ASSESSMENT CASE STUDY

APPLICATION OF CREDIT RISK MODELS IN THE BANKING INDUSTRY

GUNAVANTH MAHENDRA DUREMANTHI
SY IAQS 2022

Contents

Background

With the rise in Non-Performing Assets (NPAs), our bank as well as the industry has experienced significant losses due to the lack of a credit risk assessment framework. In accordance with the newly issued regulatory framework, we have decided to go with the 'Internal Model Approach'.

A credit rating model is going to be used to evaluate the creditworthiness of potential loan obligors by evaluating the likelihood of default based on historical data. A robust model building process to design, develop, and implement a credit risk model for loan underwriting and credit risk provisioning was undertaken. This model will be used to assess future loan disbursements and to restructure disbursed facilities of existing obligors.

Deliverables

- Data to be collected, cleaned, and analysed thoroughly.
- Credit rating model to be built using cleaned data and robust testing to be conducted to assess model assumptions and fit.
- Determination of thresholds and metrics to assist in making decisions.
- Documentation of the model methodology and assumptions.

Steps in the model building process:

1. Data Collection and Cleaning: Data was collected in liaison with the data collection team. The raw data was then inspected for discrepancies and cleaned accordingly as detailed in the report below.
2. Exploratory Data Analysis: Once the data was cleaned, Exploratory Data Analysis (EDA) was performed to gain further insight into the data we are working with. Various measures were calculated, and graphs were plotted for the same.
3. Data Preparation for Model Build: The data was then split into training and testing sets to build the model. Model specific data preparation was done wherever required.
4. Model Building: Once the data was prepared, models were built, and multiple iterations were run to check for the best fitting model. In case of ensemble learning algorithms, parameters were tuned accordingly.
5. Model Testing and Evaluation: Tests were run on models wherever required, and in-sample and out-sample analysis was performed to ensure that there is no overfitting or underfitting in the models that have been prepared.
6. Conclusion and Model Selection: The final model was selected based on computed and comparable metrics across the models. The final model was selected based on its prediction accuracy and speed of computation.

All data cleaning, analysis, and model building was performed using the R programming language and RStudio as the IDE.

Data Description

Data Overview:

As part of the data collection process, data was collected from 6000 existing obligors. The following information for each obligor was collected:

- Account Number (Numeric)
- Customer ID (Numeric)
- Checking Balance (Numeric)
- Savings Balance (Numeric)
- Loan Duration in months (Numeric)
- Amount of loan (Numeric)
- Number of Existing Loans (Numeric)
- Age (Numeric)
- Instalment as percentage of Income (Numeric)
- Number of years at current residence (Numeric)
- Credit History (Categorical: Perfect, Very Good, Good, Poor, or Critical)
- Purpose (Categorical: Business, Car, Education, Furniture/Appliances or Renovations)
- Employment Duration (Categorical: Less than 1 year, 1 to 4 years, 4 to 7 years, greater than 7 years, or Unemployed)
- Other credit (Categorical: Bank, Store, None)
- Housing (Categorical: Own, Rent, Other)
- Job (Categorical: Management, Skilled, Unskilled, Unemployed)
- Number of Dependants (Categorical: 1 or 2)
- Ownership of Phone (Categorical: Yes or No)
- Default (Categorical: Yes or No)

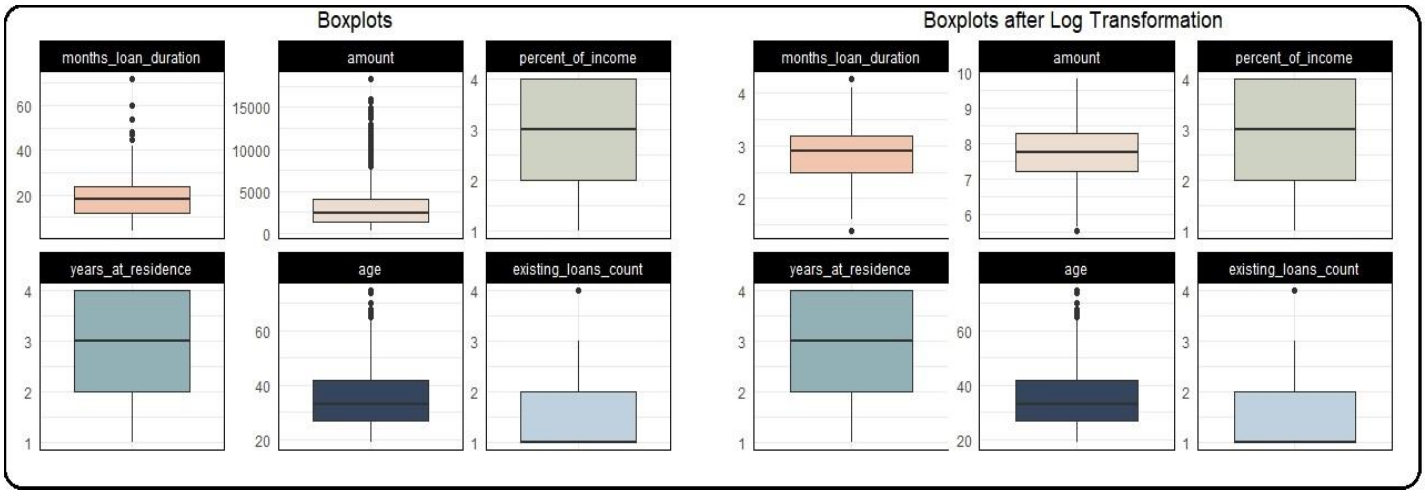
Data Management

Errors in Data:

We see that there are a lot of customers whose savings balances (18.3%) or checking balances (39.4%) are unknown. Furthermore, there are a large proportion of customers (22%) whose data for number of years at current residence are missing. Additionally, there is a spelling error in the 'purpose' variable wherein 'car' was misspelt as 'car0' in certain places.

Data Cleaning:

- Variables Customer ID and Account Number were dropped as they are redundant.
- The variable amount (USD) was renamed to amount.
- Discrepancies were checked for in categorical variables, and 'car0' was replaced by 'car' in the purpose variable.
- The missing values in 'years at residence' were replaced by the median of years at residence.
- Outlier analysis on the variables was performed and it was seen that the variables 'amount' and 'duration of loan' had significant number of outliers. Log transformations were used to overcome the effect of outliers.



- Given the high proportion of 'unknown' values in savings balance and checking balance, a gradient boosting[1] algorithm was used to predict these values. The algorithm had an in-sample accuracy of 99.1% and an out-sample accuracy of 98.36%. The variable default was not used to predict these values as it can affect the further credit modelling process.
- The variables Credit History, Purpose, Employment Duration, Other credit, Housing, Job, Number of Dependants, Ownership of Phone and Default were converted to factor type.
- The variables Checking Balance, Savings Balance, Loan Duration in months, Amount of loan, Number of Existing Loans and Age were converted to numeric type.

[1] Gradient Boosting is an algorithm used in regression and classification tasks. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.

