GERMAN CREDIT CARD DATA SET

RaviKanth Putcha

Contents

Summary

The bank receives requests for loans and must decide the customer's eligibility based on their economic profile. There are two risks that the bank is exposed to. The first is related to accepting applications from individuals who will not be able to repay the loan. The second is related to missing out on potential client who have good credit.

To help the bank make a better decision and minimize the risk based on an applicant's profile, we applied different predictive models to analyze individuals (naive bayes, decision tree and logistic regression). We explored different ways to improve the performance of our classification models by introducing different learning parameters such as SMOTE, ensemble techniques and cost sensitive analysis. Finally, we compared all the algorithms based on several criteria which concluded that Naïve Bayes is the best model to use.

In addition to predictive analytics, we also performed post predictive analysis including clustering and association techniques in order to further analyze the bank's customers and their specific characteristics in relation to the class attribute.

We recommend to the bank that creditable borrowers are categorized as those who do not need guarantors, have no other loans/credits (and no instalment plans), have fewer dependents, own their apartments, have low savings and are thus motivated to borrow, pay off loans on time and want smaller loans. We also recommended to disapprove customers whose account balances are less than 100 DM or are over drafted, have poor valuable assets, are older and less skilled within their field. These customers live paycheck to paycheck and may have trouble repaying any additional debt. More in depth analysis will be provided throughout the report.

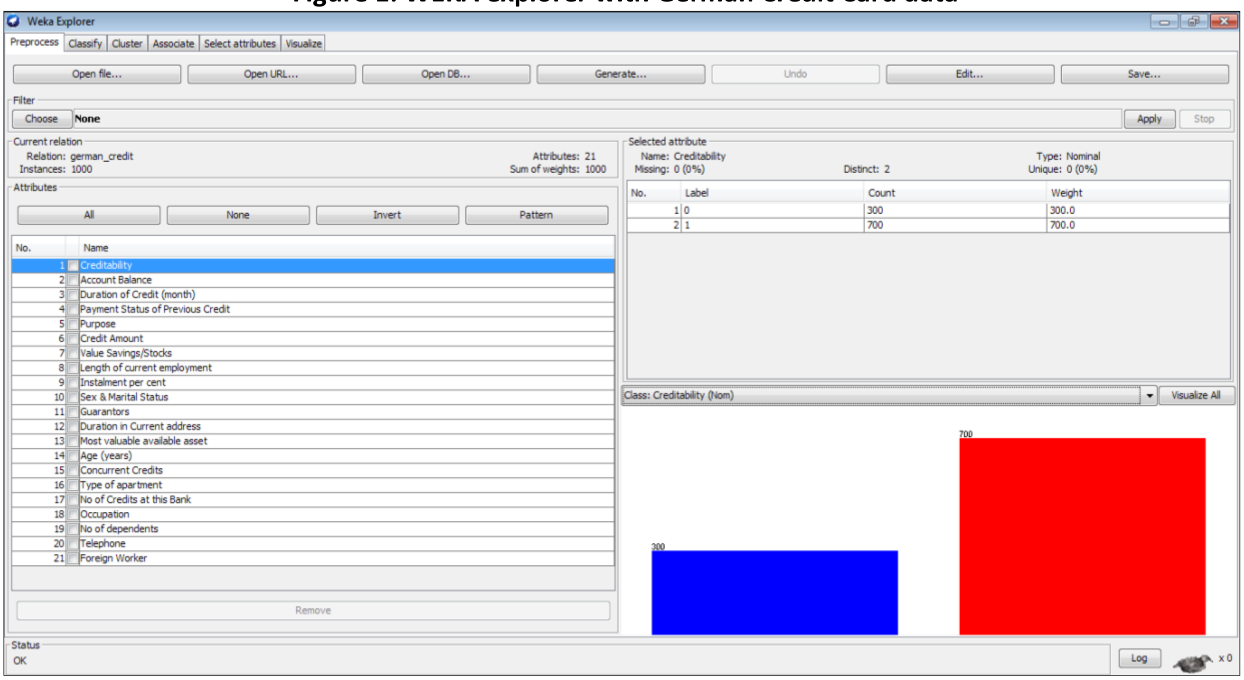
1. DATA PREPARATION

1.1. TYPES OF VARIABLES

The German Credit Card data set consisted of 1000 observations with 21 attributes. The attributes were classified into qualitative and quantitative data types. Qualitative data represents different types of data values such as gender, male and female whereas quantitative data represents numerical values such as test scores, weight etc.

For data preparation, the Credit Card dataset was loaded into Weka (Figure 1). The total number of high-risk individuals (300: shown in blue) is much lower than the amount of low risk individuals (700: shown in red). The data set contained 15 qualitative (10 Nominal and 5 Ordinal) and 6 quantitative (Numerical) data types. The attributes are described in Table 1. The new customers for credit are evaluated based on these 21 attributes.

Figure 1: WEKA explorer with German Credit Card data



1.2. ATTRIBUTE DISTRIBUTION

Table 1: Attribute Data Type

S. No	Attribute	Type	Sub Type
1	Creditability	Qualitative	Nominal
2	Purpose	Qualitative	Nominal
3	Sex & Marital Status	Qualitative	Nominal
4	Telephone	Qualitative	Nominal
5	Foreign Worker	Qualitative	Nominal
6	Occupation	Qualitative	Nominal
7	Type of apartment	Qualitative	Nominal
8	Guarantors	Qualitative	Nominal
9	Most valuable available asset	Qualitative	Nominal
10	Concurrent Credits	Qualitative	Nominal
11	Length of current employment	Qualitative	Ordinal
12	Payment Status of Previous Credit	Qualitative	Ordinal
13	Value Savings/Stocks	Qualitative	Ordinal
14	Duration in Current address	Qualitative	Ordinal
15	Account Balance	Qualitative	Ordinal
16	Duration of credit (month)	Quantitative	Numerical
17	Credit Amount	Quantitative	Numerical
18	Installment per cent	Quantitative	Numerical
19	Age (years)	Quantitative	Numerical
20	No. of Credits at this Bank	Quantitative	Numerical
21	No. of dependents	Quantitative	Numerical

1.3. NUMERICAL DATA SUMMARY

To obtain maximum, minimum, mean, median and standard deviation values, the data set was loaded into Weka and R. In Weka, the “selected attribute” panel lists the name of the selected attribute, data type, number of distinct values, missing values along with its statistics and their values (Figure 2).

In R, the “summary ()” function (> summary(data119)) was used to review each attribute in the dataset. The results are shown in Table 2.

Figure 2: Statistics for Numerical attributes in WEKA

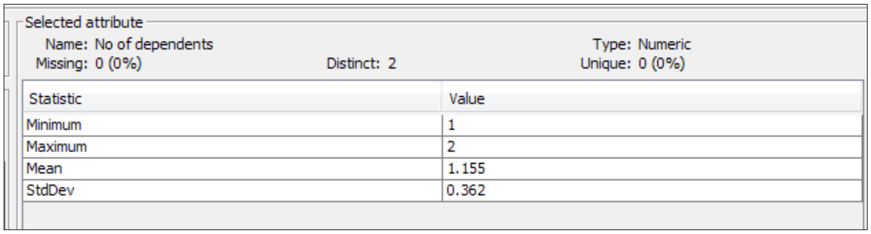
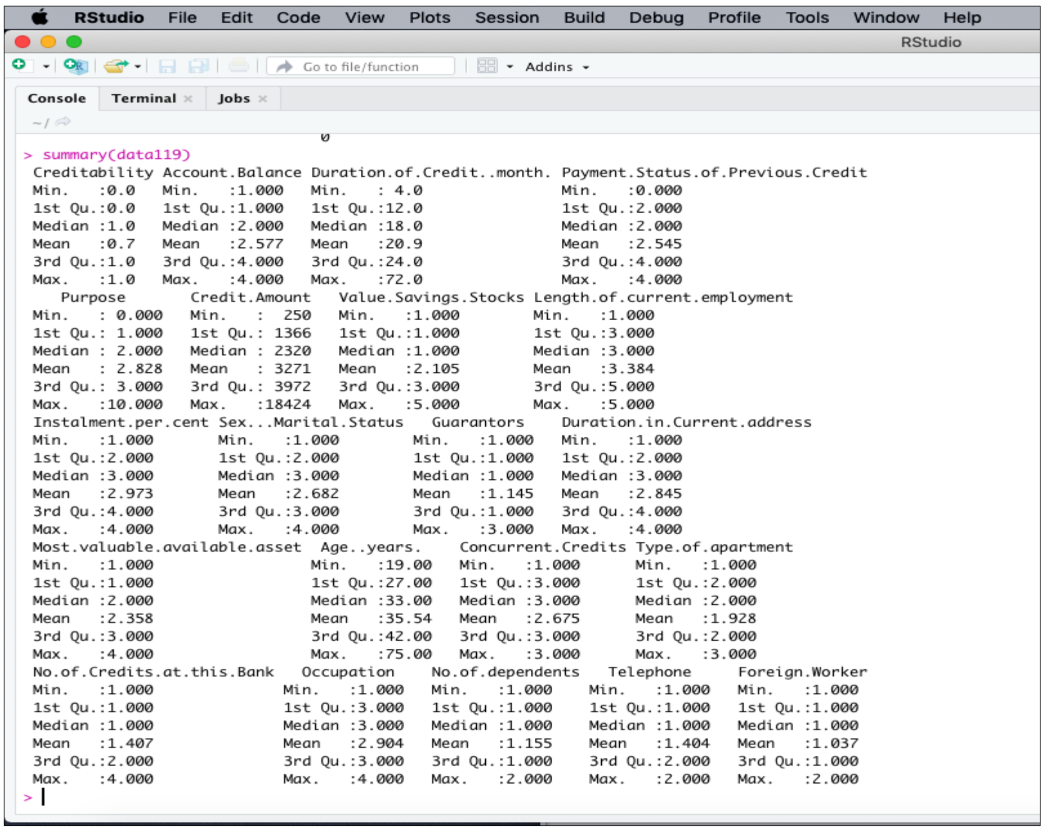


Figure 3: Summary Function in R




```
> summary(data119)
 Creditability Account.Balance Duration.of.Credit..month. Payment.Status.of.Previous.Credit
 Min.   :0.0   Min.   :1.000   Min.   : 4.0               Min.   :0.000
 1st Qu.:0.0   1st Qu.:1.000   1st Qu.:12.0               1st Qu.:2.000
 Median :1.0   Median :2.000   Median :18.0               Median :2.000
 Mean   :0.7   Mean   :2.577   Mean   :20.9               Mean   :2.545
 3rd Qu.:1.0   3rd Qu.:4.000   3rd Qu.:24.0               3rd Qu.:4.000
 Max.   :1.0   Max.   :4.000   Max.   :72.0               Max.   :4.000
    Purpose        Credit.Amount    Value.Savings.Stocks Length.of.current.employment
 Min.   : 0.000   Min.   :  250   Min.   :1.000        Min.   :1.000
 1st Qu.: 1.000   1st Qu.: 1366   1st Qu.:1.000        1st Qu.:3.000
 Median : 2.000   Median : 2320   Median :1.000        Median :3.000
 Mean   : 2.828   Mean   : 3271   Mean   :2.105        Mean   :3.384
 3rd Qu.: 3.000   3rd Qu.: 3972   3rd Qu.:3.000        3rd Qu.:5.000
 Max.   :10.000   Max.   :18424   Max.   :5.000        Max.   :5.000
 Instalment.per.cent Sex...Marital.Status   Guarantors    Duration.in.Current.address
 Min.   :1.000       Min.   :1.000        Min.   :1.000   Min.   :1.000
 1st Qu.:2.000       1st Qu.:2.000        1st Qu.:1.000   1st Qu.:2.000
 Median :3.000       Median :3.000        Median :1.000   Median :3.000
 Mean   :2.973       Mean   :2.682        Mean   :1.145   Mean   :2.845
 3rd Qu.:4.000       3rd Qu.:3.000        3rd Qu.:1.000   3rd Qu.:4.000
 Max.   :4.000       Max.   :4.000        Max.   :3.000   Max.   :4.000
 Most.valuable.available.asset  Age..years.    Concurrent.Credits Type.of.apartment
 Min.   :1.000                 Min.   :19.00   Min.   :1.000      Min.   :1.000
 1st Qu.:1.000                 1st Qu.:27.00   1st Qu.:3.000      1st Qu.:2.000
 Median :2.000                 Median :33.00   Median :3.000      Median :2.000
 Mean   :2.358                 Mean   :35.54   Mean   :2.675      Mean   :1.928
 3rd Qu.:3.000                 3rd Qu.:42.00   3rd Qu.:3.000      3rd Qu.:2.000
 Max.   :4.000                 Max.   :75.00   Max.   :3.000      Max.   :3.000
 No.of.Credits.at.this.Bank   Occupation    No.of.dependents   Telephone      Foreign.Worker
 Min.   :1.000              Min.   :1.000   Min.   :1.000    Min.   :1.000   Min.   :1.000
 1st Qu.:1.000              1st Qu.:3.000   1st Qu.:1.000    1st Qu.:1.000   1st Qu.:1.000
 Median :1.000              Median :3.000   Median :1.000    Median :1.000   Median :1.000
 Mean   :1.407              Mean   :2.904   Mean   :1.155    Mean   :1.404   Mean   :1.037
 3rd Qu.:2.000              3rd Qu.:3.000   3rd Qu.:1.000    3rd Qu.:2.000   3rd Qu.:1.000
 Max.   :4.000              Max.   :4.000   Max.   :2.000    Max.   :2.000   Max.   :2.000
> 
```

