Credit Scoring modelling

Assignment 4

Group 24

Authors:
Selma Boussfia (2645554)
Ching-Ting Ni (2654740)

Course Coordinator:
Dr. S.A. Borovkova

Master EOR
June 15, 2022

Contents

Page

1 Introduction **2**

2 Data cleaning **2**

3 Data pre-processing **3**
3.1 Missing values . . . 3
3.2 Numerical variables . . . 4
3.2.1 Variable: annual_inc . . . 5
3.2.2 Variable: earliest_cr_line . . . 5
3.2.3 Variable: int_rate . . . 6
3.2.4 Variable: loan_amount . . . 7
3.2.5 Variable: num_actv_bc_tl . . . 7
3.2.6 Variable: mort_acc . . . 8
3.2.7 Variable: tot_cur_bal . . . 9
3.2.8 Variable: pub_rec_bankruptcies . . . 9
3.2.9 Variable: revol_util . . . 10
3.2.10 Variable: total_acc . . . 11
3.3 Categorical variables . . . 11

4 Logistic Regression **13**
4.1 Under-sampling for the target variable 'loan status' . . . 13
4.2 Over-sampling for the target variable 'loan status' . . . 14
4.3 Fitting the model and results . . . 14
4.4 ROC curve and Confusion matrix . . . 15
4.5 Pearson Residuals . . . 17

5 Random Forest **18**
5.1 Training, validation and test set (60-20-20) . . . 18
5.2 Shapley Additive explanations . . . 19
5.3 ROC curve and Confusion matrix . . . 21

6 Conclusion **23**

1 Introduction

Logistic regression is a widely used method for categorical response data. This method is used to determine the probability of an event occurring based on a set of explanatory variables. The binary logistic model can be used to determine the probability that a client will be able to meet its financial obligations. A specific model with explanatory variables, such as total loan and outstanding amount of active loans, can be analyzed to determine this credit worthiness. In this assignment, the logistic model and one machine learning method will be used to determine a credit scoring model for default on loans. First, the data cleaning and pre-processing steps will be given. Second, the data will be fitted by logistic regression. Third, the Random Forest method will be performed and analyzed. The performance of these models will be analyzed by different performance measures. Finally, a conclusion will be made on which method performs best.

2 Data cleaning

In this section, the data which will be used for our analysis will be described. Furthermore, the data cleaning process will be discussed. The dataset contains the full LendingClub data which we have obtained from the website Kaggle. Lendingclub is known as a peer-to-peer lending company based in San Francisco and is one of the largest lending company in the world. It was the first peer-to-peer lending company to register its offerings as securities with the Securities and Exchange Commission, and to offer loan trading on a secondary market. The features in the dataset gives the following information about the loans: bureau data history such as loan status of the client, demographic data such as age and disbursal details of the client.
The dataset contains 2260701 records and 151 columns. Since the number of records is too large, we only select the records in which the borrower's earliest reported credit line was opened from year 1995 onward. This results in a dataset with 1760672 records. The record we consider as *defaulted* is when a customer was not able to pay the loan amount back at some point of time, and the loans that were paid back later than 30 days. The records we consider as *non-defaulted* are the records in which the loan was paid back on time and the loans which are currently being paid. After classifying the loans as defaulted or non-defaulted, the dataset contains 1748327 non-defaulted records, which is 86.85% of the dataset, and 212173 defaulted records, which is 12.14% of the dataset.

To further clean and prepare the data, the following operations are performed:

- Dropping the columns with more than 10% missing values: The features in which more than 10% of the observations are missing are deleted. This results in a dataset with 93 features. Furthermore, the features for which we do not have a clear understanding are deleted since it is not wise to include these features in our analysis without fully understanding them, which leads to final 18 features.

- Dropping records with more than 3 missing features: Also, we delete the records in which 3 or more features are missing. This results in a dataset with 1715165 records and 18 features. Table 1 shows the 18 features with their description. Variables 1 till 13 are the numerical variables and variables 14 till 18 are the categorical variables.

- Sub-sampling 30% of data: Since the dataset still contains too many records, we only take 30% of the non-defaulted records and 30% of the defaulted records. This results in a dataset with 509228 records and 18 features.

	Feature	Description
1	annual_income	The annual income provided by the borrower during registration.
2	earliest_cr _line	The month the borrower's earliest reported credit line was opened.
3	int_rate	Interest Rate on the loan.
4	loan_amnt	The listed amount of the loan applied for.
5	num_actv _bc_tl	Number of currently active bankcard accounts.
6	mort _acc	Number of mortgage accounts.
7	tot_cur_bal	Total current balance of all accounts.
8	open_acc	The number of open credit lines in the borrower's credit file.
9	pub_rec	Number of derogatory public records.
10	pub_rec_bankrup	Number of public record bankruptcies.
11	revol_util	Revolving line utilization rate.
12	total_acc	Number of credit lines currently in the borrower's credit file.
13	loan_status	Current status of the loan.
14	appl_type	Whether the loan is an individual or joint application.
15	home_owner	Home ownership status of the borrower, which are: *rent, own, mortgage and other.*
16	grade	LC assigned loan grade
17	term	The number of payments on the loan in months.
18	purpose	Category provided by the borrower for the loan request.

Table 1: Description of the features. The term LC in the table stands for LendingClub.

3 Data pre-processing

3.1 Missing values

There are in total 3 variables that have missing values, those are listed below and we provide suggestions to impute the missing values:

- revol_util: 0.068 % of the data in this variable is missing. It is noticed that there is a difference between mean of revol_util for defaulted and non-defaulted class for the non-missing observations, which are respectively 54.3, 49.3. Therefore, we replace these mean values for the missing values based on the loan status being defaulted or non-defaulted.

- num_actv_bc_tl: 0.89 % of the data in this variable is missing. There is not much difference in mean values between the defaulted and non-defaulted class, hence we replace the missing entries with the mean value of num_actv_bc_tl regardless of the loan status at 4 after rounding.

- tot_cur_bal: 0.88 % of the data in this variable is missing. It is observed that there is a difference between mean of tot_bal for defaulted and non-defaulted class for the

non-missing observations, which are respectively 112617, 135009. Therefore, we replace these mean values for the missing values based on the loan status.

3.2 Numerical variables

For the numerical variables, we will first decide whether there is redundancy in our dataset, so if the explanatory variables are correlated with each other. If two variables are highly correlated with each other, we will remove one of the variables. We do this to prevent multicollinearity when we start fitting our models. Figure 1 shows the correlation matrix. The figure shows that the correlation between the features is quite low expect for 4 variables. The variables that are highly correlated with each other are:

- open account and total account
- public records and public records bankruptcies

which is quite reasonable. We will drop the variables open account and public records. After dropping the 2 variables, the dataset contains 509228 records and 16 features.

