SOAP

This report is based on the management requirement of creating a new credit risk model and comparing it with the old model using historical lending data to address the following queries:

1- How does your new model perform compared to the one you used previously? How can it be expected to perform on new loan applications?
2- What are the important variables in this model and how do they compare to variables that are traditionally important for predicting credit risk in the banking sector?
3- Can accounting for this variation (e.g., state/zip-code and time) improve performance benchmarks?
4- Are there any surprising differences in variables that are important for predicting credit risk? This is again essential for regulations.
5- Does credit risk change over time or between states? This is not something the bank has previously investigated, and results may inform modified loan policies in the future.

The data to address the first two questions was split into training and validation sets and contained 20 variables. The exploratory data analysis led to the removal of some variables that were highly correlated to other variables. After exploring, we fitted a binomial GLM model using logit, probit, and cloglog link functions. The optimal set of variables was determined based on AIC using backward and forward selection. The three models were then compared based on their AIC, BIC, and log-likelihood. The model with the logit function was the best-fitting model. A final comparison of the full model and the logit model also revealed that the logit model was the best fit for the data.

The obtained results are illustrated by the graph below:

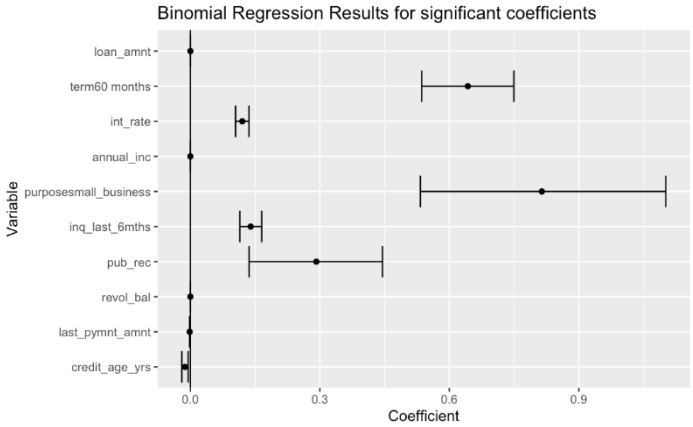

Figure 1. This visualisation shows the value of the significant regression coefficients with their confidence interval

- According to the insights obtained through the ROC curve, the Gini score and cross-validation, we can confirm that the new model performs significantly better than the previous one. The true positive rate is much higher for training and testing data, with the Gini score closer to 1. Moreover, cross-validation on the test data provides an accuracy of 86%. Hence, the bank can trust the performance of this model as a reliable source of benefits.

- The important variables of this model in terms of statistical significance of their coefficients are the ones illustrated by the above graph. Our research showcases that the traditionally important variables for predicting credit risk in the banking sector are related to the borrower's ability to service debt, including the purpose of the credit, total debt, and borrowers' financial strength. The variables purposesmall_business and pub_rec are in alignment with the variables previously mentioned. Nevertheless, loan_amnt and annual_inc have very limited explanatory power in our model, which is surprising as they reflect total debt and financial strength, respectively. Therefore, our model also has some differences in terms of important variables when compared to the ones traditionally considered.

A good example of this is the fact that according to the model, the number of payments in which the loan is paid back is significantly more important than loan amounts or annual income as a predictor.

To justify whether the variation in trends that may exist over time or between different jurisdictions can affect credit risk, we built another Generalised Linear Mixed Effect model and used the new variables (issue_d, zip_code, addr_state, earliest_cr_line) as random effects. The same covariates were selected based on the previous feature selection. The key difference was that we added the new variables as random effects to the new model. The model's performance was compared with the previous model by ROC curve and Gini score.

According to our results, there is little indication that time and jurisdiction can significantly influence credit risk. The variance of the random effects is close to 0, which means the new variables can only help us to explain small variability in the model. In terms of p-value and coefficients, all the important variables stayed almost the same. The only difference is that now the variable revol_util became highly significant for credit risk prediction. The Gini score of the new model is slightly higher than the previous one, indicating that the inclusion of time and jurisdiction-related variables improves the performance of the model to some extent.

	Training Gini score	Testing Gini score
Binomial model	0.712	0.703
Binomial Mixed Effects model	0.73	0.716

Table 2: This table illustrates the comparison of the two models by Gini score.

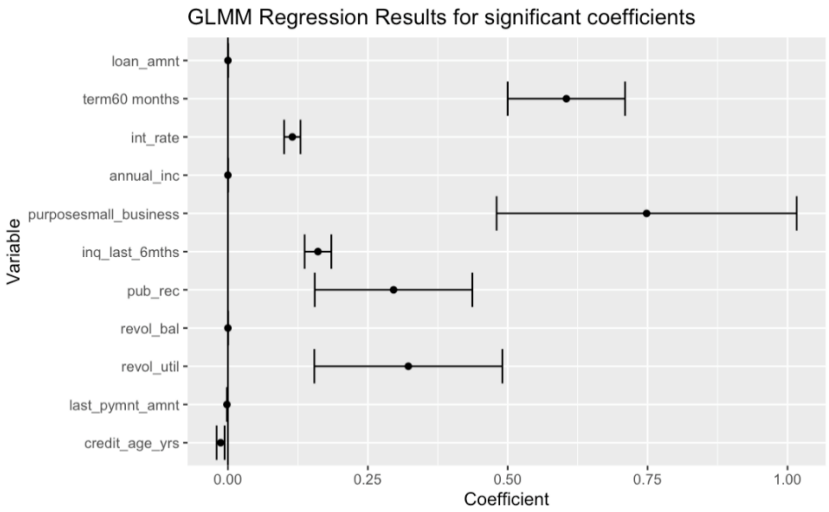

Figure 2: This visualisation illustrates the value of the significant regression coefficients with their confidence interval.

- Fitting the GLMM shows that accounting for trends that may exist over time or between different jurisdictions does slightly increase the performance of the binomial model we fit in the first section in terms of true positive rates and Gini Score.
- The key difference between the two models is that the variable revol_util becomes significant and shows a strong positive relation with repayment failure. Intuitively, it makes sense that the larger the amount of credit the borrower is using relative to all available revolving credit will decrease their ability to repay. What is striking is that it still holds true for this model that predictors like loan amount or annual income do not seem to have strong effects in terms of credit risk prediction. Indeed, these have been widely considered traditionally.
- According to the results of this analysis credit risk does not change much over time or between states. Hence, the bank does not need to worry about particular locations or times of the year in order to obtain consistent revenues from loans.