Table 2: Numerical Data Summary

Attribute	Maximum	Minimum	Mean	Median	Std. Deviation
Duration of credit (month)	72	4	20.903	18	12.059
Credit Amount	18424	250	3271.25	2320	2822.752
Installment per cent	4	1	2.973	3	1.119
Age (years)	75	19	35.542	33	11.353
No. of Credits at this Bank	4	1	1.407	1	0.578
No. of dependents	2	1	1.155	1	0.362

From the statistics summary in Table 2 and Figure 3, it can be concluded that:

1. **Duration of Credit**: The average duration of credit is 20.9 months which is 21 months approximately.
2. **Credit Amount**: Mean>Median, the data is right skewed which means there are more people with smaller amounts of credit. Maximum amount for the loan is $18,424.
3. **Instalment per cent**: Mean<Median, the data is slightly left skewed which means there are more people who have 3-4% instalment rate.
4. **Age:** Mean>Median, the data is right skewed which means younger people are more likely to apply for loans (early 30s).
5. **No. of Credits at this Bank**: The minimum and median data points are equal. The data is right skewed (Mean>median), there is more than 50% of the customers who only have 1 credit at this bank.
6. **No. of Dependents**: The Minimum, 1st quartile, median and 3rd quartile are all 1, so 75% or more people have only 1 dependent.

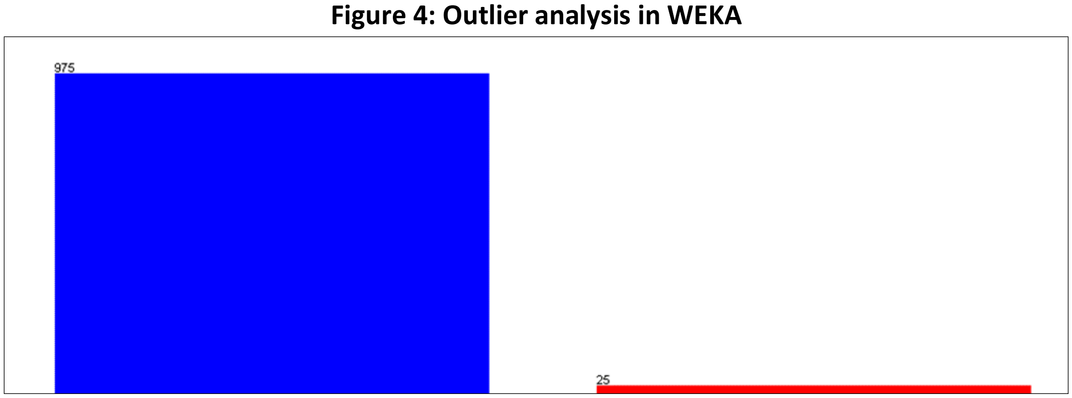
1.4. OUTLIER AND MISSING VALUE

To identify and analyze outliers and missing value, the InterQuartileRange (IQR) feature in Weka was applied on the dataset. This gave an additional attribute called Outlier (Nominal type), which had two distinct values (Yes or No). In the Credit Card dataset, there were 975 noes (if an instance was not an outlier then IQR assigned No as a label) and only 25 yeses if it was an outlier. The attributes with outliers were as follows:

1. Duration of Credit
2. Age
3. Credit Amount
4. No of Credits at this Bank
5. Instalment Per cent
6. No of dependents

For this project, the outliers were not removed as only 2.5% of the entire data set had outliers. Since the data set is small (contains only 1,000 observations) and most of the data is of nominal type, the small percentage of outliers will not affect our expirement significantly. In addition, the outliers in our data are not sampling errors. As a result, we believe the outliers still provide important information for the bank and removing them will result in losing essential data.

Figure 4: Outlier analysis in WEKA



1.5. DISTRIBUTION OF ATTRIBUTES

A Histogram is a bar chart based on the frequency distribution of the data. The x-axis shows the categories and y-axis shows the frequency of those categories. Below is a summary of the distribution of each attribute from Weka (Figure 5) and R (Figure 6,7,8 and 9). Based on these distributions it can be concluded that:

1. **Age:** People who are below the age of 40 tend to apply for a loan. The histogram is skewed to the right which means the early part of the histogram has higher frequency. **[Figure 6]**
2. **Credit Amount:** People tend to seek a credit amount less than $8,000. The graph appears to be skewed to the right. **[Figure 7]**
3. **No of dependents:** People with one dependent tend to seek loan as compared to people with more than one dependent. **[Figure 8]**
4. **No of Credits at this Bank:** Population with just one existing loan tend to apply for a new loan. **[Figure 9]**
5. **Sex & Marital Status:** Males who are single, constituted most of the population who had applied for loan.
6. **Occupation:** Skilled employees constituted most of the population who applied for a loan.

Figure 5: Histogram of all 21 attributes in WEKA

Figure 6: Age Distribution Figure

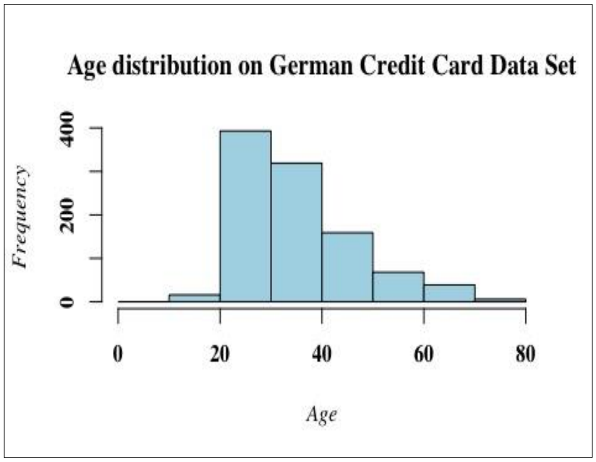


Figure 7: Credit Amount Distribution

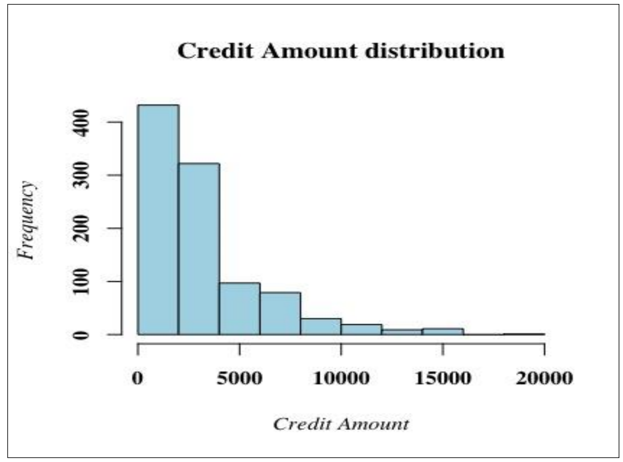


Figure 8: No. of Dependents Distribution

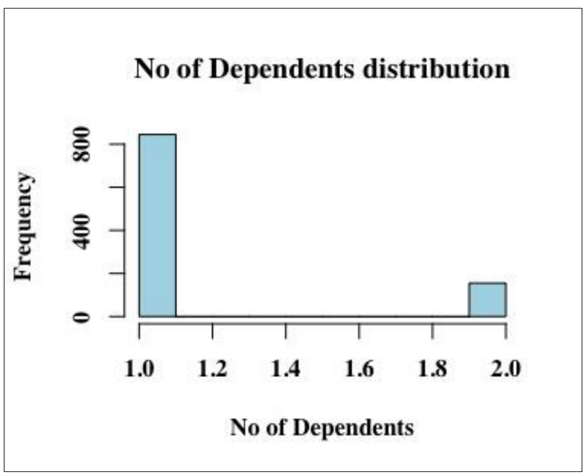
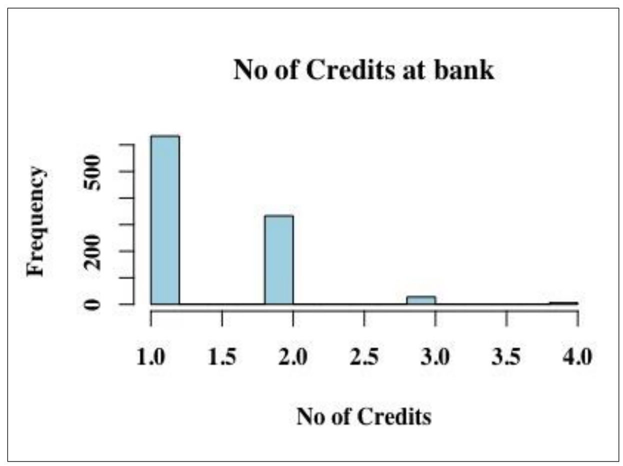


Figure 9: No. of Credits at Bank



1.6. IMBALANCE CLASS DISTRIBUTION

1.6.1. IMBALANCE FEATURE

The total number of high-risk individuals (300) is much lower than the amount of low risk individuals (700). As we apply predictive analytics techniques, we may encounter bias within our model due to the imbalanced dataset. As a result, it is important to balance the dataset in order to build a successful analytical model. Below are multiple methods that can be used to fix this issue. Each strategy has its own variations and will be analyzed in much greater depth as required in the predictive modeling section of the document.

1.6.2. RESAMPLING

Resampling is a widely used technique in data science to balance a dataset by either eliminating instances from the majority class or adding instances to the minority class.

1. Oversampling aims to add random instances to the minority class in order to achieve a more balanced dataset.

    - Random Oversampling
    - SMOTE

2. Undersampling aims to delete records from the majority class when the dataset is large enough. Through careful analysis, we decided not to pursue this method as our total number of observations are low. Removing any information will cause the model to lose important instances that will hinder the performance of the model.

