

2023-24

CMSE11428 PREDICTIVE ANALYTICS AND MODELLING OF DATA

ASSIGNMENT NAME

GROUP 4

Word count:

**Introduction**

In the dynamic world of banking and finance, credit scoring, described as gathering, evaluating, and categorising various credit aspects and characteristics in order to evaluate credit decisions is considered to be a core appraising tool (Abdou and Pointon, 2011).

Therefore, the purpose of this report is to develop a predictive classification model in Python to evaluate the default risk of bank customers in an effective manner. Additionally, the report will offer a detailed summary of how the predictive model was selected and created, providing clear insight on justification for processes such as data pre-processing, outlier removal, predictor selection as well as any other relevant methodological choices. Academic literature will be applied throughout to further supplement any decisions by providing additional domain knowledge.

Furthermore, the findings of the model will be discussed and analysed in a detailed but comprehensible manner to both banking professionals as well as a less technically proficient audience that has interest in the financial sector.

Lastly, targeted recommendations applicable to banks based on the predictive model’s output will additionally be highlighted to improve future credit scoring processes by furthering the fine tuning of models to better account for particularly influencing customer features uncovered throughout the report.

**Problem Statement**

Current or ‘traditional’ credit scoring models can be considered somewhat one-dimensional or narrow given that they mainly focus on historical financial data such as loan repayments and spending patterns. Simultaneously, they are overlooking progressively more common non-traditional variables such as residential or work location. This is important given that not all bank customers have detailed credit histories, leading to complications.

These issues can be translated into various implications. From the perspective of customers, false negatives can lead to the unjust denial of loans. For banks, this can have an impact on their decision-making processes as overestimation and underestimation of customer risk can endanger the bank’s reputation and overall portfolio health. Moreover, the growing digital economy is made up of much more complicated data that make up a borrower’s creditworthiness, unable to be entirely explained by only using traditional approaches.

Therefore, in light of these concerns, there is a need to create a predictive credit scoring model that fundamentally acts as an improved version of current existing models but also incorporates advanced analytical techniques, providing a mutual benefit to both the lender (bank) and the borrower (customer). This means that it should primarily be as accurate as possible, without falling subject to biases, and be inclusive for all demographics, while remaining adaptable to the large and constantly updated financial datasets. Moreover, the model must be able to handle large volumes of all sorts of data types, keeping pace with the increasing demands of this data-driven field.

A successful creation and implementation of a model with these aforementioned characteristics will provide the bank in question with a versatile tool to better assess their customers while also ensuring that customers that were previously rejected from traditional models are not at a disadvantage compared to others.

**Literature Review and Justification for Methodology**

The methodology employed in this report was aimed at: (a) ensuring that the appropriate procedures were followed in data handling, model creation and evaluation; and (b) providing thorough and understandable justification through literature, theory, and domain knowledge for all methodological choices. In cases of any assumptions, ample research and rational thinking were used to ensure minimal impact and sway on the model’s performance and interpretation.

Given that the dataset was already provided, data collection was not performed and instead the first task was targeted at exploring, cleaning and processing the data in an appropriate manner to be usable in the model.

Initial feature selection was based on two main criteria: relevance to the scope and data quality and availability. Various methods were taken to satisfy these criteria to enable a holistic approach that minimised the possibility of missing something of importance. To narrow down the amount of usable predictor variables, we followed a reasonable rule of thumb that any predictors with a missing value percentage >50% were deemed unusable and as Lakshminarayan, Harp and Samad (1999) point out, a solution of reducing the dataset given its ample size was therefore suitable. This was on the grounds that retaining the data or attempting to replace values can lead to bias arising from discrepancies between complete and missing data, loss of efficiency, and difficulties in managing and interpreting the data (Luengo, 2011).

In instances where a predictor had unknown encoding labels (e.g. marital status, credit card type), but was useful either due to its frequent use in other credit scoring models or due to its nature, literature was drawn upon to help understand common encoding practices and therefore make better encoding assumptions. More specifically, the marital status variable used in the model was assumed to be encoded as 0=Unknown; 1=Single; 2=Union of Fact; 3=Married without community property; 4=Married in community of purchased property; 5=Separated; 6=Divorced; 7=Widowed, following the example of another logistic regression model in credit scoring performed by Costa e Silva et al. (2020). For binary variables,

Missing valuesà Kaiser, J., 2014. Dealing with Missing Values in Data. *Journal Of Systems Integration (1804-2724)*, *5*(1).

