Capstone Project Report

Analyzing Credit Risk Using the German Credit Dataset

Name: Chandratej Kurella

Abstract

This report is all about examining credit risk using a dataset known as "German Credit," which we found in the Data Mining Book. We set out to determine if we can tell whether someone is good at handling credit (low risk) or not so good (high risk). Our main aim is to assist banks in making wise decisions that minimize potential financial losses while maximizing their profits.

In this report, we'll explain the methods we used, the things we discovered, and the conclusions we reached by studying the data. To do this, we used visual aids, decision trees, a technique called K-Nearest Neighbors (which is like finding similar people), and logistic regression (a method to predict outcomes).

In simpler terms, this project is about understanding credit risk and helping banks make choices that safeguard their finances and enhance their earnings. We examined a dataset containing information about 1000 individuals and applied various tools to identify who is a good candidate for a loan and who may not be.

Table of Contents

1. Introduction

Dataset Source

The German Credit dataset was obtained from Data Mining Book, providing information about 1000 customers and their credit risk profiles.

Source of the dataset 'German Credit'
https://www.dataminingbook.com/content/datasets-download-1stedition

Project Motivation

The primary motivation behind this study is to help banks be safer and earn more money. We want to figure out if someone is good at managing their money (low risk) or not so good (high risk). Knowing this is really important for banks to make good decisions when giving out loans. In simple words, we're doing this to help banks make sure they don't lose money and that they make more money. We want to be sure that when they say "yes" to a loan, it's to someone who will pay it back, and when they say "no," it's to someone who might not be able to pay it back.

Project Objectives

Our ultimate objective is to make lending decisions for banks more accurate and informed by understanding the most critical factors that affect a person's credit risk.

In this project, we have several main goals:

- **Data Exploration:** We want to understand the German Credit dataset better. We'll do this by creating visual representations to see which factors affect credit risk.
- **Model Development:** We plan to build various models, including K-Nearest Neighbors (KNN), classification trees, and logistic regression. These models will help us determine the key factors that influence credit risk.
- **Accurate Credit Risk Prediction:** We aim to find the best way to predict credit risk accurately. To do this, we'll use different measurements from our models.
- **Ideal Model:** Lastly, we want to suggest the best model based on how well it performs on new, unseen data (validation data) and by comparing different models to see which one works best.

2. Data Description

Data Overview

In our dataset, we have a variety of information related to individual's financial situations. What we're especially interested in is a variable called 'RESPONSE', which categorizes individuals into those with good (1) or bad (0) credit ratings.
Before we could analyze this data, we had to do some data preprocessing work. This included dealing with any missing information to make sure our data is ready for in-depth analysis.

Data Variables

The German Credit dataset contains information about 1000 customers. By using this data, and developing & assessing various models, the bank can make decision on a prospective applicant, whether to go ahead with the loan approval or not.

S.No	Variable Name	Description	Type	Comments
1.	OBS#	Observation No.	Categorical	
2.	CHK_ACCT	Checking account status	Categorical	0 : < 0 DM
				1: 0 < ...< 200 DM
				2 : => 200 DM
				3: no checking account
3.	DURATION	Duration of credit in months	Numerical	
4.	HISTORY	Credit history	Categorical	0: no credits taken
				1: all credits at this bank paid back duly
				2: existing credits paid back duly till now
				3: delay in paying off in the past
				4: critical account
5.	NEW_CAR	Purpose of credit	Binary	car (new) 0: No, 1: Yes
6.	USED_CAR	Purpose of credit	Binary	car (used) 0: No, 1: Yes
7.	FURNITURE	Purpose of credit	Binary	furniture/equipment 0: No, 1: Yes
8.	RADIO/TV	Purpose of credit	Binary	radio/television 0: No, 1: Yes
9.	EDUCATION	Purpose of credit	Binary	education 0: No, 1: Yes
10.	RETRAINING	Purpose of credit	Binary	retraining 0: No, 1: Yes
11.	AMOUNT	Credit amount	Numerical	
12.	SAV_ACCT		Categorical	0 : < 100 DM

		Average balance in savings account		1 : 100<= ... < 500 DM
				2 : 500<= ... < 1000 DM
				3 : =>1000 DM
				4 : unknown/ no savings account
13.	EMPLOYMENT	Present employment since	Categorical	0 : unemployed
				1: < 1 year
				2 : 1 <= ... < 4 years
				3 : 4 <=... < 7 years
				4 : >= 7 years
14.	INSTALL_RATE	Installment rate as % of disposable income	Numerical	
15.	MALE_DIV	Applicant is male and divorced	Binary	0: No, 1: Yes
16.	MALE_SINGLE	Applicant is male and single	Binary	0: No, 1: Yes
17.	MALE_MAR_WID	Applicant is male and married or a widower	Binary	0: No, 1: Yes
18.	CO-APPLICANT	Application has a co-applicant	Binary	0: No, 1: Yes
19.	GUARANTOR	Applicant has a guarantor	Binary	0: No, 1: Yes
20.	PRESENT_RESIDENT	Present resident since - years	Categorical	0: <= 1 year
				1<…<=2 years
				2<…<=3 years
				3:>4years
21.	REAL_ESTATE	Applicant owns real estate	Binary	0: No, 1: Yes
22.	PROP_UNKN_NONE	Applicant owns no property (or unknown)	Binary	0: No, 1: Yes
23.	AGE	Age in years	Numerical	
24.	OTHER_INSTALL	Applicant has other installment plan credit	Binary	0: No, 1: Yes
25.	RENT	Applicant rents	Binary	0: No, 1: Yes
26.	OWN_RES	Applicant owns residence	Binary	0: No, 1: Yes
27.	NUM_CREDITS	Number of existing credits at this bank	Numerical	
28.	JOB	Nature of job	Categorical	0 : unemployed/ unskilled - non-resident

				1 : unskilled - resident
				2 : skilled employee / official
				3 : management/ self-employed/highly qualified employee/ officer
29.	NUM_DEPENDENTS	Number of people for whom liable to provide maintenance	Numerical	
30.	TELEPHONE	Applicant has phone in his or her name	Binary	0: No, 1: Yes
31.	FOREIGN	Foreign worker	Binary	0: No, 1: Yes
32	RESPONSE	Credit rating is good	Binary	0: No, 1: Yes

Data Pre-Processing

Details about few fields from the given dataset,

- OBS column has been dropped, it is Observation Number(its like a serial no.)
- The Fields 'NEW_CAR, USED_CAR, FURNITURE, RADIO/TV, EDUCATION, RETRAINING belong to the purpose of taking loan. In this dataset, we can see them as dummies created for purpose of taking loan.
- The field HISTORY has 4 levels, these are classified based of the past credit history of a person.
- The field SAV_ACCT has 5 levels, these are classified based on the average balance in savings account.
- The field CHK_ACCT has 4 levels, these are classified based on the status of checking account.
- The field EMPLOYMENT has 5 levels, these are classified based on the number of years of employment a person has.
- The field job has 4 levels, these are classified based on the nature of the job.
- Our variable of interest is RESPONSE field, which holds the credit rating of a customer(person) (1-Good Credit Rating, 0-Bad Credit Rating)

There are couple more fields in the dataset where detailed explanation might not be needed and can be understood based on the description of the fields.

