**Practical Machine Learning Exercise: My Insights**

The problem statement is predicting the how much a house will sell on by using the UK government’s land registry data. Given that we have a huge data set, I took a million eighteen thousand four hundred and fifty records in a partial sample set for 2018 and 2019. where 2018 consists of 516,906 files and 2019, 501,544 files.

In the financial world, outliers are inevitable. Asset prices are varying and fluctuating in the market. Some properties in London could be ridiculously expensive, whereas some in rural areas in the UK could be unbelievably cheap. It is less accurate to represent housing prices with the mean price. we're analysing the housing market in the whole UK. Some properties could be sold at very high prices in London, whereas some properties might be sold at 1 pound in Liverpool.

Our sample data we have the maximum value is £337 million and the minimum value is £1. However, this doesn't mean there is only 1 data point for £337 million. Multiple house prices could be equal to £337 million. Next, we know the third quartile is £337K. This means in the sample dataset, 75% of housing prices are below £337K. In contrast, the first quartile is £145K meaning 25% of housing prices is below £145k.If we look into the Mean and Median price bit higher end of year 2019 compared to 2018. Also, price distribution also left Skewed.

The median price is higher for new properties than for existing properties. As you might think, leased properties are cheaper than wholly owned properties. The detached properties are much more expensive than the other.

Among the property type flat are more similar slod for London as well. London by itself sold 40,926 in the year of 2018 and 2019. Sales are on the rise in June to August 2018. Another interesting point is out of the top 10 cities, 5 places are in the North of England, i.e., Nottingham, Leeds, Liverpool, Sheffield, Leicester. I'm intrigued to know the population growth over years in these places in terms of price variation London has huge price variation.

Overall new property is 129435 and old is 889015. Between January 2018 and December 2019, the number of newly built properties was 129435 and the number of freeholds properties was 777843 and 2.63% was sold more the £1 million. If we compare January 2018 and January 2019, the volume of transactions drops by 58%, while the median price increases a little, we can also note the peak of transaction volume that occurred in Aug 2018. Most importantly old houses being sold than new houses.

**Regression Model:** Since we have categorical and numerical features, as a part of feature engendering, I have applied Label Encoder for categorical feature and log transformation for target variable. Also applied feature scaling to the target variable. In order to understand what are all feature are more important I apply feature selection with Lasso Regression model. Out of 15, 9 features are important for the model building.

As OLS method indicates coefficient, std error and P-value pretty hight for few features like record\_status, year, month etc, that indicates that that feature has highly multicollinearity also p-Value is high for those features and r2 and adjust r2 pretty low.

After doing all the feature scaling, feature selection and VIF, we have a one feature with highly correlated with the target variable. So, we can remove. Although linear regression was not performed so well. Where R- square and adjusted R-squared almost same 13%. For training and test data set also MSE 0.59 RMSE 0.77 and MAE 0.53 for Decision Tree Regression and Random Forest Regression both are same MSE 0.49, RMS 0.35 and RMSE 0.70. I also observed Decision Tree and Random Forest training R2 high but test R2 fairly low, which indicates overfitting model.

**Scope of improvement:**

* If we use the population data from 1995 to 2018, probably model accuracy will increase.
* Since regression model very sensitive towers outliers, if we use advance state-of art model that might help us to learn non-linear relationship. Hence model accuracy might increase.
* Cross validation and hyper parameter tune might help us to get better result.