Sponsored by ANZ

UC UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Predicting Financial Difficulty and Hardship with Machine Learning at ANZ

DATA601 Project 2022

Ally Hassell

University of Canterbury & ANZ Good Customer Outcomes

Date Submitted: 10/02/2023

1. INTRODUCTION

1.1 Organisation

ANZ Bank is one of Aotearoa's leading banking and financial service groups with over a million New Zealand customers (ANZ Bank). ANZ formed the Good Customer Outcomes team in response to the Financial Market Authority and Reserve Bank's 2018 review into the conduct and culture of financial institutions, which concluded that banks needed to significantly improve their approach to identifying, managing and dealing with customer welfare (Financial Markets Authority and Reserve Bank New Zealand, 2018).

1.2 Project proposal

The Good Customer Outcomes project researches and develops prototype models for measuring and improving customer outcomes to ensure ANZ is effectively prioritising its customer's needs. Good Customer Outcomes wanted to explore the use of advanced data analytics improve customer outcomes. ANZ's sponsor for this research project is Tim Cunningham, the Good Customer Outcomes project manager and Senior Manager of Initiatives in the Customer Experience team.

To reduce duplication, the project had to avoid topics which were already in research and production. This was difficult because ANZ has a sizable body of existing internal research. This meant significant subject areas could not be investigated, the most relevant was financial well-being. A range of ANZ employees were consulted to make sure the project research goals met the subject constraints and criterion. It was concluded that the most suitable topic to research was financial difficulty.

1.3 Financial difficulty and hardship

ANZ has two classes of customers struggling financially, which differ in severity; difficulty and hardship. ANZ has a proprietary system for classifying its customers into these categories which cannot be disclosed, however some example indicators can be discussed. Financial difficulty is the least severe and occurs when a customer shows behaviours such as missing repayments, regular excess arrears or a marginal transactional history. Financial hardship is more severe and occurs when a customer shows signs that they are at risk of defaulting. Customers in difficulty and hardship are actively monitored and managed by the bank. This system can only identify customers who are currently experiencing financial difficulty or hardship.

1.4 Research goal

The best way to leverage ANZ's financial data would be to create a machine learning model which can predict which customers are going to experience financial difficulty. The purpose of this model was to improve customer outcomes as it allows ANZ to provide these customers with pre-emptive support. Creating new features for financial health would maximise the benefit to ANZ, provide more information about a customer's financial situation and ensure model originality.

Therefore, the high-level research goal of this project was to use novel features and machine learning to create a classification model to predict which customers are at risk of future financial difficulty and hardship.

This goal was achieved by creating and measuring novel financial behaviours and patterns, wrangling the data, engineering features and training a model to predict financial difficulty. This approach aimed to retrospectively identify early warning indicators by looking at feature importance and their impact on recall and the F1 score.

These predictions are relevant to one of the Good Customer Outcomes teams guiding principles: identify and communicate opportunities for the benefit of customers. Identifying customers with deteriorating financial situations and proactively supporting them allows customers the opportunity to adapt their behaviour and improve their outcomes. This model will lay the foundation for ongoing work with data analytics in the Good Customer Outcomes space.

2. DATA

The following table is a summary of the ANZ data used in this project, and provides a high level overview of its characteristics. A list of the all the tables used is included in Appendix A. Count values in Table 1 are for the transactions, incomes, expenses and credit card transactions for the 6 months' worth of data used in this analysis.

Table 1: ANZ data summary information.

Domain	Class	Type	Count	Accuracy	Completeness	Input source
Customer	Dates, numerical, characters	Categorical, continuous	1,907,616	Unknown	**Variable** **Percent** **Missing** DOB 99.94% 1208 FIRST NAME 99.994% 139 LAST NAME 99.98% 258 BRANCH 100% 0 SET CODE 99.9997% 8 GENDER 97.83% 65,386 OCCUPATION 100% 0	Manually input by ANZ frontline workers
Transactions, incomes and expenses	Dates, numerical, characters	Categorical, continuous	Income 45,535,240 Expenses 394,780,000 Pay later 6,123,170	100%	100%	Banking system
Credit card transactions	Dates, numerical, characters	Categorical, continuous	227,220,000	100%	100%	Banking system
Debit card expense categories	Numerical, characters	Categorical, continuous	104	100%	100%	CSV file created by ANZ financial experts uploaded from ANZ intranet
Credit card expense categories	Numerical, characters	Categorical, continuous	829	100%	100%	CSV file created by ANZ financial experts uploaded from ANZ intranet
Financial hardship	Dates, numerical, characters	Categorical, continuous	Difficulty 12,615 Hardship 1,499	100%	100%	Difficulty and hardship branch-set combinations from CSV file uploaded from ANZ intranet

2.1 Data warehouse

All data used to create the novel features were stored in a data warehouse on a multi-cloud data platform called Teradata (Teradata). ANZ stored the raw version of this data on a SAS Application Server (SAS). Low data quality and difficult accessibility of the SAS server compared to the Teradata data warehouse meant SAS datasets were not used. Existing financial features were leveraged from one of ANZ's previous projects called ANZ Assist, this also used the Teradata data warehouse. All data from Teradata was sourced from the Enterprise Analytics data server on the ANZ Teradata server.

2.2 Boundaries and conditions

Boundary conditions were essential to ensure the relevancy of the data. This also provided the opportunity to reduce the size of the dataset which increased processing efficiency and decreased program runtime.

Customers had to be individuals with personal accounts which were all sourced from the same system. The customer had to be alive with a record which was still active, and had to have at least one product with ANZ.

Creating timescale boundary conditions was imperative to reducing the dataset to a reasonable size. The timescale was limited by the processing power required during feature engineering. To achieve a reasonable runtime different timescales were trailed to determine the computers maximum processing capacity before it crashed the computer. It was determined that the timescale should would be between 01/01/2020 to the 01/01/2023. If unlimited processing power was available, then the next limiting factor would be the availability of historical financial difficulty data which goes back to 12/04/2018.

In addition to an overarching time boundary, there was also a time interval boundary set for each individual customer. To predict financial difficulty and hardship, the customer's records for the previous six months from the day they entered were used in the model. For customers who have not experienced financial difficulty or hardship the past six months' worth of records were used.

2.3 Quality

ANZ's financial data was measured against nine quality metrics: accuracy, completeness, consistency, currency, accessibility, conformance, and timeliness.

Customer
Customer information was identified as a likely source of low data quality as it requires ANZ employees to key in the information. Due to the nature of this analysis, it would be unethical and unnecessary to use demographical information to build the model. Therefore, there was no risk of this potential source of low data quality affecting the machine learning model. Each customer has a unique customer ID number, which means that while human keyed in features may not be reliable, these codes were suitable primary keys as they are always consistent, reliable and distinct. This low data quality could affect a retrospective analysis of demographical information, but that was not within the scope of this project.

Data completeness for key demographics was satisfactory, but the accuracy of the information was unknown. Data exploration showed that consistency and completeness varied depending on the feature, most commonly for features that required written input. There was low data quality in areas with multiple possible input combinations where the format was up to the interpretation of the employee keying in the information. For example: there are prefix/first name/last name/full name columns, in some cases the prefix is included in the first name, sometimes it is in both the prefix and the first name, sometimes included in the full name, and sometimes not included at all.

It was timely and current in relation to its time of input into the ANZ system. However, this is unlikely to be reliable for variables which regularly change, such as addresses, as customers are unlikely to update their information with any consistency.

Transactions, incomes, expenses and credit cards
Compared to other business sectors, banks must adhere to strict regulations that demand high-quality data and extensive data management. Automating banking transactional systems eliminates the potential for significant errors in this domain. Complying with banking regulations means transactional, income, expenses are accurate and complete. Meticulous records meant each transaction was uniquely identifiable with a distinct transaction code. Transactional data was updated daily and therefore had a high level of currency. There was a consistent format to the ANZ

data tables in Teradata and a reliable conformance between meta-data. The data tables were easily accessible through the ANZ Teradata server.

Expense categories

Expense categories were available as a CSV on the ANZ intranet. Expense categories are consistent across New Zealand banks, and experts had been consulted to determine each category's discretionary/necessity status. The small size and CSV format mean there is flexibility for future changes. These CSV files have high data quality as they were small enough to be manually checked and peer reviewed. The small size means that high data quality is maintainable.

Financial difficulty

Each ANZ branch has a "set code" for normal, difficulty and hardship customers; when combined with the unique ANZ branch number it creates a unique branch-set code. Difficulty and hardship branch-set codes are labelled in an easily accessible CSV file on the ANZ intranet page. These unique set codes were used to label the data. The contents of the CSV file are accurate, which can be confirmed due to its small size. This information is updated regularly or upon request.

Class Imbalance

An imbalanced classification problem occurs when the distribution of observations across the classes is biased or skewed. This project was limited by an imbalance between the minority class, customers in difficulty and hardship, and the large majority class of customers who were financially stable. A class imbalance can result in a poor predictive performance for the more important minority class. The class imbalance was smaller for financial difficulty as this was more common and lower severity.

The class imbalance between the difficulty/hardship in relation to financially stable customers are:

Table 2: Class imbalance for difficulty and hardship customers.

Financial difficulty	Amount	Class imbalance (%)	Ratio
Difficulty	12,615	99.41%	1:160
Hardship	1,499	99.93%	1:1370

2.4 Data Ethics

Due to the sensitive and personal nature of financial data, privacy issues pose a significant ethical concern. ANZ has a privacy statement available on their website which is transparent about the data they collect, what it is used for and who it is given to (ANZ Data Privacy Statement). Within this statement ANZ states that they will "*use information to improve our products and services, analyse data and generate insights*", which is reflective of the data use in this project. This mitigates ethical concerns as within this statement ANZ is also ensuring data sovereignty is maintained as it shows customers how to review, obtain and correct inaccuracies in their personal data. To obtain products and services from ANZ customers must have agreed to these terms and conditions, so ANZ has met their ethical duty to provide the customer the chance to object to these terms.

