TABLE OF CONTENTS

Definitions, Acronyms and Abbreviations

Big Mart is One Stop Shopping center and Free Marketplace. Buy, sell and advertise without fee or at low cost. in Big mart sales prediction we will predict the sales of a store.For example, i want to find what drives the sales amount for a certain product in different stores and try to predict where and how I can maximize the sales for this particular product. The task is to predict the sales of a certain product at a particular store, part of a chain of stores and find out what influences that sale.

- **Introduction**

In today's world the need for new products and better-quality products are increasing at a rapid rate. Supermarket or grocery store are becoming a go to place to access these products and as a reason it highly needed that these supermarket chains be able to forecast future sales to make better decisions which would lead them to higher profits. This report focuses on predicting the sales of products which are located at 10 different outlets which belongs to the Big Mart chain.

- **Overview**

Sales prediction is a very common real life problem that each company faces at least once in its life time. If done correctly, it can have a significant impact on the success and performance of that company.in this project we will be working on the Big Mart Sales Prediction Challenge in order to predict the sales of a products in different stores.

- **Scope**

This report showcases the comparison between Random Forest,Linear Regression and Decision tree models and gives an overview of the data mining tasks that were performed.

- **Objective**

The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. The main objective is to understand whether specific properties of products and/or stores play a significant role in terms of increasing or decreasing sales volume. To achieve this goal, we will build a predictive model and find out the sales of each product at a particular store. This will help BigMart to boost their sales by learning optimised product organization inside stores.

- **LiteratureSurvey**

Sales prediction is a very common real life problem that each company faces at least

once in its life time. If done correctly, it can have a significant impact on the success and performance of that company.in this project we will be working on the Big Mart Sales Prediction Challenge in order to predict the sales of a products in different stores.

This section will showcase the implementation of Random Forest, Linear Regression and Decision tree method on the Big Mart sales data. There are two models
1. Predicts the Sales of Item for that store

2. Predicts the quantity of Item sold instead of Item Sales. The idea behind this implantation is that quantities sold might make more sense than the sale of item. In the final step while checking the accuracy Item Sold is multiplied with Item MRP.

A. Decision Trees B. Random Forest C.Linear regresion model

- **Methodology**

This section will showcase the implementation of Random Forest, Linear Regression and Decision tree method on the Big Mart sales data. There are two models
1. Predicts the Sales of Item for that store

2. Predicts the quantity of Item sold instead of Item Sales. The idea behind this implantation is that quantities sold might make more sense than the sale of item. In the final step while checking the accuracy Item Sold is multiplied with Item MRP.

A. Decision Trees

Decision trees is a machine learning technique that are used for classification and regression problems. The idea behind this algorithm goes in a top-down approach where you all the train cases at the node and then you split the tree in to branches until you the reach the leaf node. Decision tree uses Gini Index / Entropy to split the nodes. Gini Index measures the impurity of the attributes and choses the attributes which are the purest. The attribute with Gini score 0 is the purest.

B. Random Forest

Random Forest is a popular ensemble learning method. As the name suggests it creates a forest of decision trees and out of those trees the one which has the highest majority is chosen as a final model which will be used for prediction. Random forest takes N attributes form the dataset and then it splits the data into edges, just like decision trees which uses Gini Index or Entropy to determine the split points Random Forest also considers those metrics to choose the best split point. It will create N number of trees with each tree is made on subset of data and in the end it calculates the votes that each tree has and chooses the one which has the majority votes.

- ***Proposed Approach***

The decision tree is the proposed approach in this big mart sales prediction model.

The decison tree is the better current approach than other approaches because its accuracy score is more than other approaches in big mart sales prediction model.and this is the best topdown approach for classification and regresion problems.

- ***High Level System Architecture***

The main modules are:

1. Data Description
2. Data Cleaning
3. Feature Engineering
4. Creating Models

Data Description

The Big Mart sales data consists of 8523 rows and has 12 variables. The variables are described in the Table 1.1:

Variable	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particulat store. This is the outcome variable to be predicted.