1.6.3. ALGORITHMIC ENSEMBLE TECHNIQUE

Algorithmic ensemble techniques aim to modify existing algorithm classifications to ensure the model incorporates and takes into consideration an imbalanced dataset

1. Bagging Based techniques
2. Boosting-Based techniques
3. Gradient Tree Boosting techniques
4. Adaptive Boosting- Ada Boost techniques

1.7. FEATURE SELECTION

Selecting attributes refers to narrowing down the potential list of variables in order to only use the most influential ones. To determine which attributes could be eliminated or included in the analysis, the following list of feature selection methods were performed in Weka (Table 3):

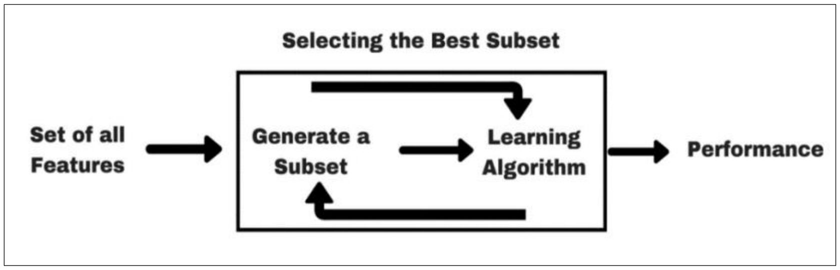
1. **Correlation Based Feature Selection (CorrelationAttributeEval):** This feature evaluates the correlation of all attributes versus the class attribute.

2. **Information Gain Based Feature Selection (InfoGainAttributeEval):** Evaluates the worth of an attribute by measuring the information gain with respect to the class attribute. The analysis is performed by using the Ranker InfoGainAttributeEval method. Table 3 illustrates the ranks of all attributes and their associated normalized values.

3. **Learner Based Feature Selection (Wrapper Method - J48):** This feature evaluates the performance of a subset of features based on the resulting performance of the applied learning algorithm (e.g. what is the gain in accuracy for a classification problem). The WrapperSubsetEval technique was chosen with BestFirst Search Method. As a result, five attributes were selected:

    1. Account Balance
    2. Value Savings/Stocks
    3. Duration of Credit (Month)
    4. Purpose
    5. Payment Status of Previous Credit

Figure 10: Wrapper Method



4. **Chi-Square Evaluator:** A Chi-Square test is used in statistics to test the independence of two events. When the observed count is close to the expected count, the Chi-Square value is small and therefore fits the null hypothesis of independence. Likewise, a higher Chi-Square (above the critical value) indicates that the null hypothesis is rejected, and the features are dependent.

5. **Symmetrical Uncertainty Evaluation:** Symmetrical uncertainty criterion overcomes the bias of information by normalizing the dataset

Table 3: Feature Selection Numerical Data Summary

Attribute	Correlation	Information Gain	Learner based - J48	Chi-Square	Symmetrical Uncertainty
Account Balance	0.23276	0.094739	1	123.7209	0.070613
Duration of Credit (month)	0.21493	0.0329	2	46.8311	0.030272
Value Savings/Stocks	0.13162	0.028115	4	36.0989	0.021887
Payment Status of Previous Credit	0.08988	0.043618	3	61.6914	0.033641
Credit Amount	0.15474	0.018333		26.3992	0.021448
Type of apartment	0.12283	0.013077		18.674	0.012963
Purpose	0.07494	0.024894		33.3564	0.014033
Length of current employment	0.0527	0.013102		18.3683	0.00863
Instalment per cent	0.0724	0		0	0
Sex & Marital Status	0.07192	0.006811		9.6052	0.005644
Most valuable available asset	0.05838	0.016985		23.7196	0.012008
Age (years)	0.09127	0.011278		16.3681	0.014251

Concurrent Credits	0.108	0.008875		12.8392	0.010284
Occupation	0.01904	0.001337	5	1.8852	0.001166
Foreign Worker	0.08208	0.005823	6	6.737	0.010495
No of dependents	0.00301	0		0	0
No of credit at this Bank	0.04573	0		0	0
Guarantors	0.00612	0.00479		6.6454	0.006758
Duration in current Address	0.01096	0.000543		0.7493	0.000398
Telephone	0.03647	0.000964		1.3298	0.001039

2. PREDICTIVE MODELING / CLASSIFICATION

2.1. INTRODUCTION

The aim of the project is to help bank managers make the right decision regarding loan applications. Predictive modeling is a process of using historical data, machine learning and AI to predict future output. There are different types of predictive models including classification and clustering, which will be further discussed.

Before conducting our experiment, the data was randomized using the randomization filter in Weka in order to remove any sampling error while training and testing our models.

2.1.1. CLASSIFICATION

Classification is a two-step process, a learning step followed by a testing step. During the learning phase, the algorithm is fed a set of inputs, which it uses to effectively build the model. During the testing phase, the model is used to make predictions.

1. **Independent variable:** An independent variable is a condition that changes in an experiment. It is the input variable.

2. **Dependent Variable:** A Dependent variable represents the output associated with changing the inputs/independent variables. In this dataset, credibility (class attribute) is the dependent variable.

The following three classification algorithms were used:

1. Naïve Bayes
2. Decision Tree
3. Logistic Regression

2.1.2. SMOTE OVERSAMPLING

The SMOTE function handles unbalanced datasets by generating synthetic observations. By using the nearest k (5) neighbours, the function is able to generate 300 additional records (synthetic data) relating to class 0. The total number of observations therefore increases to 1300 (700 good credit/600 bad credit)

Figure 11: Dataset before oversampling (Class 0 - 300 records vs Class 1 - 700 records)

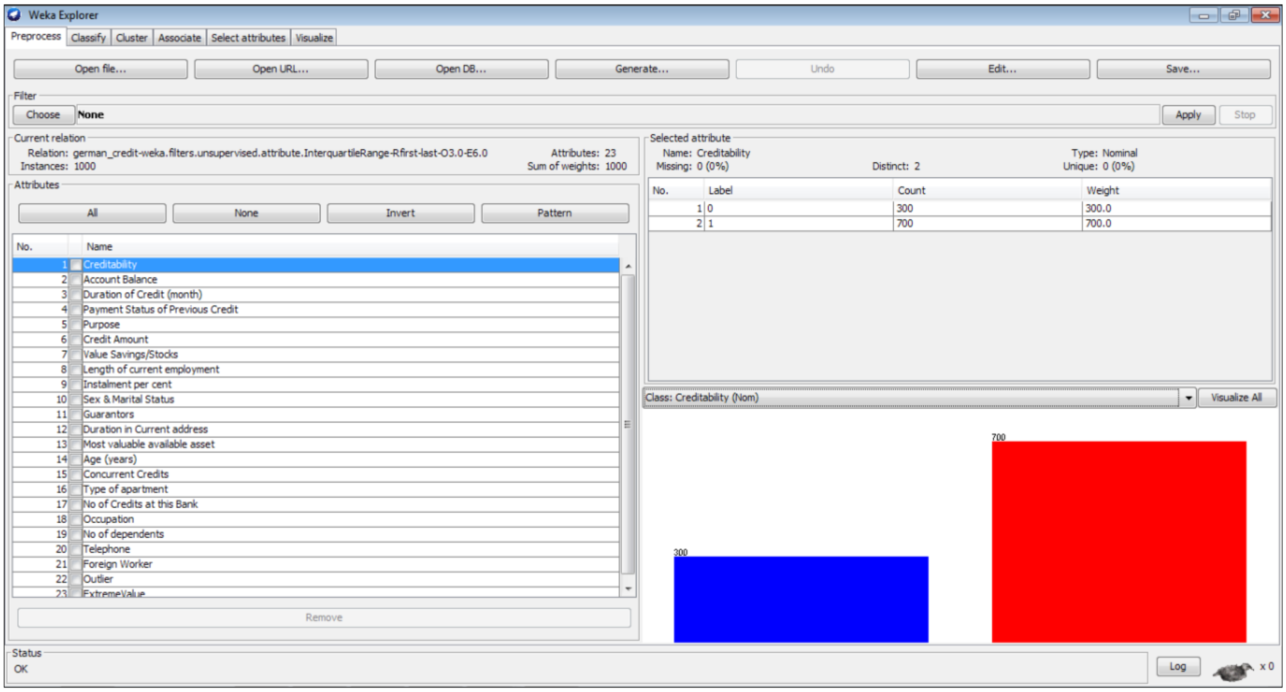


Figure 12: SMOTE Filter

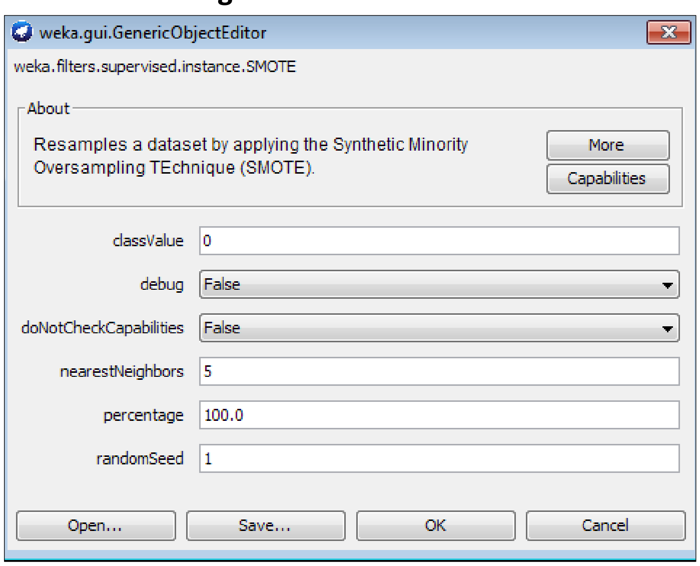
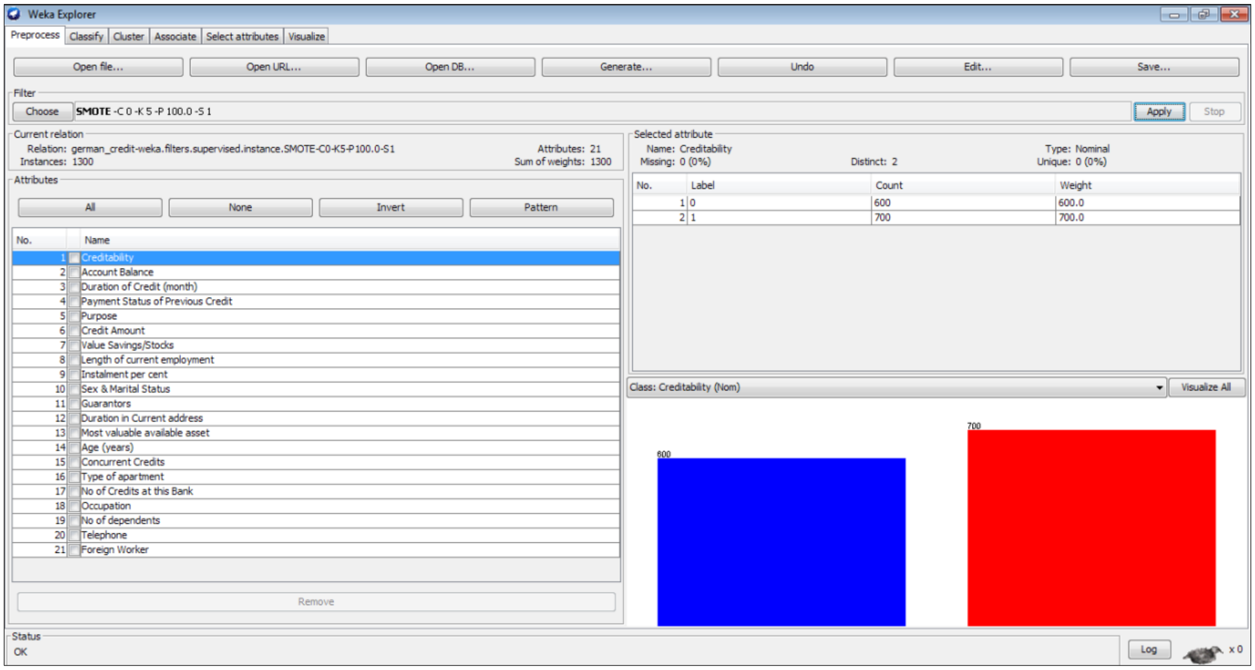


Figure 13: Oversampling using SMOTE filter (Class 0 - 600 records vs Class 1 - 700 records)



2.1.3. TRAINING AND VALIDATION OF DATASET

Cross-validation is a statistical method used to evaluate machine learning models. The general procedure is as follows:

1. Shuffle the dataset randomly
2. Split the dataset into k groups
3. For each unique group:
    - 3.1 Take the group as a hold out or test data set
    - 3.2 Take the remaining groups as a training data set
    - 3.3 Fit a model on the training set and evaluate it on the test set
    - 3.4 Retain the evaluation score and discard the model
4. Summarize performance of the model using the sample of model evaluation scores

Percentage split separates the data into two parts, one to train and the other to test. This is a good method to use when using a slow algorithm. Each set was parameterised using a random selection of 900 or 600 observations and then tested on the remaining 100 or 400 accordingly (90-10% and 60-40%.)

