**Capstone Project**   
**Project Report**

**May 2023**

Word Count: 5,000 words

**Introduction**

**Core Objectives**  
The core objectives for the Capstone project was 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.

**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.

**Cross Industry Standard Process**

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

EDA: Exploratory Data Analysis

MLA: Machine Learning Algorithm

## 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

1. 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.
2. Convert the data into a format that is useable by MLAs.
3. Perform Feature Extraction to uncover the significant factors that play a role in determining customer credit worthiness.

Week 5 - 8

1. 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.
2. Select the model(s) with the highest performance for further development.
3. Fine tune the successful models to reach peak performance in predicting credit worthiness while minimising errors and losses.
4. Create project poster template.

Week 8 - 11

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

Week 11 - 13

1. Add Phase 4 – 6 as well as Introduction and Conclusion to Project Report.
2. Add Phase 4 - 6 to Project Poster.
3. Format Python file (ensure comments are in place and is orderly and clean)

Week 13

1. 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 MachineLearning Repository website (Hofmann, 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.

### Column Definitions of the Original German Dataset

#### Fig 2.1

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | Column Name | Descripiton | Type of Variable |
| 1 | Status Existing Checking Account | Amount of Money in Checking Account | Categorical |
| 2 | Duration (months) | Duration of Credit Loan Requested | Numerical |
| 3 | Credit History | Whether a customer has had credits in the past or currently, how they are/were paid, do they have other credits not at this bank | Categorical |
| 4 | Purpose | Car, Furniture, TV, Repairs, etc | Categorical |
| 5 | Credit Amount | Amount of Credit Requested | Numerical |
| 6 | Savings Accounts / Bonds | Amount of money in savings | Categorical |
| 7 | Present Employment Since | How long the customer has been employed (currently) | Categorical |
| 8 | Instalment Rate in % of Disposable Income | Percentage of disposable income permitted to be extended in a loan | Numerical |
| 9 | Personal Status and Sex | Whether Male or Female and whether married, divorced/separated, single or widowed | Categorical |
| 10 | Other Debtors / Guarantors | Whether there is a co-applicant or guarantor on this loan or not | Categorical |
| 11 | President Resident Since | How long the customer has lived at their current residence | Numerical |
| 12 | Property | What type of property is owned or if none is owned | Categorical |
| 13 | Age (Years) | How old is the customer | Numerical |
| 14 | Other Instalment Plans | Other types of instalment plans | Categorical |
| 15 | Housing | Whether the customer owns, rents or lives freely | Categorical |
| 16 | Number of Existing Credits at This Bank | How many existing credits the customer has at this bank | Numerical |
| 17 | Job | The level of skill the customer’s current role has or whether they have a job | Categorical |
| 18 | Number of People Being Liable to Provide Maintenance For | How many people are dependent on the income of the customer | Numerical |
| 19 | Telephone | Whether the applicant has a phone or not | Categorical (binary) |
| 20 | Foreign Worker | Whether the applicant is a foreign worker or not | Categorical (binary) |
| 21 | Cost Matrix | Whether the customer is a good or bad to loan money to | Categorical (binary) |

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

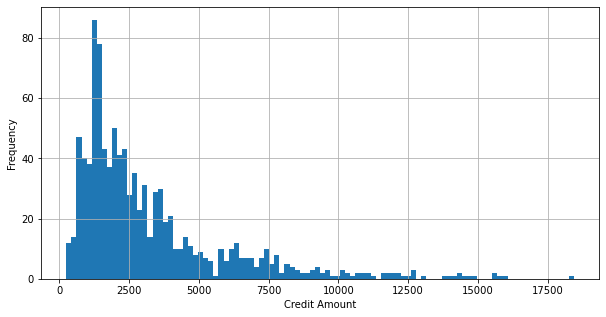


Fig 2.2 is a histogram of how frequent a credit amount was requested given the records in the German credit dataset.

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

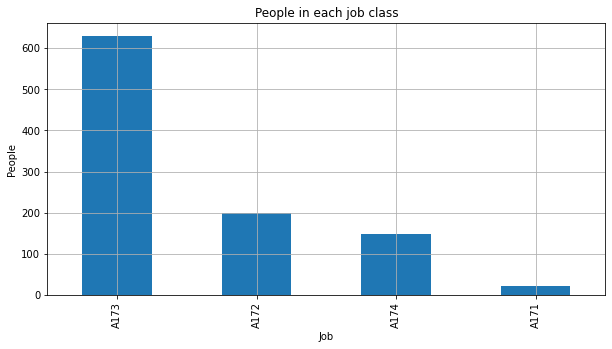


Figure 2.3 shows the amount of applicants who fall under each job category.

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

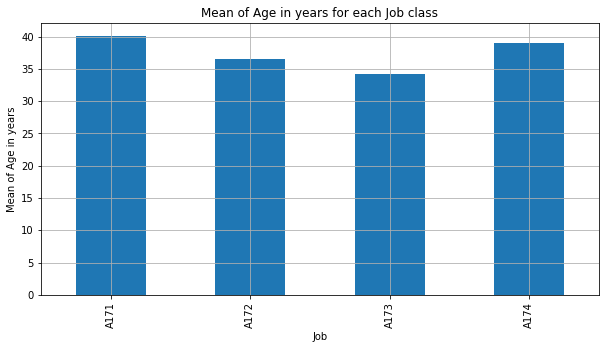


Figure 2.4 displays the average age of the people in each job class group.

#### Fig 2.5

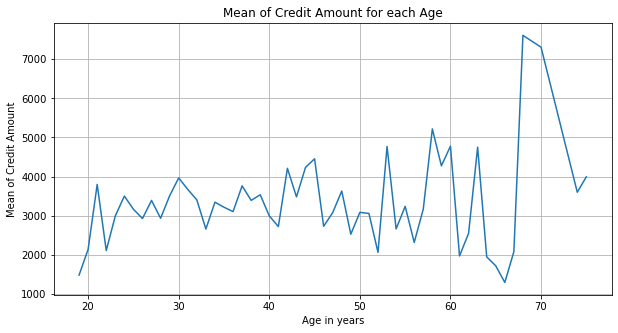


Figure 2.5 shows the average amount of credit requested by the people based on their age.

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.

## 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.

With this predictive model, customers will be able to determine the probability of being approved while also providing clarification as to why they would be a strong or weak candidate and offer tips as to how it could be improved based on the areas that were lacking.

*‘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)*.*

By widening the fields of data that is collected we also expand our data set to look at patterns to consider customers credit rating on a much wider level, not just by credit history, loan paybacks and income.

The different types of data mined can be as follows:

* Historical data
* Digital footprints
* Predictive analytics
* Alternative data (shopping history, property records, spending habits)
* Realtime transactional data

**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

### Implementing the chosen dataset

**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 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**

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 comes 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).

**Evaluation**

For the evaluation of the data we needed to check the quality and reliability of the data, sort through and then classify the data, and following this perform tests and cross check the results.

**Deployment**

Xxxxxxxxxxxx

**Conclusion**

Following the processing and visualations of our capstone project we conclude that:

* Xxxxxxxxxxxxxxx
* Xxxxxxxxxxxxxxx
* Xxxxxxxxxxxxxxx
* Xxxxxxxxxxxxxxx

**Appendix**

Evidence of group work.

**Group Reflection**

Xxxxxxxxxxxx

# References

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

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

*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/

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)

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/

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/