HOUSE PRICE EXPLORATION

Muhammed Adil Yatkın

Institute of Computer Science , University of Tartu

Introduction

One way to examine the property value of houses is to investigate house sales based on some variables.

Especially, online real-estate companies offer house valuations using *Machine Learning* techniques. Usually, house prices change according to location, condition, view, square footage of living room, number of bathrooms/bedrooms, built year and other properties of the given house. In this project I tried to understand the dependence of prices between these determining factors

Objectives:

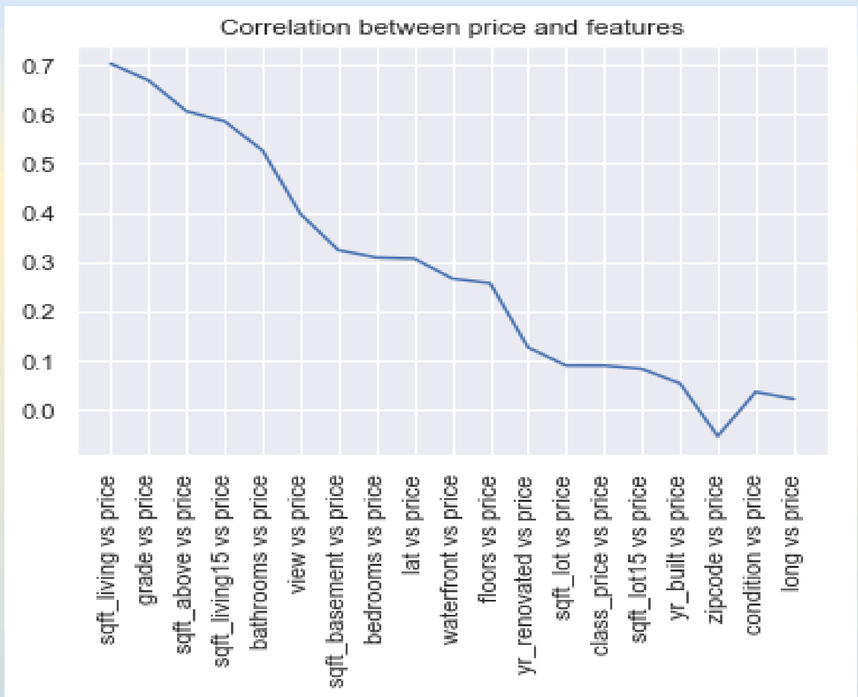
- Predict the price of house sales and find best macline learning model for predicting house prices
- Understand factors which play the most important role in predicting higher property value

Data Analysis

The dataset consists of house sales in King County, Seattle, Washington, USA, sold between May 2014 to May 2015.

In this dataset, there are 19 house features, plus the price and the id columns, along with 21613 observations. These features include the number of bedrooms/bathrooms, geographic coordinate, number of floors(level), view, condition, grade, built year etc.

I tried to understand which features play an important role in determining house prices. That's why, I checked the correlation between variables



My Approach

Firstly, I used simple Linear Regression model to predict house prices. However, predictions differed too much from original prices. This shows exact predictions is impossible in this dataset as prices are not homogenously distributed in price range.

Instead, I classified prices into 5 classes using K-Means clustering algorithm.

Price intervals

0: 78.000 – 412.500
1: 412.500-680.200
2: 680.200-1.150.000
3: 1.150.000-2.160.000
4: 2.160.000-7.700.000

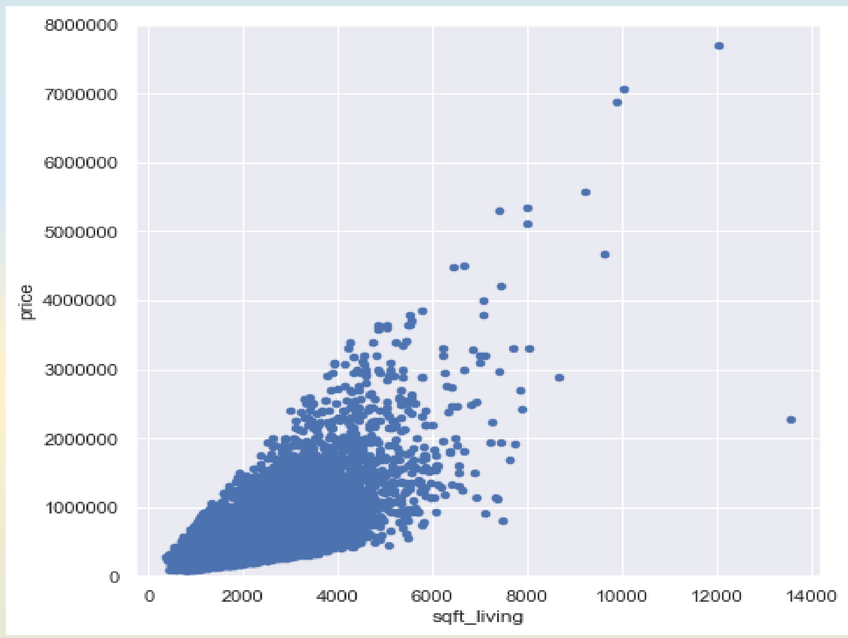
Next, I trained three classifiers - **K-Nearest Neigboors Classifier, Random Forest Classifier & AdaboostClassifier**; to predict the price classes of houses.

