**Zillow Price**

**Milestone Report 1**

**1 - Introduction**

Real estate investment is a business activity that most people are interested in, both on buying and selling demands. An accurate prediction on the house price is important to potential homeowners, real estate investors, mortgage lenders and insurers. Previous price models have been commonly used to estimates prices based on housing attributes like location, house size, number of bedrooms, age of the house, nearby school, etc. However, the traditional approaches significantly depend on human judgement that take a lot of times for analyses and can have mistaken of missing information when analyzing a large number of data.

**2 - Objective**

One of the fundamental tasks in house sale analysis is the study the relationship between the value of the houses and the market price. With the house prediction models, our goal is to assist homeowners, investors, appraises, tax assessor and other real estate agents, banks and lenders to get more accurate price prediction based on cost and sale price. The target of this project is to estimate the log error of the price prediction.

**3 - Dataset**

This project uses the dataset provided by Zillow that contains housing price data in 2017. The dataset includes 2 files, a train data and a properties data.

* The train data contains log error which is log error of the prediction price. The train data file has 77613 transactions from Jan to Sep, 2016.
* The properties file contains a list of real estate sale properties in California in 2016 (including three counties Los Angeles, Orange and Ventura). It has 2985217 rows and 58 columns.

The target value is ‘logerror’ which is listed in the train data. We merged 2 file based on the parcedid. After merging 2 files, we have the total of 77613 rows, with 61 columns.

Link to the dataset:

<https://www.kaggle.com/c/zillow-prize-1/data>

**3.1 Dataset Characteristics**Challenges with the dataset:

* There are a significant amount of missing data that make those features meaningless in terms of analysis. There are about 50% of the columns have more than 65% missing values.
* The data set has too many features, thus feature selection or elimination is crucial in order to have a good model.

**4. Data Wrangling**

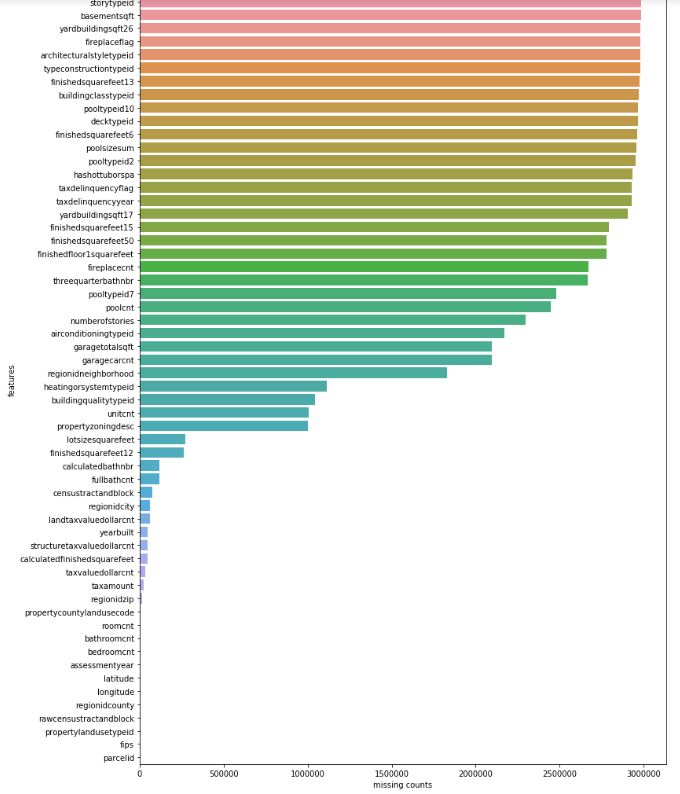
In this step, we will do data cleaning to prepare a good input dataset for the model. We will do analysis on perform data wrangling steps including dealing with missing values, data imputation, outliers’ treatment, correlation analysis and drop some unimportant variables.

**4.1 Missing values**

We remove all the columns that have more than 65% of missing values because those features do not contribute much to the model. After removing those columns, the dataset contain 77613 records with 30 features.

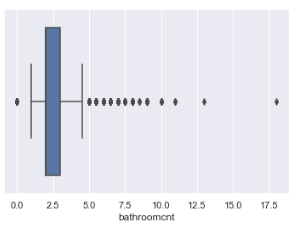
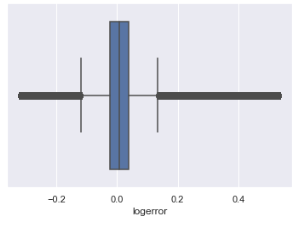
The remaining features are:

