**Utilizing Machine Learning for German Credit Analysis**

**Sarah Huerta, Jack Miller, Robert Wortelboer**

**Introduction**

Our problem statement is that we need to find certain data sets that help us understand what groups have bad and good credit based on attributes such as Income, sex, Age, Job and more. All of us are upper division college students. We care about using proper data techniques to better understand how credit works and so we can give good data to companies. Along with our professional motivation on a personal level about how other countries such as german use credit risk. The way we evaluated our data sets is by looking at other examples to see what we were not doing as a result of our mistakes during this project. Overall we feel that we provide the proper data techniques to show how credit should be judged.

**Dataset and Acquisition**

We sourced our data from Kaggle, and pulled the public data set titled German Credit Risk. This set Attributes: Age, Sex, Job Count, Housing, Savings/Checking Account, Credit Amout, Duration of Credit, Purpose and Risk. Among those attributes, we divided them into two groups, numeric and non-numeric datasets.

**Tools Used**

The Tools that we used are Orange and Jupyter Notebook. We used the Orange application to utilize its Machine Learning Functions in the Analysis of our data. Our use of Jupyter Notebook focused more on data modeling for our non-numeric attributes.

**Data Analysis and Results**

***Basic Characteristics of the Dataset***

In order to get an overview of each attribute, we looked at boxplots to view the distribution of each numeric attribute. For those attributes that are categorical, we used a bar graph to look at the distribution. Each attribute can be seen below:

| Figure 1: Basic Distribution of Credit Amount |
| --- |
|  |
| The distribution of Credit amount is skewed right, with a great number of outliers outside of the 4th quartile. |

| Figure 2: Basic Distribution of Duration |
| --- |
|  |
| The distribution of Duration is skewed right, with few outliers outside of the 3rd quartile. |

| Figure 3. Basic Distribution of Sex |
| --- |
|  |
| Within the sex attribute, roughly 70% of entries were male and 30% were female. |

| Figure 4: Basic Distribution of Age and Sex |
| --- |
| The distribution of Age is skewed right, with the large spread between the 3rd and 4th quartiles. There are seven outliers greater tn the 4th quartile. |

| Figure 5: Basic Distribution of Job |
| --- |
|  |
| Within the Job attributes, there is a mean of 1.9. |

| Figure 6:. Basic Distribution of Housing |
| --- |
|  |
| Within Housing, the majority of entries own their house. |

| Figure 7:. Basic Distribution of Checking Account |
| --- |
| Within the checking account, the average is split equally between little and moderate amounts. |

| Figure 8: Basic Distribution of Saving Accounts |
| --- |
|  |
| Most people within the data sets have a little amount in their savings. |

| Figure 9: Basic Distribution of Purpose |
| --- |
|  |
| For Purpose, most applications use their credit for Car loans. |

| Figure 10: Basic Distribution of Risk |
| --- |
|  |
| Within the Risk attribute, 30% of applicants have a high risk, and 70% are low. |

***Data Preprocessing***

When we first started our project, we removed the index column because it was throwing off our data. In class, we had worked on data visualization by using scatter plots, and so that was our first course of action to start our modeling. We quickly realized that this would not be the best fit for our data. When using the Scatterplots, we had a difficult time seeing the negative and positive correlation in the attributes. We attempted to chante visually the color key visually in order to differentiate between the attributes, however this still did not work. After deliberation, we came to the agreement that this issue was caused due to the numeric and nonnumeric attributes in our data, which we at first did not realize. After this discovery, we were able to move forward with more efficient means of modeling dependent on the attribute we were working with. For the non numeric attributes, we decided to use boxplots with bar graphs to model our data, while for the numeric attributes, we used boxplots and histograms. From there, we then moved on to our Machine Learning Modeling.

| Table 1: Confusion Matrix for both Good and Bad Models | |
| --- | --- |
| Good | Bad |
|  |  |
| The tables above represent the model comparison between each machine learning algorithm we chose.In the table above, if the number is less than 1 the difference is negligible. | |

***Data Modeling and Evaluation Discussion***

The machine learning algorithms we used within our project include the SVM, random forest, logic regression, and Naive Bayes machine learning algorithms. We used stratified 10-fold cross validation with a training set size of 66% for the datasets The predictors we focused on included: The attributes such as age, sex, occupation, housing, checking account, none of these can be used as a strong predictor, they are all equal in predicting for the credit rating within the dataset.

Table 1 shows the probability that the score model in the row is higher than that of the modeling the column. We compared the models based off of each of the risk factors, since you could either be a bad or good credit applicant we decided to separate the good applicants and bad applicants model probability. In the figure below, we show what process we used in order to generate our Classification values.

|  |
| --- |
| Figure 11: This represents our Orange Diagram that we used to produce our Classification values. |

In Table 2, We can see our output values that were given when we used a 10 fold cross validation. We can see that the Naive Bayes has the highest Accuracy value, SVM doing the worst.

| Table 2: Performance of the Models | | | | |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Recall | Precision | F1 |
| Random Forest | 72.1%% | 72.2% | 70.3% | 70.7% |
| SVM | 63.2% | 65.1% | 66.3% | 65.6% |
| Logistic Regression | 75.1% | 73.7% | 70.8% | 71.8% |
| Naive Bayes | 76.1% | 72.6% | 71.6% | 72.0% |

With picking which machine learning value model we used for our analysis, we looked at which had the highest accuracy, that being the Naive Bayes. However, we can see a difference in the recall and precision which is higher in other models. Between the four models that we chose, the Logistic Regression had a better recall at 73.7%, meaning that it had a 1.1% higher fraction of relevant instances that were retrieved than the Naive Bayes. However, the Precision of Logistic Regression was .8% lower. This means that the Naive Bayes has a slightly better fraction of relevant instances among the retrieved instances.

Moving forward with the Naives Bayes model, we then created a confusion matrix to determine the performance of its prediction values. We can see that there is a higher true negative and true positive prediction percentage than the number of False negative and false positives. With this, we can confirm that the Naive Bayes is a good Machine Learning Model for our data set.

|  |
| --- |
| Figure 12 : Pictured above is the confusion matrix for the Naive Bayes model looking at the proportion of the predictions. |

**Timeline for Completion**

In order to stay on task for our project, we worked to mirror the course structure and reflect back on what we learned in class, and how we can use these new skills, such as different methods of Machine Learning techniques like Naive Bayes, Random Forest and Logistic Regression.

**Team Workload and Roles**

Though we worked as a group to reach our overarching goal and analysis, along the way we did divide up the responsibilities in the beginning. Robert focused on the initial gathering of our data and making sure that all attributes did not have missing values, were relevant to our analysis and formatted our data during preprocessing. Sarah worked as our code writer, as she worked mainly with Jupyter Notebook to create our data models for the attributes. Jack led our team through the different Machine Learning Models on Orange. He compared the methods we used in class and used these same steps to apply it with our dataset. On all other portions of our project, we worked side by side to complete the tasks.

**References**

1. Learning, U. C. I. M. (2016, December 14). *German credit risk*. Kaggle. Retrieved April 1, 2022, from https://www.kaggle.com/datasets/uciml/german-credit