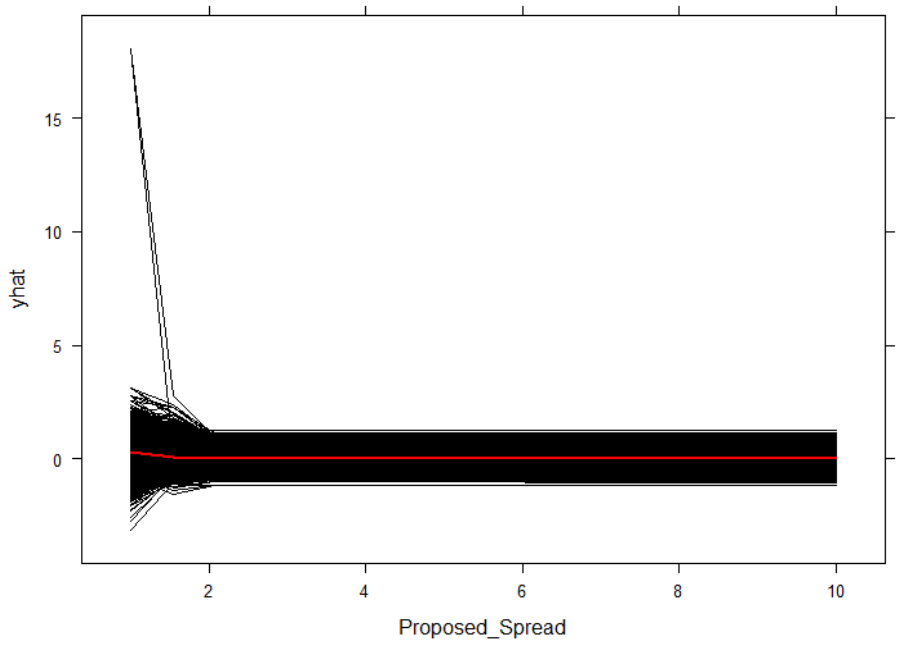


***Plot 12**: ICE plot of Age Random Forest in Screening stage.*



***Plot 13:** PDP plot of Proposed_Spread of Random Forest in Screening stage.*

The PDP plot (Plot 13) for the proposed spread shows a sharp decrease of the possibility of becoming a customer as the spread is getting close to 2. As the proposed spread stays low the probability of becoming a customer stays higher.



