Assignment Report

The task was based on Zillow Prize Challenge on Kaggle, revolving around predicting the price that a particular real estate will sell for. Here are some of the interesting part of the challenge:

- We needed to predict the log error of the house prediction, since zillow has build a good model already, it's quite difficult to improve.

- Though dataset is coming from a well known source and has been previously used for modeling by kaggle, it contains missing values, outliers and redundant values which needs cleaning before processing.

- We are given data set of 3M houses and each house has approx 50 features.

Model that I tried on the data set after cleaning process.

- OLS linear regression
- Support Vector Machine Regression
- Decision Tree Regression
- K nearest neighbour model

My Favorite Model

As we can see that, we tried out different machine learning models to see how well they predict in terms of log error and after submitting scores to kaggle I found out that linear regression works best than other complex models.

How Linear Regression Works:

1. Linear regression model basically tries to fit the given features in a line, and predicts the value of a dependent variable by using one or more predictor variables.

2. A linear equation of the form Y = mX + c is defined between the two variables, where Y = dependent variable and X= list of predictor variables. The model tries to plot the line on observed data.

3. In the regression equation, y is always the dependent variable and x is always the independent variable(s).

4. It’s good to determine the correlation between features before fitting them in linear regression model, if the variable doesn’t have correlation between them then regression model fails.

5. Since all of the data point doesn’t fit on the line for real world data, we calculate two major stats for model, R-squared value and Mean Squared Error.

6. R-squared value tells us about how well our model is able to explain the value of Y in terms of x, Higher the R2 value, better our model would be.

7. MSE (mean square error) shows the difference between predicted value and actual value, it is average of the square of the deviation. Lower the MSE better the model would be.

8. Since the real world data doesn't perfectly fit on a line, we are very likely to find that for some of the data points model perform bad than for other data points.

An evaluation of how well Linear Regression model works.

- R-squared value for our linear regression model is very low (0.006), which can be explained by the fact that zillow is already a good model and its very hard to improve on it.

- By choosing features bedroomcnt, calculatedfinishedsquarefeet, garagetotalsqft and fitting them in a linear model we found out that since the value of dependent variable(log error) is very low, higher changes in the selected features won’t affect the logerror value much.

- The linear regression model outperformed other complex models such as SVR, K nearest neighbour, decision tree model.

- By choosing mentioned features and fitting them on ols linear regression model, the submission achieved score of **0.0649067** and rank of **2033** on Kaggle.

Interesting experiences or surprises

1. By working on this data set I realized that real world data is not perfect and no matter how reputed the data source is, we always need to clean data and need to make sensible decisions in order to make data more useful for processing.

2. The concept of starting off by simple things which we understand better and then move to complex part was very helpful, in the start playing with the dataset was overwhelming but over the time my understanding for the data got better.

3. Cleaning the dataset is one of the major part of data processing. understanding the data and finding out what are the things which needs to be cleaned and which part of the data we don't need is crucial.

4. Data cleaning directly affect the performance of our model, clean and relevant data is very important for any model to perform better.

5. Its also important to find good features which explains the predictor variable better, model performs better if we choose appropriate features.

6. It's always good to test our model with testing data, by this way we can avoid overfitting and underfitting.

7. Necessarily the complex model won't give better result, as we can observe in our case, linear regression model outperformed all other complex models. So the model performance largely depends upon the data itself.