Lion's Den ING Risk Modelling Challenge

Team: Interest rate Not Guaranteed

March 14, 2024

Abstract

We were tasked with a problem to develop predictive models for estimating the probability of default on loan installments. We were supplied with the dataset comprised of various features related to borrowers' financial history, and loan characteristics.

First, logistic regression was implemented to model the relationship between the predictor variables and the binary outcome of loan default. This approach allowed for the estimation of the probability of default based on a linear combination of features, providing interpretability and ease of implementation.

Second, a challenger model using the random forest algorithm was constructed. Random forest utilizes an ensemble of decision trees to predict outcomes, offering flexibility and robustness to complex datasets. By aggregating the predictions from multiple trees, random forest provides a powerful tool for classification tasks.

The performance of both models was evaluated using metrics such as accuracy, precision, recall, and area under the curve (AUC).

Contents

1 Data

1.1 Introduction

The training dataset initially comprised 50,000 observations encompassing 36 variables. However, in adherence to the task requirements, we excluded 13,282 observations with the application status labeled as "Rejected." Consequently, our dataset was reduced to 36,718 observations. The subsequent section presents a detailed description of the variables along with their corresponding statistics.

- Numerical variables in the data set are:
  - number of applicants;
  - application amount;
  - credit duration;
  - payment frequency;
  - installment amount;
  - car value;
  - income of the main applicant;
  - income of the secondary applicant;
  - number of children of the main applicant;
  - number of dependencies of the main applicant;
  - spending's estimation;
  - number of requests during the last 3 months;
  - number of requests during the last 6 months;
  - number of requests during the last 9 months;
  - number of requests during the last 12 months;
  - limit on credit card;
  - amount on current account;
  - amount on savings account;
  - average income;

- categorical variables in the data set are:
  - default indicator – target variable;
  - loan purpose;
  - distribution channel;
  - profession of the main applicant;
  - profession of the secondary applicant;
  - material status of the main applicant;
  - property ownership for renovation;
  - vehicle type;
  - arrear in last 3 months;
  - arrear in last 12 months;
  - credit bureau score;
- date variables in the data set are:
  - application date;
  - employment date of the main applicant;
- other variables in the data set are:
  - ID;
  - customer ID;
  - application status;
  - _r_.

The descriptive statistics of the numeric variables are presented in table 1.

variable name	minimum	1st quantile	median	mean	3rd quantile	maximum	NA values
number of applicants	1.00	1.00	1.00	1.32	1.00	4.00	0
application amount	700.00	7700.00	14900.00	17729.00	25600.00	63700.00	0
credit duration (in months)	6.00	15.00	27.00	33.36	39.00	312.00	0
payment frequency	1.00	1.00	1.00	1.548	1.00	6.000	0
installment amount	53.98	597.36	1005.03	1540.38	1757.76	27075.54	0
car value	4900	21400	31400	35888	47000	132100	20538
income of the main applicant	0.00	6000.00	8880.00	10041.00	13200.00	43320.00	0
income of the secondary applicant	0.00	4400.00	6400.00	7095.00	9400.00	26700.00	28043
number of children of the main applicant	0.00	0.00	1.00	0.77	1.00	5.00	0
number of dependencies of the main applicant	0.00	0.00	1.00	1.02	2.00	8.00	0
spending's estimation	-13817	2973	4552	5504	6871	147141	32
number of requests during the last 3 months	0.00	0.00	0.00	0.55	1.00	15.00	0
number of requests during the last 6 months	0.00	0.00	1.00	0.99	1.00	18.00	0
number of requests during the last 9 months	0.00	0.00	1.00	1.35	2.00	18.00	0
number of requests during the last 12 months	0.00	0.00	1.00	1.71	2.00	26.00	0
limit on credit card	0.00	0.00	0.00	7727.00	0.00	190800.00	0
amount on current account	260.40	4925.90	8953.00	12906.60	16152.00	254161.20	7401
amount on savings account	0.00	10638.00	22876.00	30136.00	41248.00	371036.00	14649
average income	3899.00	5035.00	8569.00	8148.00	11839.00	12832.00	0

Table 1: Descriptive statistic of numeric variables.

The levels of categorical variables along with the number of observations are:

default indicator – target variable	
factor level	**number of observations**
0 (facility performing)	35591
1 (loan went into default)	1127

loan purpose	
factor level	**number of observations**
1 (car loan)	15744
2 (house renovation)	9315
3 (short cash)	10641
NA	1018

distribution channel	
factor level	**number of observations**
1 (direct)	18106
2 (broker)	12066
3 (online)	5251
Direct	248
Online	29

profession of the main applicant	
factor level	**number of observations**
1 (Pensioneer)	841
2 (Government)	1059
3 (Military)	1137
4 (Self Employed)	4345
5 (Employee)	25661
6 (Business Owner)	3245
7 (Unemployed)	430

profession of the secondary applicant	
factor level	**number of observations**
1 (Pensioneer)	183
2 (Government)	273
3 (Military)	160
4 (Self Employed)	1039
5 (Employee)	6101
6 (Business Owner)	767
7 (Unemployed)	152
NA	28043

material status of the main applicant	
factor level	**number of observations**
0 (Single)	13284
1 (Married)	9989
2 (Informal relationship)	5934
3 (Divorced)	4920
4 (Widowed)	2591

property ownership for renovation	
factor level	**number of observations**
0 (false)	1511
1 (true)	8082
NA	27125

vehicle type	
factor level	**number of observations**
0 (car)	4357
1 (motorbike)	11823
NA	20538

arrear in last 3 months	
factor level	**number of observations**
0 (false)	36259
1 (true)	459

arrear in last 12 months	
factor level	**number of observations**
0 (false)	35100
1 (true)	1618

credit score	
factor level	**number of observations**
0	19833
10	12479
20	2724
30	878
40	381
50	164
60	95
70	59
80	39
90	22
100	7
110	7
120	8
130	7
140	3
150	3
160	1
170	1
180	2
190	1
200	1
240	1
250	2

It is observed that the missing values (NA) in variables pertaining to the loan type (such as vehicle type) correspond to the specific loan type indicated. Likewise, missing values in variables associated with secondary applicants indicate the absence of a secondary applicant. However, attention is drawn to the missing values in variables related to loan purpose, amount on current account, amount on savings, and spending estimation, which require further investigation and resolution in subsequent stages of the analysis. (see section 2). Upon analyzing the minimum and maximum values of numerical variables, no significant outliers were identified, except for negative values in spending estimation. Specifically, there were 18 observations with negative

spending estimations. It is hypothesized that this may indicate a portion of their income being used to reduce spending. Consequently, no imputation was performed on these values. Additionally, discrepancies were observed in the encoding of distribution channels within the input data. For instance, the direct distribution channel was inconsistently encoded as both "1" and "Direct." To address this inconsistency, we merged these levels within this factor.

