Detailed Project Report

Stores Sales Prediction

Written By	Krishan Kumar
Email	Roaankur140@@gmail.com
Date	30-05-2024

Document Change Control Record

Version	Date	Author	Comments

Reviews

Version	Date	Reviewer	Comments

Approval Status:

Version	Review date	Reviewed By	Approved By	Comments

Index

1. Introduction

1.1 Abstract

Machine Learning is a category of algorithms that allows software applications to become more accurate in

predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build models and employ algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. These models can be applied in different areas and trained to match the expectations of management so that accurate steps can be taken to achieve the organization's target. In this paper, the case of Big Mart, a one-stop-shopping- center, has been discussed to predict the sales of different types of items and for understanding the effects of different factors on the items' sales. Taking various aspects of a dataset collected for Big Mart, and the methodology followed for building a predictive model, results with high levels of accuracy are generated, and these observations can be employed to make decisions to improve sales.

1.2 Machine Learning

The data available is increasing day by day and such a huge amount of unprocessed data is needed to be analyzed precisely, as it can give very informative and finely pure gradient results as per current standard requirements. It is not wrong to say as with the evolution of Artificial Intelligence (AI) over the past two decades, Machine Learning (ML) is also on a fast pace for its evolution. ML is an important mainstay of IT sector and with that, a rather central, albeit usually hidden, part of our life. As the technology progresses, the analysis and understanding of data to give good results will also increase as the data is very useful in current aspects.

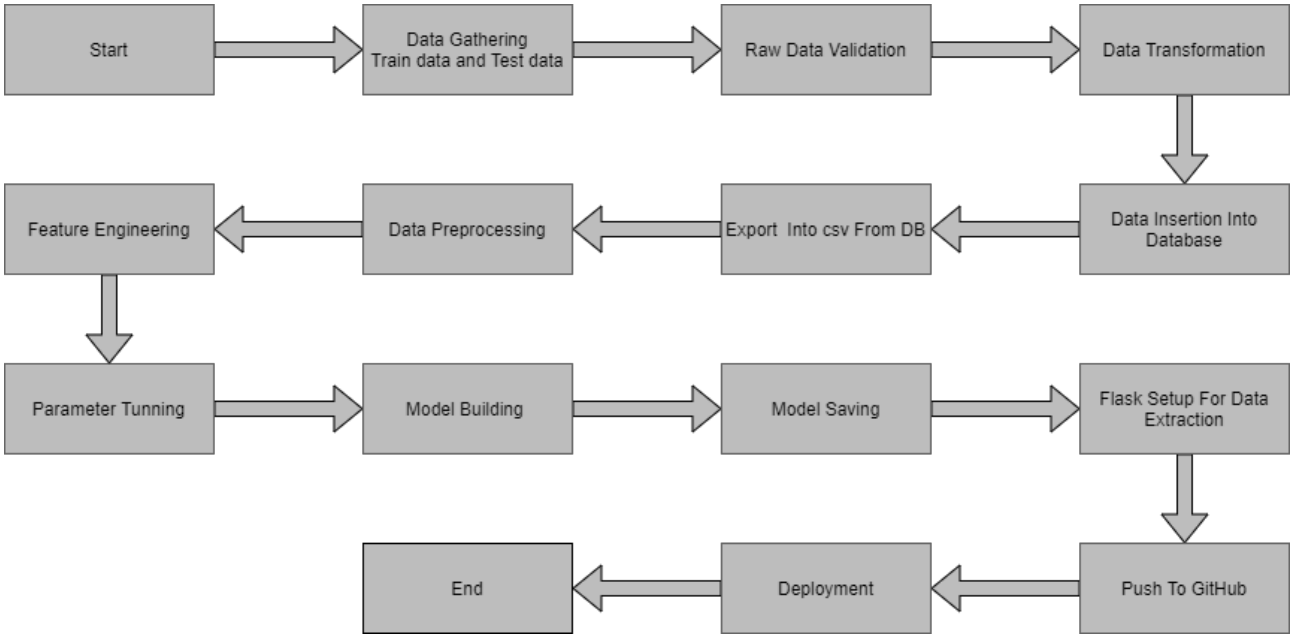
In machine learning, one deals with both supervised and unsupervised types of tasks and generally a classification type problem accounts as a resource for knowledge discovery. It generates resources and employs regression to make precise predictions about future, the main emphasis being laid on making a system self-efficient, to be able to do computations and analysis to generate much accurate and precise results. By using statistic and probabilistic tools, data can be converted into knowledge. The statistical inferencing uses sampling distributions as a conceptual key.
ML can appear in many guises. In this paper, firstly, various applications of ML and the types of data they deal with are discussed. Next, the problem statement addressed through this work is stated in a formalized way.