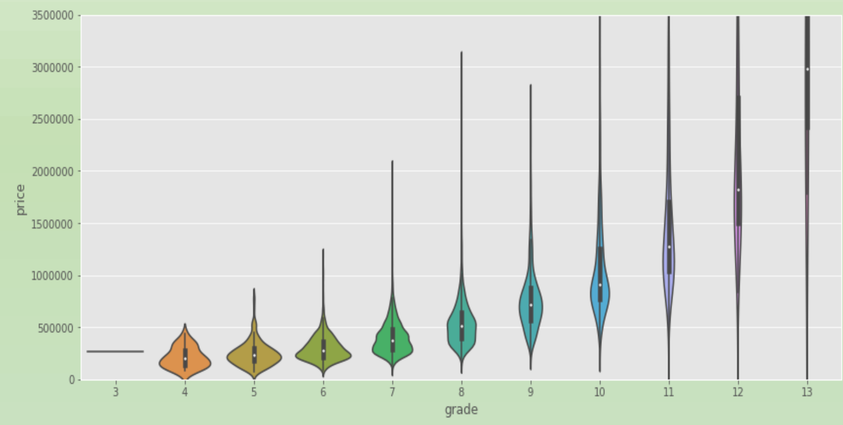
The best accuracy result (98%) is obtained using **Random Forest Classiffier**. After this experiment, we understood that at least we can predict house prices within given interval.

Explained variance score is better metric to rank models as it calculates how likely the predicted value within variance.

I cleaned less important features according to correlation matrix in order to let it converge quickly.

Square Footage of living and Price



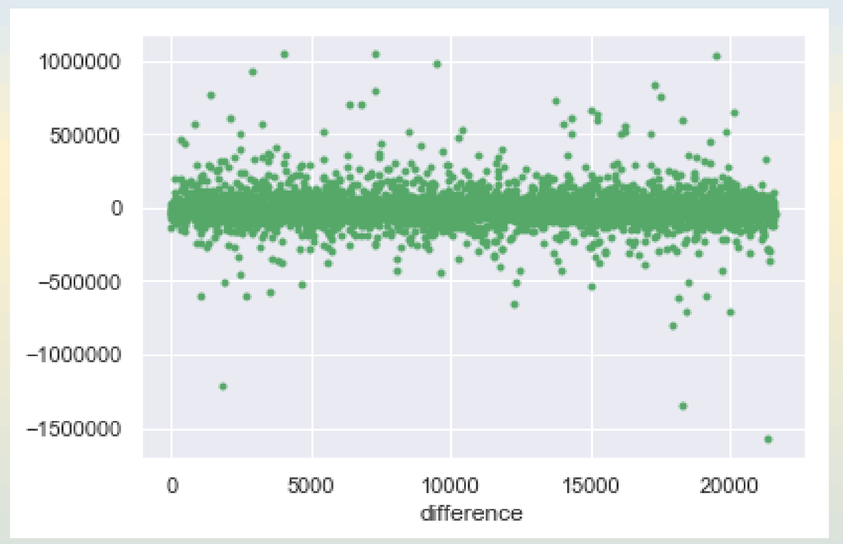


Now, I splitted whole dataset as %20 test data and %80 train data by using cross validation split and for feature importances I used 12 most important features and 3 regression techniques as RadomForestRegressor , GradientBoostingRegressor , AdaBoostRegressor ,
I achieved these explained variance scores:

Model	Variance Score
GradientBoosting	0.876718
RandomForest	0.866466
DecisionTrees	0.775933
Adaboost	0.561331

Results

So, I noticed that score of **Gradient Boosting Regressor** is nearly 87.67% and also achieved decent variance score of 0.86 which is close to 1 . Therefore, it is inferred that Gradient Boosting Regressor is the most suitable model for this dataset and when we compared difference betweeen our prediction and true prices:



Conclusion

So we can see that if we know the some important features values of houses like square footage of living , condition , number of bathrooms/bedrooms , location etc. We can make a good prediction for its expected prices or property value.

Actually our dataset was not too big to make a good predictions with regression models if it was more large we could have much better results

UNIVERSITY OF TARTU
1632