Outline

- Introduction
- Decision Trees
- Logistic Regression
- Discriminant Analysis
- K Nearest Neighbors
- Conclusion

Introduction

About the dataset

The dataset titled “**credit44.csv**” shows data obtained from a German bank containing information about loans granted to its customers. The dataset contains 900 observations, with 21 variables which includes:

- Creditability (categorical variable)
- Account Balance (categorical variable)
- Duration of Credit Month (non-categorical variable)
- Payment Status of Previous Credit (categorical variable)
- Purpose (categorical variable)
- Credit Amount (non-categorical variable)
- Value Savings Stock (categorical variable)
- Length of Current Employment (categorical variable)
- Instalment Per Cent (non-categorical variable)
- Sex…Marital Status (categorical variable)
- Guarantors (categorical variable)
- Duration in Current Address (non-categorical variable)
- Most Valuable Available Asset (categorical variable)
- Age…Years (non-categorical variable)
- Concurrent Credits (categorical variable)
- Type of Apartment (categorical variable)
- No. of Credits at this Bank (non-categorical variable)
- Occupation (categorical variable)
- No. of Dependents (non-categorical variable)
- Telephone (categorical variable)
- Foreign Worker (categorical variable)

Methodology

- Supervised learning is a learning model built to make prediction, given an unforeseen input instance (i.e. unlabelled data

- We will use labelled data, in this case "**credit 44.csv**" that has been classified, to infer a learning algorithm

- The dataset will be used as the basis for predicting the classification of other unlabelled data through the use of algorithms such as decision trees, logistics regression, discriminant analysis and *k*-NN.

Introduction – Aims & Objectives

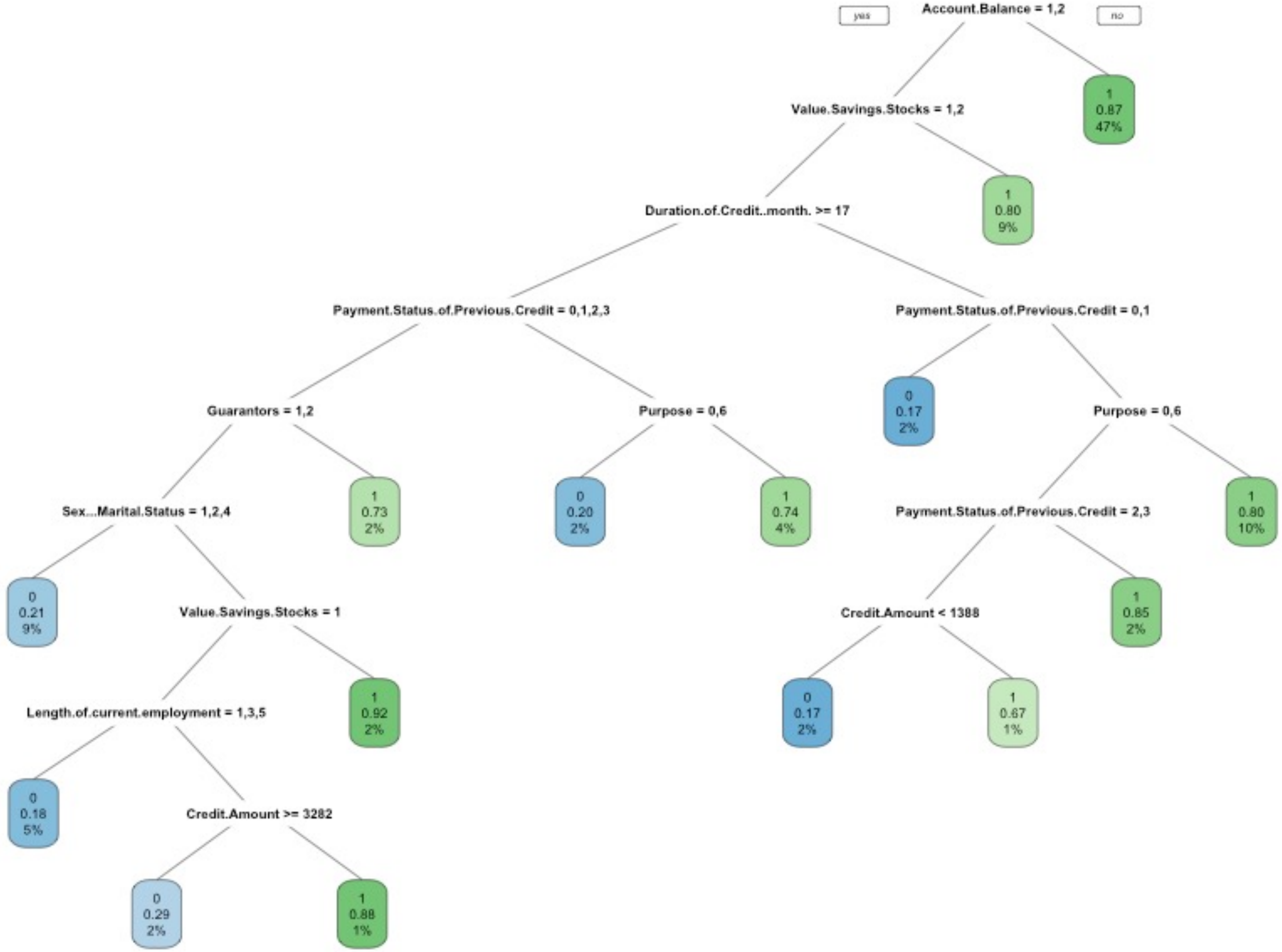
- The aims & objectives of this statistical analysis is to apply supervised learning algorithms such as decision trees, logistic regression, discriminant analysis and k-nearest neighbours to the given dataset to develop predictions models for which customers are likely to pay or default on loans and obtain any other insights lurking within the data.

