Winter 2023

Credit worthiness tool for Credit Unions

Rylee Christoffersen
CCT College Dublin

Mario Ramalho
CCT College Dublin

Recommended Citation

Christoffersen, Rylee and Ramalho, Mario, "Credit worthiness tool for Credit Unions" (2023). *ICT*. 35.
https://arc.cct.ie/ict/35

CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

Module Title:	Problem Solving For IT
Assessment Title:	Capstone Project
Lecturer Name:	Muhammad Iqbal
Student Full Name:	Rylee Christoffersen Mario Ramalho
Student Number:	2019145 2019451
Assessment Due Date:	19 May 2023
Date of Submission:	19 May 2023

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Table of Contents

Capstone Project

Project Report

Mario Ramalho, CCT College
Student Number: 2019451

Rylee Christoffersen, CCT College
Student Number: 2019145

Introduction

Word Count: 5404 (without Appendices, References and TOC)

Video Link for Findings / Solutions: https://youtu.be/Bq131jLmh44

Core Objectives

The core objectives for the Capstone project were to mine data in order to create a tool for Credit Unions (and banks) that will evaluate customers credit worthiness based on an ethical standardised criteria that is transparent to all. We explored why this was necessary and explored how important it could be to the business. Our focus is on helping Credit Unions have a stronger online presence as the banking sector has been changing rapidly and moving online and Credit Unions are currently behind in the market in this regard (Gov.ie, 2019).This tool would help automate the credit approval process, reducing the underwriting time and allow customers to get answers quicker in regards to the potential of securing a line of credit. All this in just a few clicks or taps of the finger.

'AI credit scoring decisions are based on a lot of data, such as total income, credit history, transaction analysis, work experience, and even Google Analytics. In essence, scoring represents a mathematical model based on statistical methods and accounting for a large amount of information. As a result, credit scoring using AI provides more sensitive, individualized credit score assessments based on an array of additional real-time factors, giving access to finance to more people with income potential' (datrics, 2023).

Role & Responsibilities

Before we could split out any particular roles and responsibilities we as came together as a team came together started out assessing the project in class to determine the type of business we were considering and what makes most business sense for a potential project. Wat this point we both worked together on the strategic report to define key business objectives and how we were going to achieve this and what data sets could be used. The project was done very much in a TagTeam format whereby each of us cross checked and sense checked the approach from phase to phase.

Key tasks throughout the project (who worked on what)

- We both worked on defining what the project was.
- We both worked on the strategic document (tag team effort).
- Subsequent to this we both searched for data sets that might work for this type of project however Rylee sourced the core data set from which we based our project on.

- Data preparation was undertaken by Rylee and Mario (50-50 split)
- Date processing and data visualisation – again both of us worked on this (50-50 split).
- We also had regular check-ins to ensure the project was on track but also trouble shoot issues that we were having with on the data set.
- We both worked on the MLA models.

- We both worked on the Project Report, Presentation and Poster.

CRISP-DM

The Cross Industry Standard Process for Data Mining or known as the CRISP-DM is a model or methodology that helps standardise the data science process. It has six distinct phases and are as follows:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

In the subsequent pages we will delve more into how we implemented each of these steps through our application of the CRISP-DM model in our project.

Business Understanding

This section focuses on defining the purpose of this project from the point of view of the client (in this case hypothetical, but the client would be credit unions and banks here in Ireland). It also includes the project plan, a list of definitions and terminology used throughout the report and how the business defines and measures success.

Terminologies & Definitions

Client: refers to banks and credit unions in Ireland (hypothetical)

Customer(s): refers to customers of banks and credit unions

Applicant(s): refers to the customers of the German Bank which the data is derived

EDA: Exploratory Data Analysis

MLA: Machine Learning Algorithm

XGBoost: Extreme Gradient Boost

AUC: Area Under the Curve

ROC: Receiver Operating Characteristic

FP: False Positive

TP: True Positive

FN: False Negative

TN: True Negative

FPR: False Positive Rate

TPR: True Positive Rate

Business objectives

Our goal with this project is to understand which factors play a significant role in determining credit worthiness in customers of credit unions (and banks). Our data is obtained from a German credit dataset which is available on the UCI Repository website (Hofmann D. H., 1994). Ultimately, after analysing the data, we would like to create a predictive model that is able to determine whether or not a customer would be a good candidate or a bad candidate for getting approved for credit based on the information provided in an application. This predictive tool could then be used by credit unions and banks to simplify, standardise and quicken the credit approval process across the varying bodies. This model will be a binary classification as the results will be either 'Yes' they are a good candidate or 'No' they are not.

Project Plan

Week 1 - 3

1. Perform research to discover suitable datasets that are aligned with the scope of the project.
2. Identify the source of the dataset or datasets to be used.

Week 3 - 5

3. Perform an initial EDA to check for missing data, unusual data, relationships between the data in order to get a better understanding of the data.
4. Convert the data into a format that is useable by MLAs.
5. Perform Feature Extraction to uncover the significant factors that play a role in determining customer credit worthiness.

Week 5 - 8

6. Create generic models of various MLAs to run an initial test of performance and efficiency to determine which MLA's work best with data provided.
7. Select the model(s) with the highest performance for further development.
8. Fine tune the successful models to reach peak performance in predicting credit worthiness while minimising errors and losses.
9. Create project poster template.

Week 8 - 11

10. Add Phase 1 - 3 to Project Report.
11. Add Phase 1 - 3 to Project Poster.
12. Finalise tuning the MLA model(s).

Week 11 - 13

13. Add Phase 4 – 6 as well as Introduction and Conclusion to Project Report.
14. Add Phase 4 - 6 to Project Poster.

15. Format Python file (ensure comments are in place and is orderly and clean)

Week 13

16. Prepare for submission.

Business Success Criteria

What is the definition of a successful or desired result of the project study? From the client point of view it could be:

- The predictive model has an accuracy of 92% in determining whether a customer is credit worthy or not.
- The predictive models error rate in predicting a customer is credit worthy when they are, in truth, not worthy is less than 4%.

Inventory of Resources

- Personnel working on this project consist of Mario and I, Rylee
- Our dataset we used is a CSV file gathered from the UCI Repository
- Notions, Basecamp and GitHub are used for project management.
- The programming language we used to perform the study is Python
- The Python libraries we implemented consist of:
    - NumPy
    - Matplotlib
    - Seaborn
    - Pandas
    - SciPy
    - Sklearn
    - Patsy
    - Statsmodel
- Our classification MLAs:
    - Decision Tree Classifier
    - Random Forest Classifier
    - Logistic Regression
    - XGB Classifier
    - Gradient Boosting Classifier

There is a competitive advantage to be gained by the credit unions such as:

- Provide a stronger presence in the online banking sector
- Transparency attracts more customers due to increased trust
- Increases customer satisfaction (personalised)
- Increases customer retention
- Broader customer access to credit
- Fairer loans minus bias and discrimination

Market gap here in Ireland

We also identified that there is a market gap here in Ireland. KBC and Ulster bank left the Irish market 2023 there are only three large banks now in Ireland, providing an opportunity for the credit union to grow their offering.

Future focus

AI technology can be used to help implement an unbiased approach to credit rating and in turn can optimise and speed up the process. It can result in tailored bespoke loans. This type of approach is based on inclusive-led AI technology which could be used by credit unions/banks.

A project must be viable to the business from the outset otherwise it will not make financial sense to undertake in the first place.

Data Understanding

The data that we had sourced is relevant and key to understanding and also delivering on why this project makes business sense for the credit unions.

This phase of the project was about collecting the data set, examining the data to ensure its relevant to the project but looks at the various properties through the set, an even deeper dive into the data set, documenting any issues with the data and really examining it to ensure we could work with it. We had to examine what was missing and why but also determine any attributes that may be irrelevant to us. We couldn't progress into data preparation without examining our data set in-depth.

Our dataset

https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

The Research

To begin, we searched through various dataset repositories such as Kaggle, UCI Repository, data.gov.ie, and data.gov. From the various repositories we searched, we found one dataset that was relevant and practical for our project's needs. This dataset contains data gathered from the decision making process of credit approvals provided by a bank in Germany. Our intent was to find data from the banks within Ireland, but the data didn't seem to be available so we chose a bank within the EU that used fairly general questions to determine credit worthiness that was aligned with our goal. This dataset is available on the UCI Machine Learning Repository website (Hofmann P. , 1994). After performing EDA on the data, we wanted to create a predictive model based on the approvals and denials contained within the dataset.

Initial Data Analysis of the Two Dataset Versions

In truth, the data set contains two different versions of the dataset. Both versions contain the same amount of records, 1000. The initial one (the original one) contains a combination of categorical and numerical data. The categorical data is labelled using the letter A followed by two to three digits depending on the category. A Microsoft Word document was supplied to identify the significance of each label.

