INTERNSHIP: PROJECT REPORT

Dear Intern

Project report is an inherent component of your internship. We are enclosing a reference table of content for the project report. Depending on the internship project (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections from the table of content mentioned below. You can also add additional sections. The key objective of this report is for you to systemically document the project work done.

Internship Project Title	Forecasting System - Project Demand of Products at a Retail Outlet based on Historical Data
Name of the Company	TCS iON
Name of the Industry Mentor	Himalaya Ashish
Name of the Institute	ICT ACADEMY OF KERALA

Start Date	End Date	Total Effort (hrs.)	Project Environment	Tools used
13-02-2022	20-03-2022	125	Python Environment	Google colab
Milestone #	3	Milestone:	Create data set, Clean and sanitize dataset, Preprocessing data set, Test and train the dataset, Build the regression models and fit the data in the model	

TABLE OF CONTENT

ACKNOWLEDGMENTS

I am conveying my sincere gratitude towards my industry mentor Himalaya Ashish, and academic mentor, Aswathy for helping me throughout this project till now and providing me this wonderful platform to complete this project. I am also thankful for answering my queries at every phase of the project. I also want to thank all my friends who helped me with valuable suggestions during this project.

OBJECTIVE

Our field of study is concerned with the sales of BigMart. We have the sales data of BigMart of the year 2013. Also, certain attributes of each product and store have been defined. Our aim is to determine what are the factors that are affecting the sales of bigmart and build a predictive model for sales

INTRODUCTION

Big Mart is a big supermarket chain, with stores all around the country and its current board set out a challenge to all Data Scientist out there to help them create a model that can predict the sales, per product, for each store to give accurate results. Big Mart has collected sales data from the year 2013, for 1559 products across 10 stores in different cities. With this information the corporation hopes we can identify the products and stores which play a key role in their sales and use that information to take the correct measures to ensure success of their busines In this paper, the case of Big Mart, a one-stop-shoppingcenter, 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 take decisions to improve sales. The objective of this framework is to predict the future sales from given data of the previous year's using Machine Learning Techniques

INTERNSHIP ACTIVITIES

- Watched the welcome kit videos.
- Done preparations for RIO – pre-assessment.
- Attended the RIO – pre-assessment test.
- Went through the day-wise plan.
- Read the project reference material.
- Read the industry project material.
- Watched webinar 1.
- Watched webinar 2.
- Gone through all posts in the digital discussion room.
- Went through the linear regression YouTube videos.
- Read the linear regression article.
- Watched the lectures provided and other videos for further understanding.
- Created a GitHub account.
- Searched and found out a proper data set for this project.
- Wrote activity reports.
- Checked and clarified the data set whether it has enough data for the project.
- Read articles and find out how to clean and sanitize the data.

- Cleaned the data set
- Sanitized the data set.
- Done Exploratory Data Analysis(EDA)
- Watched videos on model training
- Used Linear Regression and trained it
- Used Random Forest and trained it.
- Used XGBoost regression and trained it
- Used Decision tree
- Compare the accuracy of each regression models

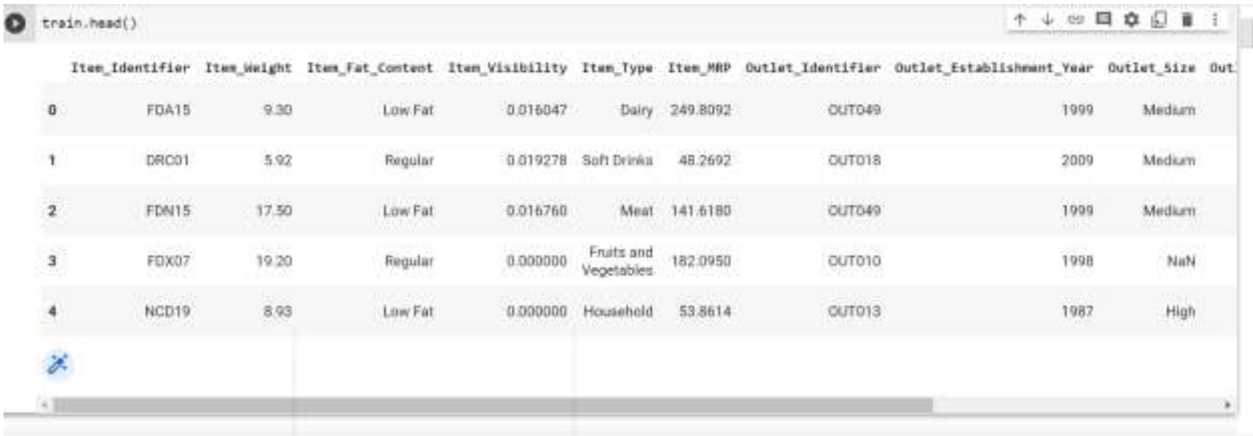
RESEARCH METHODOLOGY

The approach I took for the internship project for completing the milestones is understanding the concepts of the requirements. Reading articles and watching videos helped in achieving knowledge about the requirements.Google Colab has been used for doing the programming. A GitHub account has been created for publishing the codes.

