**Laney Qin**

**Student ID: n9632981**

**Supervisor: Dr. Guido Zuccon**

**Queensland University of Technology**

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

**Table of Contents**

1. **Project Introduction1**

Background2

Scopes2

Approach Overview2

Outcomes2

1. **Project Method4**

Statistic5

Machine Learning5

R Language5

1. **Project Management Approach4**

Project Environment5

Scrum5

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

**Reference4**

**Appendix4**

Appendix A: Oral Presentation Note5

Appendix B: Supervisor Statement5

1. Project Introduction  
     
   1.1 Background

Zillow is one of the leading real estate agents in the United States and there are 110 million properties across the nation have been served on Zillow’s information site. Zestimate is their price prediction model to estimate the market value of the properties and it is part of the online services for Zillow’s website visitors. It has been released for eleven years from 2006 to 2017. With continually improving, the median error rate of Zestimate has been reduced from 14% to 5% based on advanced machine learning algorithms and millions of statistic data. Given the statistical definition of the median error rate, it means there is still a half of the properties’ error rates go off and larger than 5%. This is a problem required to be solved so as to improve the accuracy of Zestimate and the service quality of Zillow.

1.2 Scopes

Given the product owner would be Zillow, the scope stories would be relevant to them. They require the data scientists to predict errors for six timepoints: October, November and December of both 2016 and 2017. The analysing data for this project is limited to the data supplied by Zillow only. The training data is through 2016 with 3millions of properties in three counties of California, including Las Vegas, Orange and Venus. Each property has around 57 variable data such as the room number, the square footage and the location that are the factors influencing their sales price. Then the test data of 2017 will be available on the beginning of October of 2017.

1.3 Outcomes

1.4 Approach Overview

## Background (http://www.urbanash.com/blog/zillow-seattle-zestimate-wrong/)

The news tells the reason why there is a contest with high prize.

Zillow is the most popular online real estate information site with 73 million unique visitors.

"Zillow's accuracy has a median error rate of 5%."

What does that mean?

$700,000 \* 5% = $35,000 a price disparity (difference) of $35,000

Median error rate means there will be a half amount of estimated home value if exceed 5% and the other half is closer to 5%.

## Statistic Section

<https://www.inorganicventures.com/accuracy-precision-mean-and-standard-deviation>

Accuracy: logerror = log(Zestimate) – log(Sales)

Total error = sum of systematic error + sum of random error(+/-)

Precision: deviation (mean and standard deviation)

Accuracy != Precision because systematic error

Mean is an estimate of the true values as long as there is no systematic error.

There are certain basic concepts in statistics which are helpful to the analyst when treating analytical data. This project will address median, mean and deviation as related to accuracy measurements.

To calculate the logerror is helpful to understand the reality of error.

There are quantitative measurements.

I have to make logerror estimates referring to some available data but not including the actual sale price of the home value.

Errors are classified as systematic (determinate) and random (interminate). (definition could be found on this website)

So how to measurement the error?

<http://mathbits.com/MathBits/TISection/Statistics2/logarithmic.htm>

Logarithmic Regression Model

R Studio

How to use merge to find the intersection of data

(http://www.dummies.com/programming/r/how-to-use-the-merge-function-with-data-sets-in-r/)

## R Section