## 2.Project Methodology

- Understanding the problem and project requirements, Executing Data mining, Data preparation and preprocessing.
- Performing EDA(Exploratory Data Analysis) which included feature selection ,Correlation, Data cleaning, Replacing Values,Handling outliers and understanding feature importance.
- Converting categorical values into numerical values using One hot encoding. Using normalization on feature variables.
- Separating the feature variables and target variable(Performance).
- Selecting the appropriate model for the task.
- Separating Train/Validation/Test set and Using Validation set to evaluate Training data and optimize the models hyperparameters.
- Used Cross Validation(Monte Carlo and K Fold) to check mean scores and GridSearchCV for Hyperparameter tuning .
- Finally using the test set to evaluate the performance of the model and display the results using Confusion Matrix and Accuracy Score from sklearn.metrics.

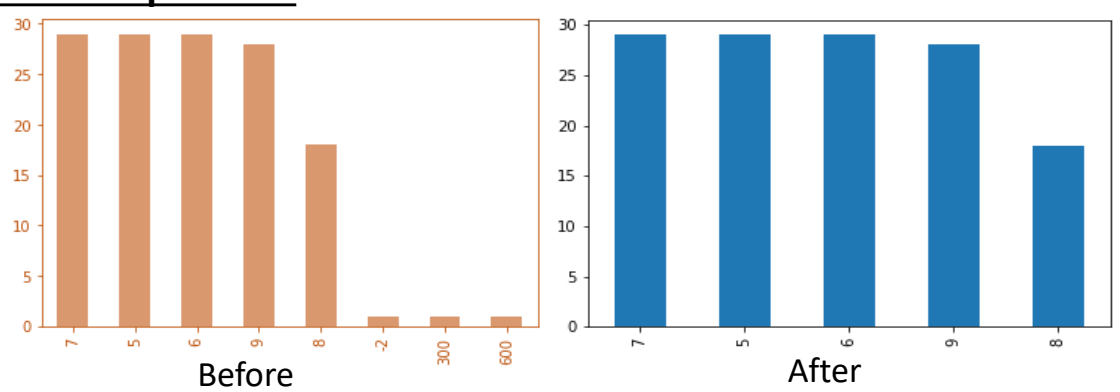
## 3.Variables

➡ Continuous ➡ Discrete ➡ Nominal ➡ Ordinal

| Variable | Type |
|---|---|
| 40min population | Numerical |
| 30 min population | Numerical |
| Floor Space | Numerical |
| Clearance space | Numerical |
| Competition score | Numerical |
| Location | Categorical |
| Staff | Numerical |
| Window | Numerical |
| 20 min population | Numerical |
| Competition number | Numerical |
| Car park | Categorical |
| Performance | Categorical |

The variables in the table are chosen after using feature selection methods with the help of Filter Feature Selection and Univariate feature selection , I had to use one hot encoding on the data frame so that the categorical variables such as “Location”, “Car park” and “Performance’” values could be converted between 0 and 1,The dataset also contained columns such as “Store Id”, ”Manager Name”, ”Country” and “Town” which had to be excluded since they had no significance with our purpose of building a ML model, By doing so I was able to solve my overfitting problem.

## 4.Data Preparation



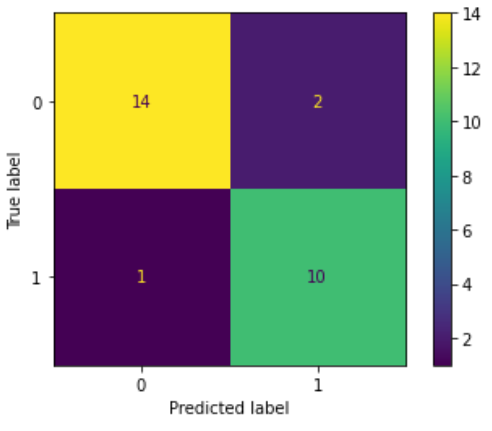

The Following Histograms show the distribution of Number of Staff of the Feature variable “Staff”, Before cleaning, We can see that we have some outliers such as -2, 300 and 600 .I decided to remove these values from my “Staff” column ,After Cleaning the feature vairable we can see that the histogram looks much better .The dataset also contained a feature named “Car park” where the duplicate values had to be replaced.

## 5.Model Training and Hyperparameters

| Model | Hyper Parameters | Parameter List used in GridSearchCV | Validation Metric |
|---|---|---|---|
| Logistic Regression | (C=40, solvers = 'saga’, penalty =l2) | {"penalty":["l1","l2"], "solver":['lbfgs','sag', 'saga' ,'newton-cg'], "C":[5,10,20,30,40,50]} | Cv_score Mean: 84% |
| Random Forest Classifier | (max_features=3,criterion='entropy',max_depth=4) | {'max_features':[2,3,4,5],'criterion':['gini','entropy'],'max_depth':[1,2,3,4,5]} | CV Score Mean: 75% |
| MLP Classifier | (solver=‘lbfgs’, hidden_layer_sizes=(10), max_iter=2000, activation=‘identity’,alpha=‘0.001') | {"hidden_layer_sizes": [(5), (10)],"activation": ['identity','tanh','relu','logistic’],” max_iter“:[100,500,1000,1500,2000],”solver":['lbfgs','sgd','adam'],"alpha": [0.01, 0.001]} | Cv_Score Mean: 82% |

Each model was trained on the training dataset and tested on the validation set .All the changes were made on the validation set. 20 % of the data was allotted to the test dataset which was used to evaluate the final model .Initially I had randomly assigned values to the hyperparameters which gave me a decent cross_val_score, However to optimize the hyperparameters I used GridsearchCV .A list of parameters were given from which GridsearchCV obtained a list of best hyperparameters that I needed to run in my model. After Hyperparameter tuning I was able to increase my cross_val_score by a good amount .Logistic Regression performed very well on the Training set as well as Validation set. The CV Score mean is the average value of CV=10.

## 6.Final Model & Results




The Model used on the Test data was Logistic Regression as it performed well on the training and validation data. The model was evaluated on the best set of Hyperparameters and was checked for its efficiency using AUC,ROC and Precision score .The accuracy obtained on the test data was 88% I.e., it correctly predicted the performance of a given store as either Good(1) or Bad(0), which was highest out of the 3 models and the Cross Validation Score for Logistic Regression gave favorable outcome which is why I chose this as my Final Model.

- There were 14 true negatives (TN): the model correctly predicted a negative outcome 14 times.
- There were 2 false positives (FP): the model incorrectly predicted a positive outcome 2 times.
- There were 1 false negative (FN): the model incorrectly predicted a negative outcome 1 time.
- There were 10 true positives (TP): the model correctly predicted a positive outcome 10 times.