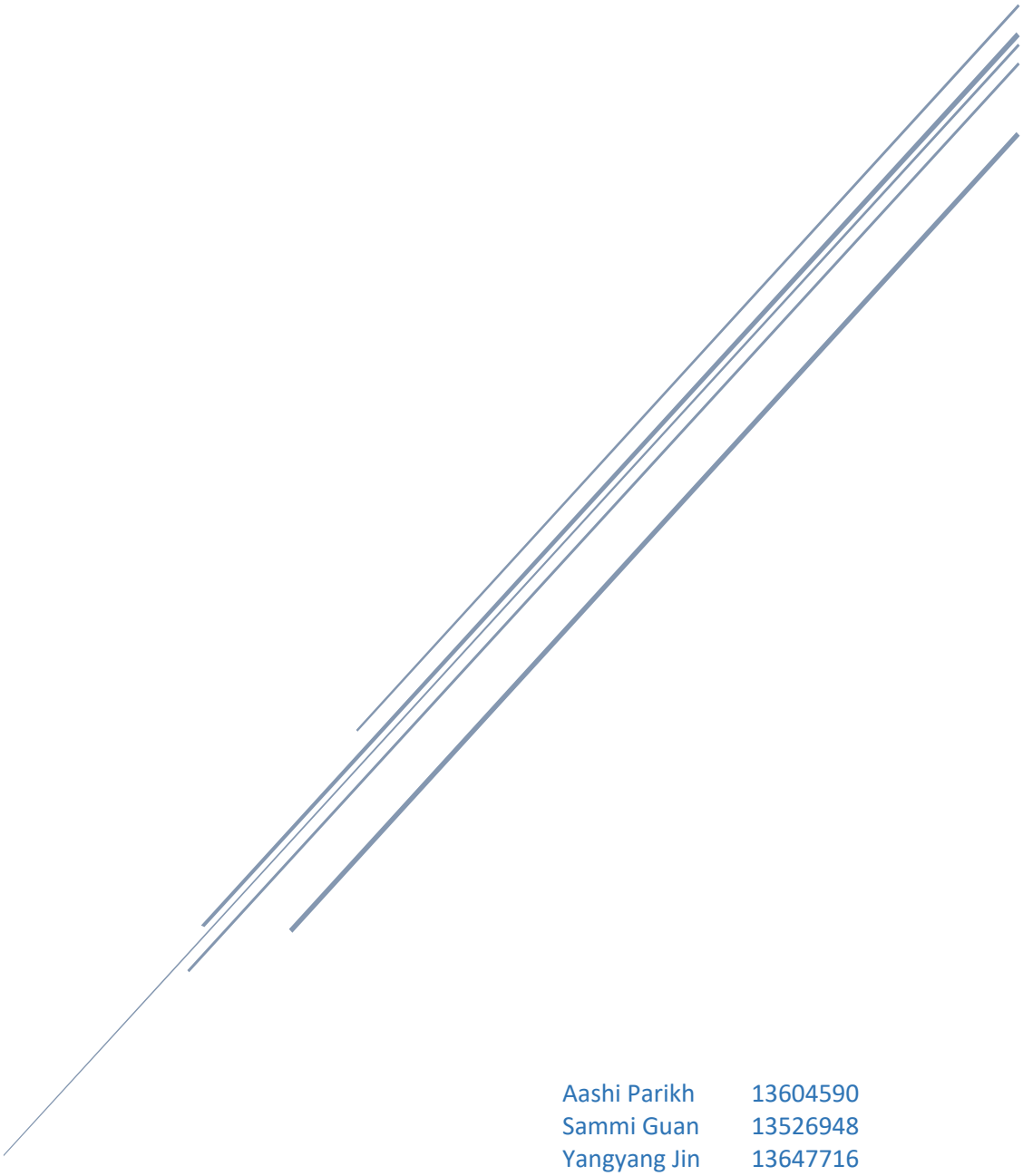
CHURN ANALYSIS IN SUBSCRIPTION-BASED BUSINESSES FROM PREDICTION TO INFERENCE

Plan and Proposal



Aashi Parikh 13604590
Sammi Guan 13526948
Yangyang Jin 13647716
Weilin Sun 13462383

Table of Contents

1.0 About Our Company

Frontop Analytic is a startup company made up of four shareholders that are enthusiastic to provide analytics solutions for a wide range of industry sectors primarily focusing on the superannuation sectors in Australia. We provide churn analysis for subscription-based businesses to evaluate our client's customer loss rate aiming to narrow it down.