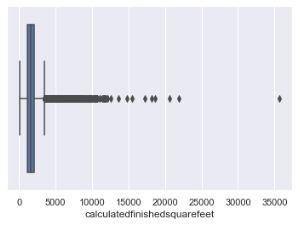
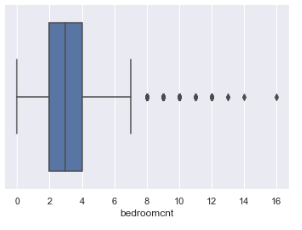
* 'logerror',
* 'transactiondate',
* 'transaction\_month',
* 'bathroomcnt',
* 'bedroomcnt',
* 'buildingqualitytypeid',
* 'calculatedfinishedsquarefeet',
* 'fips',
* 'heatingorsystemtypeid',
* 'latitude',
* 'longitude',
* 'lotsizesquarefeet',
* 'propertycountylandusecode',
* 'propertylandusetypeid',
* 'propertyzoningdesc',
* 'regionidcity',
* 'regionidcounty',
* 'regionidneighborhood',
* 'regionidzip',
* 'roomcnt',
* 'unitcnt',
* 'yearbuilt',
* 'structuretaxvaluedollarcnt',
* 'taxvaluedollarcnt',
* 'censustractandblock'

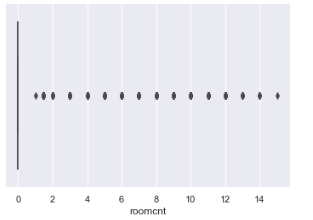
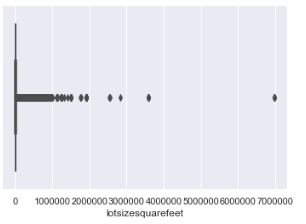


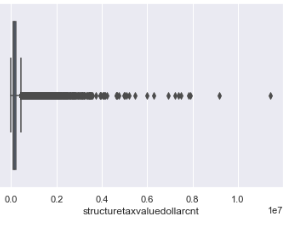
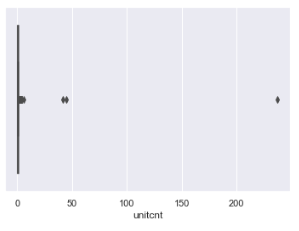
*Percentage of features that have missing values*

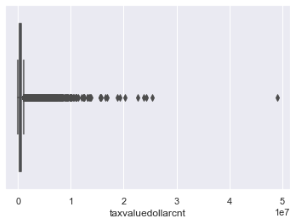
**3.3 Outlier**

We have a number of outliers in the dataset, which cause effect on the performance of the models such as the number of rooms, the number of bathrooms, bedrooms, square feet, tax amount etc. 



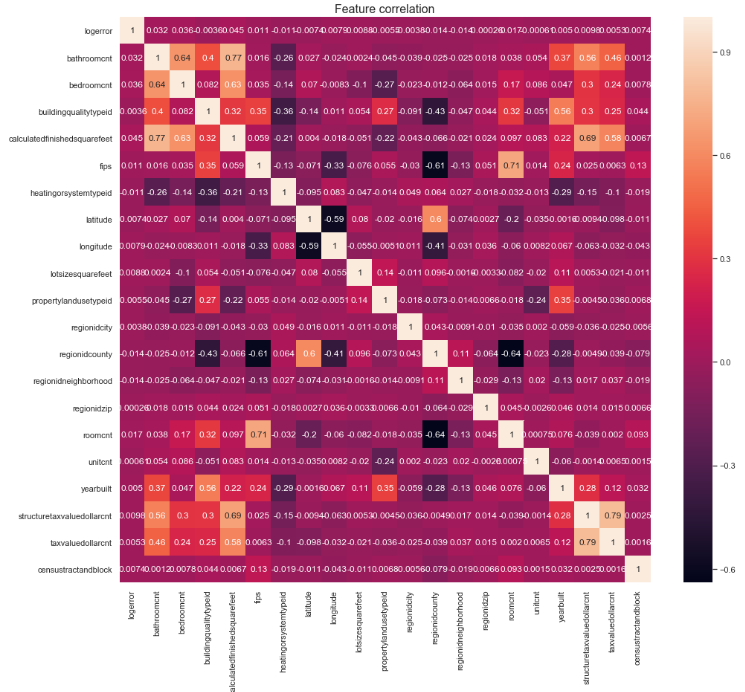






**3.4 Correlation Analysis**

The target value ‘logerror’ is not highly correlated with other features. Also, the other features are not highly correlated with each other.

****

**4. Exploratory Analysis**

We want to analyze the properties of the dataset to understand more about the house properties of house sales on the market in 2017.

**4.1. Number of house sales per month**

The data contain transactions from Jan to Sep.

