Springboard Data Science Career Track Program

Milestone Report 2

Zillow House Prediction

A case study of Machine Learning algorithm

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

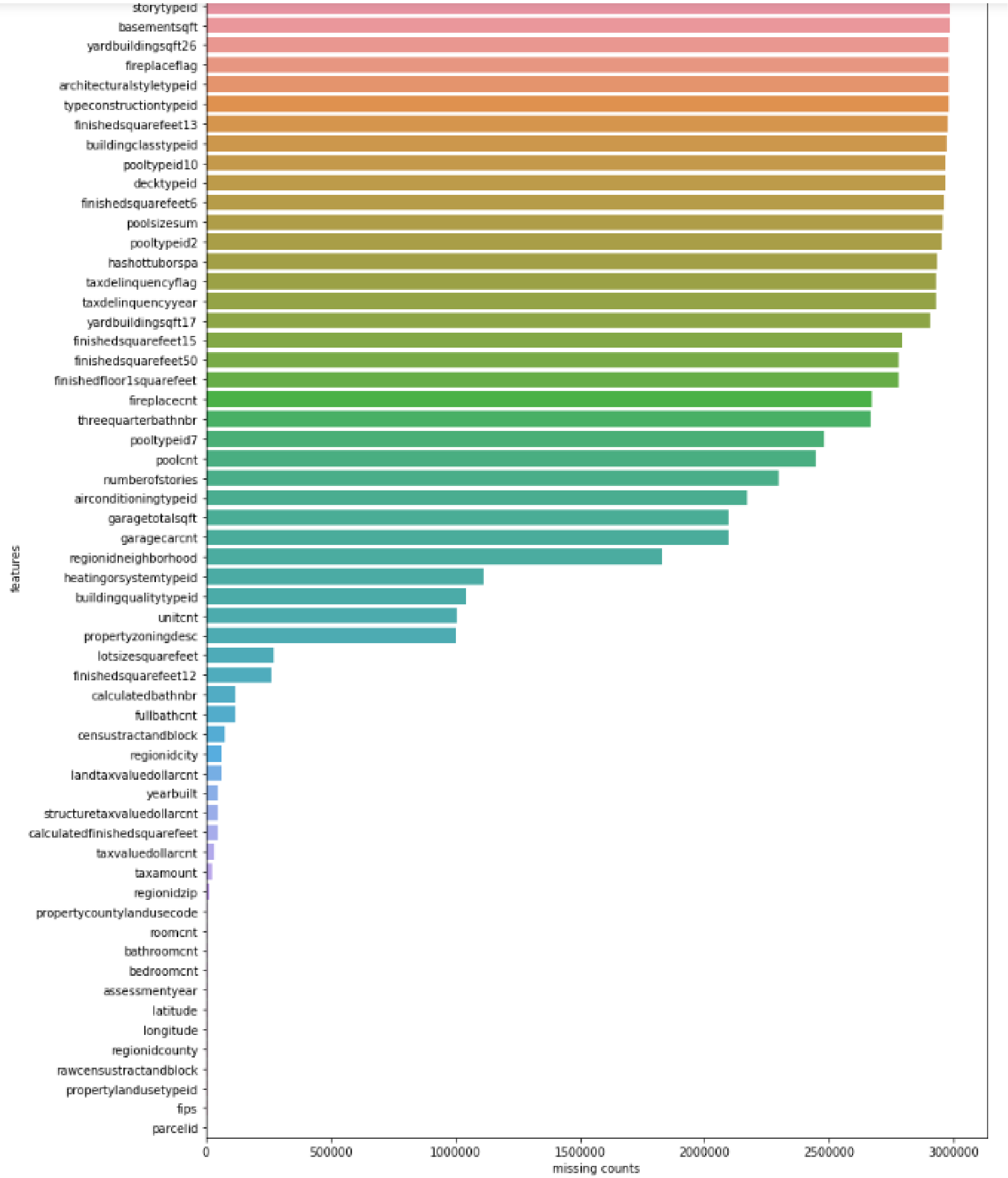
In this step, we will do data cleaning to prepare a good input dataset for the model. We will do analysis on perform data wrangling steps including dealing with missing values, data imputation, outliers' treatment, correlation analysis and drop some unimportant variables.

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:

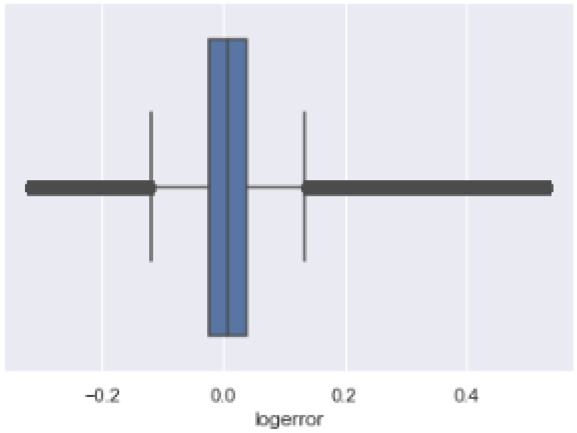
- 'logerror',
- 'transactiondate',
- 'transaction_month',
- 'bathroomcnt',
- 'bedroomcnt',
- 'buildingqualitytypeid',
- 'calculatedfinishedsquarefeet',
- 'fips',
- 'heatingorsystemtypeid',
- 'latitude',
- 'longitude',
- 'lotsizesquarefeet',
- 'propertycountylandusecode',
- 'propertylandusetypeid',
-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