1.1 Mission Statement

Frontop means innovation, but also means serious and responsible for customers. We aim to produce findings and easily adaptable analyses that allow our clients to have greater visibility and transparency to make a prompt decision. The service we provide is under a structured framework to ensure we maximise the value of data and focus on our client's requirements and objectives.

2.0 Meet the Team

2.1 Team Leader – Yangyang Jin

Yangyang is a second-year IT student majoring in data analytics at the University of Technology Sydney. He is passionate about statistics and data analysis and has a strong interest in programming Yangyang's long-term objective is to become an experienced data analyst in the technology field, and he acts as a lead manager on this project.

Team diversities in projects are common; it increases productivity, allowing for more ideas and a broader range of skills among team members. As a lead manager, Yanggang is responsible for discovering each team member's strengths and assigning them tasks that they can perform the best in completing the project. Moreover, Yangyang is also assisting in data analysis work and data visualisation to present the information for our client.

2.2 Lead Analyst – Aashi Parikh

Aashi is a final year IT student majoring in data analytics at the University of Technology Sydney, with a strong interest in marketing, Aashi has chosen his sub-major into the marketing sector. He is keen to study market conditions to access all the potential data sources for further analysis to provide quality data to increase their competitive advantage. With a passion for marketing, Aashi long-term goal is to become a market analyst in the superannuation field, and he acts as a lead analyst on this project.

As a lead analyst, Aashi conducts market research and analysis into our client's customer data to provide better insight for our clients.

2.3 Project Analyst – Xianwen (Sammi) Guan

Sammi is a final year student studying a double major in data analytics and finance at the University of Technology Sydney. With a keen interest in data analytics and finance, she recently started an internship as a business analyst in the bank's market data field. Sammi's

long-term objective is to become a data analyst in the financial service field, and she acts as a project analyst on the project.

As a project analyst, Sammi is working closely with the team to collect the research and data required, moreover, schedule project timeline to ensure the team is on track in each step of the project.

2.4 Management Analyst – Weilin Sun

Weilin is a second-year IT student majoring in data analytics at the University of Technology Sydney. He is keen to gain commercial experience in different fields relating to data analysis, with a passion for improving efficiency and providing possible solutions for our clients. He acts as a management analyst on this project. Weilin's long-term goal is to start a business with a group of people who share a similar objective.

As a management analyst, Weilin researches our client's problems and work closely with the team to gather relevant information to discover possible solutions to help our clients.

3.0 Project description

In Australia, more than 1.1 million Australians have self-managed super funds (Drury, 2021). Retirement benefits start when an employee starts working and the employer starts paying a percentage of the employee's salary into the employee's retirement account. The superannuation fund(s) will invest or manage the money for the employee until retirement (Australian Government, 2021). For most people, a super pension fund is a long-term investment.

3.1 Business Problem

Our client is one of Australia's superannuation funds. According to reports, the superannuation fund mentioned that the cost of acquiring customers and the rate of customer churn have both increased rapidly in recent years. The high churn rate of customers also led to a decline in the company's profitability. This is extremely detrimental to the operation of the company. In addition, the superannuation fund stated that they also have minimal knowledge of membership data. If the reasons and problems of customer churn are not adequately addressed, the customer churn rate will increase proportionally.

3.2 Project objectives

The objective of this project is to predict and analyse the results and causes of business churn by building a model for business churn. To begin, we will use a predictive model to score each member's churn rate in the next financial year, and then divide the churners into high-profit and low-profit members. After that, the company will offer the best incentives for high-profit members, such as discounts on services. For low-profit members, the company will encourage them to increase their Quality of Services (QoS).

