MID-PROGRESS REPORT

HOUSING PRICES PREDICTION PROBLEM

PROBLEM DESCRIPTION:

Prediction of housing prices is an important problem to prospective buyers and sellers. 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. Also, as large amount of related data is available online, house prediction models are gaining importance in academic and business circles. The goal of this project is to predict the housing prices given various factors affecting the price.

We are using the real estate data of housing prices in King County, Seattle, USA, from Kaggle. Zillow is a real estate and rental marketplace helping buyers and sellers to find their ideal home. Zillow has predicted the prices of the houses in King County from 2008 till 2018. The trend has been varying through the time and hence it would be interesting to build few models to analyze this trend. The dataset we are using has the transactions recorded from May 2014 to May 2015.

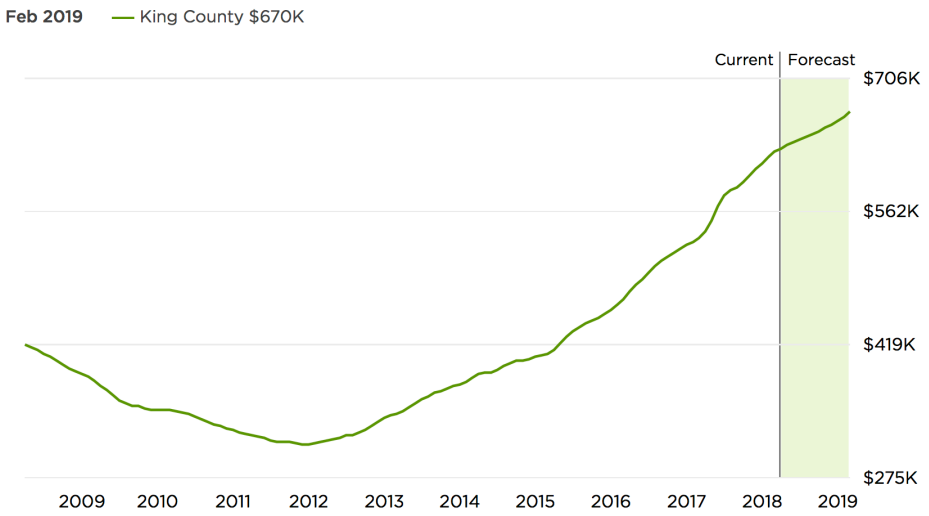

Fig 1: King County Market Overview (Source: Zillow)

PREVIOUS WORK AND IMPROVEMENTS:

Similar study was done by Victor Gan, Vaishali Agarwal and Ben Kim in the paper '**Data mining Analysis and predictions of Real estate prices**'. The dataset used by them was collected in the year 2012 and 2013. This paper includes building models on this dataset using two different techniques: Decision trees and Neural networks. While the baseline model according to this paper predicts the Mean Absolute error as 102,968, they have divided the overall dataset into categories based on the house prices. So, scores

were obtained for all the 10 categories where the minimum Mean Absolute error was 60,757 and the maximum was 343,380.
But instead, we considered the entire dataset as a single entity and split it into train and test in 80% and 20% ratio respectively.

We are planning to implement the following different techniques to improve the performance of the model:

1) Feature Engineering:
Although the paper has discussed about the various cleaning techniques like removal of outliers and others, it has not dealt with creating new set of features or presenting the old features in useful form. So, we will try to implement some feature engineering techniques with respect to this dataset. Some of them are:

- 'Date' attribute contains the date of home sale in '20141013T000000' format. This is not a useful version of the date, so instead we take only year from the given date format. A new attribute 'Year_sold' is created which signifies the year in which house was sold.
- We have two date related variables now, 'Year_sold' and 'Year_built'. As the year doesn't signify anything in the Machine learning model, we have to convert it to an accessible format. So we can calculate the age of the house by the difference between the 'Year_sold' and 'Year_built'. Hence, 'age_house' attribute is created.
- In the 'zipcode' attribute, there are 70 unique values or categories. Zipcodes in their original form doesn't make any sense, so we want to convert them into dummies by creating 70 attributes, one for each of them. Each such attribute will contain binary values 0 or 1.

These are some of the feature engineering techniques we thought as of now, we will add more as we complete the project.

2) Ensemble modeling:
As discussed previously, the paper has dealt with only decision trees and neural networks. We believe there are other models with more potential than these problems. Hence, we will try to implement bagging and boosting techniques. The two main types of ensemble modeling techniques are Bagging and boosting.

In case of bagging, as the Decision trees may underperform for a regression problem. Hence, we need an alternative or an improvement to our decision trees, which can be Random forest. Random forest can be an ideal solution for both regression and classification problems. In order to further improve our model, we will implement XG Boosting and Gradient boosting techniques. Also, hyper-parameters can be optimized to get the best performance of the model.