Plot 14: *ICE plot of Proposed_Spread of Random Forest in Screening stage.*

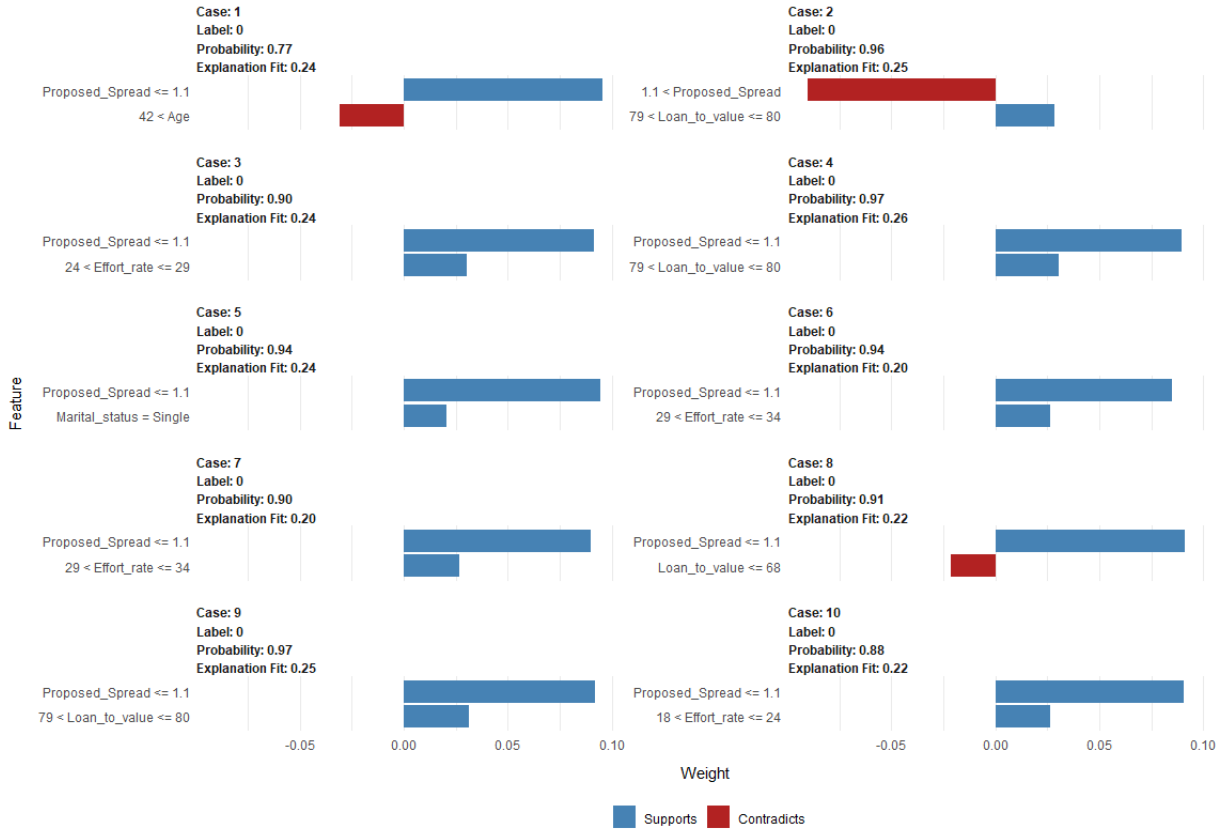
The ICE plot, Plot 14, also shows that as the proposed spread closes to 2 or more the probability of becoming a customer falls. There are also some spikes of high probability in the cases where the proposed spread is lower than 2.

5.2 Negotiation process

In Table 5 it is shown that Random Forest had again the highest accuracy in this stage (83%). The cases that the model predicted as successes, about 77.4% of them were actual successes and about 83.8% of the actual successes were predicted correctly by the model.

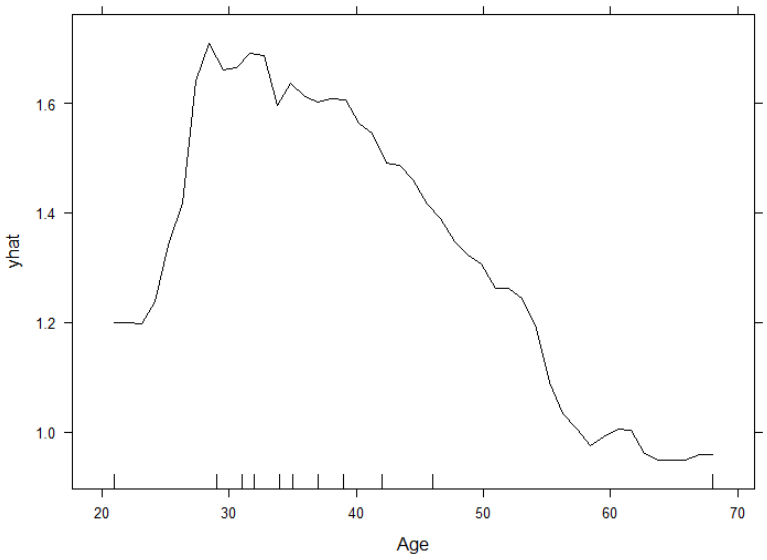
Table 5: *Negotiation stage metrics.*

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.8302735	0.7746914	0.9143898	0.8387636
Neural Network	0.7647059	0.8131635	0.6976321	0.7509804
XGBoost	0.7928062	0.7836991	0.9107468	0.8424600



