Predicting Home Values Through Random Forest

BY: ALFREDO MARTINEZ

2017

Introduction

Goal

- Build a Predictive Model System for Home Prices

Client

- Keller Williams Realty, Inc.
- Other Real Estate Investment Companies

Exploratory Data Analysis(EDA)

Housing Dataset

Categorical(Qualitative) data:

- 23 nominal (There was no natural order, e.g. Type of house)
- 23 ordinal (Order do exist, e.g. Property condition; bad, fair, good excellent)

Numerical data:

- 14 discrete(Integers, e.g. Number of rooms)
- 20 continuous(Can take on any value, e.g. Square Feet)

Weather Dataset

- Subtracted from the National Center for Environmental Information Webpage.
- Includes daily, monthly and yearly records from different weather stations in the city.

Exploratory Data Analysis(EDA)

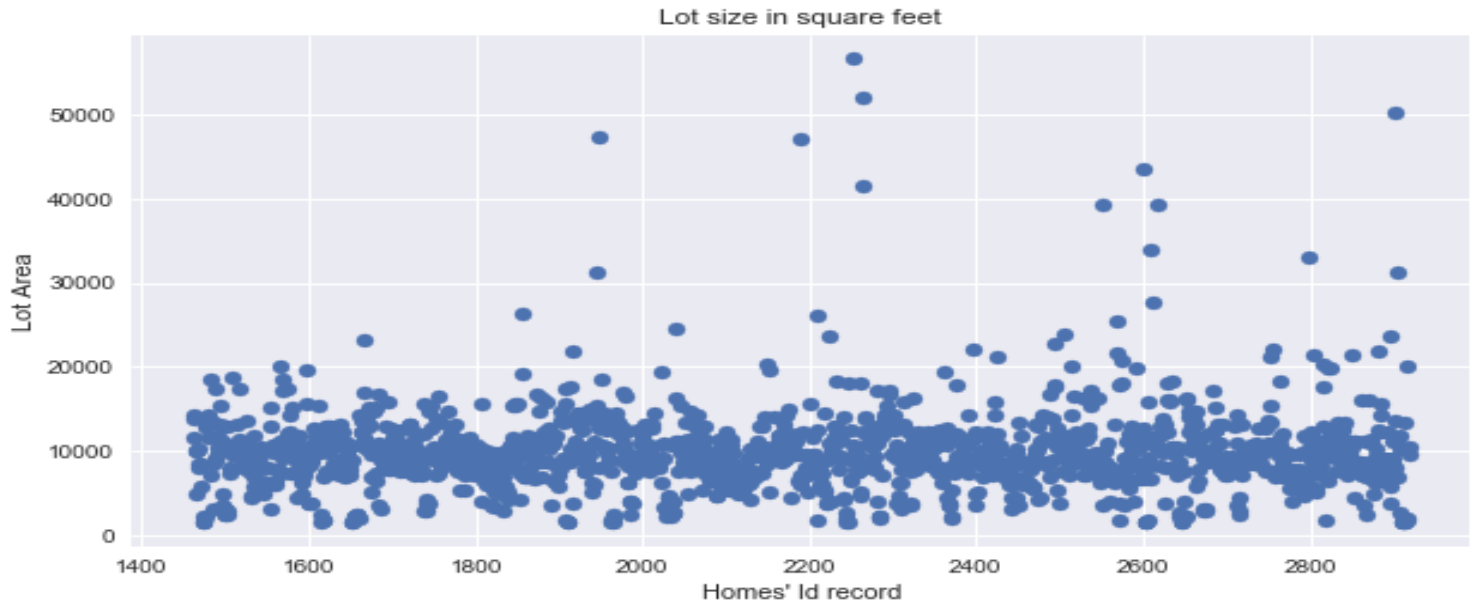
Bad Data and Outliers

Random Forest:

Different people, different arguments weather to keep all data or not

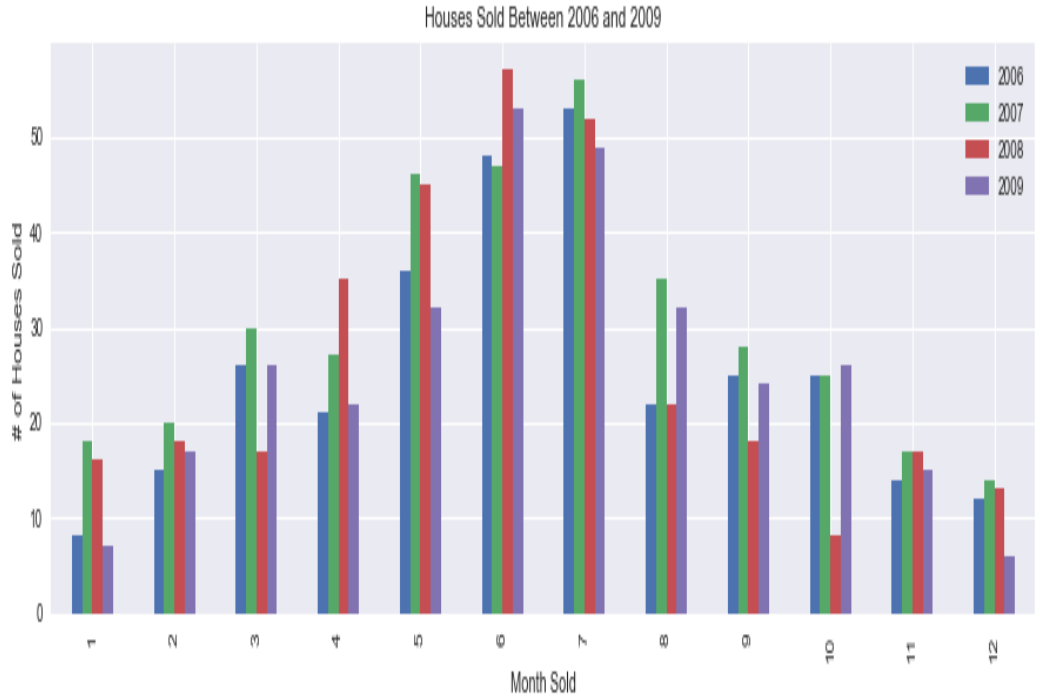
Ideas

- Not sensitive to outliers
- Sensitive

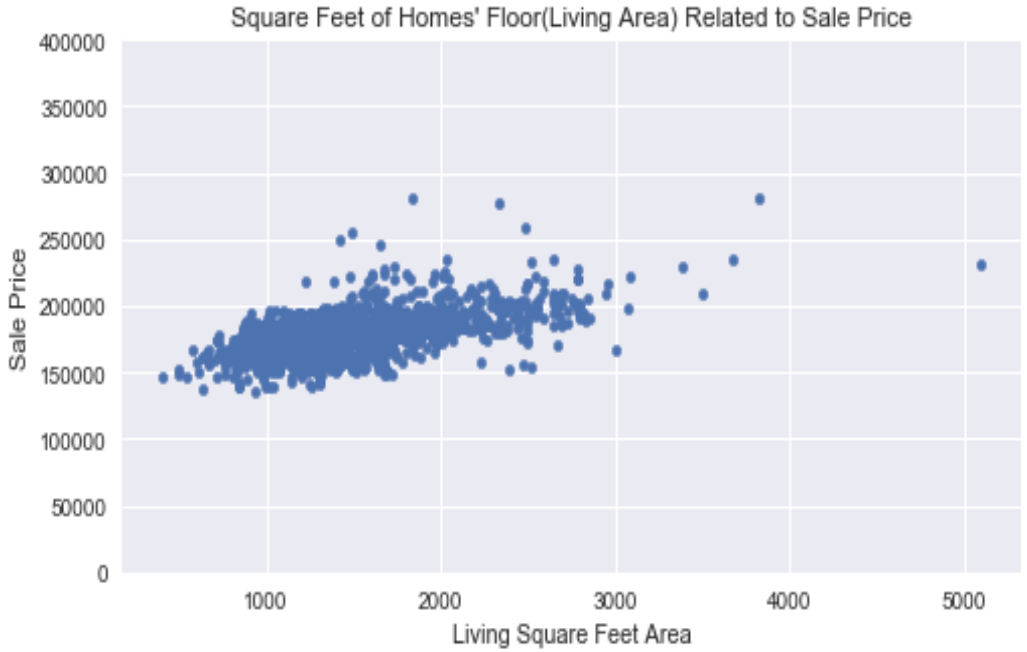


Exploratory Data Analysis

MONTHLY HOMES' SALES



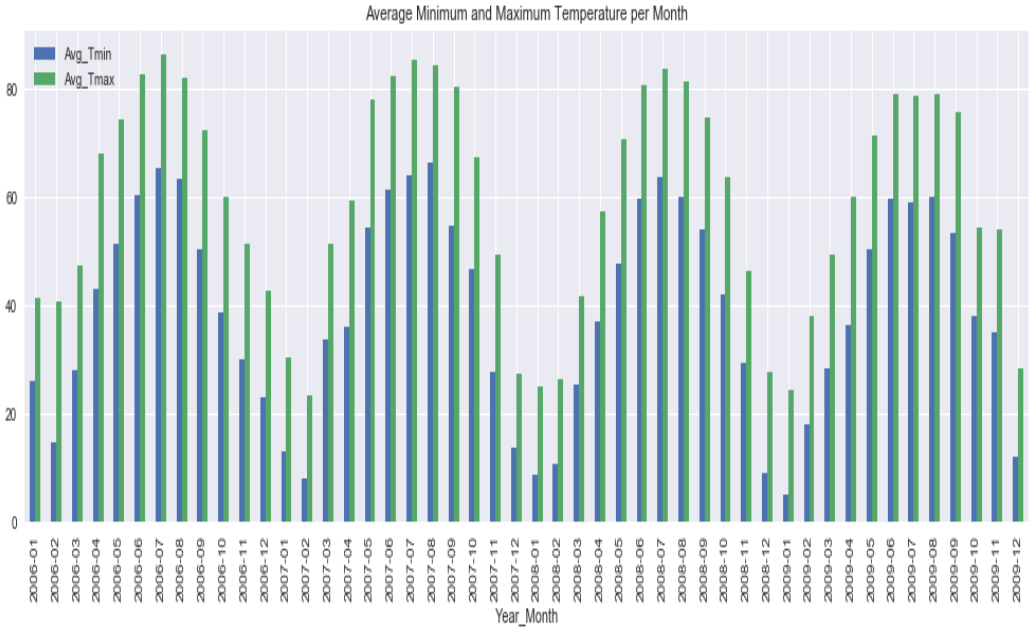
ABOVE GROUND LIVING AREA IN RELATIONSHIP TO SALE PRICE



Exploratory Data Analysis

HOME'S LOT AREA IN RELATIONSHIP TO PRICE

YEARLY WEATHER TEMPERATURE



Inferential Analysis

Alpha set at .05

Significant price difference analysis about population.

Central unit vs not a central unit

P-value: 0.0100906212171866 81 (**Significant**)

One story houses vs Two story houses

P-value: 3.5535259636604621e-13 (**Significant**)

Good privacy fence vs minimum privacy fence

P-value: 0.37003494954970051 (**Needs further research**)

Data Challenges

Missing Data

- N/A Values

Corrupted Data

- Present as (-999) values

Categorical to Numerical

- Transforming values to be able to use in predictive model(scikit-learn)

Fixed Values

- Mean, Median, Mode, and dummy variables indexing(from train to test)

Technology

Pandas

- To open and merge files
- To filter and drop cells
- To group by cells
- Etc..

Numpy

- To fill in Not a number values(NaN)
- Etc..

Datetime

- To fix and edit date times

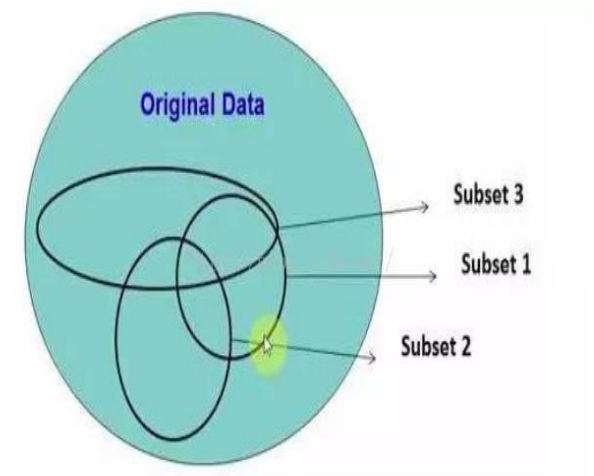
Scipy

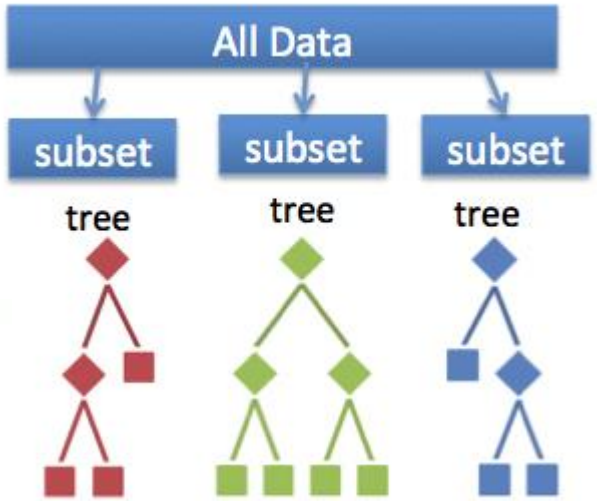
- T-test for inferential analysis

Modeling and Analysis

Random Forest Predictive Algorithm

- Continues Target variable(Regression)
- Based on Decision Trees Models
- Bootstrapping
- Root Mean Square Error Accuracy
- Feature Importance





Modeling and Analysis

Main packages and libraries

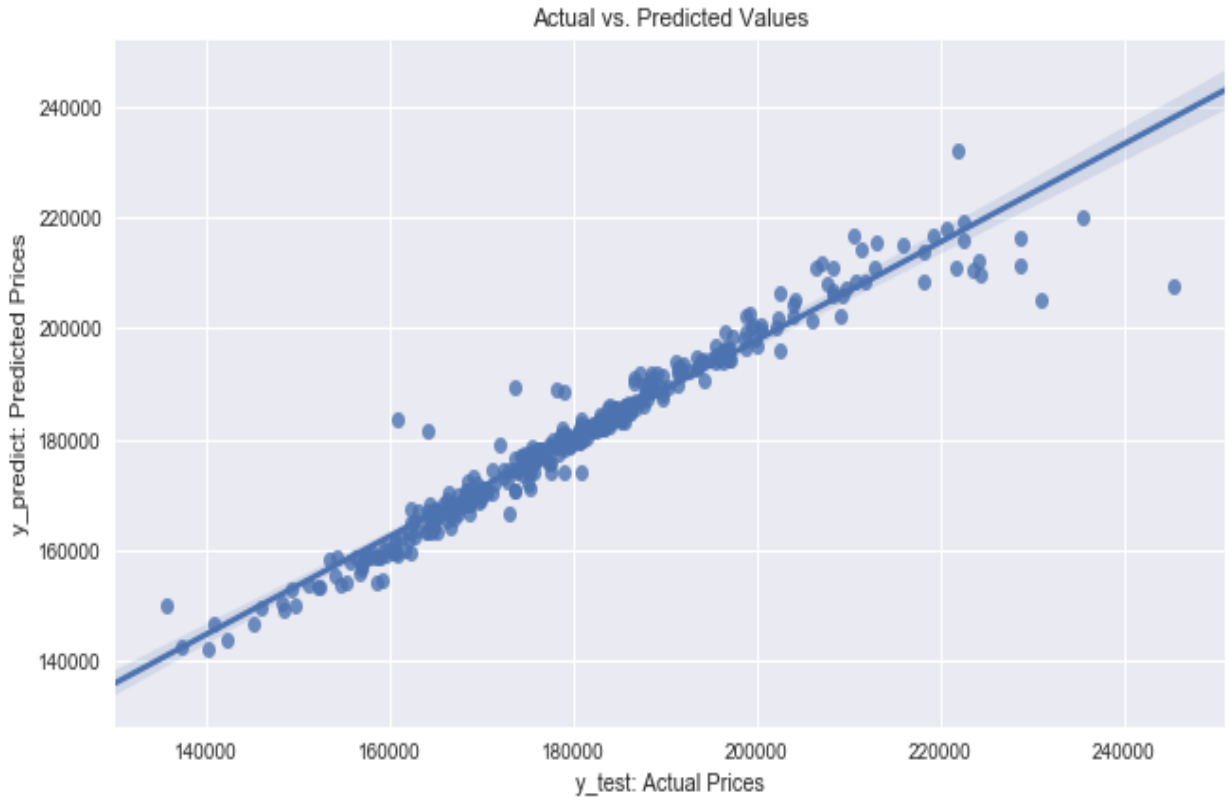
Scikit-learn and math

- RandomForestRegressor()
- Param_grid(GridSearchCV)
- Fit()
- Best_estimators()
- Predict()
- RMSE = sqrt(mean_squared_error(y_test, y_predict)

Results

Root Mean Square Error = $4,202

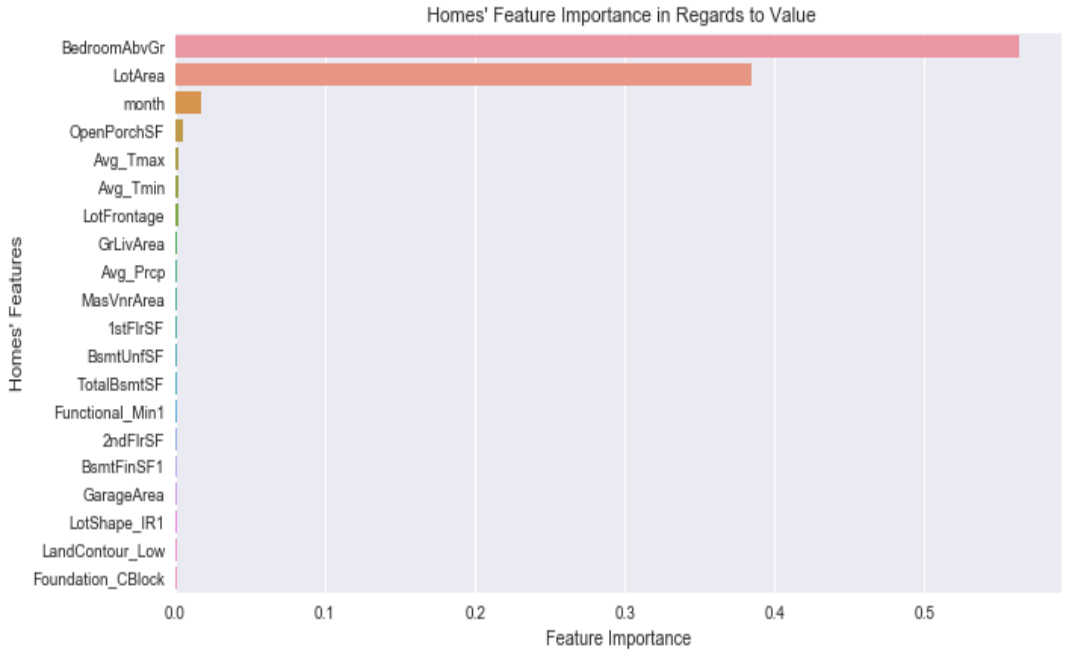
- Low Value
- Small variance
- Fitted line is close to data
- Small average distance from data points to line
- Strong predictive outcome



Modeling and Analysis

FEATURE IMPORTANCE IN RELATIONSHIP TO DETERMINE PRICE

	Features	Importance
20	BedroomAbvGr	5.67e-01
2	LotArea	3.81e-01
34	month	1.72e-02
38	Avg_Tmax	2.60e-03
28	OpenPorchSF	2.22e-03
39	Avg_Tmin	1.96e-03
1	LotFrontage	1.77e-03
12	1stFlrSF	1.55e-03
215	Functional_Min1	1.42e-03
36	Avg_Prcp	1.35e-03



Business Value

Sharper report analyses to help:

- Negotiation of sales
- Purchases
- Other strategic agreements related to real estate properties.

Thank you!

☺