Credit Risk Modelling: Summary

Findings:

- The new credit risk model performs significantly better than the previous, improving the Gini benchmark from 0.11 to 0.32.
- Loan default predictors exclude historically important ones, only including:

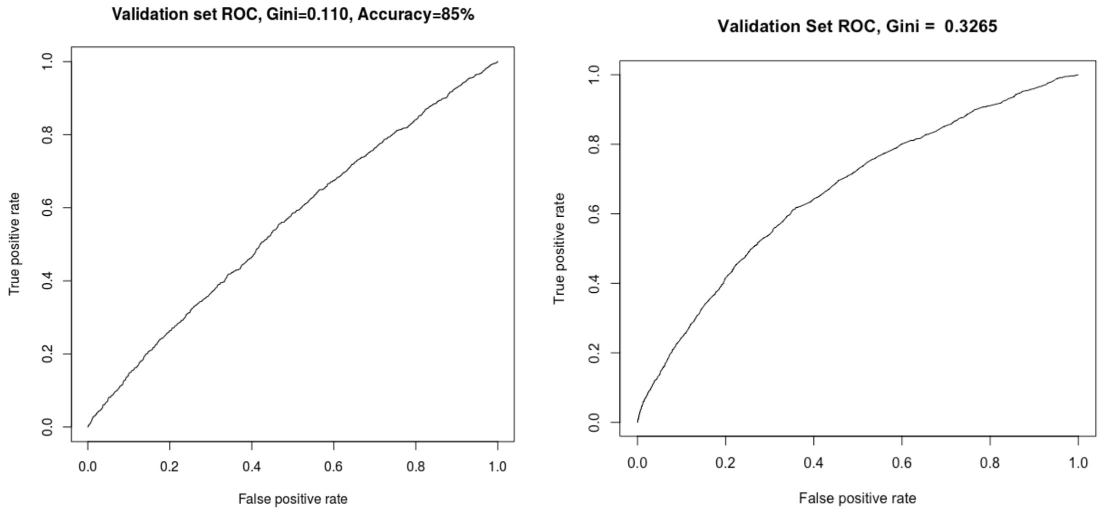
Loan Default Predictors	Effect on Probability of Default
Loan term	Longer loan = increase
Number of inquiries in last six months	More inquiries = increase
Loan Purpose	Car/wedding/major purchase = neutral All other purposes = increase
Annual income	Larger income = decrease

Context:
The risk models used previously were too ad-hoc to be suitable for a bank environment with strict regulatory requirements – therefore a 'ground-up' rebuild was required using relevant statistical measures.

Method:
Over 38000 observations of historical lending data from 2007 to 2011 were used as the basis to investigate which loan characteristics influence default. With the goal of creating a model to predict the probability that a new customer would default, variables were only considered if they were available at the time of application. By statistically analysing the data, highly correlated variables, missing information and outliers were addressed, and the number of potential variables to describe default was decreased to 14.

Using the 14 potential variables to describe default, a generalised linear binomial model was fit within a training set containing 75% of the data. Specifically using the logistic regression, variable combinations were tested and assessed based on the statistical measures of the Gini coefficient and difference of deviance test. A model was chosen that contained the four terms (shown above) and was further confirmed using cross validation and by predicting the 25% of data left unseen. The effects of the variables on the probability of default were investigated.

Conclusions:
1. How does this model perform compared to the previous? How can it be expected to perform on new loans?
Models are compared based on their Gini and ROC plots – where a larger Gini and more curved plot describes a more accurate model. With a previous Gini of approximately 0.11, the old model has a small amount of predictive power and large amounts of variation in defaults were left unexplained. In the new, four-term model, the Gini was significantly increased to approximately 0.32, and the ROC plot reflected a much more prominent curve. This occurrence proves the new model explains more variation in defaults, and the combination of the loan characteristics included are used to provide estimated of "good" and "bad" loans that are largely more accurate than the given benchmark.



Previous model on left, new model on right

2. What are the important variables in this model and how do they compare to variables the bank has found to be traditionally important in its own modelling?

The model chosen contained four characteristics of the given loan, including:

- Loan Term (36 months, 60 months)
- Loan purpose (Credit card, debt consolidation, major purchase, small business, wedding, medical, car, home improvement, vacation, educational, house, other)
- Annual income ($) {(1900,30000), (30000, 37000), (37000, 44400), (44400, 50000), (50000, 58600), (58600,65300), (65300,75300), (75300,90000), (90000,116000), (116000,6000000)}
- Number of inquiries the applicant has made in the last six months.

While there are seemingly important characteristics such as loan amount or debt to income ratio absent from this model, it is emphasised that the statistical methodology applied in this context formally concluded that the inclusion of any more terms did not increase the predictive ability of the model or explain any more variation in defaults – therefore no benefit was gained by including them.

The effects of loan term and the number of inquiries in the last six months were calculated with very narrow confidence intervals, slightly less so for the effect of annual income. The consequences of a loan's purpose were less exacting, and could be recalculated with a narrower estimation range if the fourteen purposes could be grouped into a smaller number of similar purposes by an expert.

3. What assurances and justifications can you make about the statistical rigor of your model and modelling methodology?

The statistical rigor of the chosen model is justified through the methodology itself of using the Gini and difference of deviance test to remove unnecessary variables. Furthermore, the accuracy was proven through cross validation and predicting the unseen data. The model chosen was found to be statistically simple, yet was not under fit, and the predictive power was extremely good in relation to other models and the previous benchmark.

Recommendations:

As requested, a medium was found between creating a model that had strong statistical merit, but also remained simple and interpretable for a variety of expected users within a banking environment. As a result, further statistical techniques are recommended to be applied in the future to enhance accuracy and knowledge of future defaults. It is also relevant to note that the human factor associated with loans holds a certain level of uncertainty, therefore it is encouraged that the model be used as a guide, and the possibility of unpredictable actions may occur for any new applicant.

Please see the attached Credit Risk Modelling report for all statistical reasoning and in-depth conclusions.