|  |  |
| --- | --- |
|  | |
| Python Machine Learning Project  Data and Web Mining CA – Report | |
|  | |
| Ciaran Finnegan / Dermot Madsen  Student No : 10524150 / 10522567  10/02/2020 |  |
|  |  |

List of contents

1 Project Overview 3

1.1 High Level Description 3

1.2 The CRISP-DM Methodology / Reference Model 3

1.3 Development Environments 6

2 Business Understanding 7

2.1 Determine Business Objectives 7

2.2 Assess Situation 8

2.3 Determine Data Mining Goals 9

2.4 Produce Project Plan 9

3 Data Understanding 11

3.1 Collect Initial Data 11

3.2 Descibe Data 12

3.3 Explore Data 13

3.4 Verify Data Quality 13

4 Data Preparation 14

4.1 Select Data 14

4.2 Clean Data 14

4.3 Construct Data 14

4.4 Integrate Data 14

4.5 Format Data 14

5 Modeling 15

5.1 Select Modeling Technique 15

5.2 Generate Test Design 15

5.3 Build Model 15

5.4 Assess Model 15

6 Evaluation 16

6.1 Evaluate Results 16

6.2 Review Process 16

6.3 Determne Next Steps 16

7 Deployment 17

7.1 Plan Deployment 17

7.2 Plan Monitoring and Maintenance 17

7.3 Produce Final Report 17

7.4 Review Project 17

8 Conclusion 18

8.1 Conclusion.. 18

9 Appendices and References 19

9.1 Appendix A 19

9.2 References 19

# Project Overview

## High Level Description

This document covers our planned approach and execution of a data mining analysis on a dataset relating to the assessment and prediction of wine quality.

Following the CRISP-DM model, we laid out an objective for this Continuous Assessment exercise and followed a series of steps, often iteratively to arrive at a predictive model for wine quality based on known feature characteristics.

The following sections explain the business objectives, the assessment of data, and the selection, implementation and deployment of a model to provide a predictive guide to new wine quality.

## The CRISP-DM Methodology / Reference Model

In the mid to late 1990s, business markets were showing a sharp upturn in interest into the possibilities offered by data mining practices. The need for a standard process model, widely and freely available, became quickly apparent.

By 1999/2000, a process model named CRISP-DM (Cross-Industry Standard Process for Data Mining) had been produced by leading thinkers in the industry. It was based on practical, real-world experiences and sought input across a range of business domains.

As explained in the following sections of this document, we have taken the key principles of CRISP-DM to implement our CA project.

### Methodology

The CRISP-DM methodology is described as a hierarchical process mode with four levels that transition from the generic to the specific;

1. **Phases** – process blocks consisting of several generic tasks.
2. **Generic Tasks** – so called because they are intended to be robust and stable tasks that can apply in any data mining situation.
3. **Specialised Tasks** – a description as to how the generic tasks should be applied in specific situations. Very often these tasks can be performed in multiple orders and repeated a number of times.
4. **Process Instances** – this is a record of the actions, decisions, and results of an actual data mining engagement.

### Reference Model

The life cycle of a data mining project consists of six phases, as shown in this image below.



Figure 1

The sequence of the phases is not rigid. In our Wine Quality project moving back and forth between different phases was frequently required, as expected.

The outcome of each phase determines which phase, or particular task of a phase, has to be performed next. The arrows indicate the most important and frequent dependencies between phases.

As an example, in our Wine Quality CA we needed to…<provide some actual examples – when we have them…!>

The diagram above displays the following phases.

* Business Understanding
* Data Understanding
* Data Preparation
* Modelling
* Evaluation
* Deployment

The next sections of this document elaborate on these phases a little further and the remainder of the report describes the actual implementation of the CRISP-DM against our Wine Quality project.

### Business Understanding

Understand the project objectives and the requirements from a business perspective.

Convert this knowledge into an actual data mining problem definition, along with a project plan that will provide a framework to deliver the business objectives.

### Data Understanding

Start with initial data collection.

Proceed into activities that provide familiarity with the data, including data quality issues, data insight, and possible sub-sets within the data.

### Data Preparation

Activities to construct the final dataset that will be fed into the modelling too.

