**MXN600 Assignment 2 - Summary on a Page**

**Summary of scenario**

A global top-rated bank requested to develop a statistical model for credit risk assessment, based on provided historical data of loans, repay statuses and borrowers’ personal information.

**Research task**

1. Explore the dataset and Identify a pool of variables that may potentially affect the loan repay failure.
2. Suggest a generalized linear model to identify potential loan repay failure, based on the information, available at the time of loan application. For legislative purposes, decision-making transparency is a high priority, therefore the model must remain interpretable.

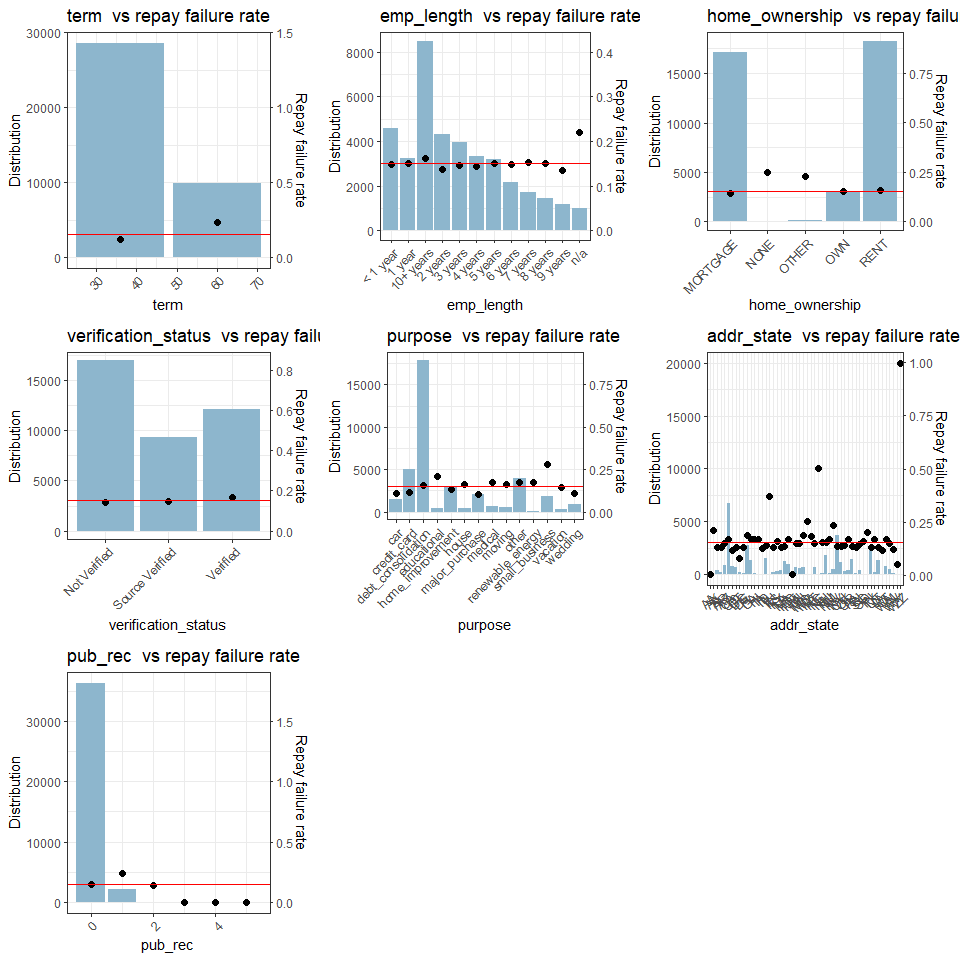
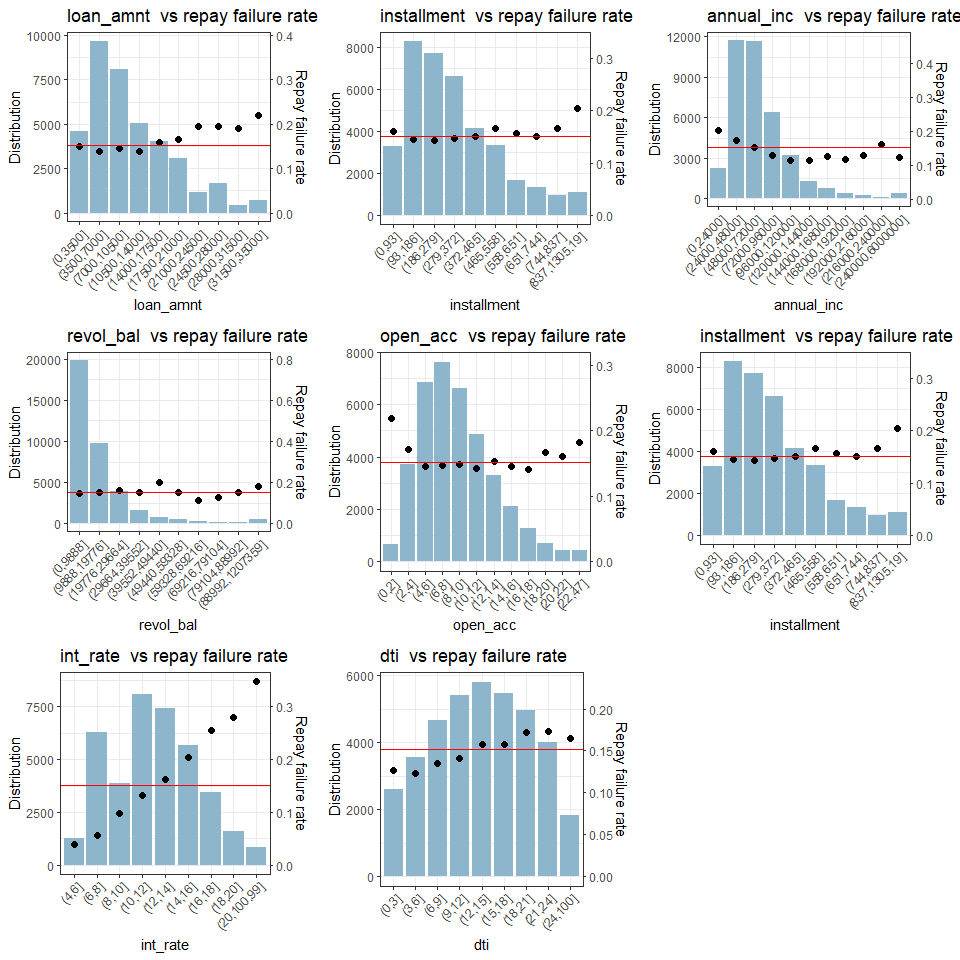
**Summary of available data**

