PREDICTIVE ANALYSIS ABOUT THE LOAN APPROVAL PROCESS

FINAL PROJECT

Using GERMAN CREDIT DATA

CIND 119 - Introduction to Big Data

Toronto Metropolitan University Chang School of Continuing Studies

***Instructor***: Dr. Davood Pirayesh Neghab

***Project Members***

EMINE UYSAL

SHARMILA NANDI

**Table of Contents………………………………………………………2**

1.Introduction (Summary)............................................................................................................... 3

2.Tools............................................................................................................................................. 3

3.Workload Distribution................................................................................................................. 3

4.Data Import and Initial Exploration (Exploratory Data Analysis (EDA))**.*…………***...............4-8

4.1 Explore the Dataset…………………………………………………………………………4

4.2. Data Preparation……………………………………………………………………………5

5.Predictive Modeling /Classification....................................................................................... 9-13

5.1. Decision Tree&Trained Version............................................................................................ 9

6. Results of Decision Tree Process…….………………………………………………………..13

7.Re-training on selected features............................................................................................14-28

7.1. Naive Bayes&Trained Naive Bayes...............................................................................14

8.Results of Naïve Bayes...............................................................................................................27

9.Finalization……………………............................................................................................28-30

9.1.Conclusions and Comparison………………………………………………………………28

9.1.1.Performans Comparison…………………………………………………………………29

9.2.Recommendations…………………………………………………………………………..29

**1.Introduction**

Our client requests support from us to assess whether their customers who apply for a loan are eligible for a loan. To minimize the risk in the application approval process, it is aimed to subject these profiles to a logical classification and to facilitate the decision process of the evaluation to be made. For these reasons, applicants' financial profiles are drawn up and their ability to repay their loans is thoroughly assessed.

In addition to the loss of income caused by the disruptions to be experienced during the collection of the debt, administrative and legal follow-up processes also create extra costs. For this reason, it is aimed to minimize the risk before initiating the debiting process for credit agreements with customers.

In this study, the customer profiles provided by the bank will be examined and a model that can be used in the evaluation process will be established. This assessment has included parameters such as household income, the regularity of payments of loans already received, the amount of accumulation etc. will be included.

**2.Tools**

We will use decision tree and naïve bayes models for our predictive analysis for Loan approval process with German Credit Dataset. These models will be built with R Studio and SAS tools.

**3.Workload Distribution**

|  |  |
| --- | --- |
| **Member Name** | **List of Tasks** |
| EMINE UYSAL, [emine.uysal@torontomu.ca](mailto:emine.uysal@torontomu.ca)  (SAS) | **Classification Part(Data Preparation, DecisionTree and Results of the Decision Tree, Recommendations) and Finalization Part** |
| SHARMILA NANDI, [sharmila.nandi@torontomu.ca](mailto:sharmila.nandi@torontomu.ca)  **(R Studio)** | **Feature Importance Part (Naïve Bayes and Result of the Naïve Bayes, Feature Importance )** |

**4. Data Import and Initial Exploration (Exploratory Data Analysis (EDA))**

**4.1 Explore the Dataset**

The dataset provided came in 2 different formats (.csv and artf). Both are suitable for analysis, but the CSV format was used for the SAS and R tools. The dataset has 1000 samples (rows) and 21 attributes (columns), with the first attribute being 'Creditability' as a classifier labeled 0 or 1; 1= Eligible for credit, 0= Not eligible for credit. In the evaluation dataset of 1000 customers, loans were approved for 700 customers, while loan applications were rejected for 300 customers. (700/1000) 70% This is the approval rate for the evaluation conditions.

We can briefly explain the situations represented by some of the variables used in the evaluation as follows:

AccountBalance, CreditAmount, PaymentStatusofPreviousCredit: Represents individual payments and investment habits metrics related to the loan approval process.

Purpose: The reason for the loan (e.g., buying a car, education) can be used to predict the financial situation by looking at the needs of individuals.

Demographic Factors: Age, gender/marital status, length of employment, dependents, and occupation.

