Road map

**Week 1 work: Dataset Creation:**

**6 step overview for machine learning modelling**

**1.Problem: classification or regression**

Classification

2.**Data : Structured or Unstructured**

We are using simple csv files in structured format

3.**Evaluation: What defines our success**

We want to tell that which type of program (Locker,Helix, HIS,3P) will have how much success rate in that location based on our training data.

**4.Features : Features:**

1. **Type of Store:** drug store,supermarket,gas station,general store
2. **No. of employees:** taken such that

Drugstore<gas<general<super

To make it more precise in terms of working emp

1. **Rating :** random
2. **Availability :** tells the availability but no role in predicting
3. **Location:** approx. 800 different places in Banglore city are taken
4. **Total sq ft:** area of the shop
5. **Dpmo(defects per million objects):** random

**What we are expecting from our ML model ?**

**Labels:**

**Type of the program: 3P(1) ,IHS(2),locker(3),Helix(4)**

**Result: 12950 training entries**

**(used Kaggles house prediction dataset for location to get some valid location which is better than random string)**

**Challenges faced: Making it !!**

5**.Model**

Logistic regression,Naïve Bayes,Random Forest can be used in [**multi-class classification**](https://developers.google.com/machine-learning/glossary#multi-class) problems

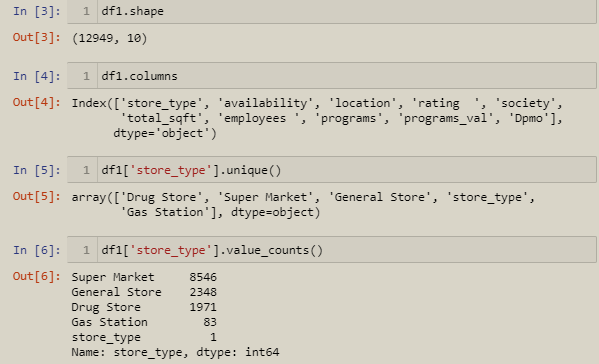
**Week 2 work**

**Lets walkthrough how my model is trained, have fun☺**

1. **Load data into dataframe:**



1. **Now some data analysis:**



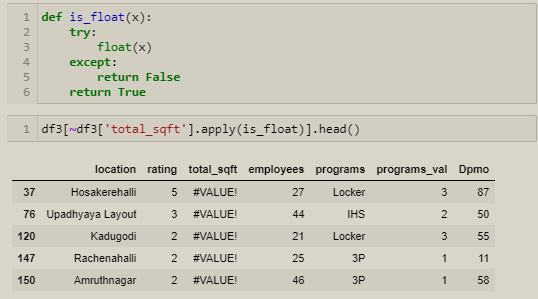
1. **So much data , now time to clean some**

Drop some unwanted features: like store type,society,availability column from 10 to 7

Remove all the null values



From the total sqft remove the values that are not numeric



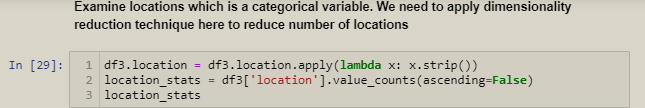
Removing

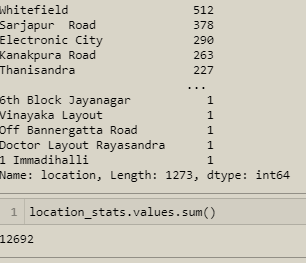


Similarly remove the rating which are greater than 5 as its not possible

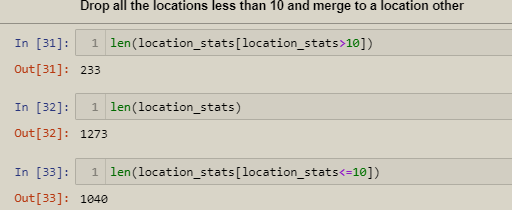


Now our data somewhat looks neat and clean to work on ☺





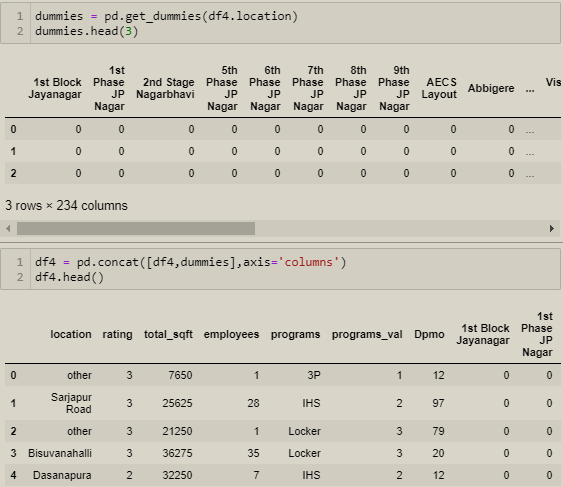
Drop all the locations less than 10



**Any location having less than 10 data points should be tagged as "other" location. This way number of categories can be reduced by huge amount. Later on when we do one hot encoding, it will help us with having fewer dummy columns**

