**Mid way report**

**Detailed Introduction of the project**

This project describes a use case where we have a client and they are planning to set up there 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.

**Data set description**

This data set which we will be using is in CSV format and we have 15 columns in the data set defining different aspect to analyse a location for opening store.

Below are the details of the column values:

* **location\_id :** Unique Id for each location to categorise they differently.
* **crime\_rate :** Crime rate of the particular area/locality.
* **proportion\_flats :** Determines the proportion of residential flats to total flats.
* **proportion\_nonretail :** Determines the proportion of non retailing stores in the localities.
* **new\_store :** Which of the retailing store is newly opened (Categorical data.)
* **commercial\_property :** Percentage of commercial property in the area.
* **household\_size :** What is an average size of the house hold in that area.
* **proportion\_newbuilds :** Determines the proportion of newly constructed buildings.
* **public\_transport\_dist :** How readly is transportation facility available in that area.
* **transport\_availability :** Categorically classification of how readily is transportation available.
* **property\_value :** How costly is the property.
* **school\_proximity :** Average number of schools available in the area
* **competitor\_density :** How competitive is the market in that area.
* **household\_affluency :** What is the household affluency of the area.
* **county :** A categorical value determining the county.

**Propose a machine learning model.**

As this a classification problem I have planned to go with RandomForest model as we have many different variables and we need to select the best out of them, taking output from different decision trees would be the perfect choice.

**Preliminary results.**

On initial analysis and test run the accuracy of the model is 89.99% which is pretty good however this result is with Classifier with zero parameters value once we get the perfect values of different parameter we can get even a better model.