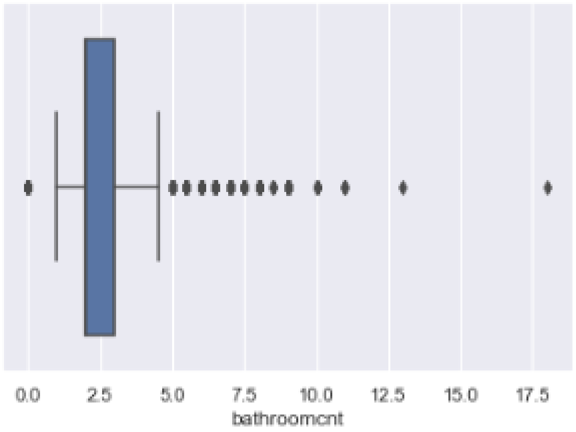


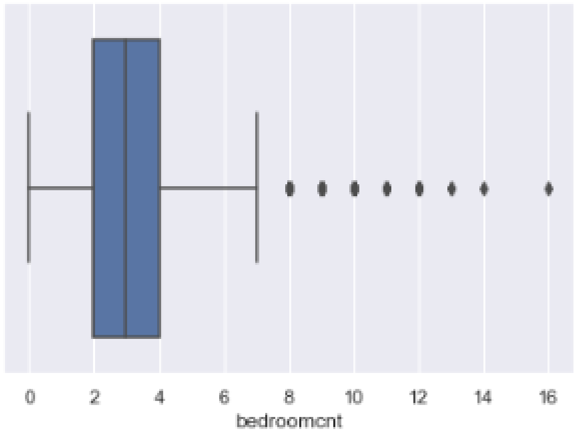
Percentage of features that have missing values

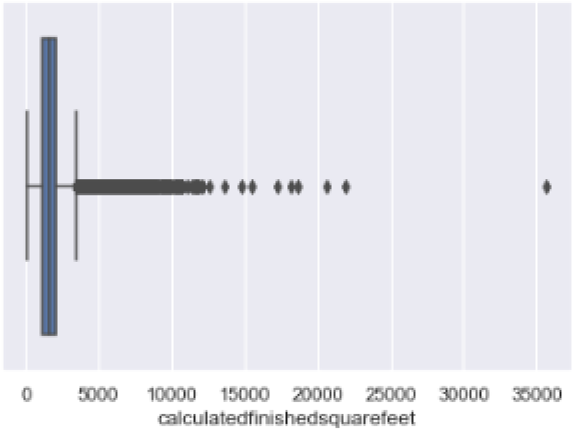
3.3 Outlier

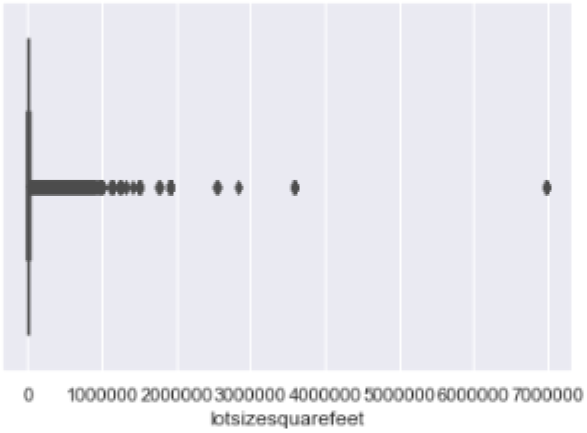
We have a number of outliers in the dataset, which cause effect on the performance of the models such as the number of rooms, the number of bathrooms, bedrooms, square feet, tax amount etc.

