HOUSING PRICES PREDICTION PROBLEM

(Real estate data from King County, Seattle, USA)

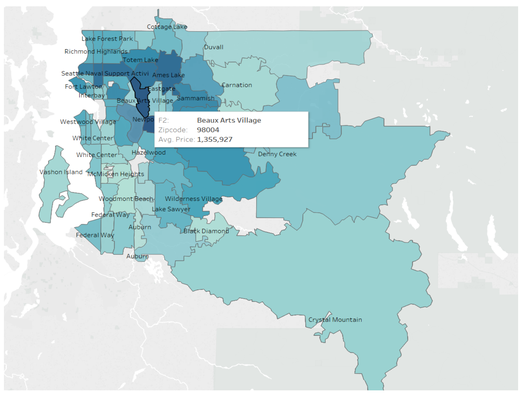
Sai Yesaswy Mylavarapu, Hemanth Kumar Koraboina, Sainandan Tummalapalli
University of North Carolina at Charlotte

INTRODUCTION

Prediction of housing prices is an important problem since the ownership of the houses keeps on changing, it would be helpful for the sellers and buyers to be able predict the approximate value/cost of the house. The ideal goal of this project is to predict the housing prices given various factors affecting the price.

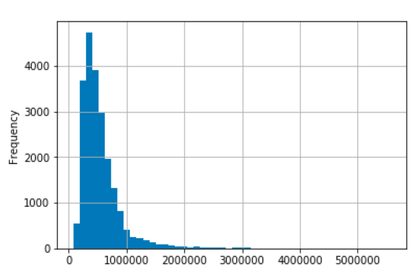
DATASET

- ✓ Real estate data from King County, Seattle for the years 2014 and 2015 is used as dataset.
- ✓ Dataset has 21613 instances and 21 variables.
- ✓ Every house has set of variables including price, bedrooms, bathrooms, sqft living, view, grade, year_renovated, lat and long.
- ✓ Reference: Similar study was done by Victor Gan, Vaishali Agarwal and Ben Kim in the paper 'Data mining Analysis and predictions of Real estate prices'.



EXPLORATION

- ✓ Statistical Analysis by calculating Mean, Standard deviation, quartiles, Min and max values for each variable.
- ✓ Univariate analysis of independent variables by plotting histograms and box plots.
- ✓ Bivariate analysis between input variables and target variable using Seaborn.
- ✓ Correlation matrix between the variables



DATA WRANGLING

- ✓ Checking for outliers and missing/NA/NAN values in the dataset – No missing values in the data.

FEATURE ENGINEERING

VARIABLE	ORIGINAL	MODIFIED
AGE OF HOUSE	'DATE' attribute is in unusable format **Eg:** '20141013T000000'	'Year_sold' is calculated from 'Date' and 'Age of house' is calculated by difference between 'Year_sold' and 'Year_built'
IS RENOVATED	'Year of Renovation' is the year in which the house was previously renovated.	Encoded it in such a way that if a house is renovated we consider it as 1 else 0.
HANDLING ZIPCODES	'zipcode' attribute has 70 unique values or categories.	Converted them into dummies by creating 70 attributes for all unique values, such that every attribute will contain binary values 0 or 1.

DIMENSION REDUCTION

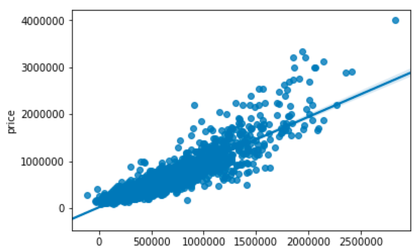
- ✓ Feature engineering step resulted in a total of 93 variables. So we applied Principal Component Analysis (PCA) in order to reduce no of attributes by capturing maximum variance in the data.