2 Data wrangling

In the data wrangling phase, we engaged in the process of transforming and enriching the provided dataset by creating new variables derived from the existing ones. This involved a systematic approach to extract additional insights and improve the dataset's comprehensiveness for subsequent analysis. Key activities included feature engineering, where we synthesized new variables based on the provided data, as well as data cleaning and preprocessing to ensure the dataset's quality and consistency. By performing these tasks, we aimed to enhance the effectiveness of our analysis and modeling efforts, ultimately contributing to more robust and reliable outcomes in our project.

Two custom functions, date_convert_1 and date_convert_2, are defined to handle date conversion tasks. These functions are designed to convert date strings into a standardized datetime format. Notably, date_convert_2 includes special handling for a specific date format, '31Dec9999', which is converted to '31Dec2200'.

The core of the data wrangling process lies within the transform function. Here's a step-by-step breakdown:

1. Reading Data: The function loads a dataset from a CSV file specified by the given path. It sets specific data types for columns Var2 and Var12 during the import process to ensure accurate data representation.

2. Column Mapping: To enhance readability and clarity, a dictionary named column_mapping is created. This dictionary serves to rename columns from their original obscure labels (e.g., Var1, Var2) to more descriptive names that reflect the data they contain (e.g., no_applicants, loan_purpose).

3. Handling Missing Values: Dealing with missing data is crucial for accurate analysis. The script employs various strategies to address missing values. Columns such as 'savings', 'spendings', and 'income_M' are filled with the median value of their respective columns. This approach helps maintain data integrity and avoids bias in subsequent analyses.

4. Feature Engineering: New features are created to provide additional insights into the dataset. For instance, the script computes the 'yearly_spend_ratio' and 'loan_to_income' ratios based on existing columns. These newly calculated features can offer valuable information for subsequent analysis tasks.

5. Data Cleaning and Preprocessing: The script ensures data cleanliness and prepares it for further analysis by performing various cleaning and preprocessing steps. It removes rows with missing target values ('y') as they are essential for supervised learning tasks. Additionally, irrelevant columns ('Application_status', 'ID', 'customer_id', 'r') are dropped to focus on relevant data. Missing values in 'account' and 'income_S' columns are filled with their respective median values. Furthermore, new columns such as 'secondary_applicant' are created to provide additional context to the dataset.

6. Further Feature Engineering: Building upon the initial feature engineering steps, additional features are generated to capture more nuanced information from the dataset. For example, the script calculates the time difference between the application date and employment start date ('empl_to_app_time'). This feature could potentially provide insights into applicant stability or employment history.

7. Categorical Variable Encoding: Categorical variables such as 'distr_channel' are encoded into numerical values to facilitate their incorporation into machine learning models. This process transforms categorical data into a format that algorithms can interpret, enabling more robust analysis and modeling.

8. Handling Missing Values (Again): The script revisits missing value handling by filling missing values for specific columns ('loan_purpose', 'profession_S', 'distr_channel') with predefined values. This step ensures that all necessary data is available for subsequent analysis tasks.

9. One-Hot Encoding: Categorical variables are further processed through one-hot encoding, a technique that converts categorical data into a binary format. This transformation expands each categorical variable into multiple binary columns, each representing a distinct category. One-hot encoding is particularly useful for machine learning algorithms that require numerical input data.

10. Returning Processed Data: Finally, the function returns the transformed DataFrame containing the processed and cleaned dataset. This prepared dataset is now ready for further analysis, exploratory data analysis (EDA), or machine learning modeling.

In summary, the provided Python script implements a comprehensive data wrangling pipeline to prepare a raw dataset for analysis or modeling. It encompasses tasks such as data cleaning, feature engineering, categorical variable encoding, and missing value handling, ensuring that the dataset is well-prepared and suitable for subsequent analytical tasks.

3 Logistic regression

As our champion model we developed logistic regression model with logit link function. It is perfectly suitable for binary classification problems, such as forecasting default of a borrower. This approach is relatively simple and is not computationally expensive compared to other options (i.e. our challenger model random forest). Moreover, logistic regression is more explainable and is easier to interpret in terms of economic intuition.

3.1 Model selection process

As a starting point we proceeded to examine one of the primary assumptions of our model: the presence of multicollinearity among our variables. This evaluation is crucial because if there are linearly dependent (or at least partially dependent) columns in our initial dataset, the outcomes derived from such a model can be misleading and even incorrect in extreme scenarios.

To ensure the statistical robustness of our model coefficients, we assessed an analysis of their Variance Inflation Factors (VIF). VIF, derived from the Tolerance vector, offers insight into the degree to which each independent variable affects the remaining ones in the model. This assessment helps ascertain any potential issues arising from multicollinearity among the predictors, which could impact the reliability and accuracy of our model's estimates.

Here is table with corresponding VIF values:

	Feature	VIF
0	no_applicants	5.410086
1	application_amount	14.174226
2	credit_duration	4.865691
3	payment_frequency	6.485652
4	installment_amount	6.144812
5	income_M	18.239379
6	income_S	15.626708
7	no_children_M	5.369603
8	no_dependencies_M	5.867079
9	spendings	4.992534
10	requests_3m	2.818109
11	requests_6m	7.029622
12	requests_9m	11.706441
13	requests_12m	8.348625
14	credit_card_limit	1.178823
15	account	2.592849
16	savings	3.191832
17	credit_score	1.378644
18	income	20.483106
19	yearly_spend_ratio	16.007042
20	loan_to_income	18.217212
21	empl_to_app_time	3.764254

Figure 1: VIF values of continuous variables

In general, a VIF value exceeding 5 is often regarded as indicative of multicollinearity. Thus, we made the decision to eliminate variables with VIF values surpassing this threshold.

To further ensure correctness of our analysis we have checked the correlogram given below, where we did not see any discrepancies.