Guarantors, Value Savings and Stocks: These are the financial characteristics that affect credit decisions.

A white sheet with many lines

Description automatically generated with medium confidence

Figure .View the First Few Rows of the Dataset

Each variable represents a specific aspect of the financial data, like credit amount, the number of dependents, occupation status, etc., and is expressed numerically to facilitate analysis and operations like statistical modeling or data analysis. And N Miss: This column indicates the number of missing values for each variable. In this table, all variables have 0 missing values, implying that the dataset is complete with no missing data for the variables shown.

A screenshot of a computer

Description automatically generated A screenshot of a computer screen

Description automatically generated

Figure 1. Variables and Attributes Figure . Missing Values Chart

**4.2.Data Preparation**

• Descriptive statistics ran for all variables:

• Generated basic descriptive statistics for numerical and categorical variables

• Identify, Count and Handling Process Implemented for missing values. There are no missings.

• Imbalance analyze did. The dataset contains an imbalanced class distribution, with a larger proportion of creditworthy cases (70%). This could influence model training, so techniques like resampling or weighting should be used to address the imbalance depending on your goal.

• In order to better understand the effect of categorical variables on the target variable, the corelation procedure was applied.

•Splitted Data into Training and Testing process implemented. The dataset underwent random sampling to create a training set with 700 entries, maintaining the original distribution of credibility. The provided settings ensure that the sampling process is reproducible.

A screenshot of a computer screen

Description automatically generated A screenshot of a survey

Description automatically generated

A screenshot of a computer

Description automatically generated

*Insights from MEANS*

* The dataset includes applicants with diverse financial profiles.
* The average credit amount is moderately high (~3271), with significant variation.
* Most individuals appear creditworthy (70%), have stable employment (~3–4 years), and are in their mid-30s.
* Categories like account balance, savings, and payment history seem crucial in determining creditability.
* According to the frequency and statistical values found by descriptive analysis, most borrowers have low Account Balance and Value Saving, which indicates potential financial constraints. The data contains a mix of numerical and categorical features with a wide range of values. The variability in characteristics such as Loan Amount and Loan Duration provides a good basis for discerning creditworthiness.

A screenshot of a computer screen

Description automatically generated

Figure .Descriptive Statistics for Numerical and and Categorical Variables

**A screenshot of a number

Description automatically generated A screenshot of a document

Description automatically generated A screenshot of a table

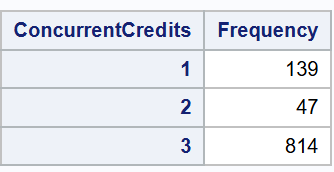
Description automatically generated A white rectangular box with blue text

Description automatically generated with medium confidence**

**A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated **

**A screenshot of a computer

Description automatically generated A white paper with black text and numbers

Description automatically generated A screenshot of a computer

Description automatically generated**

• The proc univariate function was applied to the variables to find outlier values. The Interquartile Range (IQR) formula applied (Q1 (25th percentile) and Q3 (75th percentile).

IQR = Q3 - Q1.) According to the result reached after the application, Purpose, Noofcreditatthisbank, CreditAmount have outlier values for its variables. We can also confirm this calculation from the boxplot images created.

**A graph of a diagram

Description automatically generated with medium confidence A graph of a creditability

Description automatically generated with medium confidence**

**A graph of a diagram

Description automatically generated with medium confidence**

•This distribution depicts the number of credits held at a bank. Distribution Characteristics:

* Right-Skewed: The histogram is right-skewed (positively skewed), meaning most customers have fewer credits, with a tail extending to the right for customers with more credits.

Concentration: There's a higher concentration of customers having a lower number of credits.

**A graph of a bank

Description automatically generated**

**5.Predictive Modeling /Classification**

**5.1. Decision Tree & Trained Version**

The model used all 1,000 observations.700 (70%) were used for training. 300 (30%) were used for validation to test performance on unseen data.