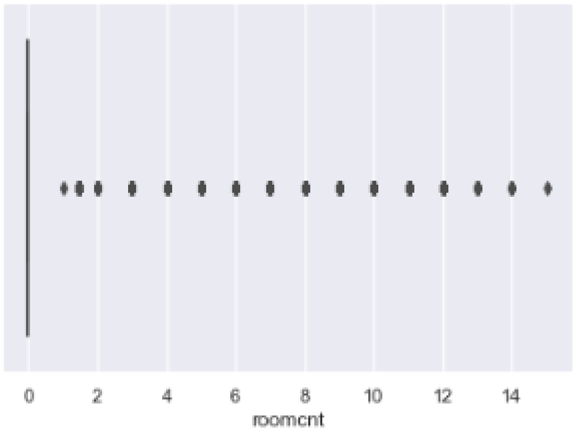


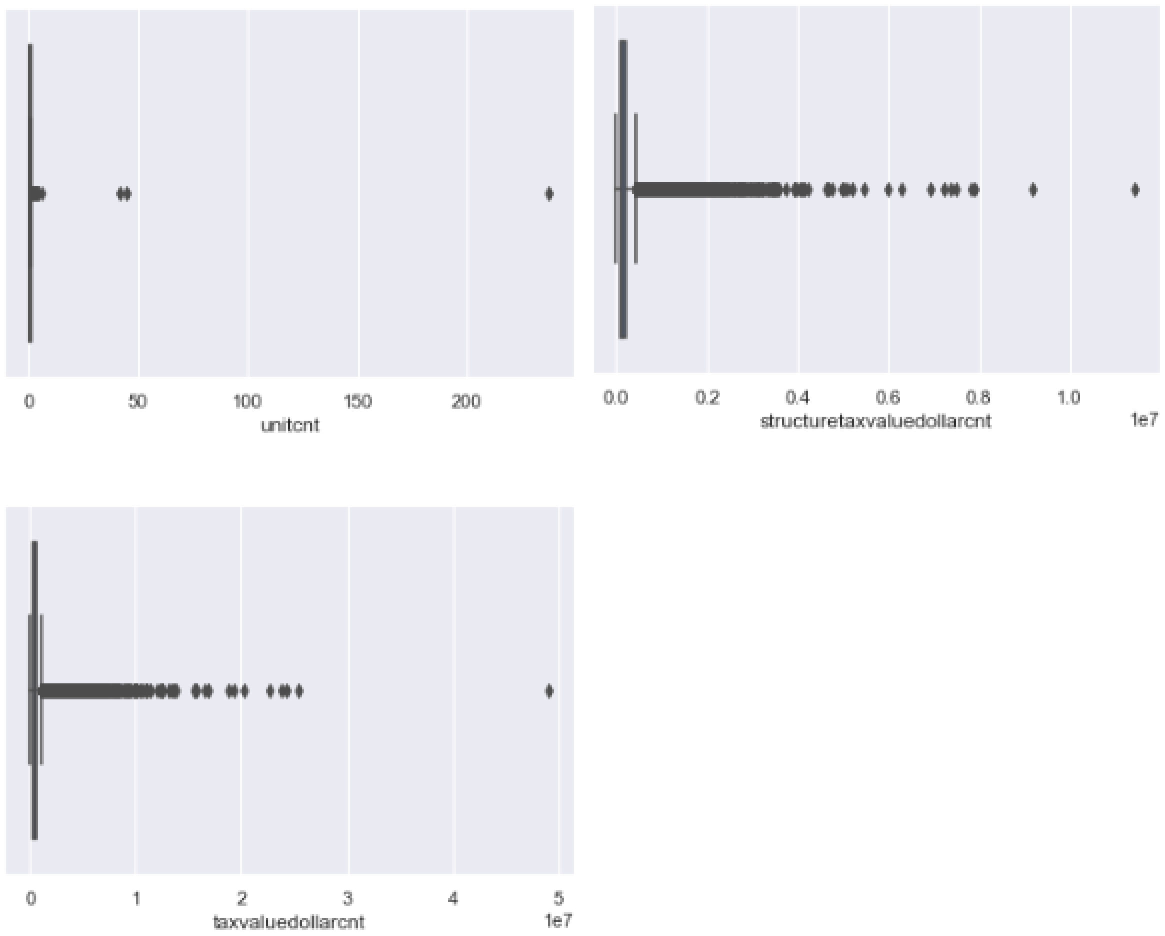






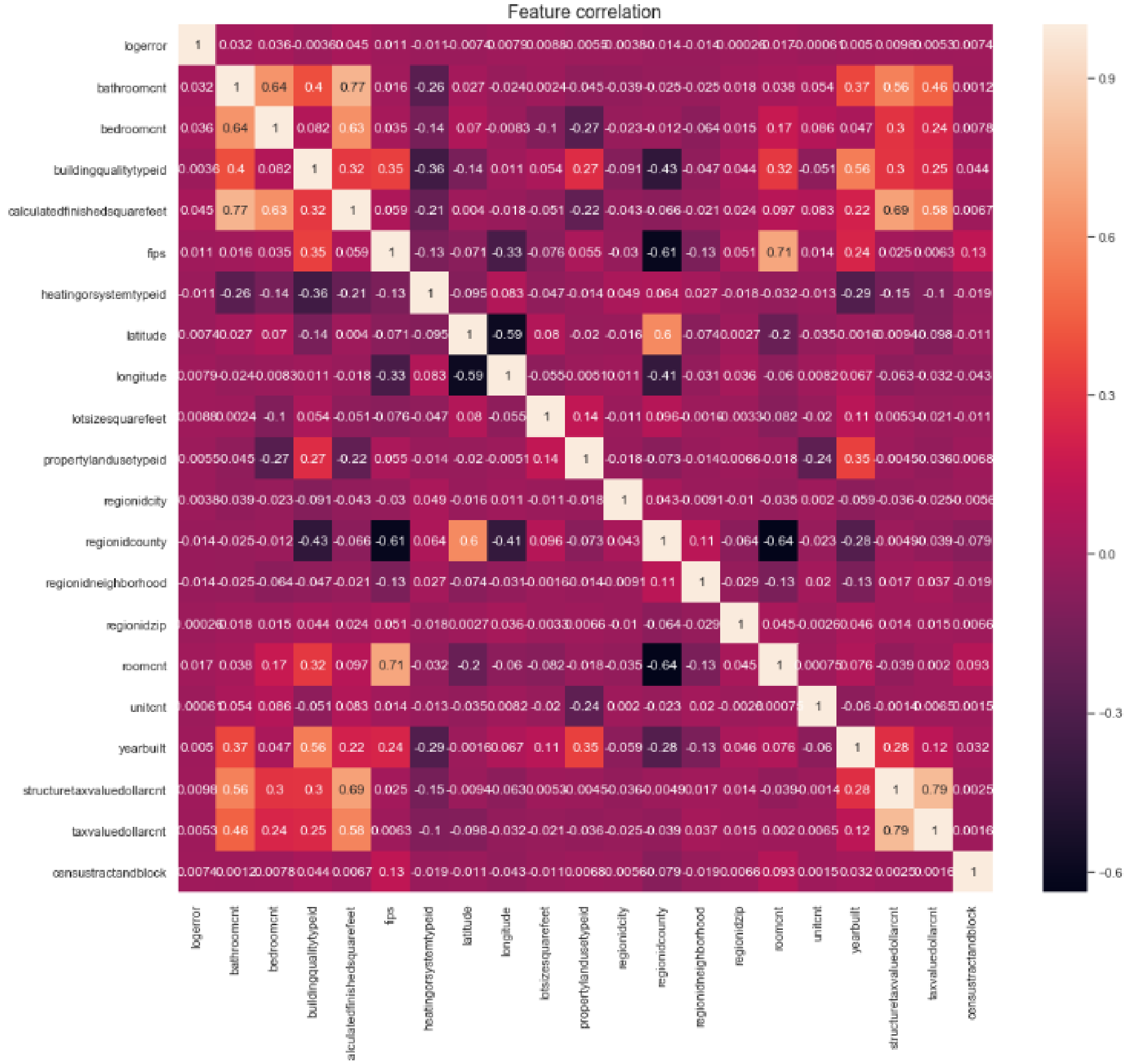






3.4 Correlation Analysis

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

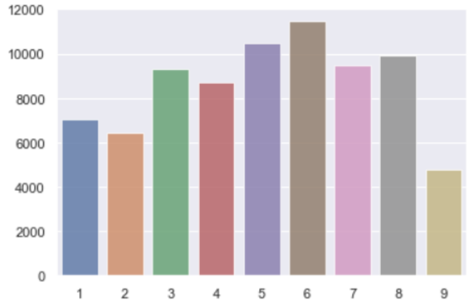
We want to analyze the properties of the dataset to understand more about the house properties of house sales on the market in 2017.

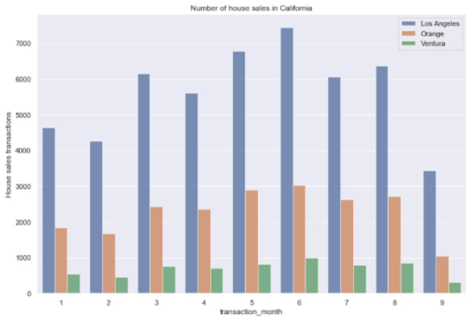
4.1. Number of house sales per month