**Data Preprocessing**

Based on previous considerations, we selected variables below to later use them in our model.

Table X

|  |  |  |
| --- | --- | --- |
| **Var\_Title** | **Var\_Description** | **Field\_Content** |
| PAYMENT\_DAY | Day of the month for bill payment, chosen by t... | 1,5,10,15,20,25 |
| APPLICATION\_SUBMISSION\_TYPE | Indicates if the application was submitted via... | Web, Carga |
| POSTAL\_ADDRESS\_TYPE | Indicates if the address for posting is the ho... | 1.2 |
| MARITAL\_STATUS | Encoding not informed | 1,2,3,4,5,6,7 |
| QUANT\_DEPENDANTS | NaN | 0, 1, 2, ... |
| CITY\_OF\_BIRTH | NaN | NaN |
| RESIDENCIAL\_STATE | State of residence | NaN |
| RESIDENCIAL\_BOROUGH | Borough of residence | NaN |
| FLAG\_RESIDENCIAL\_PHONE | Indicates if the applicant possesses a home phone | Y,N |
| MONTHS\_IN\_RESIDENCE | Time in the current residence in months | 1,2,... , NULL |
| FLAG\_MOBILE\_PHONE | Indicates if the applicant possesses a mobile ... | Y,N |
| FLAG\_EMAIL | Indicates if the applicant possesses an e-mail... | 0.1 |
| PERSONAL\_MONTHLY\_INCOME | Applicant's personal regular monthly income in... | NaN |
| OTHER\_INCOMES | Applicant's other incomes monthly averaged in ... | NaN |
| FLAG\_VISA | Flag indicating if the applicant is a VISA cre... | 0.1 |
| FLAG\_MASTERCARD | Flag indicating if the applicant is a MASTERCA... | 0.1 |
| FLAG\_DINERS | Flag indicating if the applicant is a SINERS c... | 0.1 |
| FLAG\_AMERICAN\_EXPRESS | Flag indicating if the applicant is an AMERICA... | 0.1 |
| QUANT\_BANKING\_ACCOUNTS | NaN | 0,1,2 |
| QUANT\_SPECIAL\_BANKING\_ACCOUNTS | NaN | 0,1,2 |
| PERSONAL\_ASSETS\_VALUE | Total value of the personal possessions such a... | NaN |
| QUANT\_CARS | Quantity of cars the applicant possesses | NaN |
| COMPANY | If the applicant has supplied the name of the ... | Y,N |
| FLAG\_PROFESSIONAL\_PHONE | Indicates if the professional phone number was... | Y,N |
| MONTHS\_IN\_THE\_JOB | Time in the current job in months | NaN |
| FLAG\_ACSP\_RECORD | Flag indicating if the applicant has any previ... | Y, N |
| AGE | Applicant's age at the moment of submission | NaN |

**Missing Value Handling**

1. **PAYMENT\_DAY**

In the PAYMENT\_DAY variable, there is a clear outlier of -99999, which is clearly not a valid payment day. We assumed that this is a placeholder for a missing or unknown payment day. Since we don't have additional information to determine which day might be the most appropriate replacement value, we chose to use the mode of PAYMENT\_DAY to replace the -99999 value.

1. **MONTHS\_IN\_RESIDENCE**

The MONTHS\_IN\_RESIDENCE variable has unique values ranging from 0 to 228 months. Most of the values are reasonable lengths of residence, but the maximum value of 228 months, or 19 years, while possible, may be noticeable. However, we can keep these data as they may represent the reality of long-term residents.

In this dataset, the median is 6 months, which means that half of the applicants have lived in their current residence for less than or equal to 6 months. This is a reasonable value to represent a typical length of residence.

Moreover, we find that there are 3282 of "0" value in this predictor, we are considered if this indicate collection error, so we just use "NATIONALITY" ( we assume that 1 - represent Brazilian nationality) to help testing whether they are potential new immigrants or the Brazilian people with collection error, for people not brazil, we keep zero and make assumption that they are new migrants, in constrast,we use median to replace 0

1. **APPLICATION\_SUBMISSION\_TYPE**

for the application submission type, there are value of 0, which may indicate the missing value, as there are large percentile for the 0 value, there are two reasons to replace 0 with "Unkown". This clearly communicates to anyone analysing the data that these values are missing or unknown.