A screenshot of a computer screen

Description automatically generated A screen shot of a computer

Description automatically generated

Tree Construction: The decision tree was built using the Gini Index for splits and pruned using Cost-Complexity pruning, reducing the number of leaves from 104 to 5 for a more generalized model.

Depth: The maximum tree depth requested was 10, and the final pruned tree achieved a depth of 10.

Data Usage: The dataset contained 1,000 observations, with 690 used for training and 310 for validation.Validation ensures the model works well on new, unseen data.

This is a visualization of a decision tree showing how the dataset splits at each level based on features like AccountBalance, Payment Status, Credit Amount, Age, and Length of Current Employment. This tree shows the sequence of decision rules the model uses to classify whether an individual is creditworthy or not. Each split is based on the most significant variable at that stage to improve classification accuracy amount of credit requested influences the decision.

A diagram of a tree

Description automatically generated A screenshot of a computer

Description automatically generated

The provided results contain four keys component: confusion matrices, fit statistics, performance and an ROC curve for the "Creditability" variable. The confusion matrices show the model performs better at identifying creditworthy individuals (class 1) compared to non-creditworthy individuals (class 0):

**a.Model-Based Confusion Matrix**

|  |  |
| --- | --- |
| True Negatives (0, 0) | 112 |
| False Positives (0, 1) | 104 |
| False Negatives (1, 0) | 66 |
| True Positives (1, 1) | 418 |

From this confusion matrix, we can calculate the following performance metrics:

**Error Rate:** The error rate is calculated as the proportion of incorrect predictions. The error rate for each class is provided:

* + For class 0: Error rate = 104/(112+104)=0.4815
  + For class 1: Error rate = 66/(66+418)=0.1364

This suggests that the model is more accurate in predicting class 1 (with an error rate of 0.1364) compared to class 0 (with an error rate of 0.4815).

A graph of a function

Description automatically generated with medium confidence A graph of a training curve

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated A screenshot of a computer screen

Description automatically generated

**Overall Conclusion:**

* **Top Left & Bottom Left:** The model is ineffective, performing like a random classifier.
* **Top Right & Bottom Right:** Improved model with moderate performance, better discrimination ability, and reduced error rates. This indicates a more effective model with better training outcomes using 7 leaves, indicating complexity helps this model.

**b.Model-Based Fit Statistics for Selected Tree**

\* The model performs better in classifying negatives (class 0) with a specificity of 86%, but it struggles more with classifying positives (class 1) as indicated by a sensitivity of 52%.

\* The **misclassification rate** is moderate at 24.29%, meaning that the model makes errors about a quarter of the time.

\* The **AUC** of 0.7611 suggests that the model is capable of distinguishing between classes with moderate performance.

\* The tree was pruned down to 7 leaves, which helped reduce complexity while still maintaining reasonable performance.

**c.Variable Importance**

* Most Important Variables:
* AccountBalance was the most important predictor with a relative importance of 1.0000.
* ValueSavingsandStocks and CreditAmount were also significant contributors, with lower but still notable importance.
* Less Important Variables: Features like Purpose, SexandMaritalStatus, and Typeofapartment contributed less to the model's predictions.