In this paper, we use random forest regressor, Decision tree, Linear Regressor and XG-booster approach to predict sales where data mining techniques such as discovery, data transformation, feature development, model creation and testing are used. In this technique raw data collected by a big mart will be pre-processed for missing data and outlier. An algorithm will then be trained to construct a model on that data. We will use this model to forecast the end results. It is a system in which three functions are combined. It is used to extract and transform the data from one database into an appropriate format.

Dataset Collection

BigMart's data scientists collected sales data of their 10 stores situated at different locations with each store having 1559 different products as per 2013 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 shown in Fig.on using head() function on the dataset



```
train.head()
```

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Out
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	

The data set consists of various data types from integer to float to 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 for numerical variables of our dataset

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 model building.

Exploratory Data Analysis

The exploratory data analysis consists of the descriptive properties of data. Here often using statistical graphics and other data visualization methods for analyzing the data. Graphical methods are usually qualitative and only involve a degree of subjective nature. Even a layman can understand Exploratory Data Analysis. The philosophy behind the EDA is; it is not identical to statistical graphics although the two terms are also used almost interchangeably. Statistical graphics are a collection of techniques that is all graphically based and all focusing on one data characterization aspect. But EDA encompasses a larger venue; EDA is an approach to data analysis that postpones the usual assumptions about what kind of the data follows with the more direct approach of allowing the data itself to reveal its underlying structure and model. EDA is a philosophy as to how we dissect a data set; what we look for; how we look for; and how we interpret it. Exploratory Data Analyzing Techniques Most of the EDA techniques are graphical with few quantitative techniques. The reason for heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore and the graphics give the unparalleled power to do the analyzer, enticing the data to reveal its structural secrets. The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of: Plotting raw data such as data traces, histogram, etc. Plotting simple statistics such as mean plots, standard- deviation plots, box plots and main effects of the raw data. 15 Positioning such plots so as to maximize our natural patterns recognition abilities, such as using multiple plots per page

Modeling

Algorithms employed for predicting sales for this dataset are discussed as follows:

- **LINEAR REGRESSION**: Linear regression algorithm tries to predict the results by plotting the graph between an independent variable and a dependent variable that are derived from the dataset. It is a general statistical analysis mechanism used to build machine learning models. The general equation for linear regression is

$$Z = a + bE$$

Where, Z is the dependent variable and E is independent variable.

- **RANDOM FOREST**:

  Random Forest Algorithm is used to incorporate predictions from multiple decision trees into a single model. This algorithm uses bagging mechanism to create a forest of decision trees. It the incorporates the predictions from multiple decision trees to give very accurate predictions. The Random Forest algorithm has two steps involved

  - Random forest formation.
  - Predict by Random forest classifier generated

- **XG BOOSTER APPROACH:**

  The XG Boost algorithm is developed using Decision trees and Gradient boosting. This algorithm stands on the principle of boosting other weaker algorithms placed in a gradient decent boosting framework. This approach works very accurately beating almost all other algorithms in providing accurate prediction. It can be defined as an extension to Gradient Boosting algorithm. Features of XG Boost are,

  - Parallelized tree building.
  - Efficient handling of missing data.
  - In built cross validation capability.
  - Tree pruning.
  - Cache Awareness.

- **Decision Tree**

  Decision tree is a classifier referred to as a tuple recursive in instant-space. It is a powerful way of multivariable analysis and is a powerful technique for data mining. Applications can be used in various fields, and this approach represents the variables involved in achieving a given purpose and the motives for achieving the target and the methods of execution. Let the objective be denoted as (O) and (Ci) the means of action to be followed and let (M ij) the means of action corresponding to those means, which can be indicated by qi, (i=P1 … Pn), which corresponds to the relationship.[1] n i=1 qi = 1; cuqi > 0 With this algorithm

RESULTS

Exploratory Data Analysis

The following are the charts and diagrams that I have created as part of the EDA and Visualization.

