Biqi Lin
Runyu Yan
Yuan Ling
Dajiang Liu

NYC House Sale Price Prediction Report

Agenda

- Introduction
- Methodology
- Results
- Conclusion

I.

Introduction

Figuring out how to predict the property sales is always one of the most important economical topics on the table. As we all know, House price trends are not only the concern of buyers and sellers, but it also indicates the current economic situation. However, there are various factors that might influence the property market and the house price…

“Prices of real-estate properties is critically linked with economy ”

Objective

- Predict the future sales price of NYC Property by using a year's worth of raw transaction records

Hypothesis

- There is a correlation between the sales price of property and the characteristics of the property itself and its neighborhood

Dataset

- Properties sold in New York City over a 12-month period from September 2016 to September 2017 (From Kaggle.com)
- 84548 rows and 22 columns

HBORHOOD	BUILDING CLASS CATEGORY	TAX CLASS AT	BLOCK	LOT	EASE-MENT
ABET CITY	07 RENTALS - WALKUP APARTMENTS	2A	392	6	
ABET CITY	07 RENTALS - WALKUP APARTMENTS	2	399	26	
ABET CITY	07 RENTALS -	2	399	39	

II.

Methodology

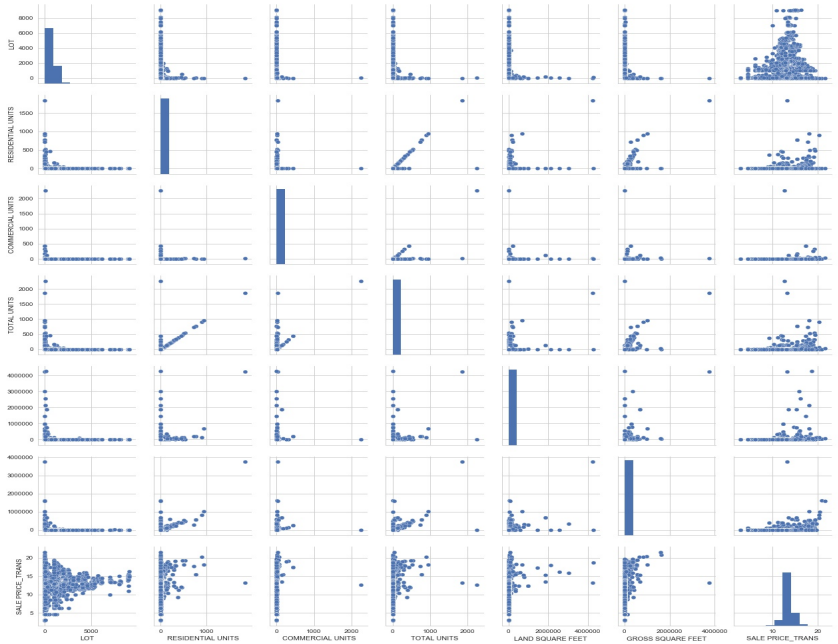
Data Pre-processing

- Missing Value
- Near-zero variance features
- Skewness

Explanation Data Analysis

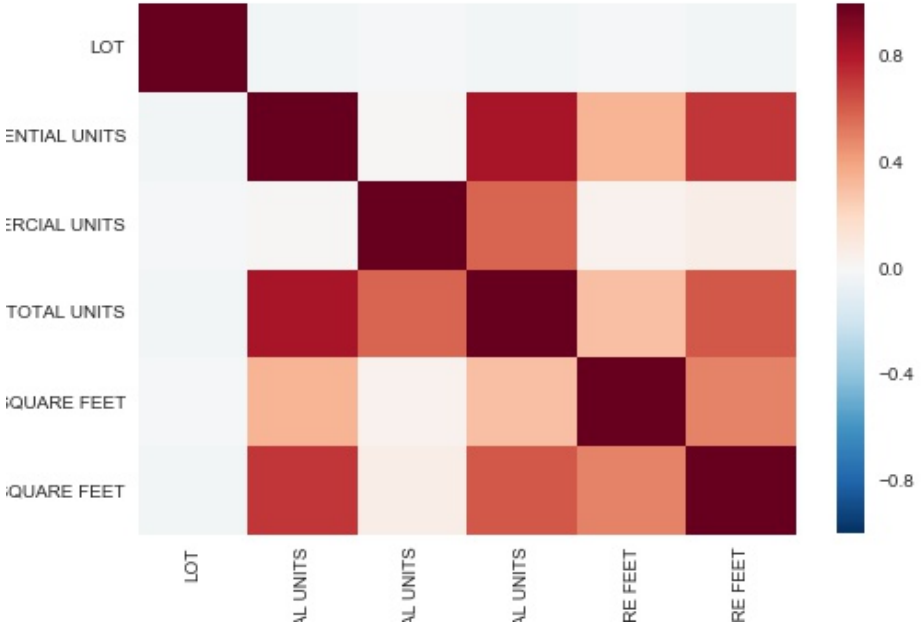
- Numerical Features
- Categorical Features

