Detailed Project Report

SALES PREDICTION TOOL

Written by	SACHIN SHARMA MUNENDRA KUMAR KUSHWAHA
Version	1.0
College Name	D.S. Degree college Aligarh

1. Introduction

1.1 Abstract

Machine Learning may be a category of calculations that permits program applications to ended up more exact in predicting results without being unequivocally modified. The fundamental preface of machine learning is to construct models and utilize calculations that can get input information and utilize measurable examination to foresee an yield whereas updating outputs as unused information gets to be accessible. These models can be connected in several regions and prepared to coordinate the desires of administration so that precise steps can be taken to attain the organization's target. In this paper, the case of Huge Shop, a one-stop-shopping- center, has been talked about to foresee the deals of diverse sorts of things and for understanding the impacts of diverse components on the items' deals. Taking different perspectives of a dataset collected for Enormous Mart, and the strategy taken after for building a prescient show, comes about with tall levels of precision are created, and these perceptions can be utilized to form choices to improvesa

1.2MachineLearning

The information accessible is expanding day by day and such a tremendous sum of natural information is required to be analyzed absolutely, because it can provide exceptionally enlightening and finely immaculate angle comes about as per current standard prerequisites. It isn't off-base to say as with the advancement of Fake Insights (AI) over the past two decades, Machine Learning (ML) is additionally on a quick pace for its advancement.

ML is an imperative pillar of IT segment and with that, a or maybe central, but ordinarily covered up, portion of our life. As the innovation advances, the examination and understanding of information togivegoodresultswillalsoincreaseasthedataisveryusefulincurrentaspects. In machine learning, one bargains with both administered and unsupervised sorts of errands and by and large a classification sort issue accounts as a resource for information disclosure. It generates resources and utilizes relapse to form exact expectations around future, the most accentuation being laid on making a framework self-efficient, to be able to do compute.