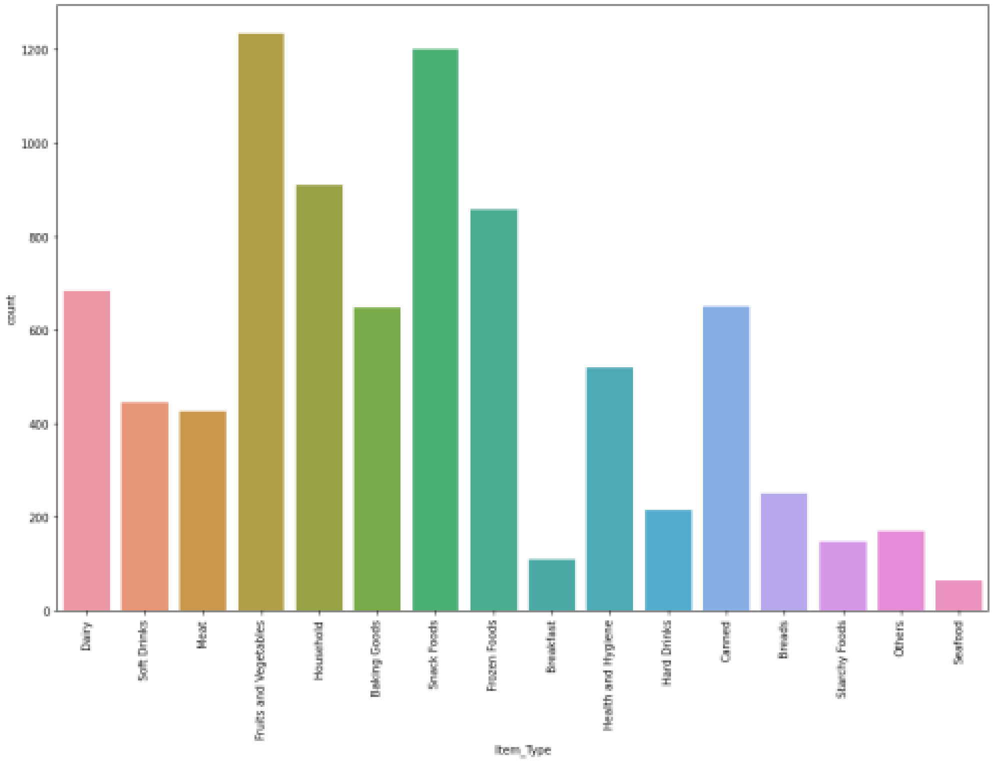


Fig 1 Distribution of Item Type

From the fig 1 its clear that Item Type we have 16 different types of unique values and it is high number for categorical variable

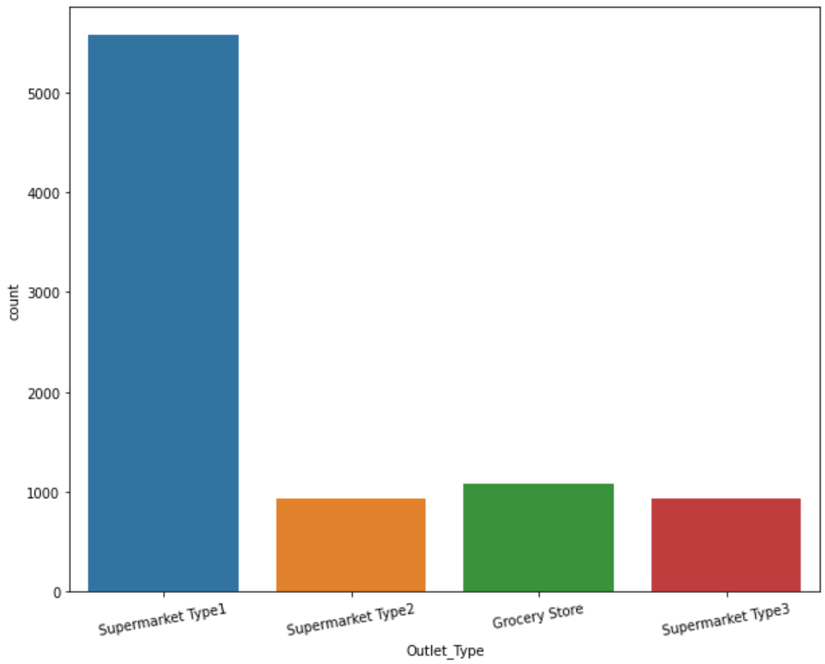


Fig 2 Distribution of variable Outlet Type

There seems like Supermarket Type2 , Grocery Store and Supermarket Type3 all have low numbers of stores