2.2. NAÏVE BAYES CLASSIFIER

2.2.1. INTRODUCTION

A Naïve Bayes classifier is a probability-based machine learning model which is used for classification. The classifier is based on the Bayes theorem.

Equation 1: Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Using Bayes theorem, we can find the probability of event A happening given that event B has occurred. Here, B is the evidence and A is the hypothesis.

The fundamental Naïve Bayes assumptions are:

1. Features are independent
2. Features have equal contribution to the outcome.

2.2.2. ADVANTAGES OF USING NAÏVE BAYES CLASSIFIER

1. Naïve Bayes is a fast and easy classifier to predict a class of a test data set.
2. Naïve Bayes Classifier performs better compared to other models assuming independence
3. It performs well in case of categorical input variables compared to numerical variables

2.2.3. RESULTS

Table 4: Unbalanced Data with 20 Attributes (without SMOTE)

Algorithm (Naive Bayes)	Cross Validation, K= 10,	Percentage Split, 60%	Percentage Split, 90%
Correctly Classified instances (%)	75.4%	74.5%	75%
Incorrectly Classified instances (%)	24.6%	25.5%	25%
F-Score	0.746	0.726	0.733
Precision rate	0.743	0.737	0.756
Recall rate	0.754	0.745	0.750

Table 5: Unbalanced Data with 6 Attributes (without SMOTE)

Algorithm (Naive Bayes)	Cross Validation, K= 10,	Percentage Split, 60%	Percentage Split, 90%
Correctly Classified instances (%)	75.4%	74.75%	73%
Incorrectly Classified instances (%)	24.6%	25.25%	27%
F-Score	0.738	0.719	0.698
Precision rate	0.739	0.750	0.754
Recall rate	0.754	0.748	0.730

Table 6: Balanced Data with 20 Attributes (with SMOTE)

Algorithm (Naive Bayes)	Cross Validation, K= 10,	Percentage Split, 60%	Percentage Split, 90%
Correctly Classified instances (%)	80%	80.9%	83.8%
Incorrectly Classified instances (%)	20%	19.8%	16.1%
F-Score	0.800	0.803	0.839
Precision rate	0.800	0.804	0.840
Recall rate	0.800	0.802	0.838

Table 7: Balanced Data with 6 Attributes (with SMOTE)

Algorithm (Naive Bayes)	Cross Validation, K= 10,	Percentage Split, 60%	Percentage Split, 90%
Correctly Classified instances (%)	78.5%	80.9%	83.8%
Incorrectly Classified instances (%)	21.4%	19.8%	16.1%
F-Score	0.786	0.803	0.838
Precision rate	0.786	0.804	0.838
Recall rate	0.785	0.802	0.838

2.2.4. CONCLUSION FOR NAÏVE BAYES CLASSIFICATION MODEL

We implemented the Naïve Bayes classifier to evaluate the credit risk in our dataset. A total of 12 tests were performed. 6 tests were performed on the unbalanced dataset and 6 were performed on the dataset that was balanced using the SMOTE function. Within these tests, half of them were performed using all 20 attributes and the other half using only the top 6 attributes.

Two types of training/testing techniques were used, the 10-fold test set and percentage split; The data was split into a 60:40 and 90:10 ratio i.e. 90% of the data was used to train the model and 10% of the data is used to validate the model.

In our first test, we used all 20 attributes and an unbalanced dataset. A maximum accuracy of 75.4% was achieved using a 10-fold test. 74.5% and 75% accuracy levels were achieved using a percentage split of 60-40% and 90-10% respectively. In the second test, we used an unbalanced dataset but with the top 6 attributes only. The maximum accuracy was again 75.4% with a 10-fold test. 74.75% and 73% was achieved using a 60-40% split and 90-10% split respectively. Overall, there was not much a difference observed in our unbalanced dataset for all 20 attributes as compared to the top 6 attributes.

In the next test, we used a balanced dataset with 20 attributes and our accuracy observed was 83.8% with a 90% percentage split strategy. This level of accuracy is the highest observed as of now. Similar results were observed when we used a balanced dataset with 6 attributes. 83.8% accuracy was observed with a 90-10% split and 78.5% accuracy using a 10-fold test set. It is clear that a 90-10% split yields a better result. However, it is important to note that this may be occurring due to the concept of overfitting.

In conclusion, a balanced dataset using SMOTE and 20 attributes is preferable. Due to the overfitting issue, it is recommended to use a 60-40% split or a 10-fold strategy. These strategies yield a slightly lower accuracy than a 90-10% split, but are a better guide for our model

2.2.5. SAMPLE OUTCOME OF NAÏVE BAYES

Figure 14: Balanced Data with 20 Attributes (Percentage Spilt 90%)

```
=== Summary ===

Correctly Classified Instances         109               83.8462 %
Incorrectly Classified Instances        21               16.1538 %
Kappa statistic                          0.6727
Mean absolute error                      0.2558
Root mean squared error                  0.3697
Relative absolute error                 51.6677 %
Root relative squared error             74.4834 %
Total Number of Instances              130

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.839    0.162    0.797      0.839    0.817      0.673    0.868     0.828     0
                 0.838    0.161    0.873      0.838    0.855      0.673    0.868     0.866     1
Weighted Avg.    0.838    0.161    0.840      0.838    0.839      0.673    0.868     0.850

=== Confusion Matrix ===

  a  b   <-- classified as
 47  9 |  a = 0
 12 62 |  b = 1
```

2.3. DECISION TREE

Decision Tree is a supervised learning algorithm. It is used for solving regression and classification problems. **J48 classifier was used in Weka** for decision tree. Data is split into nodes and leaves, where leaves can be used to predict the output. In all our decision trees, account balance was the root node and we can conclude that this parameter had the highest information gain among all the parameters.

2.3.1. ATTRIBUTE SELECTION

As explained in the feature selection section, various tests were performed to select the attributes that could yield the best results. We conducted tests using J48 classifier by choosing all 20 independent variable and also used the below attributes to predict the output.

The top 6 attributes to predict the class variable:

1. Account Balance
2. Credit Amount
3. Duration of Credit (month)
4. Value Savings/Stocks
5. Payment Status of Previous Credit
6. Most valuable available asset

2.3.2. PRUNING

Decision trees tend to perform well with training data but perform poorly on testing data. This can be corrected using pruning. In Weka the **unpruned** parameter is set to False. This means that pruning is enabled by default. **ConfidenceFactor** is also a parameter that effects the size of the tree. A small confidence value corresponds to heavy pruning and vis-versa. The table below shows the impact of changing the confidence factor on accuracy and the size of the tree.

Table 8: Accuracy and Number of leaves

Confidence factor	Accuracy	Comments
0.25	72.8	Number of Leaves: 99 Size of the tree: 136
0.5	72.1	Number of Leaves: 180 Size of the tree: 245
0.9	70.2	Number of Leaves: 336 Size of the tree: 439

2.3.3. SMOTE: Synthetic Minority Over-sampling Technique (SMOTE)

The Smote technique was used to create synthetic observations using 5 nearest neighbors. This technique was used in some of the tests to balance the dataset. There were 1300 observations in total, 600 are in class 0 and 700 in class 1

2.3.4. ENSEMBLE METHODS

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking). Decision trees may suffer from variance and bias issues and as a result we decided to include bagging within our model to improve our results.

Bagging was achieved by using **Bagging Classifier in Weka** and selecting J48 as classifier.

2.3.5. RESULTS & O/P PARAMETERS CONSIDERED TO DETERMINE PERFORMANCE OF DECISION TREE

Below is the list of parameters that were used to evaluate the model

1. Accuracy
2. Weighted Precision
3. Weighted Recall
4. Weighted F measures

Below are the results of all the tests that were conducted. We chose accuracy as the primary factor to determine the performance of our decision tree classifier.

Table 10: Decision Tree Tests and Results

No	Parameters Used	No of Attributes	Test Data Method	Test Data Size	Accuracy (%)	Weighted Precision	Weighted Recall	Weighted F Measure	Confusion Matrix
1	J48 M2 C0.25	20	Cross Validation - 10	1000	72.8	0.712	0.728	0.716	a b <-- classified as 128 172 | a = 0 100 600 | b = 1
2	J48 M2 C0.25	20	Percentage Split - 60%	400	71.75	0.706	0.718	0.687	a b <-- classified as 45 90 | a = 0 23 242 | b = 1
3	weka.classifiers.meta.**Bagging** -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2	20	Percentage Split - 60%	400	70.5	0.691	0.705	0.666	a b <-- classified as 38 97 | a = 0 21 244 | b = 1
4	weka.classifiers.meta.**Bagging** -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2	20	Cross Validation - 10	1000	74.3	0.729	0.743	0.731	a b <-- classified as 134 166 | a = 0 91 609 | b = 1
5	weka.classifiers.trees.J48 -C 0.25 -M 2 (with **SMOTE**)	20	Cross Validation - 10	1300	75.2308	0.752	0.752	0.752	a b <-- classified as 430 170 | a = 0 152 548 | b = 1
6	weka.classifiers.trees.J48 -C 0.25 -M 2 (with **SMOTE**)	20	Percentage Split - 60%	520	78.4615	0.792	0.785	0.786	a b <-- classified as 177 42 | a = 0 70 231 | b = 1
7	weka.classifiers.meta.**Bagging** -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2 (With **SMOTE**)	20	Percentage Split - 60%	520	**79.4231**	0.798	0.794	0.795	a b <-- classified as 175 44 | a = 0 63 238 | b = 1
8	weka.classifiers.meta.**Bagging** -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2 (With **SMOTE**)	20	Cross Validation - 10	1300	78.8462	0.778	0.758	0.768	a b <-- classified as 455 145 | a = 0 130 570 | b = 1
9	weka.classifiers.trees.J48 -C 0.25 -M 2	6	Cross Validation - 10	1000	72.8	0.712	0.728	0.716	a b <-- classified as 128 172 | a = 0 100 600 | b = 1
10	weka.classifiers.trees.J48 -C 0.25 -M 2	6	Percentage Split - 60%	400	70.5	0.691	0.705	0.666	a b <-- classified as 38 97 | a = 0 21 244 | b = 1
11	weka.classifiers.meta.**Bagging** -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2	6	Cross Validation - 10	1000	74.3	0.729	0.743	0.731	a b <-- classified as 134 166 | a = 0 91 609 | b = 1
12	weka.classifiers.meta.**Bagging** -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2	6	Percentage Split - 60%	400	70.5	0.691	0.705	0.666	a b <-- classified as 38 97 | a = 0 21 244 | b = 1

