Bipul Islam
SBU # 111578726

Project Report

Following is an executive summary of the current HW2. The final goal of the project was to give out error estimates of Zillow's algorithm for 2.98 million property listings. There was complete information about a property and error margins in its sale for about 90k listings-- these formed the data for model building and evaluation.

About the Task

The problem was quite challenging because the subtleties in this data are a bit hard to gauge.

- Primarily reason being, the entire task hinges on predicting the errors of a black box model that is in business for sometime.
- Secondly, for most of the intuitive features that one may think are important indicators of prices of housings are sparse.
- Third there are composite data as well as simple data. Like some columns as in Zip code are good enough to infer city, region, state, neighbourhood etc., so there is redundancy.
- Categorical variables itself makes me quite uncomfortable, so the flood of such variables was a bit jarring so to say.

Data Description

Analysis of the data showed that there are broadly 3 types of attributes in the data set.

- Categorical, both numeric as well as strings.
- Numeric but non-quantized, and,
- Numeric discrete, as in quantized like count variables.

It was observed that some of the count data correlated well like bedroom and bathroom counts, while others had very slight dependence.

Tax data - be is holistic, land, structure intuitively sounds like a good indicator but some cursory analysis shows in fact the room-count is anti-correlated to total taxamount. Interestingly enough while there is a significant correlation between bedroom and bathroom count, there is not so much correlation among them and the room count.

It would make that trying to look at correlations of the logerror with other variables, but that turned up with very low correlation values, much to my dismay.

Attribute selection

I personally leaned towards selecting:

- a majority of the count data like count of rooms, bathrooms, bedrooms;

- the descriptive elements of a house like: Total living area, total free space around the house, total garage space etc.
- Apart from this I relied heavily on Tax elements as I was convinced that predicting the actual price is probably a completely different ball game as compared to predicting error of prices against a black box.
- Apart from this I decided to take in account some numeric categorical variables that seemed interesting like Air conditioning type, building quality rating and heating system. Some of these seemed like parameters that zillow system factor as they were not as scarce as many other data.
- One decision that was consciously taken was selection of variables that have very less amount of missing data. This would ensure minimal interference on my part in massaging the data into shape.

Handling the missing data

Given my choice of variables and way I selected them the data cleaning strategy was very straight forward.

- For Numeric quantities, I imputed the missing values by the median of the same quantity observed in the collection. Then I have used the simple normalization of subtract mu and divide by Standard deviation.
- For Categorical yet numeric quantities, I decided to denote the missing data as a category as no-information. So I imputed these areas with 0. Now most models would probably interpret numeric categorical labels as some sort of rating. Better option has a higher numeric rating. Which is not always true, and also given the information available to me especially from the data-dictionary -- it was not very clear. So I decided to take up one hot encoding. Essentially for a column with n levels, we add n-1 extra columns each with a binary switch defined on each levels of the original columns.

Model building

First step was to build a linear regression model with a subset of these variables. Although the RMSE value was quite low, but so was the R^2 metric that explains the variance of the model. First attempt was to improve that model with the larger variables set and with more data cleaning operations, but there was no significant change of model performance.

The second thing that I tried out was the KNN regressor model, on the same data set. While coding I ended up setting up both training and test variables on the same data set without noticing. However hearing that most of the people were getting lousy R^2 values I decided to re-eyeball my code and then I found what has happened. Upon fixing that it was found that my model had an R^2 of -0.17. That is equivalent to saying my model does worse than a horizontal line. There was in tunable parameter with this model, I tampered with it for a bit and realised if I kept increasing the parameter which determines the number of nearest neighbours the model considers before making the decision the r^2 slowly increased without much change in RMSE value. Last estimates were R^2 of -0.005 and an RMSE close to 0.03.

The third model that I tried was a Random forest. The default implementation with the given data set yielded a R^2 of -.0.15. I however progressively tuned it with by varying the number of estimators and tree depth going up to 4. Latest model has an R^2 of close to + 0.01, with comparable RMSE as other two previous models.

The Kaggle submission Experience

This was the first time I have submitted anything on Kaggle and seeing myself on the leaderboard (although the tail end) was an interesting experience. Especially I I realised that I was not saving the normalizing metric for each columns from the training period. So I retrained my model to save those parameters. And re-generated my submission. I moved up by about 19 ranks. I would say I am hooked but, I had to finish a lot of associated tasks with this project so had to break off.

Learnings

I realise that I am not at all adept at tackling the data where it is very sparse or there are a lot of categorical variables. I need to pick up necessary skills related to that this semester. Also I need to work with Pandas more in order to be more at home with the scripting as I am generally with R.