- Additionally, to determine whether the Discriminant Analysis supervised learning algorithm is a good technique to apply to the given dataset

Decision Tree

- Decision tree is a type of supervised learning algorithm that can be used in both regression and classification problems. Each branch of the tree represents a possible decision, occurrence or reaction. Decision trees are typically drawn upside down, in the sense that the leaves are at the bottom of the tree (root, internal nodes, leaf nodes, branches)
- Given the data set, the problem statement is to predict/determine if a customer will repay a loan or not i.e. whether it is a good loan or a bad loan
- Since the problem statement is a classification problem, i.e. classifying which customers loan are good loans or bad loans, a classification decision tree model using all the independent variables will be built as the classifier generated will be highly interpretable
- To develop the model, the dataset will be split into a training and testing test in the ratio of 70% to 30% to train and test the accuracy of the model respectively.

Decision Tree – Insights/Interpretation


yes
Account.Balance = 1,2
no
Value.Savings.Stocks = 1,2
1
0.87
47%
Duration.of.Credit..month. >= 17
1
0.80
9%
Payment.Status.of.Previous.Credit = 0,1,2,3
Payment.Status.of.Previous.Credit = 0,1
Guarantors = 1,2
Purpose = 0,6
0
0.17
2%
Purpose = 0,6
Sex...Marital.Status = 1,2,4
1
0.73
2%
0
0.20
2%
1
0.74
4%
Payment.Status.of.Previous.Credit = 2,3
1
0.80
10%
0
0.21
9%
Value.Savings.Stocks = 1
Credit.Amount < 1388
1
0.85
2%
Length.of.current.employment = 1,3,5
1
0.92
2%
0
0.17
2%
1
0.67
1%
0
0.18
5%
Credit.Amount >= 3282
0
0.29
2%
1
0.88
1%

Decision Tree – Insights/Interpretation

Insights

- Using the test data, the accuracy of the model is 73.7%
- At a 95% confidence interval, the true accuracy of the model lies between 68% to 79%
- The Mcnemar's test has a p-value < 0.001 which indicates the model is statistically significant
- The relevant independent variables in order of significance for the classification model are:
  - Account Balance
  - Value Savings Stock
  - Duration of Credit Month
  - Payment Status of Previous Credit
  - Guarantors
  - Purpose
  - Sex…Marital Status
  - Length of Current Employment and
  - Credit Amount

Decision Trees – Insights/Interpretation (contd.)