Data Analysis and EDA

Variable Identification:

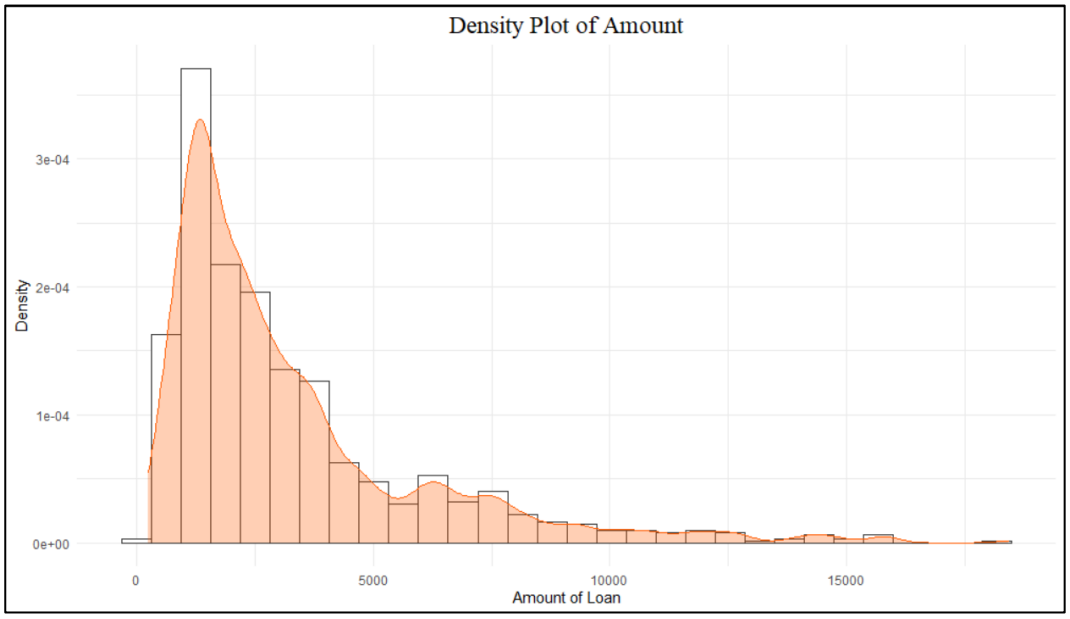
Post the data cleaning process, the response and explanatory variables were determined. 'Default' was chosen as the response variable, and the remaining 16 variables are possible explanatory variables.

Insights:

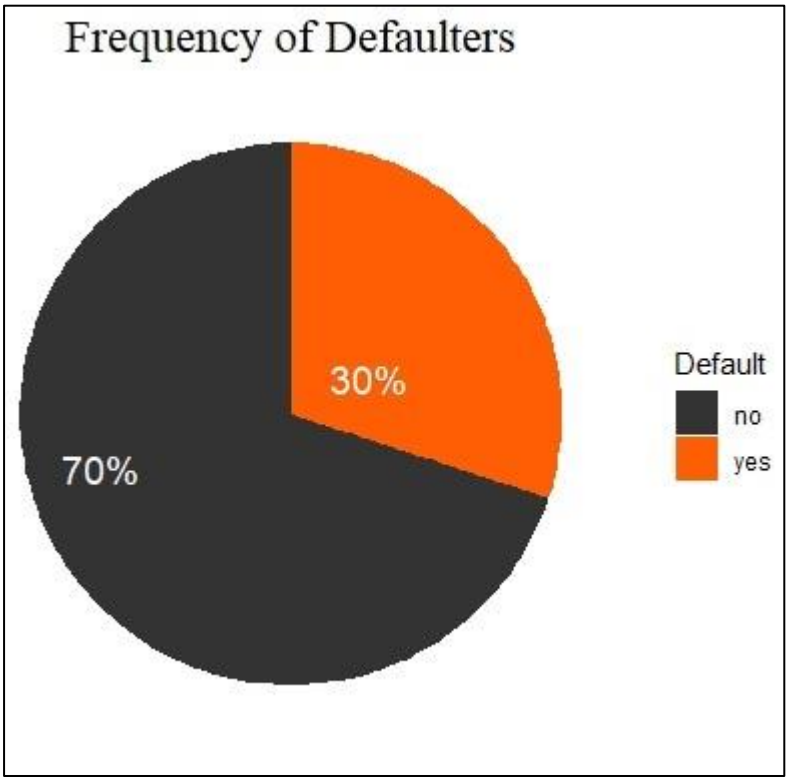
- The cleaned data set had 6000 data points across 17 columns.
- The variables duration of loan, loan amount, percent of income, years at residence, age, and existing number of loans are numeric. All other variables are factors.

As part of the exploratory data analysis exercise, several graphs were plotted, and inferences were made.

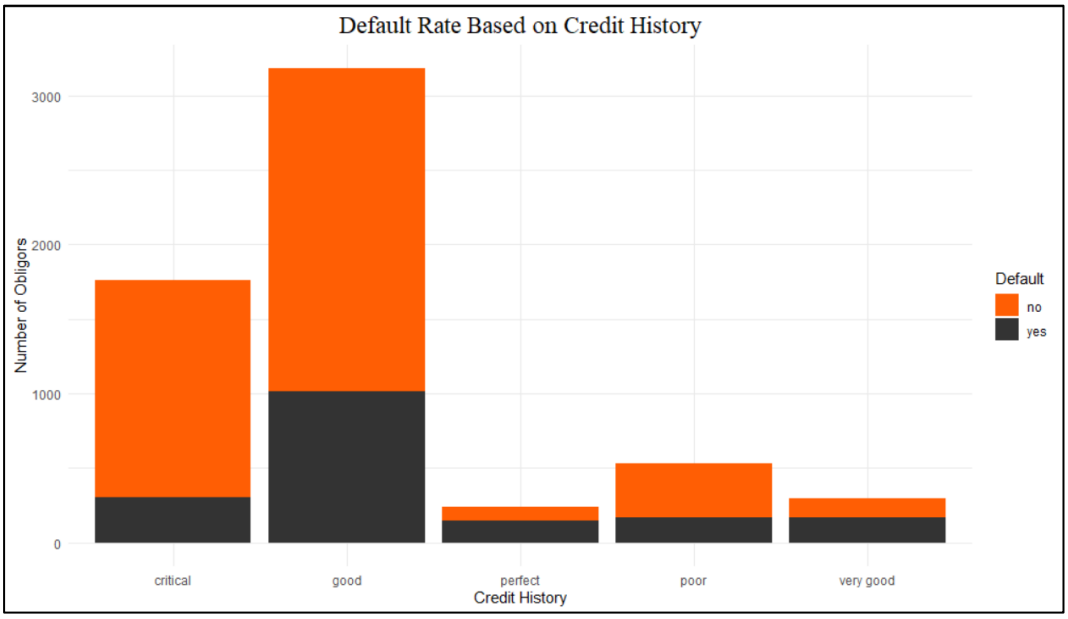
- Loan amounts and age were heavily positively skewed.



