Business Understanding Report

Ammar Hasan 150454388

20 November 2018

Contents

1 Introduction

This part of the report documents the first stage of the CRISP DM cycle, Business Understanding. In this stage the project the objectives and requirements from a business perspective are considered to form the overall data mining problem definition (specification)[1].

2 Organisation and Organisational Objectives

This data mining project needs a business rationale to be justified and to be planned. This means that the project needs to have an overall beneficial impact to the organisation problems in a way that is clearly understood.

2.1 Background

The main objective of this project is to develop a pipeline of data analytics to improve an online learning program about Cyber Security in Future Learn by Newcastle University. The data mining solution will look into data collected from users and staff to help optimise the learning solution (improve resources, user engagement, reduce dropouts, etc.).

2.2 Resources and Data

The datasets provided are collected from 7 separate runs collected at different time intervals. Each run consists a given set of data-sets consisting of related data (e.g. enrollments, video stats, etc.) from user and staff which has been anonymised by the removal of names.

2.3 Success Critera and Organisational Objectives

According to meeting with the client the following points make up the core objectives of Newcastle University for its online courses:

- Engagement
- Education
- Showcase technology
- Outreach
- Broaden type of learners

Therefore, to succeed the data mining project will need to tested against the following criterion:

- Provide insights into user engagement
- Provide ways to improve educational experience
- Help in outreaching to new communities
- Help in understanding on how to adapt the system for different learners

[1] Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. "CRISP-DM 1.0 Step-by-step data mining guide." (2000).

3 Project Specification and Plan

3.1 Requirements and Rationale

- Clean data from anomalies, redundant data and data that is not useful for the project
  - Reasoning: Ensure better join performance and remove/fix any data that can confuse the analysis
- Aggregate factors used to evaluate leaner performance, progression and learner types
  - Reasoning: Data needs to be aggregated in a form that can be investigated
- Merge aggregated data
  - Reasoning: Data needs to be joined to be have bivariate relationships explored and analysed
- Explore aggregated data using Exploratory Data Analysis and Unsupervised Learning
  - Reasoning: Exploratory Data Analysis and Unsupervised Learning give clues about data structure and relationships

3.2 Tools

All analystics will be performed using R functions from various statistical libraries and any plots based on them will be produced using ggplot library functions in R. The reports and documentation are produced using RMarkdown .rmd files that are kintted to form PDFs. All files used and produced during this project will be placed in a hierarchy generated by ProjectTemplate, which is also where the git repository used for version control lies. Tests will be conducted using TestHat.

3.3 Project Plan and Risks

The project will follow the three initial stages of the CRISP DM hierarchical cyclic process model for data mining (Business Understanding, Data Understanding and Data Preparation). The later stages are omitted from this project because the development of a model is not part of the requirements. Typically, the the CRISP DM stages are done as part of an iterative cycle in previously stated order, however this order is not strict. In this project for instance, some Data Preparation stages may occur before Data Description.

As part of the project plan, the results of this project at completed cycles will be presented to the client for feedback. This is done to avoid any risks of client requirements drifting from project specification and delivery. Moreover, literate programming is used (RMarkdown) alongside Test Driven Development where possible to reduce the risk that documentation and tests are cut or rushed when time constraints occur.