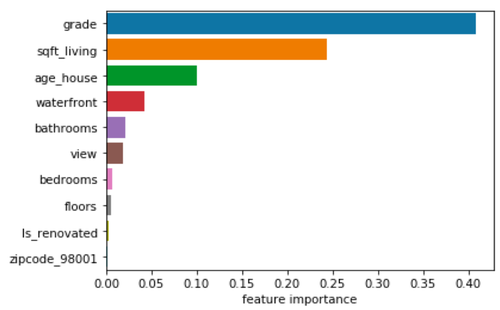
BUILDING REGRESSION MODELS

LINEAR REGRESSION

- ✓ Linear regression finds linear relationship between the target variable (PRICE) and predictors (INPUT variables). The best fit line is the one for which total prediction error (Mean Squared Error) is as small as possible.

R-squared value: 0.814





RANDOM FOREST

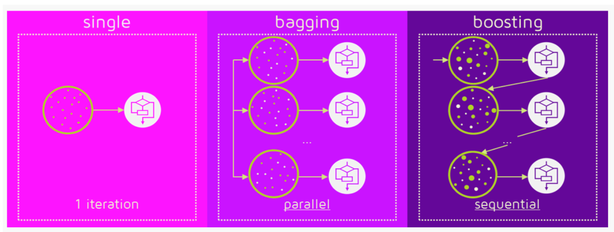
- ✓ Random forest is a bagging technique in which an ensemble of trees are built with a random subset of features.
- ✓ It is also useful in measuring relative importance of each feature in the model as shown in the figure.
- ✓ Grade has the highest importance in building the model followed by sqft_living and age of house.

R-squared value: 0.811

GRADIENT BOOSTING REGRESSOR

- ✓ Boosting always aims to improve the regular set of algorithms by giving more weightage to the weakly predicted instances.
- ✓ Loss function (MSE or RMSE) is minimized using Gradient descent principle and predictions are updated based on the learning rate.
- ✓ Feature importance graph is drawn and performance is calculated using K-fold Cross-validation.

CV Score : Mean - 0.8566363
R-squared value: 0.86

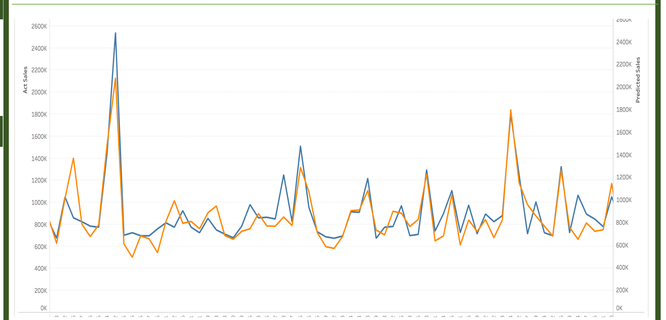


Libraries:
Scikit-learn; GridsearchCV ; XGBoost
Pandas; Numpy; Matplotlib; Seaborn;

RESULTS

- ✓ Performance of every model is evaluated by calculating Root Mean Square Error (RMSE) and R-squared values.
- ✓ Gradient boosting model gives the best performance after tuning its hyperparameters which involves learning rate, max depth, num of estimators, samples at each split etc.

Model	RMSE	R-Squared
Linear Regression	149318.4256	0.8149
Random forest	150062.3342	0.8118
XG Boost	145747.5758	0.8236
Gradient Boosting	129708.0422	0.8603



Predicted prices Vs Original prices
Learning rate - 0.1, max_depth=15, n_estimators = 100, min_samples=30, max_features=10, alpha = 0.9

CONCLUSION & FUTURE SCOPE

- Even though the dataset is limited to King County, this process can be generalized to any location.
- Future scope involves building a data science pipeline which accepts common set of parameters like rooms, grade, area and builds a model.
- Further advancements can be made by building a web or mobile based application which can accept input range from users including location and predicts price range for a house.

REFERENCES

- ✓ Victor Gan, Vaishali Agarwal, Ben Kim (2015). Data mining Analysis and predictions of Real estate prices.
- ✓ Zillow King County (2015), http://www.zillow.com/king-county-wa/home-values
- ✓ House Sales in King County, USA. – Kaggle dataset

CONTACT

- ✓ smylava1@uncc.edu
- ✓ hkoraboi@uncc.edu
- ✓ stummal3@uncc.edu