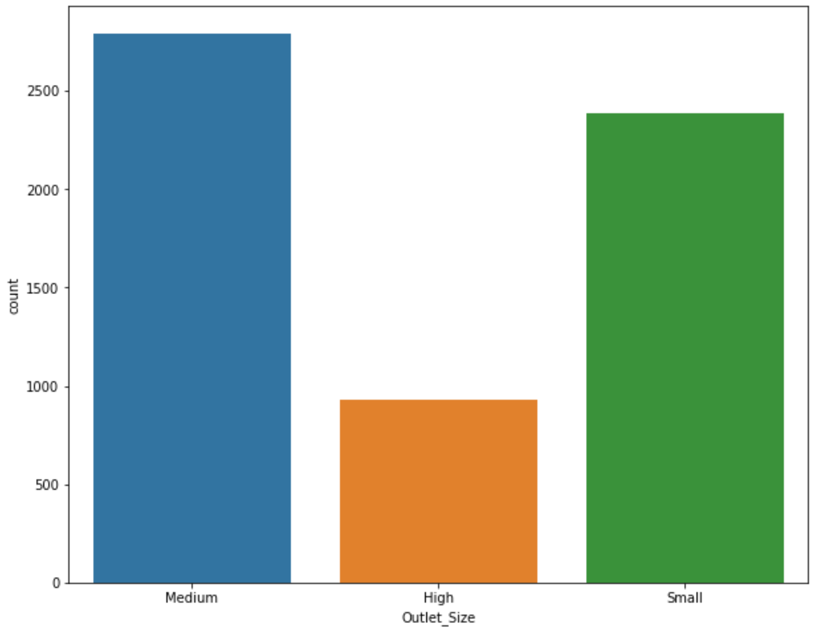


Fig 3 Distribution of variable Outlet Size

There seems to be less number of stores with size equals to “High”. It will be very interesting to see how this variable relates to our target.

