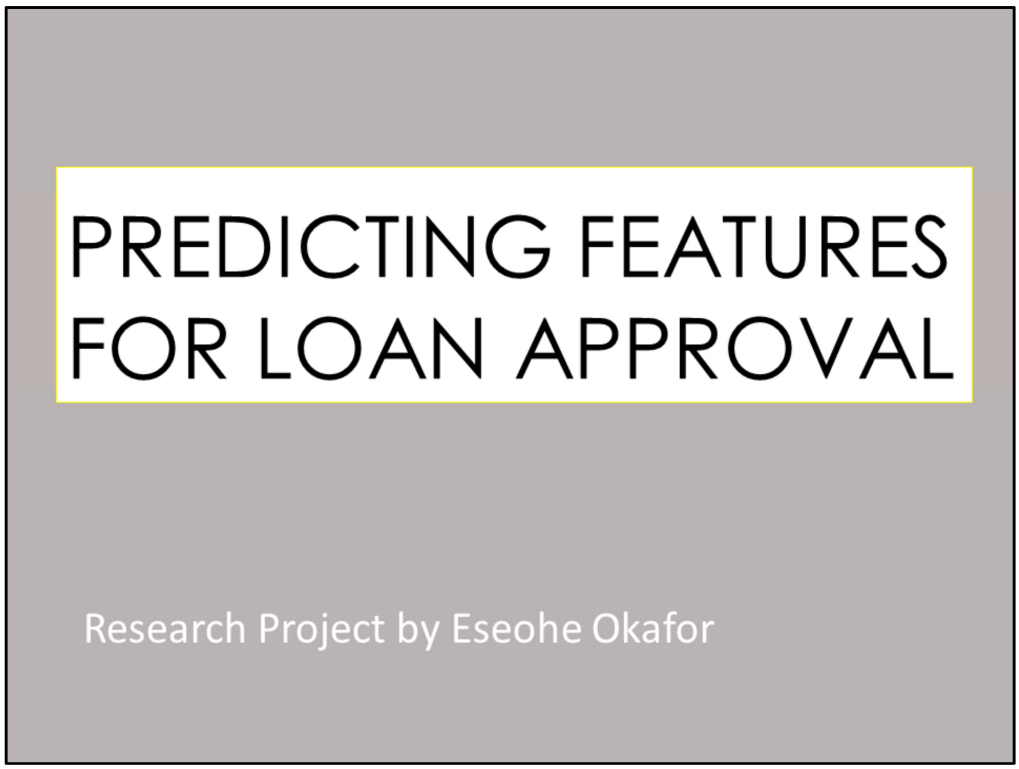
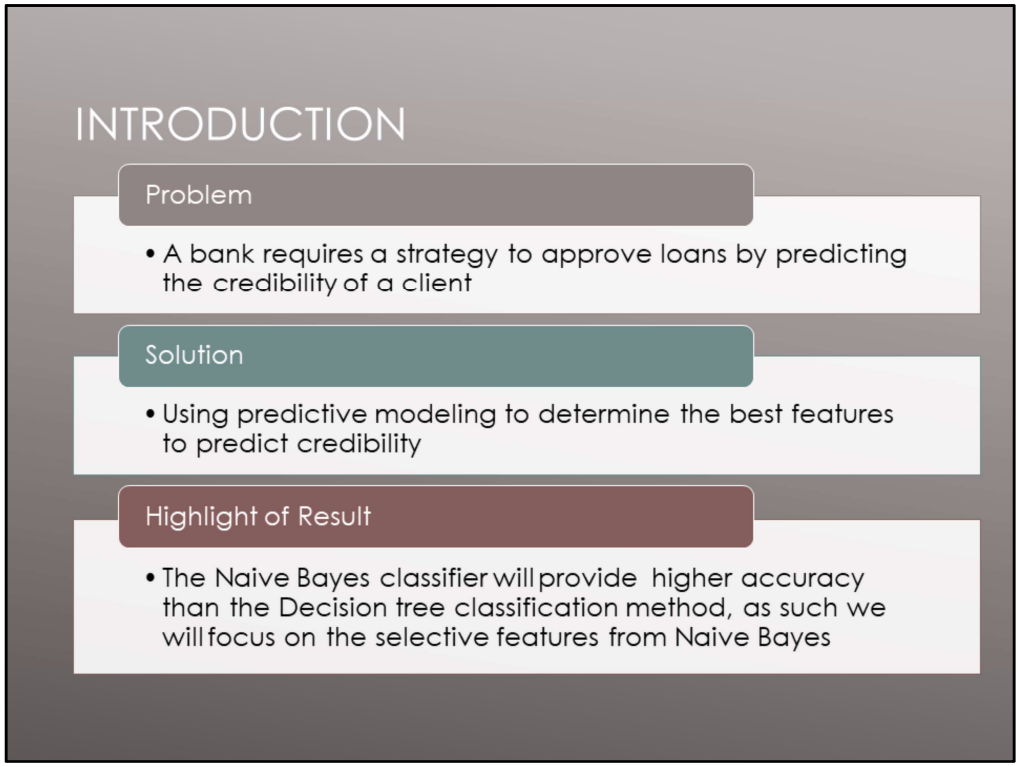