To mitigate ethical concerns about data privacy during the production of this project, customer information was only used for legitimate ends during this project and did not violate the rights of ANZ's customers. Customer data was anonymised using customer ID numbers instead identifiable personal information such as names. ANZ has rules and procedures to protect customer's personal data from unauthorized access and accidental disclosure, from advanced security software to data safety training which was undertaken at the beginning of this project

Another ethical concern was how ANZ will use the results of the financial difficulty and hardship predictions. This model were well-intentioned, but if misused, it could harm the customer. Suppose ANZ chose to use the model to reduce their potential financial risk. In that case, the model could be used to penalise customers in vulnerable financial situations by limiting their available products and services. This model is a proof of concept and is not production ready. Because the model will never be 100% accurate, if it was misused in this manner it would identify customers who were false positives. This concern has been mitigated by assurances from the Good Customer Outcomes team that it will only be used by them, and can only be used to determine who to reach out to and provide additional assistance and support, not no punitive action. The possibility of indirectly or unintentionally harming ANZ's customers was reduced by employing best practices and is only accessible by a small number of data scientists. These mitigation strategies were employed to ensure that this project does not create additional barriers and cause new or reinforce existing inequities. This model is property of ANZ and while safeguards have been put in place, ultimately they decide how it will be used.

2.5 Issues and limitations

The enormity of the data encountered in this project posed a significant obstacle. Inexperience in this area of data science meant this was an unfamiliar issue that was difficult to resolve (Scalable Data and Big Data had not studied at this point of project). This was a difficult issue to overcome because each time the code was changed, it would take a significant amount of time for it to run. This meant debugging and improving the efficiency of the code was extremely slow. Eventually this was overcome by employing a combination of strict conditions during data retrieval, using longer time periods and completing extensive feature reductions. With the help and guidance of the ANZ Data Scientist Kayle Ransby, these steps were implemented and the code was made as efficient as possible.

ANZ is a legacy company with enormous amount of sensitive data, this means that switching or updating computer systems is extremely expensive and risky. This has resulted in old software which is not commonplace in other areas of the field. To retrieve and wrangle the data the SAS language had to be learnt and applied on the job. There were limited languages that could be used to complete the machine learning models. Familiar languages like R were not available, so using machine learning models in Python also had to be learnt on the job. The setup at ANZ meant that the complex SAS code was limited by the CPU of the laptop. This slowed the coding process by hindering the ability for the SAS queries to run in a reasonable amount of time.

3. METHOD

The method is split into two sections; feature engineering and machine learning. The first stage collated all the customer features into a table with an observation per customer. The second stage used this table to train a machine learning model to predict financial difficulty and hardship.

The appendix does not contain any code due to the sensitive customer and corporate information it contains.

3.1 Feature engineering

Software, queries and code

The feature engineering program was written in SAS Enterprise Guide version 7.1 as it is the standard programming language at ANZ. This version is the latest version of SAS's proprietary software that ANZ employees can access. SAS is an old industry standard language often used by legacy companies. The program was written as a single script using a mixture of SAS data queries and SQL statements.

Exploration

ANZ's available datasets were investigated using queries to explore the variables, tables, features and dataset relationships until a reasonable level of comprehension was obtained. This exploration was used to identify features which could be incorporated into the model.

Retrieval and pre-processing

Data was retrieved using SAS SQL scripts which connect to Teradata through SAS. Data pre-processing consisted of consolidating and manipulating the data collected from Teradata and store it on the SAS server. This raw data was formatted and standardised so that it would be suitable for wrangling, feature engineering and machine learning.

Wrangling

Data wrangling was required to change the data into a more usable form which was appropriate and valuable to the machine learning stage of the project. This required transforming over a billion rows of information into a single table. This was achieved through extensive merging, transforming, joining, sorting, and data manipulation.

This SAS script took two thirds of the allocated project time to complete. It included 157 queries, 2600 lines of code, 82 intermediate tables and takes over 8 hours to run. These steps produced a clear, consistent, detailed and structured dataset with 78 features per time period.

Time period

Six months' worth of data was collected for each customer. This analysis aims to capture the change in features over time so the feature is measured, recorded and added to the customer features per time period. The time periods used to train the machine learning model was three sets of two months. For example, if a customer went into hardship on the January 1st 2023 the time intervals would be July 1st - August 31st, September 1st - October 31st, November 1st - December 31st.

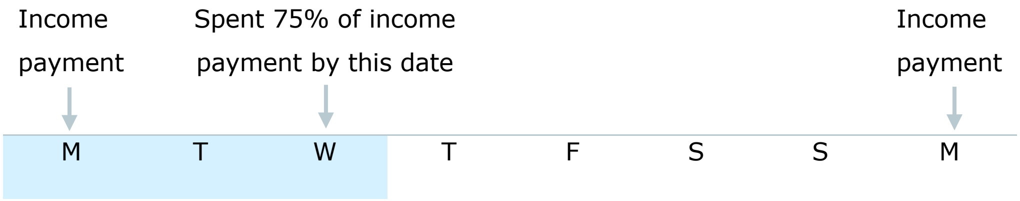
New features

Creating novel features was done to provide the model with additional information to improve its machine learning capabilities. Feature generation was limited by mathematical operations and existing data. Statistical measurements on the continuous data, such as medians or means, were carefully considered to ensure they were suitable.

New features were entirely original and had not been measured in this context at ANZ before. These were:

- Spending velocity: what percentage of the time between income payments did it take for a customer spends 75% of their income.

  For example:

  $$Velocity = \frac{Days\ till\ spent\ 75\%\ of\ income}{\left(\frac{Days\ between\ incomes}{100}\right)} = \frac{3}{\left(\frac{7}{100}\right)} = 43\%$$

  The lower the spending velocity, the faster a customer is spending their income. It had to be measured in percentage of time between income payments because customers get paid on a variety of schedules, such as weekly or monthly, and this makes the values comparable. If there are multiple velocities in a time period then the average is taken.
- Pay later: number and amount spent through pay later schemes (Afterpay and PayPal).
- Credit card: proportion of their credit spending which was discretionary or necessity.
- Expenses: proportion of their income which was discretionary or necessity spending.

Existing features

ANZ has previous research in the customer outcomes space, the main project that could be leveraged was called ANZ Assist. This previous work was used to increase the number of customer features. This met the topic constraints as ANZ Assist provided current status on customers welfare, but did not use historical information to make future predictions. Existing features could be categorised under four topics.

These topics and some examples are:

- Bank categories: credit rating
- Credit cards: late fees, number of credit cards, amount
- Overdraft: limit, average balance
- Home loans: Number of home loans, type of home loan, number of missed payments
- Personal loans: amount, amount payed off in this time period
- Financial summaries: sum funds under management at ANZ

For the full list of customer features please review the data dictionary in Appendix B.

Processing speed limitations

ANZ’s software access and availability restricted the programming language and processing speed. This resulted in using legacy software SAS that relied on the working laptops CPU to run large amounts of data processing. Extensive work was completed to make the code run quicker and more efficiently. To make the code run quicker SAS SQL queries were avoided where possible as they take longer to process than SAS data queries, particularly for join statements. The enormity of the transaction data meant that the amount of data being retrieved had to be reduced to avoid crashing the program; this was done by using extensive conditions and running it for smaller timescales on a loop.

3.2 Machine learning

Binary classification models are a supervised learning algorithms that classify observations into one of two classes. Two classification models were required to look into both stable/difficulty and stable/hardship. This included defining the performance metrics, feature reduction, comparing a variety of models and tuning the modes to maximise.

Performance metrics

To build a model which is suitable for this project it needed to maximise the number of correct financial difficulty predictions, while also reducing the number of false positives. The model performance needed to balance these metrics because this information will be used to identify customers to contact and support. Therefore each false positive is a customer that didn’t need support and wasted resources, but when a false negative occurs a customer in financial trouble has missed out on crucial support. To balance these priorities three aspects of model accuracy will be used to evaluate its performance; recall, precision and the F1 score. Accuracy was an unviable and unreliable metric for performance because the dataset is class imbalanced.

Finding a balance between precision and recall is difficult and at this stage human intuition was used. Ultimately, a production ready model would take into consideration the available resources to contact and support customers, and the consequences of customers reaching the financial difficulty or hardship stage.

Feature reduction

Dimensionality reduction is the process of reducing the number of features in an effort to its reduce computational requirements while maintaining important information. Due to the high number of customers, the large number of features and multiple time periods the customer features table produced by the SAS script was enormous. To prevent an unreasonably large computation load several steps were taken to reduce its dimensionality and increase efficiency. When beginning this process there was 234 features, which was reduced down to 16.

Firstly, the number of features was reduced by eliminating variables which were highly correlated to each other and were therefore redundant. For example, variables such as high balance, low balance, average balance and median balance contain overlapping information. In this instance, only the average amount was kept and the rest were discarded. For consistency, the average amount was always used where applicable.

Secondly, the timescale was simplified to reduce the number of times a feature would be repeated. Initially a monthly scale (6 x 1 month) was going to be used, this was changed to a two monthly interval length (3 x 2 months) as previously discussed.

The third technique was matrix factorisation. This dimensionality reduction technique reduced multicollinearity by using a spearman correlation matrix to determine which variables were highly correlated to each other.. The program then removed half of the features with a pairwise correlation coefficient of over 0.6. The value of 0.6 was chosen as this is a commonly used value in similar analyses at ANZ, however this value is flexible and could be changed in future. A visualisation of the correlation matrix can be found in Appendix D.