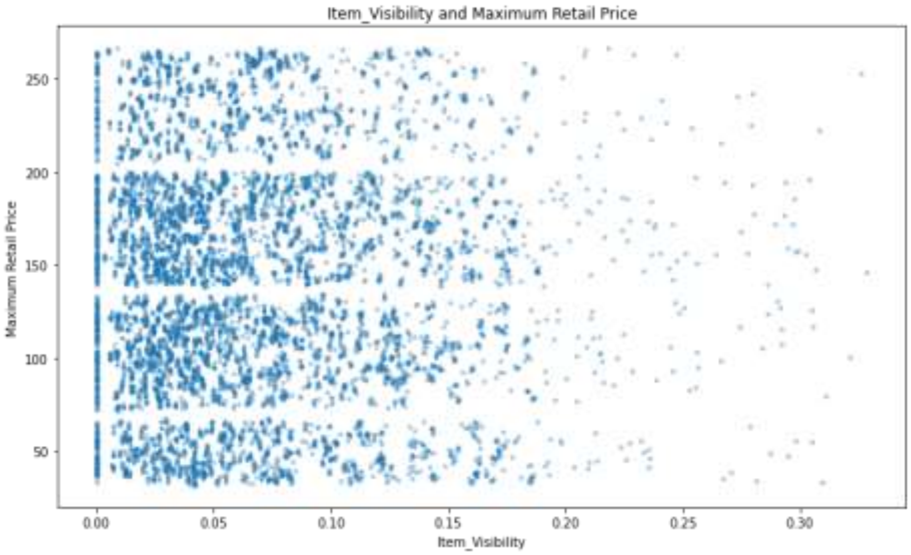


Fig. 4. Correlation between Maximum Retail Price and Item-visibility variable

Fig shows Iteam Visibility has low correlation with the target variable

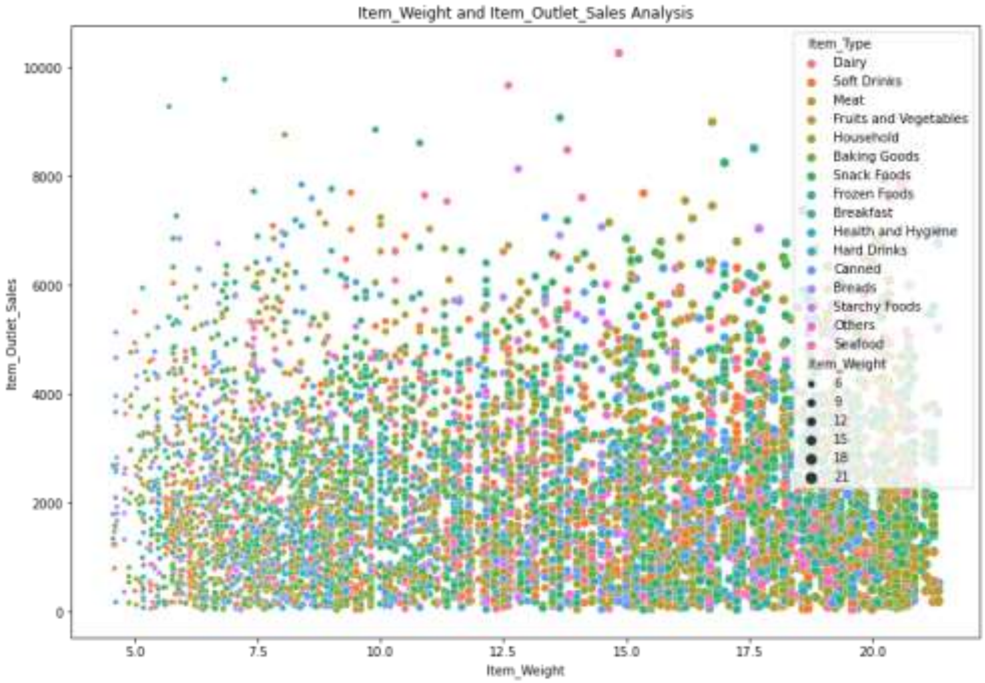


Fig 5 Relationship between Item Outlet Size and Iteam Weight

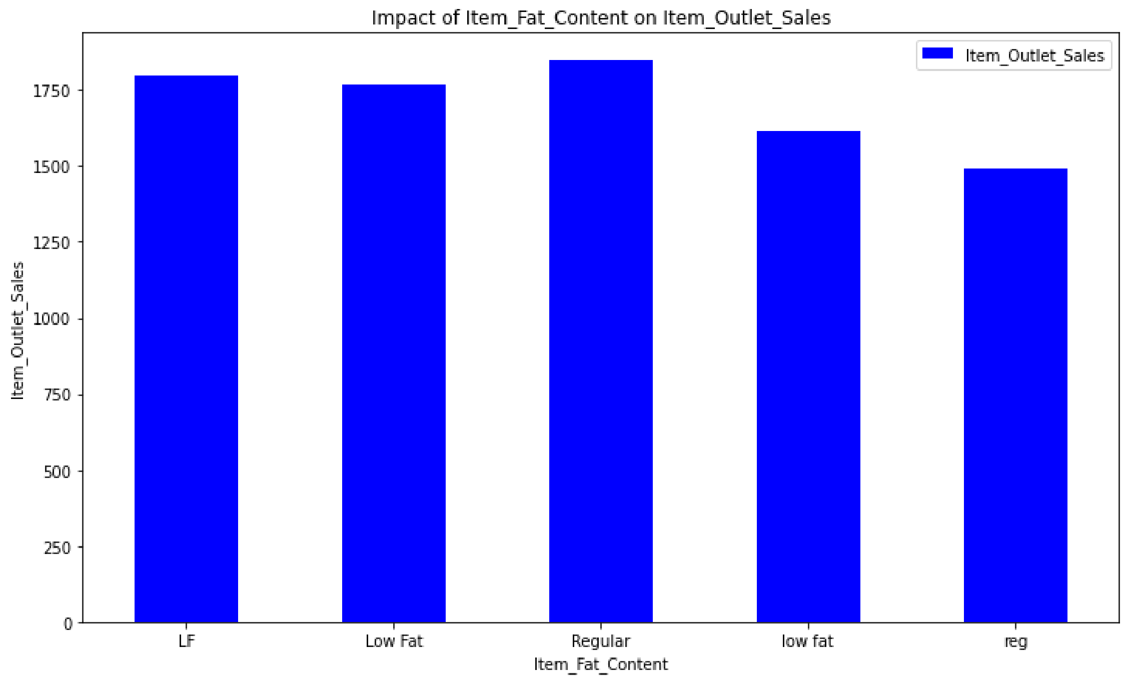


Fig 6 impact of Item content on Item Outlet Sales