Data preparation tasks can be performed in multiple orders and over many interactions.

### Modelling

Select and apply various modelling techniques.

Calibrate parameters to provide optimal values within the model.

Revert to data preparation phase, if necessary.

### Evaluation

A high quality data analysis model has been built.

Assess that the model achieves the business objective.

### Deployment

The knowledge gained by the creation of the model will need to be organised and presented in a way that it can be used by the customer.

It is important for the customer to know up front what actions need to be carried out in order to actually make use of the created models.

## Development Environments

The majority of development took place within the RapidMiner toolkit and screenshots are provided to show the step-by-step working.



Figure 2

Supplementaty data investigation was conducted using a Python project developed in Visual Studio 2019. The structure of the project was modularised around a framework similar to the CRISP-DM model.

***<Python project screen shot>***

The Python project was used to provide quick to develop validations for the RapidMiner process and the ‘Performance’ outputs provided by that tool.

# Business Understanding

## Determine Business Objectives

The first objective of the data analyst is to thoroughly understand, from a business perspective, what the customer really wants to accomplish.

In our QA, the ‘customer’ and ‘business’ are obviously a theoretical concept. However, given that our chosen dataset relates to Wine Quality, we are assuming the role in the project of a chain of Off-Licence shops, who have a particularly speciality in selling Portuguese “Vinho Verde” red wine. Part of the business USP (unique selling point) is that staff is encouraged to be knowledgeable about the quality of this wine that they may recommend to customers. Although our imaginary Off-Licence chain promotes an awareness of wine amongst staff, very few employees would aspire to the level of sommelier in this brand of Portuguese wine and therefore it is necessary to provide guidance to staff when new stocks of wine arrive in-store.

Vineyards and Wine wholesalers will presumably provide recommendations on what wines are ‘good’ but a secondary objective of our business is to have a less subjective measure of quality for new wines. Our predictive model will therefore provide a more scientific basis for a quality rating, which can be applied across the entire outlet of shops, rather than relying on a human analysis, which could be open to interpretation.

A further secondary objective is that this model may provide guidance for similar in-store marketing of other ‘niche’ wine brands, should our business wish to replicate this approach to wine promotion.

How do we define success? A model is built based to predict the quality of new stocks of red wine based on the constituent chemical properties of the liquid. Our initial dataset will contain information to train and test our model (after various algorithm selections), and we will then conduct an additional test with new ‘unseen’ data to show that the model works well to provide an employee guide to wine quality.

## Assess Situation

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors to be considered in determining the data analysis goals and project plan.

**Dataset Inventory**

We chose a dataset provided in the Kaggle website ([*https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009*](https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)) that relates to the red wine variant of the Portuguese “Vinho Verde” red wine. (The primary source of the dataset is on the UCI Machine Learning Repository - [*https://archive.ics.uci.edu/ml/datasets/wine+quality*](https://archive.ics.uci.edu/ml/datasets/wine+quality)).

The dataset is publically available and provides a series of input variables, which are based on physiochemical tests, and an output variable that is a 0 – 10 score of quality.

**Assumptions**

The ‘quality’ measure, which is obviously the key characteristic we want to assess, has been assumed to come from feedback from wine industry specialists.

**Constraints**

This an assessment of quality based on the chemical constituents of the wine. There is no data relating to year or grape type, as might be expected with an assessment of wine, so we are assuming that a chemical analysis will provide the all the data points we need for a quantifiable assessment of quality.

There is also no indication of brand or price. This dataset is deliberately excluding these factors (or is unable to include them). Therefore that type of marketing data points will not influence the prediction of ‘quality’ as produced by our model.

## Determine Data Mining Goals

A data mining goal states a project objective in technical terms.

**Goal**

Our CA project aim is to build a predictive model that provides a 0 – 10 rating for a wine based a list of 11 chemical attributes in the liquid.

**Success Criteria**

We want our model to operate with a greater than 85% accuracy in its predictions of wine quality for new “Vinho Verde” red wines.

We may extend the modelling process so that a score of ‘7’ or greater is described as ‘Very Good’, ‘4 – 6’ receives a ‘Good’ description, and anything else is ‘Poor’. Thus we refine our classification of the model outputs into simpler terms for the end user employees.