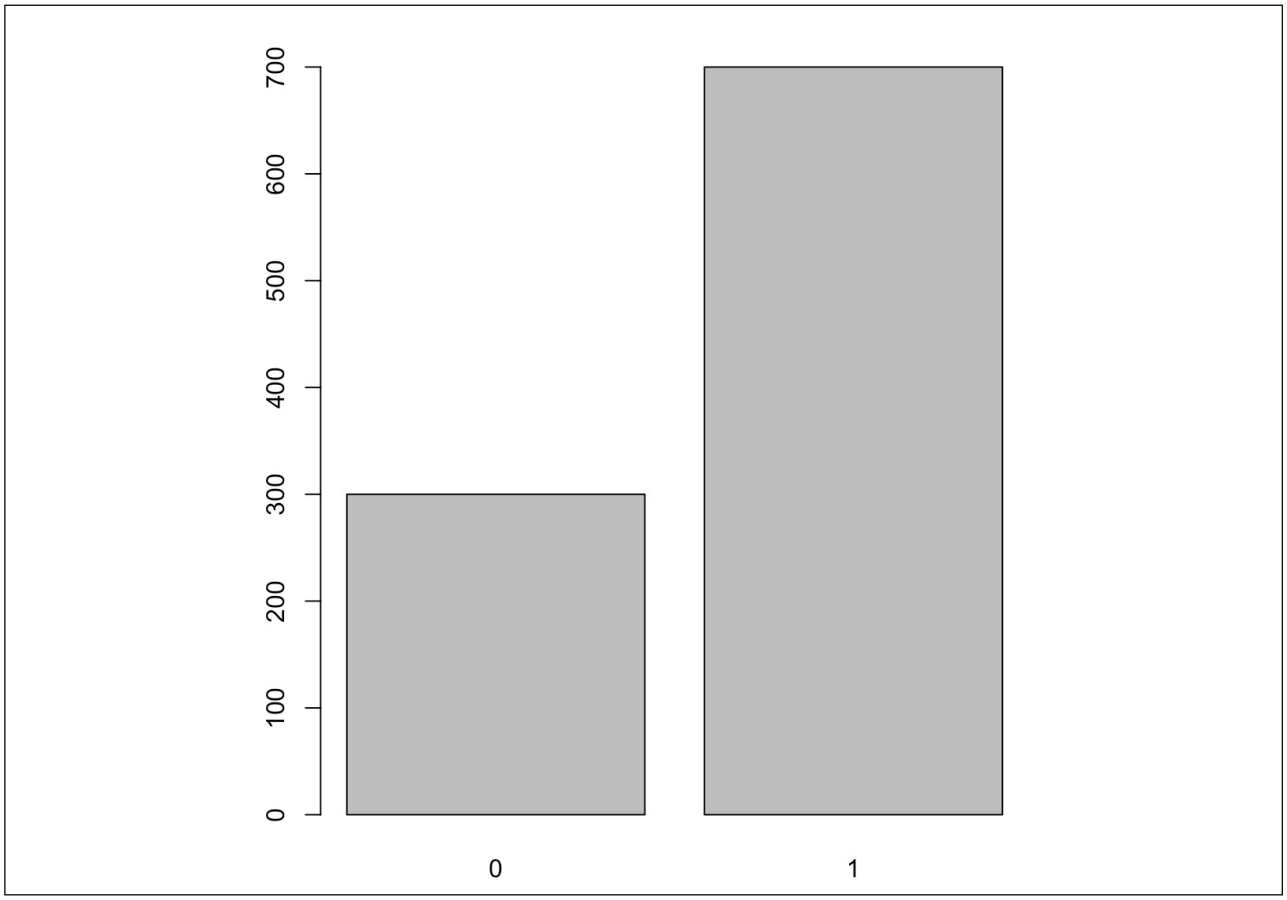
3. Exploratory Data Analysis (EDA)

EDA was conducted to gain insights into the dataset. Key visualizations were created to better understand the relationships between different features and the distribution of credit ratings. The findings from EDA helped in identifying patterns and relationships within the dataset.

Data Visualization

Credit Rating Good (1) vs Bad (0):

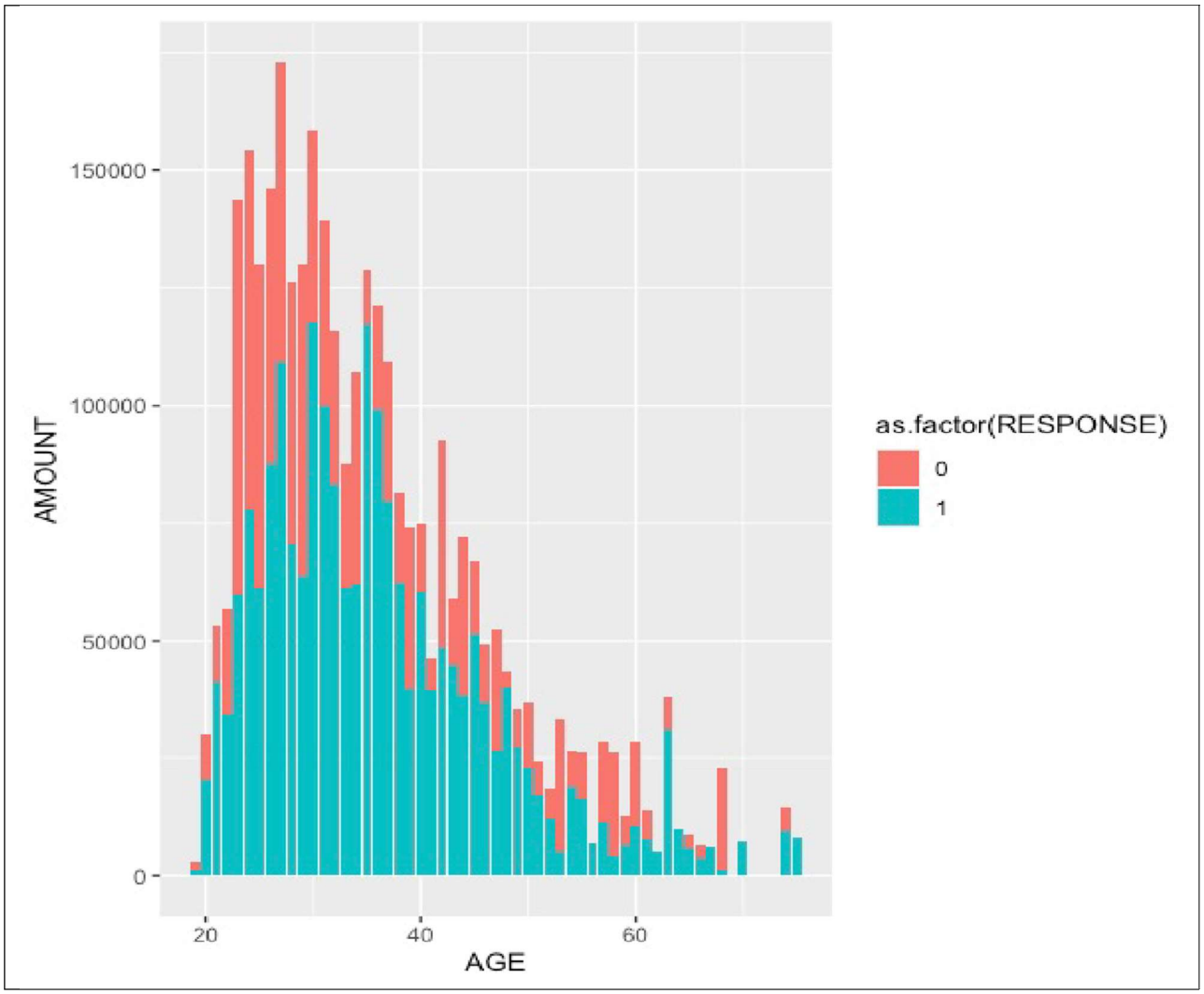
Out of 1000 observations in the given dataset, 700 observations are with Good Credit Rating and 300 observations are with Bad Credit Rating [**Good (1) vs Bad (0)],**



Age vs Credit Amount w.r.t Credit Rating Good (1) vs Bad (0):