Data Cleaning

Before we implement the algorithms on the data we have make sure that data is clean so that we can get appropriate results. In the Big mart sales data following columns shoes missing values: “Item_Weight”, “Outlet_Size”. Following steps were taken to get rid of the missing values. For “Item_Weight” we replace the missing values with mean of the column and for “Outlet_Size” we replace the values with the mode of column.

Feature Engineering

Attribute “Item_Fat_Content” is a categorical attribute which had two categories: Low Fat and Regular. However, the data had Low Fat, low fat, LF, reg and Regular which were then renamed Low Fat and Regular respectively. Attribute “Outlet_Establishment_Year” did not had much intuitive meaning and hence it was replaced with how old the store is. This might help us determine better sales because if store is

Creating Models

Lenier regresion model,Decision tree model,Random forest model

Environment Requirements

Hardware Requirements:-.Processor intel(R) Pentium(R) CPU N3710 @ 1.60GHZ

1.60,RAM 4GB,System type 64 bit operating system ,X64 based processor,Hard disk 80GB.

Software Requirements:-Operating System: Windows10,Jupytor Notebook,Anaconda Navigator

- ***Hardware Requirements***

IV. Processor intel(R) Pentium(R) CPU N3710 @ 1.60GHZ 1.60

V. RAM 4GB

VI. System type 64 bit operating system ,X64 based processor

VII. Hard disk 80GB

- ***Software Requirements***

I. Operating System: Windows10

II. Jupytor Notebook

III. Anaconda Navigator

- **Data Requirements**

The datasets are: A)Train.csv B)Test.csv

Dataset Details-

The data has 8523 rows of 12 variables.

Dataset Description -

Variable - Description

Item_Identifier- Unique product ID

Item_Weight- Weight of product

Item_Fat_Content - Whether the product is low fat or not

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

Item_Type - The category to which the product belongs

Item_MRP - Maximum Retail Price (list price) of the product

Outlet_Identifier - Unique store ID

Outlet_Establishment_Year- The year in which store was established

Outlet_Size - The size of the store in terms of ground area covered

Outlet_Location_Type- The type of city in which the store is located

Outlet_Type- Whether the outlet is just a grocery store or some sort of supermarket

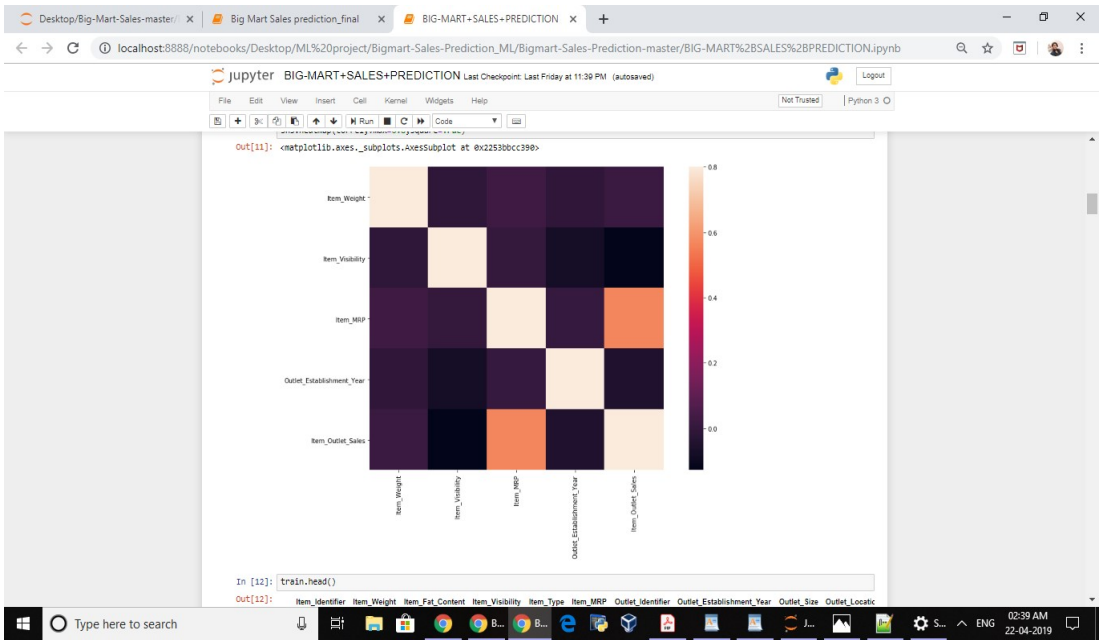
Item_Outlet_Sales - Sales of the product in the particulat store. This is the outcome variable to be predicted.

