Phuc Ton Nguyen

Prediction of pricing for Housing Data

Guided Capstone Project Report

Introduction

A local real estate company in US, asked to build a house price prediction model. The company wants to utilize the model to provide their house price estimations to their customers, house sellers and buyers. The goal of this project is collecting data with house prices and finding the best model that can predict house sale prices with the least amount of errors.

Exploratory Data Analysis

The cleaned and merged dataset contained 21,513 observations (house sales) each with 21 features. I did univariate, bivariate and multivariate analyses along with visualizations. I also performed some hypothesis testing for more statistically rigorous statements. The followings are the summary of what I found through exploratory data analysis (EDA) and inferential statistics.

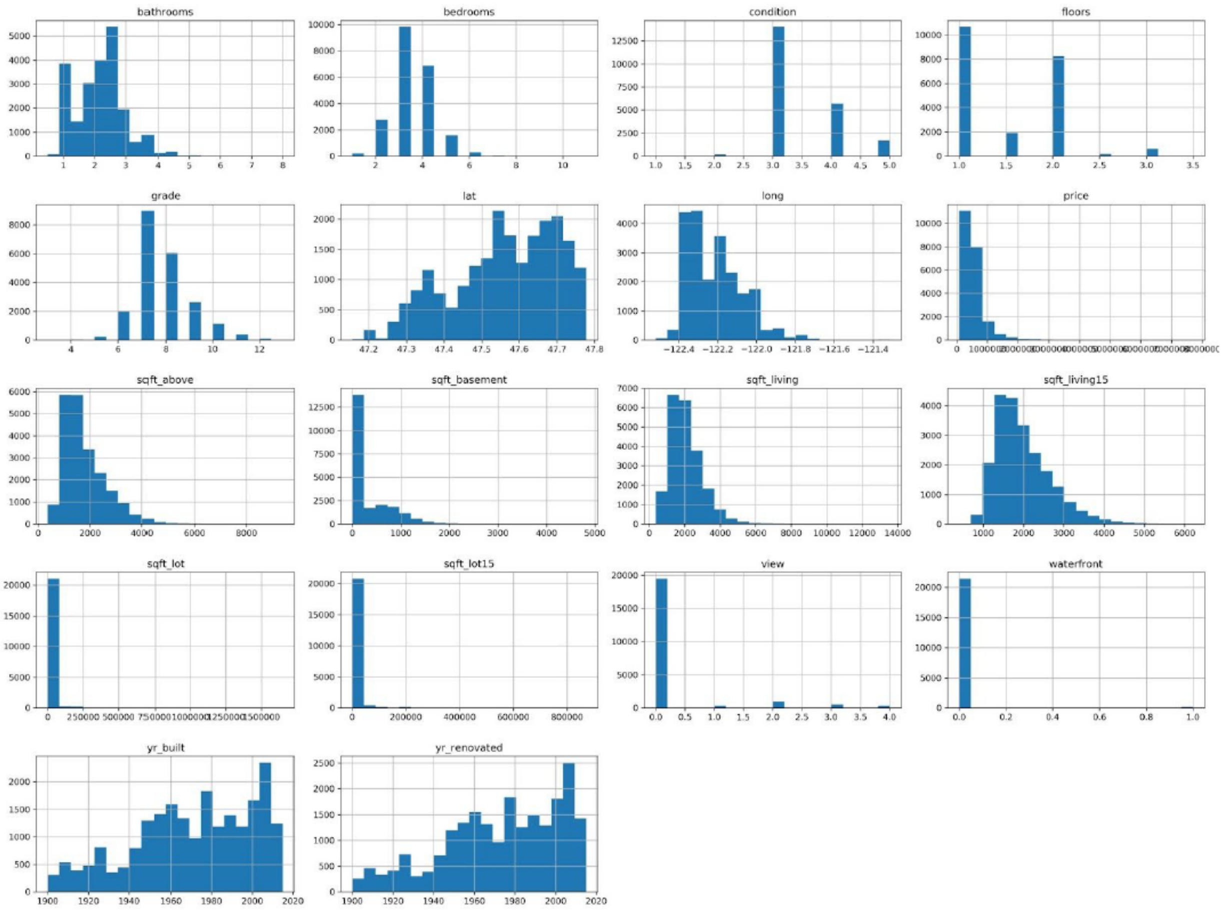
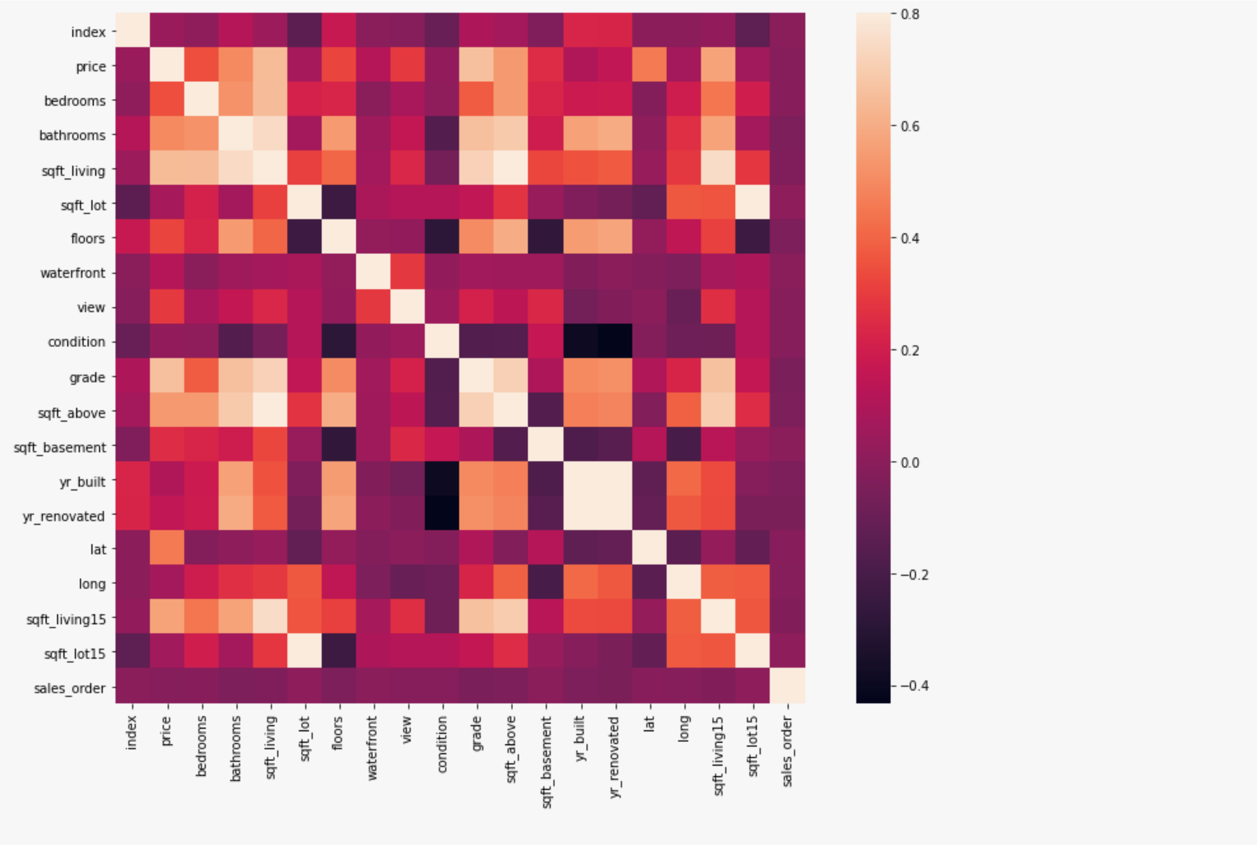