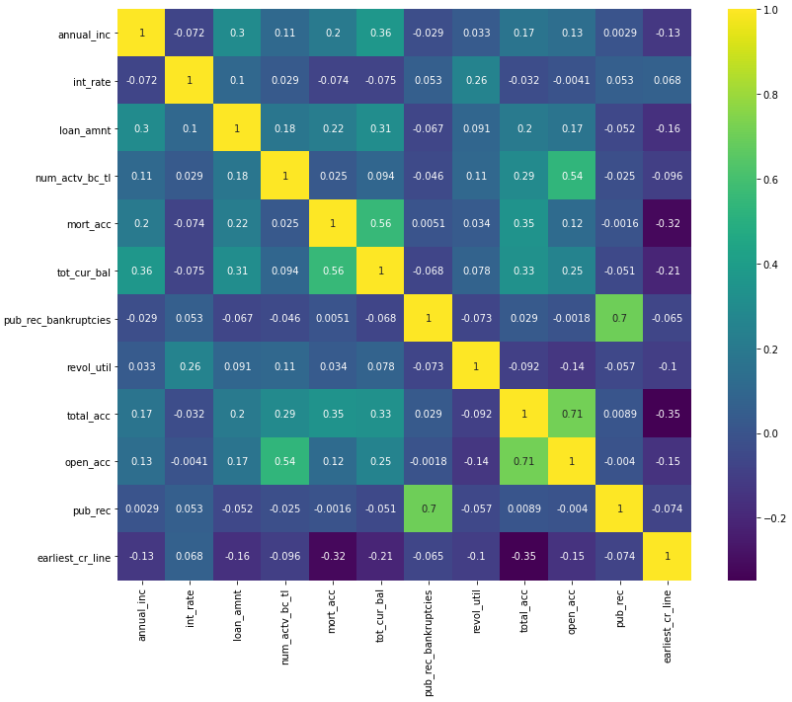

Figure 1: Correlation matrix.

Next, the existence of outliers will be analyzed for each variable in the paragraphs below. Then, the distribution plots after dropping outliers will be given. Afterwards, data transformations will be performed including log transformation and standardization, and finally the descriptive statistics are given after transformation.

3.2.1 Variable: annual_inc

This variable stands for the annual income provided by the borrower during registration. We consider values above 300000 as outliers and drop those, which accounts to 0.466% of the loans, or 509228 observations. Figure 2 gives the distribution plot after dropping the outliers, and it is noticed that the distribution is still skewed to the right, therefore we take the logarithmic transformation before standardizing the data for the latter model building phase. The descriptive statistics of the standardized variable is given in Table 2. A clear difference in standardized annual income for defaulted and non-defaulted is observed, where people who default have a lower mean value of -0.0916.

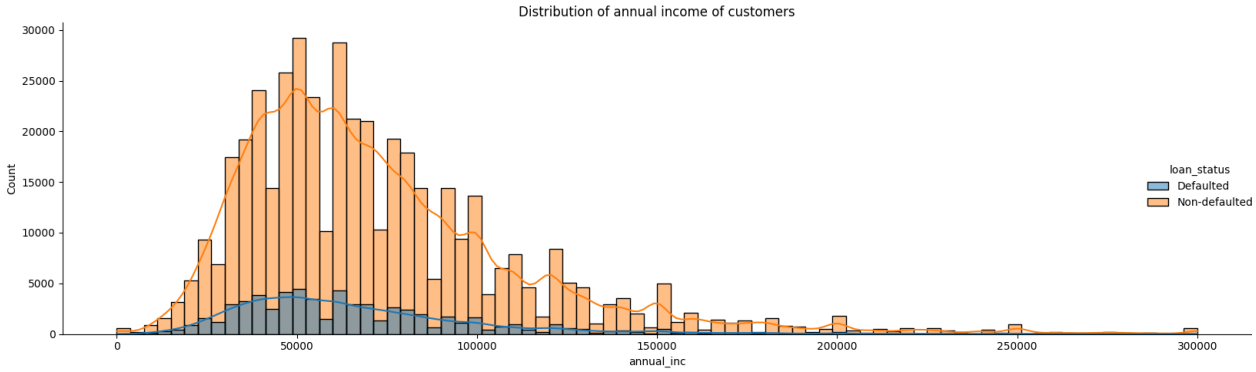

Figure 2: Distribution plot of annual_inc.

		count	mean	std	min	25%	50%	75%	max
Loan status	Defaulted	60137	-0.0916	0.87	-18.06	-0.62	-0.041	0.429	2.594
	Non-defaulted	428444	0.0128	1.016	-18.06	-0.51	0.038	0.591	2.594

Table 2: Descriptive statistics of the variable annual_inc.

3.2.2 Variable: earliest_cr_line

This variable stands for the month the borrower's earliest reported credit line was opened. We exclude the month and only take the year as numeric values. There are no significant outliers in our inspection. Figure 3 shows the distribution, and we notice that most of the credit lines were opened between 1995 to 2008. Table 3 shows the descriptive statistics after standardizing. and there is no clear difference between defaulted and non-defaulted case.

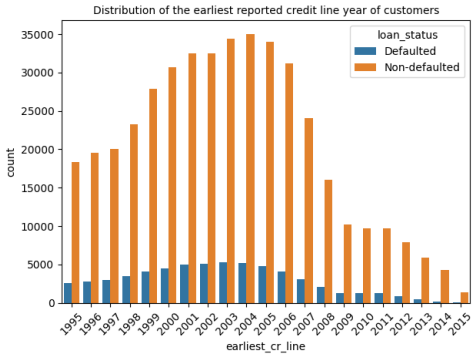

Figure 3: Distribution plot of earliest_cr_line.

		count	mean	std	min	25%	50%	75%	max
Loan status	Defaulted	60137	-0.069	0.942	-1.728	-0.847	-0.186	0.474	2.677
	Non-defaulted	428444	0.0097	1.007	-1.728	-0.847	0.034	0.694	2.677

Table 3: Descriptive statistics of the variable earliest_cr_line.

3.2.3 Variable: int_rate

This variable stands for the interest rate on the loan. There are no significant outliers in our inspection. Figure 4 shows the distribution, and we notice that most of the loans have interest below 20. Table 4 gives the descriptive statistics after standardizing the variable. A clear difference in the mean is found between defaulted and non-defaulted loans, where defaulted loans on average have higher interest rates.

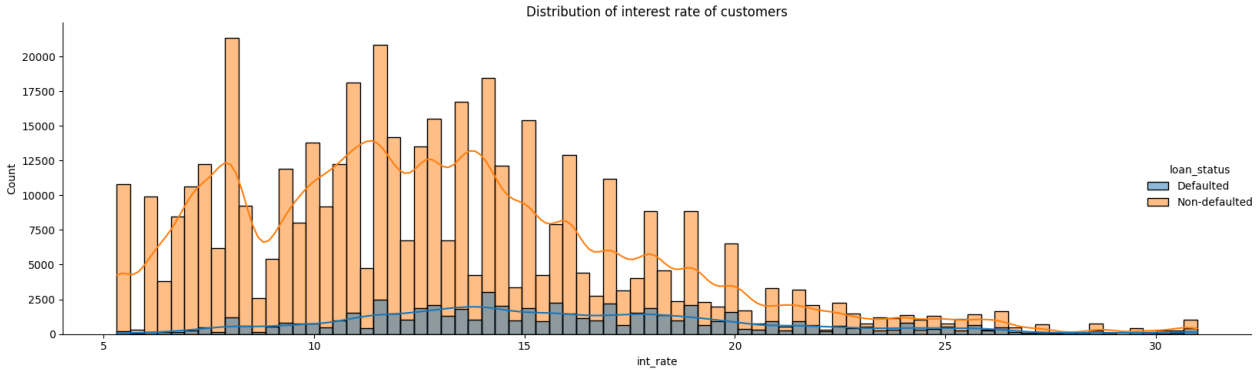

Figure 4: Distribution plot of int_rate.

		count	mean	std	min	25%	50%	75%	max
Loan status	Defaulted	60137	0.55	1.014	-1.644	-0.142	0.410	1.167	3.633
	Non-defaulted	428444	-0.077	0.973	-1.644	-0.795	-0.148	0.472	3.633

Table 4: Descriptive statistics of the variable int_rate.