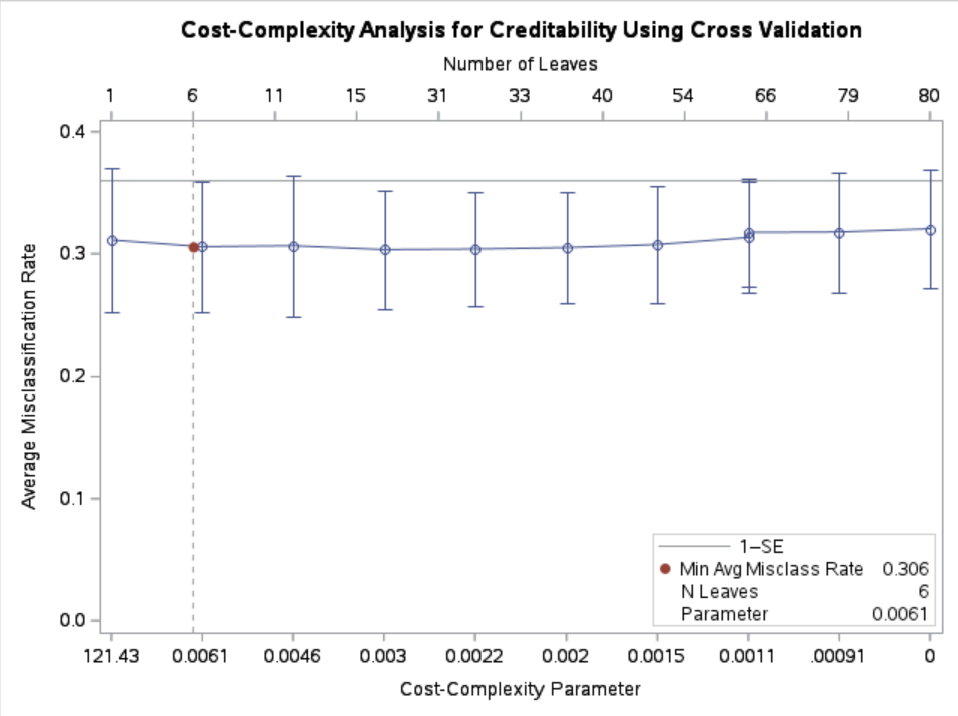
A screenshot of a graph

Description automatically generated

**d.Model Performance**

* **Initial Model:** The initial decision tree had an error rate of 1.0000, meaning it performed poorly, misclassifying almost all cases.
* **After Pruning:** The tree was pruned to 7 leaves, improving performance:
  + **Misclassification rate:** 24.29%
  + **Sensitivity:** 51.85% (ability to correctly classify positive cases)
  + **Specificity:** 86.36% (ability to correctly classify negative cases)
  + **AUC:** 0.7611 (good ability to distinguish between classes)

A screenshot of a data report

Description automatically generated 

**6.Results of the Decision Tree Process**

*Before Training (Initial Results):* The model was overly simple and ineffective;

Sensitivity (Class 1): 0.00% (it couldn't identify any approved loans).

Specificity (Class 0): 100.00% (it perfectly identified rejected loans).

AUC: 0.50 (performance was as random likeflipping a coin).

The tree was pruned to just 1 leaf, making it overly simplistic and biased toward rejecting all loans.

*After Training and Pruning:* The model improved significantly;

Sensitivity: Increased to 51.85% (the model now identifies about half of approved loans).

Specificity: Reduced slightly to 86.36%, but it still performs well for rejected loans.

AUC: Improved to 0.7611, showing the model gained moderate discriminative power.

The tree was simplified to 7 leaves after pruning, balancing complexity and generalization.

Misclassification Rate: Improved to 24.29%, though there’s still room for improvement.

*Before training*: The model was useless, rejecting all loans with no ability to classify approved cases.

*After training and pruning*: the decision tree showed improved performance, with a balanced AUC of 0.7611 and better sensitivity. However, further refinements are needed to address the remaining issues and ensure the model’s reliability for the bank's loan approval process.

**7.Re-training on selected features**

**7.1. Naive Bayes&Trained Naive Bayes**

Naives Bayes is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Similar to the the decision tree, we will evaluate based on a model containing all the attributes and a model containing attributes that influence the class attribute Creditability.

This is the Naives bayes classification with all the attributes in the model.

## Naive Bayes Classifier for Discrete Predictors

##Call:

##naiveBayes.default(x = X, y = Y, laplace = laplace)

##A-priori probabilities:

#Y

##good bad

## 0.70875 0.29125

## Conditional probabilities:

## Account.Balance

## Y <0 Deutsche Mark 0<=..<200 Deutsche Mark >200 Deutsche Mark

## good 0.20282187 0.22398589 0.07054674

## bad 0.44206009 0.36480687 0.04721030

## Account.Balance

## Y No checking account

## good 0.50264550

## bad 0.14592275

## Duration.of.Credit..month.

## Y [,1] [,2]

## good 19.21340 11.09382

## bad 24.77682 13.51666

##

## Payment.Status.of.Previous.Credit

## Y no credit all Credit Paid existing Credit delay critical Account

## good 0.01587302 0.02998236 0.52910053 0.08465608 0.34038801

## bad 0.06866953 0.07725322 0.58369099 0.08583691 0.18454936

##

## Purpose

## Y New car Used car Equipment Radio/Television Domestic Appliances

## good 0.201058201 0.123456790 0.179894180 0.313932981 0.012345679

## bad 0.287553648 0.051502146 0.201716738 0.218884120 0.008583691

## Purpose

## Y Repairs Education Vacation Retraining Business Other

## good 0.021164021 0.044091711 0.000000000 0.010582011 0.084656085 0.008818342

## bad 0.025751073 0.081545064 0.000000000 0.004291845 0.103004292 0.017167382

##

## Credit.Amount

## Y [,1] [,2]

## good 2977.734 2421.042

## bad 3810.086 3425.829

##

## Value.Savings.Stocks

## Y <100 100=<..<500 500=<..<1000 =>1000 unknown/no saving account

## good 0.54850088 0.09876543 0.08112875 0.06349206 0.20811287

## bad 0.72532189 0.11158798 0.03862661 0.02145923 0.10300429

##

## Length.of.current.employment

## Y unemployed < lyear 1 year <...< 4year 4year<=...<7years 5>=7years

## good 0.05291005 0.13756614 0.34038801 0.19929453 0.26984127

## bad 0.06437768 0.22746781 0.36480687 0.13733906 0.20600858

##

## Instalment.per.cent

## Y [,1] [,2]

## good 2.940035 1.124700

## bad 3.115880 1.066478

##

## Sex...Marital.Status

## Y male:divorced/seperated female:divorced/seperated male:single

## good 0.04585538 0.28571429 0.57848325

## bad 0.07725322 0.36051502 0.48927039

## Sex...Marital.Status

## Y male:married/widowed female:single

## good 0.08994709 0.00000000

## bad 0.07296137 0.00000000

##

## Guarantors

## Y none co-applicant guarantor

## good 0.90299824 0.03880071 0.05820106

## bad 0.92274678 0.04721030 0.03004292

##

## Duration.in.Current.address

## Y <= 1 year 1<...<= 2 years 2<..<= 3 years >3years

## good 0.1322751 0.3051146 0.1463845 0.4162257

## bad 0.1244635 0.3347639 0.1587983 0.3819742

##

## Most.valuable.available.asset

## Y real estate savings agreement/lifeinsurance car or other

## good 0.3156966 0.2222222 0.3368607

## bad 0.2060086 0.2274678 0.3562232

## Most.valuable.available.asset

## Y unknown/no property

## good 0.1252205

## bad 0.2103004

##

## Age..years.

## Y [,1] [,2]

## good 36.22046 11.33987

## bad 33.45494 11.03907

##

## Concurrent.Credits

## Y bank stores none

## good 0.10758377 0.04232804 0.85008818

## bad 0.18025751 0.06866953 0.75107296

##

## Type.of.apartment

## Y rent own free

## good 0.16225750 0.75132275 0.08641975

## bad 0.22746781 0.64806867 0.12446352

##

## No.of.Credits.at.this.Bank

## Y [,1] [,2]

## good 1.403880 0.5862251

## bad 1.360515 0.5636884

##

## Occupation

## Y unemployed/unskilled - non resident unskilled-resident

## good 0.02116402 0.20458554

## bad 0.02145923 0.19742489

## Occupation

## Y skilled employee/official

## good 0.64197531

## bad 0.61802575

## Occupation

## Y managemen/self-employed/highly qualified/officer

## good 0.13227513

## bad 0.16309013

##

## No.of.dependents

## Y [,1] [,2]

## good 1.158730 0.3657469

## bad 1.150215 0.3580508

##

## Telephone

## Y yes no

## good 0.5679012 0.4320988

## bad 0.6523605 0.3476395

##

## Foreign.Worker

## Y yes no

## good 0.95238095 0.04761905

## bad 0.98712446 0.01287554

This is the Naives Bayes classification with attributes that showed high significance to Creditability