PREDICTING FEATURES FOR LOAN APPROVAL
Research Project by Eseohe Okafor

Title slide

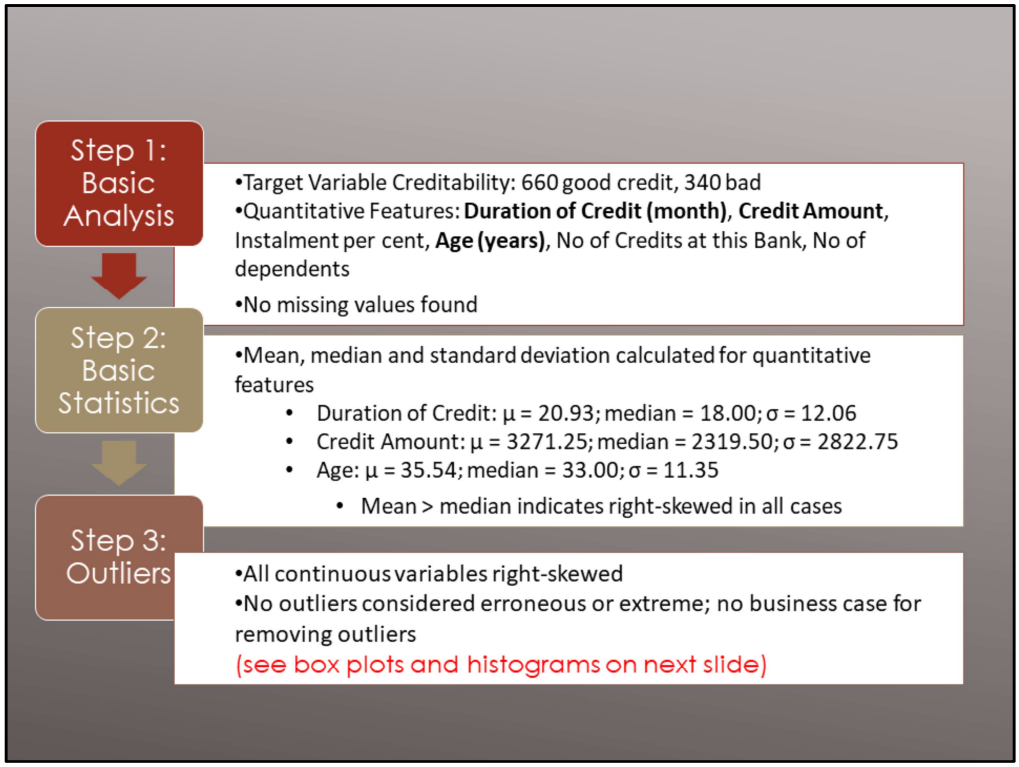


•**Problem:** The bank manager is faced with a dilemma of deciding on which loan application to approve or not. The bank manager would like a predictive model that would allow a non biased loan approval process that also has a high level of accuracy. Having the right Loan approval strategy reduces the risk of credit exposure on the bank also ensuring the business entity survival.