Moreover, we will propose a framework using a deep feed-forward neural network for classification accompanied by a sequential pattern mining method. We will also propose causal Bayesian networks to identify cause probabilities that lead to member churn.

4.0 Data mining problem

Data mining is the process of extracting hidden, unknown but potentially useful information from a large amount of data. Data mining aims to build a decision model that predicts future behavior based on past action data (Stedman, 2021). Experiments were conducted on available datasets for members holding accounts. The datasets included customer account, demographics, customer engagement, and financial data.

We define a customer as a churner if they closed their account during the subsequent 6-month time window. Therefore, we use binary outcome for each customer [0 or 1], where 1-means the account closed and 0 that it was not close in the subsequent 6-month time window. We applied two main inclusion criteria for account tenure and balance to satisfy model development. First, we only retained data for customers with more than 6-month account tenure; and second, we removed account balances below $1500 to improve prediction since predicting churn probability for inactive accounts has low value for superannuation funds.

4.1 Data understanding

In the process of data mining, it is necessary to have a detailed understanding of the data. Since there are some complex or difficult-to-understand data in the dataset, we not only need to understand the background knowledge of the project, but also understand the meaning of different attributes in the dataset. If we do not understand the content of the data, the data will have no clear meaning and will not be useful information.

4.2 Performance Issues in Data Mining

There are many factors that affect database performance. The most common of these are missing or noisy values. These factors can significantly affect the accuracy of our established data models, resulting in erroneous results. Therefore, to ensure the performance of the model in data mining, we must perform corresponding data preprocessing, which can obtain more accurate information in the dataset.

4.3 Methods of Data Mining

Recently, most of the related works in churn prediction have achieved well-practiced prediction accuracy in using deep learning methods. We proposed deep learning (DL) algorithm to deal with massive datasets, since deep learning (feature transformation) differentially weights features using historical data. The Feed-forward NNs use the output from one layer as input for the next layer, with no loops between layers. Therefore, in this business case study, the deep learning feed-forward (DFF) algorithm is employed for experiment phase to trading-off between reduced computational load and highest accuracy for different comparison algorithms. General advantages for DFF NNs can be summarized as:

- Superior accuracy from DL;
- Includes more variables than classical ML;
- DL algorithms can extract patterns while avoiding blind spots from extensive member demographics, behavioral variables, and billions of member engagement logs.

5.0 Project Proposal

We anticipate that the primary data for our analysis will be statistics on superannuation funds invested by Australian residents. The database contains records for around 270,000 members with 88 data attributes, including members' personal information, whether contact information is provided, the rate of change in the amount of the account, the percentage of the amount invested, the frequency of contributions. This is an extensive and relatively complete database and it will be used as the primary dataset for this project. We will search for other relevant data sets throughout the project and integrate some of the available parts with our data set.

We expect to use NumPy, Pandas tools in Python for data analysis, and Matplotlib and Seaborn tools are expected to be used for data visualization. These tools allow us to filter, integrate, and analyse this huge data set.

Due to the increase of customer attrition rate in recent years, resulting in a decline of the company's profitability, our goal is to build a model to predict business attrition rates and analyse their causes, this problem is classified as supervised learning. After preliminary preparation, we decided to integrate billing, demographics, service records and member engagement information for further analysis. Another analysis used a predictive model to predict attrition rates for each member, using the Churners tool to divide these high-risk members into high and low profit groups. As a result, high-profitability members will be better rewarded, while low-profitability members will receive supportive services.

These analyses will enable the company to understand the causes of churners better and resolve the problem, allowing more focus on customer needs in future work.

6.0 Project Plan

6.1 Churn Propensity Model

In this case, the Churn Propensity Model will be used for the project which is mainly on predicting individual consumers' proclivity or possibility to leave (Votava, 2021). It indicates the likelihood that we will lose a customer in the future for each customer at any given period.

However, the Churn Propensity Model is based on a Cross-industry standard process for data mining which is known as CRISP-DM. The methodology was founded in 1996 and it became one of the most widely used analysis tools within the industry nowadays (Rodrigues, 2020).