The second dataset is actually a slightly newer version of the dataset provided by Strathclyde University where the categorical data was converted to a numerical format to prepare it for use in machine learning algorithms. They used label encoding to convert the categorical data to a numeric version. They simply removed the letter and first digit, as in the case of a categorical value such as A34 and left the remaining digit so in this case 4. For values that contain the letter A and 3 following digits they removed the letter A and the first two digits, just leaving the last digit.

Initially we intended to use the second dataset that was in numerical format already as it appeared it would save us an extra step in converting the categorical variable to a corresponding numeric format. However, upon further examination we discovered that the two datasets weren't consistent with each other. The original dataset contained 21 variables in total. The numeric version contained 27. Quickly we realised that Column 1 and 27 of the numeric dataset were empty and potentially created accidently. Going through the rest of the columns, column by column, it was discovered that they weren't perfectly in order. In fact some generic variables appeared to have disappeared from the numeric version, even given its additional length. Columns 4, 8, 10, 15 and 17 of the generic dataset, don't correspond with any column in the numeric dataset. Column 4 is the purpose of the credit application. 8 is the instalment rate in percentage of disposable income. 10 is whether the applicant had other debtors, guarantors or a co-applicant. Column 15 was in regards to housing, whether they rented, owned or lived for free. 17 was their skill level in regards to a job.

The majority of the rest of the columns coincided with a numerical version in the numeric dataset. Although that still left several columns unidentified and seemingly uncorrelated with any column from the generic dataset. These columns were columns 17 to 25. The values were a binary value of 0 or 1. No document or information was provided as to how they were conceived. It looks like it could possibly be one hot encoding, but it is difficult to know where one begins and one ends as some lines have multiple 1 values in their records. It also raises the questions as to why they would use one hot coding for these variables and not some of the others.

We investigated and researched about the numeric dataset further, but could find no more evidence on how it was contrived, even after looking at the website of Strathclyde University as well as other sites (Gromping, 2019). We actually found many other studies who were also looking at this dataset and chose not to use the numeric version as many of the columns could not be explained. We decided to follow suit and abandon the idea of using the numeric version as it appeared would be far more complicated than initially perceived.

After reviewing the variables and their possible values, it was clear that some were collected and combined into one variable when it would make more sense and provide more clarity if the were collected separately. For example the Personal Status and Sex variable combines gender and marital status. For simplicity and clarity these values would be better separate, gender in one column and marital status in another as they duplicated some of the marital values since they needed them for both males and females. Another variable that could be separated into multiple is the credit history variable. Information regarding past credits could be in one variable and current credits could be in another. Two of the variable values could possibly mean the same thing. A30 is whether the applicant has had no credits taken so far or also all credits have been paid back duly. A31 is if all credits at this bank have been paid back duly. Those two values could be the same depending on the situation. Some of the values would be better off split into a binary variable of yes or no, such as “do you have existing credits”.

Exploring the Data Visually

Fig 2.2

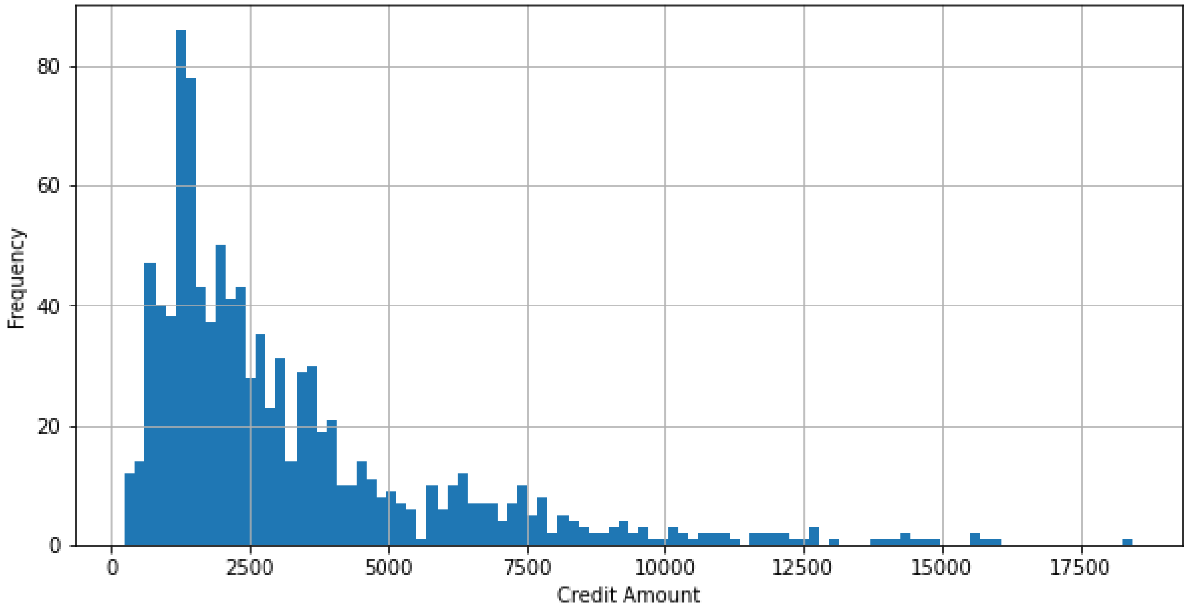


Fig 2.2 is a histogram of how frequent a credit amount was requested given the records in the German credit dataset.

By doing a count of the credit amounts requested and displaying it in a histogram, it is apparent that the amount requested varies greatly. However, the graph shows a positive skewness in the data where the majority of the credit amounts requested (75%) fall between 250 to 3972.25. This indicates that there are outliers in the dataset in regards to the credit amount requested. Given the nature of the data and the information collected, the data is not perceived as incorrect as it is possible to request this varied amount of money in a real setting such as the one supplied.

Fig 2.3

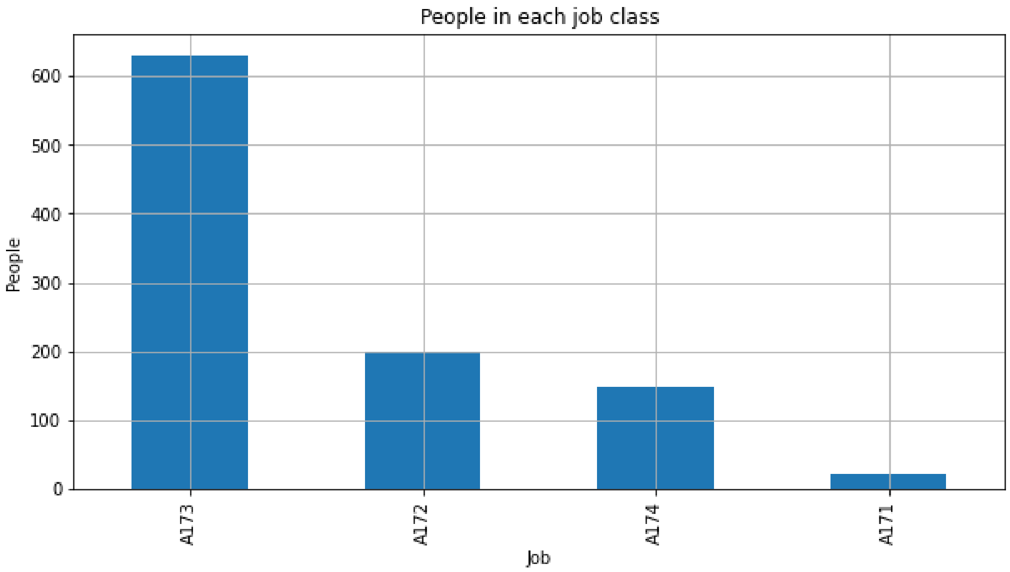


Figure 2.3 shows the amount of applicants who fall under each job category.

Legend:

A171 – Unemployed / Unskilled (non-resident)

A172 – Unskilled (resident)

A173 – Skilled Employee / Official

A174 – Management / Self-Employed / Highly Qualified Employee / Officer

As can be seen in the figure above, nearly 90% of the applicants are considered skilled employees with roughly 16% of those being highly skilled. Looking into the average age of the applicants in each job class group (Fig 2.4) it is possible to see that in general, all applicants in each group are relatively similar in age ranging from 34-41 years.