1.3 Problem Statement

"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 and Random Forest . It was found that Random Forest Regressor performs best with smallest RMSE value i.e. 781.64 and highest R2 score equals to 0.55. So ‘Random Forest’ performed well in this problem.

2.10 Model saving:

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

2.11 Git Hub

Whole project directory will be pushed into GitHub repository.

2.12 Deployment:

Cloud environment was set up and project was deployed form GitHub into Heroku cloud platform.
App link- https://bigmartsalesprediction.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. The dataset looks like 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]:
# datatype of attributes
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
 11  Item_Outlet_Sales          8523 non-null   float64
dtypes: float64(4), int64(1), object(7)
memory usage: 799.2+ KB
```

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.

```
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

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 model building.

4. Implementation and Results

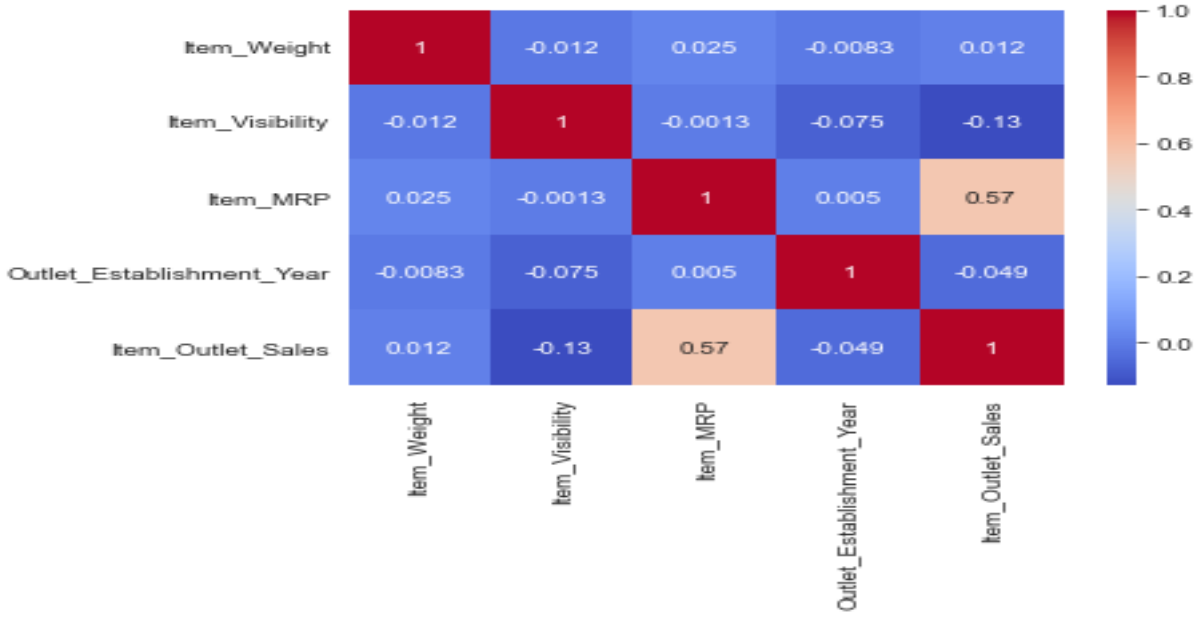
In this section, the programming language, libraries, implementation platform along with the data modeling and the observations and results obtained from it are discussed

4.1 Implementation Platform and Language

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 Correlation



- Item visibility is having nearly zero correlation with our dependent variable Item_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, more sales".
- 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 in sales.
- Variation in MRP quoted by various outlets depends on their individual sales.

4.3 Metrics for Data Modelling

- The coefficient of determination R^2 (R-squared) is a statistic that measures the goodness of a model's fit i.e., how well the real data points are approximated by the predictions of regression. Higher values of R^2 suggest higher model accomplishments in terms of prediction along with accuracy, and the value 1 of R^2 is indicative of regression predictions perfectly fitting the real data points. For further better results, the use of adjusted R^2 measures works wonders. Taking logarithmic values of the target column in the dataset proves to be significant in the prediction process. So, it can be said that on taking adjustments of columns used in prediction, better results can be deduced. One way of incorporating adjustment could also have included taking square root of the column. It also provides better visualization of the dataset and target variable as the square root of target variable is inclined to be a normal distribution.
- The error measurement is an important metric in the estimation period. Root mean squared error (RMSE) and Mean Absolute Error (MAE) are generally used for continuous variable's accuracy measurement. It can be said that the average model prediction error can be expressed in units

of the variable of interest by using both MAE and RMSE. MAE is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. The square root of the average of squared differences between prediction and actual observation can be termed as RMSE. RMSE is an absolute measure of fit, whereas R^2 is a relative measure of fit. RMSE helps in measuring the variable's average error and it is also a quadratic scoring rule. Low RMSE values obtained for linear or multiple regression corresponds to better model fitting.