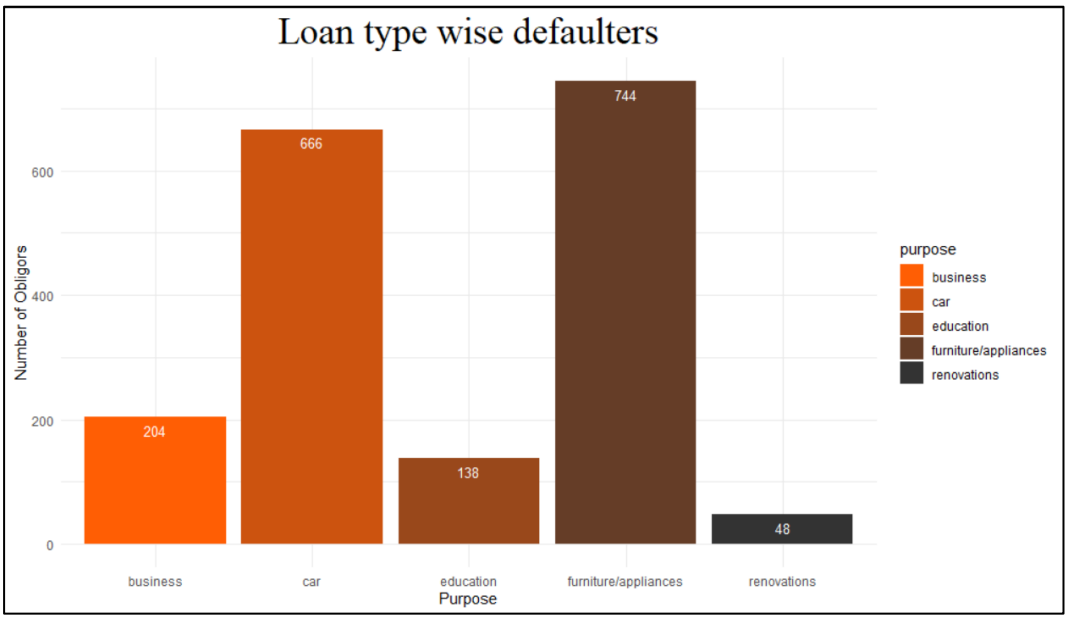
- The proportion of obligors that defaulted on their loans is 30%.



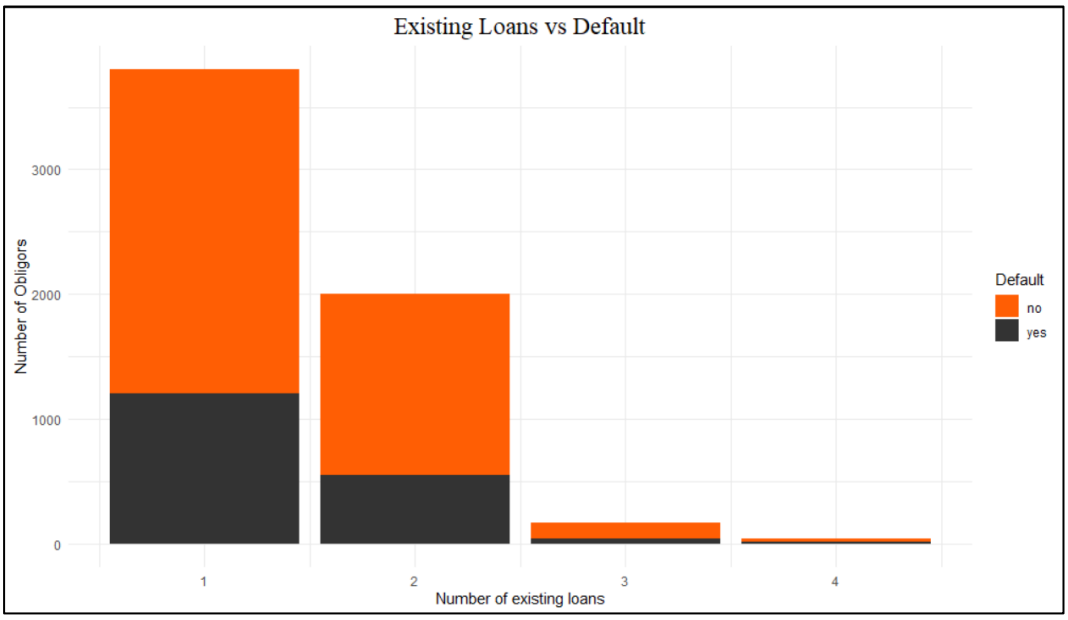
- The proportion of defaulters amongst people with 'perfect' credit history was highest, albeit the number of people with 'perfect' credit history was the lowest.
- In general, as the credit history got better the proportion of defaulters increased. This may be due to there being significantly fewer people with better credit histories.



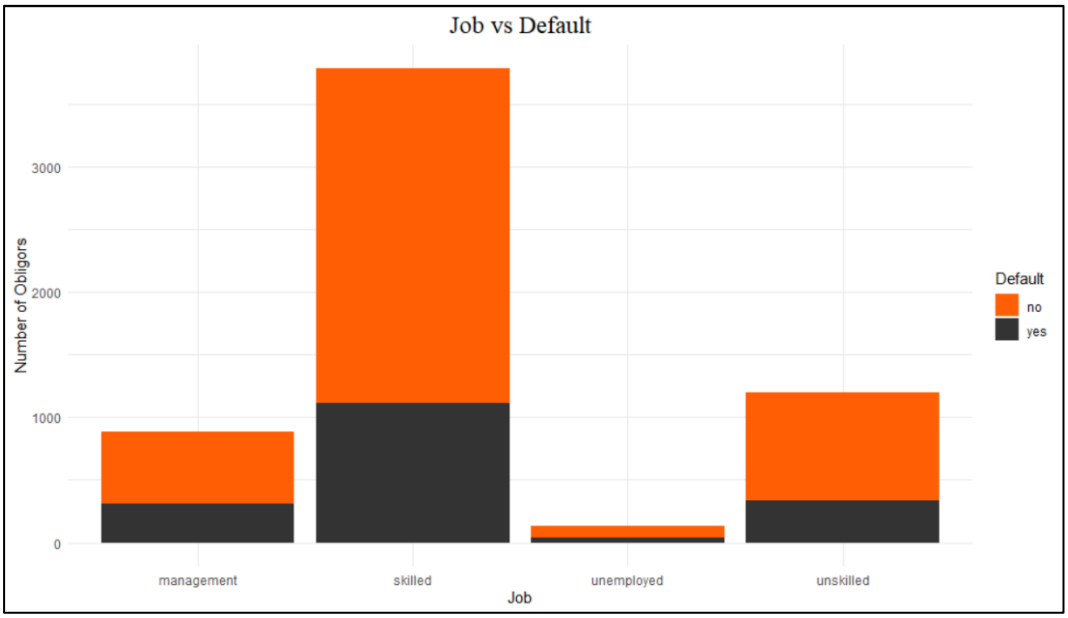
- Most loan defaults were on loans that were taken for furniture/appliance purpose. But default rates were similar across, with the default rate being lowest for furniture/appliance (this could be down to the larger number of loans for this purpose).