Fig 2.4

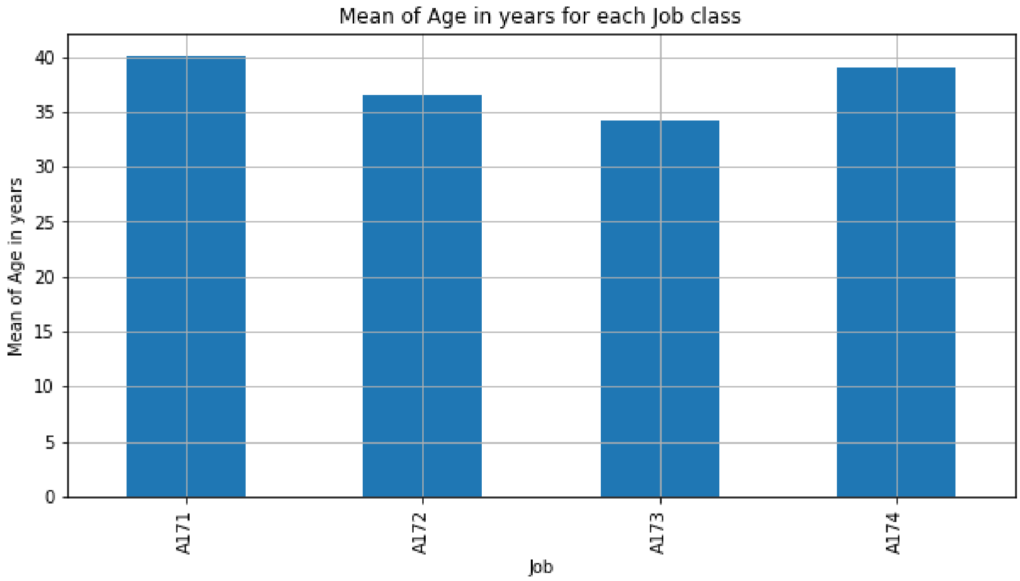


Figure 2.4 displays the average age of the people in each job class group.

Fig 2.5

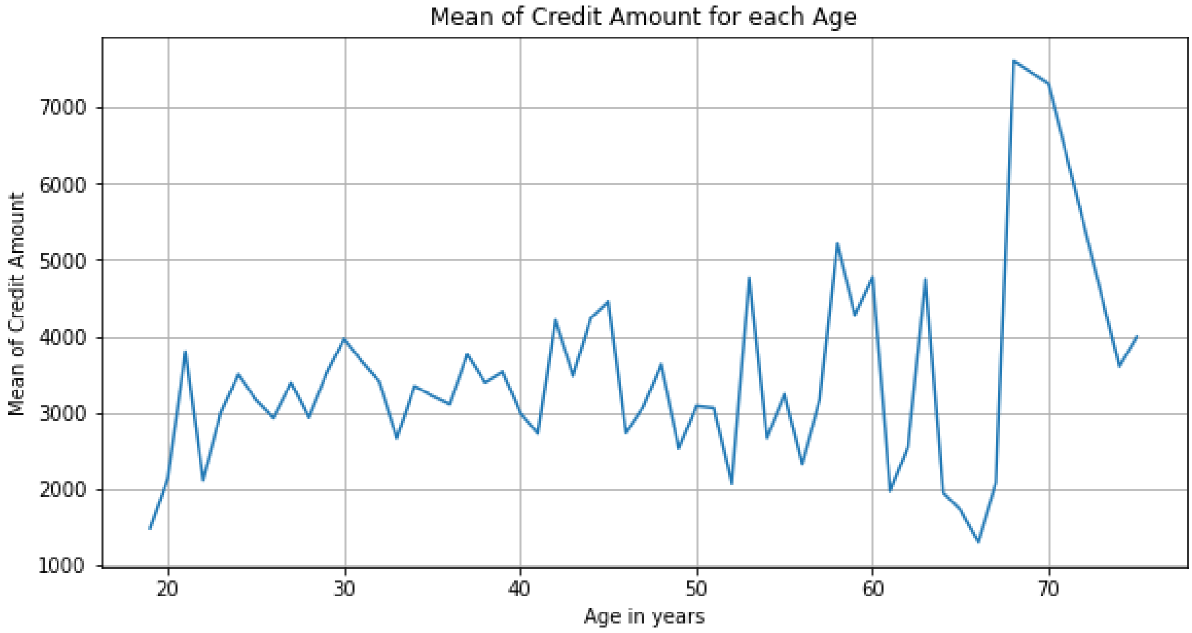
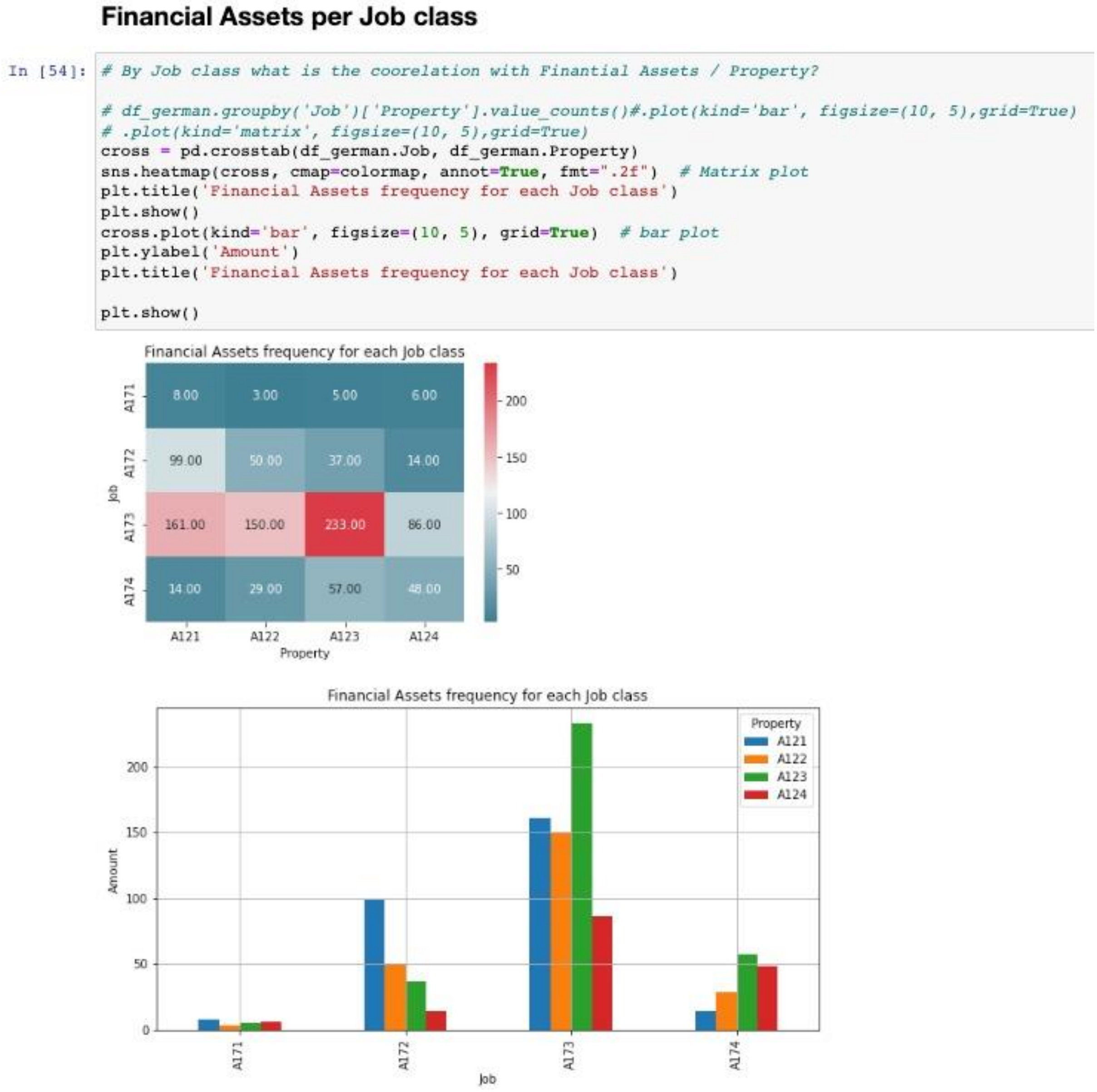


Figure 2.5 shows the average amount of credit requested by the people based on their age.

Above in Figure 2.2 we saw that there was a wide range of credit amounts requested by the applicants with the majority falling between 250 and 4000. By averaging the amount of credit requested by the people at each age we can see that the majority of the values consistently fall between this range of values. However it is possible to see a gradual increase in the amount of credit requested as the age of the person increases. As we reach 60 years of age we can see a slight decrease in the overall average credit amount requested until 68 years of age where we see a sharp increase in the average. The decrease is probable as less people in this age group are likely to be requesting loans and if so, will most likely be requesting for less money and are most likely only able to get a small amount of money due to their age being a risk factor. The sharp increase could be explained as there are less data available for the older age groups and with a few outlier values, it can drastically affect the overall average.

Fig 2.6



Financial Assets per Job class

```
In [54]: # By Job class what is the coorelation with Finantial Assets / Property?

         # df_german.groupby('Job')['Property'].value_counts()#.plot(kind='bar', figsize=(10, 5),grid=True)
         # .plot(kind='matrix', figsize=(10, 5),grid=True)
         cross = pd.crosstab(df_german.Job, df_german.Property)
         sns.heatmap(cross, cmap=colormap, annot=True, fmt=".2f")  # Matrix plot
         plt.title('Financial Assets frequency for each Job class')
         plt.show()
         cross.plot(kind='bar', figsize=(10, 5), grid=True)  # bar plot
         plt.ylabel('Amount')
         plt.title('Financial Assets frequency for each Job class')

         plt.show()
```

For the figure above we explored the financial assets of the 1000 people. Here we looked at who owned properties, and if they didn't own houses did they have other savings or cars. We also looked at who didn’t have any of these. What we found was that 233 of the A173 (skilled employee/official (blue collar worker)) had access to a car or other assets.