The fourth technique used to reduce dimensionality was feature selection using feature importance scores. Feature importance describes the usefulness of a feature in improving the models performance, features with low importance occur when they aren't useful for making predictions. Feature selection is the process of automatically or manually choosing features. When used together the features with the highest feature importance scores, and therefore contribute the most to the model, could be kept and the rest discarded. For this project any feature with an importance of 0 was selected and removed. Permutation feature importance was used as a measure of feature importance due to its versatility across models. A features permutation score is defined as the decrease in a model score when that features value are randomly shuffled.

Modelling

Two methods were used to create a machine learning model. The first tried a range of different models using ten-fold cross validation, the second tried to tune a single model and used the validation set approach with a test-train split of 30/70.

Trial 1: Trial and compare different models

This stage attempted ten different machine learning algorithms and evaluated them using the performance metrics. The Python package PyCaret was used to trial different models because it creates machine learning models quickly, easily and efficiently.

Trial 2: Tune models and improve their performance

Hyperparameters can be used to tailor the behaviour of machine learning algorithms to optimise their performance. Two methods were used for hyperparameter tuning, a methodical grid search and a random search. Once run, the suggested hyperparameters were manually set when configuring the model. This step will only be completed for financial difficulty because it is time consuming, and the small amount of financial hardship data would require significant input for a small improvement in the model.

Feature importance

Two methods for calculating feature importance were used. The permutation method was used in feature reduction, but also provides insight into which features drive the model's performance. SHAP (SHapley Additive exPlanations) feature importance was also calculated as was used to explain which features played a more important role in generating a prediction.

4. RESULTS

4.1 Trial many models

For financial difficulty and hardship, the trialled models and their performance results are in Table 3. The best performing models for each respective model were highlighted in blue. Appendix C contains the cross validation results for these models.

Table 3: Financial difficulty and hardship model performance metrics.

	Financial difficulty			Financial hardship		
Financial difficulty model	**F1**	**Recall**	**Precision**	**F1**	**Recall**	**Precision**
Decision tree	0.58	0.60	0.57	0.06	0.06	0.06
Random forest	0.69	0.57	0.87	0.05	0.03	0.32
Linear discriminant analysis	0.22	0.14	0.58	0.04	0.06	0.03
Gradient boosting classifier	0.38	0.25	0.78	0.01	0.01	0.11
Support vector machine	0.15	0.22	0.16	0.01	0.03	0.01
Naïve Bayes	0.02	0.90	0.01	0.002	0.82	0.001
Logistic regression	0.04	0.02	0.18	0.000	0.000	0.000
Quadratic discriminant classifier	0.02	0.94	0.01	0.002	0.89	0.001
Extremely randomized trees	0.37	0.28	0.52	0.05	0.03	0.17

Comparing many models was an effective method for finding a good machine learning algorithm. The best models were chosen by balancing the recall and precision scores.

4.2 Trial tuned models

Model tuning was an iterative process which aimed to find the optimal values of hyperparameters to maximize models' performance. Hyperparameter tuning was completed for the financial difficulty random forest model. The results can be found in Table 4 below, where the best performing model is highlighted in blue.

Table 4: Hyperparameter tuning process.

Model	**Tuning**	**Max depth**	**Max features**	**Max leaf nodes**	**Min leafs**	**Min splits**	**F1**	**Recall**	**Precision**
Financial difficulty: Random forest	None	-	-	-	-	-	0.42	0.33	0.58
	Grid	3	-	6	-	-	0.28	0.17	0.80
	Random grid	50	5	-	2	8	0.45	0.31	0.80
	Human trial	9	-	9	-	-	0.28	0.17	0.7

These results show that tuning can have a significant impact on the performance of a model, however the cross validation models from the Section 4.1 performed better. Random search was great for discovering hyperparameter combinations when there was no previous knowledge about which would perform well.

4.3 Best performing models

It was concluded that the cross validation models performed the best, with random forests the best model for financial difficulty and extremely randomized decision trees for financial hardship. Comparing the performance of financial difficulty and hardship, it was clear by the significantly lower results for hardship that it is much more difficult to predict.

Confusion matrices

Confusion matrices were used to define the performance of the classification algorithms:

		Actual	
		Positive	Negative
Predicted	Positive	7,153	1,049
	Negative	5,462	1,881,137

Figure 1: Financial difficulty confusion matrix.

		Actual	
		Positive	Negative
Predicted	Positive	47	226
	Negative	1,452	1,894,575

Figure 2: Financial hardship confusion matrix.

Confusion matrices show the number of correct classifications in relation to the number of false positives and false negatives. For financial difficulty, the confusion matrix shows that roughly two third of customers were able to be identified with a low false positive rate. However, the financial hardship model was only identifying a small proportion of the customers with significantly more relative false negatives.

Feature importance SHAP

Feature importance was measured using SHAP values as this is the standard method used by the PyCaret package. SHAP is based on the magnitude of feature attributions in the predictive model. It should be noted that while the model could have chosen any time period, all of these features were from four to six months ago rather than the recent time periods. The star shapes next to features (★) are to indicate the new features created for this project.

Delinquent variables occur when payments are missing. Loan arrangements are when the bank gives a customer a special interest deal to help them if they are struggling to pay their loans.

Table 5: SHAP feature importance plot for financial difficulty.

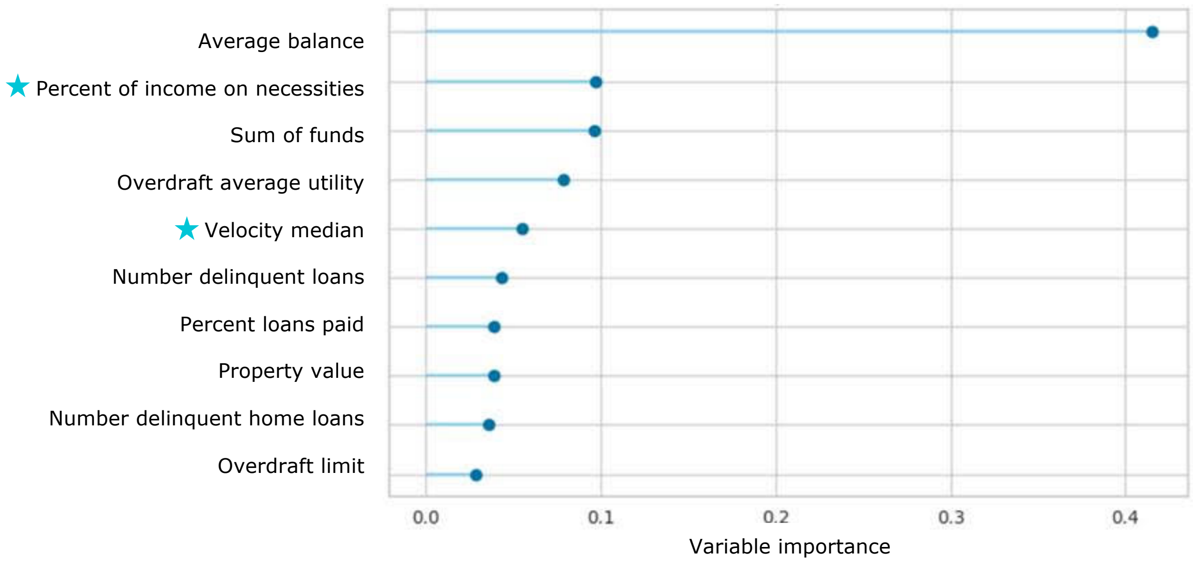
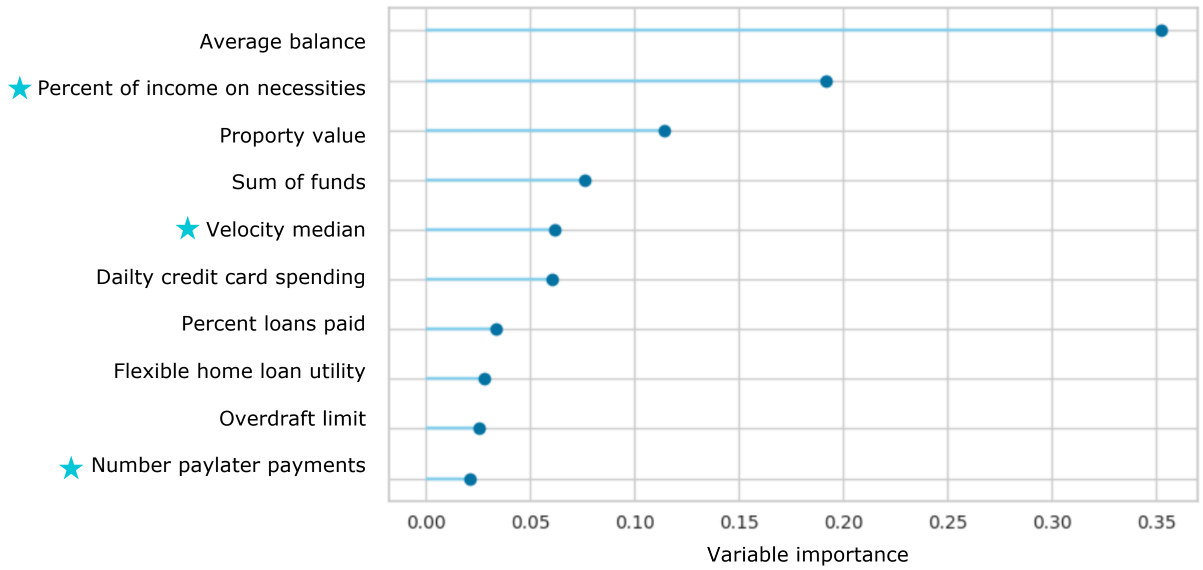


Table 6: SHAP feature importance plot for financial hardship.