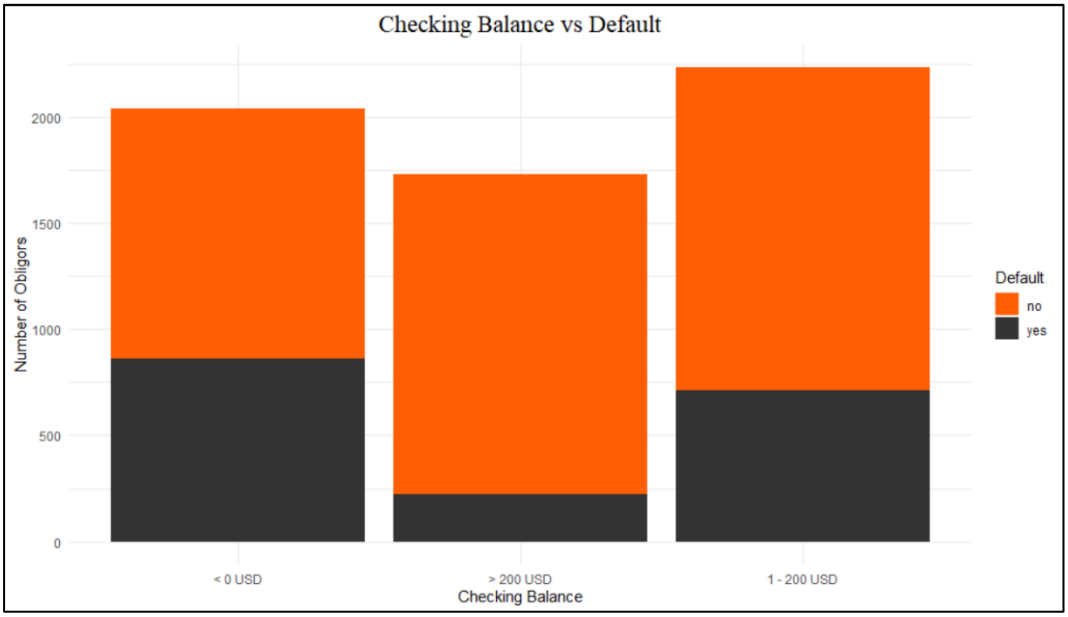
- Default rate and number of defaults among people with one existing loan was the highest.



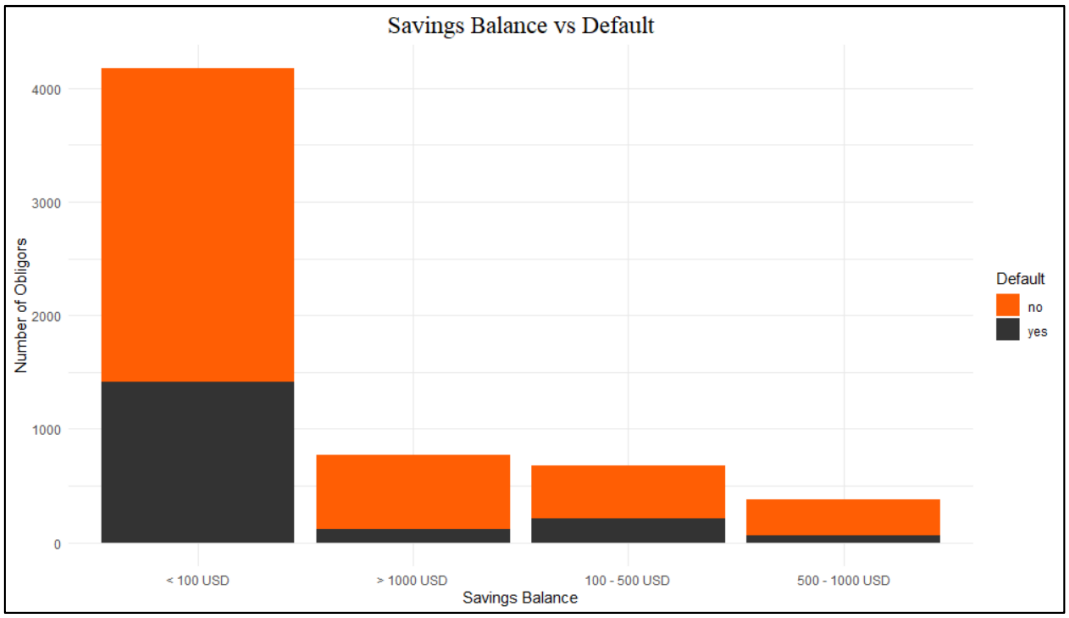
- Default rate remains fairly independent of job, but the highest default rate is amongst people who hold management positions.



- As expected with checking balance, the highest default rate is among those whose checking balance is less than $0 and lowest default rate amongst those with checking balance greater than $200.



- The trend across savings balance is comparable to the one seen in case of checking balance. As the amount held in a savings account increases, the default rate decreases.



Correlation Analysis:

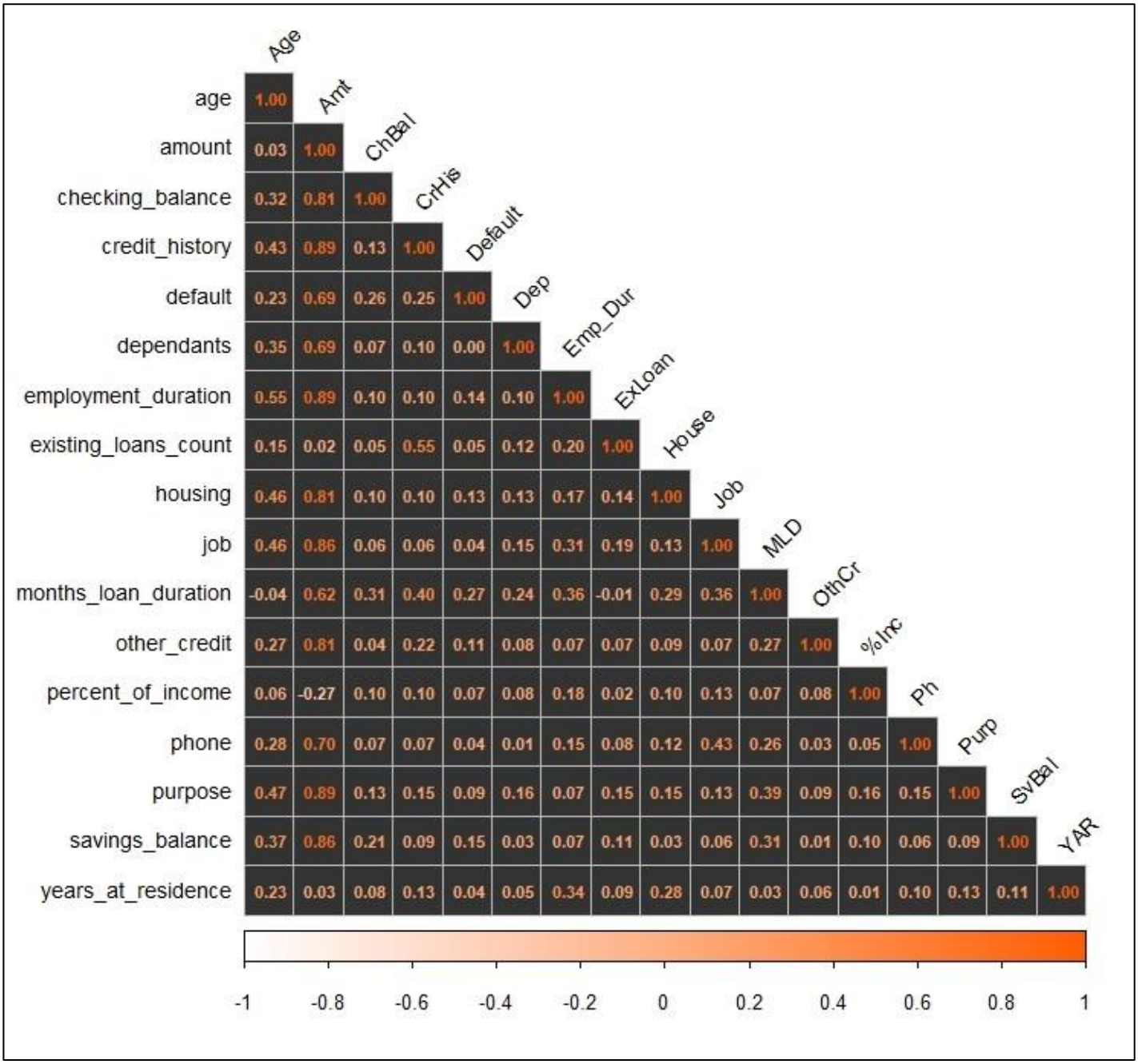
To select explanatory variables for the model build, correlation analysis of the variables was performed.

In the case of categorical variables, the Cramer's V^2 measure was considered as a measure of association between variables.

In the case of a numeric and a categorical variable, the contingency coefficient was considered as the measure of association.

In case of numerical variables, the Pearson's correlation coefficient was calculated.

[2] Cramer's V is a metric used to measure association between categorical variables. It is the mean square contingency coefficient from a χ^2 test.



Correlation Plot between variables

From the plot we see that the variables amount, duration of loan, credit history, and checking balance have the highest correlation with default.

The variables dependents, existing loans, job, percent of income, phone, purpose, and years at residence have correlations that are less than 0.1 with default. These variables will be excluded for model building.

Model Building

Data Preparation:

Before the model building process began, the data set was split into a training and testing dataset using stratified sampling. The training dataset contained 70% (4200 observations) of the values from the overall data and the testing dataset contained 30% (1800 observations) of the values from the overall data.

Logistic Regression:

Logistic regression models were built using a 5% level of significance and insignificant variables were discarded in subsequent iterations. The cut off for Variance Inflation Factors was set at 2 (Variables with VIF $>$ 2 would require revaluation).

Iteration 1: Model was built using all variables as explanatory variables. As seen in correlation analysis, the variables with low correlation (job, dependants, years at residence, and existing loans count) were deemed insignificant and discarded for the next iteration.

Iteration 2: Model was built by discarding insignificant variables. However, on inspection of the Variance Inflation Factors (VIF) of the variables, amount and duration of loan had high values. As a result, the variable duration of loan was discarded in the subsequent iteration (owing to 'amount' having highest correlation with default)

Iteration 3: The third iteration was run using the following explanatory variables,

checking balance, credit history, purpose, amount, savings balance, employment duration, percent of income, age, other credit, housing, phone

Even though the 3rd model had greater AIC than the 2nd, it showed no presence of multicollinearity and no insignificant variables. Hence, we selected this model and ran further tests to confirm the model fit.

Model Testing:

The value for McFadden's R^2 for the model came out to be 0.19 ($\approx$ 0.2) which indicates that the model has a good fit.

Additionally, the Somers' D test value for the model is 0.578 suggesting that the model has good predictive capabilities.

However, the model fails the Hosmer-Lemeshow test with a p-value that is close to zero.

Running a Likelihood ratio test between this model and model from iteration 2 we obtain a p-value that is close to zero. This indicates that model 3 is a better fit.

Model Summary:

```
summary(logistic3)
 Call:
 glm(formula = default ~ . - job - dependants - years_at_residence -
     existing_loans_count - months_loan_duration, family = binomial,
     data = training)

 Deviance Residuals:
     Min       1Q   Median       3Q      Max
 -2.1993  -0.7882  -0.4756   0.8618   2.8113

 Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
 (Intercept)                   -2.245e-01  3.257e-01  -0.689 0.490667
 checking_balance> 200 USD     -1.278e+00  1.066e-01 -11.984  < 2e-16 ***
 checking_balance1 - 200 USD   -3.104e-01  9.272e-02  -3.347 0.000816 ***
 credit_historygood             7.206e-01  9.860e-02   7.307 2.72e-13 ***
 credit_historyperfect          1.531e+00  2.003e-01   7.646 2.07e-14 ***
 credit_historypoor             7.281e-01  1.475e-01   4.934 8.04e-07 ***
 credit_historyvery good        1.603e+00  1.821e-01   8.801  < 2e-16 ***
 purposecar                    -6.748e-03  1.418e-01  -0.048 0.512308
 purposeeducation               4.960e-01  1.940e-01   2.557 0.010547 *
 purposefurniture/appliances   -3.197e-01  1.392e-01  -2.297 0.021643 *
 purposerenovations             3.266e-01  2.644e-01   1.236 0.216621
 amount                         1.740e-04  1.559e-05  11.160  < 2e-16 ***
 savings_balance> 1000 USD     -1.212e+00  1.447e-01  -8.378  < 2e-16 ***
 savings_balance100 - 500 USD  -5.981e-01  1.224e-01  -4.886 1.03e-06 ***
 savings_balance500 - 1000 USD -6.906e-01  1.871e-01  -3.691 0.000223 ***
 employment_duration> 7 years  -5.530e-01  1.284e-01  -4.305 1.67e-05 ***
 employment_duration1 - 4 years -4.677e-01  1.065e-01  -4.393 1.12e-05 ***
 employment_duration4 - 7 years -9.245e-01  1.298e-01  -7.121 1.07e-12 ***
 employment_durationunemployed -9.722e-02  1.728e-01  -0.563 0.573637
 percent_of_income              2.818e-01  3.734e-02   7.548 4.43e-14 ***
 age                           -1.754e-02  4.083e-03  -4.297 1.73e-05 ***
 other_creditnone              -3.990e-01  1.109e-01  -3.599 0.000320 ***
 other_creditstore              5.601e-03  1.885e-01   0.030 0.506291
 housingown                    -5.112e-01  1.290e-01  -3.964 7.38e-05 ***
 housingrent                   -7.085e-02  1.529e-01  -0.463 0.643174
 phoneyes                      -4.136e-01  8.550e-02  -4.838 1.31e-06 ***
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 (Dispersion parameter for binomial family taken to be 1)

 Null deviance: 5113.9 on 4189  degrees of freedom
 Residual deviance: 4179.9 on 4164  degrees of freedom
 AIC: 4231.9

 Number of Fisher Scoring iterations: 5
```

Model Calibration and Determination of Metrics:

An iterative approach was used to determine the threshold. After running over 1000 iterations for each value from 0.1 to 0.8, a threshold of 0.57 was deemed to give out the highest accuracy of results for our credit model.