And it is also the process we need to comply with for this project. According to the flow diagram of the CRISP-DM, it contains six stages which are:

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

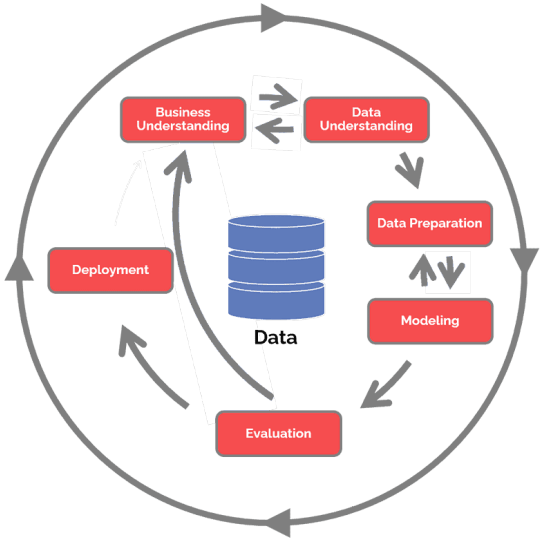


Figure 1: CRISP-DM Process Diagram

6.1.1 Business Understanding

The primary stage is business understanding which will be mainly aimed at comprehending the project's objectives and requirements. In this case, the business objectives will be defined by a deep understanding of the true needs of the customer from a business point of view. Then determine the availability of resources, project demands, risks and contingencies, and analyse the cost and benefit which can be summarised as evaluate the situation. After that, the initial project plan will be drafted which will select the suitable methods and tools, as well as create precise plans for each project stage.

6.1.2 Data Understanding

The following phase is data understanding which drives the emphasis to find, acquire, and evaluate data sets that can help to achieve the project goals, adding to the basis of business understanding. In this stage, obtaining source data, describing data, exploring data and defining the quality of data will be the four main parts. According to the project, three datasets will be used. In this case, to understand and describe each attribute and data of the datasets will be the primary, then the relationship between each attribute will be displayed by

several methods such as data visualization. The next step is to determine if the quality of the data is eligible. If not, then the dataset will be reexplored until the data quality meets standards.

6.1.3 Data Preparation

During this phase, it will be decided which dataset will be utilised in the project. Then, the extra data and unused or unrelated attributes to the project will be cleaned to obtain a more accurate result. Besides, some new features might be constructed for later Exploratory Data Analysis (EDA).

6.1.4 Modelling

In the modelling process, the initial will be chosen modelling techniques and tools. In this case, there are several methods are available to the project which are Gradient Boosting Tree (GBT), Random Forest (RF), Logistic Regression (LR) or Elastic Net (EN), Neural Network (NN), Support Vector Machine (SVM). The modelling approach we end up using will choose the one that has the highest accuracy due to the model prediction result. After modelling and programming, the output will be evaluated and the modelling techniques above with the highest accuracy will then be selected as the final solution.

6.1.5 Evaluation

In the evaluation phase, the result of modelling will be defined and discussed whether it meets the client's requirements and expectations while it compiles the criteria for commercial success. The effectiveness of the model will be measured by well-practised evaluation metrics such as Mean Error, Root Mean Squared Error, Mean Absolute Error, Area Under Curve (AUC), and confusion matrix.

6.1.6 Deployment

The last stage will be deployment, which can be divided into four main parts. Firstly, a plan for deploying the model will be created. Also, a detailed backup program of inspection and maintenance will be developed which can give a countermeasure to solve the problems during monitoring. Not the last but the most important step is finalising the project document which includes a summary of the whole project in detail with the final output. Finally, the implementation process of the project will be reviewed, and the merits and mistakes will be concluded.

6.2 Project timeline

The figure below will be the schedule of the plan which stated each task with the participants and its estimated completion time.

