Project Report
Assignment-2

Problem: The challenge was to predict the log error in comparison to Zillow predicted log errors.

Models:
I used 3 different models:

- Linear Regression Model.
- Decision Tree Regressor.
- SK Learn Linear Regressor

Out of these, I liked the Linear Regression model the most.

Working of Linear Regression Model

StatsModels API: I used statsmodel api in order to perform linear regression on the dataset. I used "**Ordinary Least squares**" method to perform the linear regression.
OLS is a method to predict the dependent or unknown variables in a linear regression model.
Goal: The goal of this model is to minimize the square of the errors between the predicted values and the given actual values. The error is calculated by summing the squares of vertical distance between the actual data point and the corresponding data point on the line.
Working: So, let's consider the equation:

$$\mathbf{Y = m_1 X + C}$$

X = Independent variable
Y = Dependent variable
Y is called the scalar response and X is called as vector of predictors.

OLS tries to find both of these constants in order to draw a regression line which tries to cover most of the given Y points or minimize the squares of the errors.
But generally, in real models we do linear regression between a set of variables and not just 1 independent variable. So, consider this eq.

$$Y = m_1X_1 + m_2X_2 + m_3X_3 + \ldots\ldots\ldots\ldots + m_iX_i$$

In this case, instead of finding the line, the model founds the equation of a plane in 3D space and then find the errors relative to that plane.
After the prediction, Statsmodel OLS method gives as a table to show the result. It calculates R-square vale and OLS value of the model.
R-square value represents that how many actual dependent values lie on the regression line. So, higher R-Square value is always good for the model. Generally, OLS should be as least as possible.

Evaluation of the OLS Linear Regression Model

1. I used the mean absolute error to compare the models. The OLS model gave the result of 0.68 while the Decision Tree Regressor gave the value of 0.90 as mean absolute error on the chosen variables. So, the OLS linear regression model performed better in this parameter.
2. The R-square value that I got from this is 0.001 and sometimes 0.006 based on the independent variables that I took which is very low. It shows that this model only explained 1 or 6 percent of the total Y values. This is because we are evaluating against Zillow's very trained model.
3. As the dependent variables are spread over different domains, for example, the "Bathroomcnt" varied from 1 to 20 while the "FinishedSquaredFeet" varied from 1K to around 6K. So, I normalized and scaled the input vector by calculating the Z-scores but the R-squared values decreased and so is the rank that I got from the Kaggle. So, I concluded that z scores made this model worse.

4. When I submitted the predicted log error values on the Kaggle, I received a score of 0.0649077 with a rank of 2061 while with z-scores I received a rank of around 2900 with a score of 0.0661867.

Interesting Experiences

1. The most important thing in Data Science is perhaps to clean the data ready to be fit in a model. An uncleaned data can make a model worse.
2. We have to appropriately handle the missing values inside the columns and rows. We can replace with Nan with Zeros or mean according to the situation or we can drop the columns if the provided values are very less and not sufficient to train the model.
3. One of the interesting thing was that I spent a lot of time on sklearn models and they were giving me low scores on Kaggle but when I submitted with the very first model that I used i.e. a linear regression model, I got a rank boost of around 1K. So advanced model is not always better.
4. When I submitted the predicted values after scaling the same independent variables by z-score method, the rank in fact decreased which was very annoying.
5. The RMSE value was always there to make me happy as it was always very low because the predicted values and actual values are very low. So RMSE is not a good method in this problem to evaluate the models.
6. The independent variables should be chosen carefully.

References:

- https://en.wikipedia.org/wiki/Ordinary_least_squares
- http://setosa.io/ev/ordinary-least-squares-regression/