We can see that,

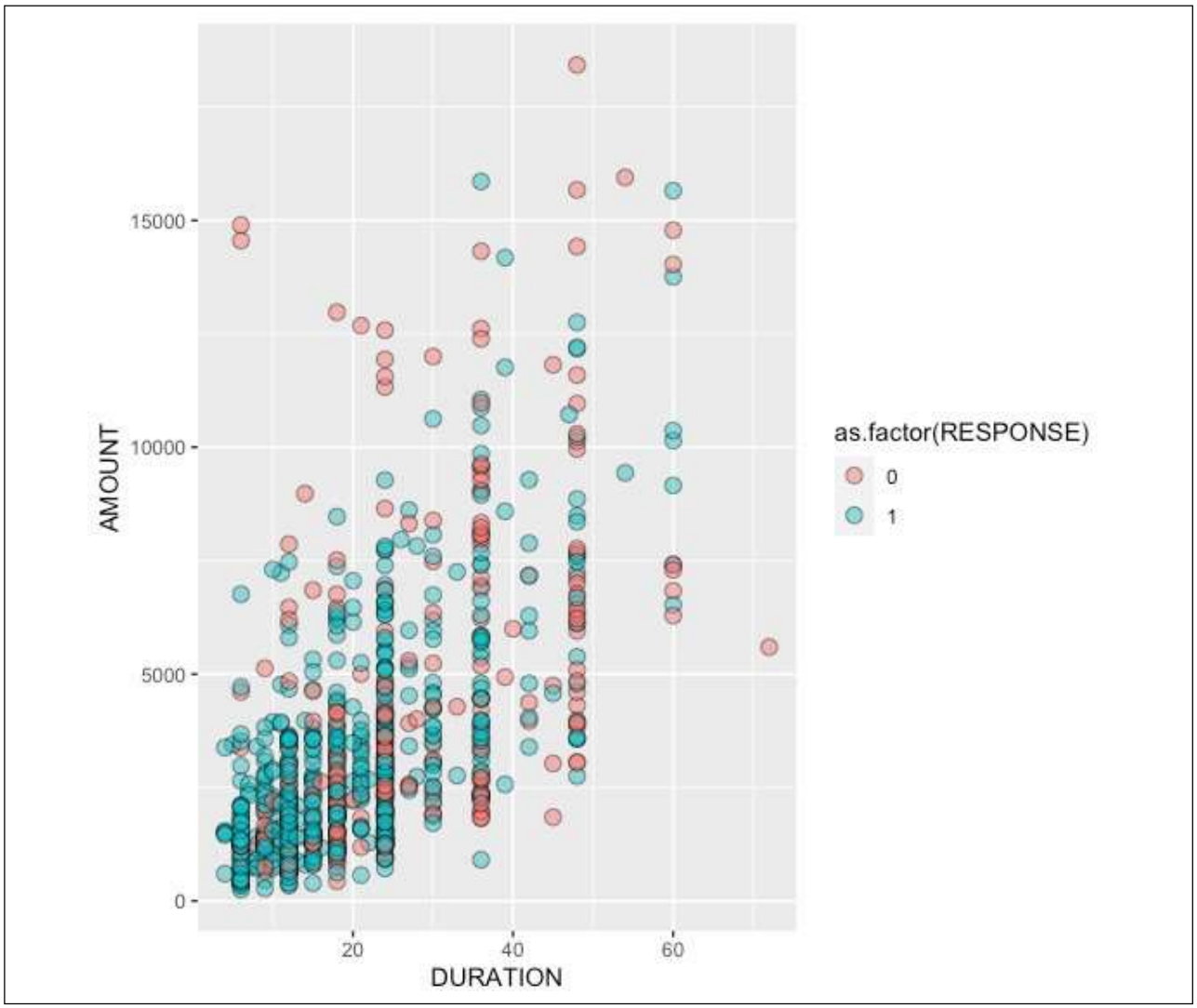
- Applicant age group between 25 to 45 and having credit amount up to 50000 has the Good Credit Rating.
- Banks offers higher credit amount to the age group between 25 to 45 having the Bad Credit Rating can be observed in the plot.
- Applicant age group between 50 to 65, offering lower credit amount to them is also somewhat risky as per the plot.



Duration vs Credit Amount w.r.t Credit Rating Good (1) vs Bad (0):

We can see that,

- Applicant Loan Duration between 4 to 25 years and having credit amount up to 5000 has the Good Credit Rating.
- Banks offers higher credit amount with higher loan duration period having the Bad Credit Rating can be observed in the plot
- Offering higher credit amount with less duration period is also having Bad Credit Rating.

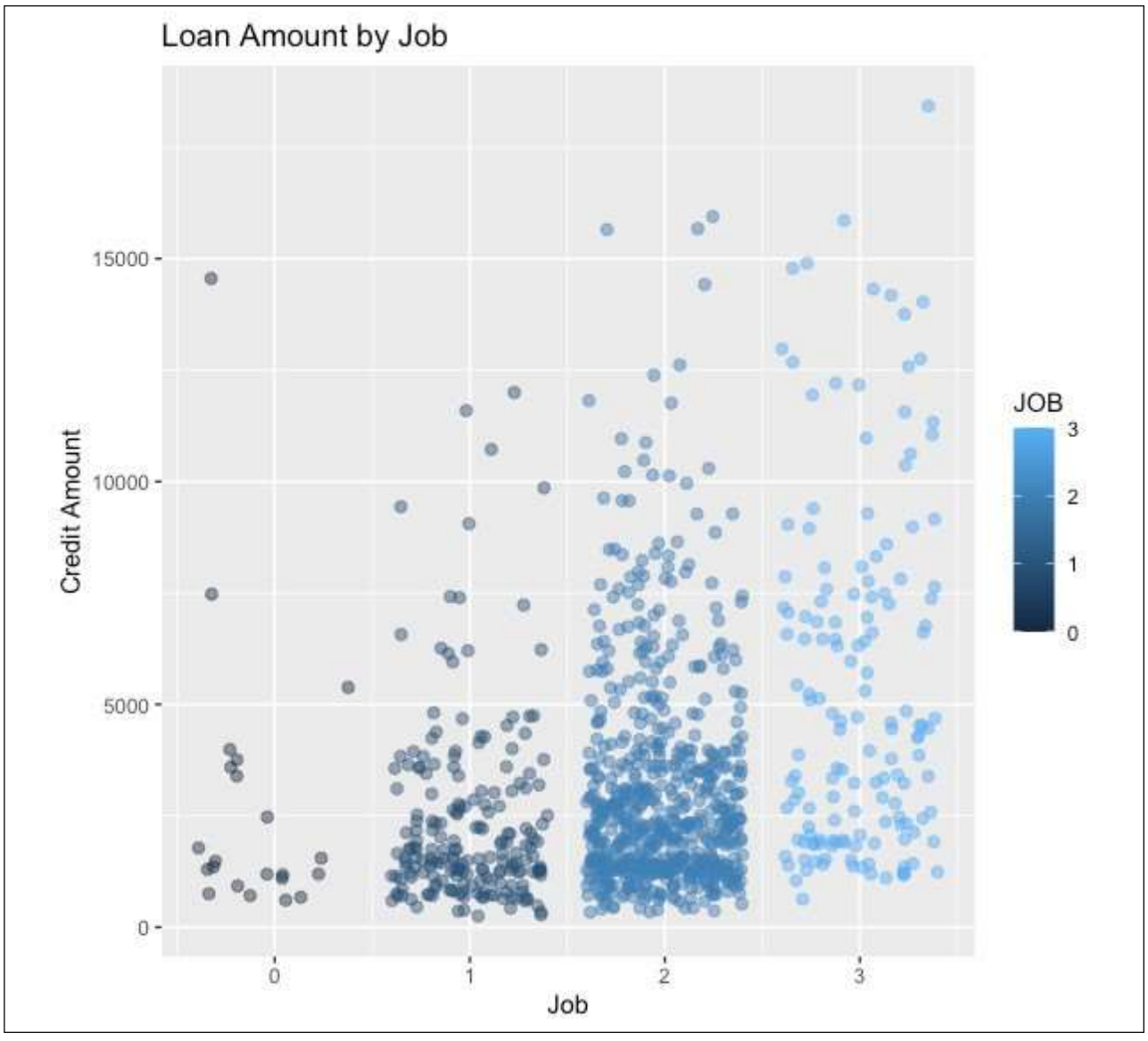


Duration vs Credit Amount w.r.t Credit Rating Good (1) vs Bad (0):

We can see that,

- Applicant having the job type 2 and 3 are offered a higher credit amount.
  Jobtype => 2 – Skilled employee, 3- Management/Self-Employed/Highly Qualified.

- Applicant having the job type 0 and 1 are offered lower credit amount.
  Jobtype => 0 – unemployed, 1- unskilled resident.