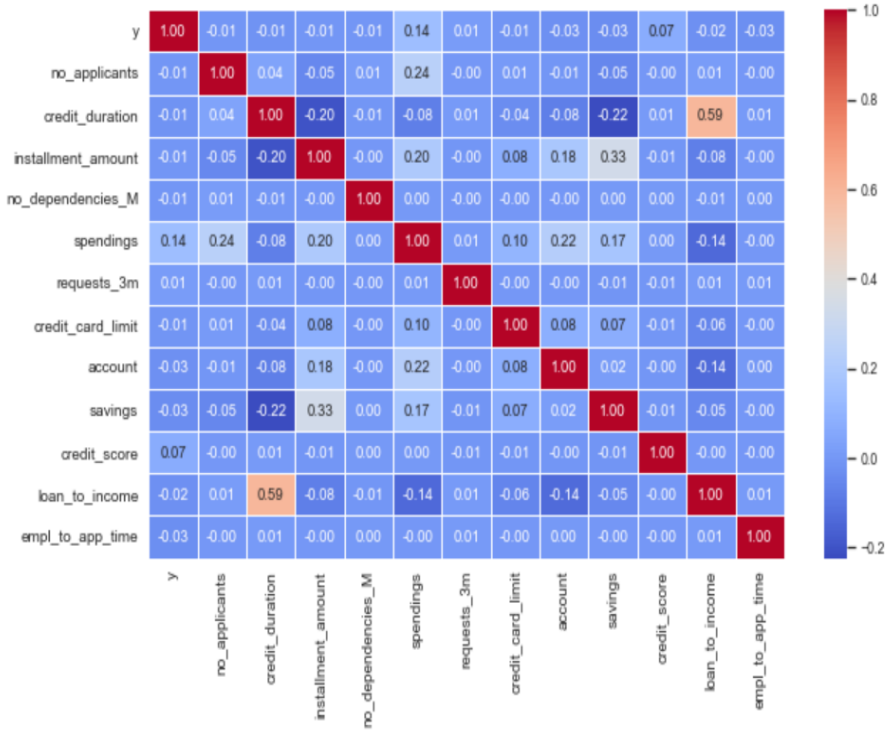

Figure 2: Correlogram of remaining variables

We then investigated the assumption of a linear relationship between independent variables and the log-odds generated by our model. To test this, we conducted a Box-Tidwell transformation. This method involves introducing a non-linear component of continuous variables into the model to assess its impact and to see if this addition makes it potentially better. Specifically, we calculated interaction terms in the form of $variable * log(variable)$, where "variable" represents a continuous predictor and "log" denotes the natural

logarithm. Subsequently, we fitted logistic model, incorporating all continuous variables along with these newly introduced non-linear interaction terms as additional independent variables, and evaluated the statistical significance of these non-linear components.

Summary of such model can be seen here:

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	514
Model:	GLM	Df Residuals:	489
Model Family:	Binomial	Df Model:	24
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-67.891
Date:	Thu, 14 Mar 2024	Deviance:	135.78
Time:	19:22:27	Pearson chi2:	454.
No. Iterations:	9	Pseudo R-squ. (CS):	0.06271
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
account	-0.0002	0.001	-0.214	0.831	-0.002	0.001
credit_card_limit	-0.0003	0.001	-0.474	0.635	-0.001	0.001
credit_duration	-0.0574	0.159	-0.361	0.718	-0.369	0.254
credit_score	0.5057	0.676	0.748	0.454	-0.819	1.830
empl_to_app_time	5.651e-05	0.003	0.019	0.985	-0.006	0.006
installment_amount	0.0002	0.003	0.063	0.950	-0.005	0.005
loan_to_income	-0.9520	1.202	-0.792	0.428	-3.308	1.404
no_applicants	-9.1583	4.052	-2.260	0.024	-17.101	-1.216
no_dependencies_M	-1.7949	1.367	-1.313	0.189	-4.474	0.884
requests_3m	1.9325	1.247	1.550	0.121	-0.511	4.376
savings	0.0002	0.001	0.430	0.667	-0.001	0.001
spendings	0.0035	0.002	1.677	0.094	-0.001	0.008
account:Log_account	1.06e-05	7.67e-05	0.138	0.890	-0.000	0.000
credit_card_limit:Log_credit_card_limit	1.871e-05	4.32e-05	0.433	0.665	-6.6e-05	0.000
credit_duration:Log_credit_duration	0.0105	0.029	0.364	0.716	-0.046	0.067
credit_score:Log_credit_score	-0.1285	0.172	-0.746	0.456	-0.466	0.209
empl_to_app_time:Log_empl_to_app_time	-1.265e-05	0.000	-0.039	0.969	-0.001	0.001
installment_amount:Log_installment_amount	6e-06	0.000	0.023	0.982	-0.001	0.001
loan_to_income:Log_loan_to_income	0.7507	0.505	1.486	0.137	-0.239	1.741
no_applicants:Log_no_applicants	5.0468	2.160	2.336	0.019	0.812	9.281
no_dependencies_M:Log_no_dependencies_M	1.0355	0.683	1.516	0.130	-0.303	2.374
requests_3m:Log_requests_3m	-0.7231	0.617	-1.172	0.241	-1.932	0.486
savings:Log_savings	-2.365e-05	4.96e-05	-0.477	0.634	-0.000	7.36e-05
spendings:Log_spendings	-0.0003	0.000	-1.662	0.096	-0.001	5.77e-05
const	3.4814	6.081	0.572	0.567	-8.438	15.401

Figure 3: Box-Tidwell transformation model

We've identified a potential issue with the variable "no_applicants", as it stands out with a statistically significant non-linear component compared to others. To address this concern, one approach is to introduce higher-order polynomial terms for the variable "no_applicants". Upon implementing

this adjustment, we observed a decrease in the significance of both terms, indicating that the addition of a single extra term is likely adequate to address the non-linearity in the model.

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	514
Model:	GLM	Df Residuals:	488
Model Family:	Binomial	Df Model:	25
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-67.883
Date:	Thu, 14 Mar 2024	Deviance:	135.77
Time:	19:24:12	Pearson chi2:	452.
No. Iterations:	9	Pseudo R-squ. (CS):	0.06274
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
account	-0.0002	0.001	-0.205	0.837	-0.002	0.001
credit_card_limit	-0.0002	0.001	-0.466	0.641	-0.001	0.001
credit_duration	-0.0573	0.159	-0.360	0.719	-0.369	0.254
credit_score	0.5078	0.676	0.751	0.452	-0.817	1.832
empl_to_app_time	4.04e-05	0.003	0.014	0.989	-0.006	0.006
installment_amount	0.0001	0.003	0.059	0.953	-0.005	0.005
loan_to_income	-0.9575	1.202	-0.796	0.426	-3.314	1.399
no_applicants	-6.6527	24.704	-0.269	0.788	-55.071	41.766
no_applicants_p2	0.6675	9.942	0.067	0.946	-18.818	20.153
no_dependencies_M	-1.8447	1.422	-1.297	0.195	-4.633	0.943
requests_3m	1.9288	1.247	1.547	0.122	-0.514	4.372
savings	0.0002	0.001	0.428	0.668	-0.001	0.001
spendings	0.0035	0.002	1.661	0.097	-0.001	0.008
account:Log_account	1.004e-05	7.69e-05	0.131	0.896	-0.000	0.000
credit_card_limit:Log_credit_card_limit	1.841e-05	4.34e-05	0.425	0.671	-6.66e-05	0.000
credit_duration:Log_credit_duration	0.0105	0.029	0.363	0.717	-0.046	0.067
credit_score:Log_credit_score	-0.1290	0.172	-0.749	0.454	-0.466	0.209
empl_to_app_time:Log_empl_to_app_time	-1.086e-05	0.000	-0.034	0.973	-0.001	0.001
installment_amount:Log_installment_amount	7.282e-06	0.000	0.028	0.978	-0.001	0.001
loan_to_income:Log_loan_to_income	0.7529	0.505	1.490	0.136	-0.237	1.743
no_applicants:Log_no_applicants	1.7657	5.681	0.311	0.756	-9.369	12.900
no_applicants_p2:Log_no_applicants_p2	0.0158	2.383	0.007	0.995	-4.656	4.687
no_dependencies_M:Log_no_dependencies_M	1.0641	0.719	1.479	0.139	-0.346	2.474
requests_3m:Log_requests_3m	-0.7219	0.617	-1.171	0.242	-1.930	0.487
savings:Log_savings	-2.358e-05	4.96e-05	-0.476	0.634	-0.000	7.36e-05
spendings:Log_spendings	-0.0003	0.000	-1.646	0.100	-0.001	6.11e-05
const	0.3582	14.938	0.024	0.981	-28.920	29.637

Figure 4: Box-Tidwell transformation model

Another thing we took a look at was if there were any strongly influential outliers in our dataset. This was done to further ensure statistical robustness of our model and to make sure our results are not skewed because of such observations.