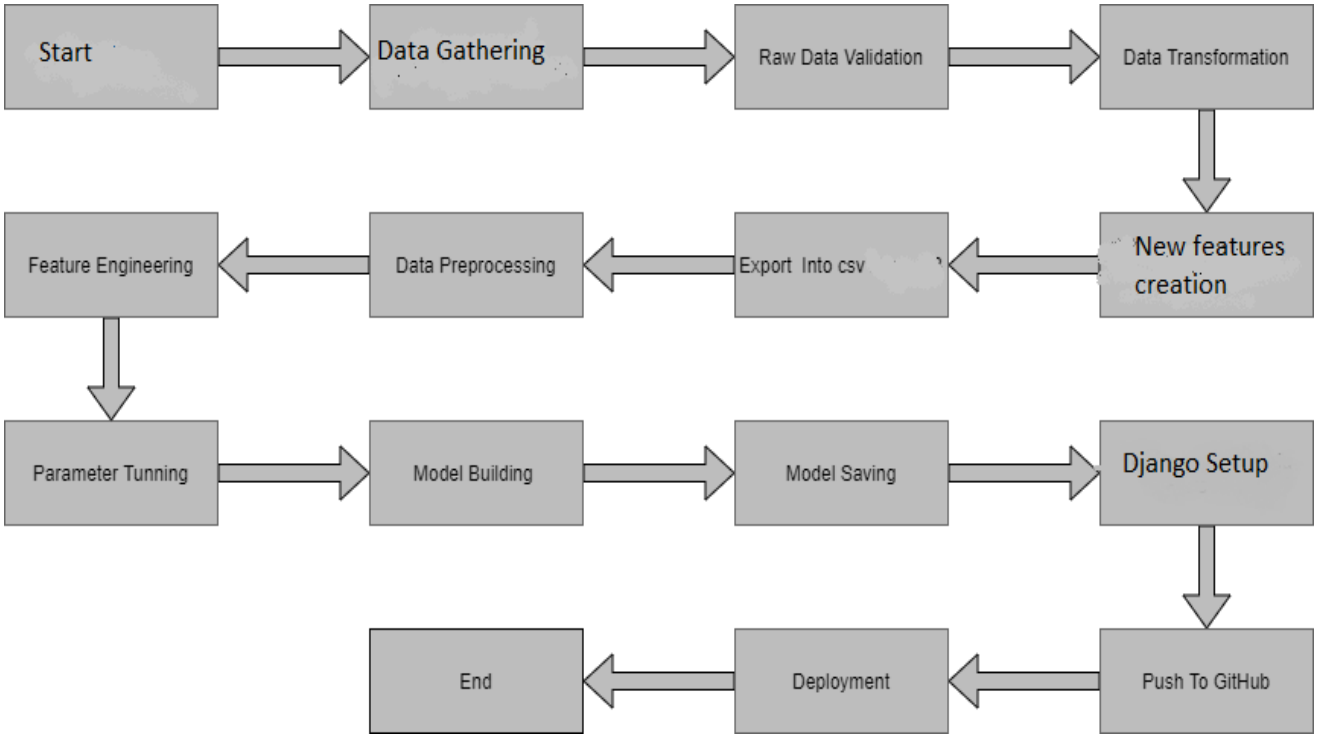
1.3ProblemStatement

"To find out what role certain properties of an item play and how they affect their sales by understanding Big Mart sales."

In order to help Big Mart, achieve this goal, a predictive model can be built to find out the sale of every item for every store. Also, the key factors that can increase their sales and what changes could be made to the product or store's characteristics.

2. Architecture:

Following workflow was followed during the entire project.



2.1 Data gathering:

Data source: https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data
Train and Test data are stored in .csv format.

2.2 Raw Data Validation:

After data is loaded, various types of validation is required before we proceed further for any operation. Validations like checking for zero standard deviation for all the columns, checking for complete missing values in any columns, etc. These are required because The attributes which contains these are of no use. It will not play role in contributing the sales of an item from respective outlets.
Like if any attribute is having zero standard deviation, it means that's all the values are same, its mean is zero. Which indicate that either the sale is increase or decrease that attribute will remain the same. Similarly, if any attribute is having full missing values, then there is no use of taking that attribute into an

account for operation. It's unnecessary increasing the chances of dimensionality curse.

2.3 Data Transformation

Before sending the data into the database, data transformation is required so that data are converted into such form with which it can easily insert into the database. Here, 'Item Weight' and "Outlet Type' attributes contain the missing values. So they are filled in both train set as well as test set with supported appropriate data types.

2.5 New Feature Generation

We can derive new item cateogory from item type and create mrp categories as mrp bin

2.6 Data preprocessing

In data preprocessing all the process required before sending the data for model building are performed. Like, here the 'Item Visibility' attributes is having some values equal to 0, which is not appropriate because if item is present in the market, then how its visibility can be 0. So, it has been replaced with the average value of the item visibility of respective 'Item Identifier' category. New attributes was added named ''Outlet years", where given establishment year is subtracted from the current year. New "Item Type" attribute was added which just take first two character of the Item Identifier which indicates the types of the items. Then mapping of "Fat content" is done based on 'Low', 'Reg' and 'Non-edible'.

2.7 Feature Engineering:

After preprocessing it was found that some of the attributes are not important to the item sales for the particular outlet. So those attributes are removed. Even one hot encoding is also performed to convert the categorical features into numerical features.

2.8 Parameter tunning:

Parameters are tunned using Randomized searchCV. Four algorithms are used in this problem, Linear Regression, Gradient boost, Random Forest and XGBoost regressor. The parameters of all these 4 algorithms are tunned and passed into the model.

2.9 Model building:

After doing all kinds of preprocessing operations mention above and performing scaling and hyper parameter tunning, data set is passed into all four models, Linear Regression, Gradient boost, Random Forest and XGBoost regressor. It was found that Gradient boost performs best with smallest RMSE value i.e. 587.0 and highestR2score equals to 0.55. So 'Gradient boost' performed well in this problem.

2.10 Model saving:

Model is then saved using pickle library in. pkl format.

2.11 Django setup for data extraction:

After saving the model, API building process started using Django. Web application creation was created here. Whatever the data user will enter and then that data will be extraction by the model to predict the prediction of sales, this is performed in this stage.

2.12 Git Hub

Whole project directory will be pushed into GitHub repository.

