BDNS End Term Project

Store Size Prediction

Gourab Saha(C23012)
Praxis Business School

•Objective

Predict the Size And type of outlet store by understanding the properties of product and outlet which play a key role in increasing sales. we will try to predict this by building a model.

•Data Set Information

Big Mart sales data contains 6135 rows and 7 feature for 1559 products across different stores.

•Library Used

- Numpy
- Pandas
- Matplotlib
- Seaborn
- sklearn

Attribute Information

•Item_Weight: Weight of product

•Item_Visibility: The % of total display area of all products in a store allocated to the particular product

•Item_MRP: Maximum Retail Price (list price) of the product

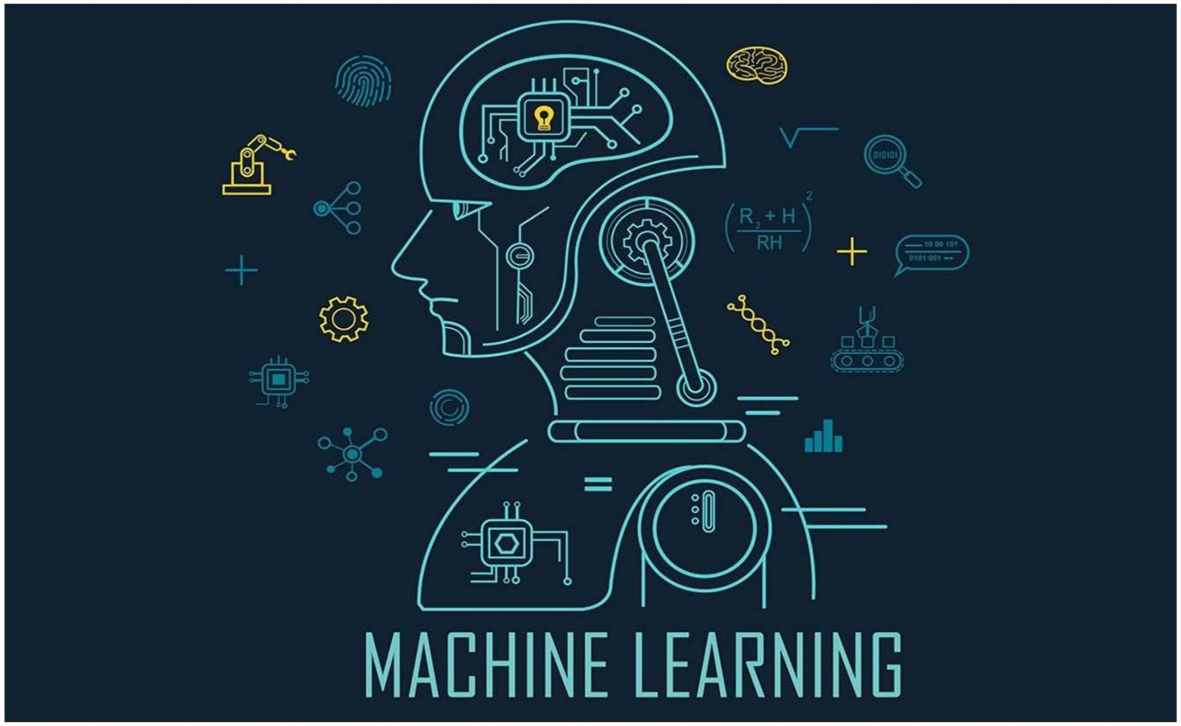
•Outlet_Establishment_Year: The year in which store was established

•Outlet_Size: The size of the store in terms of ground area covered

•Item_Outlet_Sales: Sales of the product in the particular store. This is the outcome variable to be predicted

Ml Tools-

1. Logistic Regression
2. Naive Bayes
3. Random Forest
4. Decision Tree



Procedure-

1. Uploading data in Mongodb through clever cloud credential.
2. Accessing data through spark
3. Imputing missing values through various method.
4. Using String Indexer convert categorical feature into their respective indexer.
5. Now by One hot encoding convert categorical column into numerical type.
6. By assembler divide o/p column and i/p column features.
7. Now by using pipeline we will use different algorithm to build model.
8. Random Forest is giving is best accuracy to predict the size of store.

Results

- **Logistic Regression**

  Accuracy of logistic regression is - 70.74 %

- **Naive Bayes**

  Accuracy of Naive Bayes is - 78.4

- **Decision Tree**

  Accuracy of Decision Tree is - 94

- **Random Forest**

  Accuracy of Random Forest is - 95.5

Conclusion

- Random Forest is giving highest accuracy because it is ensemble of decision tree i.e. bagging ensemble
- Logistic regression regression is giving us least accuracy because in logistic regression decision boundary should be linear and in my data points they are non linearly distributed,
- Decision Tree is almost giving Same accuracy as random forest as classes are less so it can easily divide them one class to other class.
- For further improvement some hyperparameter optimization can be done to improve accuracy.

Thank You