* The average number of transactions is about 8623 transactions.
* In September, there are lowest transactions with less than 5000 transactions while June has highest number of transactions with more than 11,000.
* We can see that the number of house transactions are very high from Spring (March) to summer (August) and lower during fall (from Sep) to winter (to the next Feb).
* Sales were down 48% in September.
* It is likely that people start to buy or sell their houses because they relocate to other area during that time because of their work relocation or their kid school transferring.

|  |  |
| --- | --- |
| *Number of house sale in 2017* | *House Sales in California Counties* |

**4.2 Which county has the highest number of house sales?**

* Most of houses were sold in Los Angeles. There are 50,730 houses sold in Los Angeles; 20,631 houses were sold in Orange, and 6252 houses were sold in Ventura.
* The average number of home sale monthly in Los Angeles is 5633 houses; in Orange County is 2,292 houses; and in Ventura is 694 houses.

**4.3 What is the number of bedrooms/bathrooms that were mostly sold in California?**

Most of houses in California have 3 rooms, and 2 bathrooms. Only a small number of houses have more than 8 rooms. There are 20,166 houses in Los Angeles; 7858 houses in Orange County and 2412 houses in Ventura have 3 rooms.

|  |  |
| --- | --- |
| *Number of bedrooms* | *Number of bathrooms* |

Most people were looking for houses with 2 bathrooms. 17% of house sales has two bathrooms, 3% has two and a half bathrooms, 9% has three bathrooms, 7% has 1 bath room, and less than 1% have more than 7 bathrooms.

#### 4.4 What is the building quality of home sale in California?

|  |  |
| --- | --- |
|  |  |

The quality of buildings were listed from the best (lowest=1) to worst (highest=12).

* 54% of house sales in California have quality of 8, which is the most common quality of building were listed. There are 20% in Los Angeles, 26% in Orange and 8% in Ventura.
* In Orange and Ventura County, the building quality all is evaluated as 8. However, there are a variety of building quality types in Los Angeles that were evaluated from the best (listed as 1) to the worst (listed as 12).

#### 4.5 What is common house lot size sales in California?

* Los Angeles houses are 2 times bigger than the house lot size in Orange and Ventura County. Orange and Ventura mostly have similar lot size.
  + LA average lot size: 38319.49
  + Orange average lot size: 14008.56
  + Ventura average lot size: 14934.37

#### 4.6 When were the houses built in California?

The houses listed in the dataset were built from 1824 to 2016. Most of them were built in 1960s to 2000s. In Los Angeles, houses were older than in Orange and Ventura. Most house sales were built from 1950s, while in Ventura, houses were built from 1960 to 1970s. And in Oranges, most of houses sale were built from 1970s to 1980s.

* In Orange, the oldest house was built in 1893
* In Ventura, the oldest house was built in 1880
* In Los Angeles, the oldest house was built 1824.

|  |  |
| --- | --- |
| *Built years of houses in California* | *Tax assessed value* |

#### 4.7 What is the property tax assessed value of house sales in California?

Los Angeles has more tax assessed value than Orange and Ventura, and it is highest in Feb 2017. Ventura has the lowest tax assessed value.

#### 5. Statistical Inferences

Though the data analysis, we have more insights about the house properties. In the next steps, we will explore and check if the data is stationary using some plots, statistics, and Dickey-Fuller test for stationary.

**5.1 How does the log error change over time?**

|  |  |
| --- | --- |
| *Log error* | *Log error distribution* |

The log error has Gaussian distribution. It is likely that the sale prediction log error does not change over time. The mean and variance of log error in Los Angeles and Orange counties are equal to 0. And p-value is equal to 0. So we can conclude that the log error is stationary.

**5.2 How does the total tax assessed value of the parcel change over time?**

|  |  |
| --- | --- |
|  | It is likely that there is no visible trend of total tax assessed value of the parcel in 2017. We compare mean and variance of Los Angeles and Orange counties and see no difference. The p-value is equal to zero. So we can conclude that there is no trend in the total tax assessed value of the parcel over time. |

**5.3 Is there any trend with the year when the houses were built?**

|  |  |
| --- | --- |
|  | The plot shows that there is no trend of the built year of houses. There is no difference between mean and variance of houses in Los Angeles and Orange counties. The p-value is equal to zero, so we reject the null hypothesis and conclude that the year built is stationary. There is no trend over time with the data. |

**6. Next steps**

In this report, we have shown the distribution and insights about the dataset. We know that there are more houses are in Los Angeles than other counties like Orange and Ventura. Houses in Los Angeles are bigger, and were built before houses in other counties. People are more likely to buy houses with 2 or 3 bedrooms, and 2 bathrooms. The quality of the houses are normally rated at 8, and the sales were increasing from March to September.

In next steps, we will implement the linear regression model on the dataset that we have prepared and compare the results with the logerror that we have in the data. Then, we will try other machine learning models and compared with the baseline model. We can also apply feature engineering to prepare a good dataset, as well as doing hyper parameter turning to improve the model performance.