To check that we decided to go with standard approach of calculating Cook's distance of each datapoint. Cook's Distance is an estimate of the influence

of a data point. It takes into account both the leverage and residual of each observation and can be interpreted as *how much a regression model changes when the ith observation is removed?* Rule of thumb for Cook's distance is to construct a threshold value $c = \frac{4}{N}$ where N is number of observations.

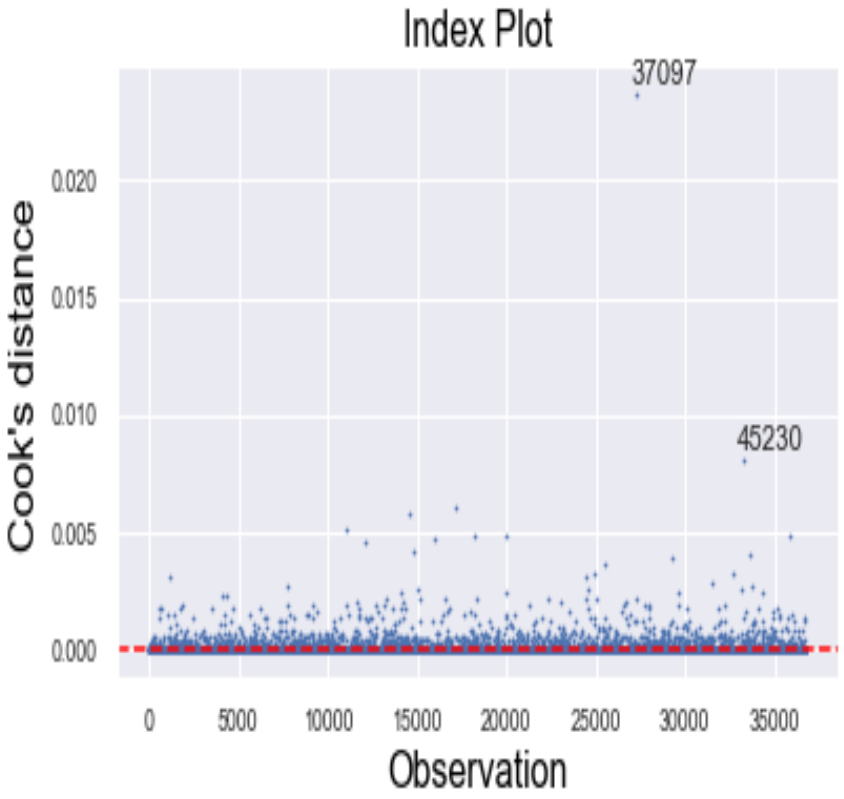

Figure 5: Cook's distance plot with indexes of most influential observations and red-dotted line indicating calculated threshold

Out of the total observations, approximately 3.3% exceeded the previously calculated threshold, indicating their high influence on the model. Despite their influential nature, these observations may still contain valuable information. To ensure that we are eliminating only the data points that significantly skew our results, we further examined the standardized residuals and raised

our Cook's distance threshold by a factor of three. Following this adjustment, the proportion of observations deemed highly influential outliers reduced to 0.2%, prompting us to remove them from the dataset.

Here we display five of the most influential outliers.

	cooks_d	std_resid
45230	0.008086	10.050279
44390	0.003246	10.242749
33199	0.003175	13.160093
45054	0.002677	12.704723
5745	0.002310	10.524859

Figure 6: Most influential observations

After above preliminary work we were ready to fit first proper model with all remaining variables. The initial step in simplifying our model involved eliminating variables that were not statistically significant. We utilized a standard t-test, assessing the significance of each variable based on 10% threshold for the p-value. The t-test evaluates the statistical significance of a variable by calculating its corresponding z-value under the assumption that the coefficient of the variable in the model equals zero. Notably, this procedure was conducted iteratively, employing backward elimination from the full model, as the removal of each variable affects the coefficients of the remaining variables.

Following the aforementioned procedure we are left only with predictor variables that are deemed statistically significant, as displayed in the figure below:

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	36662
Model:	GLM	Df Residuals:	36639
Model Family:	Binomial	Df Model:	22
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3943.7
Date:	Thu, 14 Mar 2024	Deviance:	7887.3
Time:	20:32:04	Pearson chi2:	3.20e+04
No. Iterations:	8	Pseudo R-squ. (CS):	0.04768
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.9968	0.157	-12.754	0.000	-2.304	-1.690
credit_duration	-0.0034	0.001	-2.399	0.016	-0.006	-0.001
spendings	8.094e-05	5.07e-06	15.962	0.000	7.1e-05	9.09e-05
requests_3m	0.0880	0.031	2.880	0.004	0.028	0.148
credit_card_limit	-3.091e-06	1.75e-06	-1.766	0.077	-6.52e-06	3.39e-07
account	-3.377e-05	4.21e-06	-8.026	0.000	-4.2e-05	-2.55e-05
savings	-2.102e-05	2.4e-06	-8.755	0.000	-2.57e-05	-1.63e-05
credit_score	0.0228	0.002	12.292	0.000	0.019	0.026
zero_income	1.3221	0.064	20.498	0.000	1.196	1.448
loan_purpose_2	-0.2135	0.094	-2.282	0.022	-0.397	-0.030
loan_purpose_3	0.3571	0.088	4.045	0.000	0.184	0.530
profession_M_1	1.0087	0.129	7.820	0.000	0.756	1.262
profession_M_2	-1.0739	0.254	-4.223	0.000	-1.572	-0.575
profession_M_3	-1.8423	0.337	-5.475	0.000	-2.502	-1.183
profession_M_4	-0.4810	0.118	-4.085	0.000	-0.712	-0.250
profession_M_5	-0.3660	0.081	-4.493	0.000	-0.526	-0.206
profession_M_6	-0.5644	0.132	-4.269	0.000	-0.824	-0.305
profession_M_7	1.3221	0.064	20.498	0.000	1.196	1.448
profession_S_0	0.6689	0.091	7.341	0.000	0.490	0.847
profession_S_1	1.6071	0.277	5.796	0.000	1.064	2.151
profession_S_7	1.5725	0.327	4.810	0.000	0.932	2.213
material_status_M_1	-0.3584	0.088	-4.078	0.000	-0.531	-0.186
material_status_M_3	0.4859	0.089	5.477	0.000	0.312	0.660
material_status_M_4	0.3082	0.119	2.584	0.010	0.074	0.542
arrear_12m_0	-1.9046	0.088	-21.682	0.000	-2.077	-1.732

Figure 7: Initial model after statistical significance check

As a last step, we opted to refine our model using the Akaike Information Criterion (AIC), aiming to minimize the Kullback-Leibler distance between the given model and the 'true' model. The primary objective of this step was to further streamline the number of predictor variables and ensure the optimization of our model based on a quantifiable metric. Similar to the first step, this procedure required a sequential approach, with variables being excluded or added individually in an iterative manner.