Legend:

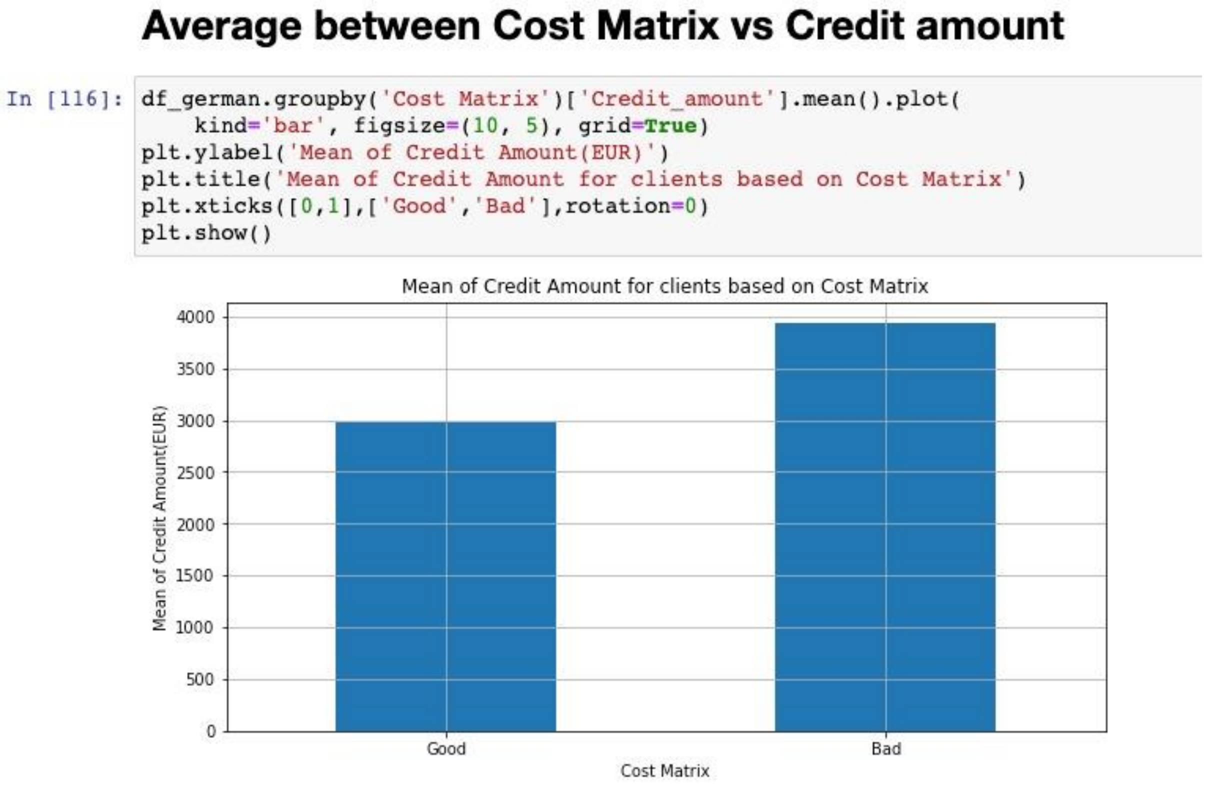
A121: Real estate

A122: if not A121 : building society savings agreement/ life insurance

A123: if not A121/A122 : car or other, not in attribute 6

A124: unknown / no property

Fig 2.7



Average between Cost Matrix vs Credit amount

```
In [116]: df_german.groupby('Cost Matrix')['Credit_amount'].mean().plot(
              kind='bar', figsize=(10, 5), grid=True)
          plt.ylabel('Mean of Credit Amount(EUR)')
          plt.title('Mean of Credit Amount for clients based on Cost Matrix')
          plt.xticks([0,1],['Good','Bad'],rotation=0)
          plt.show()
```

Mean of Credit Amount for clients based on Cost Matrix
Mean of Credit Amount(EUR)
4000
3500
3000
2500
2000
1500
1000
500
0
Good
Bad
Cost Matrix

Figure 2.7 represents the average of the total credit amount.

Data Understanding Summary

- Found a German Bank Dataset containing two datasets (generic and numeric)
- Analysed both generic and numeric datasets for applicability and ease of use in implementation of our MLA’s.
- Discovered inconsistencies and missing information between the numeric dataset and the generic one although the numeric one was contrived from the generic dataset and was used by Strathclyde University for their MLA’s.
- Column 4, 8, 10, 15 and 17 of the generic dataset are missing in the numeric one.
- Column 1 and 27 of the numeric dataset are empty (all null values).

- Column 17-25 of the numeric dataset contain binary values without labels and no information on how they were conceived.
- The insufficient information on how the numeric dataset was created and what the values mean as well as an explanation into the decision making process left the dataset unusable for our project, although initially perceived as an advantage in quickening the process.
- The generic dataset contains no missing values, no errors (it's clean).
- The generic dataset has 21 columns (7 numerical, 14 categorical)
- There are 1000 records in the dataset.
- Majority of applicants requested between 250-4,000 in credit amount.
- The credit amounts requested range from 250-18,424.
- The majority of applicants are skilled or highly skilled employees.
- The average age of applicants in each job class ranges between 34-41.
- The amount of credit requested tends to increase slightly as a person ages.

Data Preparation

This phase was all about preparing the data for modelling. This went through a number of process and phases to ensure our data could be used and definitely was the longest section of the project. This part of the project covered off a number of areas including:

- Inspecting and cleaning the data
- Constructing the data
- Integrating the data
- Formatting the data

Inspecting and Cleaning the data

After reading in the CSV file, the columns were labelled numerically so our first step was to change all of the columns labels to their proper name using the word document provided containing all the information about the dataset.

Constructing & Formatting the data

By using the original dataset, we were enabled to manipulate the format of the data in a way that suited our objective better. The numeric version that we abandoned to use, had converted the majority of the categorical variables to a numeric format via label encoding. The problem with this is that by doing so can mislead the machine learning algorithm's into thinking there is an ordinal relationship between the values and that one might carry more weight over the other when in reality, none of the values have this type of relationship. To prevent creating this false relationship within the values of the variables, we chose to use one-hot encoding on all the categorical variables. Using one-hot encoding, however, did alter the shape of our data giving it a higher dimensionality. We went from 21 columns to roughly

60. Luckily for us, the data from both the generic version and the numeric version did not contain any missing values, nor did it contain abnormal values that didn't fit the scheme.

For variables who only had a yes or no type question or either or, we converted it into a binary digit of 0 and 1, respectively. These were the last three columns: Telephone, Foreign Worker, and Cost Matrix. The Cost Matrix in the last column which represents whether an applicant is a good candidate for a loan or not was labelled as 1 if they are a good candidate and 2 if they are not. For simplicity and consistency, we converted the 2 into a 0 for bad candidates so the values were binary and easily readable by the MLA's.

Integrating the data

Since our source of data is all coming from the same source and same dataset, there was no need to integrate any other data into what we already had.

Modelling

Choosing Our Algorithms

With our data now prepared we can begin creating models through our chosen MLA's. As defined before, our problem is a binary classification problem. We are wanting to determine whether an applicant is worthy of being lent credit or not. We wanted to implement various classification algorithms using generic parameters to identify which algorithms outperform the others. The chosen MLA's are: Logistic Regression, Decision Tree, Random Forest, XGBoost and Gradient Boosting. We will define these more below and identify why we chose them.

Logistic Regression is a very popular machine learning algorithm which can be used for classification and predictive analytics. In machine learning it is considered a supervised MLA. In Binary Logistic Regression, the model estimates the probability of an event occurring, in our case being good for credit or not. If closer to 0 then it is determined not (0). If closer to 1 then it is determined as good (1). One potential downside to Logistic Regression is that it is prone to overfitting and can suffer from high dimensionality so more data cleaning and formatting may be necessary to improve the algorithm's success (IBM, 2023).

Decision Tree algorithms are very common and quite successful when it comes to classification problems such as ours. It is also considered a supervised MLA. Decision Trees work by breaking down the various predictor variables into yes or no decisions based on an if else then model that ultimately defines which class that record belongs to. Decision Trees are handy due to their ease in creating visuals of the decision making process. An advantage to Decision Trees is that they can handle both numerical and categorical variables and are less susceptible to outliers and missing values although we don't have any missing values in our data. This makes data preparation less time consuming. Decision Trees however can also be affected by over-fitting due to high variance and tend to predict slightly less accurately.