3.2.4 Variable: loan_amount

This variable stands for the amount of the loan applied for. We consider loans amount higher than 36000 as outliers and will drop those, which are 1.527% of the observations. Figure 5 shows the distribution plot, and we do not observe a clear pattern for the distribution between the loan status. Table 5 shows the standardized variable, where we see a positive mean value for defaulted loan holders, who tend to have a larger loan amount.

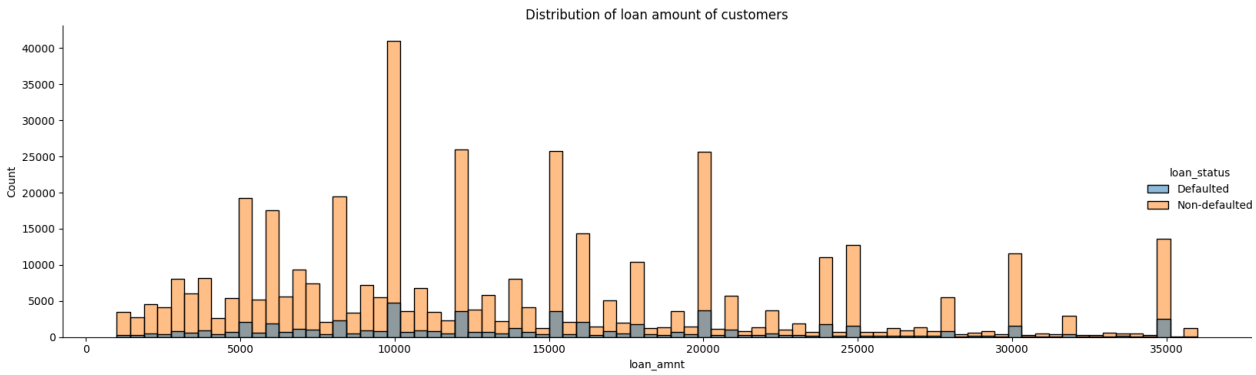

Figure 5: Distribution plot - loan_amount.

		count	mean	std	min	25%	50%	75%	max
Loan status	Defaulted	60137	0.113	1.00	-1.558	-0.641	-0.010	0.704	2.609
	Non-defaulted	428444	-0.015	0.998	-1.558	-0.784	-0.248	0.704	2.609

Table 5: Descriptive statistics of the variable loan_amount.

3.2.5 Variable: num_actv_bc_tl

This variable stands for the number of currently active bankcard accounts. We consider values above 10 as outliers and drop those which accounts for 1.133% of observations. Figure 6 shows the distribution plot, and we observe that the majority of the customers have 1 to 5 active bankcard accounts. Table 6 shows the descriptive statistics of the standardized variable. It is noticed that the defaulted loans holders have on average higher currently active bankcard accounts.

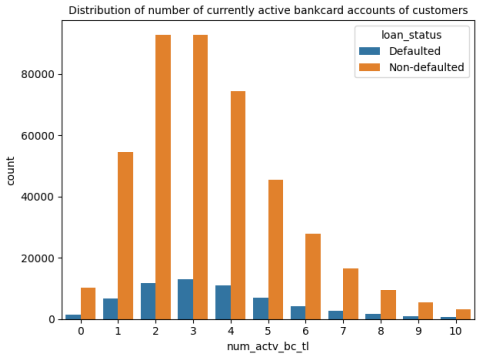

Figure 6: Distribution plot of num_actv_bc_tl.

		count	mean	std	min	25%	50%	75%	max
Loan status	Defaulted	60137	0.077	1.022	-1.741	-0.729	-0.223	0.788	3.318
	Non-defaulted	428444	-0.010	0.996	-1.741	-0.729	-0.223	0.282	3.318

Table 6: Descriptive statistics of the variable num_actv_bc_tl.

3.2.6 Variable: mort_acc

This variable stands for the number of mortgage accounts. We consider the values higher than 10 as outliers and drop those. That is we drop 0.110% of the records. Figure 7 shows the distribution plot. The figure shows that most of the customers have 1 mortgage account. Table 7 shows the descriptive statistics of the standardized variable. From the table, it can be seen that the defaulted loans holders have less mortgage accounts on average.

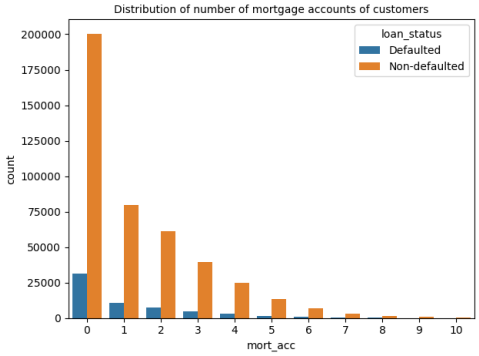

Figure 7: Distribution plot of mort_acc.

		count	mean	std	min	25%	50%	75%	max
Loan status	Defaulted	60128	-0.080	0.9659	-0.768	-0.768	-0.768	0.406	5.109
	Non-defaulted	428203	0.011	1.004	-0.768	-0.768	-0.181	0.406	5.109

Table 7: Descriptive statistics of the variable mort_acc.

3.2.7 Variable: tot_cur_bal

This variable stand for the total current balance of all accounts. We consider the values higher than 1000000 dollars as outliers and drop those. That is we drop 0.061% of the records. Figure 8 shows the distribution plot. The figure shows that most of the customers have less than 50,000 dollars as the total current balance of all accounts. Table 8 shows the descriptive statistics of the standardized variable. From the table, it can be seen that the defaulted loans holders have a smaller total current balance on average, which is quite reasonable.