By using a separate category for missing values, we can avoid potentially misleading interpretations that could arise from assigning missing values to an existing category like "Web" This approach also maintains transparency in your data handling process, which is essential for reproducibility and understanding by others.

1. **MARITAL\_STATUS**

In the MARITAL\_STATUS variable, we see that there are eight different values from 0 to 7. Based on the information you provided for the variable, we have marital statuses coded 1 through 7, but there is no mention of 0. Therefore, it is reasonable to assume that 0 represents a missing or incorrect value.

Since we don't have detailed instructions to guide us on what to do with 0, I'm inclined to take the first strategy of replacing 0 with the most common marital status.This way, we can keep as much data as possible while still having a complete variable for marital status. Let's make the substitution.

1. **RESIDENCIAL\_STATE**

Given that too many categories can lead to the problem of Data Sparsity and Curse of Dimensionality, we grouped the values in variable RESIDENCIAL\_STATE into five regions by the mapping below and transform it to a new variable called RESIDENCIAL\_REGION.

| **Direction** | **Region Codes** |
| --- | --- |
| Southeast | SP, RJ, MG, ES |
| South | RS, SC, PR |
| Northeast | BA, CE, PE, RN, AL, PB, SE, PI, MA |
| Central-West | GO, MT, MS, DF |
| North | PA, AP, AM, RO, RR, TO, AC |

**Redundant Information**

1. **RESIDENCIAL\_BOROUGH**

we may treat this as redundant informtion as residential region

**2. CITY\_OF\_BIRTH**

we may treat this as redundant informtion as residential region

**Sort Other Predictors**

1. **FLAG\_MOBILE\_PHONE**

all value are 0, so we decide to drop them

1. **QUANT\_SPECIAL\_BANKING\_ACCOUNTS & QUANT\_BANKING\_ACCOUNTS**

we can see that for those two predictor, they have exact same value, so we only try tp keep one

1. **NUMBER OF DEPENDENT**

it is reasonable to habe large value of zero, representing no dependent, so we keep this predictor

1. **MONTHS\_IN\_THE\_JOB**

most of observation report 0 in months in the job, and for other categories, most have value less that 5, which may cause problem of Curse of Dimensionality and Data Sparsity if we treat them as categories, if we treat them as numerical, it is highly right skewed and not make sense to have 32237 0 months, so we decide to drop

1. **FLAG\_ACSP\_RECORD**

all value are "N", no information, drop

1. **AGE**

we make assumption that people <18 is not reasonable in considering credit score, so we dedcide to drop <18

1. **NATIONALITY**

discrimination

1. **PERSONAL\_ASSETS\_VALUE**

33299 0, may be not that useful

**Outlier Handling**

we used box-wiskers plots and scatter plots to visualise all the numeric varibles' distributions

1. **"QUANT\_DEPENDANTS"**

Mannually remove values greater than 50, because it's unlikely to take care of 50 people for one person.

1. **"AGE"**

Approximately normal, use z-SCORE to remove outliers

1. **"MONTHS\_IN\_RESIDENCE"**

we will not considered those high value as outlier as they are reasonable and can include valuable information

1. **"PERSONAL\_MONTHLY\_INCOME"**

we will not considered those high value as outlier as they are reasonable and can include valuable information

1. **"OTHER\_INCOME"**

we will not considered those high value as outlier as they are reasonable and can include valuable information

**Feature Scaling**

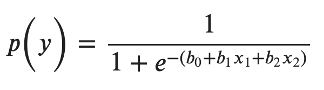
**????**

* + Explain any data preprocessing steps, like handling missing values or feature scaling, and why they are necessary for logistic regression.