Random Forest is like the next level version of Decision Tree. Instead of just using one tree, it uses multiple that are generated with random samples of training data and random variations of independent variables are taken into account in each tree. The various trees make up the forest and each tree come up with its own conclusion. All conclusions are then taken into account and the majority vote becomes the decision. This helps overcome the problem of overfitting that is common in Decision Trees. Random Forest is like having multiple people analyse the data and come to a decision individually given only random parts of the data so that way bias doesn't come into play.

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, gradient-boosted decision tree. It offers a greater computational speed than traditional GBDT which makes it great for large datasets. Our dataset isn't large so this factor isn't so important. XGBoost works by building trees in parallel like Random Forest, but instead of using a bagging technique, it uses boosting which is where it takes a single weak model and iteratively improves it by combining it with other weak models which takes into account the error from the previous model to ultimately build a stronger model. It tends to be a highly accurate model. Where Random Forest reduces variance and overfitting, XGBoost does the opposite. It reduces bias and underfitting. So depending on the outcome of the testing, one might show to be better than the other (Nvidia, 2023).

Gradient Boosting, as previously mentioned above in the XGBoost, is a form of decision tree that uses the "boosting" technique to form the trees. Gradient Boosting runs sequentially traditionally where XGBoost runs in parallel. Each tree, or week learner makes a decision based on a single split and then uses a loss function to calculate the error and adds another weak learner, one at a time to improve that error based on a weight given (Brownlee, 2016).

Training Our Models

After preparing the data in a generic format to fit each algorithm, we split the data into a training group and a test group in a 70/30 split. For our initial test on the algorithms, we created a precision recall curve (PRC) of each algorithm. A precision recall curve can be broken up into two parts. Precision is the percentage of correct positive predictions over all positive predictions, including false positives. Recall is the percentage of correct positive predictions over all predictions that should have been positive. So it is a combination of True Positives and False Negatives. Both of these measures are quite important as False Positives mean loaning to people who shouldn't qualify and inherently are higher at risk for defaulting. False Negatives are missing out on trustworthy people to offer loans to, who would overtime make the credit unions a lot of money (Radecic, 2021).

The precision value is measured between 0 and 1 (percentage) and is based on the y-axis. Recall is measured the same and is based on the x-axis. An area under the curve score (AUC) is another useful measurement that measures exactly what it says. The higher the number, the better the algorithm performed. You can see the results of creating our models and gathering their PRC scores below.

Fig 4.1

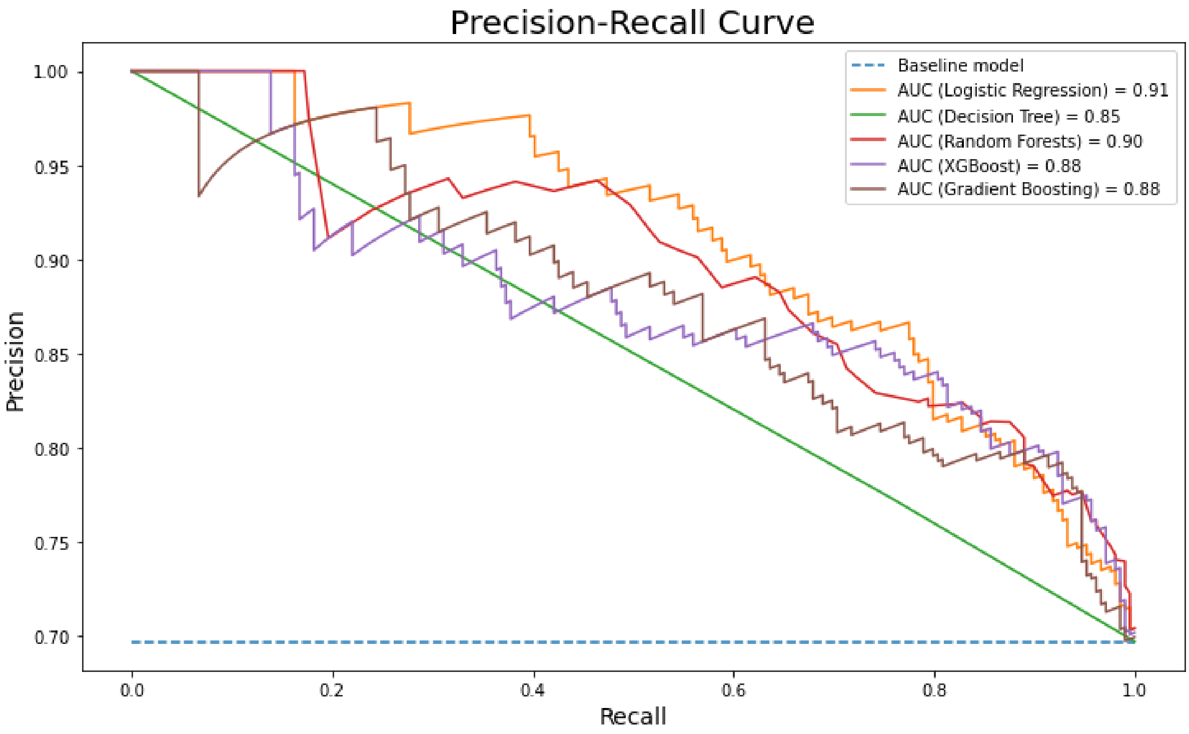


Figure 4.1 represents the results of the Precision and Recall of the generic models.

As you can see above, Logistic Regression outperformed the other values with an AUC score of 0.91. Random Forests came in a close second with .90. Both of these scores are quite good, meaning the models have high precision and high recall which is what we want. The others performed relatively similar, but not quite as good.

To get another perspective of the performance of the models, we created an ROC curve for each to compare how well each model categorises the credit worthiness of applicants. The ROC curve measures recall on the y-axis, or the True Positive Rate (TPR), over the False Positive Rate (FPR) on the x-axis. The FPR is calculated as FP / FP + TN. The AUC for the ROC measures the performance of the model. A lower score shows that the algorithm is not so good at predicting negative values, 0 in our case which corresponds to not creditworthy. As mentioned before, False Positives are risks to credit unions and banks, so the lower the score the riskier it is for a credit union to use this model as a reference for determining credit worthiness (Google, 2023).

Fig 4.2

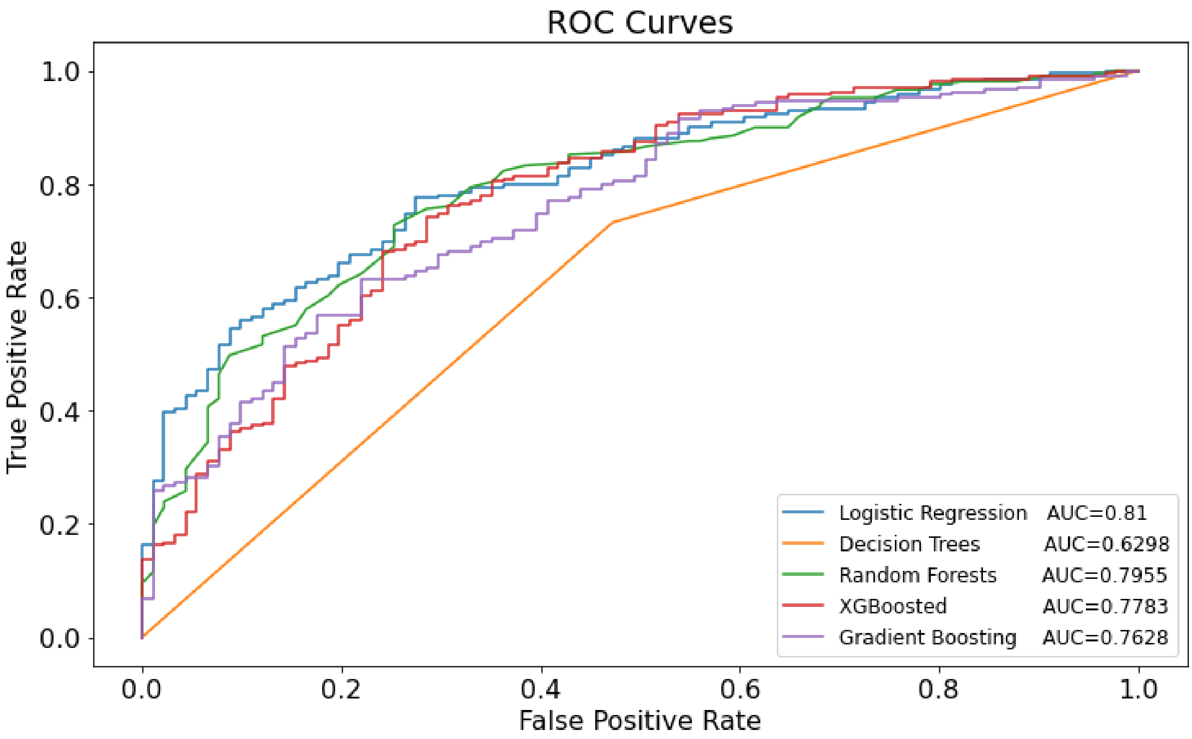


Figure 4.2 represents the results of the Receiver Operating Characteristic of the generic models

From the graph above, it is possible to see that all of the algorithms performed more poorly when measured in this way. Logistic Regression still outperformed the others with an AUC score of 0.81. Random Forests followed closely with 0.80 if we round the last digits up. Given that these two algorithms consistently performed better than the others, these will be the two algorithms that we focus on improving for our model.