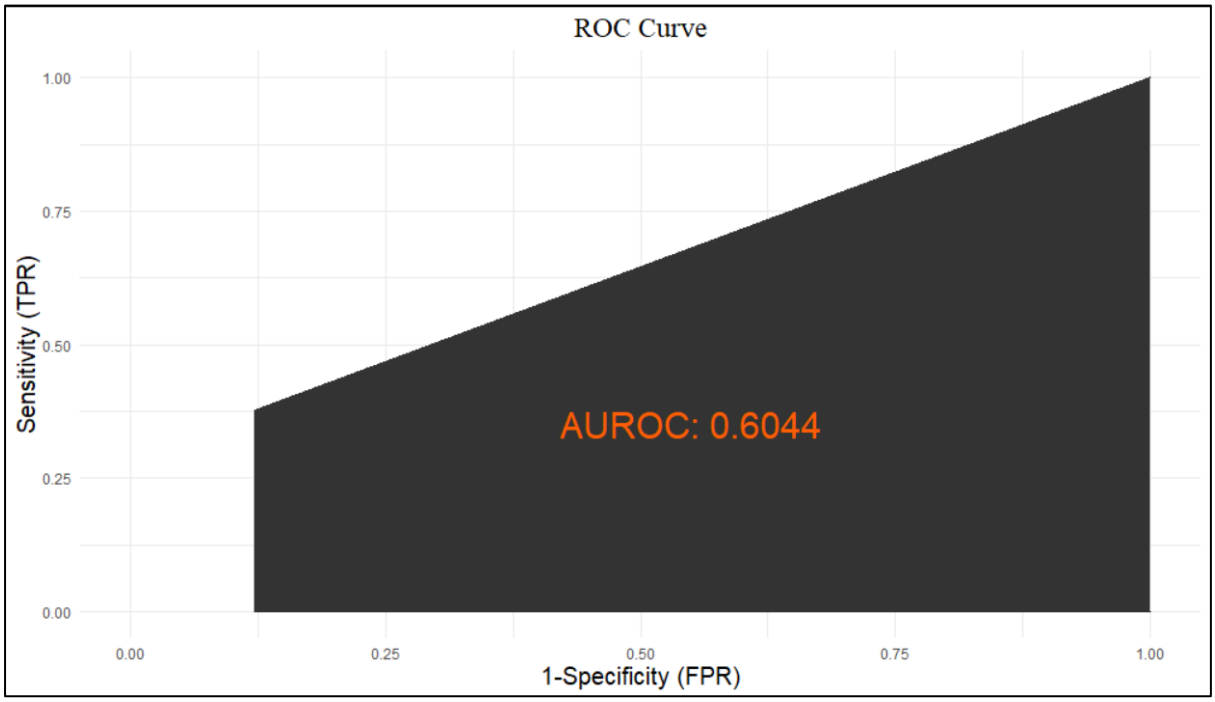
In-sample and Out-sample Analysis:

Metrics	In-sample	Out-sample
Accuracy	0.7683	0.7439
Sensitivity	0.7703	0.7556
Specificity	0.7558	0.6667
AUROC	0.6371	0.6058

As shown above, the metrics compute for both in-sample and out-sample predictions are similar and comparable, hence we conclude that there is no over-fitting or under-fitting in the model.

K-Fold Cross Validation[3]:
K-Fold cross validation was run on the model with the number of folds as 10 (i.e., the data was split into 10 sets or 'folds') and recursively trained on 9 out of the 10 sets and predictions were made on the remaining set.

The CV score obtained was 0.7504762.



Given the shortcomings of the logistic regression model built, other classification algorithms, namely – decision trees, random forest and gradient boosting models were built.

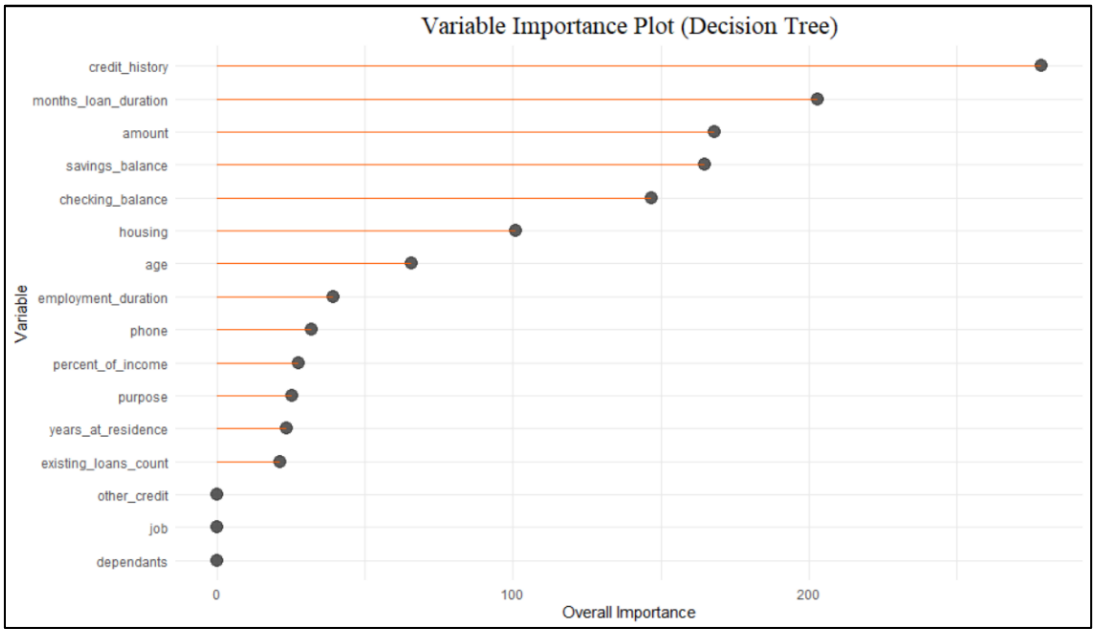
[3] K-fold cross validation is a process of splitting a dataset into 'k' number of 'folds' and training the model on 'k-n' number of sets and testing it on 'n' number of sets. It is a resampling method that determines how well the model performs.