•**Solution:** I took a look at the dataset provided on it customer base and performed some preliminary assessment on it to ensure the data set had no missing values and that all character values where in useable format. Then the baseline dataset was analyzed using the Decisions tree and Baseline line classifiers and we

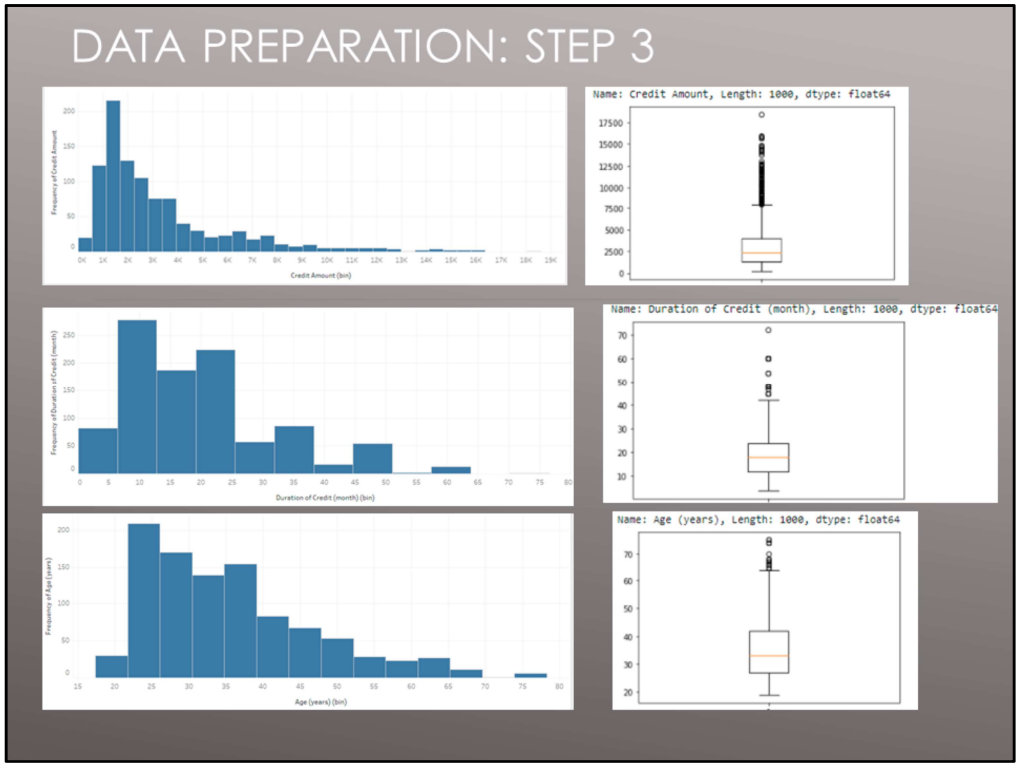
compared the results and determined the best features for accurate prediction. After that a new dataset with selected features was created and was run on decision tree and Naive Bayes in order to see which would give the best accuracy.

•**Highlights of the final results**: After comparing the four models, we decided to go with the values from the Naive Bayes classification as it provided better accuracy values and true positive values.