Low Fat products seem to higher sales than the Regular products

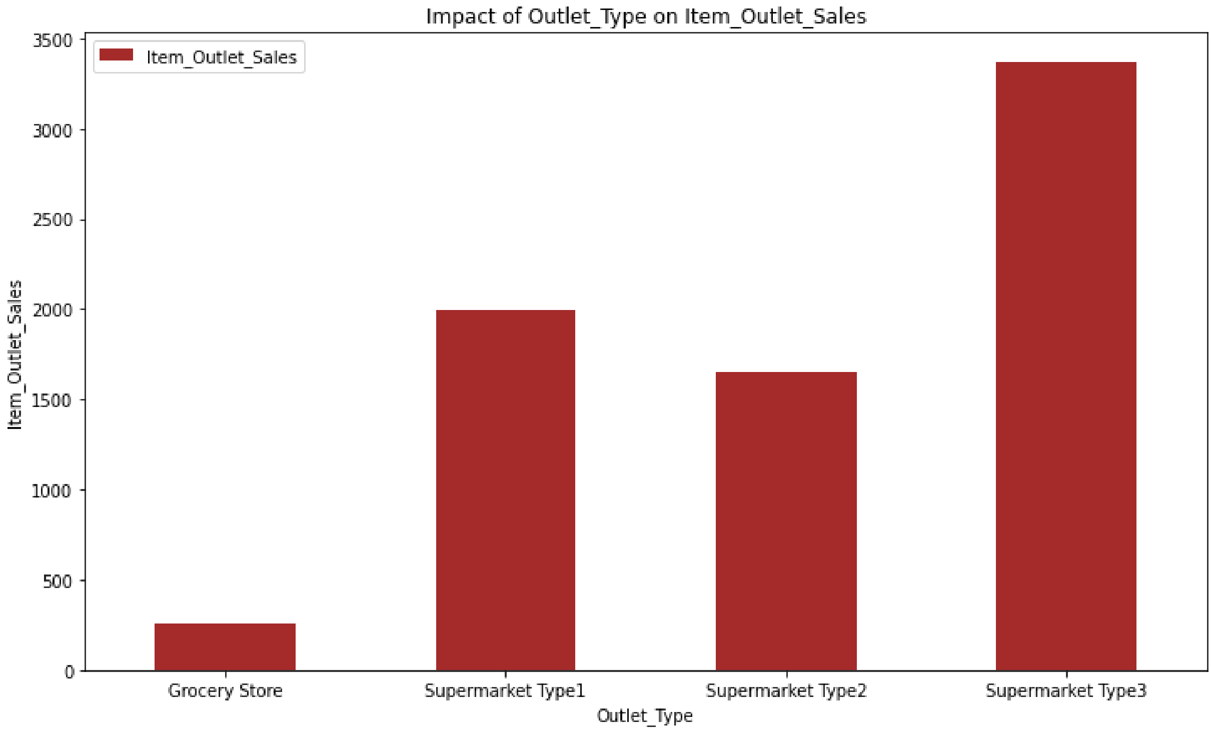


Fig 7 impact of Outlet Type on Item Outlet Sales

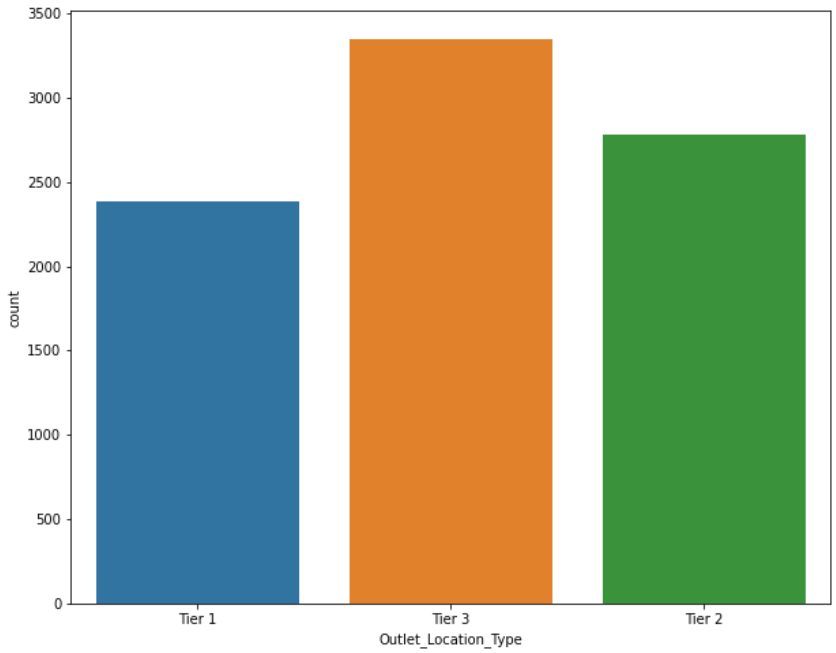


Fig 8 Distribution of variable Outlet Location Type

From the above graph we can see that Bigmart is a brand of medium and small size city compare to densely populated area.