Figure. Histograms of some numerical variables

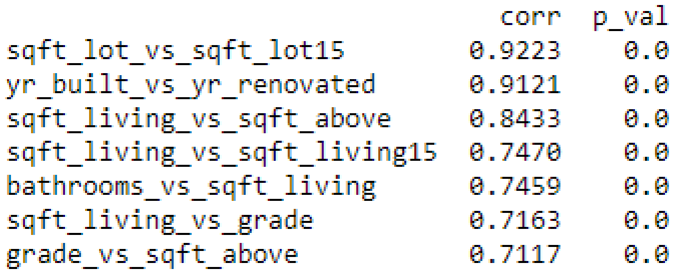
Heatmap



Correlations to price

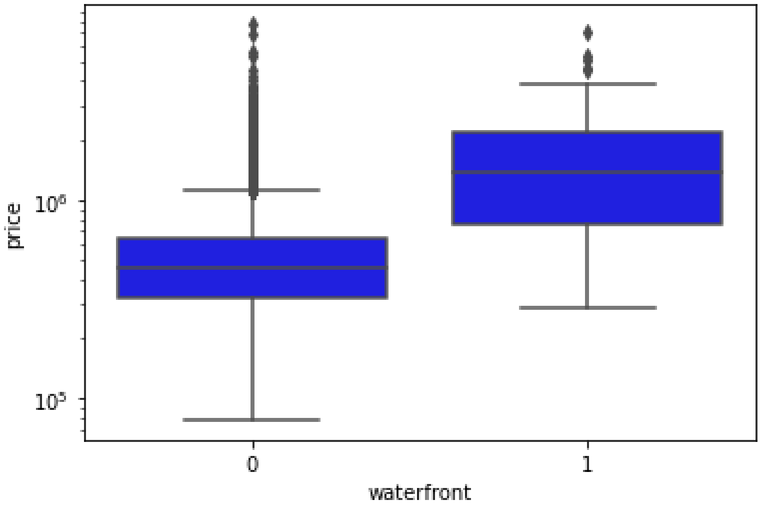
```
price           1.000000
grade            0.658150
sqft_living      0.643992
sqft_living15    0.572265
sqft_above       0.541562
bathrooms        0.497297
lat             0.456125
bedrooms         0.344176
floors           0.322480
view             0.293910
sqft_basement   0.251490
yr_renovated    0.152822
waterfront       0.115119
yr_built         0.102058
sqft_lot         0.075073
long            0.064060
sqft_lot15       0.063079
index            0.040188
condition        0.017991
sales_order     -0.014741
```

Correlation between independent variables



```
                                corr  p_val
sqft_lot_vs_sqft_lot15        0.9223    0.0
yr_built_vs_yr_renovated      0.9121    0.0
sqft_living_vs_sqft_above     0.8433    0.0
sqft_living_vs_sqft_living15  0.7470    0.0
bathrooms_vs_sqft_living      0.7459    0.0
sqft_living_vs_grade          0.7163    0.0
grade_vs_sqft_above           0.7117    0.0
```

Relationship between house price and categorical independent variables



I found the variables strongly and significantly correlated to house prices are features related to house locations square footage of house (sqft_living, sqft_above, and sqft_living15, but not sqft_basement),and house grade. These will be good predictors for a house price prediction model if high correlations between independent variables (multicolinearity) is well taken care of.

Many independent variables are highly and significantly correlated; 10 pairs have correlation over 0.7. The Highly correlated independent variable will be carefully treated in my house prediction models to avoid multicolinearity.

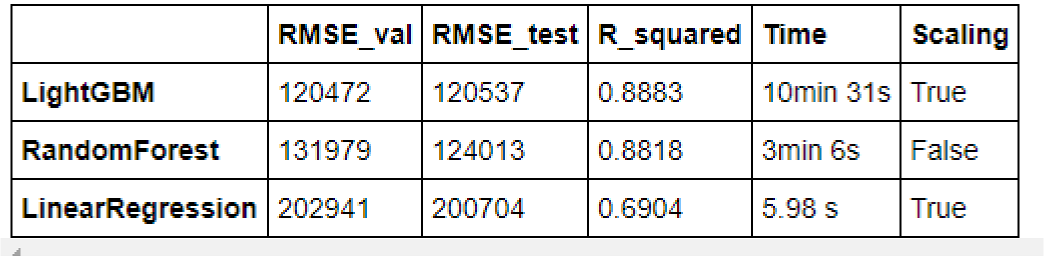
A two-sample independent t-test was used to find a significant categorical variable in a house price prediction. I used Welch's t-test which does not assume equal variance or equal sample size in two populations. I had to use it because the two groups in my categorical variables have very unbalanced sample sizes.

I found waterfront houses are significantly more expensive than not waterfront houses.