After that procedure we are left with a model that looks like this:

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	36662
Model:	GLM	Df Residuals:	36639
Model Family:	Binomial	Df Model:	22
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3943.7
Date:	Thu, 14 Mar 2024	Deviance:	7887.3
Time:	21:35:11	Pearson chi2:	3.20e+04
No. Iterations:	8	Pseudo R-squ. (CS):	0.04768
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.8316	0.157	-11.696	0.000	-2.138	-1.525
profession_M_7	2.4789	0.121	20.498	0.000	2.242	2.716
arrear_12m_0	-1.9046	0.088	-21.682	0.000	-2.077	-1.732
spendings	8.094e-05	5.07e-06	15.962	0.000	7.1e-05	9.09e-05
credit_score	0.0228	0.002	12.292	0.000	0.019	0.026
profession_M_1	0.8435	0.128	6.589	0.000	0.593	1.094
savings	-2.102e-05	2.4e-06	-8.755	0.000	-2.57e-05	-1.63e-05
account	-3.377e-05	4.21e-06	-8.026	0.000	-4.2e-05	-2.55e-05
loan_purpose_3	0.3571	0.088	4.045	0.000	0.184	0.530
material_status_M_3	0.4859	0.089	5.477	0.000	0.312	0.660
profession_S_0	0.6689	0.091	7.341	0.000	0.490	0.847
profession_S_1	1.6071	0.277	5.796	0.000	1.064	2.151
profession_S_7	1.5725	0.327	4.810	0.000	0.932	2.213
material_status_M_1	-0.3584	0.088	-4.078	0.000	-0.531	-0.186
profession_M_3	-2.0076	0.339	-5.915	0.000	-2.673	-1.342
requests_3m	0.0880	0.031	2.880	0.004	0.028	0.148
material_status_M_4	0.3082	0.119	2.584	0.010	0.074	0.542
profession_M_2	-1.2391	0.256	-4.842	0.000	-1.741	-0.738
credit_duration	-0.0034	0.001	-2.399	0.016	-0.006	-0.001
loan_purpose_2	-0.2135	0.094	-2.282	0.022	-0.397	-0.030
credit_card_limit	-3.091e-06	1.75e-06	-1.766	0.077	-6.52e-06	3.39e-07
profession_M_5	-0.5313	0.078	-6.774	0.000	-0.685	-0.378
profession_M_4	-0.6463	0.116	-5.555	0.000	-0.874	-0.418
profession_M_6	-0.7296	0.131	-5.557	0.000	-0.987	-0.472

Figure 8: Model after AIC stepwise

3.2 Final model

After all steps described above we ended up with Final model given in Figure 8.

3.2.1 Assumptions

Although some already mentioned in 3.1 subsection, here we outline the assumptions underlying our final logistic regression model.

1. Sufficiently big sample size and correct outcome type

   This assumption is met because we have 36662 observations our model was built on and our predicted variable is of binary type.

```
y.unique()
```

```
array([0., 1.])
```

```
y.value_counts()
```

```
y
0.0    35591
1.0     1071
Name: count, dtype: int64
```

2. Linear relation between continuous predictive variables and log-odds

   Same as before, we performed Box-Tidwell transformation on our final model. Summary of fitted model with additional non-linear components can be seen below.

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	2414
Model:	GLM	Df Residuals:	2403
Model Family:	Binomial	Df Model:	10
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-390.31
Date:	Thu, 14 Mar 2024	Deviance:	780.61
Time:	22:17:32	Pearson chi2:	2.35e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.01617
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
spendings	0.0003	0.000	0.751	0.453	-0.000	0.001
credit_score	-0.0376	0.049	-0.759	0.448	-0.135	0.059
savings	0.0002	0.000	0.901	0.368	-0.000	0.001
account	-0.0002	0.000	-0.942	0.346	-0.001	0.000
credit_card_limit	8.375e-05	0.000	0.438	0.661	-0.000	0.000
spendings:Log_spendings	-1.671e-05	3.03e-05	-0.552	0.581	-7.61e-05	4.26e-05
credit_score:Log_credit_score	0.0089	0.010	0.932	0.352	-0.010	0.028
savings:Log_savings	-1.756e-05	1.8e-05	-0.975	0.329	-5.29e-05	1.77e-05
account:Log_account	1.778e-05	1.94e-05	0.916	0.360	-2.03e-05	5.58e-05
credit_card_limit:Log_credit_card_limit	-7.136e-06	1.58e-05	-0.451	0.652	-3.82e-05	2.39e-05
const	-3.6701	0.969	-3.789	0.000	-5.569	-1.772

Figure 9: Box-Tidwell transformation of Final model

3. Absence of multicollinearity

   Once again, we checked VIF and corresponding correlogram.

	Feature	VIF
0	spendings	2.144141
1	credit_score	1.253134
2	savings	1.807616
3	account	1.822637
4	credit_card_limit	1.149990

Figure 10: VIF of variables used in Final model

We can clearly observe that there is no problem with multicollinearity between our predictor variables.

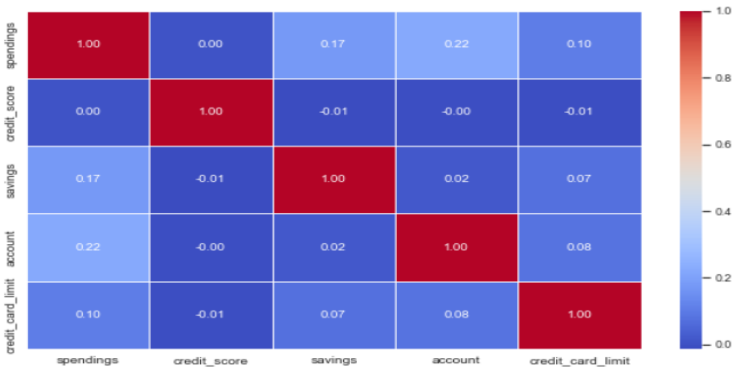

Figure 11: Correlogram of variables used in Final model

4. No strongly influential outliers

   Same as before we used Cook's distance to check this assumption, this time percentage of influential observations is just under 0.3%, so we think it is acceptable and assumption is met.