- Account Balance is the most important factor or variable in determining the Creditability of a customer i.e. whether it is a good loan or a bad loan
- For instance, a customer with his/her account balance >= 200 DM or no checking account would be considered to be a bad loan based on the classification decision tree model otherwise other variables or factors amongst others such as the *Value Savings Stock* and the *Duration of Credit Month* would be considered and so on
- Additionally, the Decision Tree model has helped in identifying the variables that are important in determining Creditability from all the other independent variables in the dataset
- It shows "*Credit Amount*" as the least significant variable for the classification as it is the variable that is farthest from the root

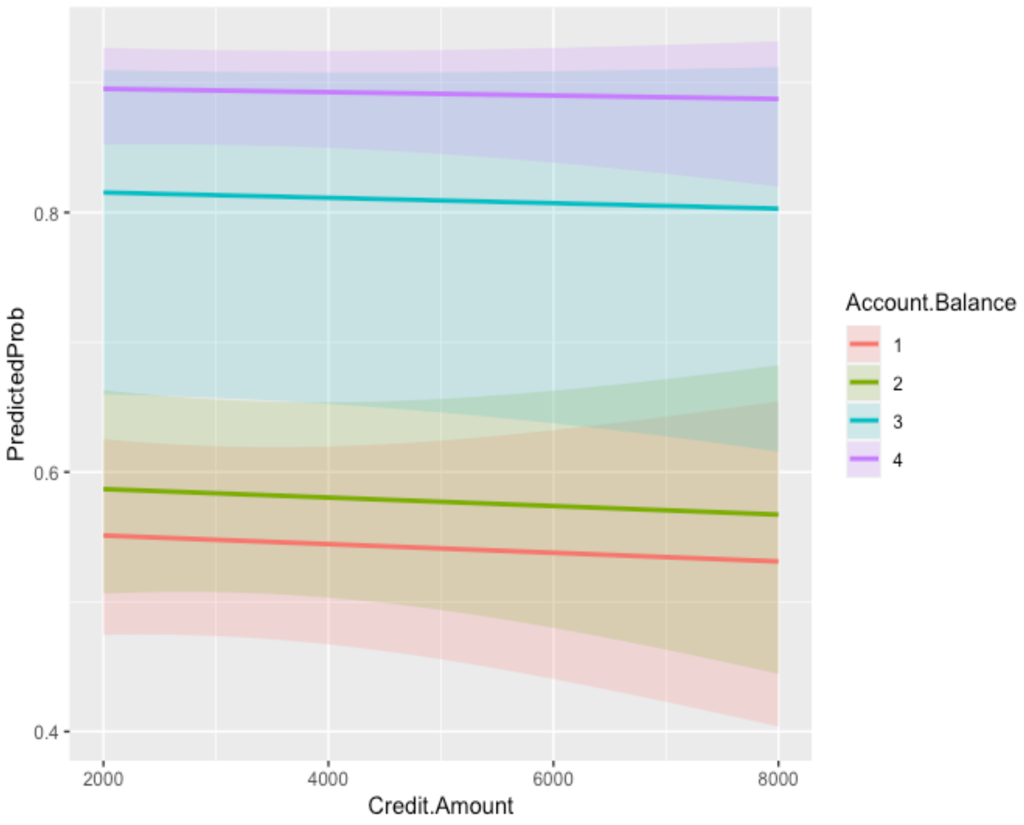
Logistic Regression

- Logistic regression is an extension of multiple regression that allows us to predict categorical outcomes
- Given the dependent variable to predict is (*Creditability*) which is a categorical variable with just two possible outcomes either a good loan or a bad loan a binary logistics regression model will be built using two continuous variable and one categorical variable from the dataset
- Variables of relevance from the decision tree will be used to develop the logistics regression model. The non-categorical variables used are *"Credit.Amount"* and *"Duration.of.Credit..month"* while the categorical variable used is *"Account.Balance"* (they are all considered as the predictor variables)
- Again, to develop a logistics regression model, we split the entire dataset into a training and test set in the ratio of 80% to 20% to train and evaluate the model respectively