##

## Naive Bayes Classifier for Discrete Predictors

##

## Call:

## naiveBayes.default(x = X, y = Y, laplace = laplace)

##

## A-priori probabilities:

## Y

## good bad

## 0.70875 0.29125

##

## Conditional probabilities:

## Account.Balance

## Y <0 Deutsche Mark 0<=..<200 Deutsche Mark >200 Deutsche Mark

## good 0.20282187 0.22398589 0.07054674

## bad 0.44206009 0.36480687 0.04721030

## Account.Balance

## Y No checking account

## good 0.50264550

## bad 0.14592275

##

## Payment.Status.of.Previous.Credit

## Y no credit all Credit Paid existing Credit delay critical Account

## good 0.01587302 0.02998236 0.52910053 0.08465608 0.34038801

## bad 0.06866953 0.07725322 0.58369099 0.08583691 0.18454936

##

## Purpose

## Y New car Used car Equipment Radio/Television Domestic Appliances

## good 0.201058201 0.123456790 0.179894180 0.313932981 0.012345679

## bad 0.287553648 0.051502146 0.201716738 0.218884120 0.008583691

## Purpose

## Y Repairs Education Vacation Retraining Business Other

## good 0.021164021 0.044091711 0.000000000 0.010582011 0.084656085 0.008818342

## bad 0.025751073 0.081545064 0.000000000 0.004291845 0.103004292 0.017167382

##

## Age..years.

## Y [,1] [,2]

## good 36.22046 11.33987

## bad 33.45494 11.03907

##

## Value.Savings.Stocks

## Y <100 100=<..<500 500=<..<1000 =>1000 unknown/no saving account

## good 0.54850088 0.09876543 0.08112875 0.06349206 0.20811287

## bad 0.72532189 0.11158798 0.03862661 0.02145923 0.10300429

##

## Most.valuable.available.asset

## Y real estate savings agreement/lifeinsurance car or other

## good 0.3156966 0.2222222 0.3368607

## bad 0.2060086 0.2274678 0.3562232

## Most.valuable.available.asset

## Y unknown/no property

## good 0.1252205

## bad 0.2103004

##

## Purpose

## Y New car Used car Equipment Radio/Television Domestic Appliances

## good 0.201058201 0.123456790 0.179894180 0.313932981 0.012345679

## bad 0.287553648 0.051502146 0.201716738 0.218884120 0.008583691

## Purpose

## Y Repairs Education Vacation Retraining Business Other

## good 0.021164021 0.044091711 0.000000000 0.010582011 0.084656085 0.008818342

## bad 0.025751073 0.081545064 0.000000000 0.004291845 0.103004292 0.017167382

##

## Credit.Amount

## Y [,1] [,2]

## good 2977.734 2421.042

## bad 3810.086 3425.829

##

## Value.Savings.Stocks

## Y <100 100=<..<500 500=<..