## Produce Project Plan

**Project Plan**

The framework of this document, even just reading from the Table of Contents, provides the general outline of activity.

In brief, our timelines are to complete the following activity by the following milestones (allowing for some iterations and back and forth before project completion);

Any project plan is a dynamic document, and this CA is no exception and we expected, and encountered, the need for many revisions.

* **Saturday January 25th**: Complete dataset selection and establish business objectives.
* **Saturday February 1st** : Complete Data Understanding, Data Preparation, and preliminary model assessment.
* **Saturday February 8th** : Complete Modelling and Evaluation, determine Production approach. Present to class.
* **Sunday February 9th** : Submit CA final report with recommendations.

**Assessment of Tools**

In order to gain an insight into the use of commonly used industry tool, the majority of the data mining approach was conducted in ***RapidMiner*** (as can be seen in the screenshots used throughout this document).

However, early stage data analysis and some preparation used ***Python*** scripting. This was partially because of familiarity with Python from earlier CA work on the course and also to provide some quick additional verification of the RapidMiner outputs.

Excel was used for part of the validation process on predictions made by the model on new data.

# Data Understanding

## Collect Initial Data

This involves the acquisition of data and loading into our chosen data mining tool kits.

**Initial Data Collection Report**

As described in Section 2.2 of this document the dataset for the CA is taken from the Kaggle website, specifically from the URL : <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>, which in turn references the original UCI Machine Learning source.



Figure n

In Appendix A of this document there is an brief guide to understanding wine types and composition, with particular relevance to the checmical data points in this dataset. (Source : *https://github.com/dipanjanS*).

Downloading the CSV file from Kaggle is a straightforward exercise and the CSV file itself is just 101 kB.

For preliminary data analysis the CVS file on Red Wine quality loads without issue into RapidMiner.



Figure n

## Describe Data

This involves an examination of the ‘gross’ (or ‘surface’) data and a report on the results.

**Data Description Report**

The file is on a CSV format, and contains 1600 row with 12 attribute columns in the following structure:

***Input variables*** (based on physicochemical tests):  
1 - fixed acidity  
2 - volatile acidity  
3 - citric acid  
4 - residual sugar  
5 - chlorides  
6 - free sulfur dioxide  
7 - total sulfur dioxide  
8 - density  
9 - pH  
10 - sulphates  
11 – alcohol

***Output*** variable (based on sensory data):  
12 - quality (score between 0 and 10)

A surface view in NotePad++ shows the following sample structure;



Figure n

In line with the description of the dataset in Kaggle, the dataset contains header information but the remainder of the dataset is purely numeric.

Given that this is a dataset aimed at relative new comers to the work of Machine Learning, it does not seem likely that there will be any invalid or missing data entries. We would expect this to be borne out in the analysis in the following sections of this document.

## Explore Data

This task addresses data mining questions using querying, visualization, and reporting techniques.

**Explore ‘Wine Quality’ Data**

Target attribute..

Relationships between data – correlations..

Simple statistical analysis..

RapidMiner and Python screenshots..

**Data Exploration Report**

Initial findings / hypotheses and their impact on the remainder of the project.

Distribution of ‘quality’ – show RapidMiner screen shot of bar chart (distribution of data). Need to balance data.

Standardise and normalise the feature attributes.

## Verify Data Quality

Is the data complete? Does it contain errors and/or missing data? If so, how common are these issues?

**Data Quality Report**

Data quality is likely to be good based on the source.

Indicate that there are no missing entries

No obvious errors.

Check for outliers..

# Data Preparation

## Select Data

The..

## Clean Data

Section..

## Construct Data

Section..

## Integrate Data

Section..

## Format Data

Section..

# Modelling

## Select Modelling Technique

For

## 

## Generate Test Design

The..

## Build Model

The..

## Assess Model

The..

# Evaluation

## Evaluate Results

We..

## Review Process

The..

## Determine Next Steps

The..

# Deployment

## Plan Deployment

Our analysis of..

.

## Plan Monitoring and Maintenance

A ...