These features show that in both cases average balance and percent of income spent of necessities are very important to these models. This reflects intuition as a low balance and only buying necessities is almost certainly going to occur is a customer is struggling financially.

It is interesting that all the variables are from three months ago, this suggests that the financial situation in the far past is more important than more recent trends.

For financial difficulty two of the ten features were novel, and for hardship three of the ten. It is great to see that the creation of features was able to add to the models performance and enrich the insights it can provide. Median spending velocity appears on both lists, this shows that how quickly a customer spends their money after being paid is an important factor in determining difficulty and hardship.

Feature importance permutation

Permutation feature importance had similar results to the SHAP features with the addition of a few other features.

Financial difficulty:

- Percentage of income on necessities ★
- Median velocity ★
- Number of pay later payments ★
- Daily credit card spending ★
- Number of floating home loans
- Average balance
- Overdraft limit
- Average flexible home loan utility
- Number of interest only home loans
- Number of delinquent home loans
- Number of loan arrangements
- Loan percent paid
- Sum of funds
- Property value

Financial hardship:

- Percentage of income on necessities ★
- Median velocity ★
- Number of pay later payments ★
- Number of floating home loans
- Average balance
- Overdraft limit
- Average flexible home loan utility
- Number of interest only home loans
- Number of delinquent home loans
- Loan percent paid
- Property value

These features are very similar to those included in the SHAP feature importance plots. This repetition of features confirms their relevance to the model. With some additional work, these could be used as a starting point for the early warning indicators requested by the Good Customer Outcomes team.

These features were also the final result of all the feature reduction steps, which went from 234 to 16 and 11 respectively. Having so many novel features included in the final features meant that they provided new information which added to the predictive power of the models.

4.4 Machine learning algorithm suitability

Differing algorithms have varying strengths and weaknesses which determine its effectiveness for any given data set.

Financial difficulty - Random forest

The simplicity of random forests means they can handle big data with numerous variables, making it suitable the complicated computation this dataset required. Random forests are particularly suited to this projects dataset because they automatically balance data sets when the data is class imbalanced. Random forests are great with high dimensional data and is therefore highly suitable to this dataset. Random forest models have fast runtimes which was important for this project.

Financial hardship – Extremely randomized trees

While a random forest build trees over random subsets of the data and chooses the optimum split, extremely randomised trees use all the data and pick a random split. The main advantage of extra trees is that the entire dataset is sampled during construction of the trees which results in a reduction in bias. This is important when using this dataset as its magnitude and class imbalance means there are many opportunities for bias and variance to affect the results. Similar to random forests, the simplicity of the machine learning algorithm means it can handle big data with lots of features.

5. CONCLUSIONS AND DISCUSSIONS

5.1 Fit-for-purpose

The original goals of this project was to provide ANZ with a predictive model so they could identify customers likely to experience financial difficulty and hardship so they could provide them with pre-emptive support. The models developed could be improved, but their current performance still provides useful information.

The machine learning model for financial difficulty is fit for purpose as it can identify roughly two-thirds of customers and only has one false positive for every seven correct identifications. Its precision means the number of false positives is small enough that if used to contact and support customers, it would waste a relatively significant amount resources. Its recall is lower than desired and this means that it does not identify all customers who will experience financial difficulty. It is fit for purpose because this machine learning model could be improved and productionized to provide customer leads to the Good Customer Outcomes team.

The machine learning model for financial hardship is less fit for purpose because it can only identify a thirtieth of these customers and has four false positives for every one correct identification. This model could be used by ANZ to contact this group because it is so small (all together about 250 people) if this trade-off is acceptable. While it is not fit the purpose of preventing large scale financial hardship, this model could still help them prevent this unfavourable outcome for 50 people. To be fit for purpose it would require extensive tuning and additional modelling.

The last desired outcome was a list of key features which can be used as early warning indicators. The SHAP feature importance’s and the permutation feature importance’s can be used as these difficulty and hardship red flags. There are many applications for these indicators, such as further research into why these are the most important features and what this means for preventing financial difficulty and hardship.

5.2 Future work

The top features for both permutation and SHAP feature importance demonstrated that the customers financial status in the far past was more important than recent trends. Therefore, the next steps can use this information to experiment with different time periods to see if the model could be improved by using data from further in the past.

Additionally, extra new features could be created to enrich the dataset and improve the models performance. Accommodation is often a customer’s biggest expense, so a new feature idea is the proportion of customers income they spend on their home loan or rent. This would provide insight into the importance of accommodation costs on financial difficulty or hardship.

Another next step would be to incorporate class imbalance techniques such as under or over sampling. This method may be particularly useful in improving the financial hardship recognition rates.

The final step would be to use the model to make predictions and use these to provide the Good Customer Outcomes team with leads on customers to contact and support. This model could then be used as a tool by the Good Customer Outcomes team to implement evidence-based strategies to prevent unfavourable outcomes.

6. REFERENCES

ANZ Bank. (2023). *ANZ About Us*. Retrieved from ANZ Bank Website: https://www.anz.co.nz/about-us/#:~:text=ANZ%20is%20one%20of%20New,New%20Zealanders%20bank%20with%20ANZ.

ANZ Data Privacy Statement.. *ANZ Data Privacy Statement*. Retrieved from https://www.anz.co.nz/about-us/privacy/

Financial Markets Authority and Reserve Bank New Zealand. (2018). *Bank Conduct and Culture: Findings from an FMA and RBNZ review of conduct and culture in New Zealand retail banks.*

SAS. *Overview of SAS Application Servers*. Retrieved from https://documentation.sas.com/doc/en/bicdc/9.4/biasag/n02001intelplatform00srvradm.htm#:~:text=A%20SAS%20Application%20Server%20is,Java%20code%20or%20MDX%20queries.

Teradata. *Teradata: Enterprise Data Analytics for a Multi-Cloud World*. Retrieved from https://www.teradata.com/

APPENDIX A. DATA TABLES

This is a list of the tables used under each category, all extracted from ANZ's Teradata system.

- Customer information
  Customer account bridge
  Customer account dimensions
- Transactional income and expense data
  Transaction fact
  Transactional income and expense
  Transactional income and expense dimension table
  Transaction particulars
  Transaction particulars dimensions
- Credit card transactions
  Credit card transaction fact
- CSV input files
  Expense categories (discretionary and necessities)
  Merchant categories (discretionary and necessities for credit card transactions)

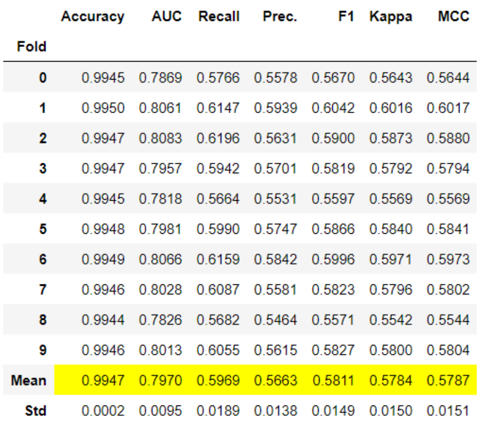
APPENDIX B. DATA DICTIONARY

Table 7: Data dictionary of data produced by SAS queries code.

Data field name	Description	Data Type	Field length	Sample
CUSTOMER_SK	Necessity percent	Numeric	8	60%
CFW_flag	Financial difficulty class label	Numeric	8	1
Hardship_flag	Financial hardship class label	Numeric	8	1
N_PERCENT	Necessity percent	Numeric	8	60%
D_PERCENT	Discretionary percent	Numeric	8	30%
VEL_AVG	Average velocity	Numeric	8	20%
VEL_MED	Median velocity	Numeric	8	34%
PAYLATER_AMT	Paylater payments amount	Numeric	8	$80
PAYLATER_NUM	Number of paylater payments	Numeric	8	3
CREDITCARD_N_DAILY	Credit card daily necessity spending	Numeric	8	$12
CREDITCARD_D_DAILY	Credit card daily discretionary spending	Numeric	8	$53
CREDITCARD_TOT_DAILY	Credit card daily total spending	Numeric	8	$65
CUS_NCR	Customer's credit card classification rating	Char	100	1B
IM_AVGBAL	Average bank balance	Currency	8	$90
CC_SEGMENT	Credit card segment	Numeric	8	1
CC_pymnt_cat	Credit card	Numeric	8	2
NUM_CC	Number of credit cards	Numeric	8	1
CC_INT_AMT	Sum of interest charged to customer	Numeric	8	$0.71
CC_LATE_FEE	Sum of late fees charged to customer	Numeric	8	$0.0
CC_LIMIT	Sum of all ANZ credit card limits	Currency	8	$9,000
CC_AVGBL	Sum of average balances across all ANZ credit cards	Numeric	8	$-483
CC_AVG_UTIL	Average utilisation of all ANZ credit cards	Numeric	8	98%
CC_CA_AMT	Dollar amount of cash advances on all ANZ credit cards	Currency	8	$1000
OD_LIMIT	Sum of all overdraft limits	Currency	8	$4000
OD_AVGBL	Sum of average balances of all overdrafts	Currency	8	$346
OD_AVG_UTIL	Average utilisation of all overdrafts	Numeric	8	54%
FLEXI_AVGBL	Sum of average balances of all flexible home loans	Currency	8	$2,304
FLEXI_AVG_UTIL	Average utilisation of all flexible home loans	Numeric	8	0%
NUM_HMLS	Number of home loans	Numeric	8	1
NUM_FIXED_HMLS	Number of fixed home loans	Numeric	8	0
NUM_FLOAT_HMLS	Number of floating home loans	Numeric	8	0
NUM_OI_HMLS	Number of interest only home loans	Numeric	8	0
NUM_DELQ_HMLS	Number of delinquent home loans (when repayments aren't paid)	Numeric	8	0
PROPERTY_VAL_TOTAL	Value of all property	Numeric	8	$500,000
NUM_PLN	Number of personal loans	Numeric	8	1
NUM_PLN_ARRANGEMENTS	Number of personal home loans with customer financial wellbeing arrangements	Numeric	8	0
NUM_DELQ_PLNS	Number of delinquent personal loans	Numeric	8	0
PLN_APP_AMT_TOT	Personal loan original drawdown amount	Currency	88	$100,000
PLN_PCT_PAYED_TOT	Percent of personal loan paid off	Numeric	8	2%
LVR	Loan to value ration	Numeric	8	45%
UNSECURED_EOM_FUM	Sum of all funds	Currency	8	$4,051