<1000 =>1000 unknown/no saving account

## good 0.54850088 0.09876543 0.08112875 0.06349206 0.20811287

## bad 0.72532189 0.11158798 0.03862661 0.02145923 0.10300429

##

## Length.of.current.employment

## Y unemployed < lyear 1 year <...< 4year 4year<=...<7years 5>=7years

## good 0.05291005 0.13756614 0.34038801 0.19929453 0.26984127

## bad 0.06437768 0.22746781 0.36480687 0.13733906 0.20600858

##

## Instalment.per.cent

## Y [,1] [,2]

## good 2.940035 1.124700

## bad 3.115880 1.066478

##

## Sex...Marital.Status

## Y male:divorced/seperated female:divorced/seperated male:single

## good 0.04585538 0.28571429 0.57848325

## bad 0.07725322 0.36051502 0.48927039

## Sex...Marital.Status

## Y male:married/widowed female:single

## good 0.08994709 0.00000000

## bad 0.07296137 0.00000000

##

## Guarantors

## Y none co-applicant guarantor

## good 0.90299824 0.03880071 0.05820106

## bad 0.92274678 0.04721030 0.03004292

##

## Duration.in.Current.address

## Y <= 1 year 1<...<= 2 years 2<..<= 3 years >3years

## good 0.1322751 0.3051146 0.1463845 0.4162257

## bad 0.1244635 0.3347639 0.1587983 0.3819742

##

## Most.valuable.available.asset

## Y real estate savings agreement/lifeinsurance car or other

## good 0.3156966 0.2222222 0.3368607

## bad 0.2060086 0.2274678 0.3562232

## Most.valuable.available.asset

## Y unknown/no property

## good 0.1252205

## bad 0.2103004

##

## Age..years.

## Y [,1] [,2]

## good 36.22046 11.33987

## bad 33.45494 11.03907

##

## Concurrent.Credits

## Y bank stores none

## good 0.10758377 0.04232804 0.85008818

## bad 0.18025751 0.06866953 0.75107296

##

## Type.of.apartment

## Y rent own free

## good 0.16225750 0.75132275 0.08641975

## bad 0.22746781 0.64806867 0.12446352

##

## No.of.Credits.at.this.Bank

## Y [,1] [,2]

## good 1.403880 0.5862251

## bad 1.360515 0.5636884

##

## Occupation

## Y unemployed/unskilled - non resident unskilled-resident

## good 0.02116402 0.20458554

## bad 0.02145923 0.19742489

## Occupation

## Y skilled employee/official

## good 0.64197531

## bad 0.61802575

## Occupation

## Y managemen/self-employed/highly qualified/officer

## good 0.13227513

## bad 0.16309013

##

## No.of.dependents

## Y [,1] [,2]

## good 1.158730 0.3657469

## bad 1.150215 0.3580508

##

## Telephone

## Y yes no

## good 0.5679012 0.4320988

## bad 0.6523605 0.3476395

##

## Foreign.Worker

## Y yes no

## good 0.95238095 0.04761905

## bad 0.98712446 0.01287554

This is the Naives Bayes classification with attributes that showed high significance to Creditability