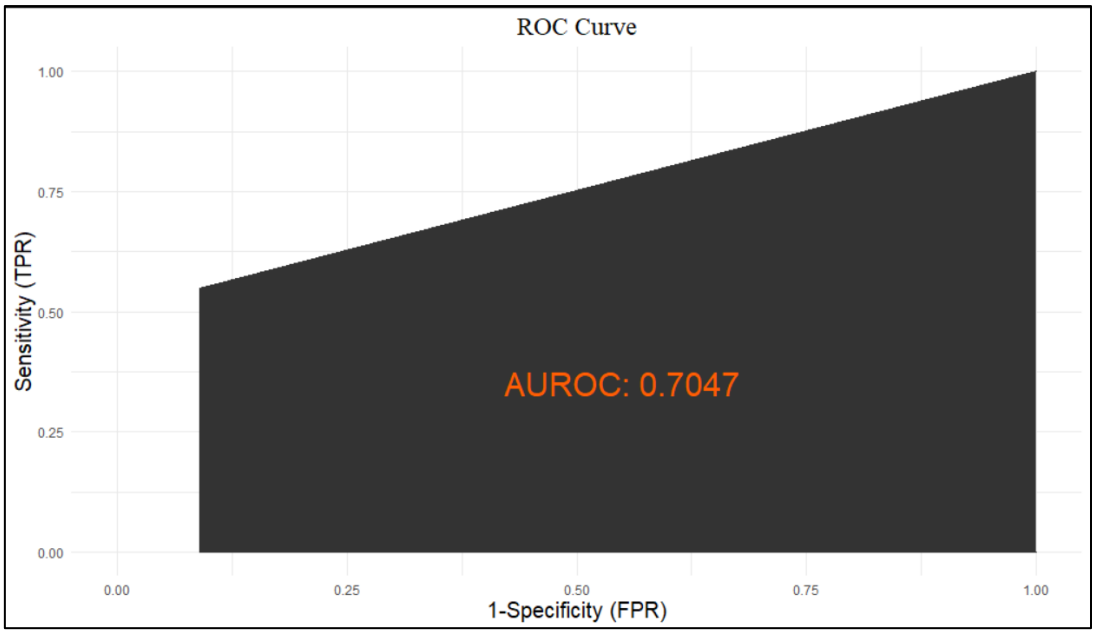
Decision Tree[4]:

A decision tree was created using the complexity parameter as 0.01 and in-sample and out-sample analysis was performed. Here are the metrics calculated:

Metrics	In-sample	Out-sample
Accuracy	0.8179	0.8017
Sensitivity	0.8288	0.8246
Specificity	0.7772	0.7237
AUROC	0.70229	0.7047

We can see that we get a better model with higher accuracy than the logistic regression.





[4] Decision Tree algorithms are classification algorithms that continuously splits data based on certain parameters.

Random Forest[5]:

To further improve on our credit model, a random forest algorithm was chosen. A random forest operates by combining multiple decision trees into a single algorithm.

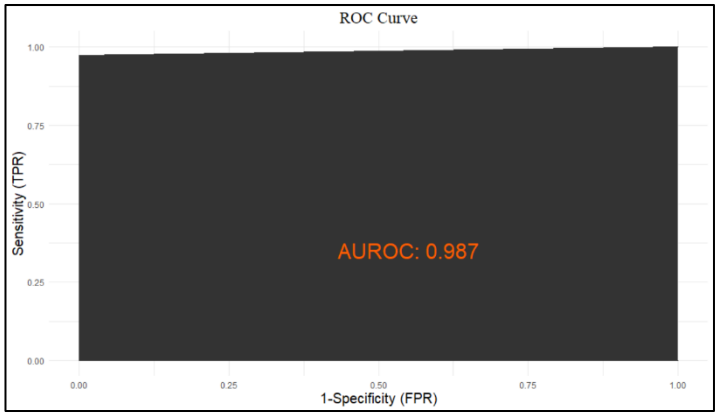
Here are the metrics calculated after creating the model:

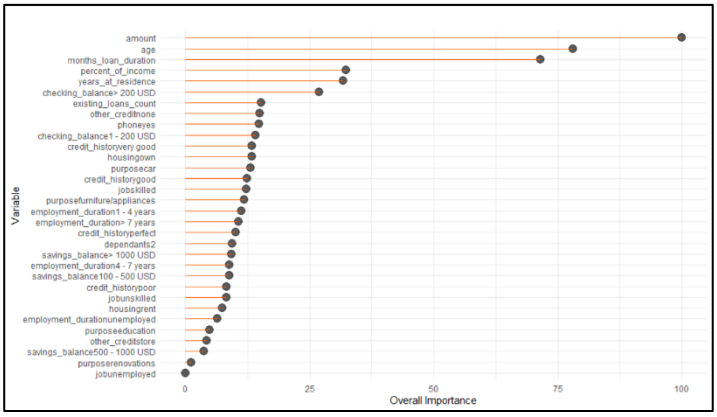
Metrics	In-sample	Out-sample
Accuracy	0.9995	0.9978
Sensitivity	1.0000	1.0000
Specificity	0.9984	0.9926
AUROC	0.9992	0.9963

The results from creating a random forest were extremely promising, as the algorithm offered exceptionally high accuracy in both training and testing scenarios.

On first glance, one might worry about overfitting of the model. However, since the in-sample and out-sample metrics are very close to each other, we have no solid evidence of the presence of overfitting in the model.

Albeit the random forest has incredibly high prediction accuracy, the time taken to build such models is usually on the higher end. To overcome this drawback, we have also considered gradient boosting algorithms.





[5] Random forests are ensemble learning algorithms that work by creating a multitude of decision trees during model training process.

Gradient Boosting:

Gradient boosting algorithms are ensemble learning algorithms that combine 'weak learners' (generally decision trees) to give out one 'strong learner' algorithm. Gradient Boosting trains many models in a gradual, additive, and sequential manner.

The package XGBoost was used to build the gradient boosting algorithm.

Data Preparation:

