**Personal Information**

Name: Duc Tran

Student ID: s3807447

Email: [S3807447@student.rmit.edu.au](mailto:S3807447@student.rmit.edu.au)

Background

My name is Duc and I am currently doing 3 subjects which are User-centred Design, Introduction to Programming and Introduction to Information Technology. A bit about myself, outside of I.T I am currently working in a bakery and love to exercise, watch anime and collect rocks. I hope to one day work somewhere in the data analytics field in I.T but am still open to other options.

Interest in IT

My interest in I.T is in the data analytics area as I love to research and make sense of things. I’m by nature a very analytical person and as a result like to solve problems by looking at the logic behind it.

My interest in I.T started when I was very young through my oldest brother who at that time was working for IBM in Sydney. I was always interested in what he was doing and when there came a time where he took computers apart I was always interested in the hardware and software aspects of it. As the years progressed I started to delve into more programming firstly starting off making a website in year 10 using Microsoft Front Page, moving on to java coding using C++, using Microsoft SQL to make databases and then to Python and Eclipse.

The reason why I chose RMIT as my preferred university to study I.T in is due to the course being offered by Open University Australia (OUA) and that this style of study compliments my busy life style.

During my bachelor degree I expect to learn what I need to know about I.T to make it into a career. I expect to further improve my understanding and knowledge on the programs I’m currently familiar with and to improve my usage of them even further.

**Ideal Job**



Figure 1: An example of my ideal job (seek 2019).

In figure 1, this is an example of my ideal job in the I.T industry which is to become a data analyst. A data analyst is responsible for analysing data and translating that data in a way that would benefit and improve the company’s understanding of that data (snag 2019). The data analyst may show the data in a number of ways including through presentations, statistical reports, graphs or other similar mediums.

To become a data analyst a bachelor’s degree is generally needed for most entry level jobs and for the more upper-levelled jobs a master’s degree is generally needed (snag 2019). Apart from obtaining a bachelor’s or master’s degree, the applicant must also have strong mathematical and analytical skills as this job requires a logical and analytical mindset.

At this current moment I believe I don’t quite have enough experience and qualification to become a data analyst. However, I do have some basic understanding of programming and creating databases and along with my analytical mind set, I believe with further progression in this I.T course I can be ready to go into being a data analyst upon graduation.

To obtain the skills, qualifications and experience necessary to begin working as a data analyst I plan to apply for a help desk position at an I.T or internet company just as my brothers have before. This way I can gain hands on experience of working in a similar environment that I may work in the future and to also boost my qualifications in the process. Academically wise, I just plan on completing this I.T course so that I can have the academic experience necessary to work this position.