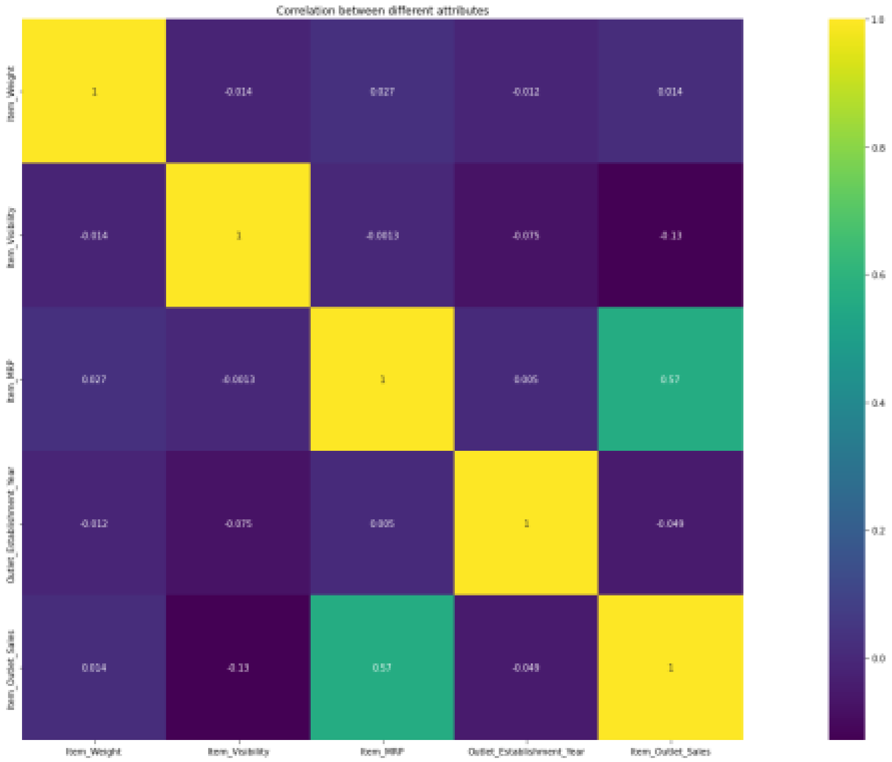


Fig 9: Graph to predict Correlation of variables with target variable.

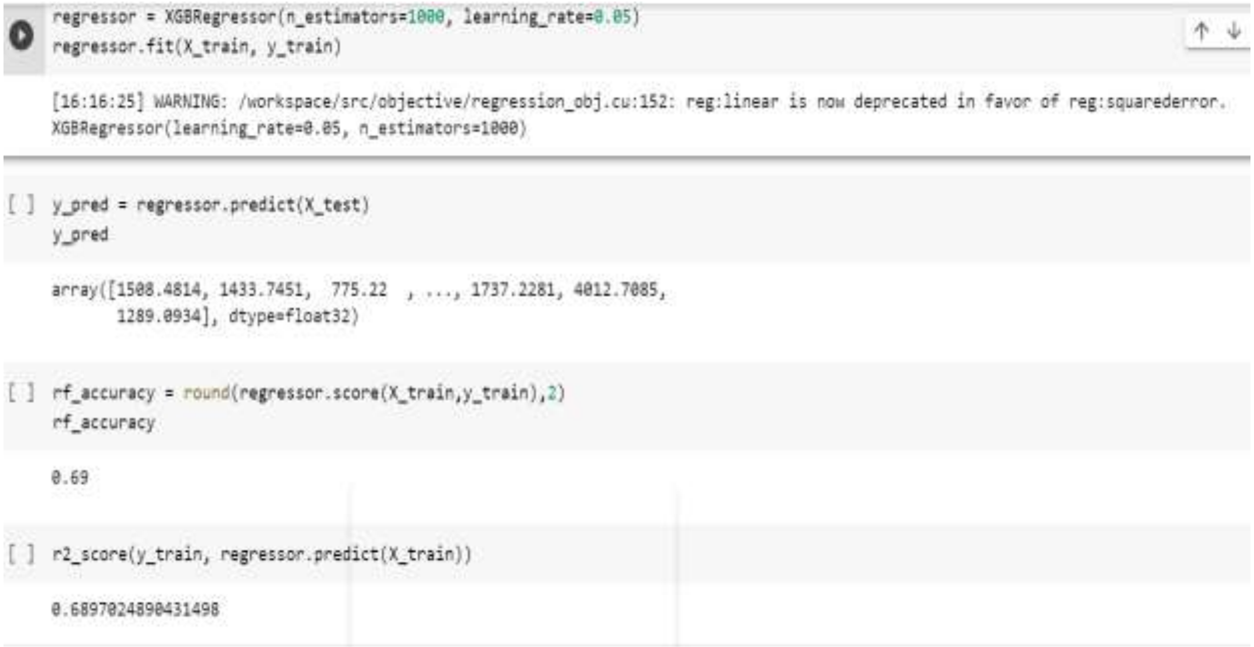
In the Correlation Heat Map, we can observe that the feature with the lowest correlation with our target variable is the Item Visibility. So, the less available the commodity is the higher the price would be in the shop.