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

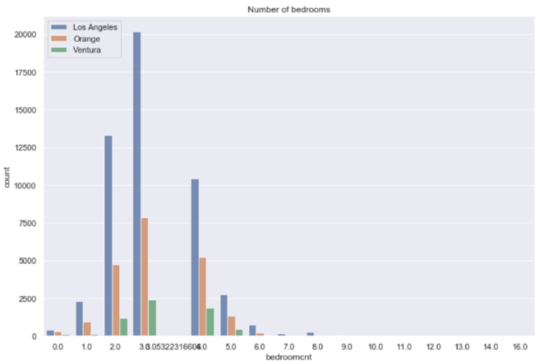
House Sales in California Counties

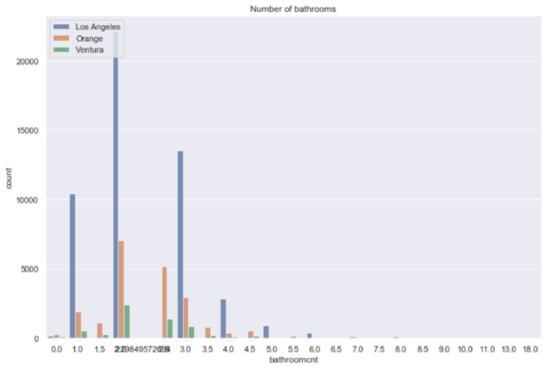
4.2 Which county has the highest number of house sales?

- 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.

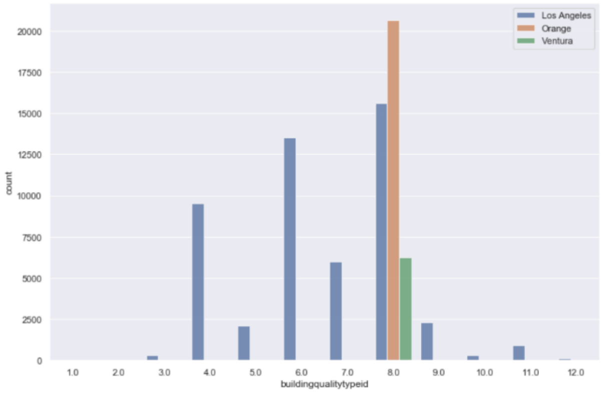


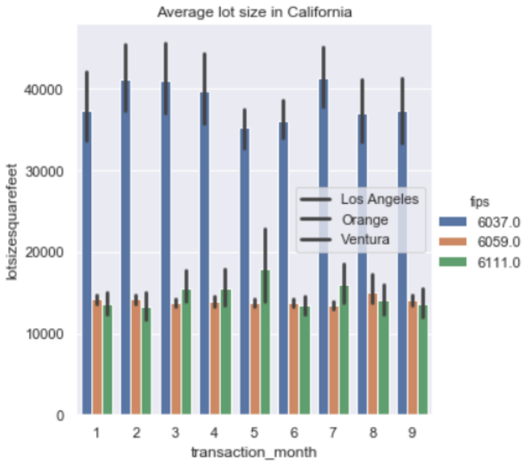


Number of bedrooms *Number of bathrooms*

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?