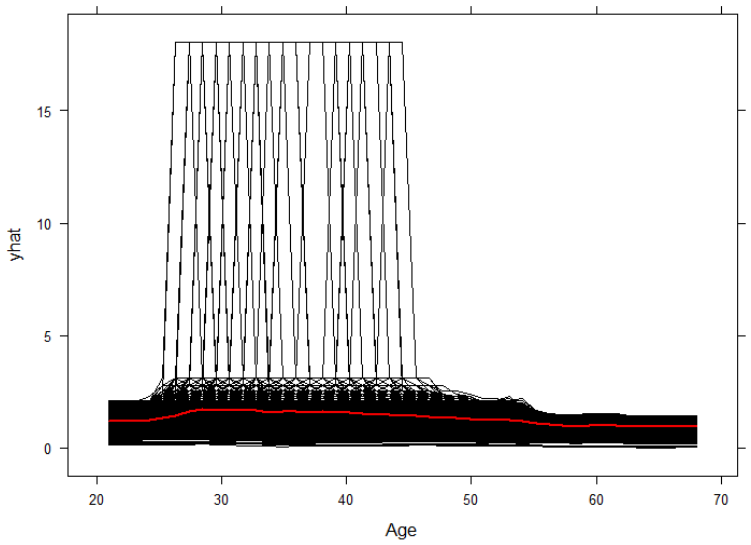
Plot 15: *Lime plot of Random Forest in Negotiation stage.*

The LIME plot above (Plot 15) again shows how the variables contributing to probability of failing, labeled as 0, or becoming a customer, labeled as 1. In the plot above we have for example 2 cases with high probability of failing. Case 2 and 4 have a probability of failing of 0.96 and 0.97 respectively. Both the value of the loan and the proposed spread, even though it is less than 1.1 in both these cases, are contributing towards predicting the failure of the person becoming a customer.



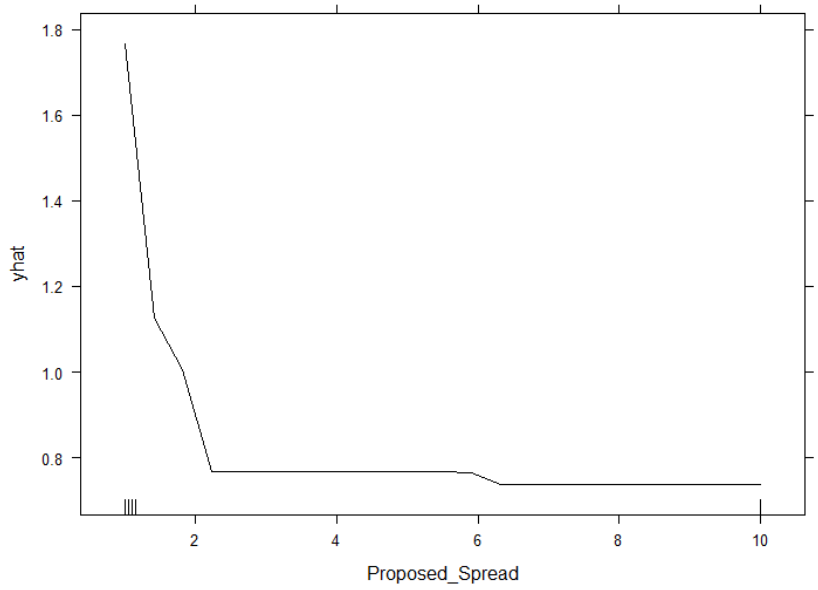
Plot 16: *PDP plot of Age of Random Forest in Negotiation stage.*

Based on the PDP plot of Age (Plot 16), young people have the highest probability of becoming customers. People in their late 20s to early 30s have a higher probability. Similar to the screening stage, the predicted probability of the model is decreasing as the ages are increasing.