## Produce Final Report

A ...

## Review Project

A ...

# Conclusion

## Conclusion..

The..

# Appendices and References

## Appendix A Understanding Wine and Types

Wine is an alcoholic beverage made from grapes which is fermented without the addition of sugars, acids, enzymes, water, or other nutrients

Red wine is made from dark red and black grapes. The color usually ranges from various shades of red, brown and violet. This is produced with whole grapes including the skin which adds to the color and flavor of red wines, giving it a rich flavor.

White wine is made from white grapes with no skins or seeds. The color is usually straw-yellow, yellow-green, or yellow-gold. Most white wines have a light and fruity flavor as compared to richer red wines.

**Understanding Wine Attributes and Properties**

* **fixed acidity:** Acids are one of the fundamental properties of wine and contribute greatly to the taste of the wine. Reducing acids significantly might lead to wines tasting flat. Fixed acids include tartaric, malic, citric, and succinic acids which are found in grapes (except succinic).
* **volatile acidity:** These acids are to be distilled out from the wine before completing the production process. It is primarily constituted of acetic acid though other acids like lactic, formic and butyric acids might also be present. Excess of volatile acids are undesirable and lead to unpleasant flavor. In the US, the legal limits of volatile acidity are 1.2 g/L for red table wine and 1.1 g/L for white table wine.
* **citric acid:** This is one of the fixed acids which gives a wine its freshness. Usually most of it is consumed during the fermentation process and sometimes it is added separately to give the wine more freshness.
* **residual sugar:** This typically refers to the natural sugar from grapes which remains after the fermentation process stops, or is stopped.
* **chlorides:** This is usually a major contributor to saltiness in wine.
* **free sulfur dioxide:** This is the part of the sulphur dioxide that when added to a wine is said to be free after the remaining part binds. Winemakers will always try to get the highest proportion of free sulphur to bind. They are also known as sulfites and too much of it is undesirable and gives a pungent odour.
* **total sulfur dioxide:** This is the sum total of the bound and the free sulfur dioxide. This is mainly added to kill harmful bacteria and preserve quality and freshness. There are usually legal limits for sulfur levels in wines and excess of it can even kill good yeast and give out undesirable odour.
* **density:** This can be represented as a comparison of the weight of a specific volume of wine to an equivalent volume of water. It is generally used as a measure of the conversion of sugar to alcohol.
* **pH:** Also known as the potential of hydrogen, this is a numeric scale to specify the acidity or basicity the wine. Fixed acidity contributes the most towards the pH of wines. You might know, solutions with a pH less than 7 are acidic, while solutions with a pH greater than 7 are basic. With a pH of 7, pure water is neutral. Most wines have a pH between 2.9 and 3.9 and are therefore acidic.
* **sulphates:** These are mineral salts containing sulfur. Sulphates are to wine as gluten is to food. They are a regular part of the winemaking around the world and are considered essential. They are connected to the fermentation process and affects the wine aroma and flavor.
* **alcohol:** Wine is an alcoholic beverage. Alcohol is formed as a result of yeast converting sugar during the fermentation process. The percentage of alcohol can vary from wine to wine. Hence it is not a surprise for this attribute to be a part of this dataset. It's usually measured in % vol or alcohol by volume (ABV).
* **quality:** Wine experts graded the wine quality between 0 (very bad) and 10 (very excellent). The eventual quality score is the median of at least three evaluations made by the same wine experts.
* **wine\_type:** Since we originally had two datasets for red and white wine, we introduced this attribute in the final merged dataset which indicates the type of wine for each data point. A wine can either be a 'red' or a 'white' wine. One of the predictive models we will build in this chapter would be such that we can predict the type of wine by looking at other wine attributes.
* **quality\_label:** This is a derived attribute from the quality attribute. We bucket or group wine quality scores into three qualitative buckets namely low, medium and high. Wines with a quality score of 3, 4 & 5 are low quality, scores of 6 & 7 are medium quality and scores of 8 & 9 are high quality wines. We will also build another model in this chapter to predict this wine quality label based on other wine attributes.

## References

The use of this Wine Quality dataset in this CA acknowledges the source publication :

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.