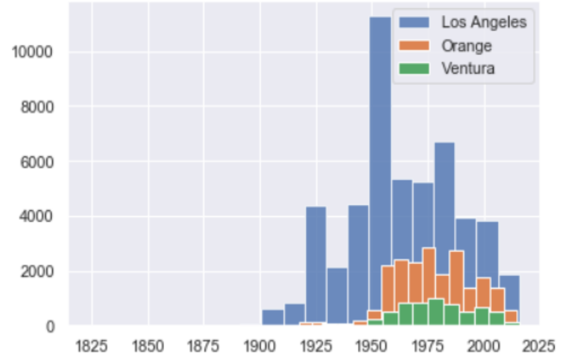
- Los Angeles houses are 2 times bigger than the house lot size in Orange and Ventura County. Orange and Ventura mostly have similar lot size.

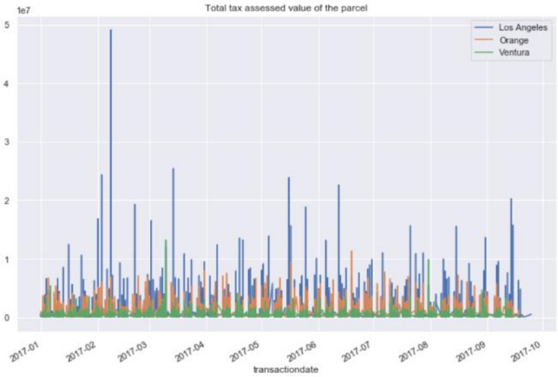
  - LA average lot size: 38319.49
  - Orange average lot size: 14008.56
  - Ventura average lot size: 14934.37

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

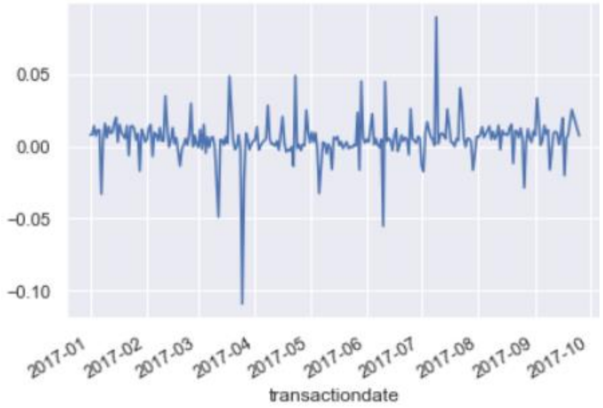
4.7 What is the property tax assessed value of house sales in California?

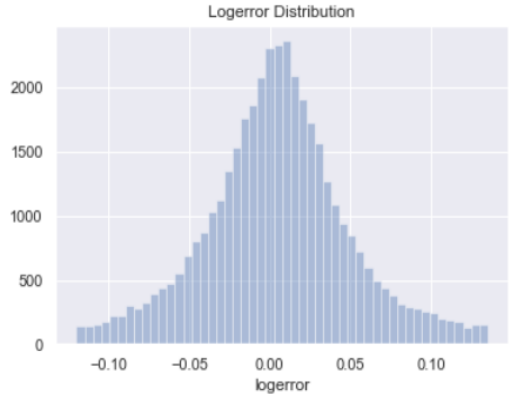
Los Angeles has more tax assessed value than Orange and Ventura, and it is highest in Feb 2017. Ventura has the lowest tax assessed value.

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?



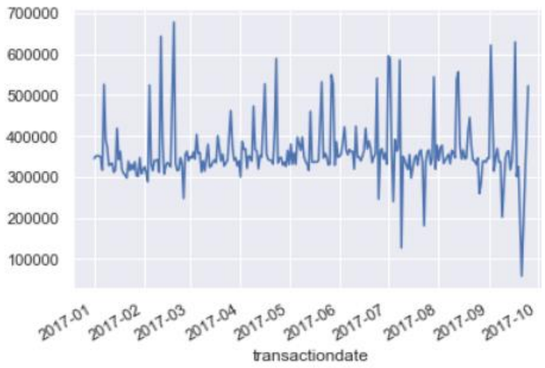


Log error

Log error distribution

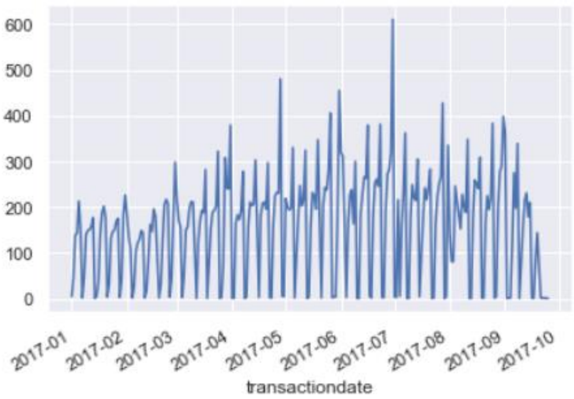
The log error has Gaussian distribution. It is likely that the sale prediction log error does not change over time. The mean and variance of log error in Los Angeles and Orange counties are equal to 0. And p-value is equal to 0. So we can conclude that the log error is stationary.

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?



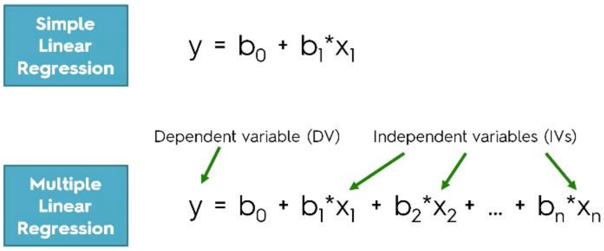
The plot shows that there is no trend of the built year of houses. There is no difference between mean and variance of houses in Los Angeles and Orange counties. The p-value is equal to zero, so we reject the null hypothesis and conclude that the year built is stationary. There is no trend over time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

Based on the dataset exploration, 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.