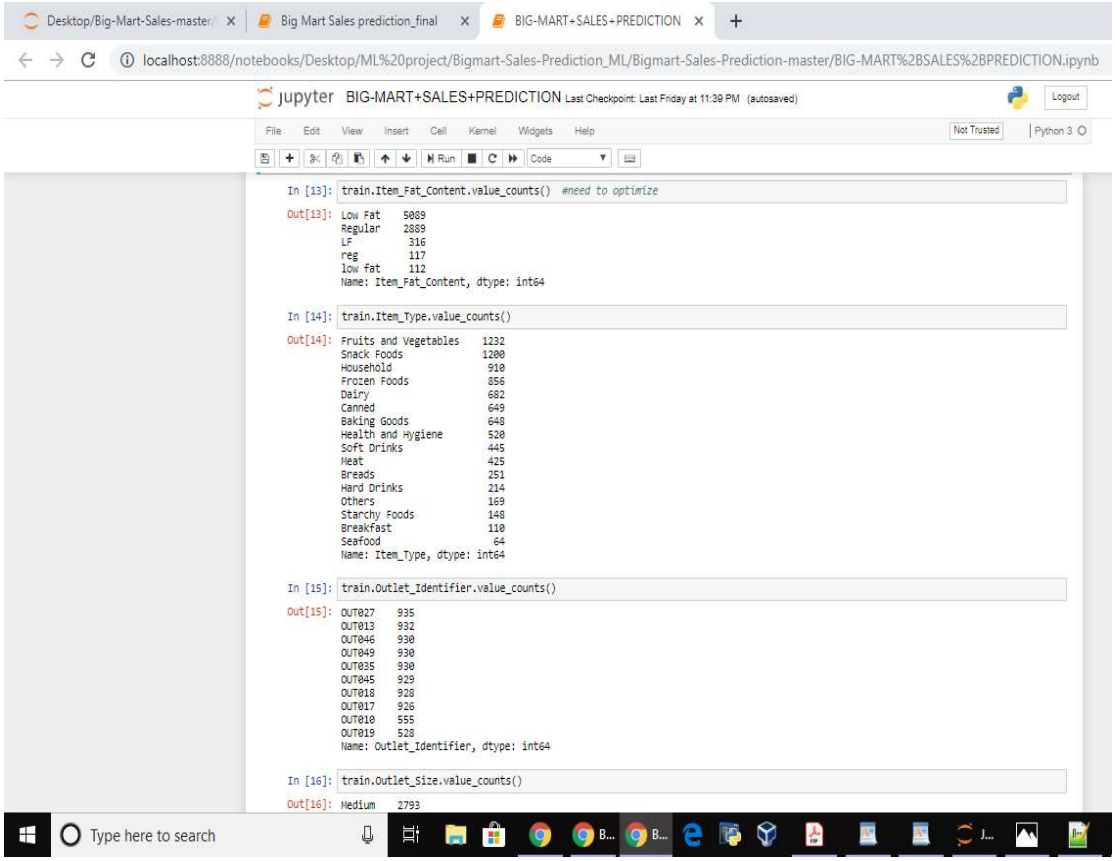
Reference: https://datahack.analyticsvidhya.com