2.13Deployment:

Cloud environment was set up and project was deployed form GitHub into Heroku cloud platform.
App link- https://salepredict.herokuapp.com/

3.Data set description

Big Mart's data scientists collected sales data of their 10 stores situated at different locations with each store having 1559 different products as per data collection. Using all the observations it is inferred what role certain properties of an item play and how they affect their sales.Thedatasetlookslike as follow:

```
1 train_df.head()
```

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Locatic
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT013	1987	High	

_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.1380
Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.2700
Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	Tier 3	Grocery Store	732.3800
Low Fat	0.000000	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052

The data set consists of various data types from integer to float to object as shown in Fig.

```
In [5]: 1 # datatype of attributes
        2 df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8523 entries, 0 to 8522
Data columns (total 12 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Item_Identifier            8523 non-null   object
 1   Item_Weight                7060 non-null   float64
 2   Item_Fat_Content           8523 non-null   object
 3   Item_Visibility            8523 non-null   float64
 4   Item_Type                  8523 non-null   object
 5   Item_MRP                   8523 non-null   float64
 6   Outlet_Identifier          8523 non-null   object
 7   Outlet_Establishment_Year  8523 non-null   int64
 8   Outlet_Size                6113 non-null   object
 9   Outlet_Location_Type       8523 non-null   object
 10  Outlet_Type                8523 non-null   object
```

In the raw data, there can be various types of underlying patterns which also gives an in-depth knowledge about subject of interest and provides insights about the problem. But caution should be observed with respect to data as it may contain null values, or redundant values, or various types of ambiguity, which also demands for pre-processing of data. Dataset should therefore be explored as much as possible.
Various factors important by statistical means like mean, standard deviation, median, count of values and maximum value etc. are shown below for numerical attributes.

```
1 train_df.describe()
```

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	8519.000000	8519.000000	8519.000000	8519.000000	8519.000000
mean	12.875420	0.069442	141.010019	1997.837892	2181.188779
std	4.646098	0.048880	62.283594	8.369105	1706.511093
min	4.555000	0.003575	31.290000	1985.000000	33.290000
25%	8.785000	0.033085	93.844900	1987.000000	834.247400
50%	12.650000	0.053925	143.047000	1999.000000	1794.331000
75%	16.850000	0.094558	185.676600	2004.000000	3100.630600
max	21.350000	0.328391	266.888400	2009.000000	13086.964800

Preprocessing of this dataset includes doing analysis on the independent variables like checking for null values in each column and then replacing or filling them with supported appropriate data types, so that analysis and model fitting is not hindered from its way to accuracy. Shown above are some of the representations obtained by using Pandas tools which tells about variable count for numerical columns and model values for categorical columns. Maximum and minimum values in numerical columns, along with their percentile values for median, plays an important factor in deciding which value to be chosen at priority for further exploration tasks and analysis. Data types of different columns are used further in label processing and one-hot encoding scheme during the modelbuilding.

4. Implementation andResults

In this section, the programming language, libraries, implementation platform along with the data modeling and the observations and results obtained from it arediscussed

4.1 Implementation Platform andLanguage

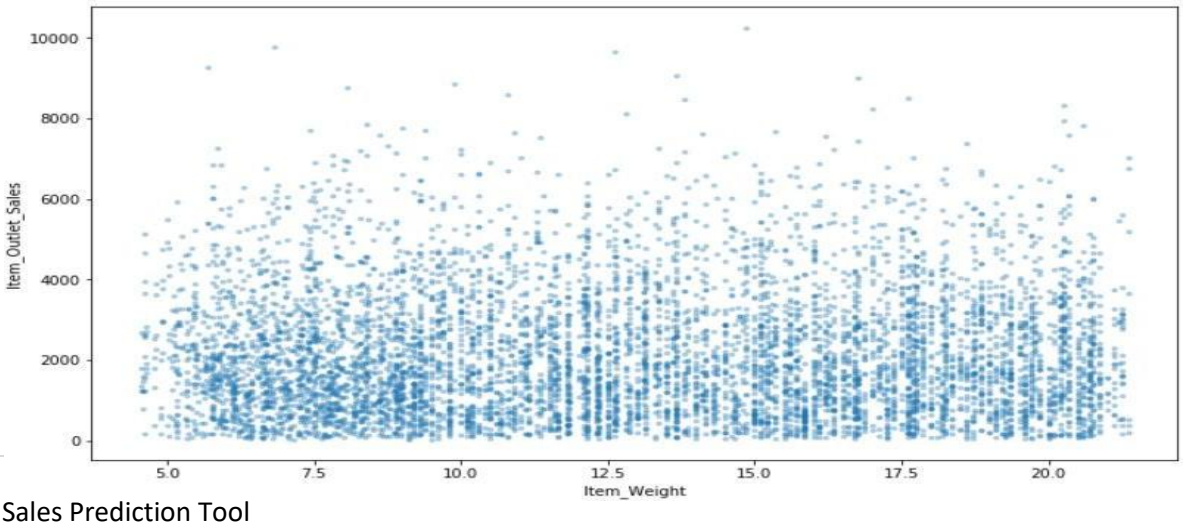
Python is a general purpose, interpreted-high level language used extensively nowadays for solving domain problems instead of dealing with complexities of a system. It is also termed as the 'batteries included language' for programming. It has various libraries used for scientific purposes and inquiries along with number of third-party libraries for making problem solving efficient.