6. Baseline model

Regression analysis is a basic method used in statistical analysis of data. It’s a statistical method which allows estimating the relationships among variables. Linear regression models assume that the relationship between a dependent continuous variable Y and one or more independent variables X is linear.

It’s used to predict values within a continuous range (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). Linear regression models can be divided into Linear Regression and Multiple Linear Regression.


Simple Linear Regression
$y = b_0 + b_1 * x_1$
Dependent variable (DV)
Independent variables (IVs)
Multiple Linear Regression
$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$

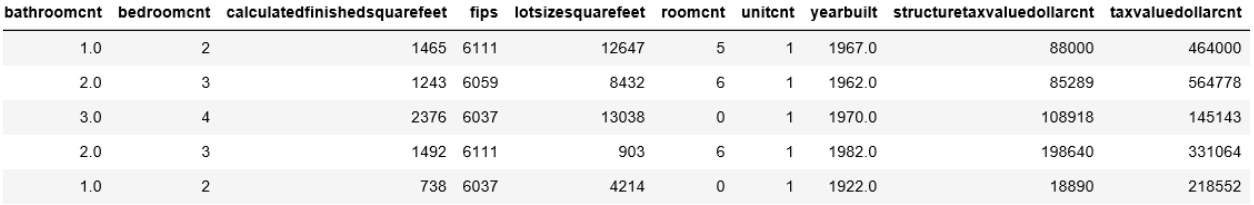
We choose Linear Regression as the base line model to predict the log error. We select set of numerical features to build the model and divide the dataset into 3 subsets based on the FIPS code. Houses in similar area code will have similar attributes like price, areas, number of rooms, etc. Since we have multi variables in the dataset, we will use the multivariable regression.

6.1 Features

We want to predict log error of prediction price for 3 counties in California including Los Angeles, Orange, and Ventura. The data was divided into a training and test set based on the transaction date. The training data contains house sale before 7/31/2017 and the test data contains house sale after that. The subsets of the data have the size as follows:

	Los Angeles	**Orange**	**Ventura**
Training set	20657	7932	2894
Test set	5690	2066	798

Each training subset data contains the following features:



bathroomcnt	bedroomcnt	calculatedfinishedsquarefeet	fips	lotsizesquarefeet	roomcnt	unitcnt	yearbuilt	structuretaxvaluedollarcnt	taxvaluedollarcnt
1.0	2	1465	6111	12647	5	1	1967.0	88000	464000
2.0	3	1243	6059	8432	6	1	1962.0	85289	564778
3.0	4	2376	6037	13038	0	1	1970.0	108918	145143
2.0	3	1492	6111	903	6	1	1982.0	198640	331064
1.0	2	738	6037	4214	0	1	1922.0	18890	218552

6.2 Evaluation metrics

In this project, we will use the following metrics used to evaluate the results of the prediction using linear regression:

- Mean Squared Error(MSE)
- Root-Mean-Squared-Error (RMSE).
- Mean-Absolute-Error (MAE).
- R² or Coefficient of Determination.

Mean Squared Error is one of the most preferred metrics for regression tasks. It is simply the average of the squared difference between the target value and the value predicted by the regression model. As it squares the differences, it penalizes even a small error which leads to over-estimation of how bad the model is. It is preferred more than other metrics because it is differentiable and hence can be optimized better. It is always a non-negative number. Values closer to zero represent a smaller error.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

Root Mean Square Error (RMSE) is the most widely used metric for regression tasks and is the square root of the averaged squared difference between the target value and the value predicted by the model. It is preferred more in some cases because the errors are first squared before averaging which poses a high penalty on large errors. This implies that RMSE is useful when large errors are undesired.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

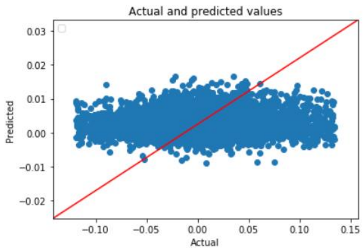
Mean Absolute Error: MAE is the absolute difference between the target value and the value predicted by the model. The MAE is more robust to outliers and does not penalize the errors as extremely as MSE. MAE is a linear score which means all the individual differences are weighted equally. It is not suitable for applications where you want to pay more attention to the outliers.

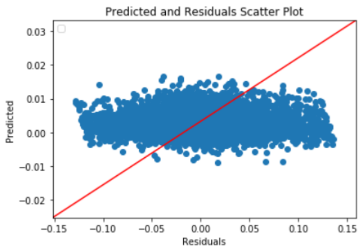
$$MAE = \overset{\text{Divide by the total number of data points}}{\boxed{\frac{1}{n}}} \underset{\text{Sum of}}{\Sigma} \underbrace{\left| \overset{\text{Actual output value}}{\boxed{y}} - \overset{\text{Predicted output value}}{\boxed{\hat{y}}} \right|}_{\text{The absolute value of the residual}}$$

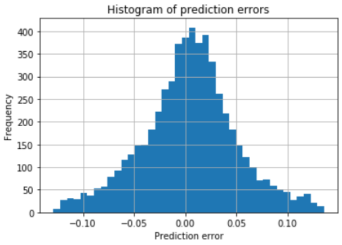
R² Error: Coefficient of Determination or R² is another metric used for evaluating the performance of a regression model. The metric helps us to compare our current model with a constant baseline and tells us how much our model is better. The constant baseline is chosen by taking the mean of the data and drawing a line at the mean. R squared value is always between 0 and 1, and that the best value is 1.0.