Average Balance in Saving Account vs Credit Amount w.r.t Credit Rating Good (1) vs Bad(0):

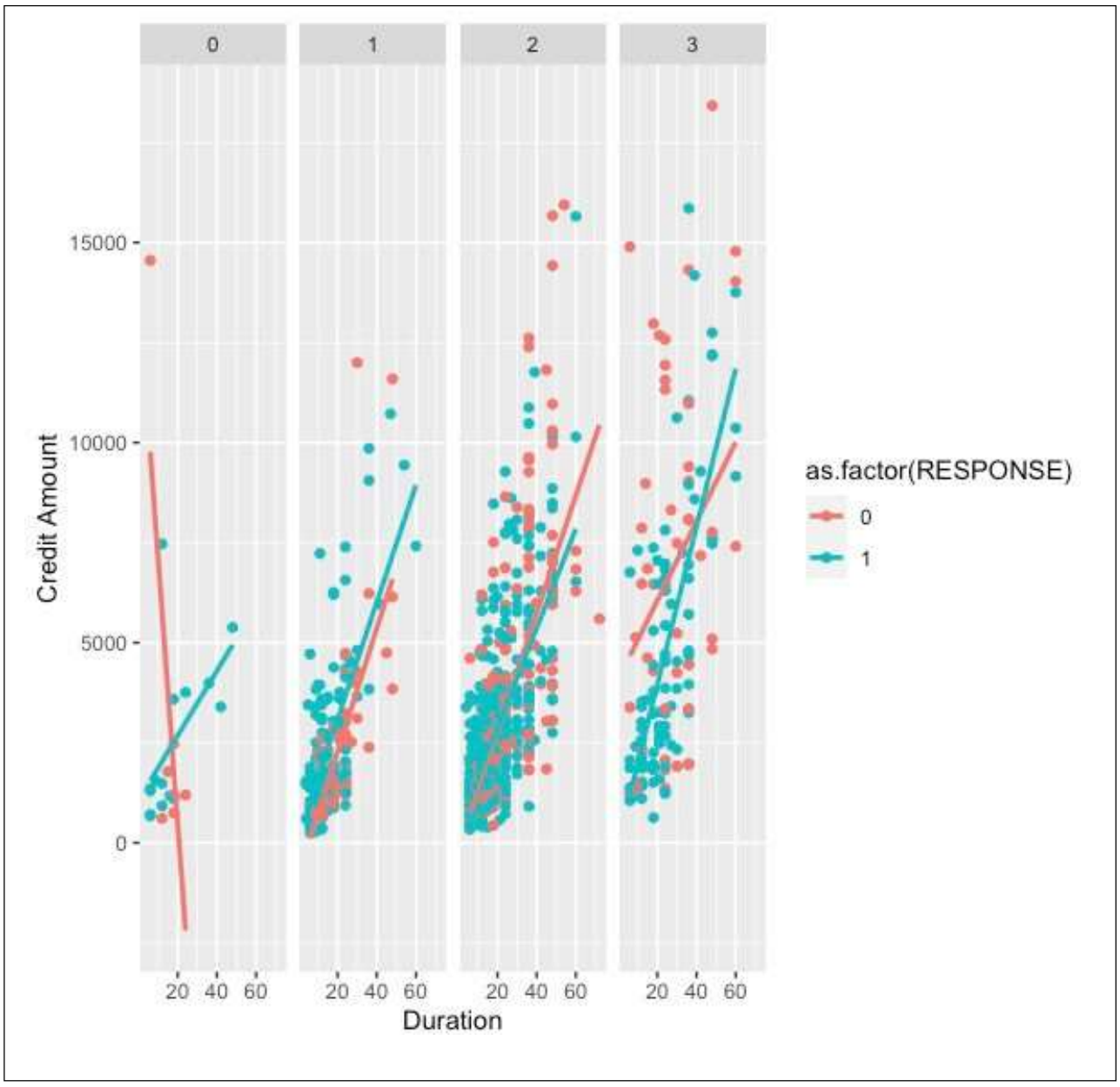
We can see that,

- Avg balance in savings account is minimum of 100DM and maximum of more than 1000DM has offered a credit amount up to 10000 are having Good Credit Rating.
- Avg balance in savings account is less than 100DM and offered a credit amount over 5000 having more Bad Credit Rating.

Loan Duration, Credit Amount, Job w.r.t Credit Rating [Good (1) vs Bad (0)]:

From this 4D plot we can see that,

- Loan duration from 4 to 25 years, given the credit amount up to 6000, having the Good Credit Rating for the job types 1, 2 and 3.
- Higher loan duration and higher loan amounts irrespective of job types are having Bad Credit Rating compared to Good Credit Rating.



4. Model Development

Three different models were developed to predict credit risk:

K-Nearest Neighbors (KNN) using Euclidean distance metric

K-nearest-neighbors algorithm can be used for classification (of a categorical outcome) or prediction (of a numerical outcome).
We defined the value of 'K' as 3. This means that the algorithm will consider the three neighbors that are the closest to the new data point in order to decide the class of this new data point.

- At K=3, we got the accuracy of 72.75%.
- AT K=11, we have got the highest accuracy i.e., 76%.

Confusion Matrix at K=3:

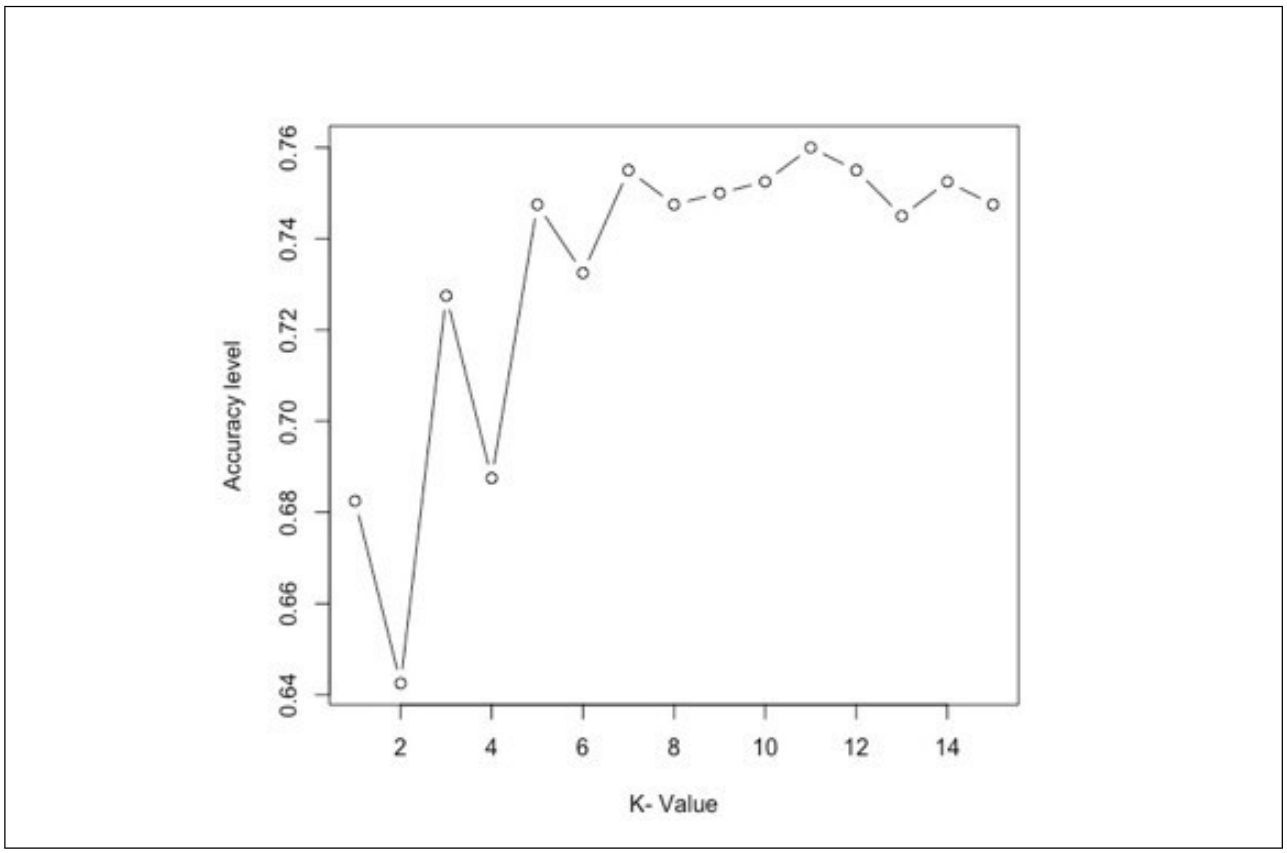
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  40  34
         1  75 251

               Accuracy : 0.7275
                 95% CI : (0.681, 0.7706)
    No Information Rate : 0.7125
    P-Value [Acc > NIR] : 0.2733962

                  Kappa : 0.2557

 Mcnemar's Test P-Value : 0.0001275

            Sensitivity : 0.3478
            Specificity : 0.8807
         Pos Pred Value : 0.5405
         Neg Pred Value : 0.7699
             Prevalence : 0.2875
         Detection Rate : 0.1000
   Detection Prevalence : 0.1850
      Balanced Accuracy : 0.6143

       'Positive' Class : 0
```

Confusion Matrix at K=11:

```
> print(conf_matrix_K11)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  29  10
         1  86 275

               Accuracy : 0.76
                 95% CI : (0.7151, 0.801)
    No Information Rate : 0.7125
    P-Value [Acc > NIR] : 0.01916