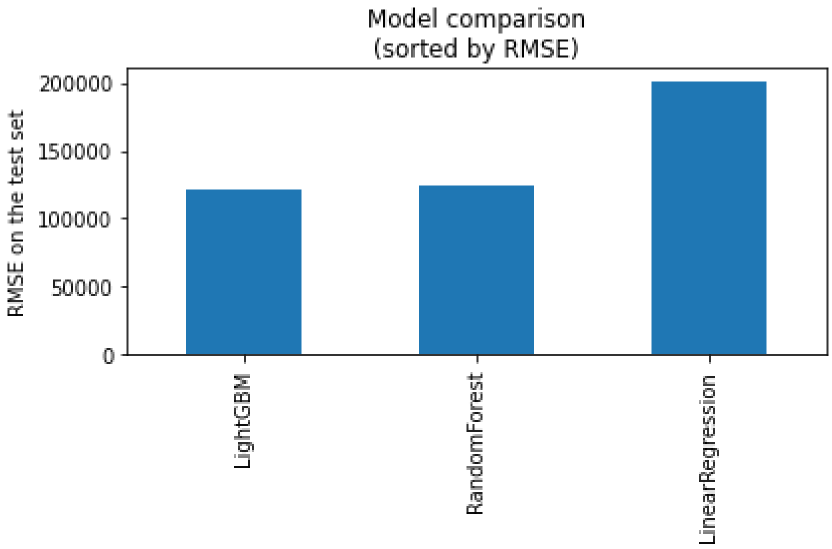
Model

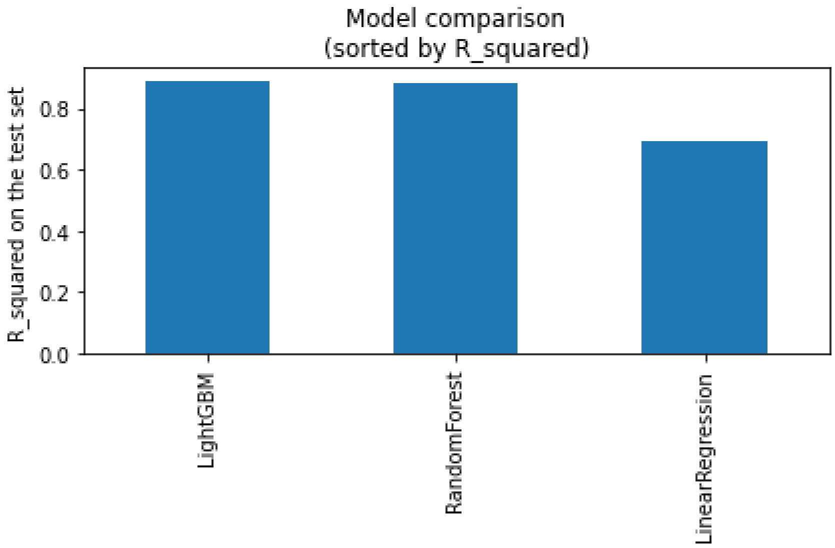
I used 3 models Linear Regression, Random Forest and Light Gradient Boosting to evaluate the models.