Logistic Regression – Results & Evaluation

- Based on the p-values, "*Credit.Amount*" variable is statistically insignificant to the regression model which reaffirms the output from the classification decision tree as it is the variable farthest from the root
- Using the Wald test to test the significance of the categorical variable, it shows a p-value < 0.001 which indicates that the Account Balance variable is statistically significant to the model
- Based on the summary statistics of the logistics regression model, for every one unit increase in Credit Amount, the log odds of Creditability decreases by -1.342e-05, additionally, for every one unit increase in Duration of Credit Month, the log odds of Creditability decreases by -3.690e-02
- The difference between the coefficient of Account Balance 2 and that of Account Balance 3 is statistically significant
- Using the test data, the accuracy of the model is 73.6%.

Logistic Regression – Results & Evaluation



Insights

- The plot shows the predicted probability of a good or bad loan of a customer whose credit amount is in the range of 2000 to 8000 at varying levels of account balance
- The plot indicates that the Credit Amount variable doesn't contribute significantly to the model, unlike the Account Balance
- Using the logistics regression model, the plot shows the varying predicted probabilities for the various levels of Account Balance with level 1 & 2 having a predicted probabilities less than 0.6 while level 3 & 4 having a predicted probability > 0.8
- Therefore, the closer the predicted probabilities is to 1, the more likely it is to be a bad loan similarly, if it is closer to 0 it is a good loan hence, the plot indicates a customer with an account balance 1 or 2 is likely to be a good loan as opposed to a customer who's account balance is 3 or 4 which is likely to be a bad loan.

Discriminant Analysis

- We are interested in a Discriminant Analysis to distinguish which variables in the dataset are best at discriminating between groups. In this case, the groups are either good loans or bad loans so that we can estimate which group any future case belongs to

- We proceed by using non-categorical variables in the dataset such as *"Duration of credit month"*, *"Credit.Amount"*, *"Instalment per cent"*, *"Duration in current address"*, *"Age..years"* and *"No of credits at this bank"*

- Additionally, we give each category equal priors i.e. the probability of a good loan 0.5 and equally a bad loan 0.5 to avoid any bias on either of the groups

- Furthermore, we split the dataset into a training and testing test in the ratio of 80% to 20% to train and test the accuracy of the model respectively.

Discriminant Analysis (contd.)

Training Data	0 (good loans)	1 (bad loans)
0 (good loans)	110	176
1 (bad loans)	104	326

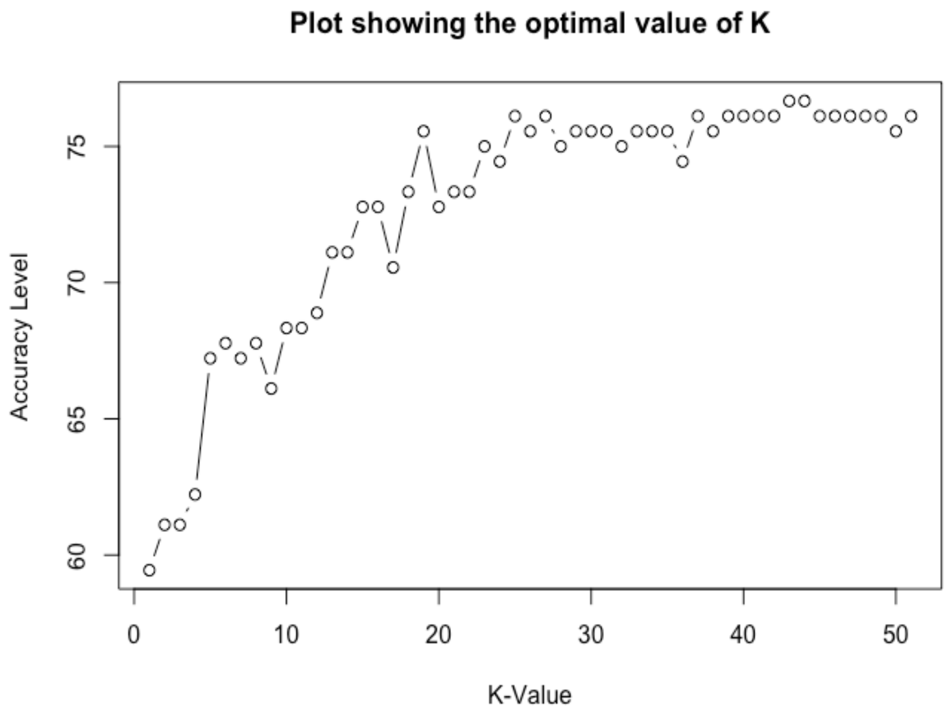
Testing Data	0 (good loans)	1 (bad loans)
0 (good loans)	28	45
1 (bad loans)	25	86