**Personal Profile**

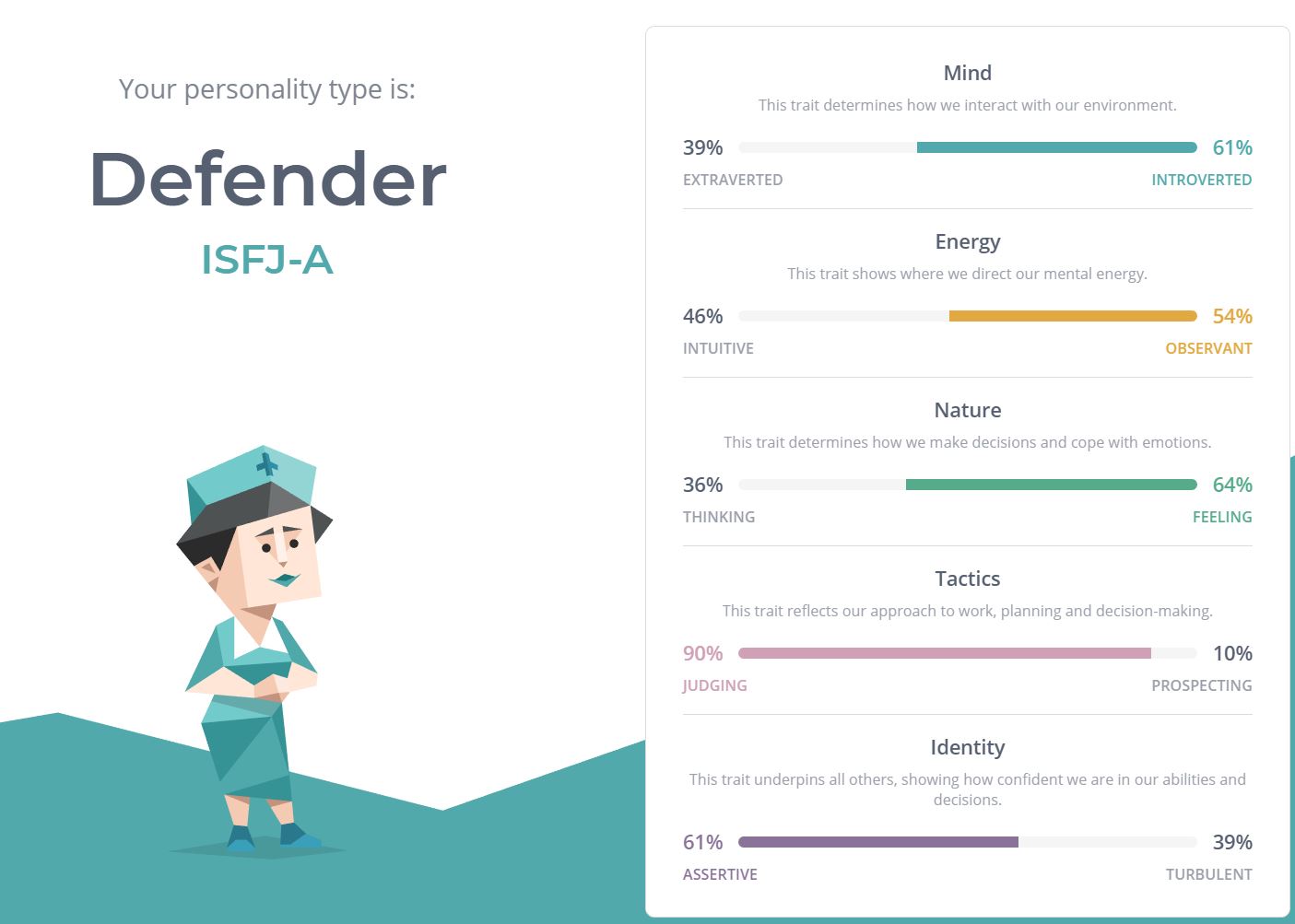


Figure 2: The results of my Myer-Briggs test (16Personalities 2019).

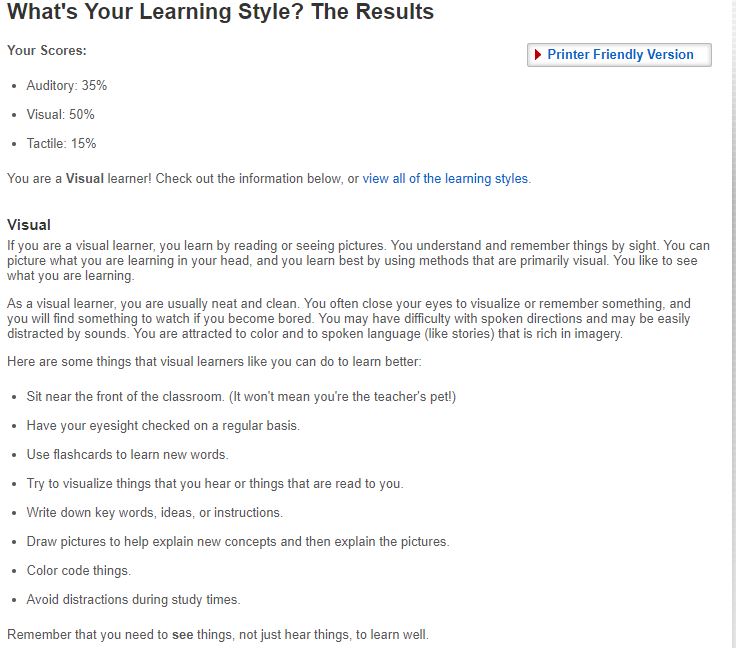


Figure 3: Here is my result from a learning style test (Education Planner.org 2019)

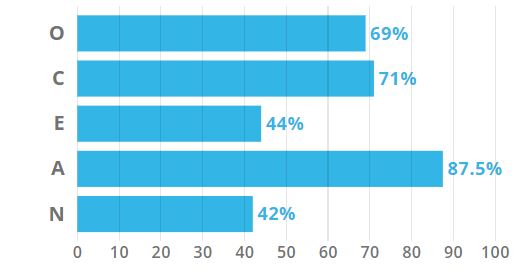


Figure 4: Here is my result from doing a big five personality test (Truity 2019).

In figures 2 to 4, the results of these test indicate who I am as a person and how I would interact with other people if I was put into a group situation. For the most part these three tests are pretty accurate to who I am and what my preferred style of learning is.

The results of these tests will show how I interact in a team, which in this case I prefer not to lead the team but rather to work together and follow directions (putting my input in as well to make sure the leader makes an informed decision on what direction our group takes).

When taking these results into account when forming a team, these results tells us what the personality of each member is like and so would help us form a team that is more in sync with each other. For example, if all members have strong desire to do well in their given tasks then overall the team has a higher probability of obtaining higher marks for the project.

**Project Idea**

**Using the gradient boosting technique to improve the reliability of mathematical models.**

**Overview**

Businesses rely on statistical data in order to gain an insight into how their market or customers or even their own employees are doing and what trends should the business be looking at in order to improve. Gradient boosting is a machine learning technique that relies on the intuition of the next best model in conjunction with previous models that as a result reduces prediction errors and improves the correlation reliability in mathematical models (Singh 2018). This report will explain how gradient boosting can be used to improve the reliability of statistical data that are used in businesses and how this can benefit businesses.