In this work, the Python libraries of Numpy, for scientific computation, and Matplotlib, for 2D plotting have been used. Along with this, Pandas tool of Python has been employed for carrying out data analysis. Random forest regressor is used to solve tasks by ensembling random forest method. As a development platform, Jupyter Notebook, which proves to work great due to its excellence in 'literate programming', where human friendly code is punctuated within code blocks, has been used

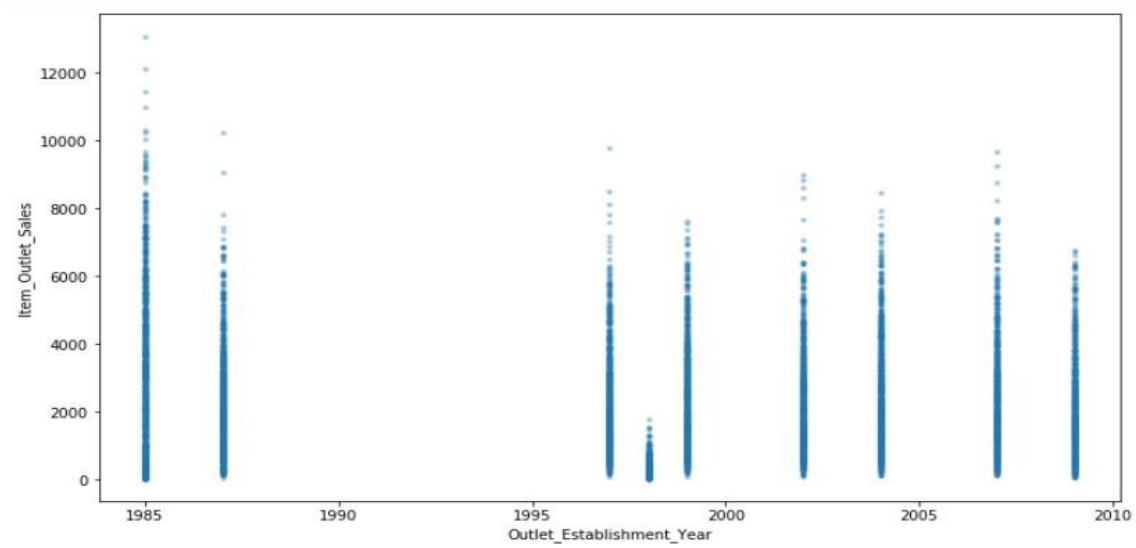
4.2 Observations

Correlation is used to understand the relation between a target variable and predictors. In this work, Item-Sales is the target variable and its correlation with other variables is observed.