Evaluation

In Figure 4.1, the PRC showed how good the algorithm was at predicting positive values. In Figure 4.2, the ROC showed how well the algorithm predicted negative values. Given that there are more applicants who were deemed good for credit than those that were deemed risky to give credit to, it makes sense why the algorithm wouldn't perform as well on the negative values as it did on the positive ones as it doesn't have as much information to go on. There are 700 creditworthy applicants and only 300 non-creditworthy applicants. To improve the performance of the algorithms on predicting negative values, it would be wise to increase the sample data of negative values to the equivalent of the positive values so there won't be a skew in the data.

Certain algorithms require different ways of preparing the data. For initially testing our models we prepared the data in a general way. However, this left the shape of our data with a high dimensionality. Not only this, but after one-hot encoding, we unintentionally created multicollinearity. Random Forest is less susceptible to this, but Logistic Regression is. Ideally it would be good to go back and prepare the data in a way that suits each algorithm best. One option for Logistic Regression is to drop some of the columns that were one-hot encoded. Scikit Learn has an option for this where you can drop the first column that is created, but this could be at the expense of losing important data. Going category by category and reducing the least important one is time consuming and requires a great deal of documenting and organising in order to keep track of what is not included.

Decision

At the current moment, both models are insufficient for meeting the business success criteria. Given the high amount of False Positive predictions, it is far too risky to be used by credit unions in this current state. The potential reasons for this could be summed up as:

- Not enough data to created accurate predictions with minimal loss
- Skewed data (more positive values than negatives)
- Multicollinearity (highly correlated predictor variables offsetting weights)
- Curse of Dimensionality

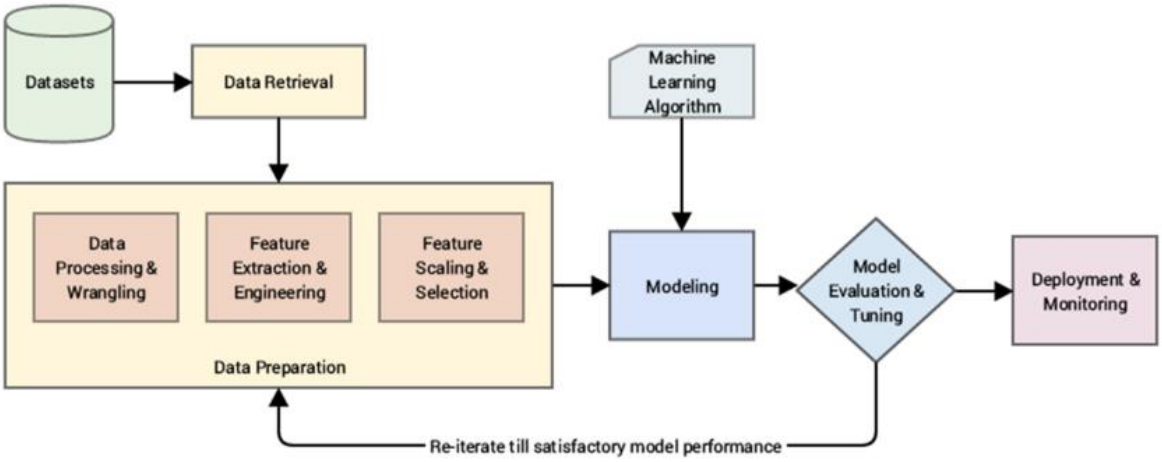
Summary of Potential Actions

- Prepare Data for each algorithm of choice for optimisation and reduction of noise
- Prime hyperparameters for each algorithm using GridSearchCV
- Supplement the skewed data with sample data to create evenness
- Reduce multicollinearity by removing highly correlated variables
- Reduce the dimensionality by selecting key features

Deployment

Conclusion

We defined success as creating a model that could predict whether or not a customer of a credit union, or bank (the applicant) is worthy of being loaned credit. In order to be successful our model would have to predict with a 92% rate of accuracy. Not only this, but the loss rate or the rate at which it falsely predicts a customer is worthy when they are not must be below 4%. After exploring the data, preparing it for our machine learning algorithms and running tests on the models, it was deemed that our model could not meet the business needs at the current moment, therefore was unsuccessful. As discussed in the evaluation, there are areas in which we could study further to potentially improve the quality of the models. The data we were able to procure is relatively quite small so even beyond implementing the suggested improvements above, there is still a chance that the lack of data could continue to produce an insufficient model. Gathering more data could play a significant role in training the model properly for a real life application. For this we need to go back and reiterate through the process until we can come up with a viable solution.

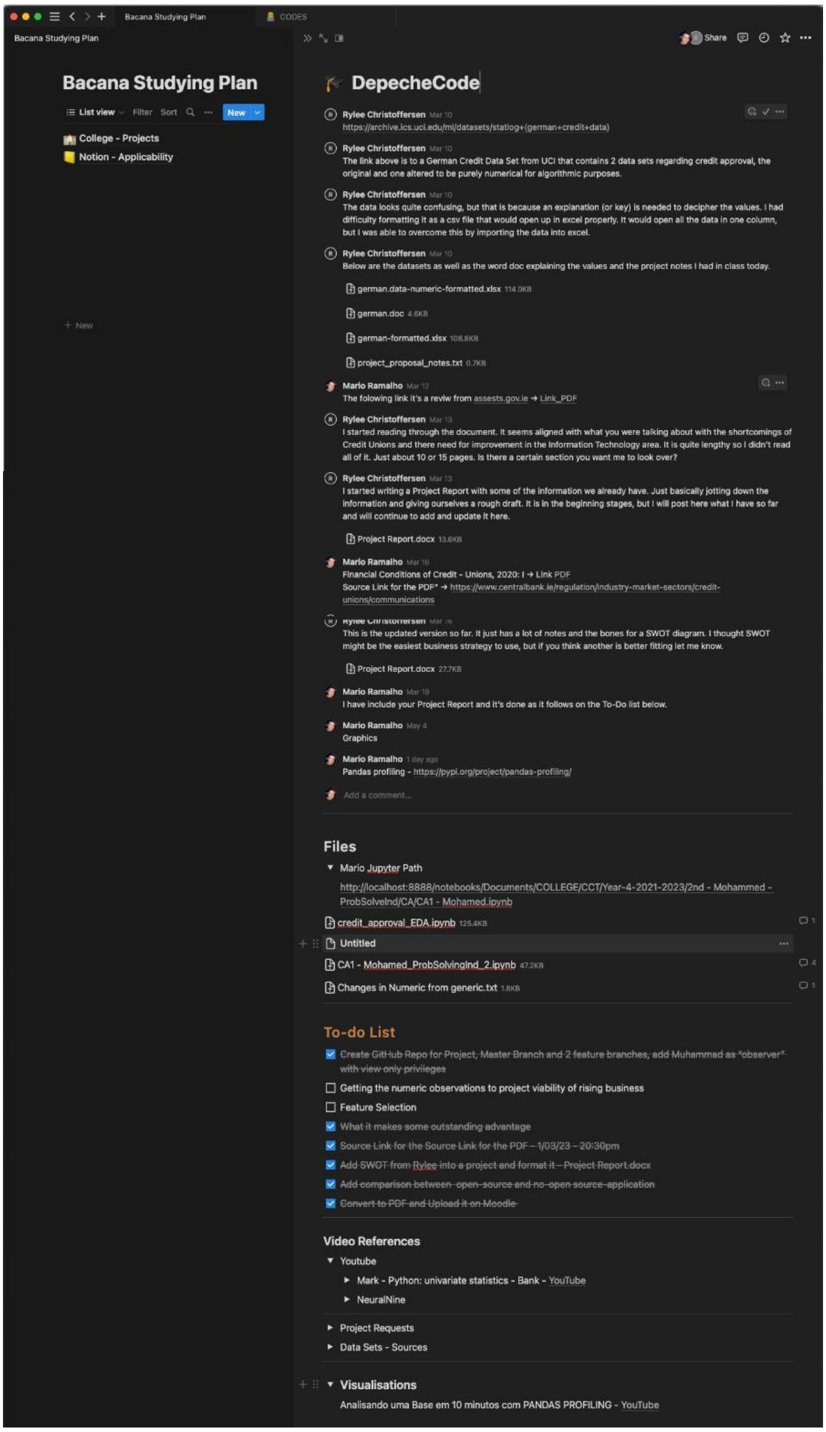

Datasets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

(Dharshini, 2021)

Appendix

Evidence of group work.