Result

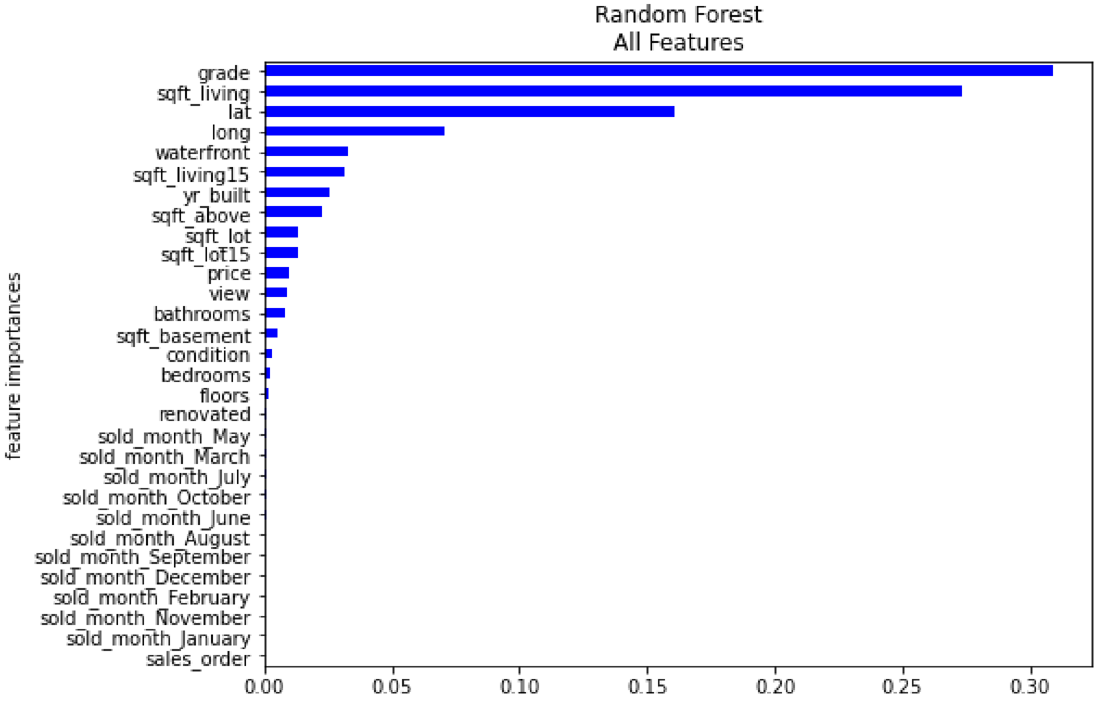


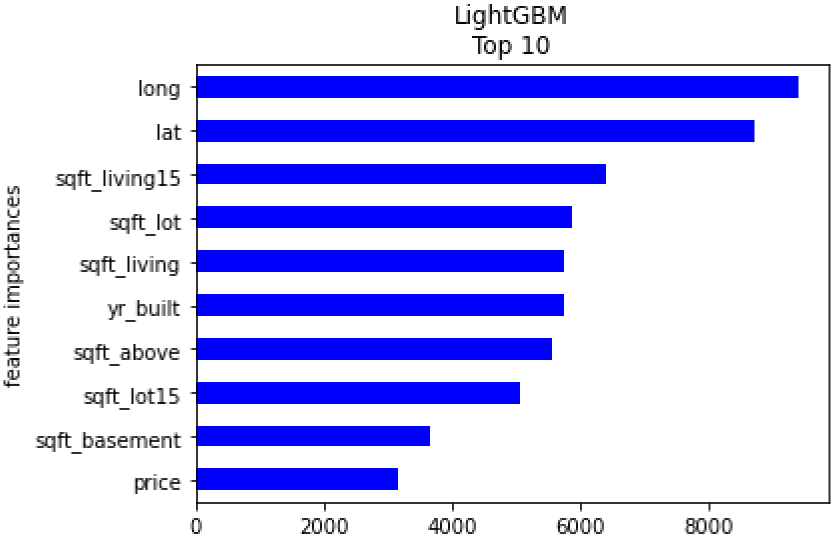
	RMSE_val	RMSE_test	R_squared	Time	Scaling
LightGBM	120472	120537	0.8883	10min 31s	True
RandomForest	131979	124013	0.8818	3min 6s	False
LinearRegression	202941	200704	0.6904	5.98 s	True

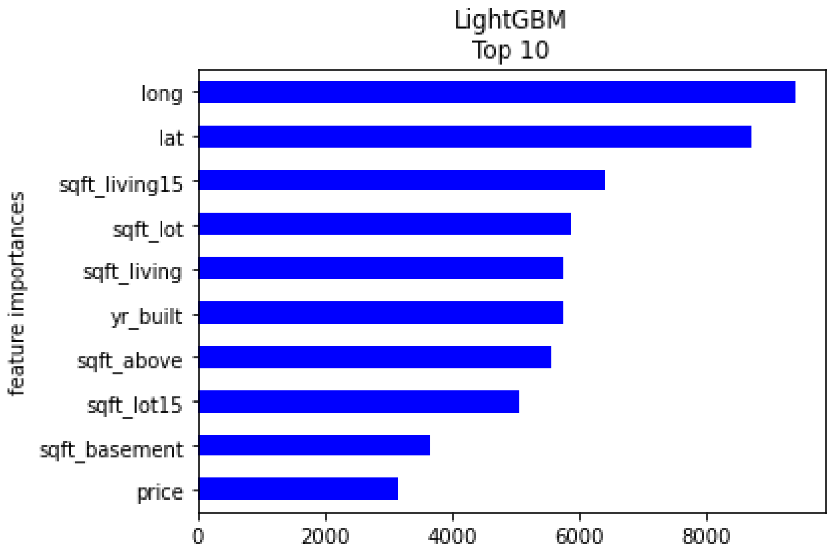




Feature Importances







Conclusion

Both EDA and machine learning showed that square footages of living area, number of bathrooms and latitudes are important features in predicting house prices and they have positive correlation with house prices. Some important features, but not linearly related to house prices, are sold months and year built.

I have found the best house prediction models is the LightGBM models which showed high speed and best performance in RMSE. If one model should be selected I would recommend to use the LightGBM model since it is faster and it makes fewer outliers.