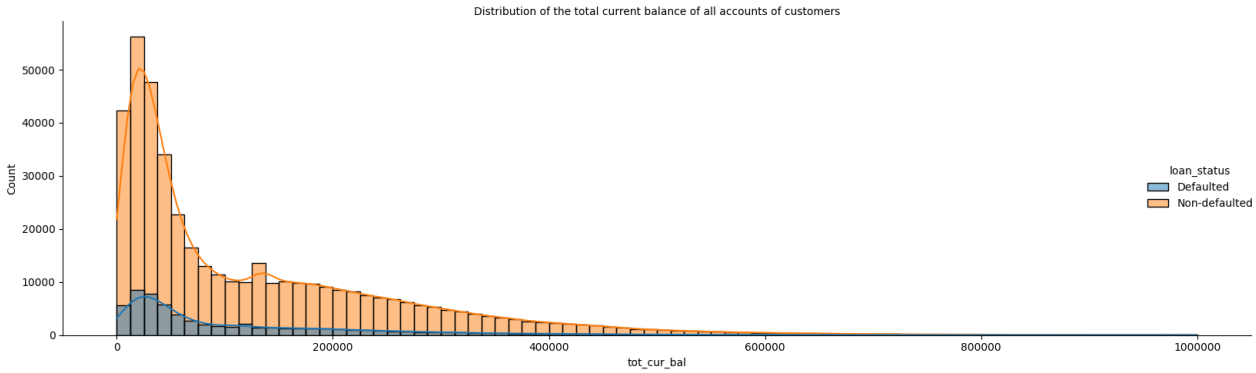

Figure 8: Distribution plot of tot_cur_bal.

		count	mean	std	min	25%	50%	75%	max
Loan status	Defaulted	60128	-0.077	0.940	-8.499	-0.696	-0.098	0.680	2.076
	Non-defaulted	428203	0.010	1.007	-8.499	-0.678	0.061	0.838	2.088

Table 8: Descriptive statistics of the variable tot_cur_bal.

3.2.8 Variable: pub_rec_bankruptcies

This variable stand for the number of public record bankruptcies. We consider the values higher than 3 as outliers and drop those. That is we drop 0.0361% of the records. Figure 9 shows the distribution plot. The figure shows that most of the customers have no public record bankruptcy. Table 9 shows the descriptive statistics of the standardized variable. The table shows that the defaulted loans holders have a higher number of public record bankruptcies on average.

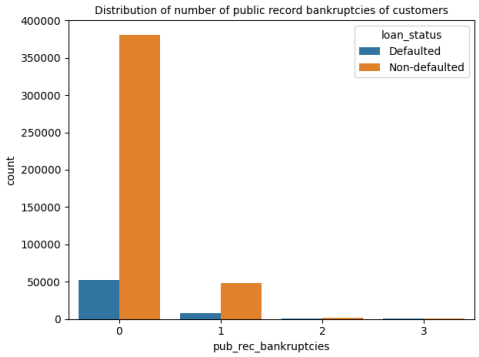

Figure 9: Distribution plot of pub_rec_bankruptcies.

		count	mean	std	min	25%	50%	75%	max
Loan status	Defaulted	60128	0.066	1.097	-0.357	-0.357	-0.357	-0.357	8.095
	Non-defaulted	428203	-0.009	0.985	-0.357	-0.357	-0.357	-0.357	8.095

Table 9: Descriptive statistics of the variable pub_rec_bankruptcies.

3.2.9 Variable: revol_util

This variable stands for the revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. We consider the values higher than 120 as outliers and drop those. That is we drop 0.008% of the records. Figure 10 shows the distribution plot. The figure shows that most of the customers have a revolving line utilization rate of 50. Table 10 shows the descriptive statistics of the standardized variable. The table shows that the defaulted loans holders have a higher revolving line utilization rate on average.

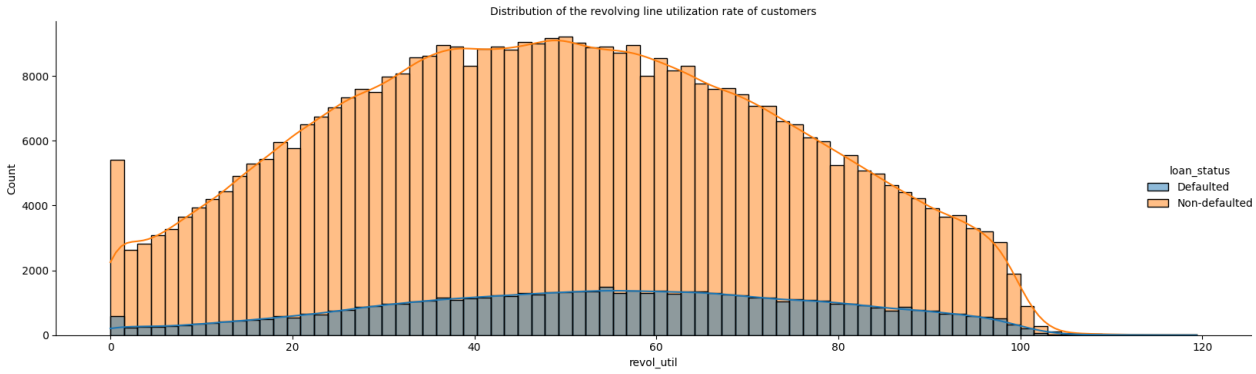

Figure 10: Distribution plot of revol_util.

		count	mean	std	min	25%	50%	75%	max
Loan status	Defaulted	60128	0.180	0.969	-2.043	-0.525	0.204	0.918	2.563
	Non-defaulted	428203	-0.025	1.001	-2.043	-0.790	-0.040	0.735	2.828

Table 10: Descriptive statistics of the variable revol_util.

3.2.10 Variable: total_acc

This variable stands for the number of credit lines currently in the borrower's credit file. We consider the values higher than 60 as outliers and drop those. That is we drop 0.773% of the records. Figure 11 shows the distribution plot. The figure shows that most of the customers have approximately 20 credit lines. Table 11 shows the descriptive statistics of the standardized variable. The table shows that the defaulted loans holders have a higher number of credit lines on average.

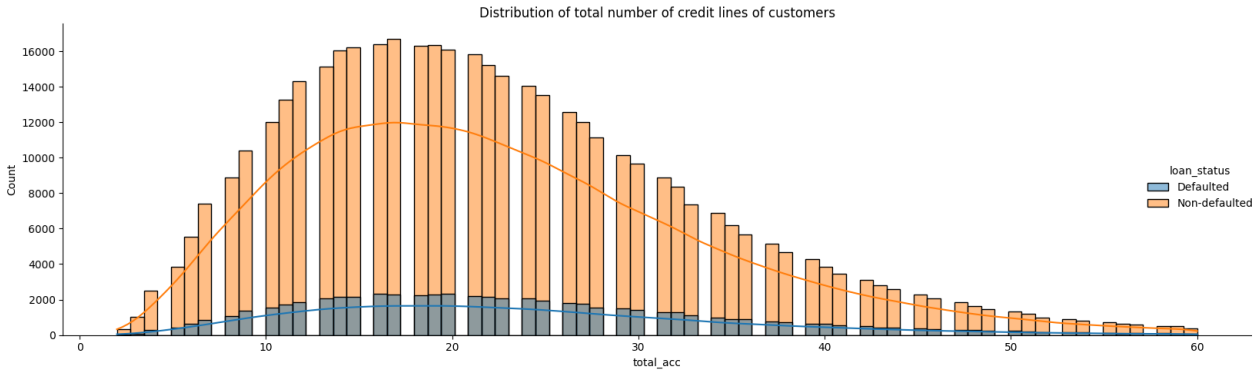

Figure 11: Distribution plot of total_acc.

		count	mean	std	min	25%	50%	75%	max
Loan status	Defaulted	60128	0.066	1.015	-1.891	-0.693	-0.049	0.687	3.450
	Non-defaulted	428203	-0.009	0.997	-1.891	-0.785	-0.141	0.595	3.4509

Table 11: Descriptive statistics of the variable total_acc.