Notions
Notions communication between Rylee and I which documented the step by step approach taken on the Capstone project.



Bacana Studying Plan

List view · Filter · Sort · New

- College - Projects
- Notion - Applicability

+ New

DepecheCode

Rylee Christoffersen Mar 10
https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Rylee Christoffersen Mar 10
The link above is to a German Credit Data Set from UCI that contains 2 data sets regarding credit approval, the original and one altered to be purely numerical for algorithmic purposes.

Rylee Christoffersen Mar 10
The data looks quite confusing, but that is because an explanation (or key) is needed to decipher the values. I had difficulty formatting it as a csv file that would open up in excel properly. It would open all the data in one column, but I was able to overcome this by importing the data into excel.

Rylee Christoffersen Mar 10
Below are the datasets as well as the word doc explaining the values and the project notes I had in class today.

german.data-numeric-formatted.xlsx 114.0KB

german.doc 4.6KB

german-formatted.xlsx 108.8KB

project_proposal_notes.txt 0.7KB

Mario Ramalho Mar 12
The folowing link it's a reviw from assests.gov.ie → Link_PDF

Rylee Christoffersen Mar 13
I started reading through the document. It seems aligned with what you were talking about with the shortcomings of Credit Unions and there need for improvement in the Information Technology area. It is quite lengthy so I didn't read all of it. Just about 10 or 15 pages. Is there a certain section you want me to look over?

Rylee Christoffersen Mar 13
I started writing a Project Report with some of the information we already have. Just basically jotting down the information and giving ourselves a rough draft. It is in the beginning stages, but I will post here what I have so far and will continue to add and update it here.

Project Report.docx 13.6KB

Mario Ramalho Mar 16
Financial Conditions of Credit - Unions, 2020: I → Link PDF
Source Link for the PDF* → https://www.centralbank.ie/regulation/industry-market-sectors/credit-unions/communications

Rylee Christoffersen Mar 16
This is the updated version so far. It just has a lot of notes and the bones for a SWOT diagram. I thought SWOT might be the easiest business strategy to use, but if you think another is better fitting let me know.

Project Report.docx 27.7KB

Mario Ramalho Mar 19
I have include your Project Report and it's done as it follows on the To-Do list below.

Mario Ramalho May 4
Graphics

Mario Ramalho 1 day ago
Pandas profiling - https://pypi.org/project/pandas-profiling/

Add a comment...

Files

▼ Mario Jupyter Path
http://localhost:8888/notebooks/Documents/COLLEGE/CCT/Year-4-2021-2023/2nd - Mohammed - ProbSolveInd/CA/CA1 - Mohamed.ipynb

credit_approval_EDA.ipynb 125.4KB 1
Untitled
CA1 - Mohamed_ProbSolvingInd_2.ipynb 47.2KB 4
Changes in Numeric from generic.txt 1.8KB 1

To-do List

- ☑ ~~Create GitHub Repo for Project, Master Branch and 2 feature branches, add Muhammad as "observer" with view only privileges~~
- ☐ Getting the numeric observations to project viability of rising business
- ☐ Feature Selection
- ☑ ~~What it makes some outstanding advantage~~
- ☑ ~~Source Link for the Source Link for the PDF – 1/03/23 – 20:30pm~~
- ☑ ~~Add SWOT from Rylee into a project and format it – Project Report.docx~~
- ☑ ~~Add comparison between open-source and no-open source-application~~
- ☑ ~~Convert to PDF and Upload it on Moodle~~

Video References

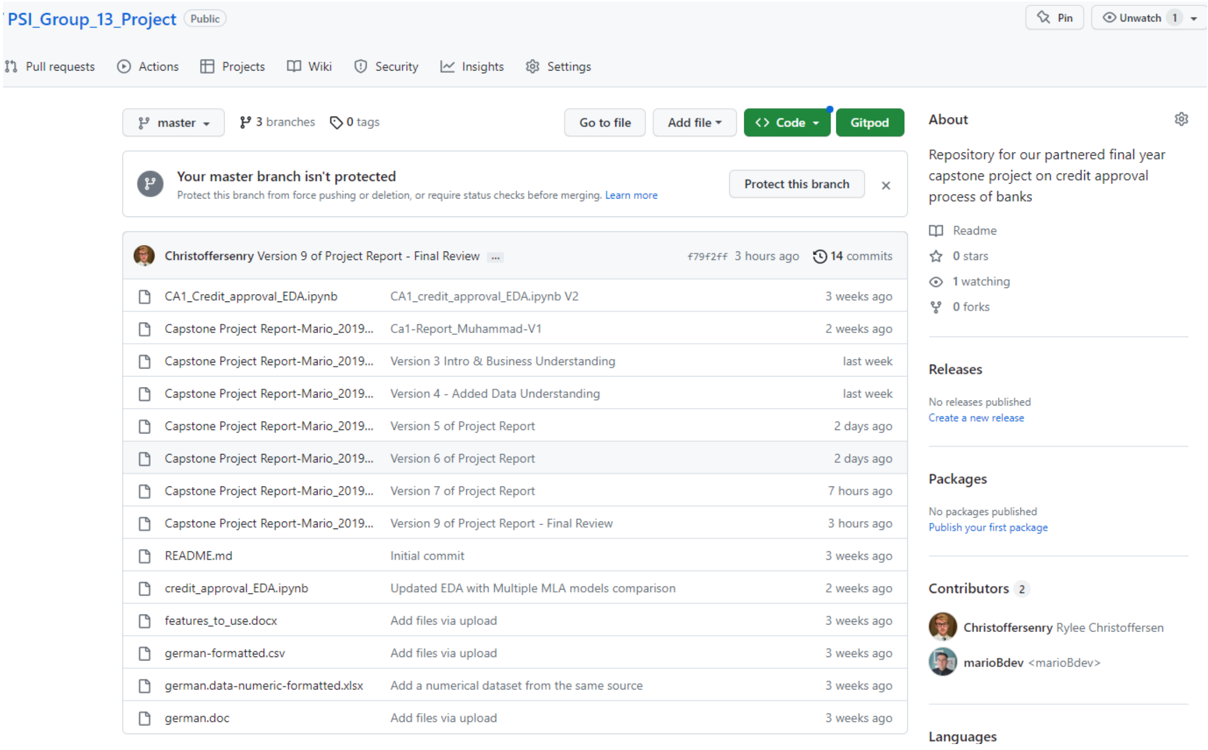
▼ Youtube
- ▸ Mark - Python: univariate statistics - Bank - YouTube
- ▸ NeuralNine

▸ Project Requests
▸ Data Sets - Sources

▼ **Visualisations**
Analisando uma Base em 10 minutos com PANDAS PROFILING - YouTube

Github

Github code repository between Rylee and me.



Your master branch isn't protected

Protect this branch from force pushing or deletion, or require status checks before merging. Learn more

Protect this branch

Christoffersenry Version 9 of Project Report - Final Review … f79f2ff 3 hours ago **14** commits

CA1_Credit_approval_EDA.ipynb	CA1_credit_approval_EDA.ipynb V2	3 weeks ago
Capstone Project Report-Mario_2019...	Ca1-Report_Muhammad-V1	2 weeks ago
Capstone Project Report-Mario_2019...	Version 3 Intro & Business Understanding	last week
Capstone Project Report-Mario_2019...	Version 4 - Added Data Understanding	last week
Capstone Project Report-Mario_2019...	Version 5 of Project Report	2 days ago
Capstone Project Report-Mario_2019...	Version 6 of Project Report	2 days ago
Capstone Project Report-Mario_2019...	Version 7 of Project Report	7 hours ago
Capstone Project Report-Mario_2019...	Version 9 of Project Report - Final Review	3 hours ago
README.md	Initial commit	3 weeks ago
credit_approval_EDA.ipynb	Updated EDA with Multiple MLA models comparison	2 weeks ago
features_to_use.docx	Add files via upload	3 weeks ago
german-formatted.csv	Add files via upload	3 weeks ago
german.data-numeric-formatted.xlsx	Add a numerical dataset from the same source	3 weeks ago
german.doc	Add files via upload	3 weeks ago

About

Repository for our partnered final year capstone project on credit approval process of banks

