# **Loan Default Rebuild SOAP (G3)**

Jesse Travaini, Ben Msambya & Nicholas Gecks

2023-06-04

Recently our peer-to-peer lending start-up has been acquired by ‘Apollo’ a regional Australian bank. This acquisition has prompted an investigation into our company’s credit risk model. Results of this investigation has concluded that a complete “ground-up’ rebuild is necessary. Thus, our team’s goal in this report is building a statistical model to predict loan default based on information known at the time of application. Regarding this model management have several concerns.

**1.0 Management Concerns**

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?

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, between your model with/without location and time information?

5. Does credit risk change over time or between states

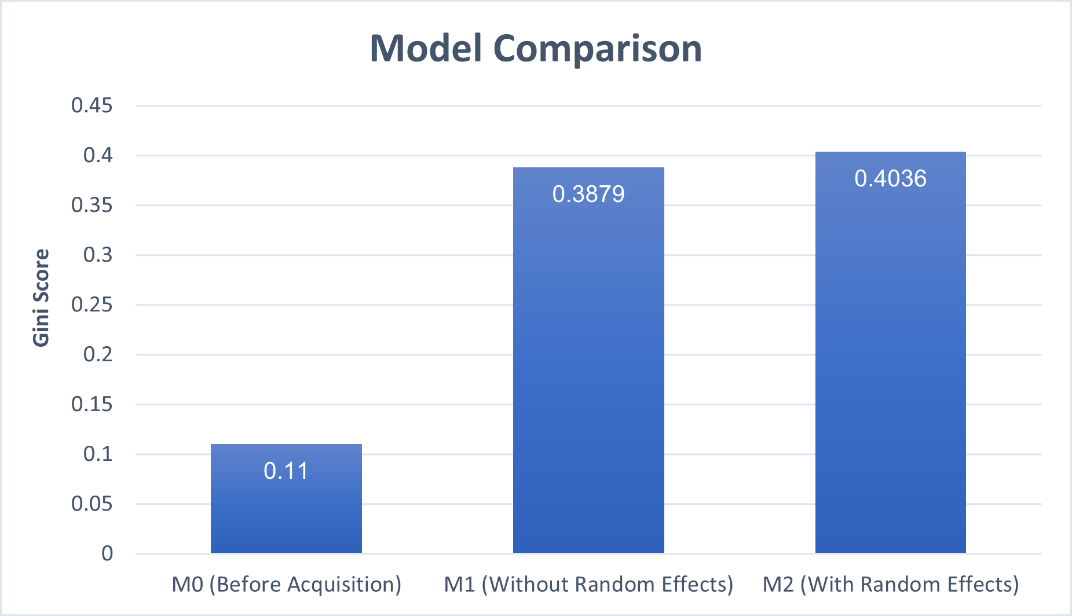
**2.0 The Data**

The standard data featured 20 different variables for each loan given with the corresponding outcome (whether the loan defaulted or not). The extended data was similar but had an extra 4 variables pertaining to the time and location of the loan. Though however, we explicitly stated our model would be based on the knowledge at the time of application and removed many variables. The variables that remain are as follows:

|  |  |  |
| --- | --- | --- |
| Applicable Variables (Standard Data) | | Extra Variables (Extended Data) |
| loan amount | annual income | Issue Date |
| term | verification status | Zip Code |
| interest rate | purpose | Address State |
| employment length | home ownership | Earliest Credit Line |

**3.0 Results**

Initial models were produced using domain knowledge and exploratory plots. These were then compared to our previous model prior to the acquisition. After assessing the new models, exploration into using an algorithmic selection approach to find important variables. This yielded better results which can be seen 1 below. During exploration discrepancies could be seen between groups in terms of both time and location. This lead to the decision to try a mixed effects model to account for these. Comparing to our previous model and final fixed effects model we can see an all-round improvement when accounting for these discrepancies. The comparison of our models can be found below.

**Figure 1: Gini Score of chosen models.**

The previous model (M0) had poor performance. The new fixed effects model (M1) has a much higher performance score than the M0 model. The M2 model has a slightly higher performance than that of the M1 model

**A picture containing text, screenshot, number, font

Description automatically generated**

**4.0 Conclusions**

1. With both our fixed effects and mixed effects models we can see significant improvement compared to our previous model.

2. Using our model we deem “List them off” important for the assessment of credit risk at the time of application.

3. A small improvement is evident when accounting for time and location-based factors.

4. When fitting the fixed effects model, we found that verification status was no longer an important factor in calculating credit risk.

5. Evidence suggests that credit changes between both time and location.

**4.0 Limitations and Recommendations**

Since our models are built off data from United States it may no translate 1-1 to the Australian market. Especially regarding state level variations. We would recommend validating our fixed effects model with Australian consumer data to verify that it is indeed applicable.