<table>
<tr><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th></tr>
<tr><td>13</td><td>weka.classifiers.trees.J48 -C 0.25 -M 2 (with SMOTE)</td><td>6</td><td>Cross Validation - 10</td><td>1300</td><td>77.6154</td><td>0.777</td><td>0.776</td><td>0.776</td><td>a b <-- classified as
464 136 | a = 0
155 545 | b = 1</td></tr>
<tr><td>14</td><td>weka.classifiers.trees.J48 -C 0.25 -M 2 (with SMOTE)</td><td>6</td><td>Percentage Split - 60%</td><td>520</td><td>77.1154</td><td>0.781</td><td>0.771</td><td>0.773</td><td>a b <-- classified as
177 42 | a = 0
77 224 | b = 1</td></tr>
<tr><td>15</td><td>weka.classifiers.meta.Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2 (With SMOTE)</td><td>6</td><td>Percentage Split - 60%</td><td>520</td><td>78.4615</td><td>0.791</td><td>0.785</td><td>0.786</td><td>a b <-- classified as
176 43 | a = 0
69 232 | b = 1</td></tr>
<tr><td>16</td><td>weka.classifiers.meta.Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2 (With SMOTE)</td><td>6</td><td>Cross Validation - 10</td><td>1300</td><td>77.6154</td><td>0.776</td><td>0.776</td><td>0.776</td><td>a b <-- classified as
448 152 | a = 0
139 561 | b = 1</td></tr>
</table>

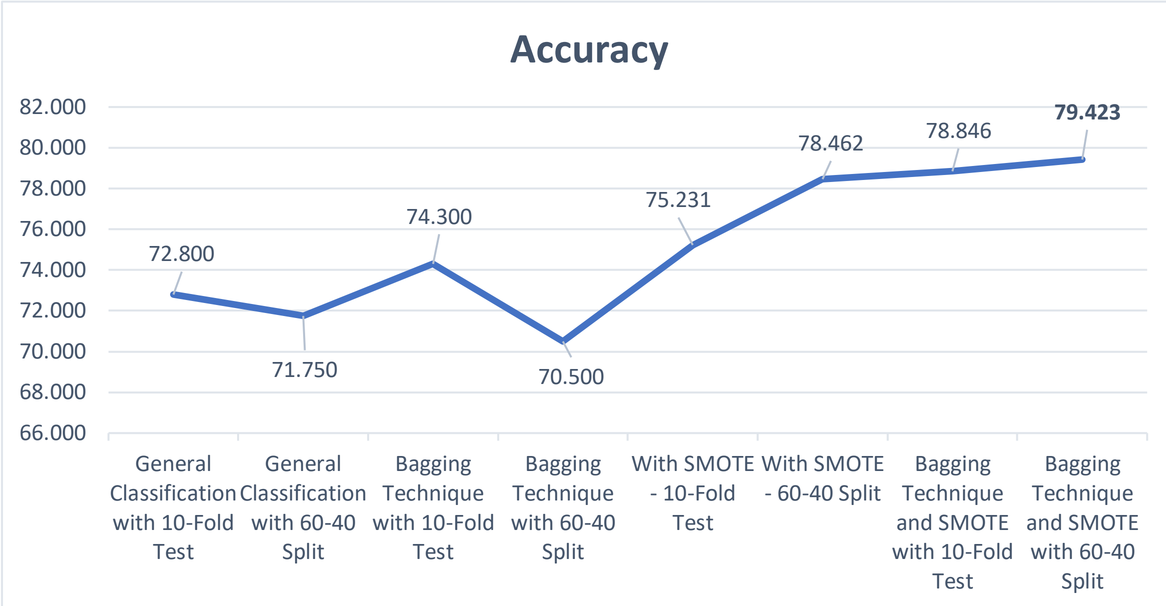
Table 10: Results of tests conducted on 20 Attributes

Parameters Used	No of Attributes	Accuracy
General Classification with 10-Fold Test	20	72.80%
General Classification with 60-40 Split	20	71.75
Bagging Technique with 10-Fold Test	20	74.3
Bagging Technique with 60-40 Split	20	70.5
With SMOTE - 10-Fold Test	20	75.2308
With SMOTE - 60-40 Split	20	78.4615
Bagging Technique and SMOTE with 10-Fold Test	20	78.8462
Bagging Technique and SMOTE with 60-40 Split	20	**79.4231**

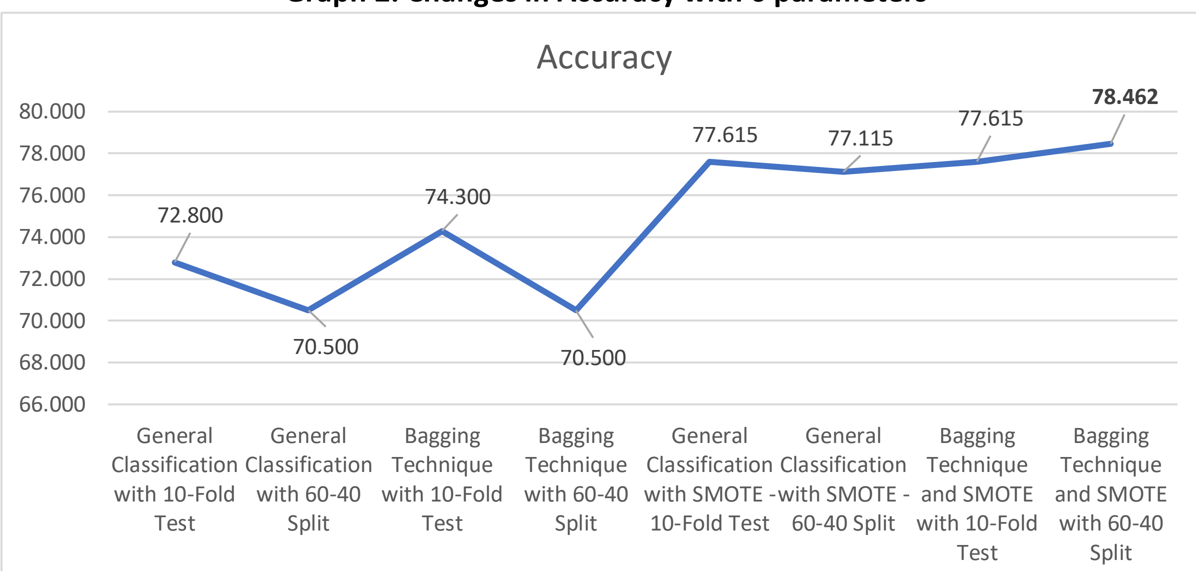
Table 11: Results of tests conducted on 6 Attributes

Parameters Used	No of Attributes	Accuracy
General Classification with 10-Fold Test	6	77.6154
General Classification with 60-40 Split	6	77.1154
Bagging Technique with 10-Fold Test	6	77.6154
Bagging Technique with 60-40 Split	6	**78.4615**
With SMOTE - 10-Fold Test	6	74.2
With SMOTE - 60-40 Split	6	74.5
Bagging Technique and SMOTE with 10-Fold Test	6	74.2
Bagging Technique and SMOTE with 60-40 Split	6	74.75

Graph 1: Changes in Accuracy with 20 parameters



Graph 2: Changes in Accuracy with 6 parameters



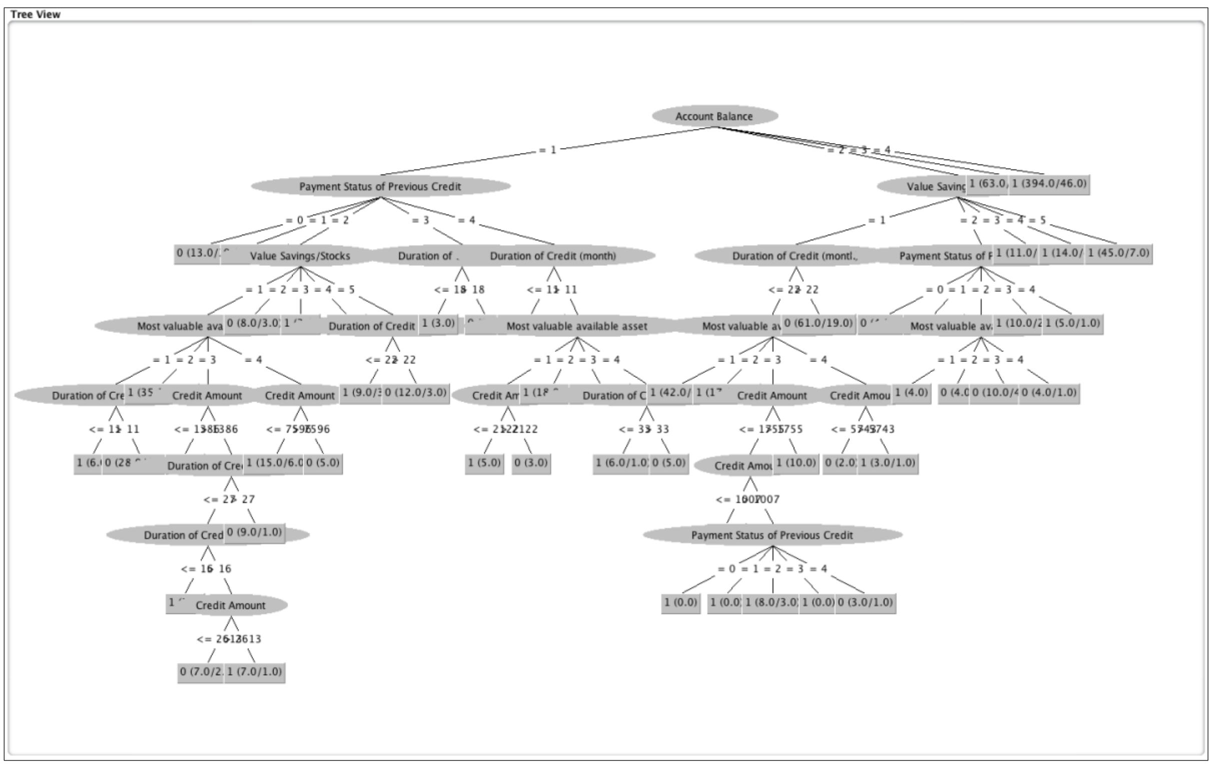
2.3.6. CONCLUSIONS FOR DECISION TREE CLASSIFICATION MODEL

A total of 16 tests were performed. Firstly, we used all 20 independent variables to construct a decision tree. Two types of testing techniques were used, the 10-fold test and percentage split, where data was split into a ratio of 60:40 i.e. 60% of the data was used to training the model and 40% of the data was used test the model. We achieved an accuracy of 72.8% with 10-fold test and 71.8% when percentage split was used. Next we carried the same tests but with a balanced dataset. The results showed that Introducing SMOTE improved the accuracy of the model. 78.4% accuracy levels were achieved using J48 algorithm on a total of 520 observations (60-40% split on a 1300 instance dataset). However, Introducing SMOTE may have caused the model to overfit. In order to fix this issue, we applied an ensemble technique (bagging). The result of applying bagging on top of SMOTE yielded the best results with an accuracy level of 79.4%.

The same tests were then carried out by reducing the attributes to 6. The top 6 attributes were chosen by performing feature selection tests as part of our data preparation. Very similar results as that of 20 attributes were observed. Highest accuracy of 78.4% was also achieved by using a combination of bagging + SMOTE + percentage split test. We can conclude that reducing the number of parameters to 6 did not cause any improvement in the classification of the model.

2.3.7. DECISION TREE SAMPLE

Figure 15: Decision Tree in Weka

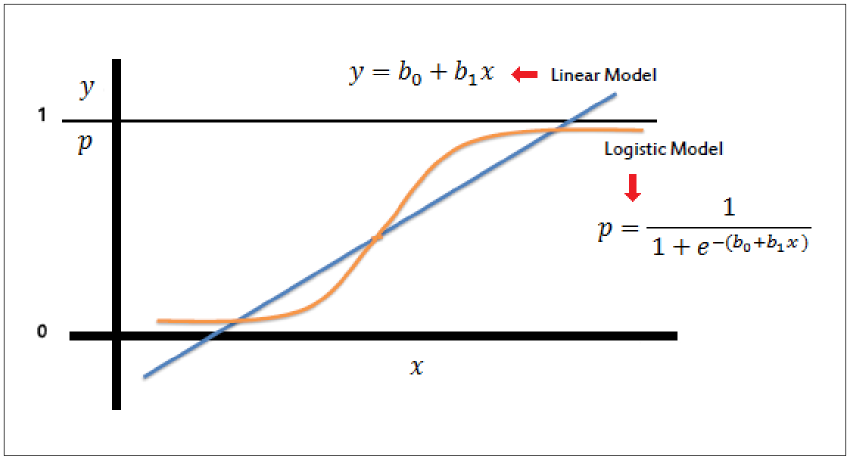


2.4. LOGISTIC REGRESSION

As a third algorithm, we decided to use logistic regression. Much like linear regression, logistic regression is a linear method, but the predictions are transformed using the logistic function in order to get a probability between 0 and 1. A threshold is then decided upon to classify each instance into either class (good credit/bad credit). We decided to keep the threshold at 0.5.