Plot 17: *ICE plot of Age of Random Forest in Negotiation stage.*

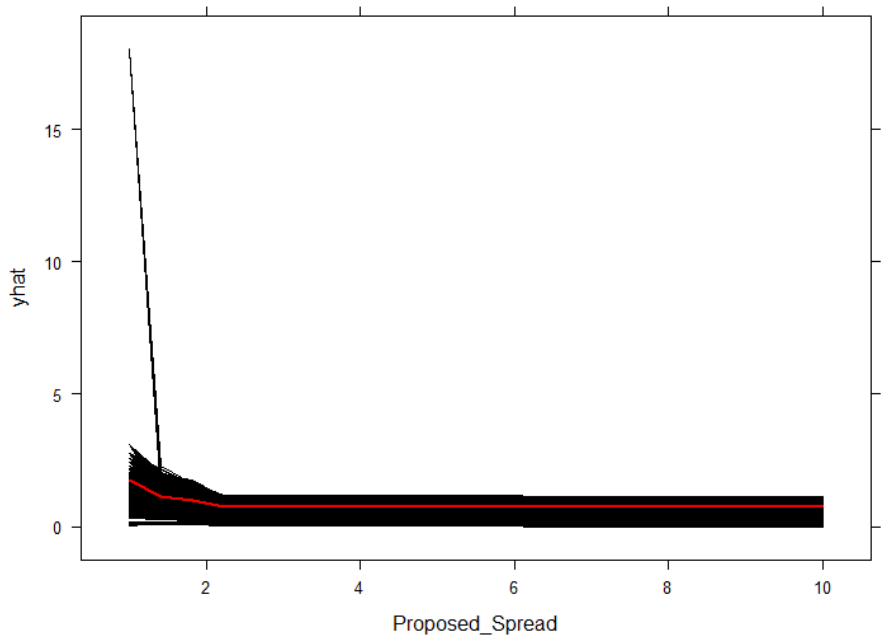
Based on the ICE plot of Age (Plot 17), there are a lot of variations in the predictions based on the age. In younger ages, there are people with higher probability of becoming customers, but it is visible that there are a lot of spikes from ages around 25 to 45.



Plot 18: *PDP plot of Proposed_Spread of Random Forest in Negotiation stage.*

Based on the PDP plot (Plot 18) of proposed spread and similar to the screening stage, as the proposed spread increases the predicted probability of people becoming customers is decreasing. The plot shows a sharp decreasing probability until after the proposed spread of 2.

Based on the ICE plot (Plot 19) for proposed spread there is a high variability between as proposed spread is between 1 and 2. After that it becomes a lot more stable. There is a higher probability of people becoming customers as the proposed spread stays low but that diminishes as the spread increases.



