Report on Proposed Credit Scoring Model

Introduction

This report outlines the proposed credit scoring model developed for the bank, describing the model itself and the testing involved in producing the model. The model was developed using 20 factors recorded on a sample of 800 bank customers as well as a measure of whether the customer was good at paying back bank credit or not. A further 200 customer sample were then used to test the effectiveness of this model. Below is a list of the recorded attributes:

1. Status of existing current account
2. Duration of the credit/loan (months)
3. Credit history
4. Purpose
5. Credit amount
6. Savings account/bonds
7. Time at current employment
8. Instalment rate in percentage of disposable income
9. Personal status and gender
10. Other debtors/guarantors
11. Amount of time in present residence (years)
12. Property type
13. Age (years)
14. Other instalment plans
15. Housing
16. Number of existing credits as this bank
17. Job
18. Number of people being liable to provide maintenance for
19. Telephone
20. Whether the individual is a foreign worker

The significance of each of the above attributes was assessed and a model was developed allowing for a probability of whether a given customer will be bad at paying back credit to be found. Below the proposed model is presented as well as a discussion on the observed relationships from the sample between each attribute and whether the credit outcome was good or bad. As being considered to have a good or bad credit score is binary, a logistical regression model is used.

Cleaning and Exploration of Data

Some categories were combined due to appearing very infrequently in the sample. In this case, any category with less than 15 occurrences out of the 1000 customers were combined, as the model is unable to accurately measure their effect on an individual's credit score due to this infrequency. The only situation where this occurred was in "Purpose", where "Retraining", which only applied to 9 customers in the sample, and "Domestic Appliances", which only applied to 12, were combined into "Others". Furthermore, the categories of "Personal Status" were not consistent between males and females, so the categories "Male: divorced/separated" and "Male: married/widowed" were combined into "Male: married/other", and the category "Female: divorced/separated/married" was changed to "Female: married/other" to match the male equivalent category.

In logistic regression, when two or more factors are highly connected, the inclusion of all these factors can lead to bias in the model produced with logistic regression, reducing the effectiveness of the model. For example, including a student's age and understanding of maths could lead to this bias, since as age increases, the student will have had more maths lessons and thus be better at maths. Correlation is used to measure how connected factors are. In this model, numeric factors with a correlation coefficient greater than 0.9 and categorical factors with a Cramér's V greater than 0.9 (two methods for measuring the correlation of numeric and categorical data respectively) were considered too highly correlated for both factors to be used in the

model without bias occurring. However, in this model, no factors numeric or categorical had a sufficiently high correlation to lead to potential bias in the model. Thus, no factors were removed due to correlation.

Data Split

The data was split into 800 records used to build the model and 200 used to test the effectiveness of the model. This allows for a large amount of data to be used to create an accurate model for creating credit scores while leaving a suitable amount of data to verify the effectiveness of the model as described in "model test". The data was randomised before being split, as if the given data was ordered by some category (e.g. by age) and the data was not split randomly, the model developed would have been bias resulting in a less effective model being built. The 800 customers were selected using stratified sampling to ensure the same proportion of customers with good and bad credit scores appeared in both the training and test data. This is to ensure the model is built with consideration for effectively scoring both good and bad customers.

The Model

A generalised linear model was built, with the first 9 attributes described above as well as attributes 12, 14 and 15 found to be significant. The proposed model is shown below (with values rounded to two decimal places aside from "credit amount" due to the coefficient being so low), and an example prediction is made to show how the model is used. In the model, a positive coefficient indicates an attribute or factor that increases the likelihood of receiving a bad credit score while a negative coefficient indicates a reduced likelihood of receiving a bad credit score (for example, X2 has a positive score, indicating that the longer the duration of credit/loan, the worse the customer will be at paying it back).

$$\ln(Y/(1-Y) = 0.59 - 0.30(X1i) - 0.89(X1ii) - 1.45(X1iii) + 0.03(X2) - 0.04(X3i) - 0.51(X3ii) - 0.44(X3iii) - 1.29(X3iv) - 1.78(X4i) - 0.92(X4ii) - 0.70(X4iii) - 0.79(X4iv) - 0.46(X4v) + 0.18(X4vi) - 0.76(X4vii) + 0.000107(X5) - 0.28(X6i) - 0.56(X6ii) - 1.07(X6iii) - 1.15(X6iv) - 0.25(X7i) - 0.16(X7ii) - 0.83(X7iii) - 0.42(X7iv) + 0.31(X8) - 0.16 (X9i) - 0.53(X9ii) + 0.21(X12i) + 0.21(X12ii) + 0.84(X12iii) + 0.20(X14i) - 0.58(X14ii) -0.58(X15i) – 0.75(X15ii)$$

Where:

Y	Probability of good credit score
X1i	1 if Status of existing current account is "€0-€500", 0 otherwise
X1ii	1 if Status of existing current account is ">€500/ salary assignments for at least 1 year", 0 otherwise
X1iii	1 if Status of existing current account is "NoCheckingAccount", 0 otherwise
X2	Duration of credit/loan in months
X3i	1 if Credit history is "all credits at this bank paid back duly till now", 0 otherwise
X3ii	1 if Credit history is "existing credits paid back duly till now", 0 otherwise
X3iii	1 if Credit history is "delay in paying off in the past", 0 otherwise
X3iv	1 if Credit history is "critical account/ other credits existing (not at this bank)", 0 otherwise
X4i	1 if Purpose is "car (used)", 0 otherwise
X4ii	1 if Purpose is "other", 0 otherwise
X4iii	1 if Purpose is "Furniture/Equipment", 0 otherwise
X4iv	1 if Purpose is "Radio/Television", 0 otherwise
X4v	1 if Purpose is "Repairs", 0 otherwise
X4vi	1 if Purpose is "Education", 0 otherwise
X4vii	1 if Purpose is "Business", 0 otherwise
X5	Credit amount
X6i	1 if Savings account/bonds is "€100-€500", 0 otherwise
X6ii	1 if Savings account/bonds is "€500-€1000", 0 otherwise
X6iii	1 if Savings account/bonds is ">€1000", 0 otherwise
X6iv	1 if Savings account/bonds is "unknown/ no savings account", 0 otherwise
X7i	1 if Present employment length is "<1 year", 0 otherwise
X7ii	1 if Present employment length is "1-4 years", 0 otherwise
X7iii	1 if Present employment length is "4-7 years", 0 otherwise
X7iv	1 if Present employment length is ">7 years", 0 otherwise

X8	Instalment rate in percentage of disposable income
X9i	1 if Personal status and sex is “Female: married/other”, 0 otherwise
X9ii	1 if Personal status and sex is “male: single”, 0 otherwise
X12i	1 if Property is “building society savings agreement/ life insurance”, 0 otherwise
X12ii	1 if Property is “car or other, not in attribute 6”, 0 otherwise
X12iii	1 if Property is “unknown / no property”, 0 otherwise
X14i	1 if Other instalment plans is “stores”, 0 otherwise
X14ii	1 if Other instalment plans is “none”, 0 otherwise
X15i	1 if Housing is “own”, 0 otherwise
X15ii	1 if Housing is “for free”, 0 otherwise

*Since none of the 1000 bank customers had a purpose of “vacation” or personal status and sex of “single female”, these factors cannot be accounted for in the model. Thus, the significance of these factors is not known, so results from this model for customers with either of these values may be less accurate.
**For categories, some factors do not have a corresponding coefficient. This is because one item of each category is accounted for when the other options equal 0 (for example, a customer with “bank” as their other instalment plan would be represented by X14i = 0 and X14ii = 0).

Example

A single male (X9ii=1) customer with a current account status of €400 (X1i=1), a loan duration of 20 months (X2=20), all credits at the bank paid back (X3i=1), taking a loan to purchase a new car (X4i->X4vii = 0), a credit amount of €2500 (X5=2500), €650 in savings account/bonds (X6ii=1), working in current place of work for 3 years (X7ii=1) with an instalment rate of 2.5% of disposable income (X8=2.5), unknown property (X12iii=1), no other instalment plans (X14ii=1), with their own housing (X15ii=1)and 1 existing credit at this bank (X16=1) would result in the following:

$$\ln(Y/(1-Y)) = 0.59-0.30+0.03(20)-0.04+0.000107(2500)-0.56-0.16+0.31(2.5)-0.16 + 0.84 + - 0.58 - 0.75$$
$$\ln(Y/(1-Y)) = 0.5225$$
$$Y = 0.63$$

In other words, the probability of this individual having a good credit score is 1-0.63=0.37. This can be interpreted as the individual’s credit score rating by the bank (i.e. a 37% score).

Model Test

A sample of 200 bank customers with known credit outcomes were tested against this model to verify its effectiveness at predicting whether a given customer was good at paying back credit or not. Any customer scoring above 50% was considered to have a bad credit score. When compared to the given classification of these 200 customers, the model gave a different score to 42 (21%) of these customers. Thus, the misclassification rate of this model is 21%. In other words, predicted the credit outcome of the customer with 79% accuracy. However, due to the significantly larger number of positive credit score data (of the 1000 sampled, 700 had a good credit score), the model correctly assigns a good credit score to a customer with a good credit rating with 88% accuracy, and correctly assigns a bad credit score to customers with a bad credit rating with 58% accuracy. Since the bank is likely at more risk when giving good credit scores to customers who are more likely to result in a bad credit outcome, the model’s lower accuracy when it comes to identifying bad customers may be a problem. To build a model that is more effective at recognising customers more likely to result in a bad credit outcome, more data on bad credit customers would be needed.