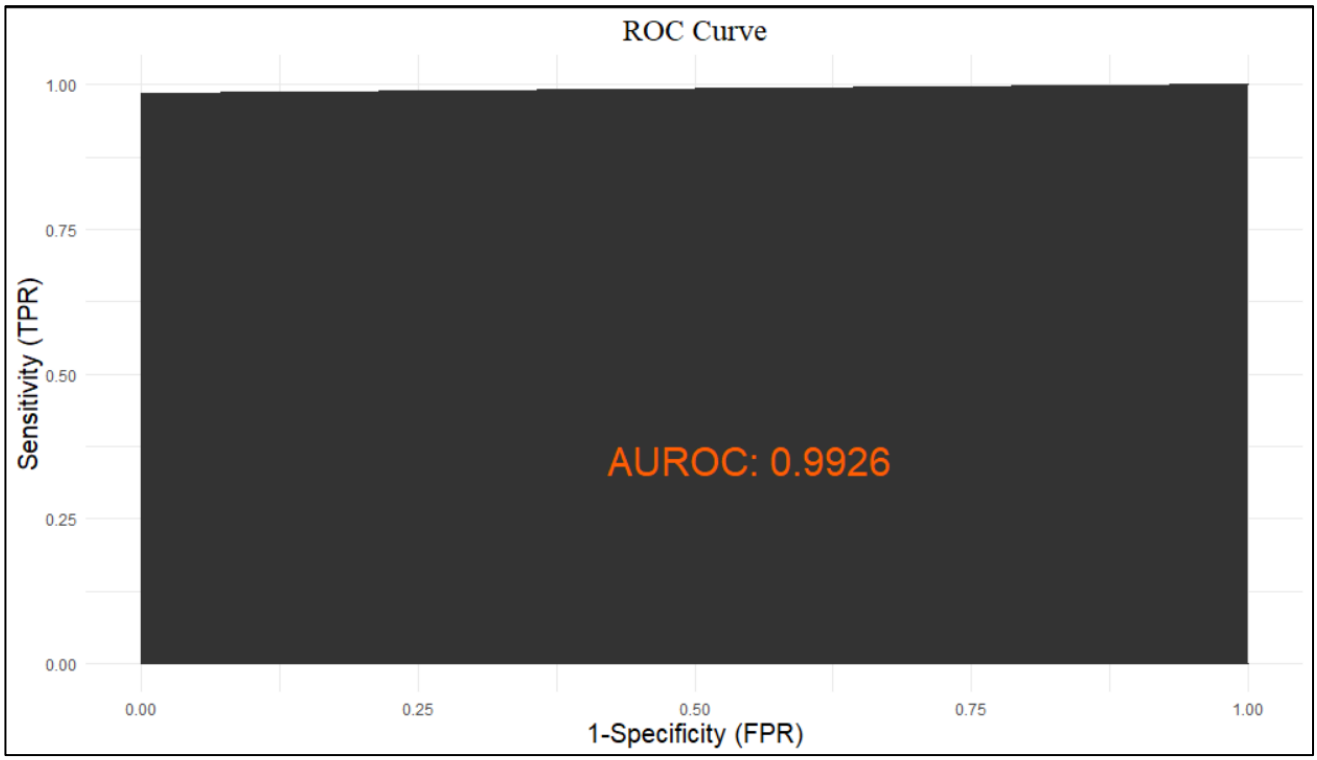
The data had to be transformed into matrix format to run the algorithm. The training as well as testing data was converted to matrix format. This was done by the process of 'one hot encoding'. Model was built and predictions were made accordingly.

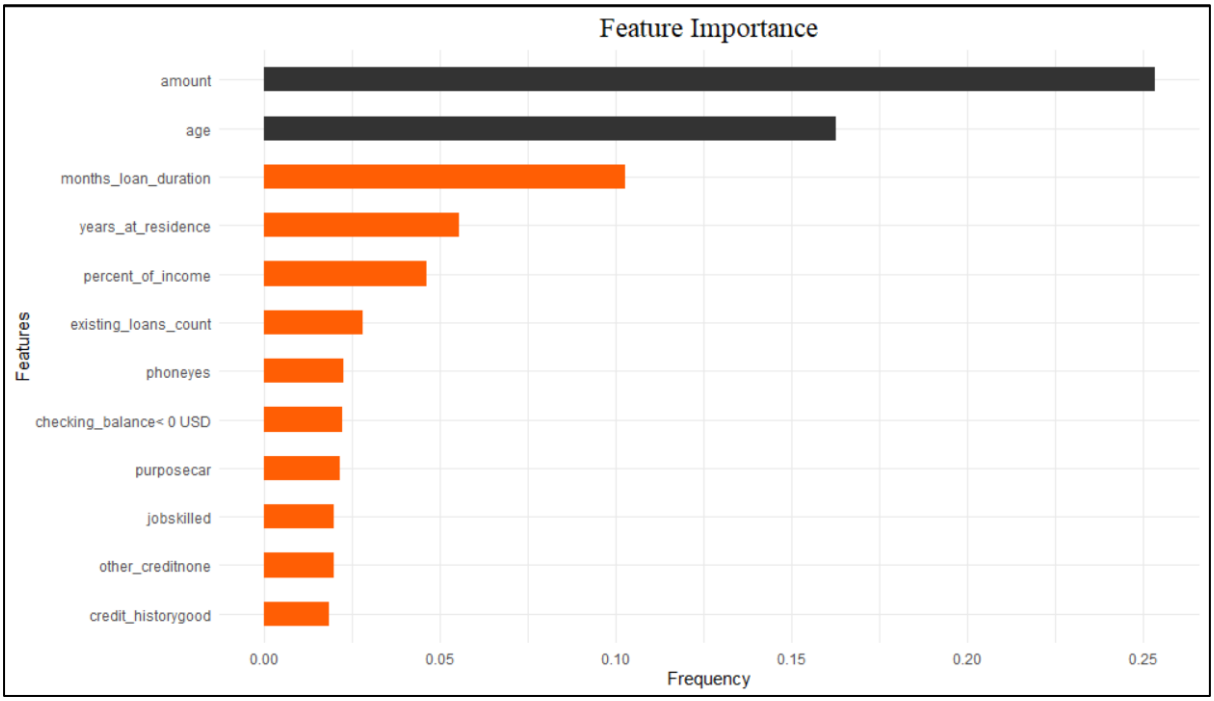
Model Build:

The model was built, and the following metrics were calculated:

Metrics	In-sample	Out-sample
Accuracy	0.999	0.9956
Sensitivity	1.0000	1.0000
Specificity	0.9984	0.9852
AUROC	0.9992	0.9926

The model has excellent predictive capabilities and shows no signs of over or under fitting. The K-fold cross validation score for the model was 0.999524 which was achieved on the 162nd iteration.





As expected, we see that the variables with highest correlation have the greatest importance in the predictive model.

Conclusion

Model Comparison:

Model	Accuracy (Out-sample)	AUROC (Out-sample)
Logistic Regression	0.7439	0.6058
Decision Tree	0.8017	0.7047
Random Forest	0.9978	0.9963
Gradient Boosting	0.9956	0.9926

Considering the data that was collected, we believe that the final, gradient boosting model is the best suited to our use case. This model makes up for the inconsistencies faced with logistic regression and makes up for the shortcomings of the decision trees and random forest models (by being significantly faster).

Additionally, the models that were built after the logistic regression have no extra data requirements and would therefore give better results for the same data.

Uses of Credit Scoring Models:

Our model has been prepared based on data that pertains to individuals. Apart from assisting in the retail banking segment, the model can also easily be extended to commercial banking operations.

Apart from being used to underwrite loans to new obligors, credit scoring models can be used extensively by the loan servicing department of the bank to help them assess the varying repayment capabilities of existing obligors and adjust the cost of loan accordingly. These models can also help greatly in reducing the loan processing time.

It can also be used by the risk management team to assess the credit risk that the bank faces and manage its credit losses to adhere to the bank's policies and regulatory reporting requirements.

Credit scoring models can also help tailor the type of demographic to target for specific loan types. For example, a person with a very good credit history, who has a management position, living on rent may be more interested in a housing loan and have the capability to repay the loan over time.