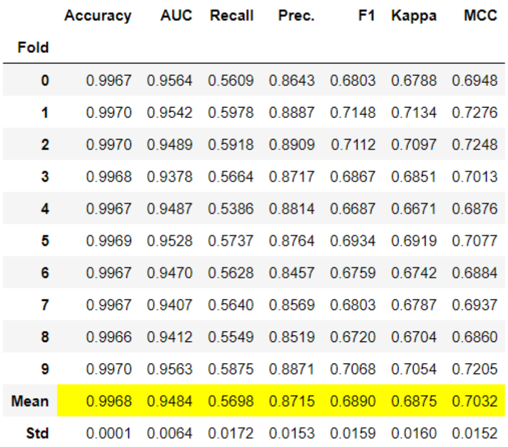
APPENDIX C. ALL TRIAL MODEL RESULTS

7. FINANCIAL DIFFICULTY



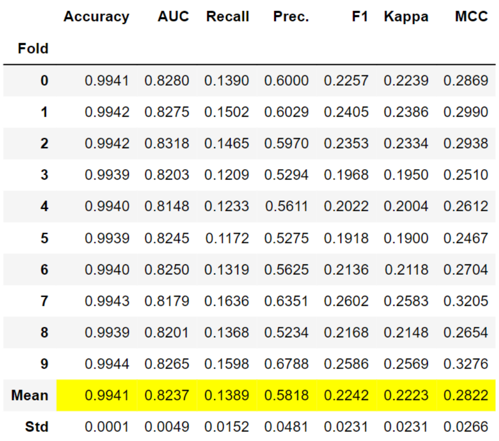
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9945	0.7869	0.5766	0.5578	0.5670	0.5643	0.5644
1	0.9950	0.8061	0.6147	0.5939	0.6042	0.6016	0.6017
2	0.9947	0.8083	0.6196	0.5631	0.5900	0.5873	0.5880
3	0.9947	0.7957	0.5942	0.5701	0.5819	0.5792	0.5794
4	0.9945	0.7818	0.5664	0.5531	0.5597	0.5569	0.5569
5	0.9948	0.7981	0.5990	0.5747	0.5866	0.5840	0.5841
6	0.9949	0.8066	0.6159	0.5842	0.5996	0.5971	0.5973
7	0.9946	0.8028	0.6087	0.5581	0.5823	0.5796	0.5802
8	0.9944	0.7826	0.5682	0.5464	0.5571	0.5542	0.5544
9	0.9946	0.8013	0.6055	0.5615	0.5827	0.5800	0.5804
Mean	0.9947	0.7970	0.5969	0.5663	0.5811	0.5784	0.5787
Std	0.0002	0.0095	0.0189	0.0138	0.0149	0.0150	0.0151

Figure 3: Decision tree model raw results.



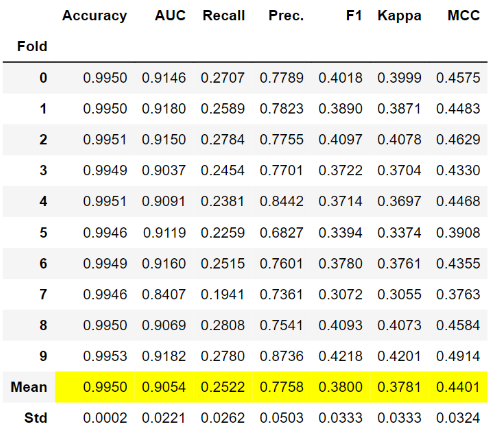
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9967	0.9564	0.5609	0.8643	0.6803	0.6788	0.6948
1	0.9970	0.9542	0.5978	0.8887	0.7148	0.7134	0.7276
2	0.9970	0.9489	0.5918	0.8909	0.7112	0.7097	0.7248
3	0.9968	0.9378	0.5664	0.8717	0.6867	0.6851	0.7013
4	0.9967	0.9487	0.5386	0.8814	0.6687	0.6671	0.6876
5	0.9969	0.9528	0.5737	0.8764	0.6934	0.6919	0.7077
6	0.9967	0.9470	0.5628	0.8457	0.6759	0.6742	0.6884
7	0.9967	0.9407	0.5640	0.8569	0.6803	0.6787	0.6937
8	0.9966	0.9412	0.5549	0.8519	0.6720	0.6704	0.6860
9	0.9970	0.9563	0.5875	0.8871	0.7068	0.7054	0.7205
Mean	0.9968	0.9484	0.5698	0.8715	0.6890	0.6875	0.7032
Std	0.0001	0.0064	0.0172	0.0153	0.0159	0.0160	0.0152

Figure 4: Random forest model raw results.



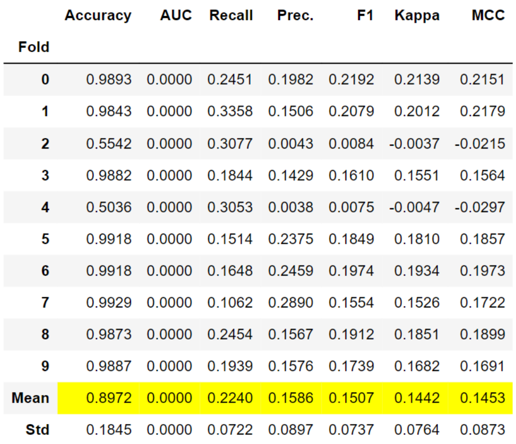
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9941	0.8280	0.1390	0.6000	0.2257	0.2239	0.2869
1	0.9942	0.8275	0.1502	0.6029	0.2405	0.2386	0.2990
2	0.9942	0.8318	0.1465	0.5970	0.2353	0.2334	0.2938
3	0.9939	0.8203	0.1209	0.5294	0.1968	0.1950	0.2510
4	0.9940	0.8148	0.1233	0.5611	0.2022	0.2004	0.2612
5	0.9939	0.8245	0.1172	0.5275	0.1918	0.1900	0.2467
6	0.9940	0.8250	0.1319	0.5625	0.2136	0.2118	0.2704
7	0.9943	0.8179	0.1636	0.6351	0.2602	0.2583	0.3205
8	0.9939	0.8201	0.1368	0.5234	0.2168	0.2148	0.2654
9	0.9944	0.8265	0.1598	0.6788	0.2586	0.2569	0.3276
Mean	0.9941	0.8237	0.1389	0.5818	0.2242	0.2223	0.2822
Std	0.0001	0.0049	0.0152	0.0481	0.0231	0.0231	0.0266

Figure 5: Linear discriminant analysis (LDA) model raw results.



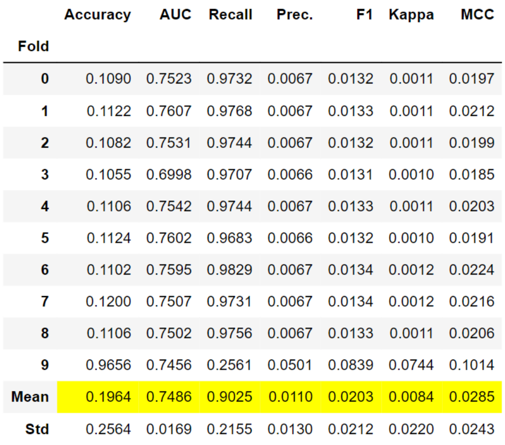
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9950	0.9146	0.2707	0.7789	0.4018	0.3999	0.4575
1	0.9950	0.9180	0.2589	0.7823	0.3890	0.3871	0.4483
2	0.9951	0.9150	0.2784	0.7755	0.4097	0.4078	0.4629
3	0.9949	0.9037	0.2454	0.7701	0.3722	0.3704	0.4330
4	0.9951	0.9091	0.2381	0.8442	0.3714	0.3697	0.4468
5	0.9946	0.9119	0.2259	0.6827	0.3394	0.3374	0.3908
6	0.9949	0.9160	0.2515	0.7601	0.3780	0.3761	0.4355
7	0.9946	0.8407	0.1941	0.7361	0.3072	0.3055	0.3763
8	0.9950	0.9069	0.2808	0.7541	0.4093	0.4073	0.4584
9	0.9953	0.9182	0.2780	0.8736	0.4218	0.4201	0.4914
Mean	0.9950	0.9054	0.2522	0.7758	0.3800	0.3781	0.4401
Std	0.0002	0.0221	0.0262	0.0503	0.0333	0.0333	0.0324

Figure 6: Gradient boosting classifier model raw results.