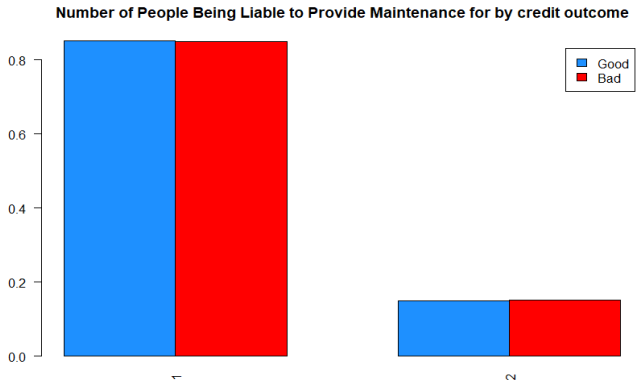
Graphs

The below graphs compare the relationships between each attribute for customers with good and bad credit outcomes for the 800 observations used to create the model. To compare categorical data, bar charts showing the proportion of customers with good and bad credit outcomes for each attribute were constructed. A large difference between the proportion of good and bad credit customers for any given attribute of a factor would indicate said factor to be of interest when attempting to determine a customer’s credit score. For numeric data, box plots for the values of good and bad credit customers were constructed. Again, a large difference

between median values or the distribution of values would indicate that the corresponding attribute differs for good and bad customers when it comes to paying back credit.

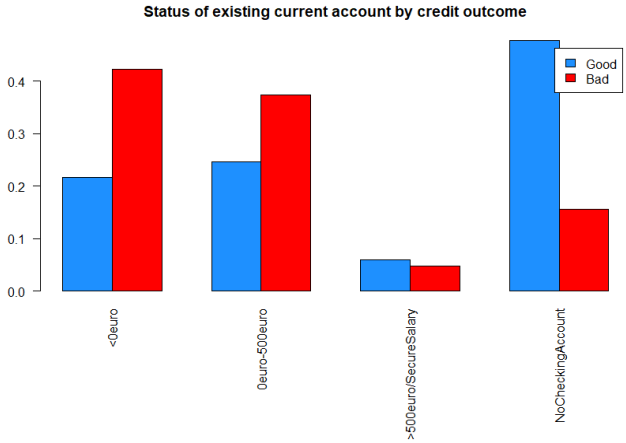
Unused Factors – Bar Chart

Eight of the recorded factors were found to have an insignificant effect on the model. Thus, they have been excluded from the model. Graphs of the unused factors showed little difference between customers with good and bad credit scores, as shown in the example to the left, where the difference between the number of people being liable to provide maintenance for customers with good and bad credit outcomes is almost imperceptible on the graph.



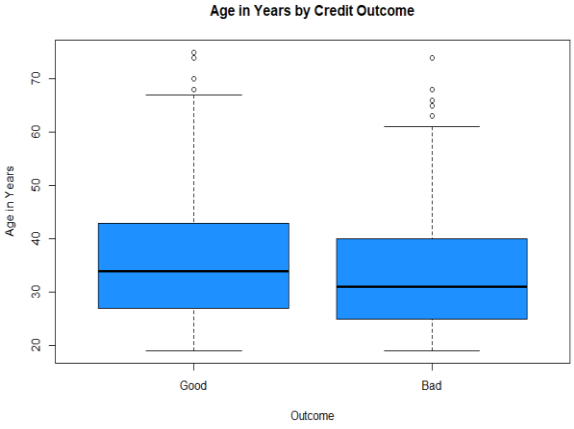
Relevant Factors – Bar Chart

Twelve factors were found to have a significant effect on the model. The bar chart on the right is an example of a relevant factor, and compared to the above graph, the difference between good and bad credit scores are far larger. A relevant difference between customers with good and bad credit outcomes in at least one attribute indicating these factors are relevant in determining customer credit scores.