1. **Introduction to Literature Review**:
   * Start with a brief overview of the importance of credit risk modeling in the banking sector.
   * Highlight the common methodologies used in the field, focusing on logistic regression.
2. **Review of Relevant Literature**:
   * Discuss previous studies that have used logistic regression in credit scoring or similar contexts.
   * Highlight their findings, particularly any strengths and weaknesses noted in these approaches.
   * Include recent studies to demonstrate the current relevance of logistic regression in this field.
3. **Justification for Using Logistic Regression**:
   * Explain why logistic regression is a suitable choice for your analysis. This could include its effectiveness in handling binary outcomes (like default/no default), its interpretability, and how it can handle a mix of numerical and categorical variables.
   * Address any limitations of logistic regression and how you plan to mitigate them in your analysis.
4. **Comparative Analysis**:
   * Compare logistic regression with other common methods used in credit scoring (like decision trees, random forests, neural networks).
   * Discuss why logistic regression is preferred over these alternatives for your specific context or data.
5. **Methodological Framework**:
   * Outline how you will implement logistic regression in your study.
   * Discuss the variables you plan to include in the model and why they are relevant.
   * Explain any data preprocessing steps, like handling missing values or feature scaling, and why they are necessary for logistic regression.
6. **Expected Outcomes and Relevance**:
   * Discuss what you expect to find from your logistic regression model.
   * Explain how these findings could be beneficial for the bank, such as improving credit scoring accuracy, reducing default risk, or enhancing customer segmentation.
7. **Conclusion**:
   * Summarize the key points from your literature review and methodology justification.
   * Reiterate the potential impact of your study on the bank’s credit scoring process.

# Logistic Regression Model Formulation

According to the following formula, the binary dependent variable for logistic regression, namely “Non-Default (0)” and “Default (1)”is transformed using the logit transformation into a probability ranging from 0 to 1, indicated by y (1).



In developing the credit scoring model, it is crucial to identify key parameters. An challenge in credit scoring lies in the imbalanced dataset, with only 26.1% falling into the 'Default' category. This bias necessitates the use of a 'balanced' class weight, adjusting weights based on class frequencies to enhance the model's ability to predict 'Default' (2). To ensure reliability, the 'newton-cholesky' solver is chosen for two main reasons. Firstly, given the data characteristics—34931 observations in the training set compared to 23 predictors—the 'newton-cholesky' method proves robust for large datasets with less predictors. It also efficiently handles the 14 binary predictors resulting from one-hot encoding (2). Secondly, the method's incorporation of Newton's approach with Cholesky Factorization, subject to a sparsity constraint, addresses data imbalance, reinforcing the model's reliability in binary classification (3).

The regularization penalty is omitted in modelling process. Most application employed regularization to mitigate overfitting risk, there is an extensive number of features relative to a limited sample size (4). Our case with only 23 predictors allows for a comprehensive analysis of all features, ensuring a robust and inclusive credit scoring model.

|  |  |  |
| --- | --- | --- |
| Solver | Penalty | class weight |
| newton-cholesky | None | balanced |

# Random Forest Model Formulation

For the random forest model, a tree-based approach is employed to recursively partition the dataset into two groups based on a specified criterion until a predetermined stopping condition is satisfied (7). Similar to logistic regression, it is crucial to acknowledge that our imbalanced dependent variable may introduce a bias favouring majority classes ("Non-default") at the expense of the minority class ("Default"), which holds particular relevance for the bank's interest. To address this, we incorporate class weights to rectify the imbalance.

In the pursuit of optimal model performance, we leverage RandomizedSearchCV with 50 iterations and a 5-fold cross-validation to identify parameter values that maximize balanced accuracy scores. Notably, this approach outperforms grid search and manual search methods by demonstrating efficacy in identifying models of comparable quality within a reduced computation time. The parameter ranges were constrained, and unselected parameters retained their default values (table1).

# Evaluation

As using inappropriate evaluation metrics can lead to unsuitable preference of model selection, especially for imbalanced class, fit in our case (11). Thus, we choose to use balanced accuracy score, false positive rate F1 scores and ROC curve and interpretability to reflect these problems.

|  |  |  |
| --- | --- | --- |
| Evaluation Metrics | Logistic Regression | Random Forest |
| Balanced accuracy score (Train) | 0.5768 | 0.6621 |
| Balanced accuracy score (Test) | 0.5749 | 0.5882 |
| False Positive Rate | 0.4512 | 0.4485 |
| Weighted F1 Scores | 0.5891 | 0.5968 |
| AUC - ROC | 0.5577 | 0.5882 |

## Balanced accuracy score

Balanced accuracy score ensures equitable consideration of both minority and majority classes during model evaluation, aligning with our objective of predicting “Default” (9). In this context, the random forest exhibits superior training set balanced accuracy at 0.6621 compared to logistic regression at 0.5768. This discrepancy aligns with mathematical proofs indicating that the error rate consistently converges with an increasing number of trees, and in our case, we employ 155 trees for model training (5).

However, from the test scores, a larger gap in the random forest at approximately 0.0739 compared to logistic regression at 0.0019, suggesting random forest has higher potential to overfitting and less generalizability on unseen dataset. Despite this, both models achieve similar test scores, with a balanced accuracy of 0.5882 for random forest and 0.5749 for logistic regression. Considering these comparable test scores, logistic regression emerges as a more suitable model for our predictive objectives.

## False Negative Rate

After consideration of the impacts associated with inaccurate predictions, we recognize that mitigating the impact of false negative important in credit scoring models. False negative indicates when model suggest approving a loan, for an individual who will actually defaults. This scenario can result in financial losses for the lending institution and fewer acceptable in credit scoring process, emphasizing our commitment alignment with the bank's objectives (10). It is evident that logistic regression outperforms in this critical metric, with rate of 0.4512 compared to the random forest's rate (0.4485). Consequently, logistic regression is deemed more aligned with our modelling objectives.

## Weighted F1 Score / AUC - ROC

Alternative measures proposed for evaluating imbalanced scenarios include the ROC curve and weighted F1 scores (Micro F1), incorporating precision and recall considerations (8, 12). For ROC curve, better model performance is indicated by a more convex shape toward the upper left, with the area under the ROC curve (AUC-ROC) serving as a comprehensive measure of overall performance—larger values being indicative of better performance (8). Weighted F1 scores, derived by averaging F1 scores per class and adjusting for class support, reflect better performance with higher F1 scores (12). Through above table, we can conclude that random forest performs better in both metrics, 0.0077 higher in weighted F1 score and 0.0305 in areas under ROC, indicating random forest are more suitable in this case.

## Interpretability

As we aim to provide recommendations for the credit scoring process, the interpretability of the model holds significance. Banks express interest in features related to defaults, necessitating the identification of feature importance. Both logistic regression and random forest allow for the determination of feature importance, utilizing beta coefficients or the Gini index. However, the random forest, while enhancing prediction accuracy, sacrifices interpretability by aggregating numerous decision trees. This renders the random forest a "black box" model, offering limited insights into the mechanics of predictions (5). Consequently, although feature importance scores can be visualized, the depth of insights is comparatively limited compared to logistic regression. Therefore, from this perspective, logistic regression is the more preferable choice.

In conclusion, based on the evaluation of above metrics, logistic regression are more suitable in our credit scoring prediction model, and we will use it to make further analysis and recommendations.

# Appendix A Potential Predictors and Feature Importance Scores

## 

## APPENDIX A.1 Logistic Regression

|  |  |
| --- | --- |
| **Variables Name** | **Beta Value** |
| PAYMENT\_DAY | 0.097763 |
| QUANT\_DEPENDANTS | 0.0303 |
| MONTHS\_IN\_RESIDENCE | -0.004 |
| PERSONAL\_MONTHLY\_INCOME | 0.0150 |
| OTHER\_INCOMES | 0.0134 |
| QUANT\_CARS | 0.0878 |
| AGE | -0.3220 |
| POSTAL\_ADDRESS\_TYPE | -0.0897 |
| FLAG\_EMAIL | -0.0417 |
| FLAG\_VISA | 0.0664 |
| FLAG\_MASTERCARD | -0.1876 |
| FLAG\_DINERS | 0.1409 |
| FLAG\_AMERICAN\_EXPRESS | -0.0675 |
| MARITAL\_STATUS | -0.0128 |
| FLAG\_RESIDENCIAL\_PHONE\_Y | -0.6021 |
| COMPANY\_Y | -0.0632 |
| FLAG\_PROFESSIONAL\_PHONE\_Y | -0.1074 |
| APPLICATION\_SUBMISSION\_TYPE\_Unknown | 0.0538 |
| APPLICATION\_SUBMISSION\_TYPE\_Web | 0.0407 |
| RESIDENCIAL\_REGION\_North | -0.0956 |
| RESIDENCIAL\_REGION\_Northeast | 0.0314 |
| RESIDENCIAL\_REGION\_South | -0.2854 |
| RESIDENCIAL\_REGION\_Southeast | -0.0329 |