Consequently, any output of the logistic function that returns a value greater than 0.5 will be classified as good credit individual

Figure 16: Logistic Model



2.4.1. ATTRIBUTE SELECTION AND TRAINING STRATEGY

We decided to run our model multiple times, each time by altering a different parameter. Firstly, the experiment included all 20 attributes followed by the main 6 attributes chosen. Each time, the data was split using two strategies. Finally, cost sensitive learning will be applied as a measure to reduce the imbalance feature of our dataset.

Table 12: Training Strategy and top 6 Attributes

Top 6 attributes	Training Strategy
Account Balance	10-fold cross validation
Duration of Credit (month)	60-40 percentage split
Payment Status of Previous Credit	
Credit Amount	
Value Savings/Stocks	
Most valuable available asset	

2.4.2. COST SENSITIVE LEARNING

In machine learning, most learning algorithms assume that miss classification costs made by a classifier are equal. However, in our situation, this is not the case. As a bank, the cost of lending out a loan to a bad customer (false positive) can result in much greater losses than denying a loan to a good customer (false negative). Cost sensitive learning is a subfield of machine learning that takes into consideration the cost of prediction errors when training the model. When training our model, we will be changing our cost matrix by altering the false positive error weight from 1 to 2. With a goal to **minimize the bank's error costs**, our training algorithm will be much more careful in identifying bad credit individuals. Below are the results split into two tables:

Table 13: Results of Logistic Regression on 20 Attributes

S. No	Parameters Used	No of Attributes	Cost Matrix	Test Data Method	Test Data Size	Accuracy (Correctly Classified Instances)	Weighted Precision	Weighted Recall	Weighted F Measure	Confusion Matrix
1	Logistic regression	20	NA	10-fold	1000	75.20%	0.742	0.752	0.745	a b <-- classified as 150 150 | a = 0 98 602 | b = 1
2	Logistic Regression	20	NA	60-40%	400	73.50%	0.725	0.735	0.725	a b <-- classified as 66 69 | a = 0 37 228 | b = 1
3	Logistic regression with cost sensitive learning	20	Cost Matrix 0 2 1 0	10-fold	1000	72.40%	0.745	0.724	0.731	a b <-- classified as 196 104 | a = 0 172 528 | b = 1
4	Logistic regression with cost sensitive learning	20	Cost Matrix 0 2 1 0	60-40%	400	70.50%	0.713	0.705	0.708	a b <-- classified as 83 52 | a = 0 66 199 | b = 1

Table 14: Results of Logistic Regression on Top 6 Attributes

S. No	Parameters Used	No of Attributes	Cost Matrix	Test Data Method	Test Data Size	Accuracy	Weighted Precision	Weighted Recall	Weighted F Measure	Confusion Matrix
1	Logistic Regression	6	NA	10-fold	1000	74.20%	0.725	0.742	0.725	a b <-- classified as 122 178 | a = 0 80 620 | b = 1
2	Logistic Regression	6	NA	60-40%	400	72.67%	0.725	0.738	0.718	a b <-- classified as 43 63 | a = 0 13 181 | b = 1
3	Logistic regression with cost sensitive learning	6	Cost Matrix 0 2 1 0	10-fold	1000	71.70%	0.747	0.717	0.726	a b <-- classified as 205 95 | a = 0 188 512 | b = 1
4	Logistic regression with cost sensitive learning	6	Cost Matrix 0 2 1 0	60-40%	400	71.67%	0.721	0.708	0.721	a b <-- classified as 72 47 | a = 0 42 152 | b = 1

2.4.3. MODEL EVALUATION CRITERIA

There are a variety of different methods to evaluate a machine learning model. We decided to look at several different criteria in combination rather than using one of them exclusively.

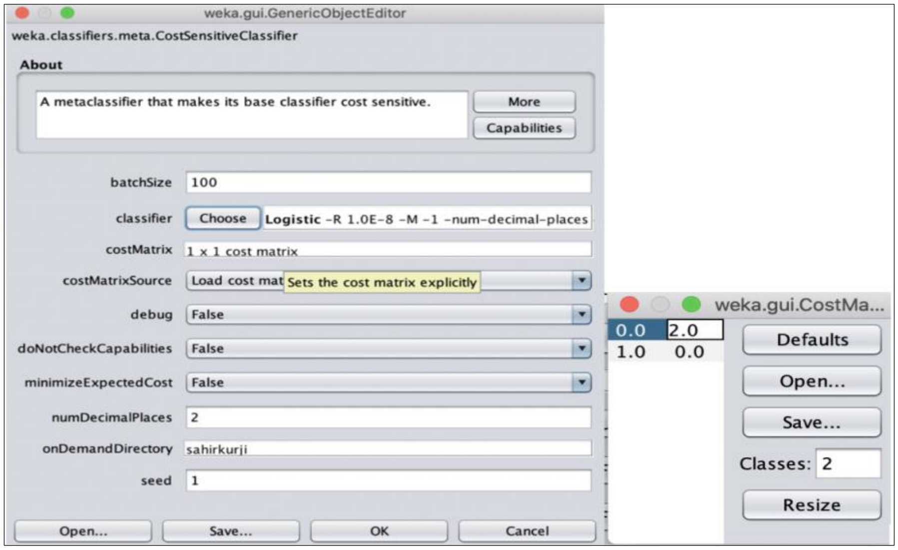
1. Accuracy (correctly classified instances)
2. Weighted Recall
3. Weighted Precision
4. Weighted F Measure

2.4.4. WEKA PARAMETERS

In order to perform logistic regression, the following steps were performed to achieve our results. Firstly, the dataset was uploaded onto Weka. In the classifier section, under the function dropdown, logistic was chosen. All other parameters remained the same. The test was then run with both training strategies and sets of attributes.

In order to make use of the cost sensitive function, the metaclassifier **CostSensitiveClassifier** was chosen. In the parameters section, the logistic algorithm was chosen, and the cost matrix was updated.

Figure 17: CostSensitiveClassifier in Weka



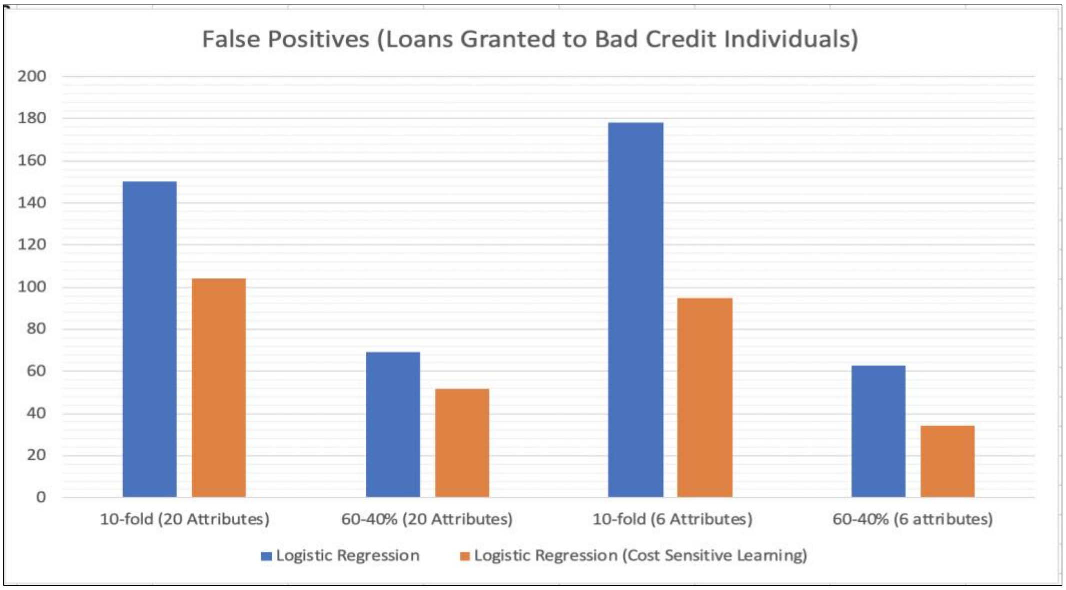
2.4.5. MODEL EVALUATION

From Table 13 and 14, we notice that all instances and measuring criteria range within a small window between 70 and 75%. The yellow boxes illustrate the highest performance for each decision criteria. As we can see, there is no model that clearly outperforms the other. However, when analyzing the results more carefully, there are multiple observations worth noting. Firstly, using our evaluation criteria, we can see that the best strategy to split the dataset is 10-fold cross validation rather than a simple percentage split. This is shown in the table where all instances have a higher accuracy using the cross-validation training strategy. The f-score also confirms the same.

Another important observation lies within the confusion matrices. Although we don't see an improvement in the 5 criteria using a cost sensitive strategy, the number of false positives does reduce (instances classified as b=1 but are actually a=0; loans given out to bad credit customers).

The aim of using a cost sensitive model is to minimize the overall **error cost** and not optimize the correctness of classifications. This is a function that includes the probability of default as well as the cost of default to the bank. When analyzing the ROC chart and Area Under Curve, there is relatively no difference when applying a cost sensitive filter. As a result, we can come to the conclusion that applying a cost sensitive strategy will not improve the **weighted** accuracy, precision and recall of the model but does improve the false positive rate of the model. In simple terms, this means that out of all the bad credit individuals, the number of misclassifications is lower. For example, comparing the first experiment to the 3rd, the false positive rate goes down from 50% to 34%. However, with this improvement, we also see a reduction in the true positive rate from 86% to 75%. This highlights the inverse relationship between optimizing the false positive rate and true positive rate. When reducing the false positive rate, we tend to reduce the true positive rate as well. Ultimately, by using the cost sensitive classifier, we are not necessarily improving the performance of the model. We are shifting the risk away from giving loans to bad customers at the expense of missing out on potential good clients. This is done because the cost of default is much greater to the bank than the cost of denying a good client. Ultimately, the cost sensitive model should prove to be more profitable.

Graph 3: Logistic Regression Results Comparison



2.5. CLASSIFICATION ALGORITHM COMPARISION

Table 15: Comparison of All Classification Algorithms

No	Algorithm	Testing Method	Number of Instances	Accuracy	Precision	Recall	F-Measure
1	Naive Bayes	10-Fold Test	1000	**75.4**	0.743	0.754	0.746
2	Naive Bayes	60-40 Split	400	74.5	0.737	0.745	0.726
3	Decision Tree	10-Fold Test	1000	72.8	0.712	0.728	0.716
4	Decision Tree	60-40 Split	400	71.75	0.706	0.718	0.687
5	Logistic Regression	10-Fold Test	1000	75.2	0.742	0.752	0.745
6	Logistic Regression	60-40 Split	400	73.5	0.725	0.735	0.725

Although we tested the dataset on various techniques and classification algorithms, we decided to draw a general conclusion on performance by comparing the results generated by Naïve Bayes, logistic regression and decision tree on all 20 attributes. The table above summarizes the performance using both training/testing techniques. For simplicity reasons, we used accuracy as our main measure of comparison. We can conclude that Naïve Bayes classification method (10-fold test) provided the highest accuracy. Another important observation to make is that a 10-fold training/testing strategy yielded better results than a simple percentage split throughout a significant portion of our experiments.

