**Linni Qin**

**Student ID: n9632981**

**Supervisor: Dr. Guido Zuccon**

**Unit Code: IFN701**

**Project Category: A Development Project**

**Queensland University of Technology**

**Forecasting zestimate error for zillow: a data analysis project Proposal**

**Table of Contents**

1. **Project Introduction1**

1.1 Background1

1.2 Scopes1

1.3 Approach Overview3

1.4 Outcomes3

1. **Project Method4**
2. **Project Management Approach4**

Project Environment5

Scrum5

1. **Communication Plan4**
2. **Constrain and Risk Management Plan4**

**Reference4**

**Appendices4**

Appendix A: Oral Presentation Note5

Appendix B: Supervisor Statement5

1. **Project Introduction  
   1.1 Background**

Zillow (zillow.com) is one of the popular information sites for real estate in the United States. It plays the role as realestate.com.au and domain.com.au in Australia to offer an online ecosystem for real estate professionals such as agents and mortgage bankers, home buyers, sellers and renters. There are 110 million properties across the nation have been served on Zillow’s platform (Reference). Zestimate is their price prediction model to estimate the market value of the properties and it is also an online tool to assist their website visitors making home-relevant decisions. According to the official data, the current median error rate of Zestimate is 5% (Reference). This percentage seems that there is solely a minor disparity between the forecasting value and the actual sales price. However, it means greatly for the most expensive property people purchases in their lives. For example, a home with actual value of $700,000, its 5% will be $35,000 that approximately equals to a whole year net income for a junior employee. The prediction value with this error rate might cause the potential home buyers to change their purchase decision or to be overconfident in their purchase capability. Furthermore, given the statistical concept, median error means there are a half of the properties’ estimation error rates exceed 5%. Consequently, the less visitor resulting from the worse home value estimate service will lead to the decrease of revenue of Zillow resulting from the less commercial advertiser on its site. Therefore, it is a critical issue for Zillow to improve the accuracy of Zestimate. This is also the reason why Zillow posted this issue a challenge with a prize up to $12 million on data science platform Kaggle.com to allow global data scientists to compete. Likewise, this project largely denotes the value of research relevant to value prediction models as the current Zestimate model is based on more than 7millions of machine learning algorithms and statics data points (Reference).

**1.2 Scopes**

Given the competition rules of Kaggle, the expectation of Zillow, the table *“MoSCow Prioritised Requirement List”* below will represent the significance of the requirements through the whole project with applying prioritisation tool MoSCoW. “Must”, “Should” and “Could” requirements would be determined in the scope and “Would” items should be out of the scope as the limited time factor for this project.