3.3 Categorical variables

For categorical variables, we will first determine if there are variables that are not significant. We will do this by using the Information Value criteria which can be used to understand the predictive power of a variable. The information value for a categorical variable can be calculated as follows:

$$IV(x) = \sum_{i=1}^{N(x)} \left(\frac{g_i}{g} - \frac{b_i}{b} \right) \cdot \log \left(\frac{\frac{g_i}{g}}{\frac{b_i}{b}} \right)$$

where $N(x)$ is the number of levels in the variable x, g_i represents the number of non-defaults in category i of variable x_i, b_i represents the number of defaults in category i of variable x_i, g represents the number of non-defaults in the entire dataset, b represents the number

of defaults in the entire dataset. We consider a variable with information value less than 0.15 as a variable with poor predictive power, so we exclude this variable. Table 12 shows the variables with their information value. The table shows that we can exclude all variables expect for the variables 'grade' and 'term' since these variables have information values higher than 0.15.

Variable	Information Value
grade	0.5156
home_ownership	0.0197
application_type	0.0284
purpose	0.0245
term	0.2014

Table 12: Information Values of the categorical variables.

Next, the frequency of values in each potential explanatory variable will be analyzed. We do this because we want to have balanced categories in each one of the variables. This is important because otherwise we cannot estimate the effect of the category on the likelihood of default. To examine the frequencies, the bar charts of the variables 'grade' and 'term' are given. Figure 12 shows the frequencies of each category of the variable 'grade'. The figure shows that the categories are not balanced. The frequency of grade D till G are very low compared to other grades. Figure 13 shows the frequencies of the categories of the variable 'grade' per default category. From the figure, it can be seen that frequencies per default category are quite balanced except for the grades D, E, F, and G. Therefore, we aggregate these grades groups to only one class called grade. Figure 14 shows the frequencies of each category of the variable 'term'. Also for this variable, the categories are not really balanced. Figure 15 shows that the frequencies of the categories of the variable 'term' per default category are quite balanced.

Lastly, we create dummy variables for these two categorical variables and drop the original variables. To avoid multicollinearity issue later in the model estimation section in this assignment, the first class is dropped and treated as the reference group. This leads to dummy variables for loan grade [grade_B, grade_C, grade_D] and dummy variable for term [term_60_months].

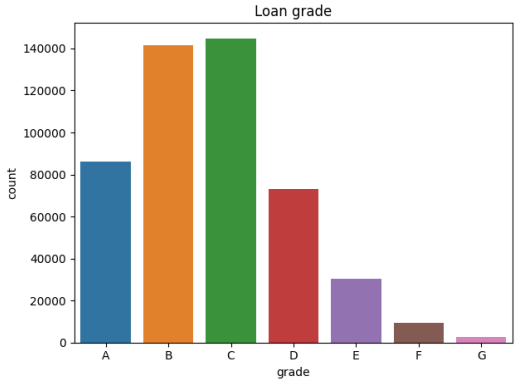

Figure 12: The frequencies of each category of the variable 'grade'.

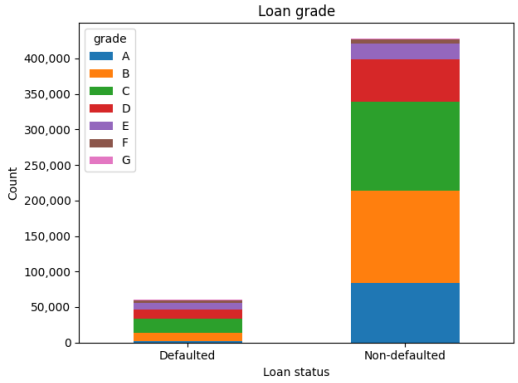

Figure 13: The frequencies of the categories of the variable 'grade' per default category.

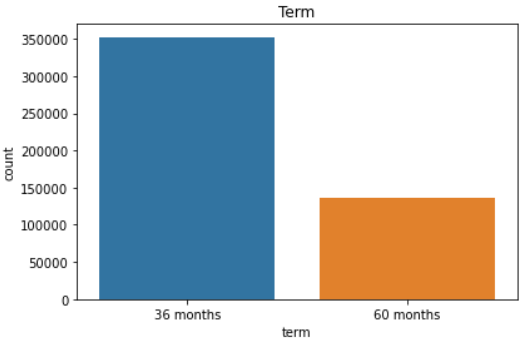

Figure 14: The frequencies of each category of the variable 'term'.

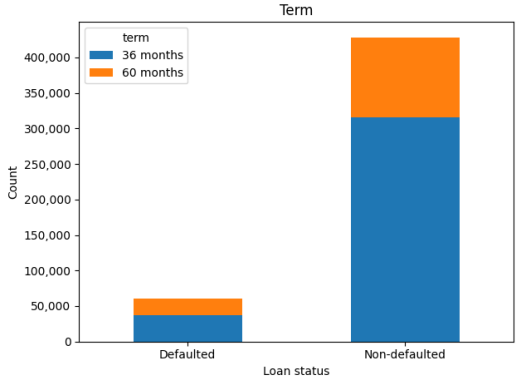

Figure 15: The frequencies of the categories of the variable 'term' per default category.

4 Logistic Regression

In this section, the data will be analyzed by performing logistic regression. The results will be interpreted by means of Odds ratios. The performance of the logistic model will be measured by means of the optimal cut-off value for probability of default (PoD), the ROC curve, the confusion matrix and by analyzing the Pearson Residuals. The dependent variable is Bernoulli distributed and is formulated as follows:

$$Y_i = \begin{cases} 1 & \text{if loan of the customer } i \text{ is defaulted} \\ 0 & \text{otherwise} \end{cases}$$

4.1 Under-sampling for the target variable 'loan status'

The portion of the non-defaulted loans is 87.7% and the defaulted loans is 12.3%. To proceed with the logistics regression model, under-sampling for the non-defaulted loans is performed

where we randomly sample 60% of the non-defaulted loans. This results in the final dataset of 317203 loan records, and 15 features considering the dummy variables for categorical variables. This more balanced dataset consists of 19% defaulted loans and 81% non-defaulted loans.

4.2 Over-sampling for the target variable 'loan status'

Furthermore, we over-sample the defaulted loans to 257066, which is the same size as the non-defaulted loans. This practice is called borderline SMOTE, where we simulate more defaulted observations that share similar characteristics as the existing non-defaulted loans, which is in fact harder to determine or classify. Now the dataset set contains 50% of defaulted observations and 50% of non-defaulted observations.

4.3 Fitting the model and results

We split our data as training and test set, for which 70% is for training the logistic regression and 30% is for testing.

Table 13 shows the estimations results of the variables as well as the standard errors, z value, p-value and OR. We use a significance level of 0.05. The table shows that all estimates are significantly different from 0. Figure 16 shows the bar charts of the coefficient estimates. This figure shows the relationship between the variables and the probability of default and gives a nice visualisation whether the effect on the likelihood of defaulting is positive or negative. The figure shows that a customer with Grade D has a higher effect on the likelihood of defaulting than those with Grades a and b, which is reasonable since a customer with grade d has a higher risk for defaulting.

Furthermore, the values of significant coefficients will be interpreted in terms of Odds Ratios (OR) of default. From table 13, it can be seen that the variable annual_inc has an OR of 0.879, which indicates that we expect a decrease of 12.1% in the odds of defaulting, for a 1 unit increase in annual_inc. Thus, having a higher income decrease the odds of defaulting. The OR of variable int_rate is equal to 1.138, which indicates that we expect an increase of 13.8% in the odds of defaulting, for a unit increase in int_rate. So, if the interest rate on the loan increase, then the odds of defaulting also increases. Figure 17 also shows the OR of the variables. The figure shows that the variable grade_D has the highest OR.

	Estimate	Std. error	z value	Pr($>|z|$)	OR
intercept	-1.607	0.018	-90.148	0.000	0.200
annual_inc	-0.128	0.005	-25.198	0.000	0.879
int_rate	0.129	0.008	15.540	0.000	1.138
loan_amnt	0.083	0.005	18.358	0.000	1.087
num_actv_bc_tl	0.055	0.004	14.386	0.000	1.056
mort_acc	-0.105	0.005	-23.019	0.000	0.899
tot_cur _bal	-0.102	0.005	-20.214	0.000	0.902
pub_rec_bankruptcies	0.034	0.003	9.795	0.000	1.034
revol_util	0.087	0.004	21.663	0.000	1.091
total_acc	0.140	0.005	31.147	0.000	1.150
earliest_cr_line	-0.128	0.004	-30.527	0.000	0.879
grade_B	1.076	0.017	62.310	0.000	2.934
grade_C	1.724	0.020	87.924	0.000	5.610
grade_D	2.1768	0.027	81.332	0.000	8.817
term_60_months	-0.066	0.009	-7.318	0.000	0.935

Table 13: Estimation results.

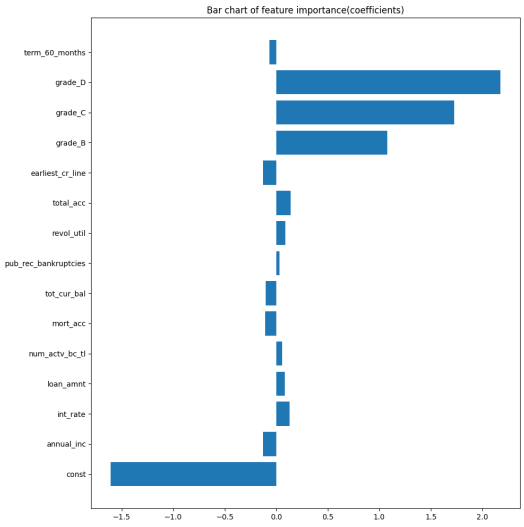

Figure 16: Estimates of the variables.

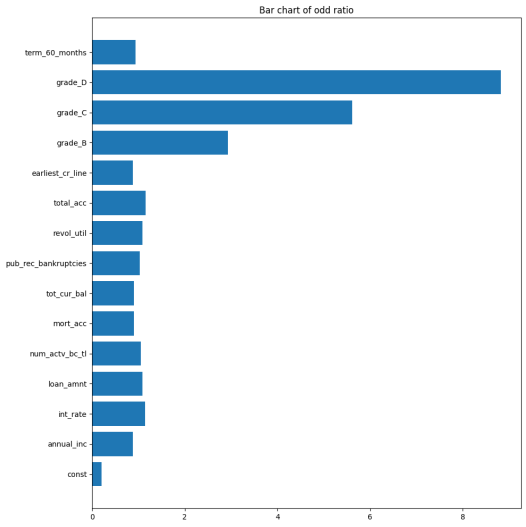

Figure 17: Odd Ratios of the variables.

4.4 ROC curve and Confusion matrix

Figure 18 shows the ROC curve for the logistic model. We measure the area under the curve to compare it to the area of a perfect model. The area under the curve of a perfect model is equal to 1, so the closer the area is to 1, the higher the classification power. The area under the curve is equal to 0.71, which is an acceptable value, this indicates that the model does not have a high classification power but also not really low.

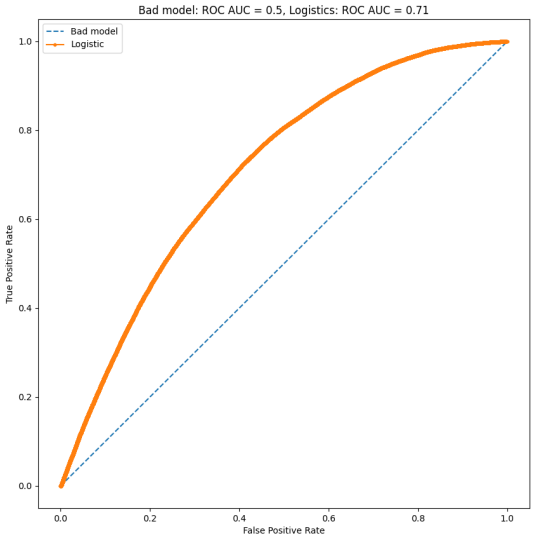

Figure 18: ROC curve.

The optimal cut off point would be where "true positive rate" is high and the "false positive rate" is low. Optimizing this value based on the predictions in the training dataset gives a value of 0.534. Thus, the model classifies any individual with a probability of defaulting of 0.534 or higher as default. While any individual with a probability less than 0.534 is classified as non-default.

Next, the confusion matrix will be used on the test set to evaluate the performance of the classification model. The purpose of a confusion matrix is to sum the number of correct and incorrect predictions per class (default or non-default). This is calculated based on the actual observations (binary) and predictions classified as default or non-default according to the optimal cut-off probability from predicted probability the testing data. Table 14 shows the confusion matrix. Using this confusion matrix, we calculated the recall, precision and accuracy to determine the performance of the model with the F1 score which are shown in Table 15. The F1 score is the singular metric summarizing the confusion matrix and so model performance. The F1 score is equal to 64.3%, which is an acceptable score, so by means of the F1 score we can say that the model performs well. The confusion matrix was also done for the training set. The quantities based on the confusion matrix for the training set were very close to that for the test set.

	Predicted non-default	Predicted default
Observed non-default	50222	26748
Observed default	26820	50450

Table 14: Confusion matrix for the test set.

Recall	0.652
Precision	0.653
Accuracy	0.652
F1 score	0.653

Table 15: Quantities based on the confusion matrix for the test set.