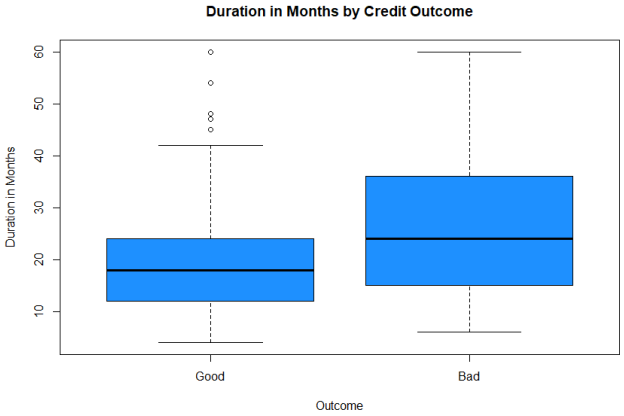
Unused Factors –Box Plots

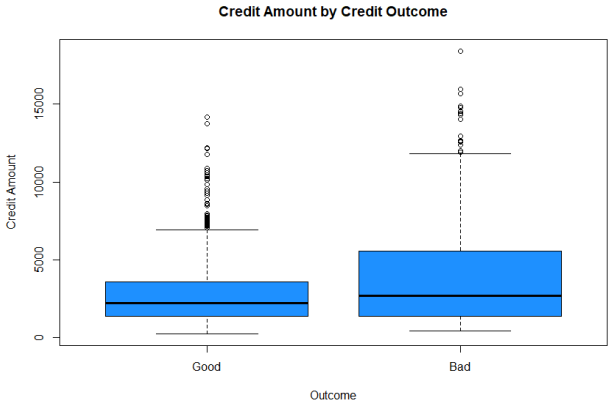
Only one of the four numeric factors recorded, age in years, was not found to be a significant indicator of a customer's credit quality. As with the bar chart, the difference between good and bad customers' ages were relatively small.



Relevant Factors – Box Plots

For the box plots of relevant factors (see below), Credit amount and duration of the credit loan both show clear differences between the values obtained by customers with good and bad credit outcomes. Once again, compared to age in years, the magnitude of this difference becomes much clearer.





Discussion of Factors Relevant to Model

1.Status of existing current account: The model indicates that a higher current account status indicates a higher credit score should be awarded, with no checking account resulting in the best score. The more money held in a current account, the more money the customer has and thus the less likely they are to default on a repayment. Since customers without a checking account are less likely to receive loans, they are thus less likely to default on them.

2.Duration of the credit/loan (months): In the model, a higher duration of credit indicates a greater likelihood of defaulting on repayments, which makes sense as longer loans require more commitment to pay off.

3.Credit history: Here, a poor credit history results in a higher credit score. This is likely due to the bank already being aware of warning signs about the customer's ability to repay credit, making them less likely to receive credit.

4.Purpose: The model's indication that purpose is a relevant factor when determining credit scores likely relates the importance of the purpose to the likelihood of paying a loan back. If the object is of little importance, a customer will likely take the loan less seriously and be more likely to default on it.

5.Credit amount: The model indicates that a higher credit amount results in a greater likelihood of having a poor credit score. Like with 2, larger credit is more difficult to pay off, leading to this assertion.

6.Savings account/bonds: This factor mirrors 1, and the same conclusion can be drawn; customers with greater wealth are less likely to default on repayments, and customers with no savings account/bonds will have a harder time receiving credit in the first place.

7.Time at current employment: Initially, the greater the time at current employer, the better the credit score. However, >7 years is worse than those between 4 and 7 years. This may be due to customers with low employment time having little economic stability, while those staying too long at a job may be experiencing stagnation in their career, leading to less growth in income and a reduced ability to meet credit requirements.

8.Instalment rate in percentage of disposable income: A higher value here indicates a greater likelihood of a poor credit outcome for a customer, likely due to paying higher instalments relative to income meaning that the customer has less income to cover other expenses, and should those expenses be a higher priority, they will end up defaulting on their loan.

9.Personal status and gender: The model indicates single males to receive the highest credit score, as they likely have less obligations than someone who is married. Married/other females also received better credit ratings from the model compared to married/other males, however assuming why may be considered sexism.

12.Property type: Individuals with higher value property, such as real estate, were considered more likely to have a good credit score by the model. This is likely due to greater stability compared to those owning a car or no property at all.

14. Other instalment plans: Customers with stores as other instalment plans receive a lower credit rating than those with bank, however those with none received a higher rating. Those with other instalment plans have other credit obligations to meet, likely causing them to be less reliable at paying credit.

15. Housing: Customers housed for free received higher credit scores than those with their own homes, while those paying rent received a lower credit score. Customers paying rent have monthly obligations to repay, potentially leading to lower credit outcomes, while those housed for free have fewer repayment obligations.

Conclusion

The above model can be used as a measure of a customer's credit score based on the 13 required attributes. The model predicted the credit outcomes of the 200 additional observations with 79% accuracy, however the model is more accurate when it comes to correctly predicting good credit quality customers than bad ones, which may be an issue for the bank. As shown by the above graphs, most factors deemed significant by the model showed more variety in results between customers with good and bad credit outcomes, acting as further evidence that the attributes used in the model differ between good and bad customers in relation to credit. The Central Credit Register, operated by the Central Bank of Ireland, have collected and centralised Irish loan information, and lenders can request credit reports from them on customers they are offering loans to. A worthwhile consideration for the bank may be to compare these reports to the output of this model for a more complete view of a customer's credit status. More information can be found here.