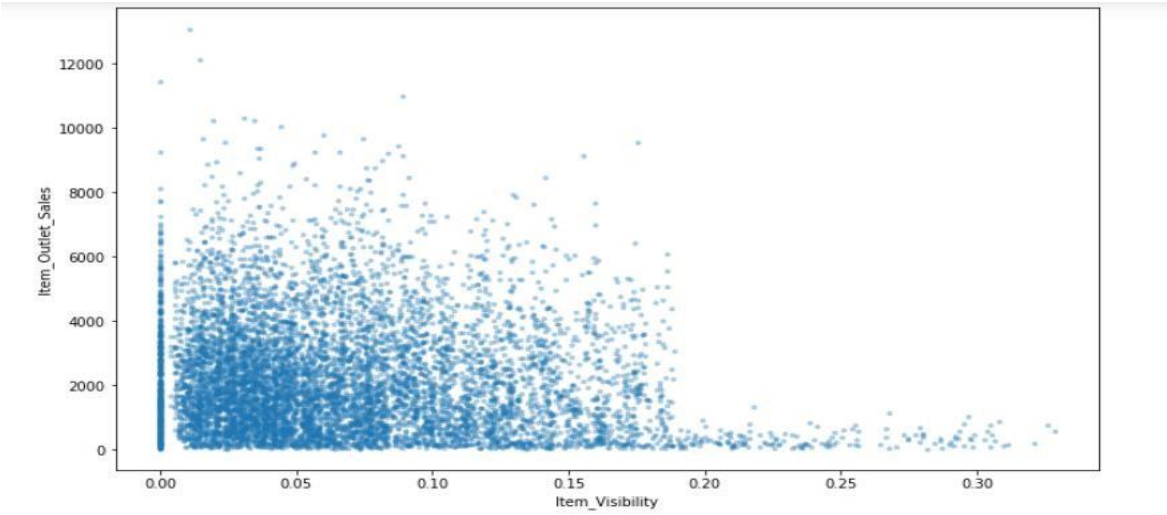
Considering the case of Item-Weight, the feature item weight is shown to have a low correlation with the target variable Item-Outlet-Sales in below Fig.



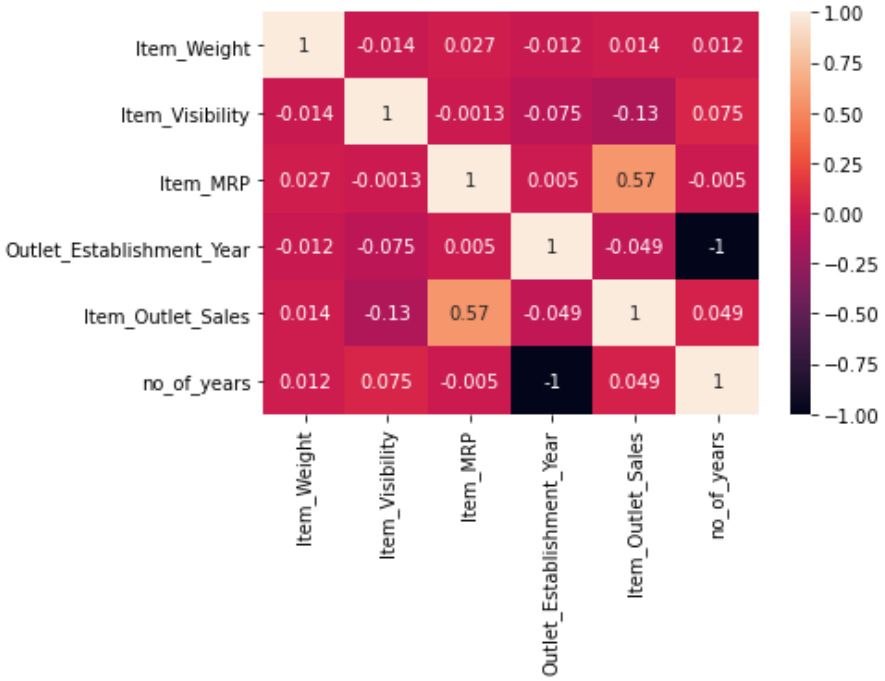
As can be seen from below Fig. there is no significant relation found between the year of store establishment and the sales for the items. Values can also be combined into variables thatclassify them into periods and give meaningfulresults.



The place where an item is placed in a store, referred to as Item visibility, definitely affects the sales. However, the plot chart show that the flow is in opposite side. One of the reasons might be that daily used products don't need high visibility. However, there is an issue that some products have zero visibility, which is quite impossible.



4.3 Correlation



- Item visibility is having nearly zero correlation with our dependent variableItem_Outlet_Sales and grocery store outlet type. This means that the sales are not affected by visibility of item which is a contradiction to the general assumption of “more visibility thus, moresales”.
- Item_MRP (maximum retail price) is positively correlated with sales at an outlet, which indicates that the price quoted by an outlet plays an important factor insales.
- Variation in MRP quoted by various outlets depends on their individualsales.