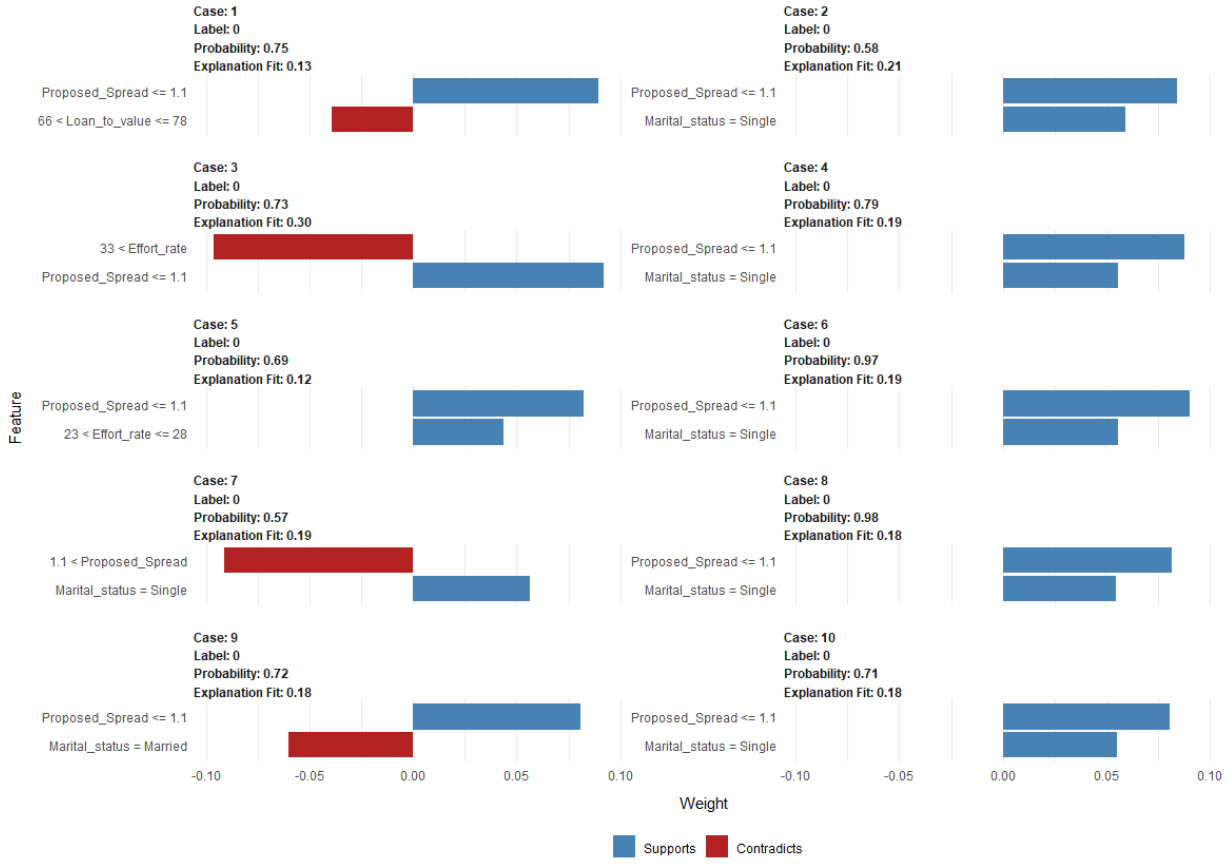
Plot 19: *ICE plot of Proposed_Spread of Random Forest in Negotiation stage.*

5.3 Formalization process

Table 6 shows that Random forest had also in this stage the highest accuracy (76.1%). The cases that the model predicted as successes, about 72.3% of them were actual successes and about 70.6% of the actual successes were predicted correctly by the model.

Table 6: *Finalization stage metrics.*

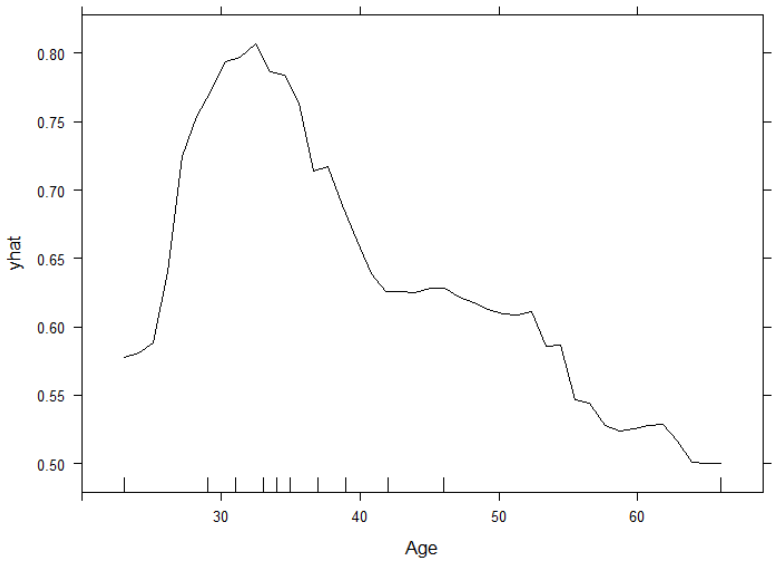
Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.7015782	0.6117381	0.9854545	0.7548747
Neural Network	0.5494978	0.6124722	1.0000000	0.7596685
XGBoost	0.6829268	0.6690909	0.6690909	0.6690909