4.5 Pearson Residuals

For diagnosing the logistic regression model, we will also plot the Pearson residuals against the predicted values. For a well fitted model, there should be no correlation between the residuals and predictors. At best, the trend should be a horizontal straight line without curvature. Figures 19 and 20 show the plots of the Pearson residuals against the predicted values for each variable. From Figure 19, it can be seen that for most of the variables the trend is a horizontal straight line without curvature expect for the variables annual_income, mort_acc and pub_rec_bankruptcies. So for all variables expect for these three, the model fits the data well.

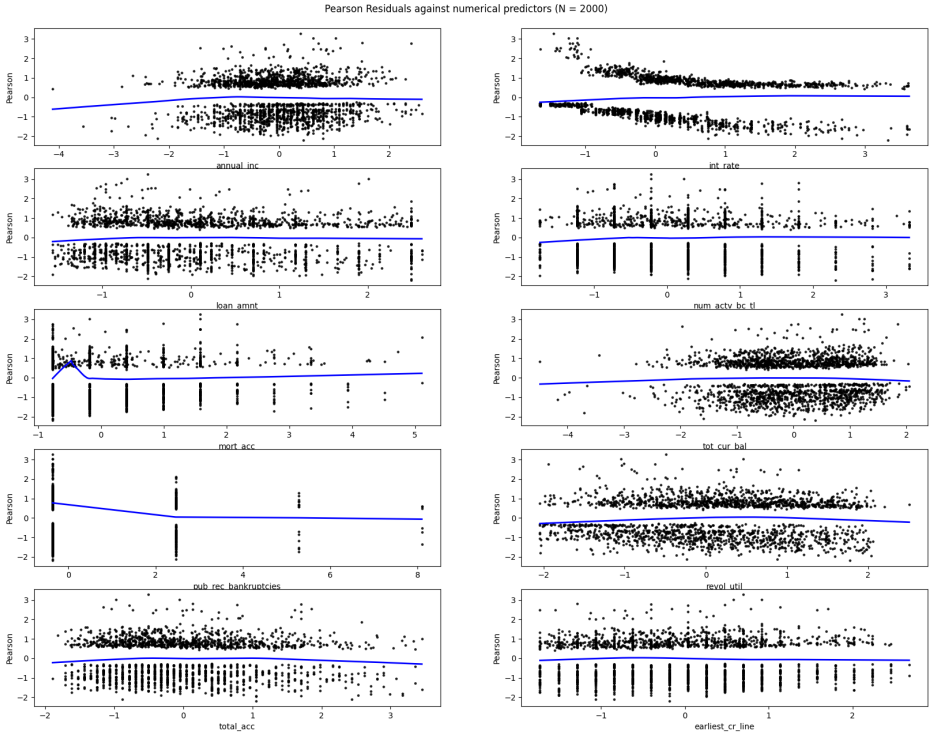

Figure 19: Plots of Pearson residuals against predicted values for each variable.

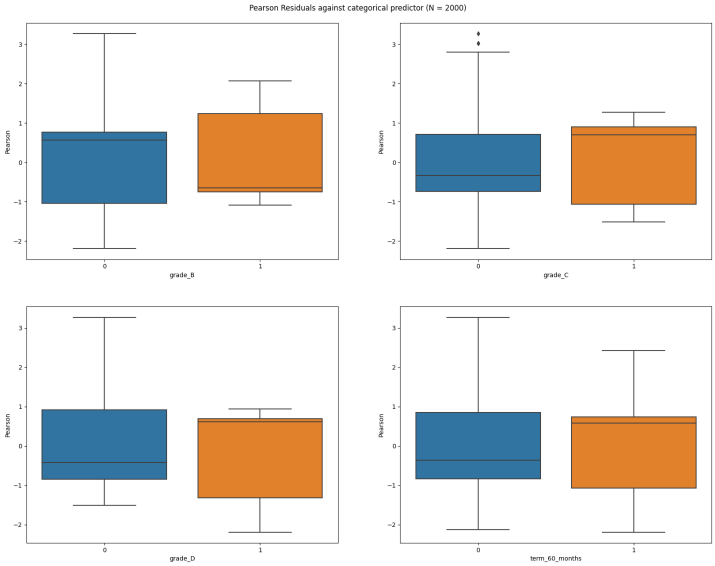

Figure 20: Boxplots of Pearson residuals against predicted values for each variable.

5 Random Forest

In this section, the Random Forest (RF) method will be used to classify our dataset. Furthermore, the results will be interpreted by means of Shapley Additive explanations (SHAP) values. The performance of the RF method will be measured by means the ROC curve, the confusion matrix and quantities obtained from the confusion matrix.

The Random Forest method is a classification algorithm that uses numerous decision trees to classify data. In order to generate an uncorrelated forest of trees, this method uses bootstrapping and feature randomness when building each individual tree. The resulting forest of tree has a higher prediction power than any individual tree.

5.1 Training, validation and test set (60-20-20)

We split the data to training set, validation set, and test set. The first operation we perform is tune in the hyperparameters using the validation set before fitting the mode. The results are shown in Table 16 using the grid search method by cross-validated search over pre-defined parameter spaces. Short description of each hyperparameter is illustrated below:

- n_estimators: number of trees in Random Forest.
- min_samples_split: minimum number of samples required to split a node.
- min_samples_leaf: minimum number of samples that should be present in the leaf node after splitting a node.
- max_samples: fraction of the original dataset is given to any individual tree.

- max_leaf_nodes: maximum number of terminal nodes after splitting, if reached, the tree stops growing.
- max_features: maximum number of features to consider at every split.
- max_depth: maximum number of levels in tree.

	n_estimators	min_samples_split	min_samples_leaf	max_samples	max_leaf_nodes	max_features	max_depth
Values	200	14	12	0.3	20	sqrt	30

Table 16: Hyperparameters values using grid search.

5.2 Shapley Additive explanations

Next, the results are interpreted by means of Shapley values, which give us the marginal contributions of each explanatory variable across all observations. Two important measurements Shapley values provide are:

- Global interpretability: how much each predictor contributes, either positively or negatively, to the target variable.
- Local interpretability: each observation gets its own set of SHAP values, which helps us explain why a case receives its prediction and the contributions of the predictors.

Figure 21 shows the mean positive or negative effect of each predictor on the target variable. Due to the large test dataset, we only sample $10,000$ observations from the test set. The impacts are ordered where explanatory variables int_rate, grade_D, grade_B are the top three important predictors. In the same plot, a predictor associated with a red bar indicates a positive impact on the target variable, while a predictor associated with a blue bar indicates a negative impact.
Figure 22 shows the SHAP values of each observation, where every dot is one observation. Again the variables are ordered in order of importance. The horizontal location shows whether the effect of that value is associated with a higher or lower probability of default. The color shows whether that variable is high (in red) or low (in blue) for that observation. The figure shows that a high interest rate on the loan is associated with a high probability of default and a low interest rate on the loan is associated with a low probability of default. This is also the case for the variable grade_D. For the variable grade _B it is the opposite, a high grade _B is associated with a low probability of default and a low grade _B is associated with a low probability of default.