3. POST PREDICTIVE ANALYSIS

Post predictive analysis was performed to identify the characteristics of customers whose loans were approved as well as the characteristics of those who got disapproved. Below are the different strategies that were used:

1. K-Means
2. Apriori

3.1. K-MEANS CLUSTERING

Clustering is an unsupervised technique used to find homogeneous subgroups within the data set such that data points in each cluster are as similar as possible. K-means in one of the algorithms that is used to cluster the data. This algorithm tries to partition the dataset into k pre-defined clusters. It attempts to group the points within a cluster that are as similar as possible while trying to keep the inter-cluster distance as far as possible. It uses the sum of least squares where the sum of distances between data points and the cluster centroids are minimized.

SimpleKmeans is available under the Cluster category in Weka. There are three different tests that we performed using this algorithm.

1. Class to Cluster evaluation using all 1000 instances
2. Test on only those who were credit approved (700)
3. Test on only those who were disapproved (300)

3.1.1. CLASS TO CLUSTER EVALUATION

This test was performed by selecting the **'Class to cluster evaluation'** in Weka to verify how the k-means algorithms would segregate the data based on the class attribute (Credibility). The ***distanceFunction*** was defaulted to EuclideanDistance and the ***numClusters*** parameters was set to 2.

Figure 18: Results of K Means segregation on class attribute

Final cluster centroids:

Attribute	Full Data (1000.0)	Cluster# 0 (576.0)	1 (424.0)
Account Balance	4	1	4
Duration of Credit (month)	20.903	20.6476	21.25
Payment Status of Previous Credit	2	2	4
Purpose	3	3	3
Credit Amount	3271.248	3004.3299	3633.8538
Value Savings/Stocks	1	1	1
Length of current employment	3	3	3
Instalment per cent	2.973	3.1563	2.7241
Sex & Marital Status	3	3	3
Guarantors	1	1	1
Duration in Current address	4	4	2
Most valuable available asset	3	1	3
Age (years)	35.542	36.3507	34.4434
Concurrent Credits	3	3	3
Type of apartment	2	2	2
No of Credits at this Bank	1.407	1.2691	1.5943
Occupation	3	3	3
No of dependents	1.155	1.1528	1.158
Telephone	1	1	1
Foreign Worker	1	1	1

The algorithm was able to split the data into two clusters based on the credibility class, but the accuracy of this classification was very low as shown in Figure 19. **46%** of the instances were classified incorrectly.

Based on the classification that was performed, we observed that the following characteristics are different between the two clusters; Account balance, payment status of previous credit, credit amount, instalment per cent, duration in current address and most valuable available asset. We can therefore infer that these characteristics are the differentiating factors within our experiment. It is important to note that the accuracy of this experiment was low at 55%. As a result, we decided to run multiple tests that will be explained below in order to get a better holistic view.

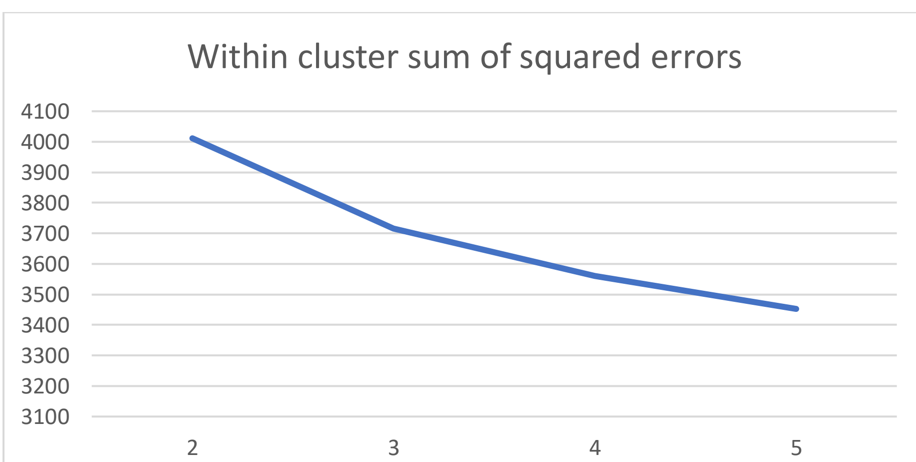
Figure 19: Classification Metrics of K means

```
=== Model and evaluation on training set ===

Clustered Instances

0      576 ( 58%)
1      424 ( 42%)


Class attribute: Creditability
Classes to Clusters:

   0   1  <-- assigned to cluster
 208  92 | 0
 368 332 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1

Incorrectly clustered instances :       460.0    46      %
```

3.1.2. TEST TO DETERMINE THE CHARACTERISTICS OF GOOD CREDIT RATING

We leveraged the **RemoveWithValues** feature in Weka to retain only the 700 records (credit approved). The number of clusters were determined by using "Elbow method".

Graph 4: Elbow method to determine number of clusters



For simplicity measures, we have decided to segregate the good credit individuals into 3 different clusters.

Figure 20: K Means Cluster output for Good Credit data

Final cluster centroids:

Attribute	Full Data (700.0)	Cluster# 0 (184.0)	1 (223.0)	2 (293.0)
Account Balance	4	4	4	4
Duration of Credit (month)	19.2071	16.5761	19.157	20.8976
Payment Status of Previous Credit	2	2	4	2
Purpose	3	3	0	3
Credit Amount	2985.4429	2460.1196	3146.5695	3192.7065
Value Savings/Stocks	1	1	1	1
Length of current employment	3	2	5	3
Instalment per cent	2.92	2.5978	3.0493	3.0239
Sex & Marital Status	3	2	3	3
Guarantors	1	1	1	1
Duration in Current address	4	4	4	2
Most valuable available asset	3	1	2	3
Age (years)	36.22	34.1576	42.148	33.0034
Concurrent Credits	3	3	3	3
Type of apartment	2	2	2	2
No of Credits at this Bank	1.4243	1.2228	1.722	1.3242
Occupation	3	3	3	3
No of dependents	1.1557	1.0435	1.3004	1.116
Telephone	1	1	1	1
Foreign Worker	1	1	1	1
Creditability	1	1	1	1

There are several observations that can be made when analyzing the results. Firstly, we see that all individuals who were granted a loan do not have a checking account at the bank. Perhaps the bank is trying to attract new customers by approving their loans. Next, across most instances, credit approved individuals had existing credits paid back duly. The bank therefore considers these individuals to have the capacity to remain in good credit over the duration of the loan.

In addition, we see that individuals who don't need guarantors are less risky and therefore more likely to get approved. Finally, we see that the bank considers the occupation of an individual as important. Across all clusters, credit approved individuals are highly skilled.

When looking across each cluster, we observe the following attributes to be different: Age, duration of credit, most valuable asset, credit amount and purpose. The mean value of all other attributes remained consistent across all clusters and are therefore not differentiating factors.

Cluster 0 consists of middle-aged people who seek lower credit amount and seek loans for a short duration of time. Another characteristic of this population is that they had one valuable asset.

Cluster 1 consists of older population who applied for a higher credit amount. people in this cluster also required a longer duration of credit. The credit purpose for these individuals leans towards a mean of 0. This means that these clients used the loan to buy a car.

Cluster 2 consisted of middle-aged clients who had applied for higher credit. The people in this cluster also applied for a longer duration of credit. An important trait of these customers consisted that they had 3 or more valuable assets.

3.1.3. CHARACTERISTICS OF DISAPPROVED CLIENTS (300 INSTANCES)

The third and final test was conducted to determine the characteristics of bad credit customers. While analyzing the output [Figure 21] of both clusters, we extract and infer several significant observations. Firstly, we see that both clusters who have a checking account balance under zero were considered high risk individuals and were disapproved. The bank also considered a longer duration of credit as being risky. This is evident in the difference between the individuals who received a loan (short duration of credit between 16-20 months) and those who were rejected the loan (long duration of credit >20 months). In addition, we observe that the bank was more reluctant in granting loans with higher credit amount. Those who were approved requested a mean amount of around $3,000 while those who were denied requested a larger sum of around $4,000. Lastly, individuals who had assets that are worth less were also less likely to be given a loan. These individuals either had a car or no assets at all.

Figure 21: K Means Cluster output for Bad Credit data

Final cluster centroids:

Attribute	Full Data (300.0)	Cluster# 0 (131.0)	1 (169.0)
Account Balance	1	1	1
Duration of Credit (month)	24.86	26.542	23.5562
Payment Status of Previous Credit	2	2	2
Purpose	0	0	3
Credit Amount	3938.1267	4576.1603	3443.5562
Value Savings/Stocks	1	1	1
Length of current employment	3	5	3
Instalment per cent	3.0967	3.1832	3.0296
Sex & Marital Status	3	3	2
Guarantors	1	1	1
Duration in Current address	4	4	2
Most valuable available asset	3	4	3
Age (years)	33.96	40.1756	29.142
Concurrent Credits	3	3	3
Type of apartment	2	2	2
No of Credits at this Bank	1.3667	1.4427	1.3077
Occupation	3	3	3
No of dependents	1.1533	1.2901	1.0473
Telephone	1	1	1
Foreign Worker	1	1	1
Creditability	0	0	0

Within each cluster, we can build a profile of each type of consumer as follows:

We can observe that loan applications of people in Cluster 0 were denied because they were old people who applied for very high credit amount. In addition, the purpose of the loan for these is to buy a car. As a result, the bank regarded these individuals as being too risky.

In contrast to Cluster 0, Cluster 1 consisted of young individuals who had relatively less job experience but wanted long duration of credit. It is possible that their loan application was declined due to the high amount asked as well as lower job security.

3.2. APRIORI

Association rules analysis is a technique to uncover how items are associated to each other. As part of post predictive analysis, we used Apriori – an association algorithm, to determine the association between attributes provided in the German Credit Card Data set. Apriori algorithm is applicable only to nominal, binary and unary data. All numerical attributes in the data set were converted to Nominal using multiple techniques mentioned further on in the analysis.

Below are the parameters that we considered when using the algorithm in Weka:

1. **Support**: Measures the popularity of an attribute in the dataset. The **lowerBoundMinSupport** (lower interval) and **upperBoundMinSupport** (upper interval) parameters can be adjusted in Weka. The Apriori algorithm works between these two intervals and increments by a delta value (increment level).

2. **Confidence**: Measures the probability of attribute X being chosen given attribute Y. Confidence can be changed by choosing **metricType** as Confidence and **minMetric** parameter.

3.2.1. NUMERIC TO NOMIAL PARAMETER CONVERSION

There are two ways to convert Numeric values to Nominal values in Weka

1. **NumericToNominal:** This converts all the numeric attributes to nominal in the dataset.

2. **Discretize:** In this method, we get to choose the attribute and the numbers of bins to be used when converting numeric data.

    - Duration of Credit – 8 bins
    - Credit Amount – 6 bins
    - Installment percent – 4 bins
    - Age – 8 bins
    - Number of Credits at this Bank – 3 bins
    - No of Dependents – 2 bins

3.2.2. RESULTS OF APRIORI ALGORITHM

To begin with, Apriori algorithm was applied on observations for which credit was approved (700 instances). We set the **metricType** to Confidence, ***car*** parameter was set to True, so that associations would be made on class the attribute. Support value was set to 0.6.

