ST3002 – CREDIT SCORING MODEL

INTRODUCTION

The purpose of this report is to explore the attributes that describe whether an individual customer is a good or bad credit risk. Based on the applicant's profile, the bank has to decide whether to make the loan approval or not. If the applicant is a good credit risk, not approving the loan results in a loss of business to the bank. If the applicant is a bad credit risk, approving the loan results in a financial loss for the bank. To minimize loss from the bank's perspective, and to maximize profitability - the bank needs an accurate credit scoring model to determine who to give a loan to.

The dataset contained 1000 customers who had borrowed money from banks. There are 20 attributes describing each customer who had borrowed money, and the "Outcome" variable describing whether a customer was defaulted or not.

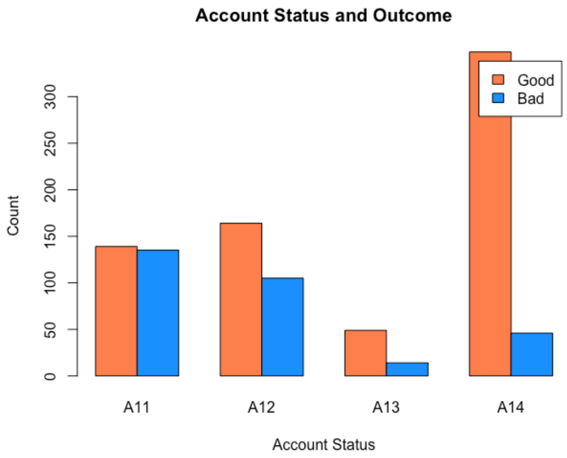
The 20 attributes are described below:

A1: Account Status	A2: Credit Duration	A3: Credit History	A4: Purpose	A5: Credit Amount
A6: Savings accounts/ bonds	A7: Employment duration	A8: Instalment rate	A9: Sex & Marital Status	A10: Debtors/Guarantors
A11: Residency length	A12: Property	A13: Age	A14: Installment Plans	A15: Housing
A16: Existing credits	A17: Job	A18: Dependants	A19: Telephone	A20: foreign worker

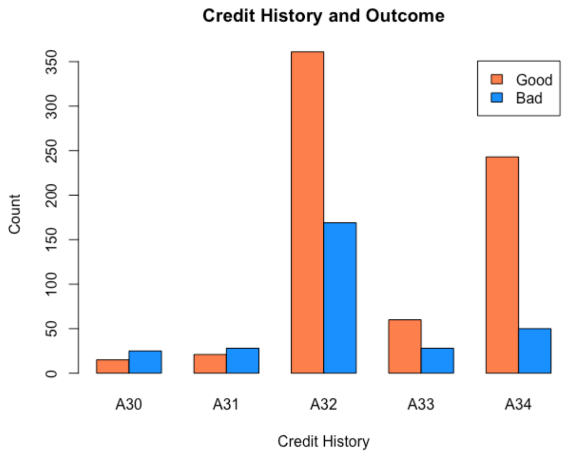
EXPLORING THE DATA

The data consisted of continuous, numerical and categorical data. Using the R package we analysed the initial data and plotted various visualizations. Below, examples of data visualizations can be seen which were created from the dataset along with interesting observations.

CATEGORICAL DATA

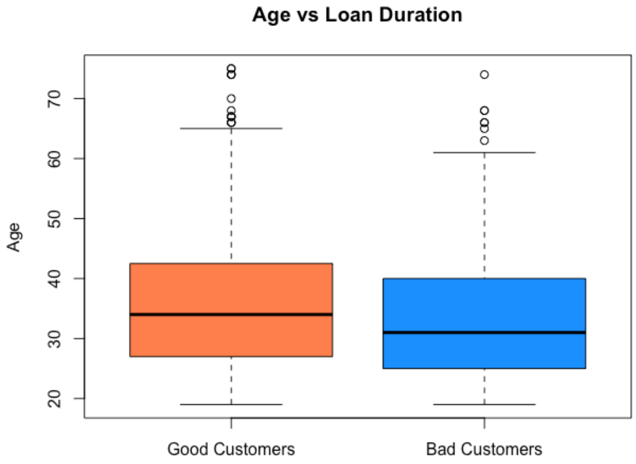


The bar chart shows current account status plotted against the Outcome. It can be seen that A11: < €100 Has the highest number of bad credit holders and A14: No checking account has the lowest number of bad credit holders. This came as a surprise as intuitively one would assume a person with no current checking account would not have a credit history, or a good one.