                  Kappa : 0.2704

 Mcnemar's Test P-Value : 0.00000000000001938

            Sensitivity : 0.2522
            Specificity : 0.9649
         Pos Pred Value : 0.7436
         Neg Pred Value : 0.7618
             Prevalence : 0.2875
         Detection Rate : 0.0725
   Detection Prevalence : 0.0975
      Balanced Accuracy : 0.6085

       'Positive' Class : 0
```

K=3 Observations	K=11 Observations
Lower accuracy(72.75%) with higher sensitivity (34.78%) indicates a decent ability to identify positive instances.	Improved overall accuracy (76%) with higher specificity (96.49%).
Balanced Accuracy = (Sensitivity + Specificity) / 2 Balanced Accuracy = (34.78 + 88.07) / 2 = 61.425%	Balanced Accuracy = (Sensitivity + Specificity) / 2 Balanced Accuracy = (25.22 + 96.49) / 2 = 60.855%
Balanced accuracy is 61.43%, with a trade-off between sensitivity and specificity.	Sacrifice in sensitivity (25.22%), but better positive predictive value (74.36%).

Accuracy graph for 15 iterations (K ranges from 1 to 15) can be seen below:



We have plotted the accuracy graph for 15 iterations (K ranges from 1 to 15), and we can see the highest accuracy at K=11, and the accuracy is not stable for other K-values, as we can see the variation in the plot.

Take away Conclusions:
Choose K=3 for a more balanced approach between sensitivity and specificity.
Choose K=11 for higher overall accuracy and a focus on correctly identifying negative instances.

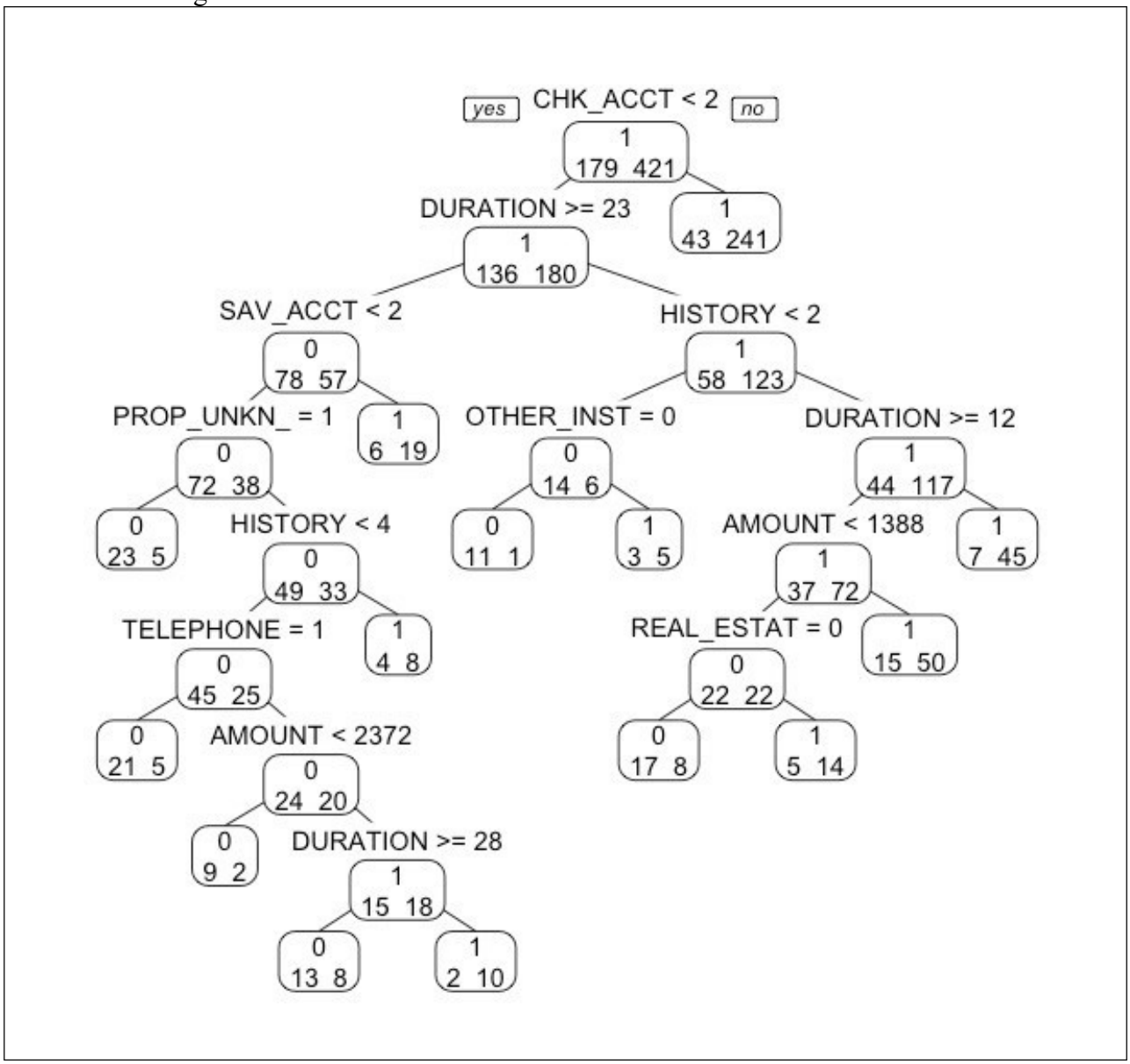
Classification Trees

If we do not know our classifiers, the decision tree will choose those classifiers for us

Default Decision Tree:

Observations from Default Decision Tree:

- The first split of the tree is on Checking Account balance and second split is on Duration.
- Checking Account and Duration seems to be the most important variables in the decision tree. Savings Account & History are also equally important depending on the condition check on CHK_ACCT & Duration.
- The rules/conditions can be observed in the below diagram and these can be reduced by removing redundancies



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  50  32
         1  71 247

               Accuracy : 0.7425
                 95% CI : (0.6967, 0.7847)
    No Information Rate : 0.6975
    P-Value [Acc > NIR] : 0.0270698

                  Kappa : 0.3285

 Mcnemar's Test P-Value : 0.0001809

            Sensitivity : 0.4132
            Specificity : 0.8853
         Pos Pred Value : 0.6098
         Neg Pred Value : 0.7767
             Prevalence : 0.3025
         Detection Rate : 0.1250
   Detection Prevalence : 0.2050
      Balanced Accuracy : 0.6493

       'Positive' Class : 0
```

The accuracy we got from default tree is **74.25%** which can be observed from the confusion matrix provided.

Validation of decision tree using the 'Complexity Parameter(CP)' & Cross Validated Error:

Here the optimal CP value associated with the minimum error is 0.01675978.

Graphical representation to the cross validated error summary:

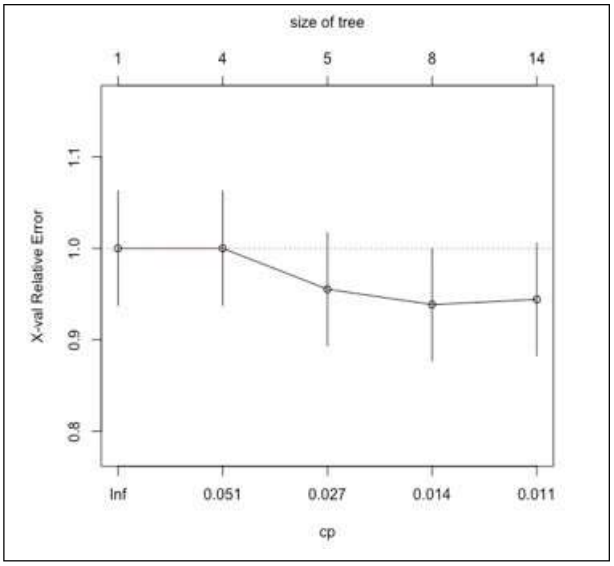
Here, The CP values are plotted against the geometric mean to depict the deviation until the minimum value is reached.

We prune the tree to avoid any overfitting of the data. The convention is to have a small tree and the one with least cross validated error given by printcp() function i.e. 'xerror'.