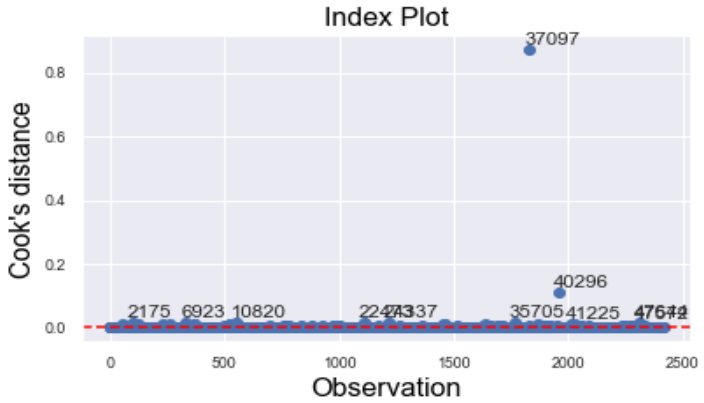

Figure 12: Cook's distance of observations from training data used to fit Final model

3.2.2 Economic interpretation of model

The analysis of signs of the coefficients in the final model provides insight into the driving factors behind the probability of default. If the coefficient corresponding to given variable is positive, the increase of this coefficient causes the increase of the probability of default. If the coefficient corresponding to given variable is negative, the increase of this coefficient causes the decrease of the probability of default. note that this analysis assumes that other coefficient do not change, therefore true impact may differ due to correlations.

The variables with positive coefficient are:

- spending's;
- profession of main applicant == unemployed;
- profession of main applicant == pensioneer;
- credit score;
- loan purpose == short cash;
- material status of main applicant == divorced;
- material status of main applicant == widowed;
- lack of secondary applicant;
- profession of secondary applicant == unemployed;
- profession of secondary applicant == pensioneer;
- request in last 3 months.

The variables with positive coefficient are:

- arrear in last 12 months==0;
- savings;
- account;
- material status of main applicant==Married;

- profession of main applicant == Military;
- profession of main applicant == Government;
- profession of main applicant == Self employed;
- profession of main applicant == Employee;
- profession of main applicant == Business owner;
- credit duration;
- loan purpose==House Renovation;
- credit card limit;

We note that the analysis of coefficient signs is consistent with our intuition regarding the driving factors behind the probability of default.

3.2.3 Model performance

In this subsection we present descriptive statistics of our Final regression model both on Training and Testing data sets respectively.

The cutoff we used to calculate descriptive statistics of our model was set to 0.04 as the minimizor of:

$$\underset{0<c<1}{\operatorname{argmin}}(1 - FPR(c))^2 + (TPR(c))^2,$$

where TPR is the true positive rate on the trainig data and FPR is the false positive rate in the training data.

Training data:

- Accuracy: 0.88
- Precision: y = 1: 0.98, y = 0: 0.12
- Recall: y = 1: 0.89, y = 0: 0.50
- F1: y = 1: 0.93, y = 0: 0.20

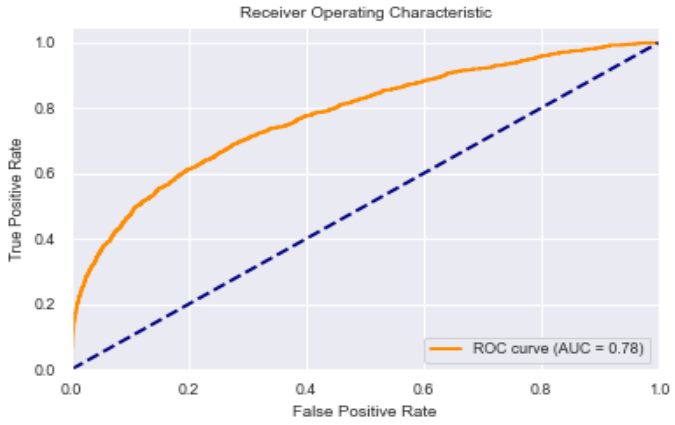

Figure 13: ROC curve on training data for Final model

Testing data:

- Accuracy: 0.87
- Precision: y = 1: 0.98, y = 0: 0.12
- Recall: y = 1: 0.88, y = 0: 0.47
- F1: y = 1: 0.93, y = 0: 0.19

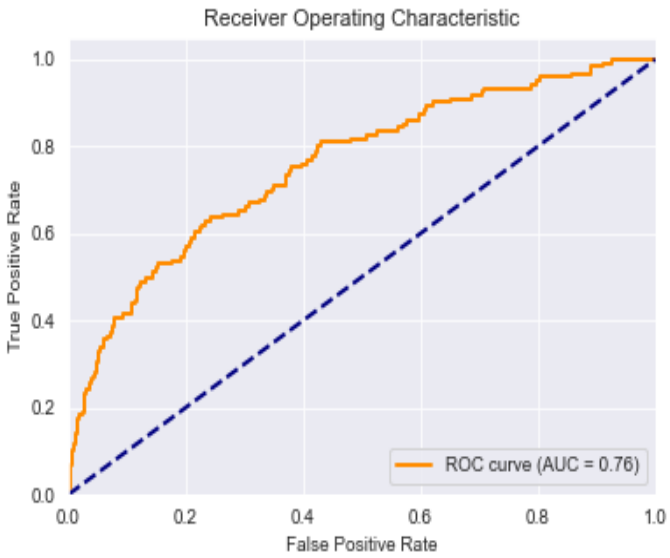

Figure 14: ROC curve on testing data for Final model

In summary, our analysis indicates that the results obtained on both the training and testing datasets are satisfactory. Through rigorous data preprocessing, variable transformation, and model training, we have achieved outcomes that align with our expectations. The performance metrics and evaluation criteria applied to both datasets demonstrate the effectiveness and generalizability of our model. Overall, our findings suggest that the model exhibits robustness and reliability in predicting the desired outcome.

4 Challenger model

4.1 Introduction

We have decided to use random forest as the challenger model. Random forests are known to be robust for the classification problems as they tend to ignore outliers. Furthermore random forests are able to capture non-linear dependencies in the data. It should be noted that random forests are not optimizing the AUC nor confusion matrix on the training data due to the fact that each tree in the forest is trained at the subset of data both in terms of observations and predictors. It should be noted that the random forest falls into black-box category of models. We note that the lack of interoperability is a significant drawback when it comes to the default prediction model, however it is a reasonable challenger to the linear regression as the better performance of the random forest may indicate that the treatment of outliers and nonlinearities in the data is poorly handled in the logistic regression model. For the detailed description of random forest algorithm please refer to bibliography.

For the purpose of the building of random forest model we are using all variables in imputed data set described in Section 2, as the random forests have tha capability to marginalise the influence of variables with qestionable predictive power.

In order to calibrate random forest parameters we have decided to perform empirical tuning for

- **Maximum number of nodes in each tree** – In order to achieve robustness in the random forest algorithm, each node does not consider all variables for the optimal split search, but limits itself to the random subset of variables. Where size of this set is the optimized parameter. The rule of thumb in the literature suggest that the square root of total number of variables should be used.
- **Number of trees in the forest** – Number of trees in the forest parameter tends to increase both calibration time and predictive power. With more trees in the forest, predictions from individual trees are more averaged and random noise in the prediction is decreased. The

rule of thumb in the literature is to use the square root of the number of explanatory variables. In the case of our data that umber is equal to 6.