With respect to the results obtained in this work, it can be said that there is no big difference between our train and test sample since the metric RMSE ratio is calculated to be equal to the ratio between train and test sample. The results related to how accurately responses are predicted by our model can be inferred from RMSE as it is a good measure along with measuring precision and other required capabilities. A considerable improvement could be made by further data exploration incorporated with outlier detection and high leverage points. Another approach, which is conceptually easier, is to combine several sub-models which are low dimensional and easily verifiable by domain experts, i.e., ensemble learning can be exploited.

4.4 Prediction results

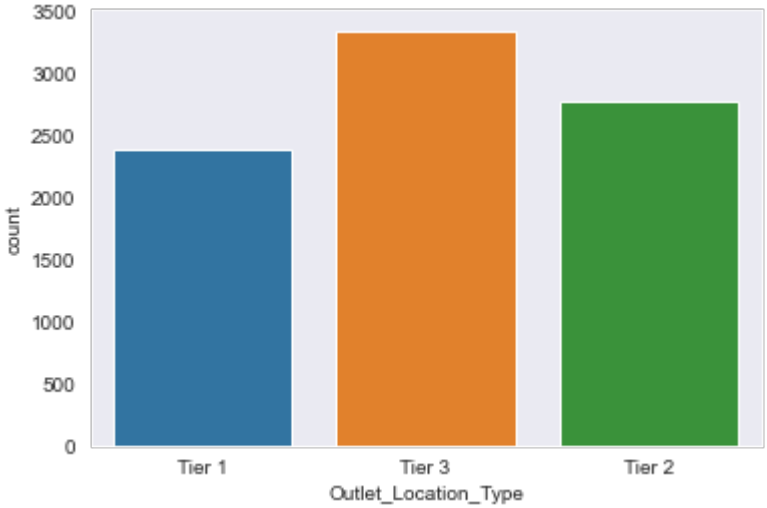
- The largest location did not produce the highest sales. The location that produced the highest sales was the OUT027 location, which was in turn a Supermarket Type3, having its size recorded as medium in our dataset. It can be said that this outlet's performance was much better than any other outlet location with any size provided in the considered dataset.
- The median of the target variable Item_Outlet_Sales was calculated to be 3364.95 for OUT027 location. The location with second highest median score (OUT035) had a median value of 2109.25.
- Adjusted R-squared and R-squared values are higher for Gradient boost model than average. Also its RMSE value is low as compared to other model with highest CV score. Therefore, the gradient boost model fits better and exhibits accuracy

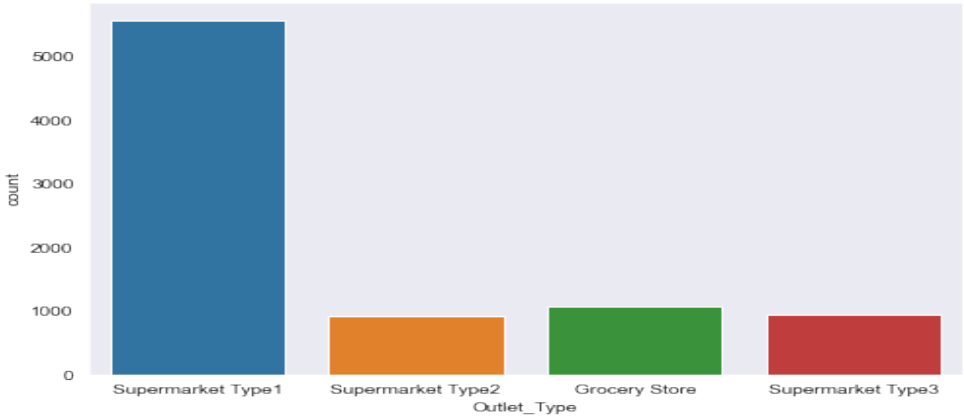
5. Conclusion

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. Future Scope

Multiple instances parameters and various factors can be used to make this sales prediction more innovative and successful. Accuracy, which plays a key role in prediction-based systems, can be significantly increased as the number of parameters used are increased. Also, a look 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 device supported with an in-built intelligence by virtue of Internet of Things (IoT), to be more feasible for use. Various stakeholders concerned with sales information can also provide more inputs to help in hypothesis generation and more instances can be taken into consideration such that more precise results that are closer to real world situations are generated. When combined with effective data mining methods and properties, the traditional means could be seen to make a higher and positive effect on the overall development of corporation's tasks on the whole. 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 section for 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. We pass 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.