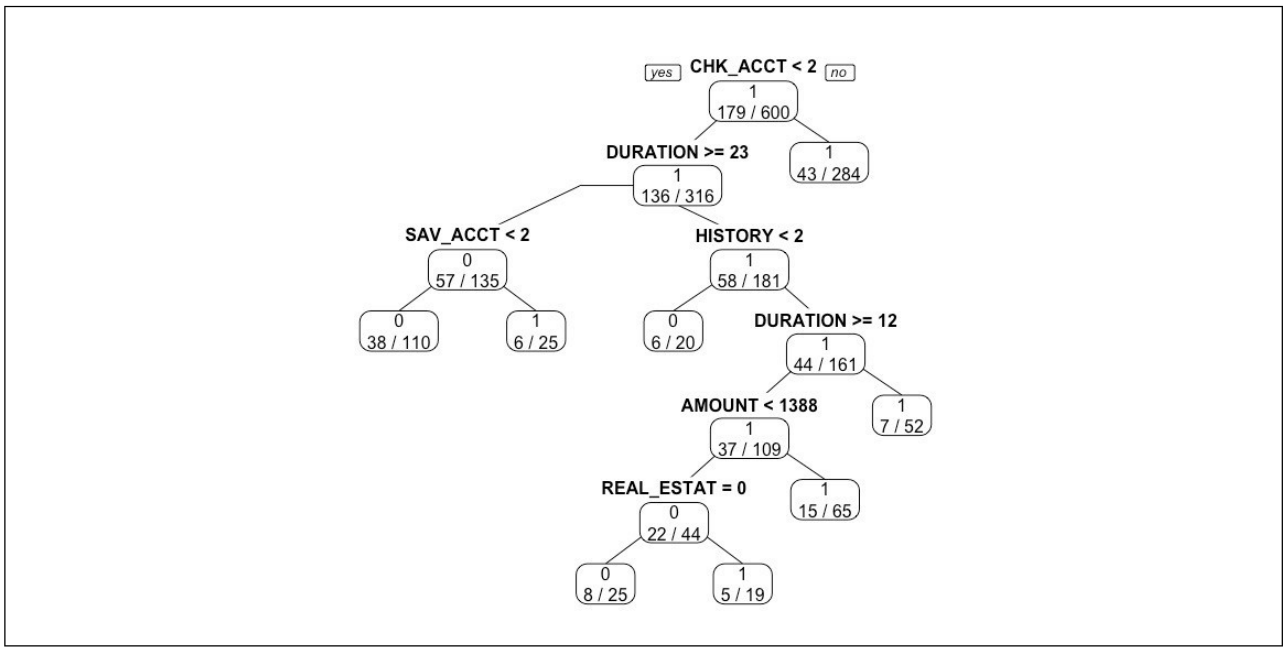
From the list of cp values, we need to select the one having the least cross-validated error(xerror) and we use it to prune the tree.

The value of cp should be least, so that the cross-validated error rate is minimum.

Here the optimal CP value associated with the minimum error is 0.01



Pruned Tree :



The accuracy we got from pruned tree is 75.75%

Random Forest:

The accuracy we got from random forest is 76%.

Logistic Regression

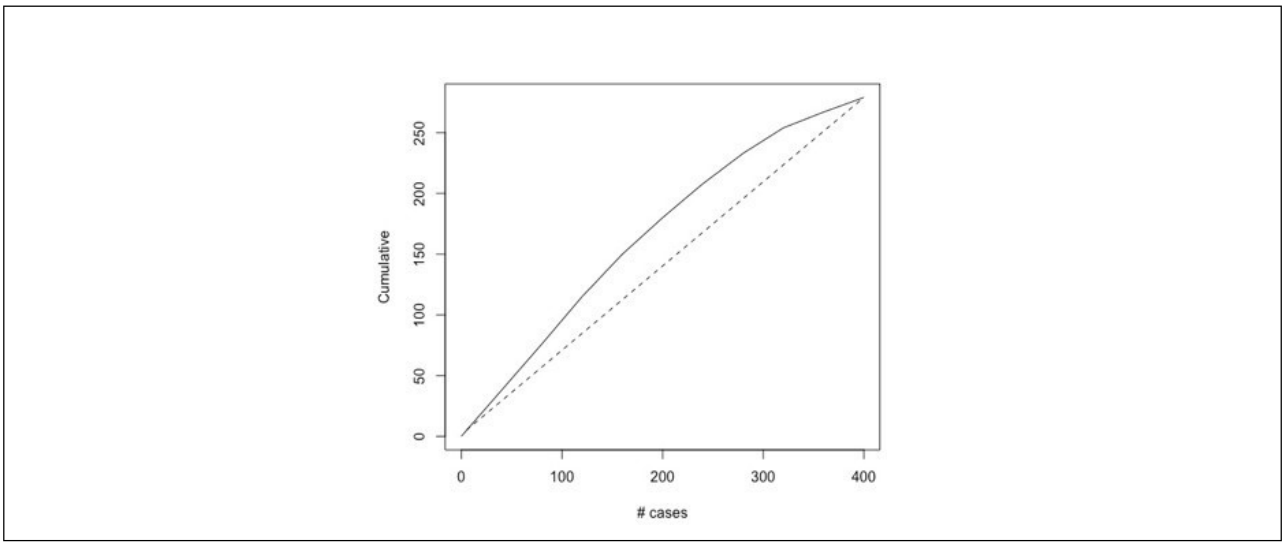
For logistic regression, we can use maximum likelihood. It does not assume a linear relationship between dependent and independent variables. However, it assumes a linear relationship between link function and independent variables in logit model.

- For logistic regression: The predicted Y lies with in 0 and 1 range.
- To interpret logistic regression results, we first transform the coefficients to odds ratios by raising e - or Euler's constant - to the coefficients.
- As we have so many columns, the result of GLM function for logit model can be observed in R code.
- We can see P-values and Z-values for all the columns in R. As we know that high P-value and low Z-value columns are not a significant predictor columns.

Interpreting the meaning of the logit coefficients for at least two influential variables.

- For 1 unit increase in CHK_ACCT holding others constant, there is 0.522 times increase in magnitude by which log of odds of belonging to class 1 changes.
- For 1 unit increase in SAV_ACCT, there is 0.2522 increase in magnitude by which log of odds of belonging to class 1 changes.

Here in our logit model, the lift chart having the lower area under the curve, so we feel it may not be the better model.



Deciles plot:

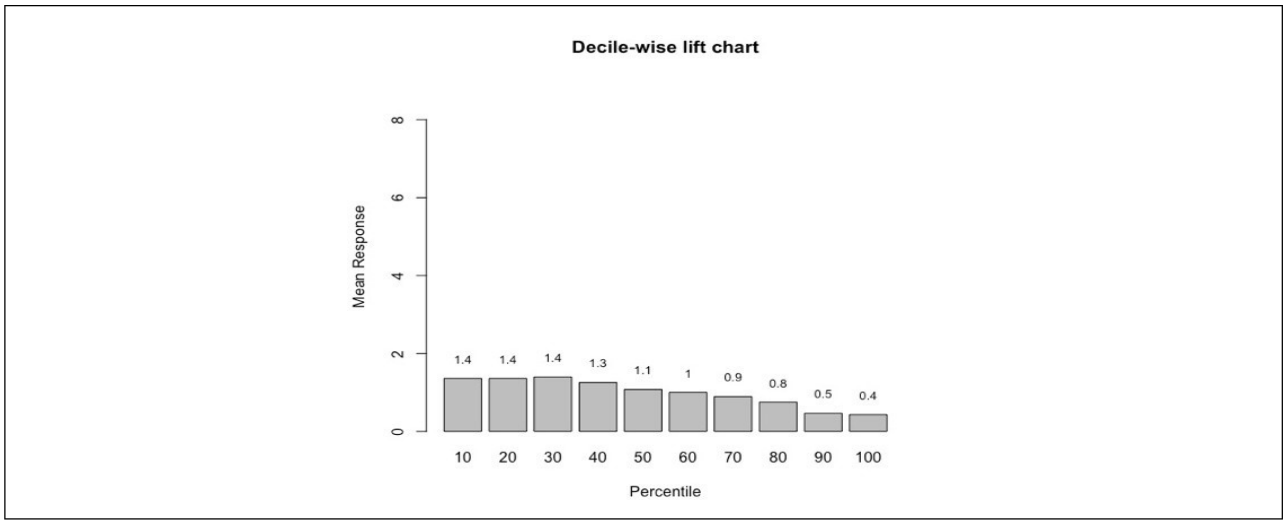
The deciles plot tells that,

- If we target the first 10% of the customers from the predictions made by the model, approximately 13.6% percent of them will be default.
- If we target the first 60% of the customers from the predictions made by the model, approximately 74.6% percent of them will be default.