Plot 20: *LIME plot of Random Forest in Finalization stage.*

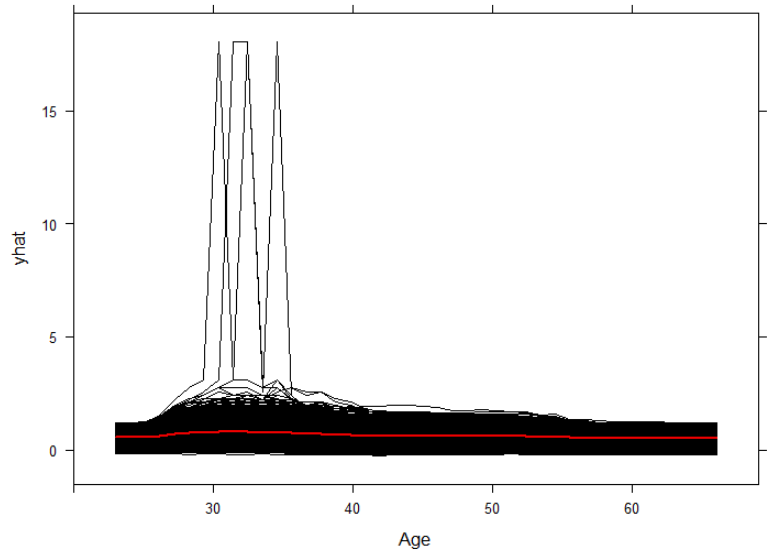
Based on the LIME plot (Plot 20) shows the contributing effect of two variables in each case. The case 10 of the plot shows that the model predicted probability 0.61 for the person to become a customer with Effort_Rate higher than 32 which contributes to failure and the person being Married contributing to the probability of eventually becoming a customer. On the other hand, in case 2 where the predicted probability of failing to become a customer is 0.88, the variables Effort_rate being higher than 32 is contributing towards success and Marital_status being single are contributing towards failure.

Based on the PDP plot of Age (Plot 21), the probability of becoming a customer rises until 35 years of age and then declines relatively sharply until the 40 years mark. Then it decreases with a lower rate until a bit after 50 years of age and after that point, it decreases with a higher rate.

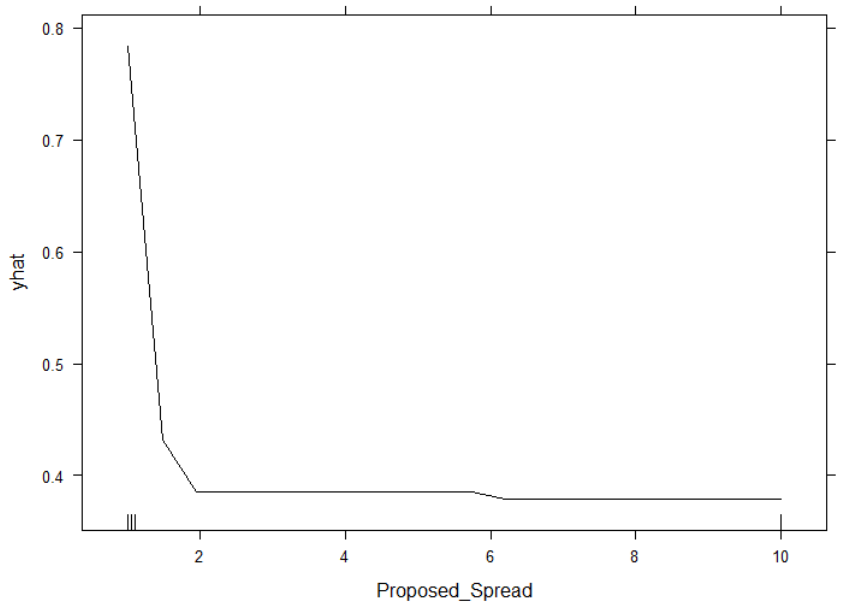


Plot 21: *PDP plot of Age of Random Forest in Finalization stage.*

Based on the ICE plot for Age (Plot 22), in ages below around 37 the probability of becoming a customer is higher with a lot of spikes in the predictions which means that the model predictions are more sensitive in these ages. After around 37 years of age the predictions become more stable.



