**Final Report Submission**

**Introduction to the project**

This project describes a used case where we have a client and they are planning to set up their upcoming store in a big city. They are new to the city and have very less idea on which area would be best to setup there store. For this we can help them with some data of different localities of the city, analyse those data and predict which would be the ideal locality they should invest in to get the best out of their store. The data set is generally giving us the sales of few already set up stores in different localities and we can use up that data to predict which market condition or which location is best suited to set up the store.

**Project Problem**

The problem that we are dealing in this situation is that we have detailed data of number of localities in a city and we have numerous features of different stores situated in that localities. We have challenge to make sense out of this unclean data and make a predictive model out of this, which can help us predict best location suitable to deploy store in a huge city. The another challenge we have with this problem statement is that the model which we are proposing should be very much accurate and should be believable.

**Key Contribution**

There is a huge contribution of Amity team who have helped me achieve completion of this project. I am thankful to internet resources like medium, you tube, towards Data science and statsmodel.

**Machine Learning**

Machine learning techniques/algorithms are used to make prediction out of learning from the existing data. We can apply machine learning algorithm on data using python which is most suited when it comes to machine learning and deep learning.

* **Data set resource:** The data set that we are dealing in this problem statement is in a CSV formatted data with two files namely train.csv gathering data for training and test.csv file gathering data for testing the model’s accuracy on.
* **Data set pre-processing:** Data columns namely “*commercial\_property*” and “*school\_proximity*” have null values in it and have we have replaced the null values with the columns means values for both the columns. We have also made categorical columns “*new\_store*”, “*transport\_availability*” and “*country*” have replaced these data data with numerical encoding data to make the model more accurate as the model only understands data in numerical level.
* **Finding the best machine learning model for the problem objective:** As the class variable is a continues data, that it is not be classified to different groups it is a value of sales so we have to use a regressor rather than a classifier. We have came up with RandomForestRegressor() regressor from present in sklearn library of python language. This regressor gives the best value of accuracy of whopping **97.60 % of accuracy**. The reason of selecting this model is that Random Forest has the least variance on testing data and has a decent value for the bais, in short the bais-varience trade is best taken care of by Random Forest.
* **Training and testing the model:** The training of the model is done using a hyperparameters optimazation technique which uses tells which parameters should we use in our model to make the prediction more accurate, one of such hyperparameter optimizer is RandomizedSearchCV which we have used in out application with RandomForestRegressor().
* **Model Evaluation:** The model which we created and trained gave a very good result. The accuracy of prediction for our model is as close as 98%, which is very good. This means that what ever result out model is presenting, there is a 98% chance that the prediction of the model is accurate.

**Analysis of result**

As a result of the project we can conclude that the locations with Id 148, 400, 199 are the top 3 best picks by our model which would give the best sales results. However after evaluating a little more deeper we realized that if we sort the top 10 location holding best sales, with top 5 feature values than we get a more accurate result.

Finally the output which we present is location ID 199, 477, 227.