- **Maximum depth of the tree** – Predictive power of the random forest tends to increase with the maximum depth of the tree, however in some cases it is possible that vast trees may tend to overfit. It should be noted that this parameter have significant impact on the training speed.

4.2 Building process

Initial random forest

Firstly, we have built the random forest using 2000 trees with maximum depth of 5 and 6 variables considered at each split. The ROC curve along with AUC metric on both training and test data of this model is presented in figure 15.

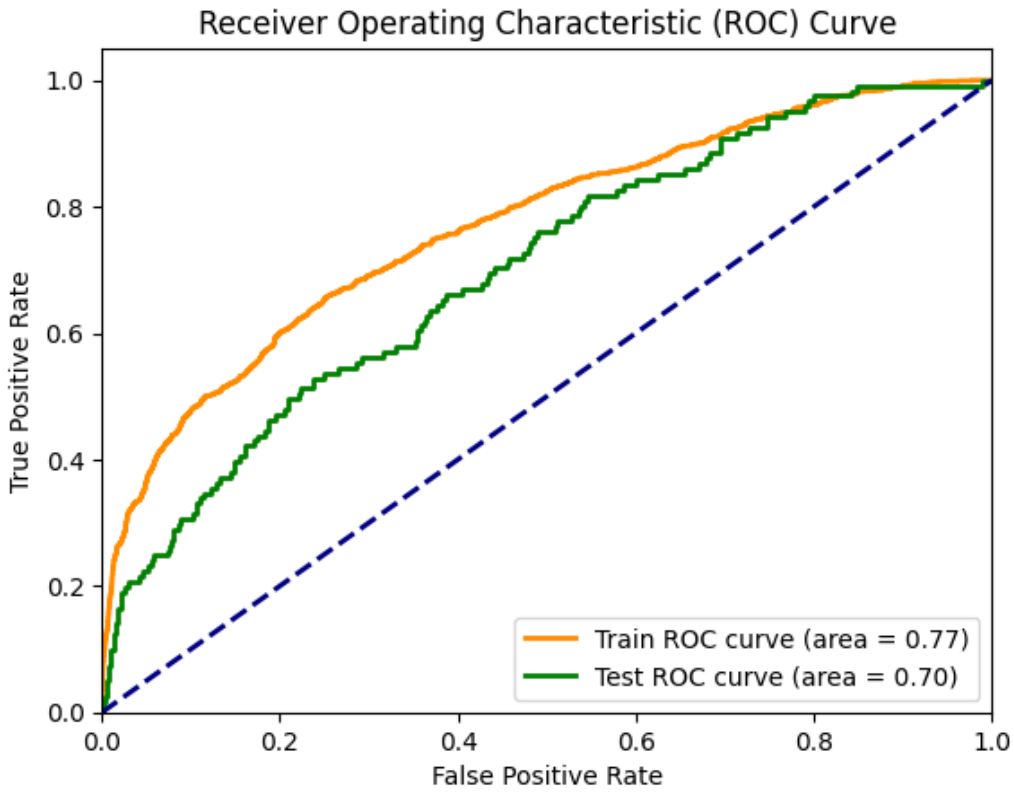

Figure 15: Performance of the initial random forest model.

Second random forest

Secondly, we tried to increase the maximum depth of the trees. The second model is using 2000 trees with depth of 8 and 6 variables considered at each split. The ROC curve along with AUC metric on both training and test

data of this model is presented in figure 16. One can observe that the AUC increased slightly for both training and test data.

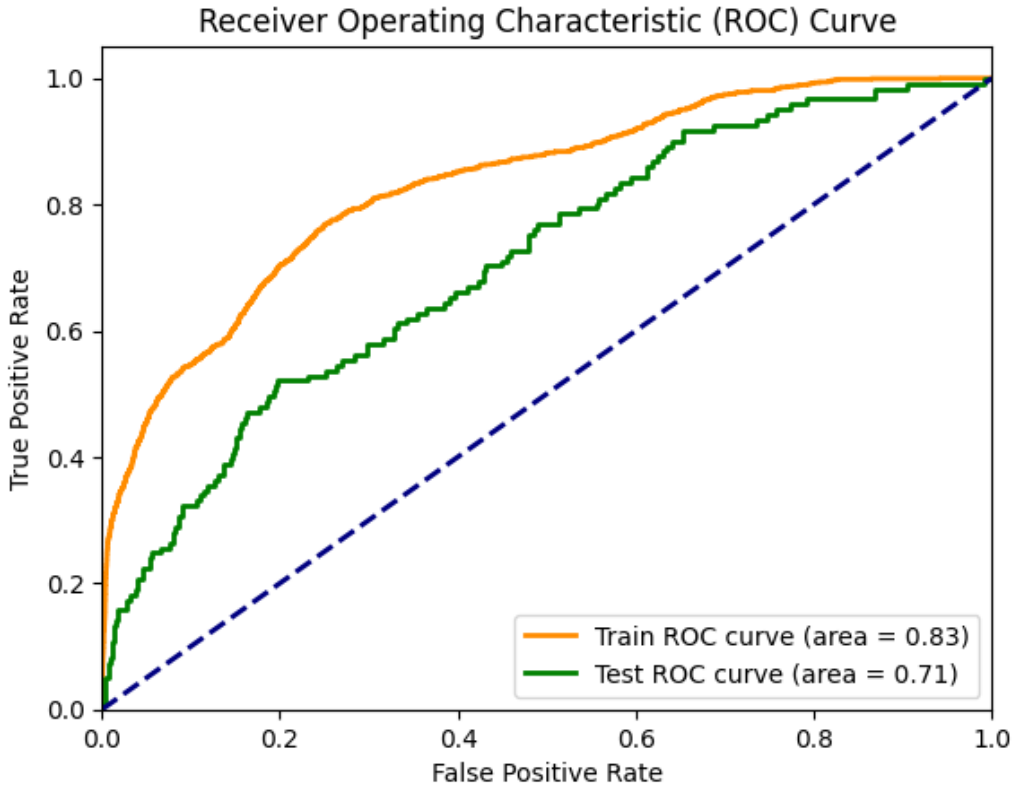

Figure 16: Performance of the second random forest model.

Third random forest

Next, we considered a family of 18 random forests with N trees, maximum depth of M and 6 variables considered at each split, where:

- $N \in \{50, 100, 200, 400, 1000, 2000\}$;
- $M \in \{5, 10, \infty\}$;

and search for the model which maximises the AUC on training data. It should be noted that when the M is equal to ∞ the nodes are expanded until all leaves are pure or until all leaves contain less than 2.

The best model found, was the model with 1000 trees, maximum depth of 10 and 6 variables considered at each split. The ROC curve along with AUC metric on both training and test data of this model is presented in figure 17.

We have decided to not extend our search of better number of trees and maximum depth as the increase of the AUC was not significant and the optimising process is computationally expensive.