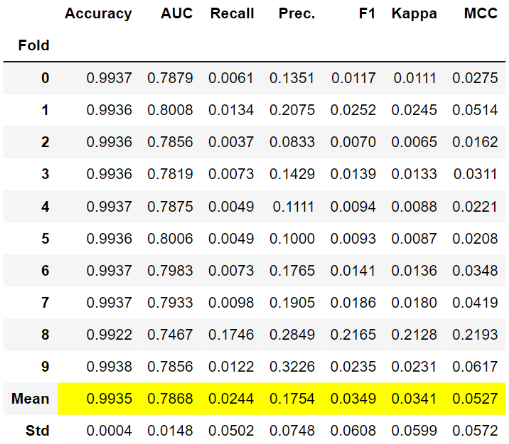
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9893	0.0000	0.2451	0.1982	0.2192	0.2139	0.2151
1	0.9843	0.0000	0.3358	0.1506	0.2079	0.2012	0.2179
2	0.5542	0.0000	0.3077	0.0043	0.0084	-0.0037	-0.0215
3	0.9882	0.0000	0.1844	0.1429	0.1610	0.1551	0.1564
4	0.5036	0.0000	0.3053	0.0038	0.0075	-0.0047	-0.0297
5	0.9918	0.0000	0.1514	0.2375	0.1849	0.1810	0.1857
6	0.9918	0.0000	0.1648	0.2459	0.1974	0.1934	0.1973
7	0.9929	0.0000	0.1062	0.2890	0.1554	0.1526	0.1722
8	0.9873	0.0000	0.2454	0.1567	0.1912	0.1851	0.1899
9	0.9887	0.0000	0.1939	0.1576	0.1739	0.1682	0.1691
Mean	0.8972	0.0000	0.2240	0.1586	0.1507	0.1442	0.1453
Std	0.1845	0.0000	0.0722	0.0897	0.0737	0.0764	0.0873

Figure 7: Support vector machine model raw results.



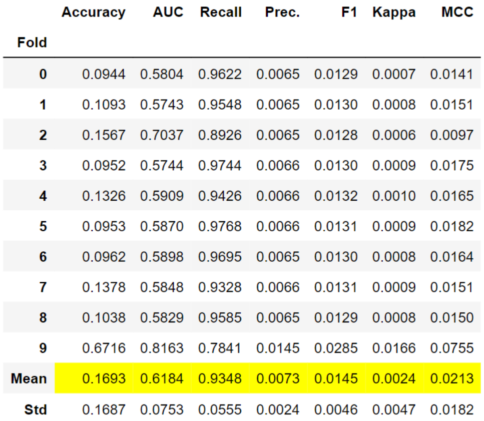
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.1090	0.7523	0.9732	0.0067	0.0132	0.0011	0.0197
1	0.1122	0.7607	0.9768	0.0067	0.0133	0.0011	0.0212
2	0.1082	0.7531	0.9744	0.0067	0.0132	0.0011	0.0199
3	0.1055	0.6998	0.9707	0.0066	0.0131	0.0010	0.0185
4	0.1106	0.7542	0.9744	0.0067	0.0133	0.0011	0.0203
5	0.1124	0.7602	0.9683	0.0066	0.0132	0.0010	0.0191
6	0.1102	0.7595	0.9829	0.0067	0.0134	0.0012	0.0224
7	0.1200	0.7507	0.9731	0.0067	0.0134	0.0012	0.0216
8	0.1106	0.7502	0.9756	0.0067	0.0133	0.0011	0.0206
9	0.9656	0.7456	0.2561	0.0501	0.0839	0.0744	0.1014
Mean	0.1964	0.7486	0.9025	0.0110	0.0203	0.0084	0.0285
Std	0.2564	0.0169	0.2155	0.0130	0.0212	0.0220	0.0243

Figure 8: Naïve Bayes model raw results.



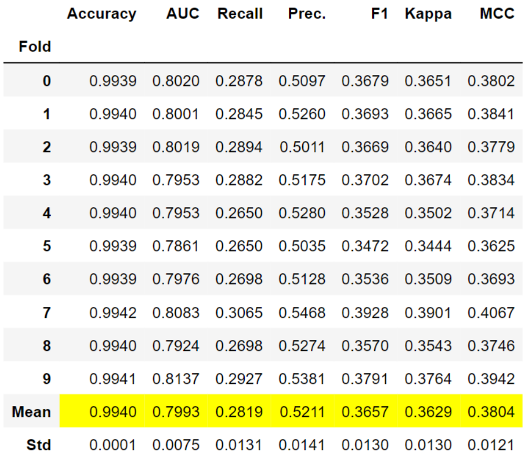
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9937	0.7879	0.0061	0.1351	0.0117	0.0111	0.0275
1	0.9936	0.8008	0.0134	0.2075	0.0252	0.0245	0.0514
2	0.9936	0.7856	0.0037	0.0833	0.0070	0.0065	0.0162
3	0.9936	0.7819	0.0073	0.1429	0.0139	0.0133	0.0311
4	0.9937	0.7875	0.0049	0.1111	0.0094	0.0088	0.0221
5	0.9936	0.8006	0.0049	0.1000	0.0093	0.0087	0.0208
6	0.9937	0.7983	0.0073	0.1765	0.0141	0.0136	0.0348
7	0.9937	0.7933	0.0098	0.1905	0.0186	0.0180	0.0419
8	0.9922	0.7467	0.1746	0.2849	0.2165	0.2128	0.2193
9	0.9938	0.7856	0.0122	0.3226	0.0235	0.0231	0.0617
Mean	0.9935	0.7868	0.0244	0.1754	0.0349	0.0341	0.0527
Std	0.0004	0.0148	0.0502	0.0748	0.0608	0.0599	0.0572

Figure 9: Logistic regression model raw results.



	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.0944	0.5804	0.9622	0.0065	0.0129	0.0007	0.0141
1	0.1093	0.5743	0.9548	0.0065	0.0130	0.0008	0.0151
2	0.1567	0.7037	0.8926	0.0065	0.0128	0.0006	0.0097
3	0.0952	0.5744	0.9744	0.0066	0.0130	0.0009	0.0175
4	0.1326	0.5909	0.9426	0.0066	0.0132	0.0010	0.0165
5	0.0953	0.5870	0.9768	0.0066	0.0131	0.0009	0.0182
6	0.0962	0.5898	0.9695	0.0065	0.0130	0.0008	0.0164
7	0.1378	0.5848	0.9328	0.0066	0.0131	0.0009	0.0151
8	0.1038	0.5829	0.9585	0.0065	0.0129	0.0008	0.0150
9	0.6716	0.8163	0.7841	0.0145	0.0285	0.0166	0.0755
Mean	0.1693	0.6184	0.9348	0.0073	0.0145	0.0024	0.0213
Std	0.1687	0.0753	0.0555	0.0024	0.0046	0.0047	0.0182

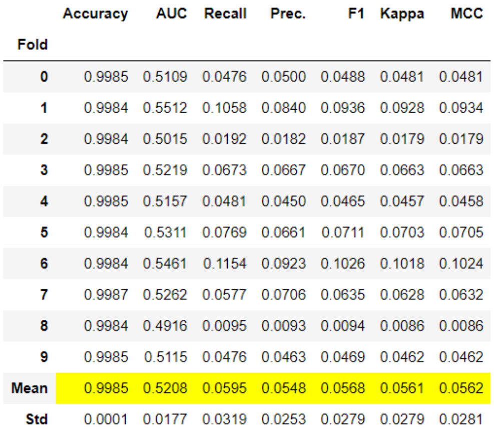
Figure 10: Quadratic discriminant classifier model raw results.



	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9939	0.8020	0.2878	0.5097	0.3679	0.3651	0.3802
1	0.9940	0.8001	0.2845	0.5260	0.3693	0.3665	0.3841
2	0.9939	0.8019	0.2894	0.5011	0.3669	0.3640	0.3779
3	0.9940	0.7953	0.2882	0.5175	0.3702	0.3674	0.3834
4	0.9940	0.7953	0.2650	0.5280	0.3528	0.3502	0.3714
5	0.9939	0.7861	0.2650	0.5035	0.3472	0.3444	0.3625
6	0.9939	0.7976	0.2698	0.5128	0.3536	0.3509	0.3693
7	0.9942	0.8083	0.3065	0.5468	0.3928	0.3901	0.4067
8	0.9940	0.7924	0.2698	0.5274	0.3570	0.3543	0.3746
9	0.9941	0.8137	0.2927	0.5381	0.3791	0.3764	0.3942
Mean	0.9940	0.7993	0.2819	0.5211	0.3657	0.3629	0.3804
Std	0.0001	0.0075	0.0131	0.0141	0.0130	0.0130	0.0121

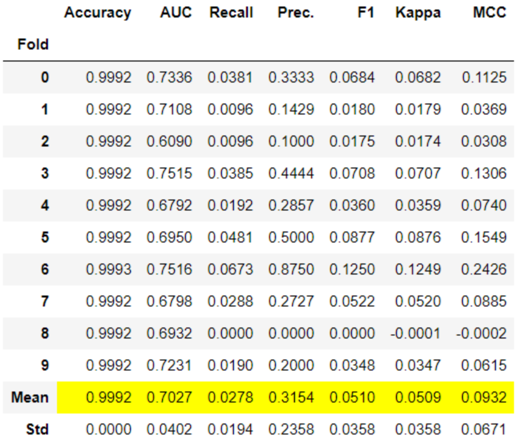
Figure 11: Extra trees classifier model raw results.