**Motivation**

The usage of the gradient boosting technique to improve statistical reliability would help to further strengthen statistical data and allow the business to make better decisions on how they can go forward. An example show casing the strength of the gradient boosting technique is shown by a study that was conducted by Cortez et.al in the University of Minho. This study involved creating two datasets with one set allocated red Vinho Verde Wine samples and the other set allocated white Vinho Verde Wine samples. The goal of this study was to judge the wine quality based on various physiochemical tests (Cortez et al 2009). Now using this study Hoare (n.d) decided to compare the gradient boosting technique and the Classification and Regression Trees (CART) technique to see which of the two returned a more accurate prediction. The result was that using the gradient boosting technique the accuracy of the correlation of the study was up at 64.74% whereas using the CART method the predictability was only at 50% (Hoare n.d). Having a higher predictability percentage means that when doing statistical work in the future this technique would provide a more accurate statistical report with which the business can work with and therefore make better decisions when looking how to move forward.

**Description**

Gradient boosting is useful in areas where data is often unbalanced and that the environment is a supervised learning type such as studying DNA sequences to cyber security (as that environment is constantly changing) (Ravvanshad 2018). This is due to the fact that gradient boosting can be used to detect anomalies in information with high precision. The gradient boosting technique can also be used to construct ranking models that are used in systems designed to retrieve information (Ravvanshad 2018). In a nutshell gradient boosting has two main features that allows it to be a powerful technique on addressing the learning to rank problems on real life datasets and there are as follows: one that it performs its optimization functions within a functional space rather than a parameter type of space which allows for the easier use of custom loss functions (functions that are designed to help the machine to predict data in the correct direction and place heavy penalties on the wrong decision). Secondly, gradient boosting allows us to deal with unbalanced datasets through the strengthening of the positive class (predicting data in the correct direction) (Ravvanshad 2018). However, with these features we must not forget about the downsides to using the gradient boosting technique. The gradient boosting method tends to be too closely fit to data that is not necessary and training of the machine tends to take longer as each tree of data is sequentially built and read and therefore it takes the machine longer to learn to predict the dataset in a positive way (Ravvanshad 2018).

**Tools and Technologies**

In order to conduct the gradient boosting technique in terms of hardware we can just use a normal computer or a laptop that is capable of running the Windows, Linux or OS X (Apple) operating systems and of course have a reliable internet connection (JP 2019). Software wise, the app that uses the gradient boosting technique supports all of the major programming languages which include C++, Python, Java and many more (JP 2019).The portability of this technique means that the user can easily access it and implement it into their work when required.

**Skills Required**

Skill wise in order to use the gradient boosting technique the user must first have an understanding of mathematical models, correlation and generally statistical mathematics. The user must also have knowledge of programming in the programming languages that the app mentioned above supports. The software and hardware itself is very easy to come by as everyone doing this course should have access to a computer that runs one of the aforementioned operating systems. Accessing the above programming programs is also very simple as they can be downloaded freely on the internet. In terms of the mathematical aspect everyone would have learnt what a regression model is in school and what correlation and causation is.

**Outcome**

Upon the successful completion of this project we should be able to develop a more precise method of prediction using regression models. This will improve further on the original problem of have a lower rate of predictability and provide a stronger statistical interpretation of data. This development will allow businesses to better predict how the market and employees are going and what action to take into the future.

**Reference**

16Personalities 2019, *Free Personality Test*, viewed 14 June 2019, <<https://www.16personalities.com/isfj-personality>>.

Cortez et al., 2009, *Wine Quality Data Set*, viewed 16 June 2019, <<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>>.

Education planner.org 2019, *What’s Your Learning Style? 20 Questions*, viewed 14 June 2019<<http://www.educationplanner.org/students/self-assessments/learning-styles-quiz.shtml>>.

Hoare, J n.d., *Gradient Boosting Explained – The Coolest Kid on The Machine Learning Block*, viewed 16 June 2019, <<https://aibusiness.com/2019-trends-data-science-innovation-acceleration/>>.

JP, M 2019, *XGBoost Algorithm: Long May She Reign!*, viewed 16 June 2019, <<https://morioh.com/p/32c506939ad5/xgboost-algorithm-long-may-she-reign>>.

Ravvanshad, A 2018, *Gradient Boosting vs Random Forest*, viewed 16 June 2019, <<https://medium.com/@aravanshad/gradient-boosting-versus-random-forest-cfa3fa8f0d80>>.

Seek 2019, *Data Analyst*, viewed 12 June 2019, <<https://www.seek.com.au/job/39244648?searchrequesttoken=473e08fd-6665-40c7-b54a-e82cb248185e&type=standout>>.

Singh, H 2018, *Understanding Gradient Boosting Machines*, viewed 16 June 2019, <<https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab>>.

Snag. 2019, *Data Analyst Job Description*, viewed 12 June 2019, < <https://www.snagajob.com/job-descriptions/data-analyst/>>.

Truity 2019, *The Big Five Personality Test*, viewed 13 June 2019, <<https://www.truity.com/test/big-five-personality-test>>