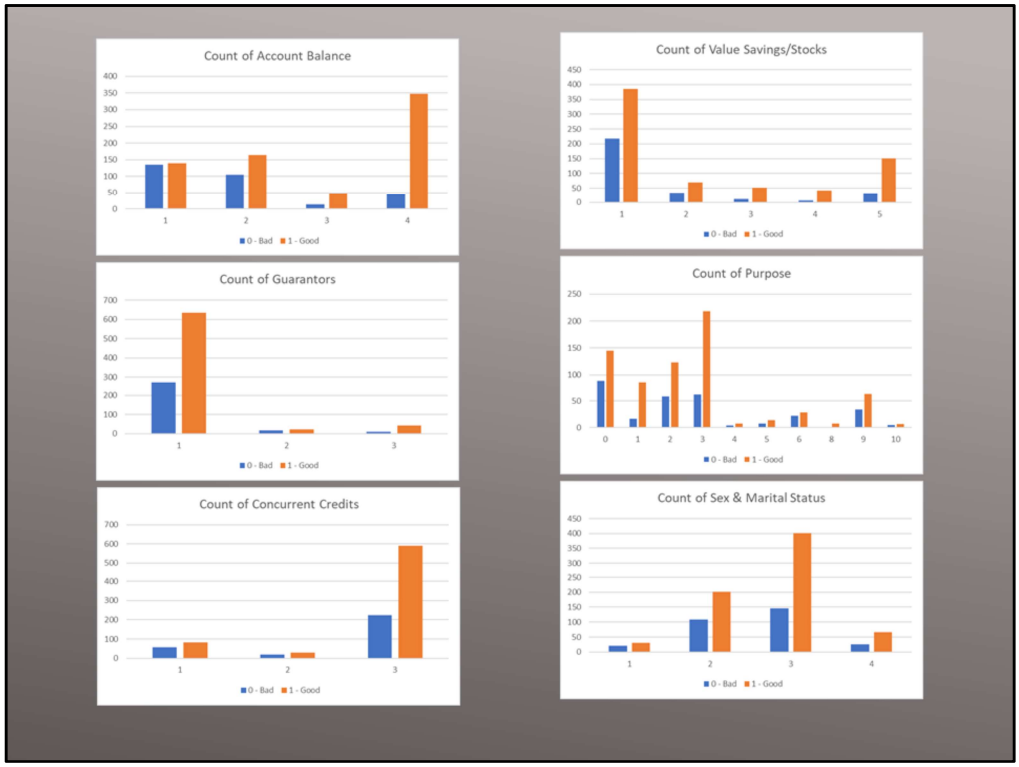


The German credit data set has 21 features. Creditability, the target feature is a categorical variable. 1000 records were present in the dataset. 700 are marked with a creditability of 1 for good credit and 300 are marked with 0 for bad credit. The quantitative features in bold are continuous variables. All other features not listed are categorical variables.

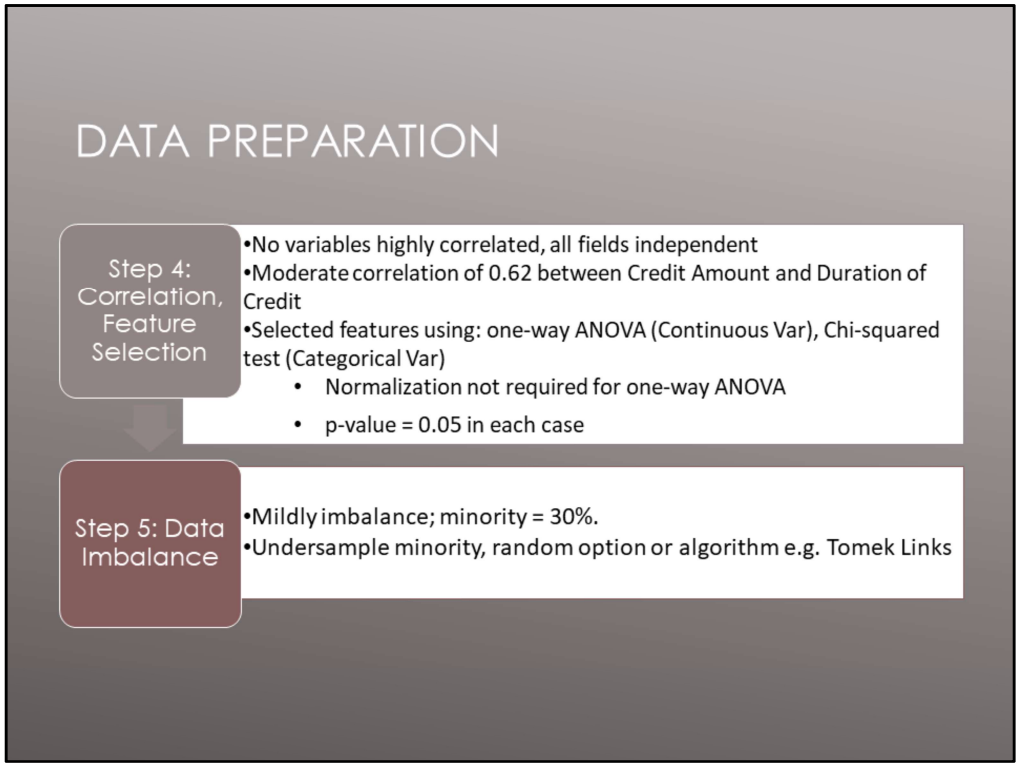
Descriptive statistics were calculated for all numerical variables. Notably, the mean exceeds the median for all continuous numerical values, indicating a right- skewed distribution. Box plots and histograms below bear this out visually.