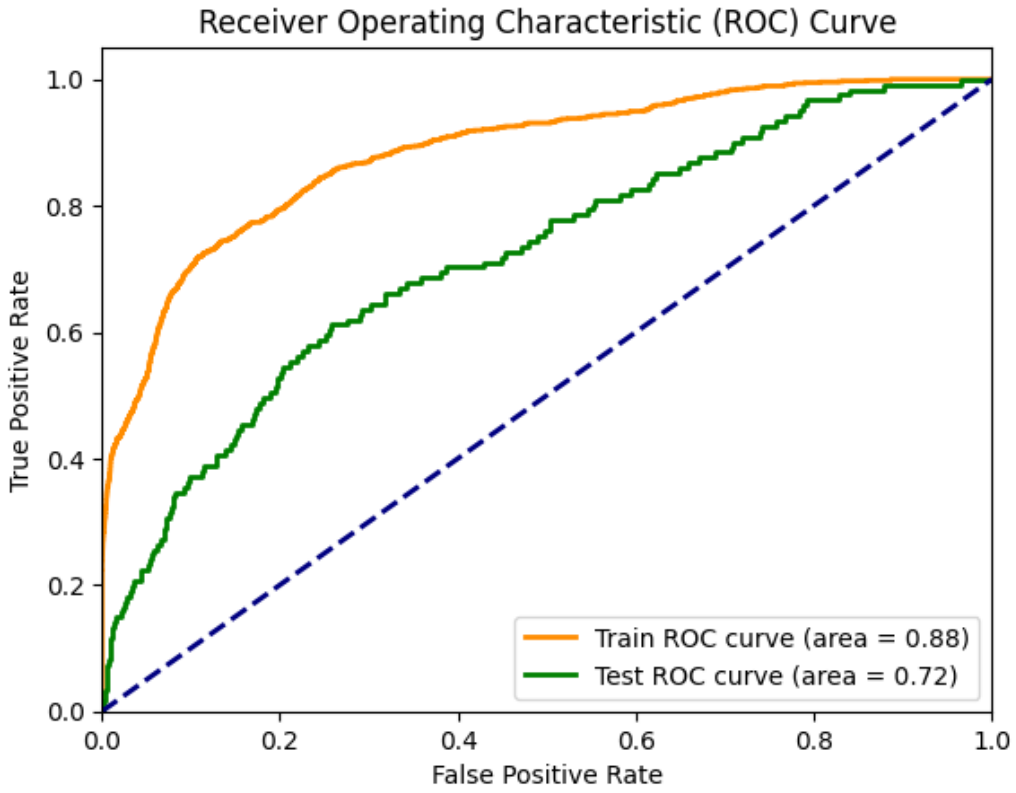

Figure 17: Performance of the third random forest model.

Final random forest

Lastly, we optimised the model across different numbers of variables considered at each split. We considered values of 1, 2, 3, 5, 6, 10 and 20. The model which performed the best in terms of training set AUC was the model

with 2 variables considered at each split. The ROC curve along with AUC metric on both training and test data of the final model is presented in figure 18. We decided to stop our search for best model parameters at this stage as the AUC metric is on the satisfactory level and the training process for random forest is computationally expensive.

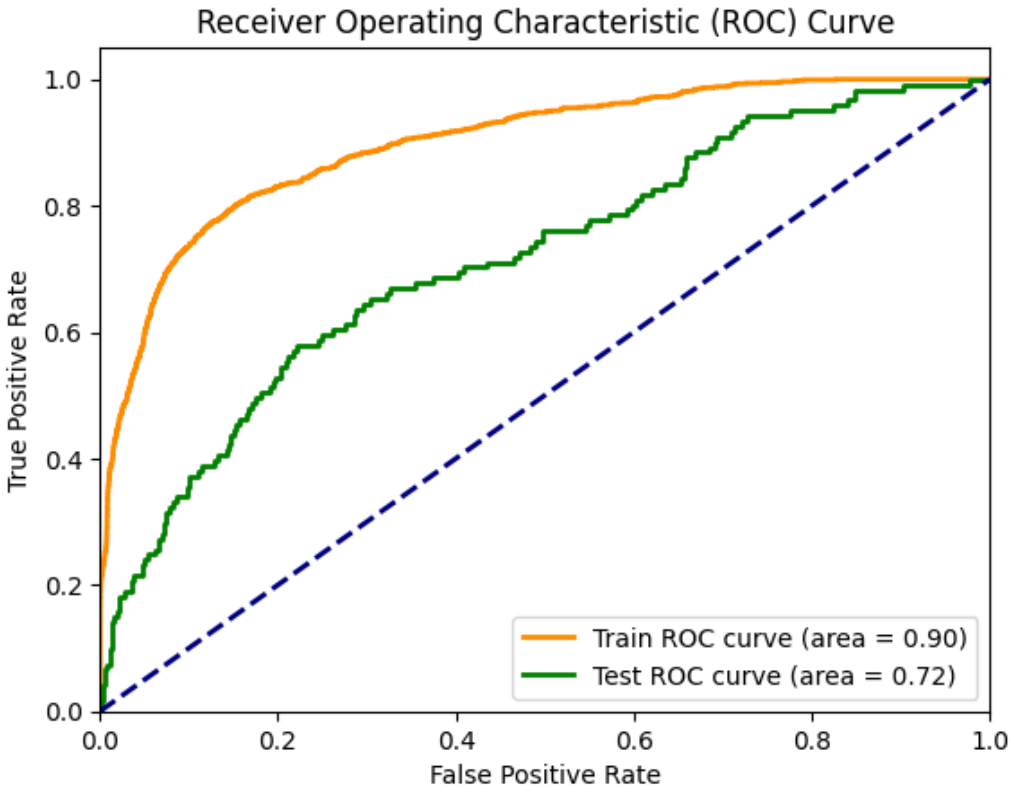

Figure 18: Performance of the final random forest model.

4.3 Conclusion

Random forest algorithm with empirical tuning has strong predictive power, however the model is inferior to the logistic regression model build in Section 3 in terms of AUC on test data. We note that the AUC on training set favours the random foest, but this may indicat the overfit of random forest model. Moreover, the black box approaches may not be suitable in the credit default classification problems, as they lack intractability and do not provide insight into the dependencies between the default events and the quantitative characteristics of the customers. Furthermore, computational speed of the random forest is far worse than the computational speed of the logistic regression. For those reason we decided to reject the random forest model in favor of the logistic regression. That said, this exercise provided comfort around the treatment of outliers and nonlinearities in the data, as the random forest which is known to be robust in terms of this problems did not outperform the logistic regression. The comparison of the ROC curves of random forest and logistic regression model on both training and test data are presented in table 2.

	Training AUC	Test AUC
logistic regression	0.78	0.76
random forest	0.90	0.72

Table 2: AUC on training of training and test data of logistic regiression and random forest.

5 Bibliography

[1] L. Breiman. Random forests. Machine Learning, 45:5–32, 2001
[2] https://github.com/kennethleungty/Logistic-Regression-Assumptions/blob/main/Logistic_Regression_Assumptions.ipynb