Readme
0 stars
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

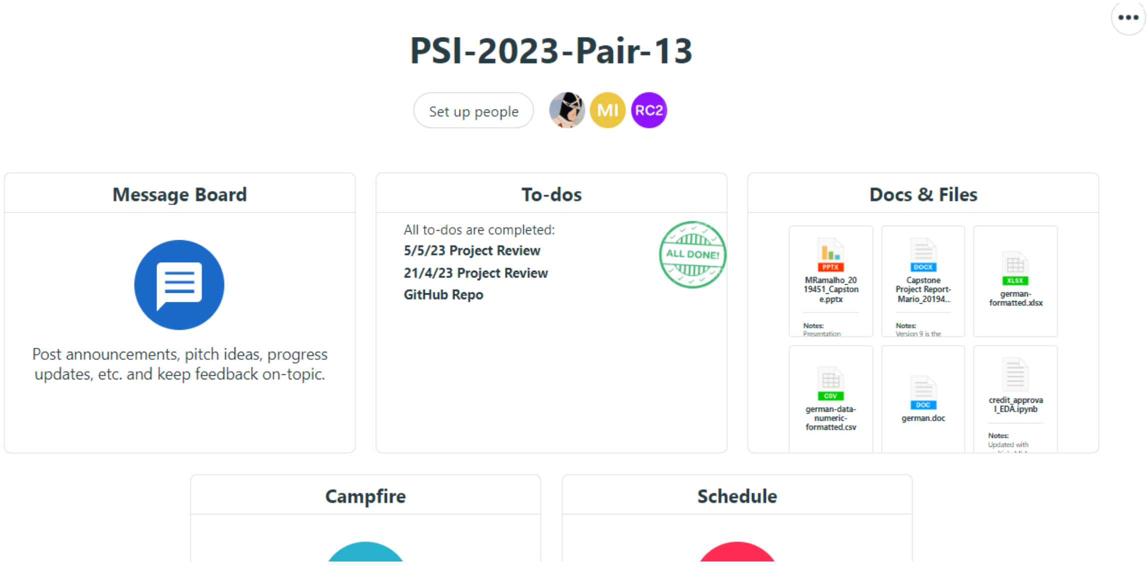
No packages published
Publish your first package

Contributors 2

Christoffersenry Rylee Christoffersen
marioBdev <marioBdev>

Languages

BaseCamp



PSI-2023-Pair-13

Message Board

Post announcements, pitch ideas, progress updates, etc. and keep feedback on-topic.

To-dos

All to-dos are completed:
5/5/23 Project Review
21/4/23 Project Review
GitHub Repo

Docs & Files

PPTX MRamalho_2019451_Capstone.pptx
Notes: Presentation

DOCX Capstone Project Report-Mario_20194...
Notes: Version 9 is the

XLSX german-formatted.xlsx

CSV german-data-numeric-formatted.csv

DOC german.doc

credit_approval_EDA.ipynb
Notes: Updated with

Group Reflection

Rylee’s Review:

For me, I feel like I learned a lot from this project. For the most part the subjects we have studied this year were quite new for me and it was definitely a challenge. For this project, I found it difficult to assess how much time should be given to each task as I haven’t done something quite like this before. Often times I felt like I got caught up in things that I should have spent less time on. As always working in groups or partnerships, like in this case, bring it's own set of challenges. Communication is key and I felt like we could have benefited from more frequent check-ins. I also felt like we could have written down and outlined tasks to be completed (with deadlines) more often. I was happy to see that we both contributed a lot to this project and were patient with the other as with completely different schedules meant we were working on the project at different times. Definitely one of the biggest unforeseen challenges was in finding the data. I didn’t realise how difficult it would be. Overall, I am pretty proud of the efforts that we both put, especially given all the obstacles that we faced.

Mario’s Review:

What was the objective you set out to achieve?

Essentially Rylee and me needed to create a data visualisation set that would help a business to automate the credit approval process, reducing the underwriting time and allow customers to get answers quicker in regards to the potential of securing a line of credit. So as part of the project had to source the data set, clean and format the data, model it, deploy and ultimately create meaningful results that could be used by the business.

Did you achieve this objective?

We feel like we did achieve the objective in that we prepared the datasets, clean and format it, model it. Unfortunately after modelling it the dataset did not provide enough information to deploy. It wasn’t the answer we wanted but I still feel I achieved the objective.

Did the team achieve/not achieve your objectives?

As mentioned earlier in the report we split two datasets and in regards to achieving the objective I feel we both were successful in this. As mentioned it wasn’t exactly the answered we were looking for but we felt that we approached the project in a methodical way.

What challenges did you face?

The one area I felt was weakest was identifying the entities and languages (for example one attribute was called the instalment percentage definition). I don’t work in a bank and trying to understand this type of financial language and that words used in datasets was tough. In a real life world we would also be working embedded in the industry and have a better understanding of a data set so it felt forced.

Sourcing the right type of dataset was tough, then also trying to figure out if we could work with a dataset or discard it. In total we had identified at least 3 different data sets to work from. Out of these 3 we initially thought we could work with 1 only but in the end we worked with two data sets. Again this did add to the complexity of the project.

Working separately was also tough as we had to split the initial process, I started work on the object based one and Rylee work on the numerical one. Then we kept having to liaise with each other over a number of weeks. This was also challenging as we both worked at different times and only met with each other twice a week in college. Plus we both had other projects that needed to be worked on in this time frame.

The other challenge I faced was as the dataset found it hard to condense the output down to a summary. There was so much data mined that I struggled with this.

How did you overcome these challenges?

Online tools help enormously especially for the language being used in the dataset, every entities needed to be researched and cross checked to ensure it was relevant. This added a huge amount of time to the project.

In sourcing this data set we used to validate each other choices in what dataset to keep and what to discard. Teamwork was used to solve this challenge.

In working separately – I tried to ensure there was constant reviews and check-ins with Rylee but also with our lecturer. I tried to do as many face to face on location in college campus.

Condensing the data for the PPT and poster – this took time and tried to select what we thought as most relevant to the original brief.

What have the team learned from this?

Communication is key. We used Notebook and lecture provided a github, bandcamps but meeting face to face was really useful to really work out and tease through the project. Also data preparation and machine learning is a process, the end output isn't necessarily what you hope for and process is key.

References

Brownlee, J. (2016, September 9). *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

Congdon, L. (2015, February 3). *8 advantages of using open source in the enterprise.* Retrieved from The Enterprisers Project: https://enterprisersproject.com/article/2015/1/top-advantages-open-source-offers-over-proprietary-solutions

datrics. (2023, March 19). *Credit Scoring Using Maching Learning.* Retrieved from datrics: https://www.datrics.ai/credit-scoring-using-machine-learning

Deloitte. (2023, March 19). *Artificial Intelligence for Credit Risk Management.* Retrieved from Deloitte: https://www2.deloitte.com/cn/en/pages/risk/articles/artificial-intelligence-for-credit-risk-management.html

Dharshini, P. (2021, June 13). *4 Ways to Handle Insufficient Data In Machine Learning!* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/06/4-ways-to-handle-insufficient-data-in-machine-learning/

E. R., S. (2021, June 17). *Understand Random Forest Algorithms With Examples (Updated 2023)*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Google. (2022, February 2). *Why Open Source?* Retrieved from Google Open Source: https://opensource.google/documentation/reference/why

Google. (2023, May 19). *Classification: ROC Curve and AUC*. Retrieved from Machine Learning: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:text=An%20ROC%20curve%20(receiver%20operating,True%20Positive%20Rate

Gov.ie. (2019, February 7). Retrieved from Review of Implementation of the Recommendations in the Commission on Credit Unions Report: https://www.gov.ie/en/publication/ee48d7-review-of-implementation-of-the-recommendations-in-the-commission-on/

Gromping, U. (2019, November 29). SouthGermanCreditData:CorrectingaWidely UsedDataSet. Berlin, Germany: Facbereich II.

Hofmann, D. H. (1994, November 17). *Statlog (German Credit Data) Data Set.* Retrieved from UCI : https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Hofmann, P. (1994, November 17). *Statlog (German Credit Data) Data Set.* Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

IBM. (2023, May 17). *What is logistic regression?* Retrieved from IBM: https://www.ibm.com/topics/logistic-regression

Klein, A. (2019, April 11). *Credit denial in the age of AI.* Retrieved from Brookings: https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/

McQuaid, D. (2020, February 26). Random Forest in Python. Dublin, Ireland.

Nvidia. (2023, May 17). *XGBoost*. Retrieved from Nvidia: https://www.nvidia.com/en-us/glossary/data-science/xgboost/

Radecic, D. (2021, January 4). *Precision-Recall Curves: How to Easily Evaluate Machine Learning Models in No Time*. Retrieved from Towards Data Science: https://towardsdatascience.com/precision-recall-curves-how-to-easily-evaluate-machine-learning-models-in-no-time-435b3dd8939b

Schaefer, L. (2022, February 12). *The Top 4 Reasons Why You Should Use MongoDB.* Retrieved from MongoDB: https://www.mongodb.com/developer/products/mongodb/top-4-reasons-to-use-mongodb/

Stojiljkovic, M. (2020, January 1). *Logistic Regression in Python*. Retrieved from Real Python: https://realpython.com/logistic-regression-python/

Zach. (2022, April 6). *How to Plot Multiple ROC Curves in Python (With Example)*. Retrieved from Statology: https://www.statology.org/plot-multiple-roc-curves-python/

Zach. (2022, October 12). *How to Test for Multicollinearity in Python*. Retrieved from Statology: https://www.statology.org/multicollinearity-in-python/#:~:text=The%20most%20straightforward%20way%20to,between%201%20and%20positive%20infinity.