6.3 Results of Linear Regression

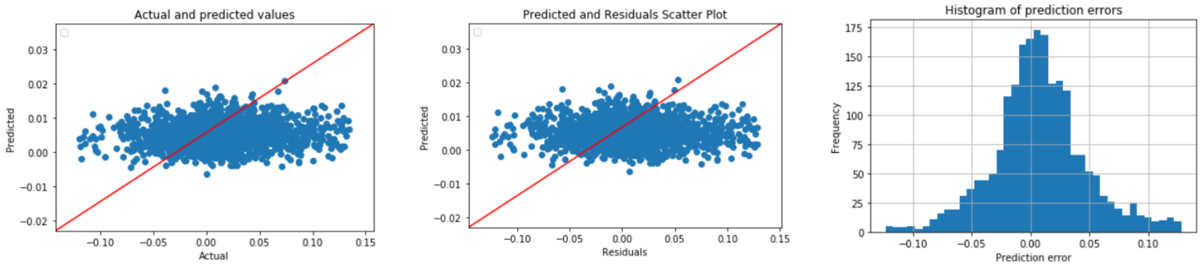
The plots show that actual and predicted values are not really good. They are far from the diagonal line. However, the prediction errors are normally distributed.



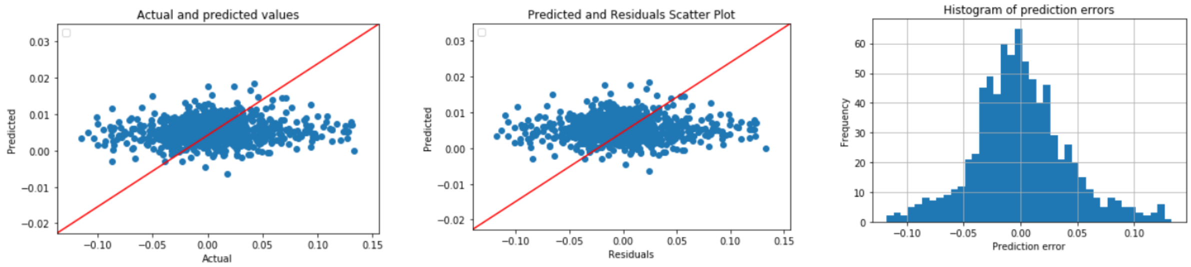




Los Angeles



Orange County



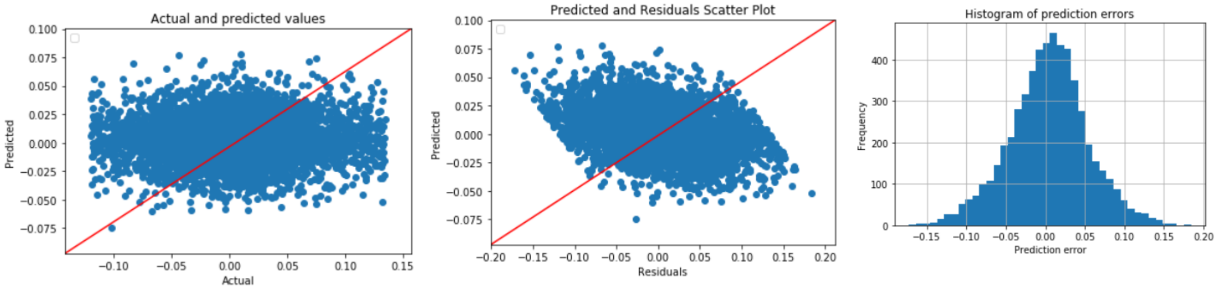
Ventura County

From the results of linear regression, we can see that the R square is very low, which mean that nearly 0% of the variance of log error can be explained by the input features. R squared is even negative for Los Angeles and Orange counties, model does not predict well on the data, it even worse than using the mean of variables. The MSE and RMSE is small which mean that the data values are closely to the central moment (mean). However, the dataset of Ventura have negative MAPE, this means that the dominator (the actual values) are negative.

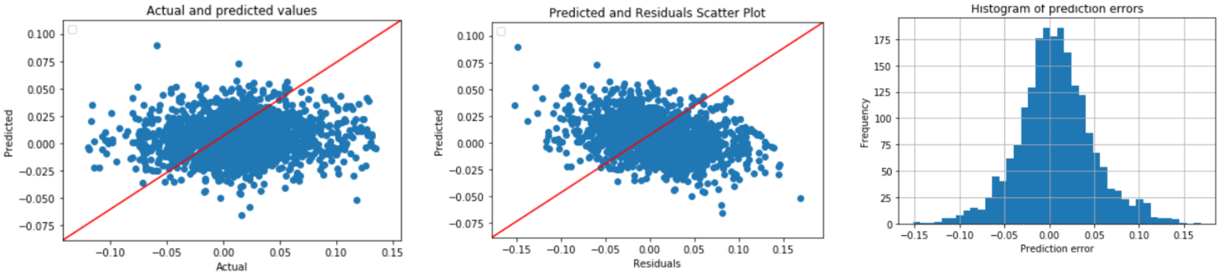
	R2	**MSE**	**RMSE**	**MAE**	**MAPE**
LA	-0.0028	0.002	0.04	0.0356	48
Orange	-0.0278	0.002	0.04	0.0298	7.2
Ventura	0.0079	0.002	0.04	0.0312	-6.5