**MoSCow Prioritised Requirement List**

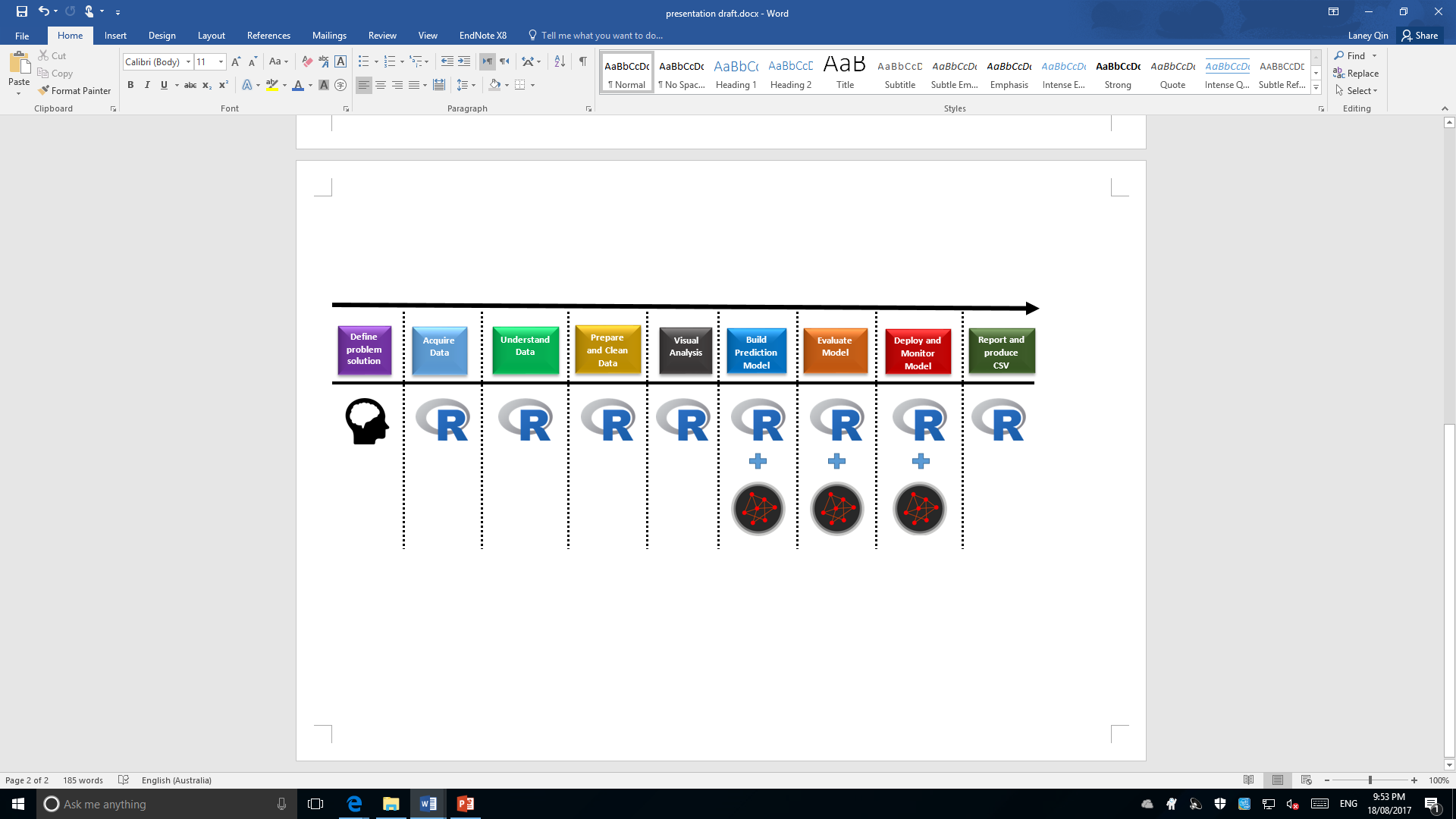
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***ID*** | ***Requirement List (Reference)*** | ***Priority*** | ***Points*** | ***Reason*** |
| 1 | As a contest sponsor, Zillow requires the participants to use their supplied data only. Any external data is prohibited. | Must | 2 | This is compulsory requirement. |
| 2 | As a competition organiser, Kaggle requires the participants to predict the error rate and store the value in a CSV file for 6 time points: October, November and December for both 2016 and 2017. R markdown file or Python file is required for indicating the data analysis. | Must | 28 | This is compulsory requirement. Without these outcomes, the progress of data analysis and error rate prediction cannot be shown. |
| 3 | Kaggle wants the participants to submit the CSV and data analysis files on October 16, 2017. | Should | 2 | Outcomes should be submitted as per requirement. |
| 4 | As a data supplier, Zillow wants the participants to analyse the value of their collected data. | Should | 23 | Data analysis is aimed to find out the valuable data and their correlations. |
| 5 | As a real estate service provider, Zillow wants the participants to find out the effective prediction model to increase Zestimate’s accuracy. | Could | 25 | The most effective prediction model could improve Zestimate’s performance definitely. However, the level of effectiveness should be evaluated according to the time factor of this project. |
| ***In-Scope Points:*** | | | ***80*** | |
| 6 | Kaggle wants the participants to do contribution on their coding sharing platform and forum. | Would | 2 | This is based on the willing of participants. |
| 7 | Kaggle expects the participants to predict the home values for Zillow. | Would | 18 | Data analysis and error prediction are two current core section. Home value prediction could be done in next round of competition. |
| ***Out-Scope Points:*** | | | ***20*** | |
| ***Project Total Points:*** | | | ***100*** | |

1.3 Approach Overviews

1.4 Outcome

Data analysis report represents the insight of the datasets.

Prediction model that generate the estimated value.

1. Project Method

**Reference**

Kaggle Inc. (n.d.). Zillow Prize: Zillow’s Home Value Prediction (Zestimate). Retrieved from https://www.kaggle.com/c/zillow-prize-1#Competition Overview

Zillow Inc. (n.d.). Zestimate. Retrieved from https://www.zillow.com/zestimate/