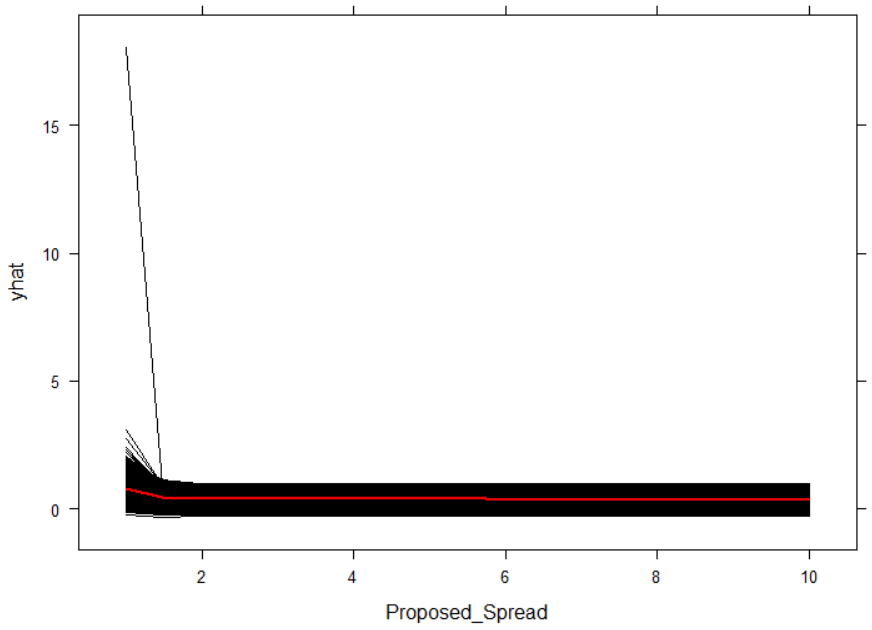
Plot 22: *ICE plot of Age of Random Forest in Finalization stage.*



Plot 23: *PDP plot of Proposed_Spread of Random Forest in Finalization stage.*

The PDP plot of Proposed_Spread (Plot 23) shows a sharp decrease of the probability of becoming a customer until the proposed spread of 2 that stabilizes after that point. There is a small decrease of the probability just before the proposed spread of 6. In general, the lower proposed spread, the higher the probability of the person becoming a customer.

The ICE plot of Proposed_Spread (Plot 24) again shows a sharp decrease of predicted probability before the proposed spread of 2 and in general lower proposed spread means higher predicted probability of becoming a customer.



Plot 24: *ICE plot of Proposed_Spread of Random Forest in Finalization stage.*

5.4 Comparison and insights

In all three stages the machine learning model with the highest accuracy is Random Forest. It also had the highest average F1 score across the three stages and it showed the most balanced performance overall. After this model, the XGBOOST came second in accuracy and Neural Network third. Even after trying to improve the accuracy of the neural network in all three stages, the accuracy stayed very low.

In all three stages, Age and Proposed_Spread play a significant role for the random forest model in predicting if a person is going to eventually become a customer. The probability seems to be higher if the person is in his early 30s and the proposed spread is low. The ICE plots showed spikes in the probability predictions in young ages, below 30 years of age, and low proposed spread in all 3 stages.

In both Screening and Negotiation stages, Random Forest and XGBOOST performed quite well in general with balanced Precision and good Recall. The highest accuracy achieved was by

Random Forest and it was a bit more than 83% in the negotiation stage. The Neural Network had perfect recall which means that the model did not miss any successes but still the other two models scored better in all the other measures.

6. Conclusion

In conclusion, Random Forest had the highest accuracy compared to Neural Network and XGBOOST in all 3 stages. Even after a lot of trial and error, the Neural Network model's accuracy stayed low. In cases where accuracy is the most needed metric then Random Forest looks like the best choice. For instances where the goal is to capture as many people that are going to become customers eventually, the companies should consider using Neural Network because of its really high Recall in the Finalization stage or Random Forest for the rest of the stages. Also, in the finalization stage, XGBOOST had better precision so it can be used if sacrificing a bit of accuracy compared to Random Forest. So the decision makers could use a combination of models depending on the current goal or the stage that is being monitored.