Numerical Features
Scatter Plot Matrix



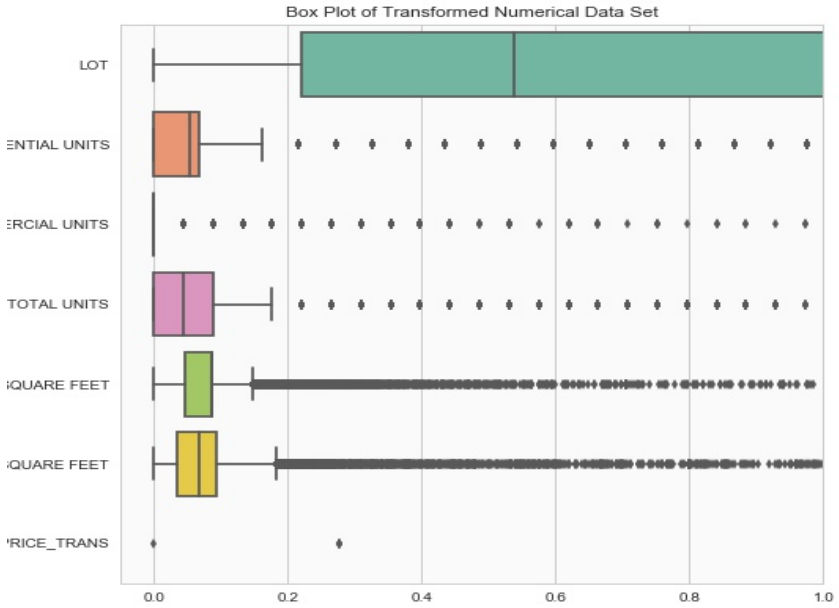
No clear pattern between all features and sale price

Numerical Features
Correlation Matrix



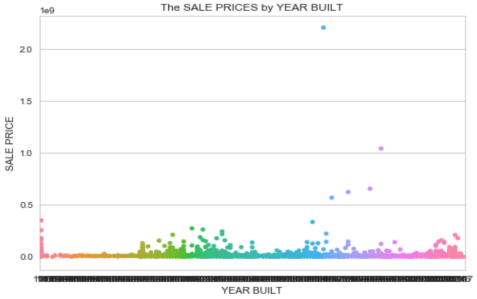
Total Units Residential Units and Residential Units & Gross Square Feet had relative high correlations

Numerical Features
Boxplot



Both Land Square Feet and Gross Square Feet had a lot of outliers

Categorical Features
Bar Plots



The sale price would increase as the property's tax class increase

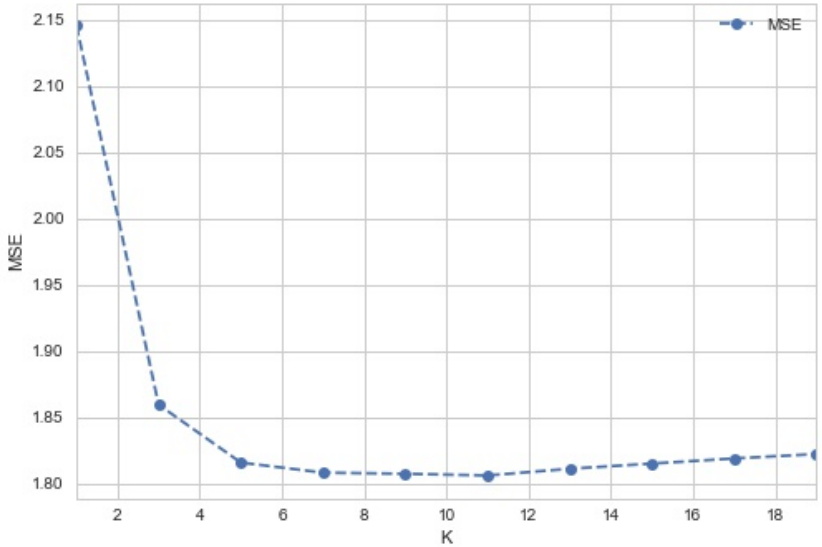
Models

III.

Results

The whole data set was randomly divided into the training data and the testing data with the ratio 7:3. The training set was used to apply 10-fold cross-validation with the Mean Standard Error (MSE) for selecting the optimal tuning parameter for each model.

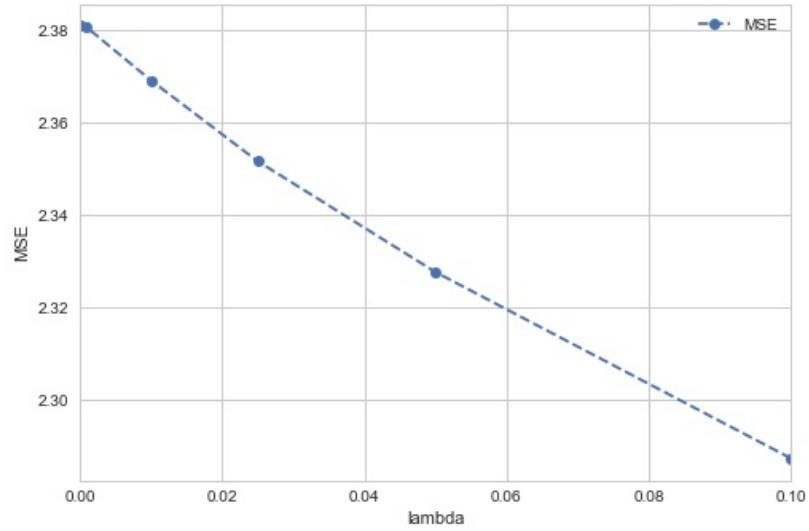
KNN



Mean Square Error as a function of the number of nearest neighbors (K) for KNN
Results are from 10-fold cross-validation

»The optimal k with minimal MSE was determined to be 11

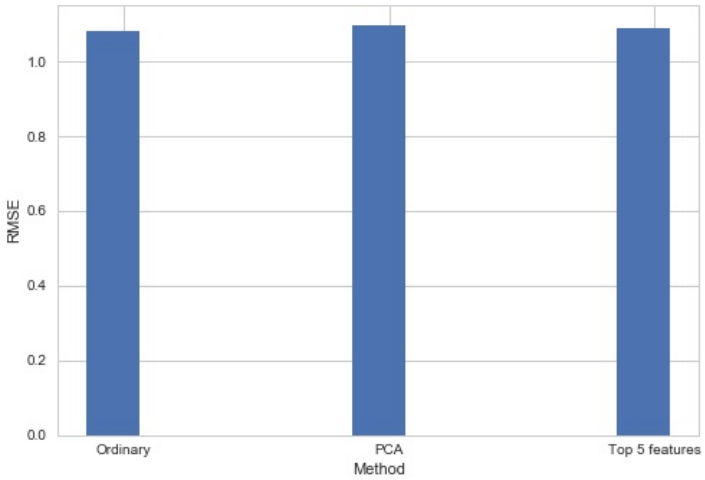
LASSO



Mean Square Error as a function of (λ) for LASSO
Results are from 10-fold cross-validation

»The larger λ, the smaller value of MSE. The optimal λ is determined to be 0.1

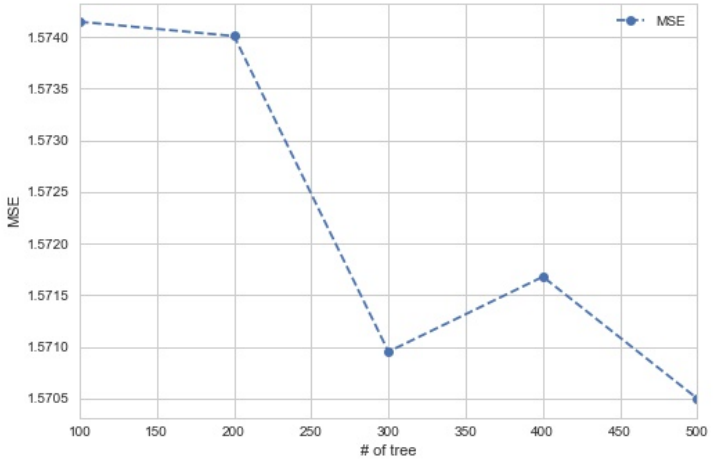
LASSO



Root Mean Square Error with different training set for LASSO

»No performance improvement of either methods

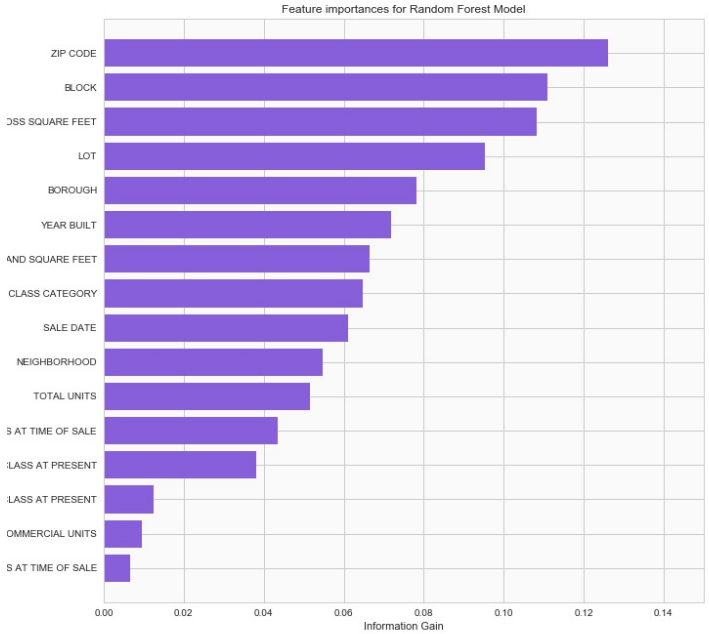
Random Forest



Mean Square Error as a function of the number of trees RF
Results are from 10-fold cross-validatio

»The optimal B was determined to be 500

Random Forest



Feature Importance Ranking by Information Gain for RF

»The location of the building and the total area of the property are the most significant factors to the sale price

Model Comparison

RMSE:	0.8695	0.7297	1.0806

» **The model with the minimal RMSE was selected as the best model**

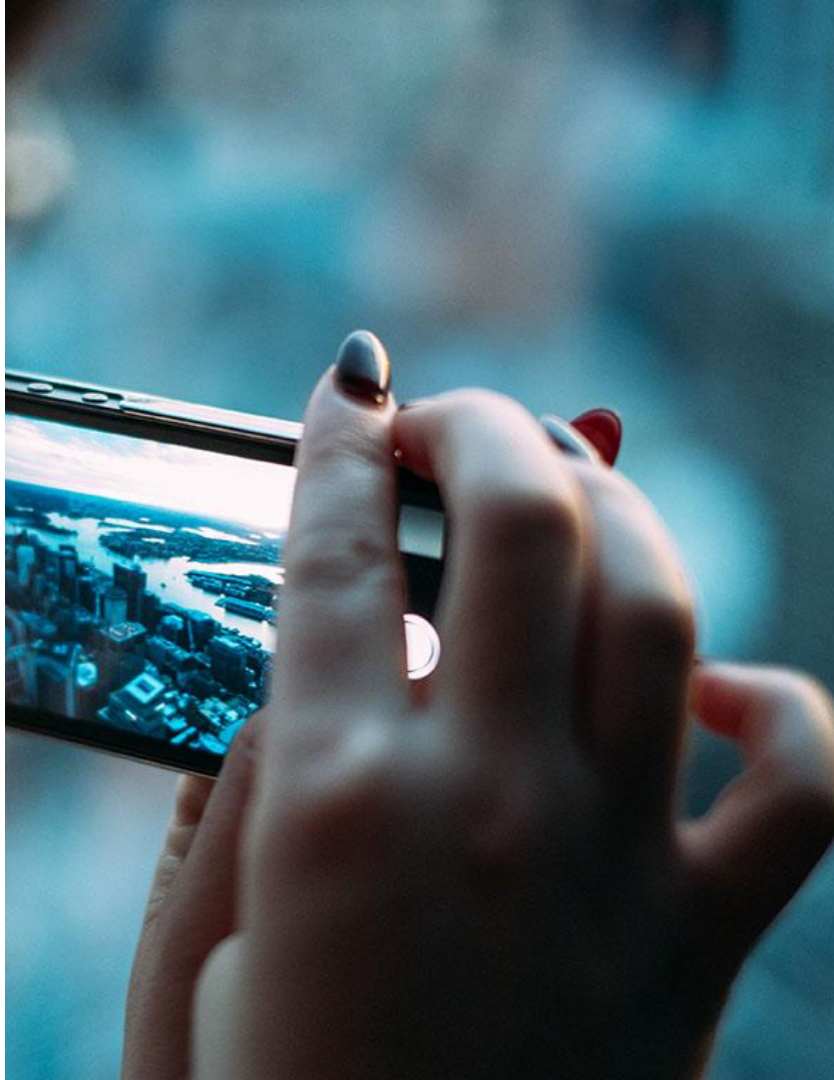
IV.

Conclusion

We apply three regression models to predict New York house sale price: LASSO, KNN, Random Forest Regression. The Random Forest Regression gives the best performance, with RMSE= 0.7297. We believe it is due to the nonlinearity interaction between the target and certain features, as well as the noise in our data set. Our result can serve as a reference for house sellers.

Future Work

- Exploring other models
- Better dataset
- Replace the sale price by asking price



THANKS!