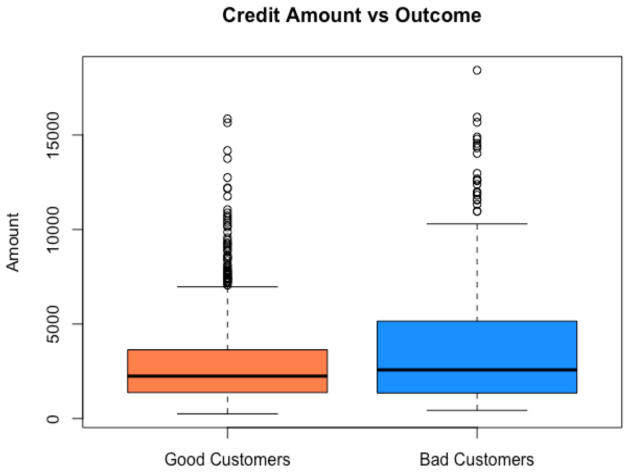
Here, we have credit history plotted against their outcome. In the sub-categories A32: existing credits paid back till now, and A34: critical amount – we can see the record of overall credit holders are higher.

CONTINUOUS DATA



This boxplot shows the age range within customers who have good credit vs bad credit. The median for the bad customers is lesser than the good customers, but it may be premature to say younger people tend to have bad credit records.

NUMERICAL DATA



Lastly, as a visualization for numerical data, we have the credit amount borrowed compared to whether the individual is a good credit customer or a bad one. From the graph it can be observed that customers with a bad credit rating tend to take out a larger amount of credit. Intuitively, this makes sense as many of them underestimate the size of the loan they are taking out and ultimately being unable to repay it to the bank.

DATA CLEANING AND TRANSFORMATION

Before starting work on the data, the number of observations and variables were recorded, and the possibility of missing values. The following figures were recorded.

Questions	Results
Total number of observations	1000
Total number of variables	21
Total missing data	0

After making sure that no missing data was present, the data was then transformed. A number of attributes were grouped together – for reasons including having similar frequencies, similar characteristics, and an effort to reduce the variation within variables to allow for an easier to use credit scoring model. Below it can be seen which attributes were grouped.

Variables	Transformation performed
Account Status	A12 and A13 were grouped to represent positive account balances. A11 represented a negative balance, and A14 representing no checking account.
Credit History	A30, A31, and A32 were grouped to represent no previous credit issues. A33 represents credit issues and A34 represents a critical credit account.
Purpose	A42, A43, A44, A45, A410 were grouped to represent domestics. A46 and A48 were grouped to represent education. A40 represents a new car, A41 represents a used car, A49 represents business.
Savings	A63 and A64 were grouped to represent savings accounts equal to 500 or more. A61 represents savings accounts under 100. A62 represents savings accounts between 100 and 500. A65 represents no savings account.
Present employment	A74 and A75 were grouped to represent employment greater than 4 years. A71 represents unemployment. A72 represents less than a year of employment. A73 represents between 1 and 4 years of employment.
Personal Status and Sex	A91 and A94 were grouped to represent males who are not single. A93 represents single males, A92 represents females.
Debtors and Guarantors	A102 and A103 were grouped to represent guarantors. A101 represents no guarantors,

TRAINING AND TESTING DATA

The cleaned data was split into a 7:3 ratio using stratified sampling. The reason for creating training dataset is to allow for the model to be as accurate as possible. Data is fed into the training dataset allowing the model to learn, and produce complex results. When the model is finalised from the training dataset, it is then passed to the test data set to measure the accuracy of the model. Stratified sampling was used to keep the same ratio of good and bad credit customers from the cleaning dataset, in the training dataset, allowing for greater accuracy. The ratios can be seen below.