4.4Metrics for DataModelling

- The coefficient of assurance R2 (R-squared) may be a measurement that measures the goodness of a model’s fit i.e., how well the genuine information focuses are approximated by the expectations of relapse. Higher values of R2 recommend higher model accomplishments in terms of forecast at the side exactness, and the esteem 1 of R2 is demonstrative of relapse forecasts impeccably fitting the genuine information focuses. For further better comes about, the utilize of balanced R2measures works ponders. Taking logarithmic values of the target column within the dataset demonstrates to be noteworthy within the expectation handle. So, it can be said that on taking alterations of columns utilized in expectation, superior comes about

can be derived. One way of consolidating alteration seem too have included taking square root of the column. It moreover gives way better visualization of the dataset and target variable as the square root of target variable is slanted to be a normaldistribution. The blunder estimation is an critical metric within the estimation period. Root cruel squared mistake (RMSE) and Cruel Outright Mistake (MAE) are by and large utilized for nonstop variable's exactness estimation. It can be said that the normal show expectation blunder can be communicated in units of the variable of intrigued by utilizing both MAE and RMSE. MAE is the normal over the test test of the supreme contrasts between forecast and genuine perception where all person contrasts have rise to weight. The square root of the normal of squared contrasts between forecast and genuine perception can be named as RMSE. RMSE is an outright measureoffit,whereasR2isarelativemeasureoffit.RMSEhelpsinmeasuringthevariable's average error and it is additionally a quadratic scoring run the show. Moo RMSE values gotten for direct or numerous relapse compares to way better show fitting. With regard to the comes about gotten in this work, it can be said that there's no enormous contrast between our prepare and test test

4.5 Prediction results

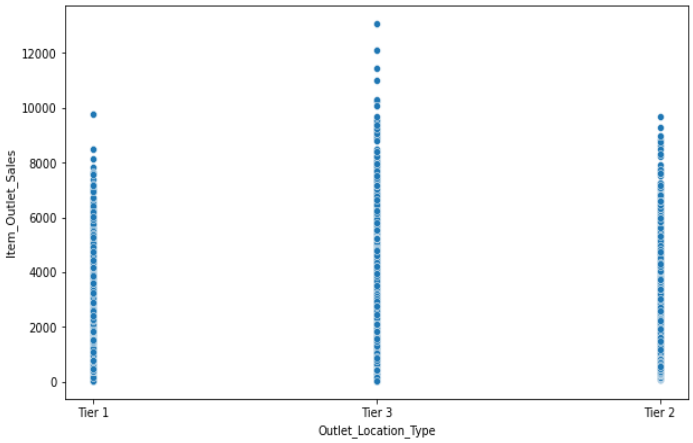
- The biggest area did not create the most noteworthy deals. The area that delivered the most elevated deals was the OUT027 area, which was in turn a Grocery store Type3, having its estimate recorded as medium in our dataset. It can be said that this outlet's execution was much way better than any other outlet area with any measure given within the considereddataset.
- The middle of the target variable Item_Outlet_Sales was calculated to be 3364.95 for OUT027 area. The area with moment most elevated middle score (OUT035) had a middle esteem of 2109.25.
- Adjusted R-squared and R-squared values are higher for Angle boost demonstrate than normal. Too its RMSE esteem is moo as compared to other show with most elevated CV score. Hence, the angle boost demonstrate fits way better and shows accuracy

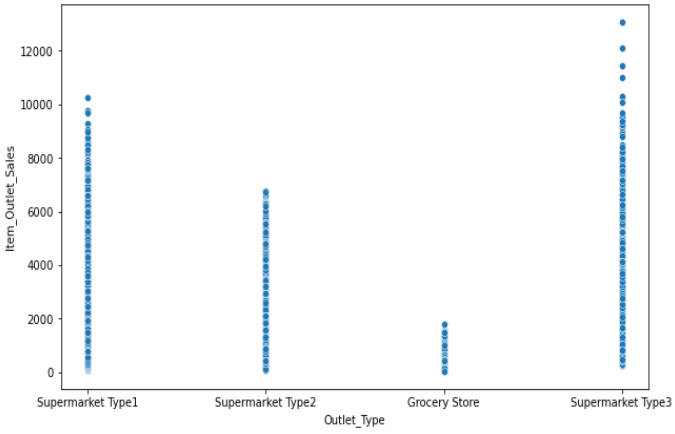
5. Conclusion

In this project, basics of machine learning and the associated data processing and modeling algorithms have been described, followed by their application for the task of sales prediction in Big Mart shopping centers at different locations. On implementation, the prediction results show the correlation among different attributes considered and how a particular location of medium size