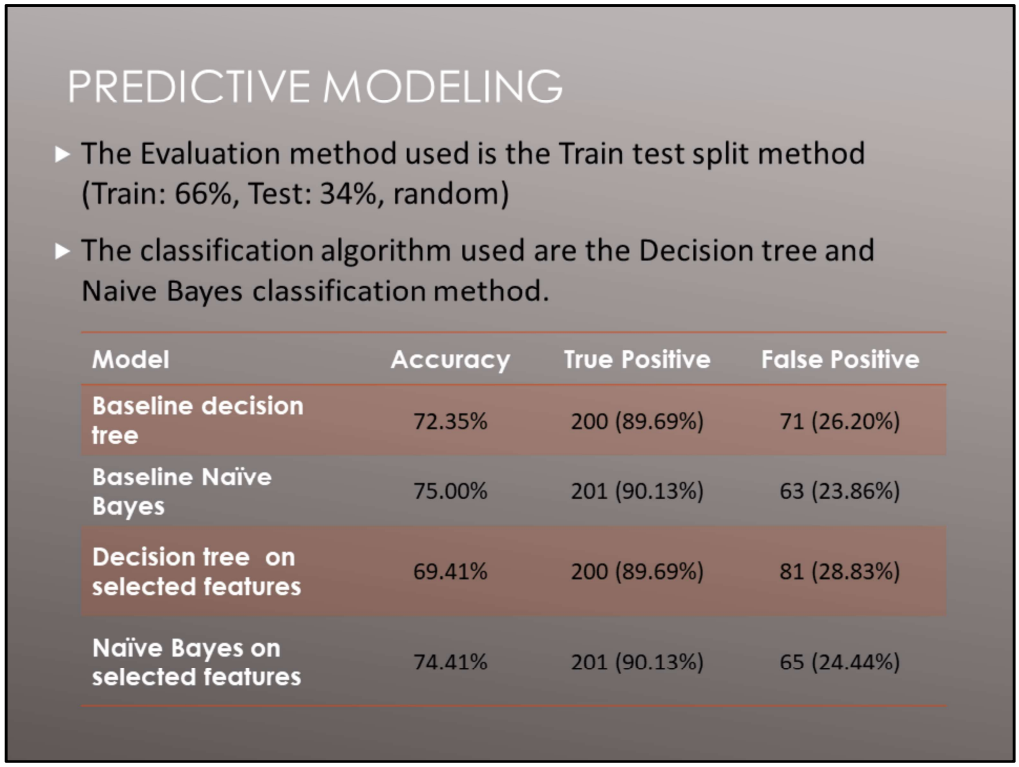
Histograms created using tableau, box plots in python assuming a normal distribution, again we see a right skew.



Some categorical variables showed a difference in frequency of good and bad credit between category options. This includes Value Savings, Account Balance, Payment Status of Previous Credit (not shown), Purpose, Sex/Marital Status, Concurrent Credits, Foreign Worker (not shown), Guarantor.



Highly correlated variables indicate a relationship between two variables (i.e. the fields are not independent). In such cases, one of the variables predicts the other and may be removed from analysis. We do see moderate correlation between Credit Amount and Duration. No other variables appear to highly correlated. Features for this project were selected using a one-way Anova for continuous variables and Chi-squared test for categorical variables. The minority class exists at the 30% level and is considered mildly imbalanced. Undersampling the minority class either randomly or using an algorithm such as the Tomek Link method may create a 50:50 dataset.



Evaluation Method: To evaluate the dataset, we will be using the Train test set spit method (random), which is easy to implement and less time consuming as we will only have to train 66% of the dataset (randomly selected) in creating the model. For the purpose of this analysis, we will be looking at the following results from the classification algorithm; Accuracy, True Positive, False Positive and finally precision and recall values. The two classification algorithm used will be:

-The Decision tree classification method which is a decision support tool that represents entire dataset population being analyzed. It usually begins with a root node (Account Balance in this case) being split into multiple decision sub-node.
-The Naïve Bayes classification method is a probabilistic statistical analysis tool aiming to show the probability of a label given observable features. In this case the label is Credibility with the label indicating whether the customer has a good or bad credit rating.

What do we mean by positive?
we define positive as a client with good credit. credibility = 1

False positive is the more important value here because we don't want to give loan to people who are not credible. This is to minimize risk/loss. Hence, we want high precision value.