Notes:

With 700 observations labeled as Good Credit Rating (1) and 300 observations as Bad Credit Rating (0) in the dataset, we have a clearer breakdown of the credit ratings within the 1000 observations.

The 13.6% and 74.6% corresponds to the percentage of customers with a Bad Credit Rating, totaling 300 individuals. For the initial 10% of customers (10% of 1000), there would be 13.6% of 300 individuals with a Bad Credit Rating. Similarly, for the first 60% of customers (60% of 1000), the number of defaulters would be 74.6% of the 300 individuals with a Bad Credit Rating



Null Deviance and Residual Deviance:

- Null deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model.
- Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.

AIC: The analogous metric of adjusted RÇ in logistic regression is AIC.

- AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value.

The Null Deviance (the deviance just for the mean) and Residual Deviance (the deviance for the model with all the predictors) with degrees of freedom, and AIC from our logit model can be observed in the below screenshot:

```
    Null deviance: 731.33  on 599  degrees of freedom
Residual deviance: 553.79  on 569  degrees of freedom
AIC: 615.79

Number of Fisher Scoring iterations: 6
```

We can see that there is difference in Null deviance and Residual deviance, and also in their degrees of freedom.

Confusion Matrix and Statistics:

The accuracy we got from logit model is 77%, which can be observed in the below screenshot:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  63  34
         1  58 245

               Accuracy : 0.77
                 95% CI : (0.7256, 0.8104)
    No Information Rate : 0.6975
    P-Value [Acc > NIR] : 0.0007502

                  Kappa : 0.4225

 Mcnemar's Test P-Value : 0.0164887

            Sensitivity : 0.5207
            Specificity : 0.8781
         Pos Pred Value : 0.6495
         Neg Pred Value : 0.8086
             Prevalence : 0.3025
         Detection Rate : 0.1575
   Detection Prevalence : 0.2425
      Balanced Accuracy : 0.6994