Figure 22: Output of Apriori for Good Credit Data

```
Best rules found:

 1. Foreign Worker=1 667 ==> Creditability=1 667    conf:(1)
 2. Guarantors=1 635 ==> Creditability=1 635    conf:(1)
 3. Guarantors=1 Foreign Worker=1 611 ==> Creditability=1 611    conf:(1)
 4. No of dependents=1 591 ==> Creditability=1 591    conf:(1)
 5. Concurrent Credits=3 590 ==> Creditability=1 590    conf:(1)
 6. No of dependents=1 Foreign Worker=1 569 ==> Creditability=1 569    conf:(1)
 7. Concurrent Credits=3 Foreign Worker=1 561 ==> Creditability=1 561    conf:(1)
 8. Guarantors=1 No of dependents=1 539 ==> Creditability=1 539    conf:(1)
 9. Guarantors=1 Concurrent Credits=3 538 ==> Creditability=1 538    conf:(1)
10. Type of apartment=2 528 ==> Creditability=1 528    conf:(1)
11. Guarantors=1 No of dependents=1 Foreign Worker=1 524 ==> Creditability=1 524    conf:(1)
12. Guarantors=1 Concurrent Credits=3 Foreign Worker=1 517 ==> Creditability=1 517    conf:(1)
13. Type of apartment=2 Foreign Worker=1 504 ==> Creditability=1 504    conf:(1)
14. Concurrent Credits=3 No of dependents=1 503 ==> Creditability=1 503    conf:(1)
15. Concurrent Credits=3 No of dependents=1 Foreign Worker=1 485 ==> Creditability=1 485    conf:(1)
16. Guarantors=1 Type of apartment=2 476 ==> Creditability=1 476    conf:(1)
17. Guarantors=1 Concurrent Credits=3 No of dependents=1 461 ==> Creditability=1 461    conf:(1)
18. Guarantors=1 Type of apartment=2 Foreign Worker=1 458 ==> Creditability=1 458    conf:(1)
19. Guarantors=1 Concurrent Credits=3 No of dependents=1 Foreign Worker=1 449 ==> Creditability=1 449    conf:(1)
20. Type of apartment=2 No of dependents=1 446 ==> Creditability=1 446    conf:(1)
21. Occupation=3 444 ==> Creditability=1 444    conf:(1)
22. Concurrent Credits=3 Type of apartment=2 444 ==> Creditability=1 444    conf:(1)
23. No of Credits at this Bank=1 433 ==> Creditability=1 433    conf:(1)
24. Type of apartment=2 No of dependents=1 Foreign Worker=1 429 ==> Creditability=1 429    conf:(1)
25. Occupation=3 Foreign Worker=1 428 ==> Creditability=1 428    conf:(1)
```

Some of the rules and their interpretation are as follows:

1.Foreign Worker=1 667 ==> Creditability=1 667

Loan seeking population mostly constituted of foreign works

2. Guarantors=1 635 ==> Creditability=1 635

Most of the loan applications which were approved, did not require a guarantor

3. Guarantors=1 Foreign Worker=1 611 ==> Creditability=1 611

Debtor who were foreign workers required no guarantors in 611 cases

4. No of dependents=1 591 ==> Creditability=1 591

In most cases, debtor has only 1 dependent

10. Type of apartment=2 528 ==> Creditability=1 528

Debtor owns his/her apartment in 528 instances

21. Occupation=3 444 ==> Creditability=1 444

Skilled employees were debtors

Next, we tested the algorithm on the top 6 attributes. Initially, we tried running this test with high values of support, but association rules were generated only when support value was 0.4 or lower.

Figure 23: Output of Apriori for Good Credit and top 6 attributes

```
Best rules found:

 1. Credit Amount='(-inf-3279]' 483 ==> Creditability=1 483    conf:(1)
 2. Value Savings/Stocks=1 386 ==> Creditability=1 386    conf:(1)
 3. Payment Status of Previous Credit=2 361 ==> Creditability=1 361    conf:(1)
 4. Account Balance=4 348 ==> Creditability=1 348    conf:(1)
 5. Duration of Credit (month)='(-inf-15.333333]' 342 ==> Creditability=1 342    conf:(1)
 6. Duration of Credit (month)='(-inf-15.333333]' Credit Amount='(-inf-3279]' 303 ==> Creditability=1 303
```

Some association rules and their interpretation were as follows:

1. Credit Amount='(-inf-3279]' 483 ==> Creditability=1 483

   Debtors seek smaller credit amount

2. Value Savings/Stocks=1 386 ==> Creditability=1 386;

   Good debtors have lower saving

3. Payment Status of Previous Credit=2 361 ==> Creditability=1 361

   Good debtors paid back their old loans

Lastly, we tried to implement the algorithm on all 20 attributes and all 1000 records in the dataset. For this particular test, we started with default support value of 0.5, there were no association rules in this setting. Results were observed only when the support value was 0.3. We concluded that results obtained using this test could not be used to figure out associations as the support value was too low.

3.2.3. APRIORI CONCLUSION

We recommend to the bank that creditable debtors are characterized by:

1. Do not require guarantors
2. Have no additional loans
3. Have less dependents
4. Own an apartment
5. Have low savings and are interested to borrow
6. Pay off loans in a timely manner

In addition, bank managers should not approve loans of the people whose checking account balance is < 0 DM (Deutsche Mark) and average balance in savings and stocks is <100 DM as they would have no means to repay the loan.

3.3. RECOMMENDATIONS FOR BANK

After running k-means clustering, there are several patterns that are useful for the bank. Factors such as age, most valuable available asset, purpose, credit history and employment skill level were some of the traits we observed to be important. For instance, we can conclude that a middle-aged person who applies for lower credit amount for a short duration of time is more likely to be issued a loan than a middle-aged person who applied for loan for high amount and for longer duration of time. In addition, an individual who is more skilled, doesn't need a guarantor and has a better credit history will also be more likely to earn a loan.

The Apriori algorithm also yielded important and similar observations for the bank. Creditable borrowers don't need guarantors, have no other loans/credits with no existing instalment plans, have fewer dependents, own their own apartments, have low savings, pay off loans on time, and want smaller loans. Based on these recommendations, prospective borrowers not profiled as creditable should have markers that deviate from the profile above.

It is also important to note that **not** only the top attributes that were identified as part of feature selection contributed to the eligibility of applications. Some factors like value savings/stocks actually yielded common results between good credit and bad credit individuals. As a result, additional features beyond 'the top 6' are highly recommended to be used by the bank in order to make the right decisions.

4. REFERENCES

Brownlee, Jason. "How to Perform Feature Selection With Machine Learning Data in Weka." *Machine Learning Mastery*, 13 Dec. 2019, machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/.

Brownlee, Jason. "How to Transform Your Machine Learning Data in Weka." *Machine Learning Mastery*, 13 Dec. 2019, machinelearningmastery.com/transform-machine-learning-data-weka/.

Chakrabarty, Navoneel. "Application of Synthetic Minority Over-Sampling Technique (SMOTe) for Imbalanced Datasets." *Medium*, Towards AI - Best Artificial Intelligence Publication, 16 July 2019, medium.com/towards-artificial-intelligence/application-of-synthetic-minority-over-sampling-technique-smote-for-imbalanced-data-sets-509ab55cfdaf.

Chakure, Afroz. "Decision Tree Classification." *Medium*, Towards Data Science, 7 July 2019, towardsdatascience.com/decision-tree-classification-de64fc4d5aac.

Ciortan, Madalina. "Overview of Feature Selection Methods." *Medium*, Towards Data Science, 26 July 2019, towardsdatascience.com/overview-of-feature-selection-methods-a2d115c7a8f7.

Dabbura, Imad. "K-Means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks." *Medium*, Towards Data Science, 3 Sept. 2019, towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a.

"Decision Tree Learning." *Wikipedia*, Wikimedia Foundation, en.wikipedia.org/wiki/Decision_tree_learning.

"Feature Selection." *Wikipedia*, Wikimedia Foundation, en.wikipedia.org/wiki/Feature_selection.

Gajawada, Sampath Kumar. "Chi-Square Test for Feature Selection in Machine Learning." *Medium*, Towards Data Science, 20 Oct. 2019, towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223.

Janabi, Dr. Kadhim B. S. Al. "Data Reduction Techniques: A Comparative Study for Attribute Selection Methods." (2018).

Joshi, Renuka. "Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures." *Exsilio Blog*, 9 Sept. 2016, blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.

Lutins, Evan. "Ensemble Methods in Machine Learning: What Are They and Why Use Them?" *Medium*, Towards Data Science, 2 Aug. 2017, towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f.

Ng, Annalyn. “Association Rules and the Apriori Algorithm: A Tutorial.” *KDnuggets*, Apr. 2016, www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html.

Paul, Sayak. “Diving Deep with Imbalanced Data.” *DataCamp Community*, 4 Oct. 2018, www.datacamp.com/community/tutorials/diving-deep-imbalanced-data.

Shubham, Jinde. “Ensemble Learning - Bagging and Boosting.” *Medium*, Becoming Human: Artificial Intelligence Magazine, 6 July 2018, becominghuman.ai/ensemble-learning-bagging-and-boosting-d20f38be9b1e.

Smolyakov, Vadim. “Ensemble Learning to Improve Machine Learning Results.” *Medium*, Stats and Bots, 7 Mar. 2019, blog.statsbot.co/ensemble-learning-d1dcd548e936.

Brownlee, Jason. “Cost-Sensitive Logistic Regression for Imbalanced Classification.” *Machine Learning Mastery*, 31 Jan. 2020, machinelearningmastery.com/cost-sensitive-logistic-regression/.

Brownlee, Jason. “How to Perform Feature Selection With Machine Learning Data in Weka.” *Machine Learning Mastery*, 13 Dec. 2019, machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/.

Brownlee, Jason. “How to Transform Your Machine Learning Data in Weka.” *Machine Learning Mastery*, 13 Dec. 2019, machinelearningmastery.com/transform-machine-learning-data-weka/.

Brownlee, Jason. “How To Use Classification Machine Learning Algorithms in Weka.” *Machine Learning Mastery*, 22 Aug. 2019, machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/.

Chakrabarty, Navoneel. “Application of Synthetic Minority Over-Sampling Technique (SMOTe) for Imbalanced Datasets.” *Medium*, Towards AI - Best Artificial Intelligence Publication, 16 July 2019, medium.com/towards-artificial-intelligence/application-of-synthetic-minority-over-sampling-technique-smote-for-imbalanced-data-sets-509ab55cfdaf.

Chakure, Afroz. “Decision Tree Classification.” *Medium*, Towards Data Science, 7 July 2019, towardsdatascience.com/decision-tree-classification-de64fc4d5aac.

Ciortan, Madalina. “Overview of Feature Selection Methods.” *Medium*, Towards Data Science, 26 July 2019, towardsdatascience.com/overview-of-feature-selection-methods-a2d115c7a8f7.

Dabbura, Imad. “K-Means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks.” *Medium*, Towards Data Science, 3 Sept. 2019, towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a.

“Decision Tree Learning.” *Wikipedia*, Wikimedia Foundation, en.wikipedia.org/wiki/Decision_tree_learning.

“Feature Selection.” *Wikipedia*, Wikimedia Foundation, en.wikipedia.org/wiki/Feature_selection.

Gajawada, Sampath Kumar. “Chi-Square Test for Feature Selection in Machine Learning.” *Medium*, Towards Data Science, 20 Oct. 2019, towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223.

Joshi, Renuka. “Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures.” *Exsilio Blog*, 9 Sept. 2016, blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.

Lutins, Evan. “Ensemble Methods in Machine Learning: What Are They and Why Use Them?” *Medium*, Towards Data Science, 2 Aug. 2017, towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f.

Ng, Annalyn. “Association Rules and the Apriori Algorithm: A Tutorial.” *KDnuggets*, Apr. 2016, www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html.

Paul, Sayak. “Diving Deep with Imbalanced Data.” *DataCamp Community*, 4 Oct. 2018, www.datacamp.com/community/tutorials/diving-deep-imbalanced-data.

Shubham, Jinde. “Ensemble Learning - Bagging and Boosting.” *Medium*, Becoming Human: Artificial Intelligence Magazine, 6 July 2018, becominghuman.ai/ensemble-learning-bagging-and-boosting-d20f38be9b1e.

Smolyakov, Vadim. “Ensemble Learning to Improve Machine Learning Results.” *Medium*, Stats and Bots, 7 Mar. 2019, blog.statsbot.co/ensemble-learning-d1dcd548e936.

Swaminathan, Saishruthi. “Logistic Regression - Detailed Overview.” *Medium*, Towards Data Science, 18 Jan. 2019, towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc.