HOUSING PRICE PREDICTION

ABSTRACT

People are careful when they are trying to buy a new house with their budgets and market strategies. Real estate is the least transparent industry in our ecosystem. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Predicting housing prices with real factors is the main crux of our research project. Here we aim to make our evaluations based on every basic parameter that is considered while determining the price. We use various regression techniques in this pathway, and our results are not sole determination of one technique rather it is the weighted mean of various techniques to give most accurate results.

Introduction:

Residential Property Prices (RPPs) – also named House price indices (HPIs), are index numbers that measure the prices of residential properties over time. RPPIs are key statistics not only for citizens and households across the world, but also for economic and monetary policy makers. They can help, for example, to monitor potential macroeconomic imbalances and the risk exposure of the household and financial sectors. Hence, housing price ranges are of great interest for both buyers and sellers.

In this project, house prices will be predicted based on the given explanatory variables that cover many aspects of residential houses. The goal of this project is to create a regression model that is able to accurately estimate the price of the house given the features. We empirically study how the various factors influence the house prices. Our regression analysis revealed the best fit model to predict the price of the house.

Problem Statement:

The problem requires us to predict the price of a house, in a particular location, based on the different features of the house. This will help the buyers to make accurate decisions, based on the pricing trends in the neighbourhood. It will also help sellers set a realistic target for the sale that they anticipate.

Objective:

To analyse all the features provided in the data set and arrive at a predictive solution for the price of a house.

Roadmap:

We will follow the below roadmap, to arrive at a model, to predict prices for new data.

1. Analyse Data: Use descriptive statistics and visualization to better understand the data available.
2. Prepare Data: Use data transforms in order to better expose the structure of the prediction problem to modelling algorithms
3. Evaluate Algorithms: Design a test harness to evaluate a number of standard algorithms on the data and select the top few to investigate further.
4. Improve Results: Use algorithm tuning and ensemble methods to get the most out of well-performing algorithms on the data.
5. Present Results.

1. Analyse data:

- The data for this project is provided to us in the form of a .csv file.

- Preliminary analysis of the data indicates that this data is from the house sales from May 1st 2014 to May 30th 2015 in Kings County, USA.
- This data set has 21613 rows and 23 columns.

Data type observations:

Column Name	Description	Observation
cid	Unique value for each record	▪ Categorical ▪ Qualitative ▪ Nominal variable
Dayhours	Date of sale	▪ Date type object
Price	Target Variable	▪ Numerical ▪ Continuous ▪ Quantitative
room_bed	Total number of bedrooms	▪ Numerical ▪ Discrete
room_bath	Total number of bathrooms	▪ Numerical ▪ Discrete
living_measure	Total area of the house	▪ Numerical ▪ Continuous ▪ Quantitative
lot_measure	Total area of the lot	▪ Numerical ▪ Continuous ▪ Quantitative
Ceil	Number of levels in the house	▪ Numerical ▪ Discrete
Coast	Water front view	▪ Categorical (binary) ▪ Qualitative ▪ Nominal variable
Sight	Number of times viewed	▪ Numerical ▪ Discrete
Condition	Rating of the house condition	▪ Categorical ▪ Qualitative ▪ Ordinal
Quality	Rating based on the house quality	▪ Categorical ▪ Qualitative ▪ Ordinal
ceil_measure	Total area of the house excluding the basement	▪ Numerical ▪ Continuous ▪ Quantitative
basement_measure	Total are of the basement	▪ Numerical ▪ Continuous ▪ Quantitative
yr_built	The year the house was built	▪ Date type object
yr_renovated	The latest year of renovation	▪ Date type object
Zipcode	Zipcode	▪ Categorical ▪ Qualitative ▪ Nominal.

Lat	Latitude coordinate	
Long	Longitude coordinate	
living_measure15	Living room area in 2015	▪ Numerical ▪ Continuous ▪ Quantitative
lot_measure15	lotSizearea in 2015	▪ Numerical ▪ Continuous ▪ Quantitative
Furnished	Is the house furnished	▪ Categorical(binary) ▪ Qualitative ▪ Nominal.
total_area	Living area+Lot area	▪ Numerical ▪ Continuous ▪ Quantitative

Observations:

- No null values in the given data set.
- There are 356 duplicate records found in 'cid' column. Hence, we can safely assume that this column represents the customer id and the duplicates are indicative of customers with different houses.

Data Visualizations:

Univariate Analysis:

Descriptive statistics:

Column Name	**Range**	**Observation**
Price	Price ranges from 75000 to7700000	Since the mean is greater than median, we can say that the distribution is right positively skewed
room_bed	Number of bedrooms range from 0 to 33.	75% of data values are around 4. This either means the data is highly skewed towards right or there could be outliers.
room_bath	Number of bathrooms range from 0 to 8.	75% of data values are around 2.5 whereas the maximum value is 8.The data is highly skewed towards right.
living_measure	Range is from 918 to 13540	
lot_measure	Range is from 15106 to 165135.	Mean is closer to the 75% range. The data is highly skewed towards right.
Ceil	The values range from 1 to 3.5	Mean is equal to median. Normal distribution.
Sight	Has been viewed from 0 to 4 times.	Houses till the third quartile have been viewed 0 times. Data is right skewed
ceil_measure	square footage of house ranges from 290 to 9140	
basement_measure	square footage of the basement 0 to 4820	Highly skewed data.

yr_built	Values range from 1900 to 2015	Data follows normal distribution, with mean closer to median.
yr_renovated	Range is from 0 - 2015.	Zero indicates that the house is not renovated. Highly skewed data, indicating that the houses in the first two quartiles are not renovated.
living_measure15	Range is from 399 -6210	Close to normal distribution.
lot_measure15	Range is from 651 -871200.	
total_area	Range is from 1423 - 1652659	Data is highly left skewed.

Data Visualization:

UnivariatePlots: We started with plotting our target variable and made the following observations.

- Price is right skewed, with most concentration between 0 - 1000000.
- Skewness: 4.021716
- Kurtosis: 34.522444
- The values are quite significant and the data needs to be standardized.

- We then proceeded to plot all the numerical variables and categorical variables, made a note of the distribution type and the presence of outliers, where applicable

Bi-variate Analysis:

- We used joint plots to visualize bi-variate distribution for numeric variables, and further analysed the associations using the pair grid function.
- We used point biserial analysis for categorical variables.
- The associations with ordinal variables are assessed through box-plots.
- We have used spearman coefficient and person coefficient to arrive at the correlation values.
- We have also performed geospatial analysis to understand the data distribution with respect to latitude and longitude.

3. **Prepare Data**

Derived columns:

- Creating a column house_age that is derived from yr_built.
- Created a binary is_renovated column.
- Created a binary basement_present column.
- Coast, condition, quality, furnished, sight needs to be encoded.
- Decomposed dayhours column into yr_sold, month_sold, time_sold.

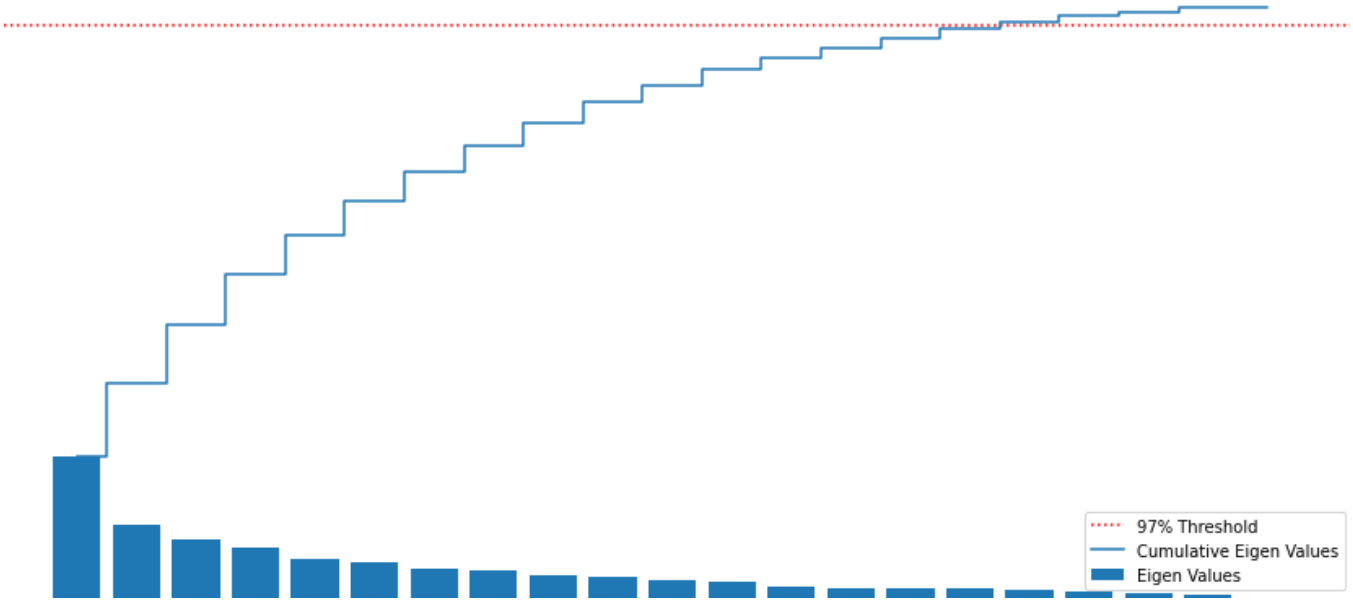
Data Cleaning:

- Drop the yr_built column.
- Drop the yr_renovated column.
- Dropping cid, dayhours, time_sold as they are irrelevant for our analysis and redundant.

Data Transforms:

- The differing scales of the raw data may impact the algorithms.
- Part of a requirement for a standardised data set is to have each attribute have a mean value of zero and a standard deviation of 1.
- Implement standardisation using pipelines.
- Use cross-validation to validate performance of algorithms.

From below Cumulative Eigen Plot we concluded that there are 16 features which explains more than 97% of variance cumulatively in the dataset



4. Evaluate Algorithms

a) Split-out validation dataset:

- We start by splitting out data into train test and a leave-out validation set, to test the efficiency later on.

b) Test options and evaluation metric:

Regression Evaluation Metrics:

Here are three common evaluation metrics for regression problems:

- Mean Absolute Error (MAE) is the mean of the absolute value of the errors
- Mean Squared Error (MSE) is the mean of the squared errors
- Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors

Comparing these metrics:

- MAE is the easiest to understand, because it's the average error.

- MSE is more popular than MAE, because MSE "punishes" larger errors, which tends to be useful in the real world.
- RMSE is even more popular than MSE, because RMSE is interpretable in the "y" units. All of these are loss functions. Hence, we need to minimize them.

c) Spot Check Algorithms:

- We have considered the below algorithms to do a spot check and have arrived at the corresponding metrics

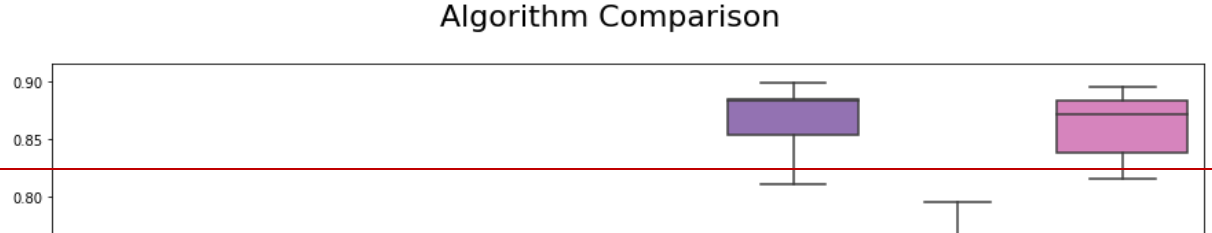
Model	**MAE**	**MSE**	**RMSE**	**R Squared**	**Cross validation**	**Train Accuracy**	**Test accuracy**
Linear Regression	127063.85	44737018177.73	52.33	0.71	0.70	70.06	70.96
Ridge Regression	127039.51	44742906985.13	52.33	0.71	0.70	63.09	63.82
Lasso Regression	127063.18	44737394309.75	58.41	0.71	0.70	20.80	20.90
Elastic Net Regression	146786.49	56768070697.46	32.23	0.63	0.62	98.28	88.98
Random Forest Regressor	146786.49	56768070697.46	32.32	0.63	0.88	98.28	88.90
Decision Tree Regressor	102822.80	42794034747.19	49.33	0.72	0.62	77.86	74.19
Gradient Boosting Regressor	102822.80	42794034747.19	33.55	0.72	0.62	90.12	88.06

d) Compare Algorithms:

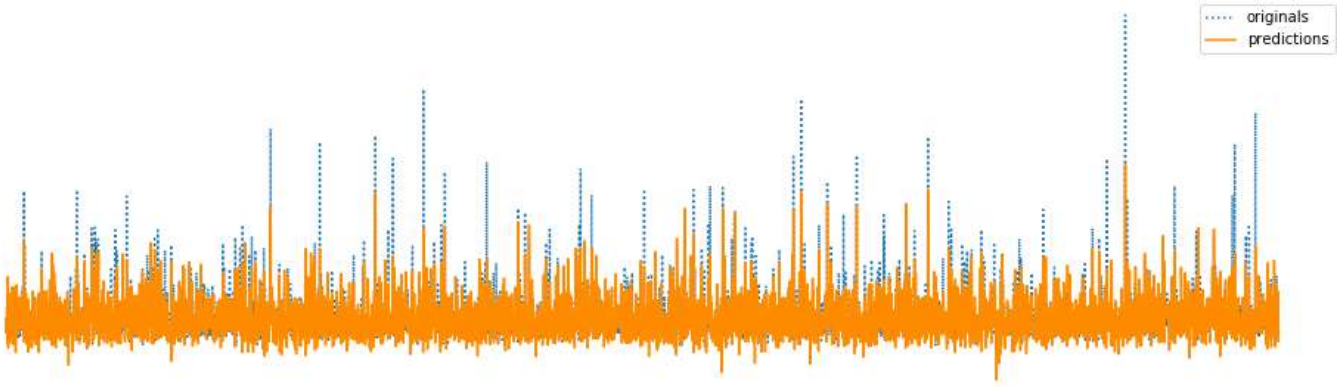
- We have arrived at the following results, evaluating the algorithms, after using standard scalar.

Scaled LR: 0.694831 (0.014187)
Scaled RR: 0.694898 (0.014206)
Scaled LASSO: 0.694897 (0.014208)
ScaledEN: 0.668394 (0.014665)
ScaledRF: 0.869880 (0.025525)
ScaledDT: 0.719429 (0.066930)
ScaledGBM: 0.861635 (0.026955)

- Gradient Boost and Random Forest seems to be performing better.



ORIGINALS VS PREDICTION



6. Improve Accuracy:

Feature Engineering:

- Feature selection is performed on univariate statistical tests, by using f-test, and by applying the decision tree regressor for feature selection.
- We also performed RFE on the best performing models to get model based feature selection.
- RFE on GradientBoostingRegressor returns top 18 features that are important(features with score 1)
- After doing feature selection we have list of features to be dropped from the dataset. We will check the model performance, after dropping the features.

a) Algorithm Tuning
A grid search algorithm measured by cross-validation on the training set and evaluation on a held-out validation set is used.

b) Ensembles
Gradient Boosting ensemble algorithm is used with the following hyper parameters

- Number of Trees
- Number of Samples
- Number of Features
- Learning Rate
- Tree Depth

7. **Finalize Model:**

Final model is based on the following values:

Model: Gradient Boosting Regressor

Training score : 0.89

Testing score : 0.87

Root Mean Squared Error (RMSE): 0.3426

Conclusion:

Gradient Boosting ensemble technique stood pretty well in predicting the house prices. The model that is fined tuned with hyper parameter tuning is able to predict with 85.9% to 91.2% accuracy with 95% confidence level.

Limitations:

The accuracy of the model is constrained by the limited features available. We could have been able to better predict the housing price, had we more features like

- Proximity to schools and business centres which explains the motivation behind the sale.
- Quality or ratings for the schools in the area.
- Age of the home buyers, which in turn dictates the functionality of the house.
- Number of houses available in the area, at the time of sale, which is indicative of the demand and supply in the area
- Interest rate during the year of sale, which determines the cost of variable mortgage payments
- proximity to parks and playgrounds, noise pollution, light pollution, crime, zoning laws, air quality, internet connection quality, traffic volume, road quality

Also, this data is highly specific to King County in USA and is therefore not indicative of the market and economy elsewhere. This model is not generalized to be applied to other markets.