Next, we are going to analyze the outcome of 1 particular client, or one observation. Table 17 shows the predictor values and mean of the first and tenth observation. Figure 23 shows the result of the first observation which defaulted. The base value is equal to 1. Annual _income has a negative impact on the probability of default. The annual _income for this observation is equal to -0.11 which is higher than the average value -0.042, so the prediction is pushed to the right. Figure 24 shows the result of the tenth observation which was non-defaulted. Int _rate has a positive impact on the probability of default. Int _rate for this observation is equal to -0.77 which is lower than the average value 0.260, so the prediction is pushed to the left.

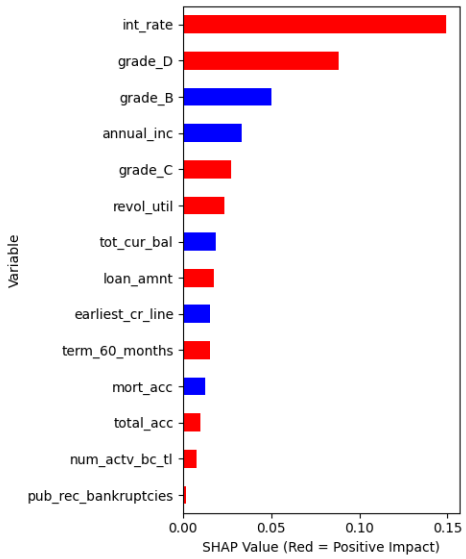

Figure 21: Mean of marginal impact for each predictor on output.

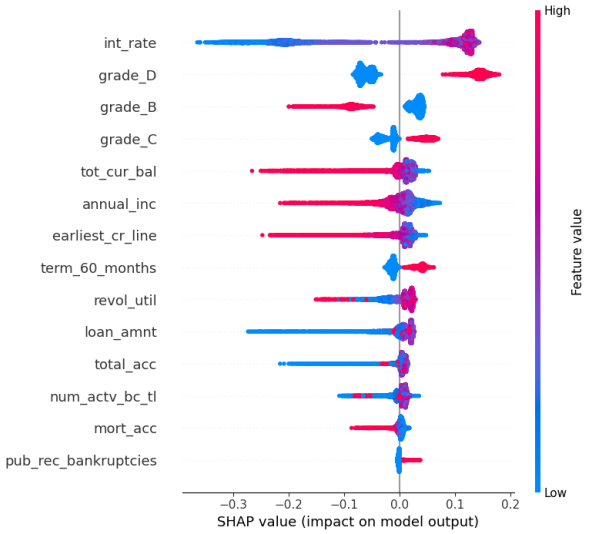

Figure 22: Marginal impact for each observation on output.

	Predictor mean	First observation	Tenth observation
annual_inc	-0.042	0.819	0.951
int_rate	0.260	0.368	-0.766
loan_amnt	0.044	0.046	-1.082
num_actv_bc_tl	0.036	-0.223	1.294
mort_acc	-0.045	0.406	-0.180
tot_cur_bal	-0.022	1.064	0.975
pub_rec_bankruptcies	0.040	-0.357	-0.357
revol_util	0.091	1.781	-0.227
total_acc	0.016	-0.811	1.332
earliest_cr_line	-0.018	-0.186	0.694
grade_B	0.237	0.00	1.00
grade_C	0.315	1.00	0.00
grade_D	0.336	0.00	0.00
term_60_months	0.321	1.00	0.00

Table 17: Predictor values - mean and observations.

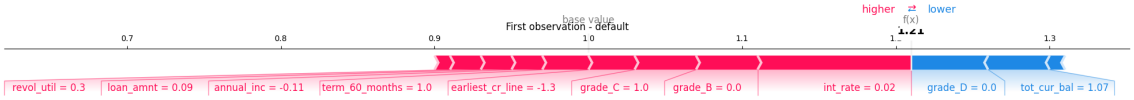

Figure 23: Local interpretation for defaulted record: force plot.

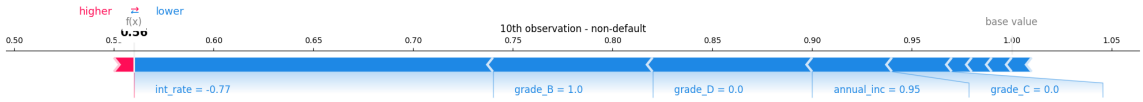

Figure 24: Local interpretation for non-defaulted record: force plot.

5.3 ROC curve and Confusion matrix

Next, the performance of the model is measured by means of the ROC curve and confusion matrix. Figure 25 shows the ROC curve for the machine learning model. The area under the curve is equal to 0.752, which is an acceptable value. This indicates that the classification power of our model is high.

Next, the confusion matrix will be used on the test set to evaluate the performance of the machine learning model. Table 18 shows the confusion matrix. Using this confusion matrix, we calculated the recall, precision and accuracy to determine the performance of the model with the F1 score which are shown in Table 19. The F1 score is equal to 68.6%, which is an

acceptable score, so by means of the F1 score we can say that the model performs well. The confusion matrix was also done for our training set. The quantities based on the confusion matrix for the training set were very close to that for the test set.

The area under the ROC curve and the F1 score obtained by the logistic model were equal to 0.71 and 0.653, respectively. The area under the curve and the F1 score obtained by the RF method are both higher than that of the logistic model, so we can say that the RF method performs better than the logistic model.

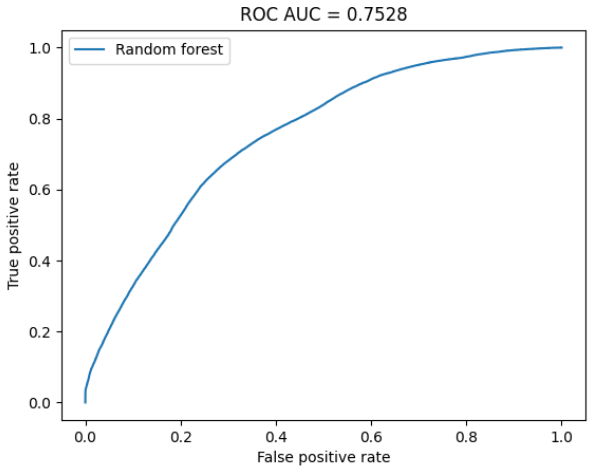

Figure 25: ROC curve.

	Predicted non-default	Predicted default
Observed non-default	32116	19200
Observed default	12940	38571

Table 18: Confusion matrix for the test set.

Recall	0.687
Precision	0.690
Accuracy	0.687
F1 score	0.686

Table 19: Quantities based on the confusion matrix for the test set.

6 Conclusion

In this assignment, we have predicted the probability of default on a loan on the due date based on customer's and loan's characteristics. Those relationships are modeled by means of a logistic model and the Random Forest method to determine which model performs best. The performance of these models are measured by means of the ROC curve, the confusion matrix and quantities obtained by the confusion matrix. Based on the area under the ROC curve and F1 score obtained by the logistic model and RF method, we can say that both models perform well. We conclude that the Random Forest method performs better than the logistic model since the area under the curve and the F1 score obtained by the RF method are both higher than the ones obtained by the logistic model. However, the difference in performance is not very high since the values do not differ much using the same dataset.