8. FINANCIAL HARDSHIP



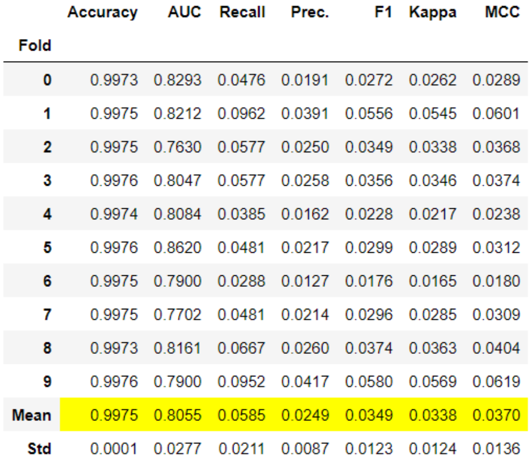
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9985	0.5109	0.0476	0.0500	0.0488	0.0481	0.0481
1	0.9984	0.5512	0.1058	0.0840	0.0936	0.0928	0.0934
2	0.9984	0.5015	0.0192	0.0182	0.0187	0.0179	0.0179
3	0.9985	0.5219	0.0673	0.0667	0.0670	0.0663	0.0663
4	0.9985	0.5157	0.0481	0.0450	0.0465	0.0457	0.0458
5	0.9984	0.5311	0.0769	0.0661	0.0711	0.0703	0.0705
6	0.9984	0.5461	0.1154	0.0923	0.1026	0.1018	0.1024
7	0.9987	0.5262	0.0577	0.0706	0.0635	0.0628	0.0632
8	0.9984	0.4916	0.0095	0.0093	0.0094	0.0086	0.0086
9	0.9985	0.5115	0.0476	0.0463	0.0469	0.0462	0.0462
Mean	0.9985	0.5208	0.0595	0.0548	0.0568	0.0561	0.0562
Std	0.0001	0.0177	0.0319	0.0253	0.0279	0.0279	0.0281

Figure 12: Decision tree model raw results.



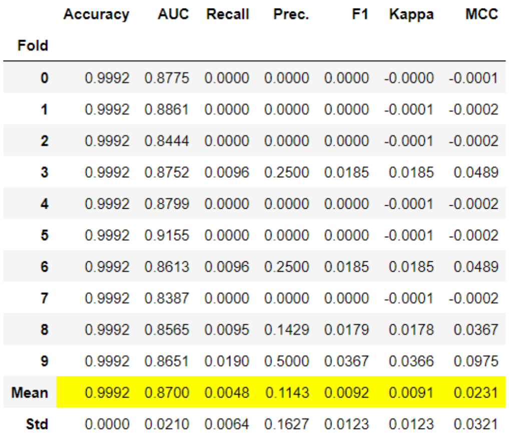
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9992	0.7336	0.0381	0.3333	0.0684	0.0682	0.1125
1	0.9992	0.7108	0.0096	0.1429	0.0180	0.0179	0.0369
2	0.9992	0.6090	0.0096	0.1000	0.0175	0.0174	0.0308
3	0.9992	0.7515	0.0385	0.4444	0.0708	0.0707	0.1306
4	0.9992	0.6792	0.0192	0.2857	0.0360	0.0359	0.0740
5	0.9992	0.6950	0.0481	0.5000	0.0877	0.0876	0.1549
6	0.9993	0.7516	0.0673	0.8750	0.1250	0.1249	0.2426
7	0.9992	0.6798	0.0288	0.2727	0.0522	0.0520	0.0885
8	0.9992	0.6932	0.0000	0.0000	0.0000	-0.0001	-0.0002
9	0.9992	0.7231	0.0190	0.2000	0.0348	0.0347	0.0615
Mean	0.9992	0.7027	0.0278	0.3154	0.0510	0.0509	0.0932
Std	0.0000	0.0402	0.0194	0.2358	0.0358	0.0358	0.0671

Figure 13: Random forest model raw results.



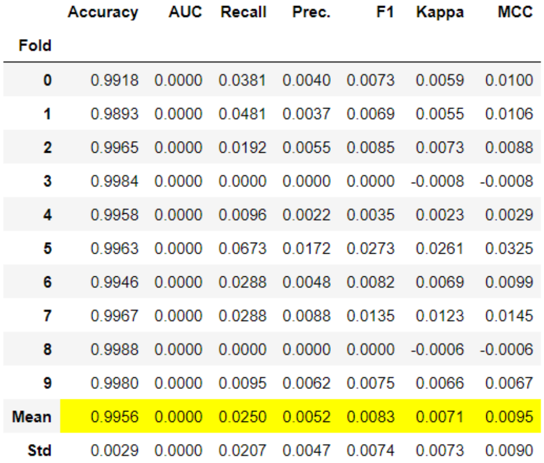
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9973	0.8293	0.0476	0.0191	0.0272	0.0262	0.0289
1	0.9975	0.8212	0.0962	0.0391	0.0556	0.0545	0.0601
2	0.9975	0.7630	0.0577	0.0250	0.0349	0.0338	0.0368
3	0.9976	0.8047	0.0577	0.0258	0.0356	0.0346	0.0374
4	0.9974	0.8084	0.0385	0.0162	0.0228	0.0217	0.0238
5	0.9976	0.8620	0.0481	0.0217	0.0299	0.0289	0.0312
6	0.9975	0.7900	0.0288	0.0127	0.0176	0.0165	0.0180
7	0.9975	0.7702	0.0481	0.0214	0.0296	0.0285	0.0309
8	0.9973	0.8161	0.0667	0.0260	0.0374	0.0363	0.0404
9	0.9976	0.7900	0.0952	0.0417	0.0580	0.0569	0.0619
Mean	0.9975	0.8055	0.0585	0.0249	0.0349	0.0338	0.0370
Std	0.0001	0.0277	0.0211	0.0087	0.0123	0.0124	0.0136

Figure 14: Linear discriminant analysis (LDA) model raw results.



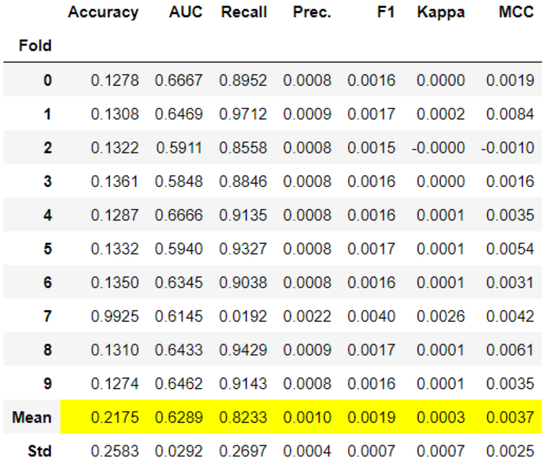
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9992	0.8775	0.0000	0.0000	0.0000	-0.0000	-0.0001
1	0.9992	0.8861	0.0000	0.0000	0.0000	-0.0001	-0.0002
2	0.9992	0.8444	0.0000	0.0000	0.0000	-0.0001	-0.0002
3	0.9992	0.8752	0.0096	0.2500	0.0185	0.0185	0.0489
4	0.9992	0.8799	0.0000	0.0000	0.0000	-0.0001	-0.0002
5	0.9992	0.9155	0.0000	0.0000	0.0000	-0.0001	-0.0002
6	0.9992	0.8613	0.0096	0.2500	0.0185	0.0185	0.0489
7	0.9992	0.8387	0.0000	0.0000	0.0000	-0.0001	-0.0002
8	0.9992	0.8565	0.0095	0.1429	0.0179	0.0178	0.0367
9	0.9992	0.8651	0.0190	0.5000	0.0367	0.0366	0.0975
Mean	0.9992	0.8700	0.0048	0.1143	0.0092	0.0091	0.0231
Std	0.0000	0.0210	0.0064	0.1627	0.0123	0.0123	0.0321

Figure 15: Gradient boosting classifier model raw results.



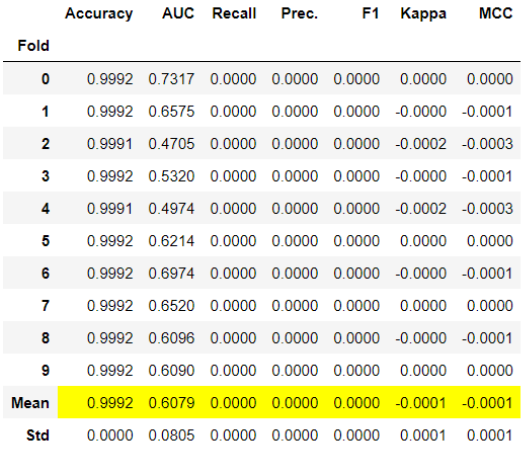
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9918	0.0000	0.0381	0.0040	0.0073	0.0059	0.0100
1	0.9893	0.0000	0.0481	0.0037	0.0069	0.0055	0.0106
2	0.9965	0.0000	0.0192	0.0055	0.0085	0.0073	0.0088
3	0.9984	0.0000	0.0000	0.0000	0.0000	-0.0008	-0.0008
4	0.9958	0.0000	0.0096	0.0022	0.0035	0.0023	0.0029
5	0.9963	0.0000	0.0673	0.0172	0.0273	0.0261	0.0325
6	0.9946	0.0000	0.0288	0.0048	0.0082	0.0069	0.0099
7	0.9967	0.0000	0.0288	0.0088	0.0135	0.0123	0.0145
8	0.9988	0.0000	0.0000	0.0000	0.0000	-0.0006	-0.0006
9	0.9980	0.0000	0.0095	0.0062	0.0075	0.0066	0.0067
Mean	0.9956	0.0000	0.0250	0.0052	0.0083	0.0071	0.0095
Std	0.0029	0.0000	0.0207	0.0047	0.0074	0.0073	0.0090

Figure 16:Support vector machine model raw results.