- ## Proposed Approach

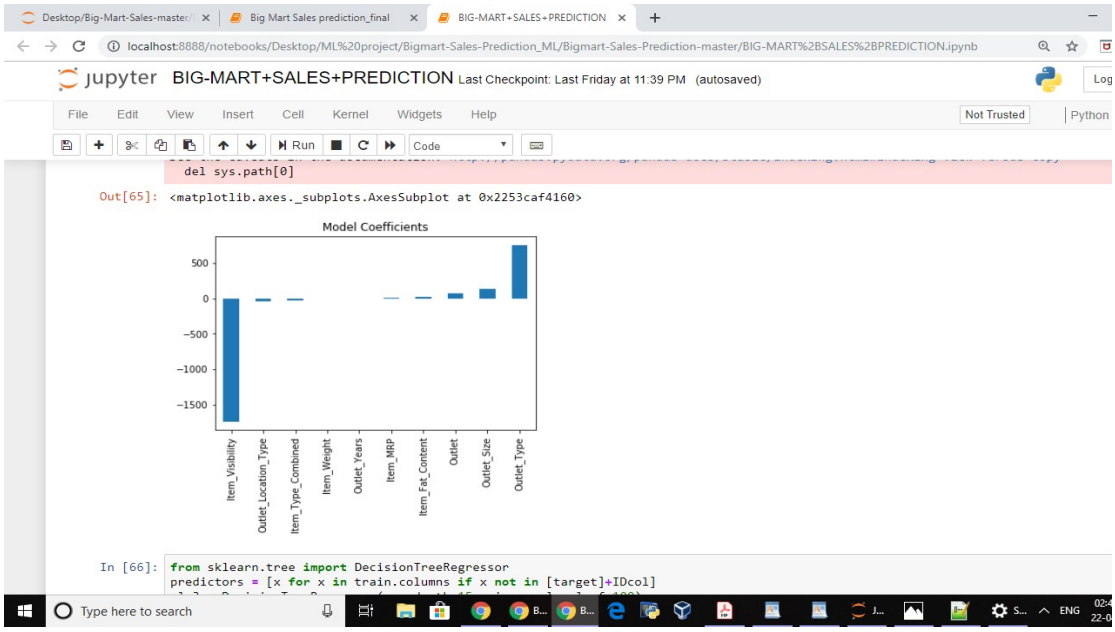
The decision tree is the proposed approach in this big mart sales prediction model.The decison tree is the better current approach than other approaches because its accuracy score is more than other approaches in big mart sales prediction model.and this is the best topdown approach for classification and regresion problems.Decision trees is a machine learning technique that are used for classification and regression problems. The idea behind this algorithm goes in a top-down approach where you all the train cases at the node and then you split the tree in to branches until you the reach the leaf node. Decision tree uses Gini Index / Entropy to split the nodes. Gini Index measures the impurity of the attributes and choses the attributes which are the purest. The attribute with Gini score 0 is the purest.