##

## Naive Bayes Classifier for Discrete Predictors

##

## Call:

## naiveBayes.default(x = X, y = Y, laplace = laplace)

##

## A-priori probabilities:

## Y

## good bad

## 0.70875 0.29125

##

## Conditional probabilities:

## Account.Balance

## Y <0 Deutsche Mark 0<=..<200 Deutsche Mark >200 Deutsche Mark

## good 0.20282187 0.22398589 0.07054674

## bad 0.44206009 0.36480687 0.04721030

## Account.Balance

## Y No checking account

## good 0.50264550

## bad 0.14592275

##

## Payment.Status.of.Previous.Credit

## Y no credit all Credit Paid existing Credit delay critical Account

## good 0.01587302 0.02998236 0.52910053 0.08465608 0.34038801

## bad 0.06866953 0.07725322 0.58369099 0.08583691 0.18454936

##

## Purpose

## Y New car Used car Equipment Radio/Television Domestic Appliances

## good 0.201058201 0.123456790 0.179894180 0.313932981 0.012345679

## bad 0.287553648 0.051502146 0.201716738 0.218884120 0.008583691

## Purpose

## Y Repairs Education Vacation Retraining Business Other

## good 0.021164021 0.044091711 0.000000000 0.010582011 0.084656085 0.008818342

## bad 0.025751073 0.081545064 0.000000000 0.004291845 0.103004292 0.017167382

##

## Age..years.

## Y [,1] [,2]

## good 36.22046 11.33987

## bad 33.45494 11.03907

##

## Value.Savings.Stocks

## Y <100 100=<..<500 500=<..<1000 =>1000 unknown/no saving account

## good 0.54850088 0.09876543 0.08112875 0.06349206 0.20811287

## bad 0.72532189 0.11158798 0.03862661 0.02145923 0.10300429

##

## Most.valuable.available.asset

## Y real estate savings agreement/lifeinsurance car or other

## good 0.3156966 0.2222222 0.3368607

## bad 0.2060086 0.2274678 0.3562232

## Most.valuable.available.asset

## Y unknown/no property

## good 0.1252205

## bad 0.2103004

**8.Results of Naïve Bayes**

**8.1.Performance Metrics for Naives Bayes Classification**

**Confusion Matrix**

## pred3 good bad

## good 115 26

## bad 18 41

##

## pred4 good bad

## good 112 33

## bad 21 34

We can see that 115/200 of the credit was labeled as good credit and 26 good credit were mislabled as bad credit for the model with all attributes

We can see that 112/200 of the credit was labeled as good credit and 33 good credit were mislabled as bad credit for the model that highly

influence Creditability.

## [1] "Accuracy for test with all attributes 0.78"

## [1] "Accuracy for test with attributes that influence Creditability 0.76"

Naive Bayes is a collection of classification algorithms commonly used in Machine Learning. In this project we used the  Naive Bayes algorithm to predict if the customer should receive or not a loan. We used a simple train-test set split strategy for this dataset. The main attribute was ‘Creditability”. The final output shows that we build a Naive Bayes classifier that can predict if the customer should receive or not a loan with an accuracy of 78%. In addition for the model including the attributes that influence creditability has an accuracy of 76%. Thus the model using all attributes is slightly better.

**A diagram of a credit balance

Description automatically generated** **A table with numbers and text

Description automatically generated**

**9.Finalization**

**9.1. Conclusions and Comparison**

Based on the finding in the data preperations section, one major finding that I thought was interesting was during regression analysis of the models. The r-squared of the additative model was 0.268 which implies that on 26.8 or around 27 % of the data from the regression fits the data, after applying stepwise regression which elimanted the attributes which were redundant by looking at AIC scores. The R-squared value for the step.model was 26% which indicates that less data from the regression was fitted onto the data. Even though the R-squared values were low for the model the AIC which is estimator of prediction error indicated it was the best model by choice. This model represented the attributes that influence Creditability. Looking at the p-values of the attributes we saw that some were more significant than the others. For instance, Credit amount is signficant which makes sense because you know whether you have good or bad credit depending on the amount of credit you have. For the Correlation Graph we saw that Creditability, is highly correlated to Account Balance, Payment Status of Previous Credit, Value in Saving Stocks, Age and there is weaker correlation with Duration of Credit, Credit Amount, Most Valuable available asset. We can also see that there is a strong correlation between other attributes, such as Duration of Credits per month and Credit Amount. This makes sense because the credit amount is dependent on how long the credit lasts. This also shows that that there exists some aspects of multicollinearity which must be considered when working with the data set.

Based on the findings in our Predictive modeling using both naives bayes classification and decision tree classification. The decision tree showed the outcomes of good or bad credit going through a series of decisions and their possible consequences, including the probability of good or bad credit occuring going through that path. Like the tree method Naive Bayes showed that We also saw that around 78% of the data was accurately classified compared to the decision tree method which was 76%

**9.1.1.Performance Comparison**

Based on the performance metric of Naives Bayes Classifier and Decision Tree Classifier. We can see that Naives Bayes Classifier has a slight edge since its accuracy of retrieving data is 78% accurate compared to the 76% accuracy of the Decision Tree Classifier.

**9.2. Recommendations**

**A graph with a bar and a bar chart

Description automatically generated with medium confidence**

1. Focus on Key Predictors in the decision-making process, such as Account Balance, Value Savings and Stocks, Loan Amount, Purpose, Gender and Marital Status, and Apartment Type. As can be seen in the image, these are the variables that have the most impact on creditability.

2. The following variables can be added to the variables typically collected in the loan debiting process to improve the creditability estimation:

• Income

• Credit score

3. You can be simplified the model by removing the following variables, as the dataset shows zero correlation with credibility based on the Phi coefficient:

• Number of dependents

• Number of loans in this bank

• Time spent at current address

•Profession