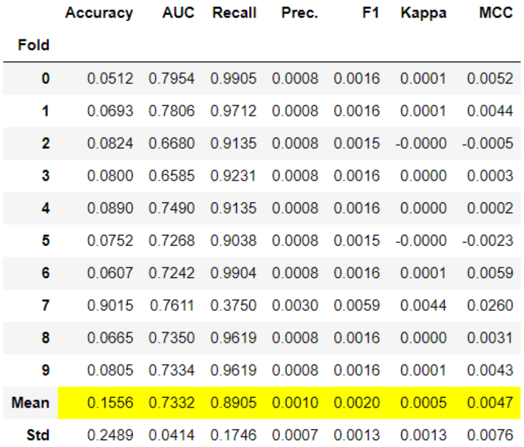
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.1278	0.6667	0.8952	0.0008	0.0016	0.0000	0.0019
1	0.1308	0.6469	0.9712	0.0009	0.0017	0.0002	0.0084
2	0.1322	0.5911	0.8558	0.0008	0.0015	-0.0000	-0.0010
3	0.1361	0.5848	0.8846	0.0008	0.0016	0.0000	0.0016
4	0.1287	0.6666	0.9135	0.0008	0.0016	0.0001	0.0035
5	0.1332	0.5940	0.9327	0.0008	0.0017	0.0001	0.0054
6	0.1350	0.6345	0.9038	0.0008	0.0016	0.0001	0.0031
7	0.9925	0.6145	0.0192	0.0022	0.0040	0.0026	0.0042
8	0.1310	0.6433	0.9429	0.0009	0.0017	0.0001	0.0061
9	0.1274	0.6462	0.9143	0.0008	0.0016	0.0001	0.0035
Mean	0.2175	0.6289	0.8233	0.0010	0.0019	0.0003	0.0037
Std	0.2583	0.0292	0.2697	0.0004	0.0007	0.0007	0.0025

Figure 17: Naïve Bayes model raw results.



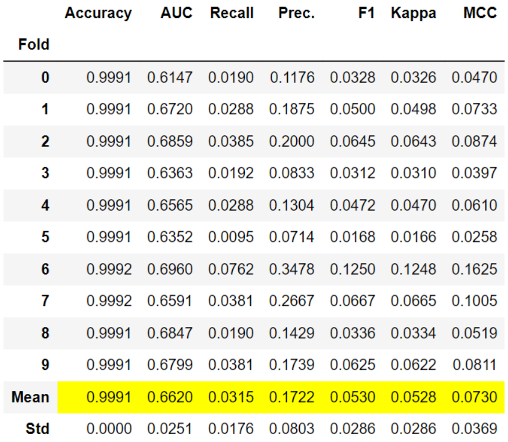
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9992	0.7317	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.9992	0.6575	0.0000	0.0000	0.0000	-0.0000	-0.0001
2	0.9991	0.4705	0.0000	0.0000	0.0000	-0.0002	-0.0003
3	0.9992	0.5320	0.0000	0.0000	0.0000	-0.0000	-0.0001
4	0.9991	0.4974	0.0000	0.0000	0.0000	-0.0002	-0.0003
5	0.9992	0.6214	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.9992	0.6974	0.0000	0.0000	0.0000	-0.0000	-0.0001
7	0.9992	0.6520	0.0000	0.0000	0.0000	0.0000	0.0000
8	0.9992	0.6096	0.0000	0.0000	0.0000	-0.0000	-0.0001
9	0.9992	0.6090	0.0000	0.0000	0.0000	0.0000	0.0000
Mean	0.9992	0.6079	0.0000	0.0000	0.0000	-0.0001	-0.0001
Std	0.0000	0.0805	0.0000	0.0000	0.0000	0.0001	0.0001

Figure 18: Logistic regression model raw results.



	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.0512	0.7954	0.9905	0.0008	0.0016	0.0001	0.0052
1	0.0693	0.7806	0.9712	0.0008	0.0016	0.0001	0.0044
2	0.0824	0.6680	0.9135	0.0008	0.0015	-0.0000	-0.0005
3	0.0800	0.6585	0.9231	0.0008	0.0016	0.0000	0.0003
4	0.0890	0.7490	0.9135	0.0008	0.0016	0.0000	0.0002
5	0.0752	0.7268	0.9038	0.0008	0.0015	-0.0000	-0.0023
6	0.0607	0.7242	0.9904	0.0008	0.0016	0.0001	0.0059
7	0.9015	0.7611	0.3750	0.0030	0.0059	0.0044	0.0260
8	0.0665	0.7350	0.9619	0.0008	0.0016	0.0000	0.0031
9	0.0805	0.7334	0.9619	0.0008	0.0016	0.0001	0.0043
Mean	0.1556	0.7332	0.8905	0.0010	0.0020	0.0005	0.0047
Std	0.2489	0.0414	0.1746	0.0007	0.0013	0.0013	0.0076

Figure 19: Quadratic discriminant classifier model raw results.



	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9991	0.6147	0.0190	0.1176	0.0328	0.0326	0.0470
1	0.9991	0.6720	0.0288	0.1875	0.0500	0.0498	0.0733
2	0.9991	0.6859	0.0385	0.2000	0.0645	0.0643	0.0874
3	0.9991	0.6363	0.0192	0.0833	0.0312	0.0310	0.0397
4	0.9991	0.6565	0.0288	0.1304	0.0472	0.0470	0.0610
5	0.9991	0.6352	0.0095	0.0714	0.0168	0.0166	0.0258
6	0.9992	0.6960	0.0762	0.3478	0.1250	0.1248	0.1625
7	0.9992	0.6591	0.0381	0.2667	0.0667	0.0665	0.1005
8	0.9991	0.6847	0.0190	0.1429	0.0336	0.0334	0.0519
9	0.9991	0.6799	0.0381	0.1739	0.0625	0.0622	0.0811
Mean	0.9991	0.6620	0.0315	0.1722	0.0530	0.0528	0.0730
Std	0.0000	0.0251	0.0176	0.0803	0.0286	0.0286	0.0369

Figure 20: Extra trees classifier model raw results.

APPENDIX D. CORRELATION MATRICES FOR FEATURE REDUCTION

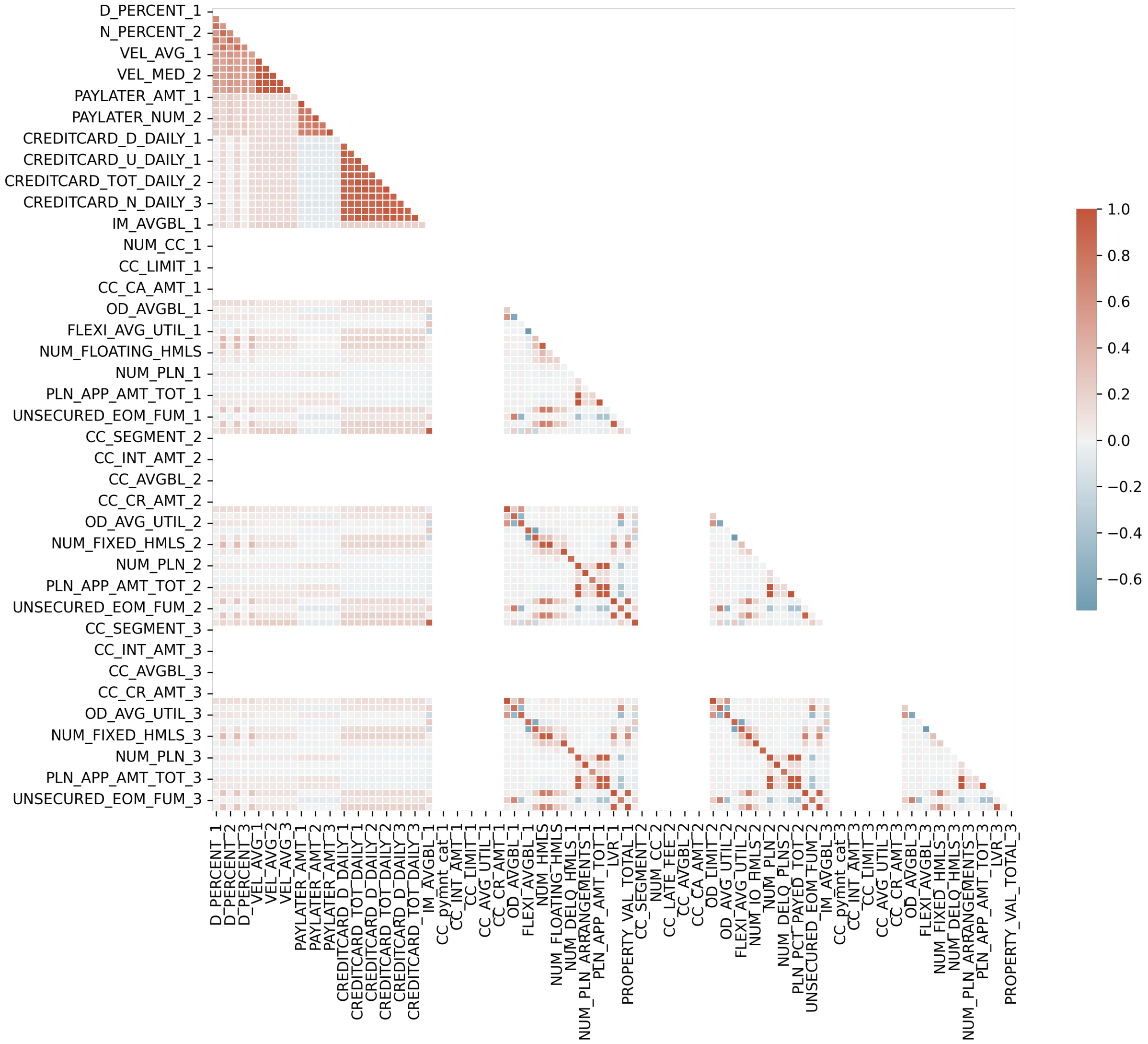


Figure 21: Correlation value between all variables in original dataset before feature reduction techniques.