- ## Results

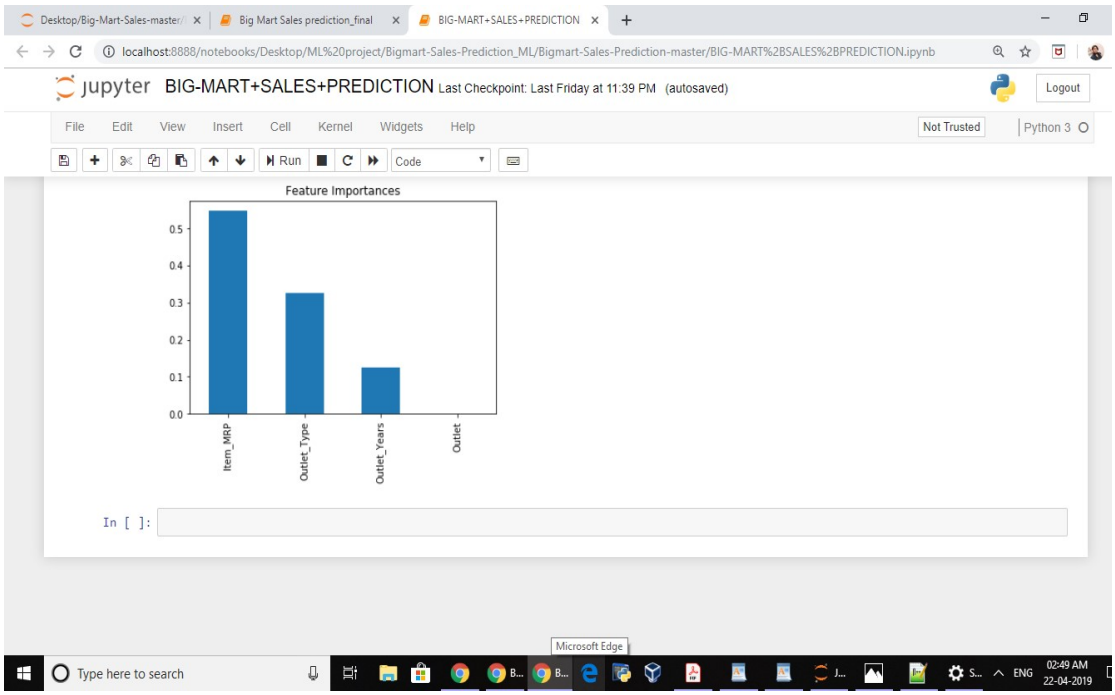

Desktop/Big-Mart-Sales-master/
Big Mart Sales prediction_final
BIG-MART+SALES+PREDICTION
localhost:8888/notebooks/Desktop/ML%20project/Bigmart-Sales-Prediction_ML/Bigmart-Sales-Prediction-master/BIG-MART%2BSALES%2BPREDICTION.ipynb
jupyter BIG-MART+SALES+PREDICTION Last Checkpoint: Last Friday at 11:39 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x2253bbcc390>
Item_Weight
Item_Visibility
Item_MRP
Outlet_Establishment_Year
Item_Outlet_Sales
0.8
0.6
0.4
0.2
0.0
In [12]: train.head()
Out[12]: Item_Identifier Item_Weight Item_Fat_Content Item_Visibility Item_Type Item_MRP Outlet_Identifier Outlet_Establishment_Year Outlet_Size Outlet_Locatic
Type here to search
ENG
02:39 AM
22-04-2019


Desktop/Big-Mart-Sales-master/
Big Mart Sales prediction_final
BIG-MART+SALES+PREDICTION
localhost:8888/notebooks/Desktop/ML%20project/Bigmart-Sales-Prediction_ML/Bigmart-Sales-Prediction-master/BIG-MART%2BSALES%2BPREDICTION.ipynb
jupyter BIG-MART+SALES+PREDICTION Last Checkpoint: Last Friday at 11:39 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
In [13]: train.Item_Fat_Content.value_counts() #need to optimize
Out[13]: Low Fat 5089
Regular 2889
LF 316
reg 117
low fat 112
Name: Item_Fat_Content, dtype: int64
In [14]: train.Item_Type.value_counts()
Out[14]: Fruits and Vegetables 1232
Snack Foods 1200
Household 910
Frozen Foods 856
Dairy 682
Canned 649
Baking Goods 648
Health and Hygiene 520
Soft Drinks 445
Meat 425
Breads 251
Hard Drinks 214
Others 169
Starchy Foods 148
Breakfast 110
Seafood 64
Name: Item_Type, dtype: int64
In [15]: train.Outlet_Identifier.value_counts()
Out[15]: OUT027 935
OUT013 932
OUT046 930
OUT049 930
OUT035 930
OUT045 929
OUT018 928
OUT017 926
OUT010 555
OUT019 528
Name: Outlet_Identifier, dtype: int64
In [16]: train.Outlet_Size.value_counts()
Out[16]: Medium 2793
Type here to search

Lenier Regresion:



Decision Tree regresion:



- ## Conclusions

In this Big mart sales prediction model we have used a various methodologies in order to predict the product sales of diffrent stores.In this the decision tree model will be the better approach because its accuracy is more when compare to other approaches.

- ## Future Work

Implementing the Big mart sales prediction model for predicting the product sales for more stores by using the various approaches.

References

[1] https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3518362/

[3] https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674

[4] Hands-On Machine Learning with Scikit-Learn and TensorFlow By Aurelien Geron. ISBN 9781491962299

[5] https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python

[6] https://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/2/

[7] https://www.datascience.com/resources/notebooks/random-forest-intro