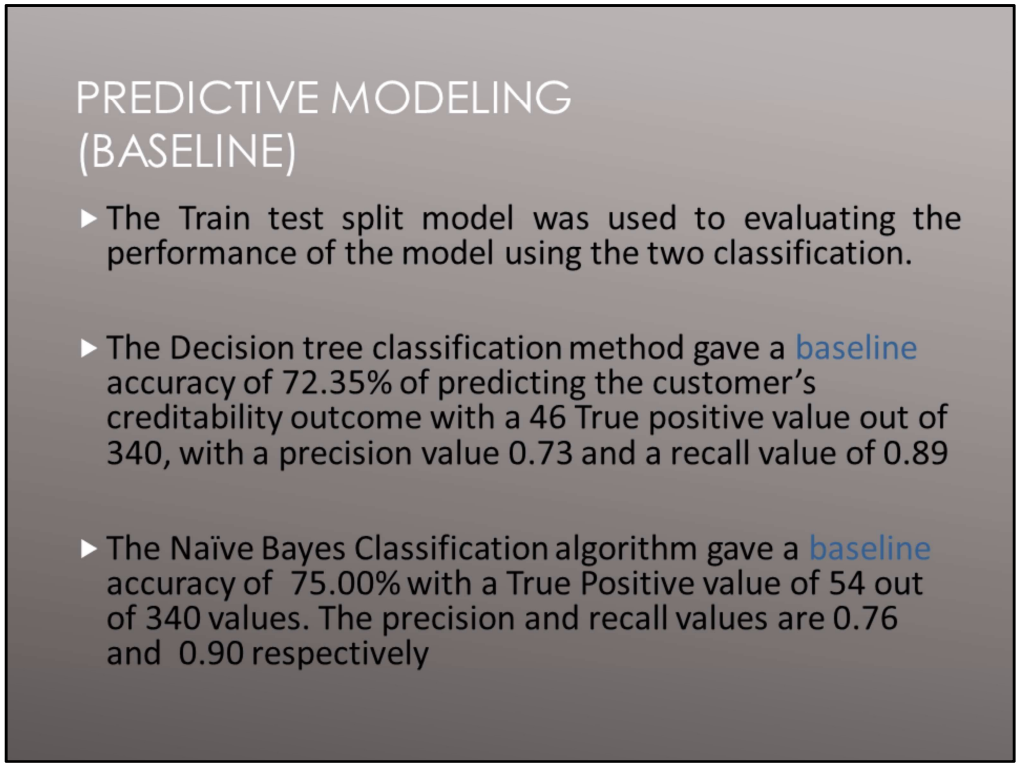
True positive rate (TPR), recall, **sensitivity** = TP / (TP + FN)
False positive rate (FPR), fall-out, probability of false alarm = FP / (FP + TN)
Positive predictive value (PPV), precision = TP / (TP + FP)

Our baseline model perform better than the selected features as it has higher accuracy and lower number of false positive.

We did predictive modeling using weka and sklearn. We choose weka classification algorithm because it produces higher accuracy. Sklearn data:

•using 14 selected features:

- DT accuracy: 73%
- NB accuracy: 62%

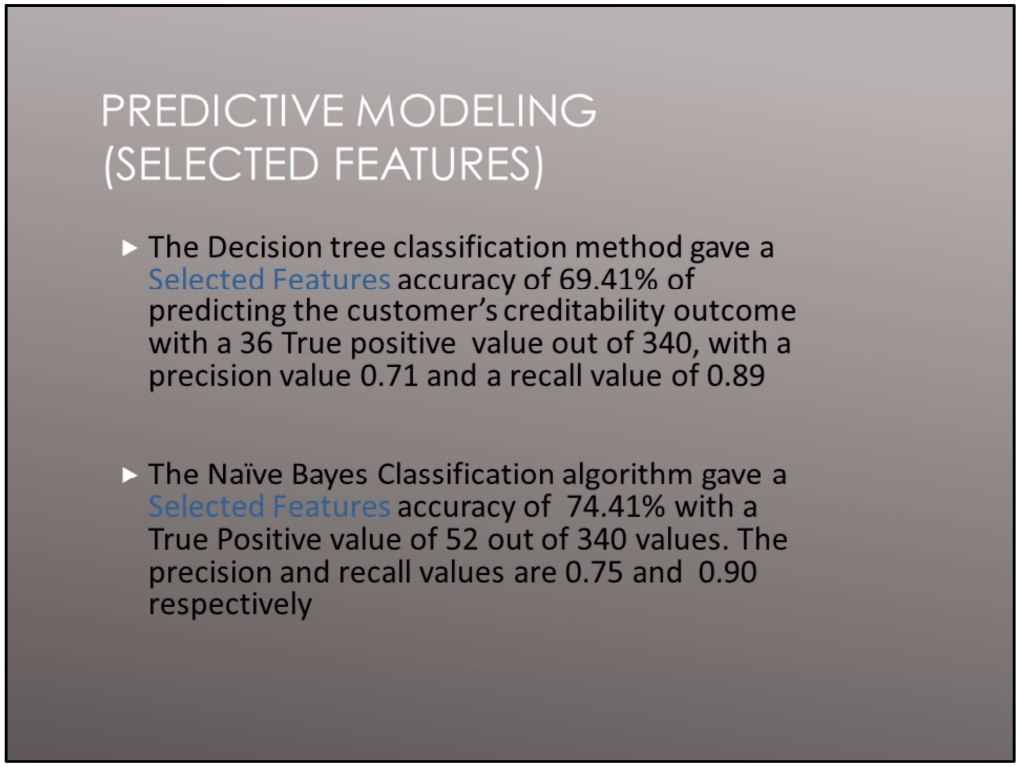


For the Baseline model, the Naïve Bayes classifier outperformed the Decision tree classifier. With prediction accuracy at 75%, it is an improvement of 2.65% over the Decision tree classifier. The True Positive Value indicates the model will be able to correctly assign the right class label of good credit 54 out of 76 times. The difference 22 represents the False Positive value. The precision value of 0.761 and recall value of 0.90 for the good credit classification is also an improvement form the Decision tree classifier which gives us a precision value of 0.73 and recall value of 0.89in comparison.

• *Note 1: Precision value indicates the ratio of correct positive predictions to the total values predicted positives.*

Recall values indicates ratio of correct positive predictions to the total actual positives

- *Note 2: Predictive Modelling analysis performed on Weka*

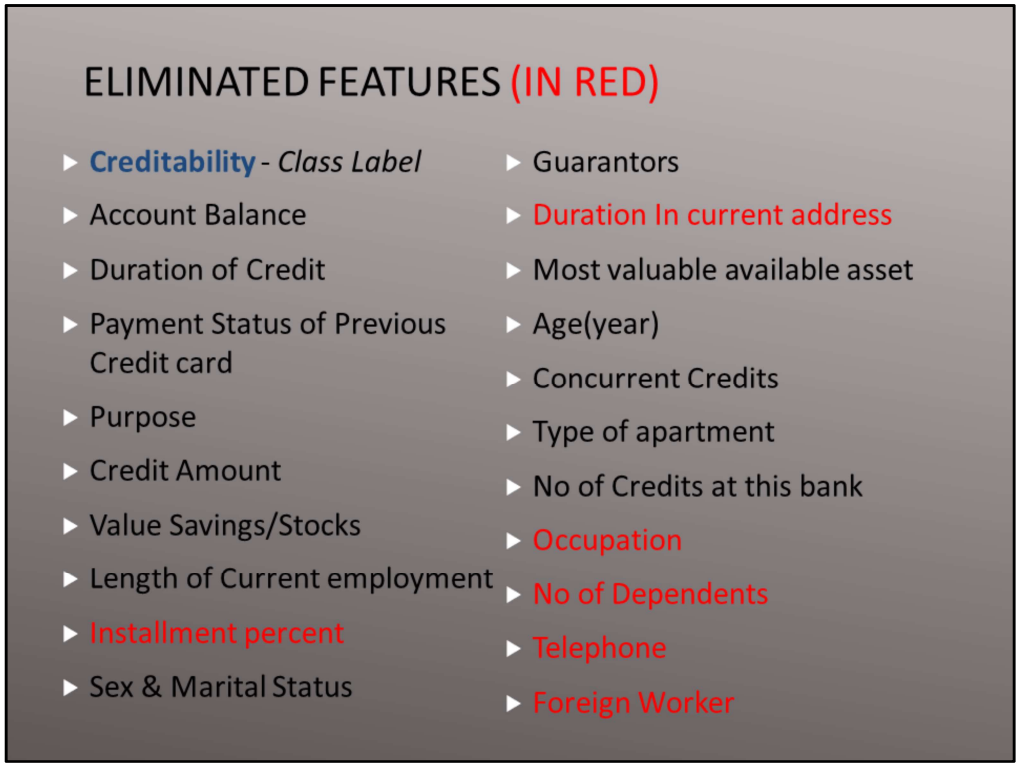


For the Selected features model, the Naïve Bayes classifier again outperformed the Decision tree classifier. With prediction accuracy at 74.41%, it is an improvement of 5% over the Decision tree classifier. The True Positive Value indicates the model will be able to correctly assign the right class label of good credit 52 out of 74 times. The difference 22 represents the False Positive value. The precision value of 0.71 and recall value of 0.89 for the good credit classification is also an improvement form the Decision tree classifier which gives us a precision value of 0.75 and recall value of 0.90 in comparison.