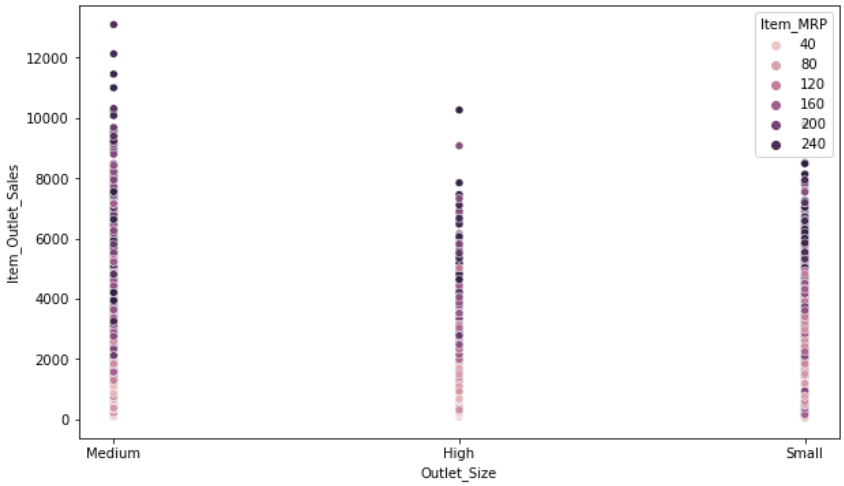
recorded the highest sales, suggesting that other shopping locations should follow similar patterns for improved sales.

Also it can be concluded that more locations should be switched or shifted to Tier-3 in outlet type “Supermarket Type3” to increase the sales of products at Big Mart. Any one-stop-shopping-center like Big Mart can benefit from this model by being able to predict its items’ future sales at





different locations.



6. FutureScope

Different occurrences parameters and different components can be utilized to create this deals expectation more imaginative and effective. Exactness, which plays a key part in prediction-based frameworks, can be altogether expanded as the number of parameters utilized are expanded. Moreover,a see into how the sub models work can lead to increase in productivity of system.The

project can be further collaborated in a web-based application or in any gadget upheld with an in-built insights by ethicalness of Web of Things (IoT), to be more doable for utilize. Different partners concerned with deals data can also give more inputs to assist in theory era and more occasions can be taken into thought such that more exact comes about that are closer to genuine world circumstances are created. When combined with effective data mining strategies and properties, the traditional implies may well be seen to form the next and positive impact on the generally advancement of corporation's assignments on the total.

One of the main highlights is more expressive regression outputs, which are more understandable bounded with some of accuracy. Moreover, the flexibility of the proposed approach can be increased with variants at a very appropriate stage of regression model-building. There is a further need of experiments for proper measurements of both accuracy and resource efficiency to assess and optimize correctly.

7. Q & A:

Q1) What's the source of data?

Ans. The data for training is provided by the client from:
https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data

Q 2) What was the type of data?

Ans. The data was the combination of numerical and Categorical values.

Q 3) What's the complete flow you followed in this Project?

Ans. Refer the Architecture sectionfor this.

Q 4) After the File validation what you do with incompatible file or files which didn't pass the validation?

Ans. Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder.

Q 5) What techniques were you using for data pre-processing?

- Removing unwanted attributes
- Visualizing relation of independent variables with each other and output variables
- Checking and changing Distribution of continuous values
- Removing outliers
- Cleaning data and imputing if null values are present.
- Converting categorical data into numeric values.
- Scaling the data

Q 6) How training was done or what models were used?

- Before diving the data in training and validation set we performed clustering over fit to divide the data into clusters.
- As per cluster the training and validation data were divided.
- The scaling was performed over training and validation data
- Algorithms like Linear regression, Gradient boost, Random forest and XGBoost were used .

Q 7) How Prediction was done?

Ans. The testing files are shared by the client. Wepass its data to the best model which we have saved in pickle format and get the prediction.

Q 8) Where the model was deployed?

Ans. When the model is ready, we deploy it in Heroku platform. This model is an web application where user can enter the data and these data gets extracted in the backend and user gets the prediction result.