**Use One Hot Encoding For Location**

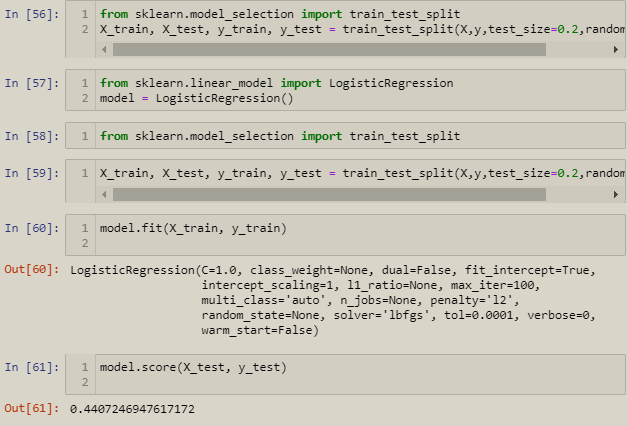
ML algo doesn't work with char so do hot encoding to convert it into numerical



All set to go now time to build the model

Remember : I used 80 :20 ratio for test train

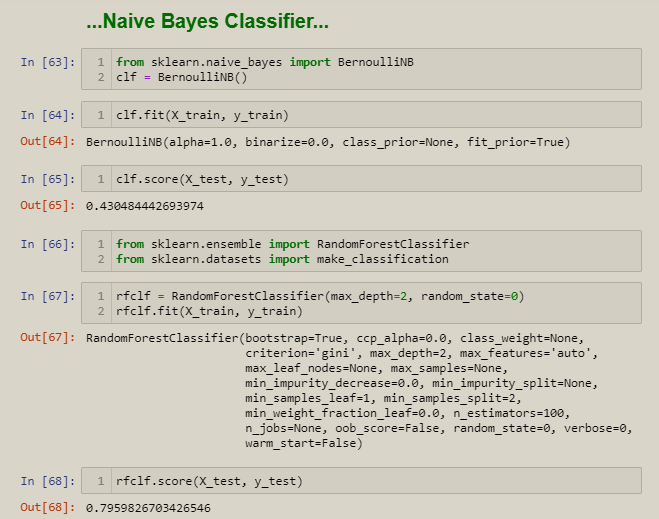
**Logistic Regression**



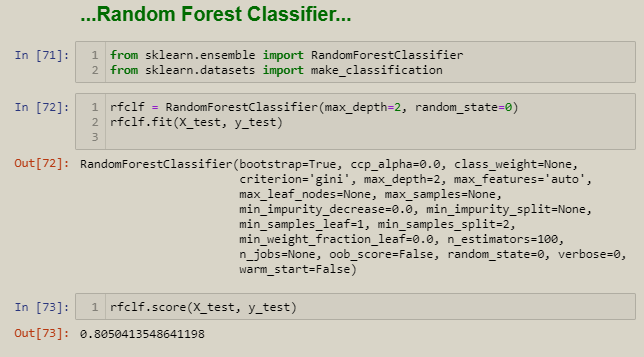
Accuracy of 44.07%

Its not nice lets check from other classifier algorithms .As it is mostly of random data so don’t know in actual which one is more suitable

**Naïve Bayes Classifier**



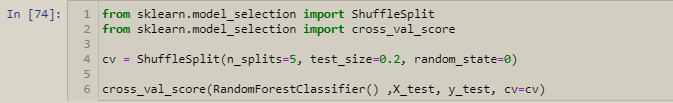
Seems like it is more accurate for our data.



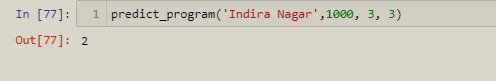
**K fold cross validation**

K-fold cross validation works by breaking your training data into K equal-sized “folds.” It iterates through each fold, treating that fold as holdout data, training a model on all the other K-1 folds, and evaluating the model’s performance on the one holdout fold.

Cross validation is useful because it provides a lower-variance estimate of the model’s true out of sample score than if you had only used a single train-test split.

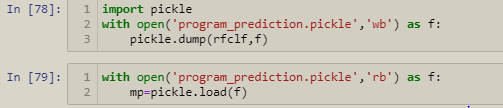


Lets check one input if we are getting output

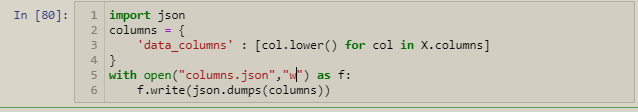


So its working fine

## Export the tested model to a pickle file



## Export location and column information to a file that will be useful later on in our prediction application



Tadda all set to go :-!!

Now use this pickle file in your frontend part and see the prediction.