       'Positive' Class : 0
```

Specificity and Sensitivity plays a crucial role in deriving ROC curve.

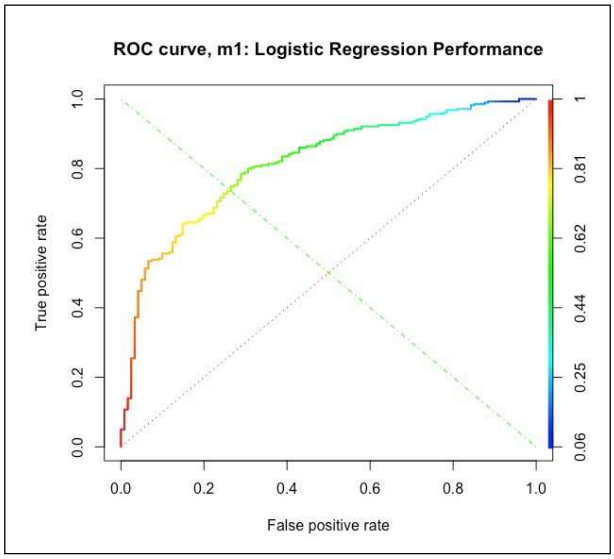
ROC Curve:
Receiver Operating Characteristic (ROC) summarizes the model's performance by evaluating the trade-offs between true positive rate (sensitivity) and false positive rate (1- specificity).

For plotting the ROC, it is advisable to assume p > 0.5 since we are more concerned about success rate. ROC summarizes the predictive power for all possible values of p > 0.5. The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model.

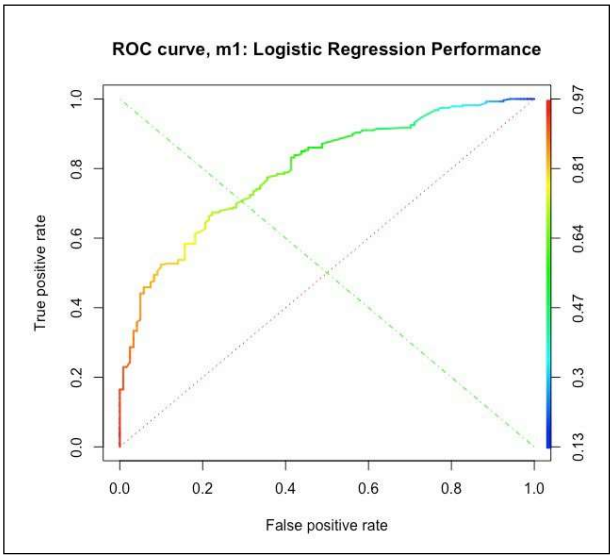
The ROC of a perfect predictive model has TP (true positive) equals 1 and FP (false positive) equals 0. This curve will touch the top left corner of the graph.

We developed two logit models, and compared the ROC curve between the two models, and we feel 'model1' is better than 'model2' as 'model1' has higher area under the curve when compared to 'model2'.

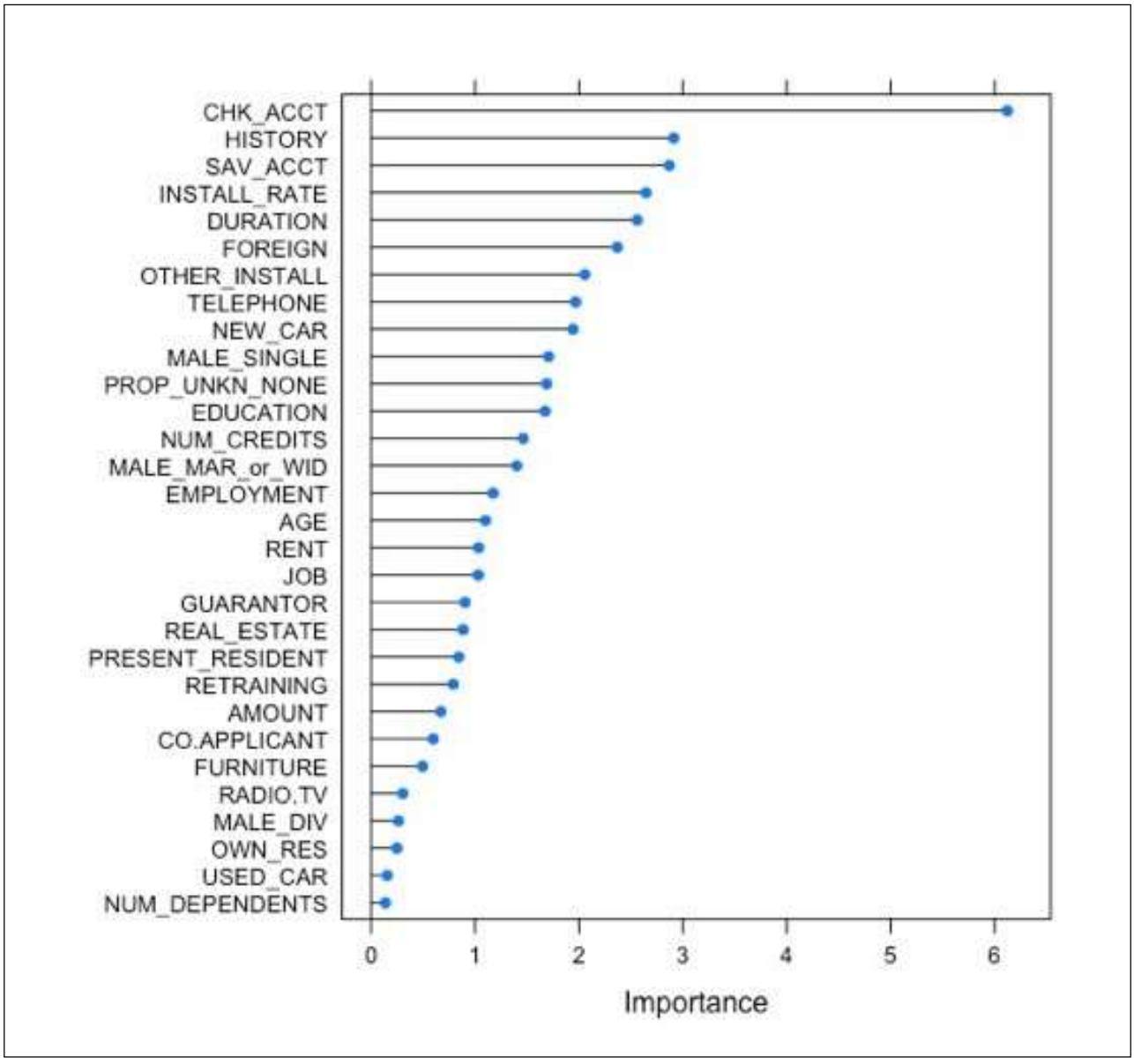
Model 1:



Model 2:



The **Variable of importance plot from logit model** can be observed below:



The Variables CHK_ACCT, HISTORY, SAV_ACCT, INSTALL_RATE, DURATION looks like important variables.

5. Results and Findings

The performance of each model was evaluated using relevant metrics such as confusion matrices, accuracy, type-1 and type-2 errors, AUC curves, sensitivity, and specificity. The findings provided insights into which variables effectively predict credit ratings and influence credit risk.

Comparing the models developed:

Comparing the predictive accuracy in your validation data for all the models developed above are given below:

- The predictive accuracy we got from K-NN,
  - At K=3, we got the accuracy of 72.75%.
  - AT K=11, we have got the Accuracy of 76%.
- The Predictive Accuracy we got from the Default Decision tree is 74.25%.
- The Predictive Accuracy we got from the Pruned tree is 75.75%.
- The Predictive Accuracy we got from the Random Forest is 76%.
- The Predictive accuracy we have got from Logit is 75.94%.

Logit model has given the highest accuracy when compared to KNN and Trees.

Variable Influences on Credit Risk

The predictors we would use based on predictive accuracy and the analysis above:

We will consider the columns which has low P-values (p<=0.05) those are significant predictor columns. Few of them are listed below,

- CHK_ACCT – Checking account status with the bank.
- HISTORY – Credit history.
- SAV_ACCT – Avg balance in savings account.
- INSTALL_RATE – Installment rate %.
- OTHER_INSTALL – Other installments if any.
- FOREIGN – Either foreign worker or not.
- TELEPHONE – Applicant having phone number in this name.

We can consider these important attributes from this model, as these variables play an important role in classifying whether a person has a Bad Credit Rating or Good Credit Rating.

6. Discussion

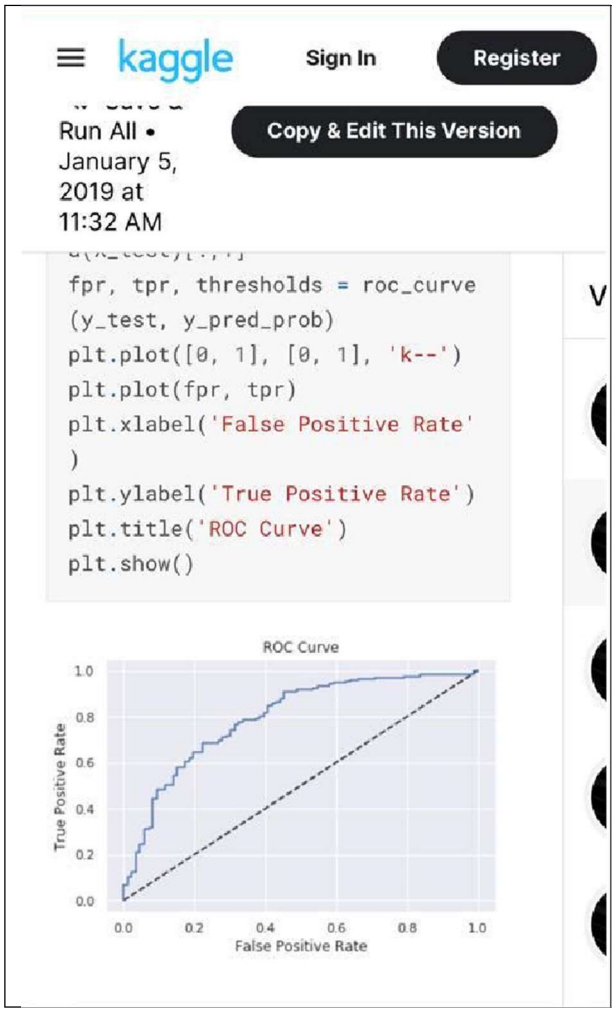
The results were interpreted and their implications for the banking industry were discussed. We considered how these findings could be applied to make informed lending decisions and minimize risk.

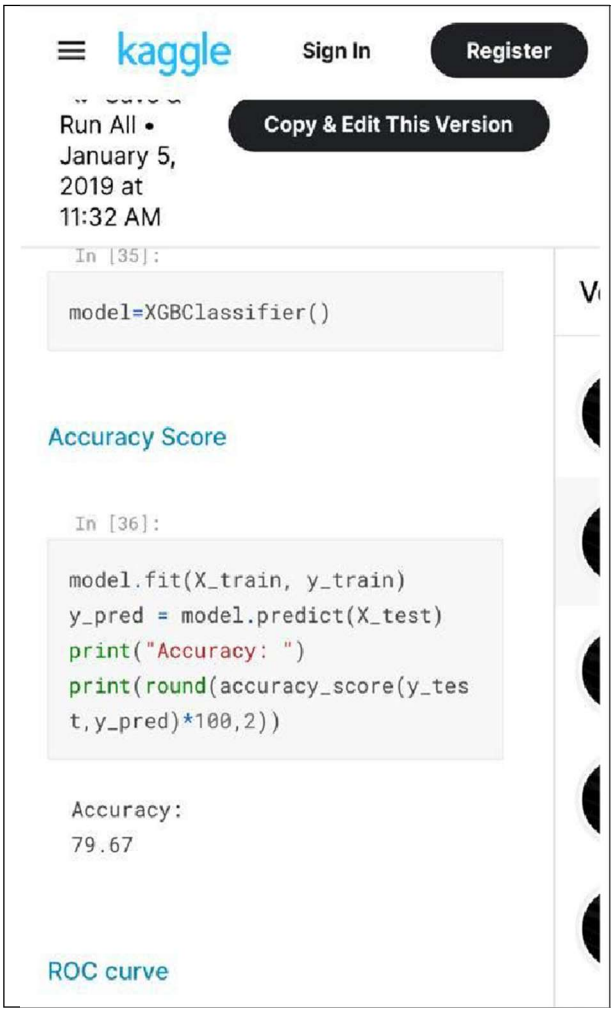
We compared our results with the existing analysis available online and listed below, and we can see that **we got the better accuracy's than the model analysis available online.**

Comparison with Existing Analysis (available online)

a) Kaggle

- Kaggle has various models for this data set.
- These models include Principal Component Analysis, and the best accuracy is 79.67%.





b) Penn State analysis

- Penn state has the analysis listed as part of its course work.
- Models used: Decision Trees, Logistic Regression, Linear Discriminant Analysis (LDA).
- Accuracy: 73.4% (Random Forest); 60% (Logit); 58.2% (LDA).

Link: https://online.stat.psu.edu/stat857/node/220/

7. Conclusion

1) **K-NN Model:**
   At K=3, the accuracy is 72.75%, with a balanced accuracy of 61.43%.
   At K=11, the accuracy improves to 76%, with a balanced accuracy of 60.86%.
   Higher sensitivity at K=3 (34.78%) compared to K=11 (25.22%).
2) **Decision Trees:**
   Default Decision Tree: Accuracy of 74.25%.
   Pruned Tree: Improved accuracy to 75.75%.
3) **Random Forest:**
   Achieved an accuracy of 76%.
4) **Logit Model:**
   Highest accuracy among all models, with a predictive accuracy of 75.94%.

Overall Comparison:

- The Logit model out performs both K-NN and Decision Trees in terms of accuracy.
- K-NN at K=11 and Random Forest share the highest accuracy at 76%.
- Decision Trees show improvement with pruning but are still slightly behind other models.
- Sensitivity and specificity considerations: K-NN at K=3 favors sensitivity, while Logit offers a balanced approach.

8. Recommendations

For Balanced Performance:
Choose the Logit model with 75.94% accuracy, providing a well-balanced predictive capability.

Alternative Options:
Consider K-NN at K=11 and Random Forest, both offering competitive accuracies at 76%.

Specific Application Focus:
Tailor the model choice based on application needs, emphasizing sensitivity, specificity, or overall accuracy.

9. References

Textbook:
Title: Data Mining for Business Analytics: Concepts, Techniques, and Applications in R,
Authors: Galit Shmueli, Peter C. Bruce, Inbal Yahav, Nitin R. Patel, and Kenneth C. Lichtendahl Jr.
Publisher: Wiley.
ISBN: 10: 1118879368 ISBN 13: 978-1118879368