	Good credit customer	Bad credit customer
Cleaned	700	300
Training	490	210
Test	210	90

DATA MODELLING

To analyse which attributes contributed to whether a customer would default their loan or not, a regression model was used of Binomial distribution. The initial model was built using the glm function, with the Outcome as the predicted variable and all the other variables as predictor variables. AIC & BIC measures the quality of a model relative to other models. After running the first model, by using the summary function we could analyse the output from R to see which variables were significant and which were not. Telephone, Dependants, Job, Age, residence duration, debtors, sex and status were considered insignificant. However, I included the Job variable in the second glm model as I thought that there was a possibility insight could still be gained.

Running the second refined glm model reduced the AIC & BIC, however, the job variable was still being showed as insignificant – subsequently it was removed from the model. The “Property” variable was also removed as it represented a small fraction of significance compared to the other variables.

Running the refined glm model for a third time decreased the AIC, with all variables being represented as significant.

A step model was the fourth model also ran on the training dataset, and gave the following variables as significant: Account Status, Duration, Credit History, Purpose, Credit Amount, Savings, Employment, Instalment Rate , Other Instalments, Housing, Existing Credits, Telephone, Foreign worker. AIC was reduced fractionally, however BIC increased. Below a summary of each model can be seen:

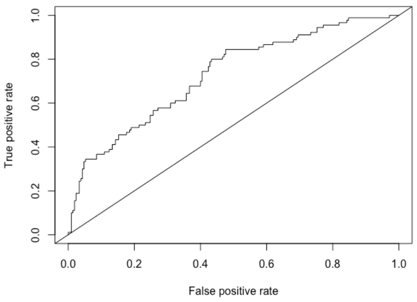
	AIC	BIC
Model 1	685.64	840.37
Model 2	678.19	801.07
Model 3	672.34	772.46
Model 4	672.11	776.78

Model number 3 was picked as it generated the lowest BIC, and compared to Model 4's AIC – the difference is trivial. Below are the variables and attributes that are most significant in our credit scoring model.

Significant variables final model	Coefficients value
Account Status A14	-2.116e+00
Account Status A12	-5.783e-01
Credit Duration	2.697e-02
CreditHistoryA34	-8.573e-01
CreditHistoryA33	-7.936e-01
PurposeA43	-5.022e-01
PurposeA41	-1.512e+00
CreditAmount	1.446e-04
SavingsA65	-7.210e-01
SavingsA63	-6.818e-01
SavingsA62	2.165e-01
EmployementA74	5.200e-01
EmployementA73	9.772e-01
EmployementA72	1.561e+00
InstallmentRate	3.343e-01
OtherInstallmentA143	-9.291e-01
OtherInstallmentA142	-3.401e-01
ExistingCredits	5.008e-01
HousingA153	-3.353e-01
HousingA152	-7.647e-01
ForeignWorkerA202	-2.530e+00

Each of the coefficients associated with each attribute can be used to predict whether a customer will default on a credit Loan or not. For example, for Account status A14(no checking account) – a no checking account status reduces the chances of a default.

MODEL EVALUATION



The plot to the left is the AUC (Area under the curve) – ROC (receiver operating characteristics) curve. It is a highly used evaluation metric when examining a models performance. The AUC-ROC tells you how much the model is capable of distinguishing between classes (ie True positives, false positives etc) in this case – how well our credit scoring model can predict good or bad customers. The closer the AUC number is to one, the better the model. Our model has a 73% chance of predicting a good or bad credit customer.

CONCLUSION

The model we chose performs reasonably well with the data presented. However, it is the banks decision in deciding what probability it is happy with. It's important that the bank has a model that identifies false positives as this presents a considerable loss to the bank. It is also recommended that the bank uses a larger dataset sample to increase the models success in accurately predicting Outcomes, using the significant variables from above as a guide will also promote increase accuracy.