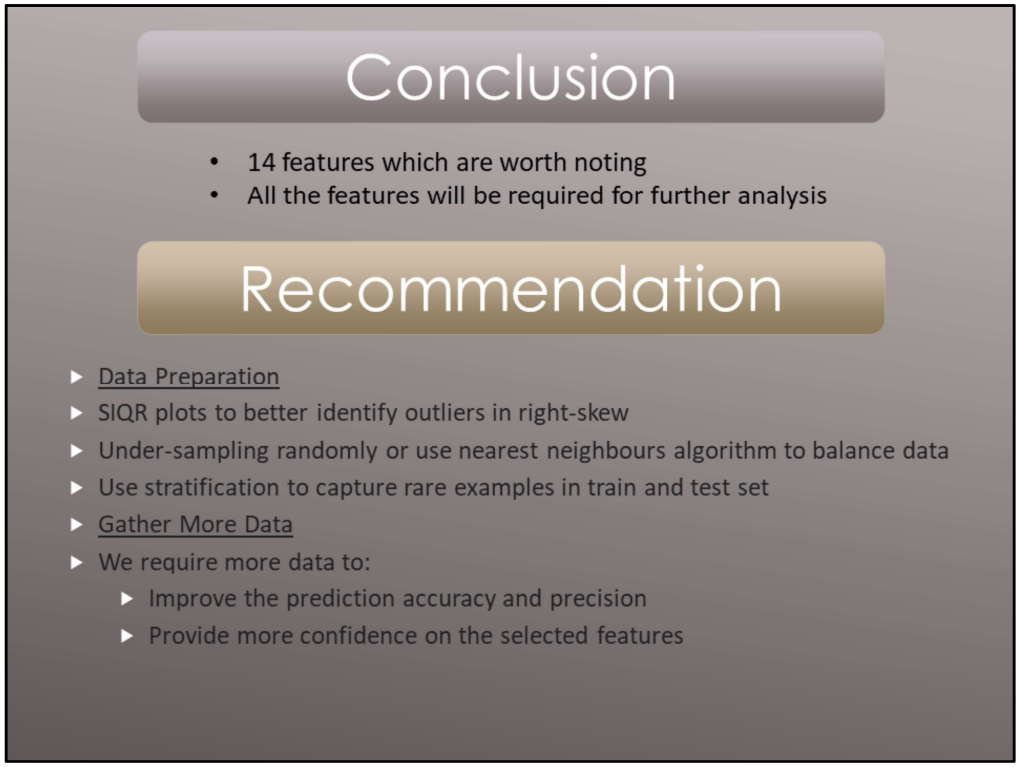
•*Note 1: Precision value indicates the ratio of correct positive predictions to the total values predicted positives.*

Recall values indicates ratio of correct positive predictions to the total actual positives

- *Note 2: Predictive Modelling analysis performed on Weka*



•Attributes such as; Installment percent, Duration in current address etc. were manually eliminated.
•These attributes were eliminated because they have no significant contribution in determining the customer creditability rating.
•Using Weka, both eliminated attributes were filtered out, leaving 14 attributes remaining in the selected features dataset.
•We made elimination decisions based on one-way ANOVA and Chi-Square method



Conclusion:

•The 14 features previously mentioned are worth noting, but because the accuracy calculated from using selected features is lower than using all the features, we suggest that all the features will be required for further analysis.

We used 2 method for feature selection; 1) statistical analysis 2) weka .AttributeSelection().

weka .AttributeSelection(), produced 3 selected features which overlaps with the 14 features using one-way ANOVA and Chi-Sq:

•Account Balance

•Duration of credit

•Payment Status of Previous Credit

Future Action:

•The accuracy is ~70%. We require more data to better predict the credibility of a client for loan.

- There is a slight imbalance in the data but we fear that the accuracy is only reflecting the underlying class distribution