**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**

## 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.

In truth, the data set contains two different versions of the dataset. 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 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. Going through 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 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.

### Implementing the chosen dataset

By using the original dataset, we were enabled to manipulate the format of the data in a way that suits 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 than 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. 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. 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.

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

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 datasets**  
The first data set that we had sourced was useful to the project but we quickly discovered that it was an object based dataset and when we examined it closely we just couldn’t work with it. It simply would not provide an accurate data visualisation project for us to work from. As a result we had to locate a second data set. This was taken from the same source but was a numerical version which had 27 columns.

This particular dataset was something we could work with because XXXXXXXXXXXXXXXXXXX.

**Our dataset findings were as follows:**

1. On the numerical dataset as the whole first column was not a number NAN
2. There was no correlation between the last five attributes.
3. There was no missing data
4. XXXxxxxxxxxxxxx
5. Xxxxxxxxxxxxx
6. Xxxxxxxxxxxxxxx
7. XXxxxxxxxxxxxxxx

**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**

As mentioned on inspection of the data we removed the first column as that was NAN.

We also dropped the last five attributes (column 21–27) as there was no way of filling this information as it may skew the data.. Then we started the process of looking for correlations, and looking for which columns will be useful to help generate visualisations from. We need to ascertain which columns to related to different columns to give some useful data to mine. These are as follows:

* XXXXXXXXXXXXX
* XXXXXXXXXXXXX
* IXXXXXXXXXXX
* XXXXXXXXXXXX
* XXXXXXXXXXXX
* XXXXXXXXXXXX
* XXXXXXXXXXXX

**Constructing the data**

We also then constructed the data into a format that we know would provide optimal results. XXXXXXXXXXXXXXXXXXXXX

**Integrating the data**

We also integrated another data set XXXXXXXX

**Formatting the data**

Lastly we formatted the data XXXXXXXXX

**Modelling**

Once our data set was ready to go we could then start the modelling phase or creating diagrams of the data. We did this by XXXXXXXXXXXXXXXXXX

**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/