The dataset provided consisted of 38,480 observations. Each observation represented a loan granted by the bank and contained 37 variables grouped into two categories: data provided by the borrower and historical data describing their financial behaviour. The response variable was a binary value representing the status of the loan repayment: 0 for successful payment and 1 for failed repayment. The average repayment failure of all loans in the dataset is 15.1%.

**Data exploration and variables selection**

After the cleansing of the provided dataset, exploratory data analysis was performed, the output of which is shown below. Each diagram represents the density histogram of one independent variable and repayment failure rate for each category or bin (for numeric variables). The red horizontal line reflectsthe average repay failureacrossthe entire dataset**.**

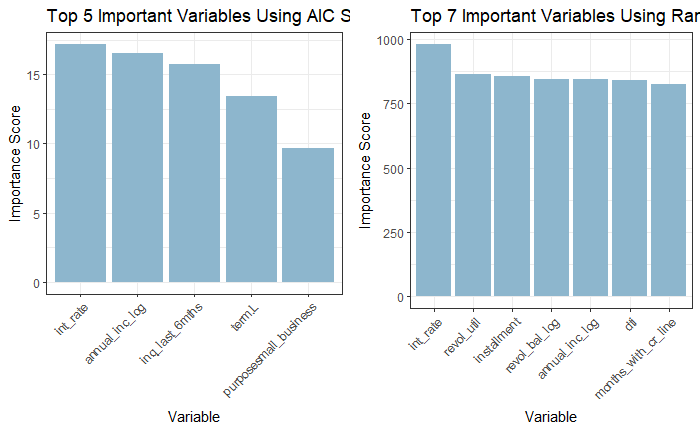
Credit term vs Repay failure Interest rate vs Repay failure DTI vs Repay failure

Interest Rate DTI

AIC step-wise regression Random Forest

Feature selection Feature selection

The visual exploration facilitated a reduction in number of variables for further investigation down to 13. To analyze the significance of the selected variables, two techniques were applied: step-wise regression feature selection and random forest feature selection. The diagrams on the left illustrate the top 5 features based on AIC score and top 7 using RFFS.

Variable Variable

**Modelling approach**

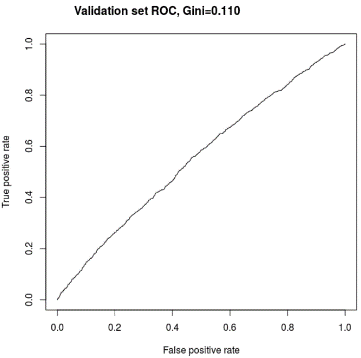
To avoid the model bias, we applied a class balancing technique so the training data consisted of 50% observations with failed repayments and 50% of successful repayments. We compared the performance of three similar generalized linear models, trained on the different sets of variables:

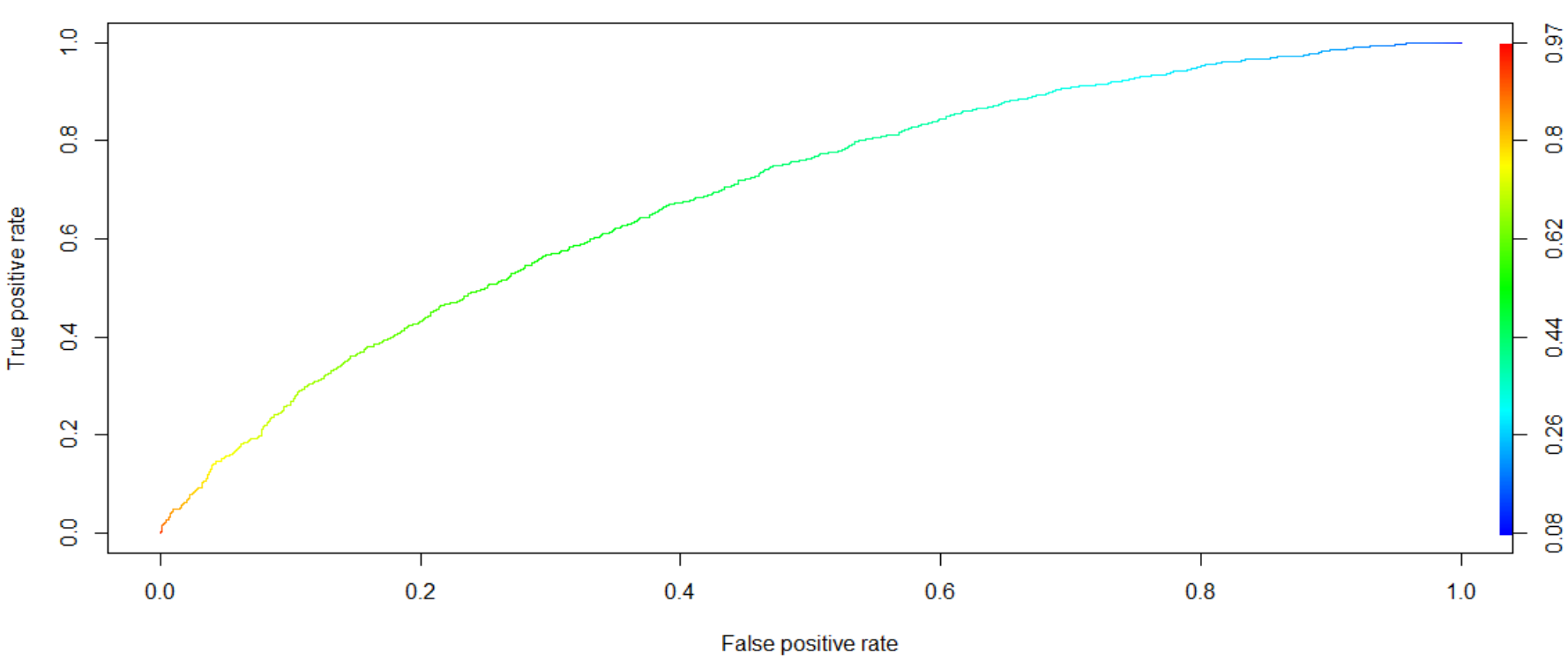
* Derived from stepAIC function *(`int\_rate`, `annual\_inc\_log`, `inq\_last\_6mths`, `term` and `purpose`)*
* Derived from the Random Forest *(`int\_rate`, `annual\_inc\_log`, `revol\_util`, `installment`, `revol\_bal\_log`, `dti` and `months\_with\_cr\_line`)*
* Combination of the remaining variables *(`int\_rate`, `annual\_inc\_log`, `inq\_last\_6mths`, `term`, `purpose`, `revol\_util`, `installment`, `revol\_bal\_log`, `dti` and `months\_with\_cr\_line`)*

Three different link functions were also applied: Logit, Probit and Cloglog.

**Modelling results and evaluations**

The models were evaluated with Gini score. Each demonstrated similar performance, however the combined model performs slightly better: 0.389 combined against 0.382 and 0.3330 for StepAIC and Random Forest derived respectively. Tuning the models by utilising different link functions did not reveal an increase in model performance. The model’s goodness-of-fit was evaluated by investigating the distribution of the residuals which did not reveal any significant deviation. The best model demonstrated 64% accuracy while accurately capturing 61.5% of repay failures on the balanced test dataset, which is reasonable considering a lack of positive (repayment failures) observations in the dataset and the limitations of the model, i.e. no interactions and no polynomial dependencies.

The below ROC curve shows the performance of the model at different prediction thresholds along with Gini scores.



Gini = 0.110

Gini = 0.389

**Answers to research questions and recommendations**

Based on the interpretation of the most successful model, it is recommended to consider the following 10 variables for estimating the risk of borrower's repayment failure: First of all interest rate (showed a strong effect over failure repayment), revolving line utilization rate, annual income (including a transformation over its effects), instalment, number of inquiries during last six months, revolving balance, loan term, DTI rate, loan purpose and number of months since a credit line was opened. The suggested model demonstrated 64% accuracy on detecting the potential repayment failures on the balanced dataset. As a downside of the approach of the balanced dataset, the model shows relatively high false-positive rate. Before implementing the model, it is highly recommended to tune the threshold depending on the institution’s current credit policy.

Further analysis is required to evaluate the effect of interactions among the variables. Moreover, recommendations coming from the business should be considered in future efforts when determining variables to be discarded due to being highly correlated. Finally, the business must consider the level of acceptable risk before considering the final threshold for the model. In other words, the bank must allow their risk policy to inform their judgement on assessing customer suitability; a choice between offering more loans to customers who may be at risk of repayment failure versus offering fewer loans to applicants less likely to default.