7. Random Forest

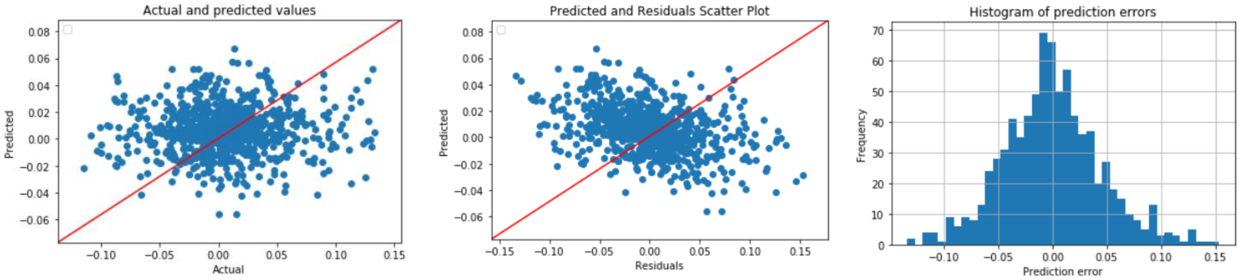
We have implemented linear regression model as the baseline model for the dataset. The results are low and the model does not perform well on prediction the data. In the next steps, we will try Random Forest and hyper parameter tuning to improve the performance of the predictions.



Los Angeles



Orange



Ventura

We have run experiments with Random Forest for the three counties including Los Angeles, Orange and Ventura. WE have R square is negative, which mean that none of the variance of log error can be explained by the input features. MSE and RMSE is small which mean that the data values are closely to the central moment (mean). However, MAPE is too high, which show the model is bad at predicting the actual values.

	R2	MSE	RMSE	MAE	MAPE
LA	- 0.15	0.002	0.04	0.0387	178
Orange	-0.15	0.002	0.04	0.0321	92
Ventura	-0.16	0.002	0.04	0.0344	96

Random Forest Results

7.1 Hyper parameter Tuning

We've built a random forest model to predict the log errors. However, the results are not very good. We will move on to model hyper parameter tuning, to optimize the random forest model. Hyper parameter is different from a parameter is that a parameter can be learned automatically from the data but a hyper parameter does not. Hyper parameters must be set manually or chosen before training process. In the case of a random forest, hyper parameters include the number of decision trees in the forest and the number of features considers by each tree when splitting a node.

Using Scikit-Learn's RandomizedSearchCV method, we can define a grid of hyperparameter ranges, and randomly sample from the grid, performing K-Fold CV with each combination of values.

As a brief recap before we get into model tuning, we are dealing with a supervised regression machine learning problem.

To find the best hyperparameters, we try out a wide range of values and see what works! We will try adjusting the following set of hyperparameters:

- n_estimators = number of trees in the foreset
- max_features = max number of features considered for splitting a node
- max_depth = max number of levels in each decision tree
- min_samples_split = min number of data points placed in a node before the node is split
- min_samples_leaf = min number of data points allowed in a leaf node
- bootstrap = method for sampling data points (with or without replacement)

We run the random forest through 3 folds, and the best hyper parameters we got are:

```
{'n_estimators': 400, 'min_samples_split': 10,
'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': 10,
'bootstrap': True}
```

To evaluate the model using hyper parameter tuning, we compare the model with Accuracy, which is equal to

```
mape = 100 * np.mean(errors / test_labels)
```

Accuracy = 100 - MAPE

		Accuracy
Linear Regression	LA	51%
Random Forest		-78%
Hyper Parameter		53%
Linear Regression	Orange	92%
Random Forest		7%
Hyper Parameter		76%
Linear Regression	Ventura	106%
Random Forest		3%
Hyper Parameter		95%

The hyper parameter tuning improve the prediction results compared with Linear Regression and Random Forest. However, for Ventura dataset, the Linear Regression have accuracy of 106% that is abnormal. It is because we have negative MAPE.

8. Recommendations for the Clients

With the results of the regression models we have implemented, it shows that these models are not suitable for accurate prediction on house price errors. It is possible to automate a lot of work that necessary to give for prediction. With a sufficient amount of datasets and reliable model, it is conceivable to develop a system which automates a lot of human works, and offer a greater level of accuracy.

9. Limitations

We have developed our model using the attributes from the dataset for houses listed on Zillow website for 2017. There are still limitations in the model in regards to the terms of timeline, percentage of growth, time decay, and seasonal factors therefore the model results are not very good at prediction of house price log error. There are several approach that consider time series forecasting like ARIMA model. It is based on a description of the trend and seasonality in the data.

10. Conclusions and Future Work

The research on real estate require a lot of analysis and efforts regarding to many different economic factors, that can bring beneifts to investors, house sellers or buyers, banks, lender, etc. Our model can be utilized to automate the forecast of house sale price and make prediction if the house is overpriced or underpriced, therefore can help client's decision making in house sale market.

The model we have developed in this report have automated some level of human manual analysis. With sufficient data, the model will be easily to automate higher level analysis and reduce a lot of time and efforts for real estate experts.

We do not explore further the idea of doing time series analysis method like ARIMA in this report and leave it for the future work.