The Negotiation and Finalization stages look similar and the plots in those stages show the same effects and spikes of the variables on the predicted probability and the Screening stage differentiated a bit compared to the two stages that followed. In the screening stage Net_amount_available appears more compared to the other stages. In the Negotiation and Finalization stages Proposed_Spread, Loan_to_value, Effot_rate and Marital_status appeared to influence the model predicted probabilities consistently.

Utilizing the information given by the plots, it is possible to provide some insights to marketing managers and decision makers of financial services businesses. Marketing managers could focus on marketing campaigns and customer support that are tailored for people that the model with highest accuracy shows the highest predicted probability to become actual customers. That means focusing on people in their early 30s and below which are already married or in a marital union. Also, looking into their financial situation or the net amount of money available is going to help focusing on people that have a higher probability of becoming customers.

Furthermore, decision makers of the company could divide the customers into segmentations. Those segments would be based on age, financial situation, proposed spread available for them and marital status. Using this route, they could adjust their focus and manpower to customer segments with people that have higher expected probability to pass through all the stages and become customers.

6.1 Limitations and Future Research

There were some limitations in the process of this thesis. The data used was from a single product in financial services (loans). Additionally, the data needed some cleaning as there were some issues with outliers and some data entries that were not making sense in real life. This issue made the amount of data for each of the three stages smaller. Also, the thesis also focused only on 3 machine learning methods as there should be a limit to make the thesis comprehensive.

These limitations could be reduced by additional research. In the future, more data could be collected from financial services businesses, including different services and products, and used in similar research. Furthermore, more machine learning methods could be used and compared on their accuracy and on the insights that can be drawn from interpretation techniques. To take it a step further, it would be interesting to use other interpretation methods that are available.

7. Bibliography

Coussement, K. (2014). Improving customer retention management through cost-sensitive learning. *European Journal of Marketing*, 48(3/4), 477–495. https://doi.org/10.1108/ejm-03-2012-0180

Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the Customer Base of Service Providers: An Examination of the Differences between Switchers and Stayers. *Journal of Marketing*, 64(3), 65–87. https://doi.org/10.1509/jmkg.64.3.65.18028

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Gordini, N., & Veglio, V. (2017). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Industrial Marketing Management*, 62, 100–107. https://doi.org/10.1016/j.indmarman.2016.08.003

Kumar, V., & Reinartz, W. (2016). Creating enduring customer value. *Journal of Marketing*, 80(6), 36–68. https://doi.org/10.1509/jm.15.0414

Kumar, V., Venkatesan, R., & Reinartz, W. (2008). Performance Implications of adopting a Customer-Focused sales Campaign. *Journal of Marketing*, 72(5), 50–68. https://doi.org/10.1509/jmkg.72.5.50

Lee, H., Lee, Y., Cho, H., Im, K., & Kim, Y. S. (2011). Mining churning behaviors and developing retention strategies based on a partial least squares (PLS) model. *Decision Support Systems*, 52(1), 207–216. https://doi.org/10.1016/j.dss.2011.07.005

Lee, J., Jung, O., Lee, Y., Kim, O., & Park, C. (2021). A comparison and interpretation of machine learning algorithm for the prediction of online purchase conversion. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(5), 1472–1491. https://doi.org/10.3390/jtaer16050083

Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., & Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, 281(3), 588–596. https://doi.org/10.1016/j.ejor.2018.04.034

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" *Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-Agnostic interpretability of machine learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1606.05386

Tillmanns, S., Ter Hofstede, F., Krafft, M., & Goetz, O. (2017). How to Separate the Wheat from the Chaff: Improved Variable Selection for New Customer Acquisition. *Journal of Marketing*, 81(2), 99–113. https://doi.org/10.1509/jm.15.0398

www.thebusinessresearchcompany.com/

https://finance.yahoo.com/