## Appendix A.2 Random Forest

|  |  |
| --- | --- |
| **Variables Name** | **Beta Value** |
| PAYMENT\_DAY | 0.1014 |
| QUANT\_DEPENDANTS | 0.0306 |
| MONTHS\_IN\_RESIDENCE | 0.0792 |
| PERSONAL\_MONTHLY\_INCOME | 0.1196 |
| OTHER\_INCOMES | 0.0091 |
| QUANT\_CARS | 0.0131 |
| AGE | 0.3176 |
| POSTAL\_ADDRESS\_TYPE | 0 |
| FLAG\_EMAIL | 0.0079 |
| FLAG\_VISA | 0.0048 |
| FLAG\_MASTERCARD | 0.0081 |
| FLAG\_DINERS | 0 |
| FLAG\_AMERICAN\_EXPRESS | 0 |
| MARITAL\_STATUS | 0.0678 |
| FLAG\_RESIDENCIAL\_PHONE\_Y | 0.1036 |
| COMPANY\_Y | 0.0151 |
| FLAG\_PROFESSIONAL\_PHONE\_Y | 0.0179 |
| APPLICATION\_SUBMISSION\_TYPE\_Unknown | 0.0131 |
| APPLICATION\_SUBMISSION\_TYPE\_Web | 0.0152 |
| RESIDENCIAL\_REGION\_North | 0.0043 |
| RESIDENCIAL\_REGION\_Northeast | 0.0192 |
| RESIDENCIAL\_REGION\_South | 0.0288 |
| RESIDENCIAL\_REGION\_Southeast | 0.0127 |
| QUANT\_BANKING\_ACCOUNTS | 0.0110 |

# Appendix B Tuning Parameter for Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
| Hyperparameter | Description | Range of Trail | Optimal Value |
| n\_estimators | The number of trees in the forest. | 10 - 200 | 155 |
| min\_samples\_leaf | The minimum number of samples required to be at a leaf node. | 5, 10, 20, 50, 100 | 50 |
| max\_features | The number of features to consider when looking for the best split: | auto, sqrt, log2 | log2 |
| max\_depth | The maximum depth of the tree | None, 5, 10, 20, 50, 100 | 50 |
| class\_weight | balanced adjust weights inversely proportional to class frequencies in the input    balanced\_subsample adjust weights based on the bootstrap sample | balanced\_subsample, balanced | balanced\_subsample |
| min\_samples\_split | The minimum number of samples required to split an internal node: | 2, 5, 10, 20 | 2 |
| bootstrap | Whether bootstrap samples are used when building trees. If False, the whole dataset is used to build each tree. | True, False | False |

# Appendix C Evaluation Metrics

## Appendix C.1 Confusion Metrix

### Appendix C.1.1 Logistic Regression

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Prediction | |
|  |  | Non-Default | Default |
| ***True*** | Non-Default | 6074 (TN) | 4994 (FP) |
| Default | 1555 (FN) | 2342 (TP) |
|  |  |  |  |

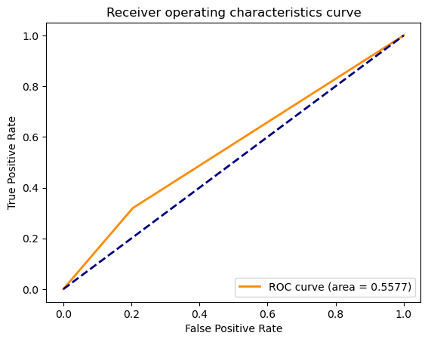
### Appendix C.1.2 Random Forest

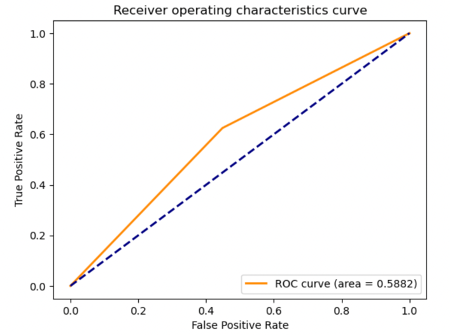
|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Prediction |  |
|  |  | Non-Default | Default |
| True | Non-Default | 6104 (TN) | 4964 (FP) |
|  | Default | 1462 (FN) | 2435 (TP) |

## Appendix C.2 Classification Report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Recall | Precision | Specificity |
| Logistic Regression | 0.5624 | 0.5488 | 0.7962 | 0.6010 |
| Random Forest | 0.5706 | 0.5515 | 0.8068 | 0.6248 |

## Appendix C.3 AUC-ROC Curve





Appendix C.4 Feature Importance Visualisation