- Based on the training data, the model predicts 436 correctly out of 716 which is about 60.9%. 110 out of 286 good loans were classified correctly while 326 out of 430 bad loans were classified/predicted correctly.
- Similarly, based on the testing data, the model predicts 114 correctly out of 184 which indicates a 61.9% accuracy. 28 out of 73 good loans were classified correctly while 86 out of 111 bad loans were classified correctly
- Since the dataset has no missing observations, the model therefore has no ungrouped or unclassified cases
- As the accuracy of the model drops when compared to that of the Decision Tree and Logistics Regression models, we can conclude that the Discriminant Analysis is not ideal for the given dataset.

K-Nearest Neighbours

- K Nearest Neighbour is a supervised learning algorithm that classifies a new data point into the target class, depending on the features of it's neighbouring data points
- The instance based learning (*k*-NN) is based on the memorization of the dataset thus the classification is obtained by looking into the memorized examples
- Given the dataset, using *k*-NN algorithm, we separate and categorise the outcomes in order to predict the classification of a new point (i.e. whether it is a good loan or a bad loan) to future or potential credit applicants
- We apply the k-NN algorithm to classify the *Creditability* to the same set of non-categorical variables used in the Discriminant Analysis such which are *Duration of Credit Month, Credit Amount, Instalment Percent, Duration in Current Address, Age…Years and No of Credits at this Bank*

K-Nearest Neighbours – Choosing the Optimal Value of k



- The plot shows the value of k and it respective accuracy
- From the plot, the highest accuracy of k ranges from when k is equal to 19 and then only increases or decreases marginally as k increases
- The k-nearest neighbours algorithm uses a very simple approach to perform classification. When tested with a new example, it looks through the training data and finds the k training examples that are closest to the new example. It then assigns the most common class label (among those k training examples) to the test example.
- k is therefore just the number of neighbors "voting" on the test example's class.
- Good practice suggests that the value of k can be selected by taking the square root of the total number of observations in the dataset

K-Nearest Neighbours – Results & Evaluation

- The dataset is split into a training and test in the ratio of 80% to 20% to train and test the algorithm respectively
- The chosen value of the k parameter is 29 (to avoid overfitting to the this particular dataset and an odd number to avoid ties between the two groups we aim to classify)
- Using the test data, the accuracy of the model is 75.6%, from the test data, the algorithm classified 3 out of 4 good loans correctly and 133 out of 146 bad loans correctly
- The McNemar test p-value < 0.001 which indicates that the model is statistically significant

Conclusion

- Using all the independent variables in the dataset, the Decision Tree algorithm helped in identifying the variables that are relevant to the response variable (Creditability) and produced an accuracy of 73.7%. The Classification Decision Tree is highly interpretable, intuitive and easy to understand when making a decision about a customer/applicant creditability

- In the logistics regression model, we selected three variables as produced by the decision tree; two non-categorical variables and one categorical variable. The summary statistics of the model shows the coefficient estimate (log odds) of each variable, it significance and its contribution to the model. Overall, the model is statistically significant and has an accuracy of 73.6%

Conclusion

- Using the Discriminant Analysis algorithm on six non-categorical variables in the dataset and giving equal priors to the two groups we aim to classify, the model has an accuracy of approximately 61% based on the training data while an accuracy of approximately 62% on the testing data (i.e. data the model hasn't seen)

- Similarly, using the *k*-NN algorithm on the same six non-categorical variables and setting the value of k to 29, the model has an accuracy of approximately 76% therefore, we can conclude that the k-NN algorithm is preferred/better compared to the Discriminant Analysis in classifying/predicting a good loans or bad loans (*Creditability*).