3) Dimension reduction:
Along with feature engineering, we are also trying to implement dimension reduction in our project. As we can see, by converting the categorical features into dummy and making any other necessary changes, the count of total number of variables may go up to 95 or above. Building a model with 95 attributes can be difficult sometimes, so we apply some basic to advanced dimension reduction techniques. They are:

- Attributes like Id and date are not useful in their original formats. So they are dropped from the dataset.
- Some attributes have huge correlation with other input variables. For example, sqft_living has correlation with sqft_lot and sqft_above. So such variables can be neglected.
- Finally, feature importance graph is drawn in order to point out the significance of each feature based on our model. All the features with zero or no significance can be neglected.

4) Creating Data Science Pipeline:

Even though the dataset is limited to King County, we can see that every housing prices prediction problem will have similar parameters. Apart from that, we are planning to create a Data Science pipeline which can be generalized for any regression problem. Any predictive analytics problem involves the same sequence of steps to achieve the best model, and hence we are trying to create a pipeline for all such problems. It involves a concrete set of steps starting from importing the required libraries, examining the dataset, performing univariate and bivariate analysis and finally building all the available models so that their accuracy can be compared.

TASKS ACCOMPLISHED:

- Reading and analyzing the reference paper – The first step we have completed is analyzing the paper, by understanding the approach in the cited paper. After which, we came up with some interesting additions to this approach. Based on the set of novel ideas suggested by group members, we finalized on four options which can be considered as considerable improvements over the original approach.
- Understanding the variables - Next step in the process was to understand the variables and their importance. We assessed each variable individually by plotting its histogram, understanding its mean and variance.
- Univariate and Bivariate analysis – Every variable is checked for outliers by plotting a box plot. Also, bivariate analysis is performed using 'Seaborn' package. We have removed some outliers to make the work easier.
- Feature Engineering - As discussed above, we have completed the feature engineering process by creating the age of house attribute from years. And zipcodes are converted to dummy binary variables.
- Linear Regression – We performed linear regression by considering various sets of variables. And the performance of model after applying feature engineering is much better compared to the original dataset. The best RMSE value so far is 149,318.

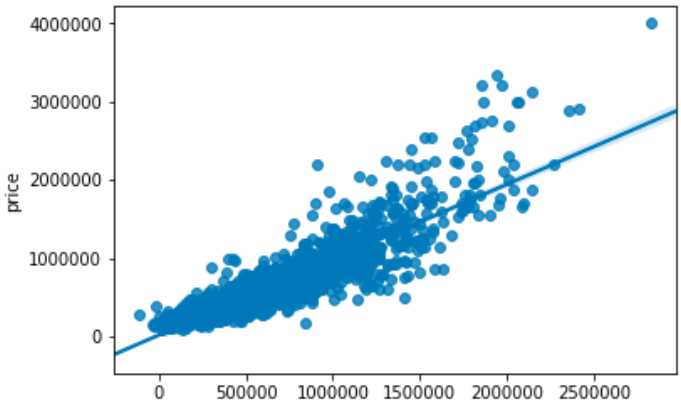

Fig 2: Linear Regression model (After feature engineering)

RESPONSE TO THE FEEDBACK:

"This is what I told you not to do for a project. This project simply repeats the process that we are doing in our assignment. Working with 3 people, the project ask you to do more than typical assignment. If you are using published dataset, you must have an access to the published paper. See how others work on this. Based on the survey, figure out what you can do differently from them."

So, based on the above feedback given by you, we have considered the following changes:

- As a reference we have considered previous work on housing data by Victor Gan, Vaishali Agarwal and Ben Kim in the paper '**Data mining Analysis and predictions of Real estate prices**'. They have achieved decent results by applying Decision trees and Neural networks.
- Based on this study, we have provided commendable changes or improvements in our project. Some of the improvements suggested are:
  1) Feature Engineering
  2) Ensembling methods
  3) Dimension reduction.
  4) Creating a data science pipeline for solving almost any type of prediction problem.
  (All the improvements are clearly explained above)
- We are thinking about other implementations involving complex Neural network structures for best possible representation of houses pricing data.
- Even though the dataset is limited to King County, USA, the housing prices problem is a generalized problem which is useful everywhere. So, our final step would be building some action rules from this data, which can be generalized and implemented in all the use cases involving housing prices.

TIMELINE FOR REMANING PART:

- Building various Ensemble models (Including hyperparameters optimization)
    - Bagging
    - XG Boost
    - Gradient boosting - By April 15

- Applying dimension reduction and other techniques to improve model performance further - By April 20

- Creating Data science pipeline for generalized implementation - By April 25

- Finalizing the project and preparing report - By April 30

TEAM MEMBERS:

- Sai Yesaswy Mylavarapu
- Sainandan Tummalapalli
- Hemanth kumar Koraboina