MODELING RESULTS

Accuracy of each model obtained is shown in the below table

ALGORITHAM	**ACCURACY**
LINEAR REGRESSOR	56%
DECISION TREE	59%
RANDOM FOREST REGRESSOR	61%
XGBOOSTER REGRESSOR	68%

Adjusted R-squared and R-squared values are higher for XGbooster regression model than average. Therefore, the used model fits better and exhibits accuracy. Also, model accuracy and score of regression model can reach nearly 68% if built with more hypothesis consideration and analysis, as shown by code snippet in the Figure



```
regressor = XGBRegressor(n_estimators=1000, learning_rate=0.05)
regressor.fit(X_train, y_train)
```

```
[16:16:25] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
XGBRegressor(learning_rate=0.05, n_estimators=1000)
```

```
[ ] y_pred = regressor.predict(X_test)
    y_pred
```

```
array([1508.4814, 1433.7451,  775.22  , ..., 1737.2281, 4012.7085,
       1289.0934], dtype=float32)
```

```
[ ] rf_accuracy = round(regressor.score(X_train,y_train),2)
    rf_accuracy
```

```
0.69
```

```
[ ] r2_score(y_train, regressor.predict(X_train))
```

```
0.6897024890431498
```

CHALLENGES & OPPORTUNITIES

During this project, the challenges that I had faced was on handling Outliers and developing a model for the classification. I have the concern on removing outliers in the datasets will reduce size of data and inaccuracy of reults. Also I only had a rough idea of these classification techniques from my academic background. But these daily activities made me much more knowledgeable on these techniques. So I had the opportunities to build a good and conceptual knowledge of these models. Also, I learned about XGBooster Regression Model which I wasn't aware of it earlier.

REFLECTIONS ON THE INTERNSHIP

It was my first internship that I had done in my academic career. So everything was new to me. the digital discussion room helped in connecting various people who are from different backgrounds and cultures. This helped me to develop a systematic approach to doing the project. The activity reports and interim reports helped me to analyze my process and doings. This helped in refurbishing some of the concepts throughout the project.

RECOMMENDATIONS

I felt that much more resources regarding the project can be made available in the project reference part. The reference materials were a bit small. I had to do some research out of the given project references.

OUTCOME / CONCLUSION

The objective of this framework is to predict the future sales from given data of the previous year's using machine Learning techniques. In this paper, discussed how different machine learning models are built using different algorithms like Linear regression, Decision Tree,Random forest regressor, and XG booster algorithms. These algorithms have been applied to predict the final result of sales. We have addressed in detail about how the noisy data is been removed and the algorithms used to predict the result. Based on the accuracy predicted by different models we conclude that the random forest approach and XG Booster approach are best models. Our predictions help big marts to refine their methodologies and strategies which in turn helps them to increase their profit.

ENHANCEMENT SCOPE

This industry project has a wide scope. Using the resume or CV of an individual, one can actually predict the salary. Some of the Natural language processing techniques will help in developing this application

LINK TO CODE

https://colab.research.google.com/drive/1U8JZunKwMJcSOa49TMxxQ0lqqW4J9Cgw?usp=sharing