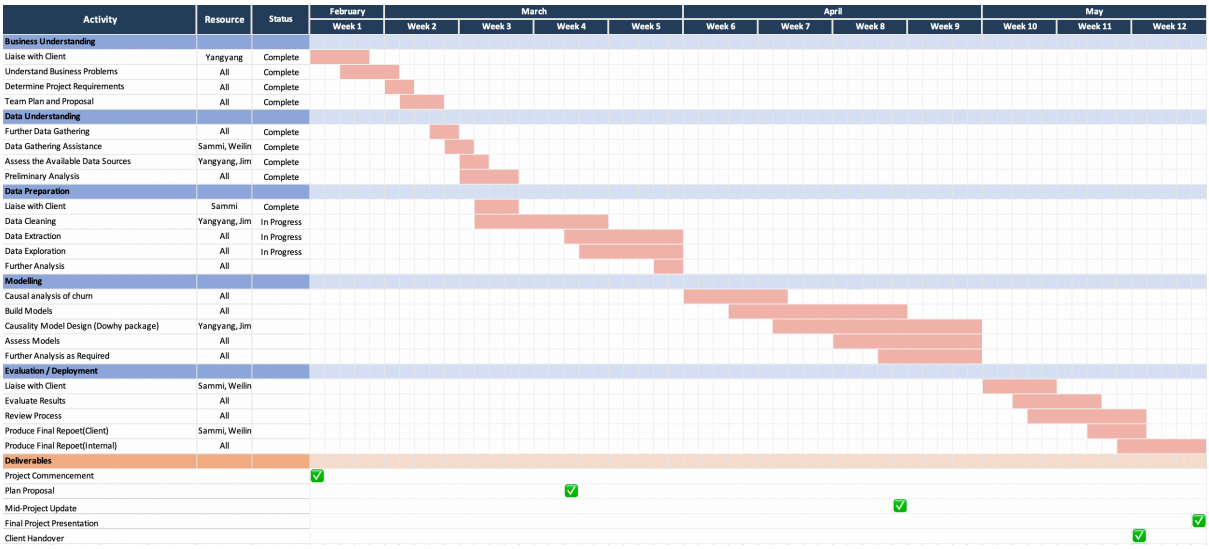


Figure 2: Gantt chart of project #20

6.3 Milestone

According to the Gantt chart above, four milestones will be setup which will be introduced below.

6.3.1 Project Understanding

14.02.2022 - 13.03.2022

In this section, five different tasks will be included which are mainly focused on determining the project plan, being familiar with the dataset, and separate tasks of the assignments for each group member. A draft of the project proposal will be the output of this stage.

6.3.2 Data Preparation

13.03.2022 – 17.03.2022

For the data preparation phase, data processing and cleaning will be contained as well as the problem statement and project significance. In this case, to communicate with the tutor in real-time about the progress of the assignment, two group meetings are arranged in the period. Moreover, the finalised assignment one which is the project proposal and initial data analysis will be completed in this section.

6.3.3 Modelling

24.03.2022 – 21.04.2022

The third milestone is modelling which is also the most important one of the projects which will focus on assignment 2. In this phase, the Churn Propensity Model with causal analysis of churn and causality model will be developed. Because of the complexity of this milestone, five group meetings will be scheduled during this section.

6.3.4 Evaluation

28.04.2022 - 19.05.2022

The last phase is the evaluation which will mainly on finalized the project. In this case, the project result and process will be reviewed. Besides, the final report will contain the tested model and a description of each phase will be summarised. In this case, four meetings will be expected to be interspersed in the final phase to achieve a closer collaboration.

Reference

AutrialianGoverment. (2021, October 26). Super. Australian Taxation Office. https://www.ato.gov.au/individuals/super/#Whatissuper

Drury. (2021, November 15). SMSF statistics: 1.1 million members with $822bn in super. SuperGuide. https://www.superguide.com.au/smsfs/smsf-statistics

Rodrigues, I. (2020, February 17). *CRISP-DM methodology leader in data mining and big data*. TowardsDataScience. https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining and-big-data-467efd3d378

Stedman, C. (2021, September 7). What is data mining? SearchBusinessAnalytics. https://www.techtarget.com/searchbusinessanalytics/definition/data-mining

Votava, A. (2021, September 7). *Churn prediction model*. TowardsDataScience. https://towardsdatascience.com/churn-prediction-model-8a3f669cc760