Detail Project Report (DPR)

Store Sales Prediction

Written By:

Mohamed Afaque Mulla

Version: 1.0

Date: 16-02-2023

Index

1. Introduction

1.1 Abstract

The goal of the project is to develop a system that can forecast future consumer demand for specific products in retail settings. by looking into a product's previous sales data. The sales of products at large shopping centers are tracked in order to predict future demand. Manufacturing and product warehouses are useful for storing a lot of things. The major objective of this study is to examine historical data, identify significant relationships among various characteristics, and develop a system that can make predictions about how much a certain product will be in demand. This technology will assist in controlling the storage capacity of warehouses.

1.2 Machine Learning

The amount of data available is growing daily, and this massive volume of unprocessed data must be carefully evaluated in order to produce outcomes that meet the current standards for being highly useful and finely pure. It is accurate to claim that Machine Learning (ML) is evolving at a rapid rate, just as Artificial Intelligence (AI) has over the previous 20 years. ML is a significant pillar of the IT industry and, as a result, a significant, if typically unnoticed, aspect of our lives. Data is incredibly useful in current aspects, therefore as technology advances, so will the analysis and interpretation of data to produce effective results.

Both supervised and unsupervised types of tasks are dealt with in machine learning, and often a classification-type problem serves as a source for knowledge acquisition. The main focus is on developing a system self-efficient so that it can perform computations and analysis to produce much more accurate and precise results. It creates resources and uses regression to make precise predictions about the future. Data can be transformed into knowledge by applying statistical and probabilistic algorithms. Sampling distributions are used as a conceptual foundation for statistical inference.

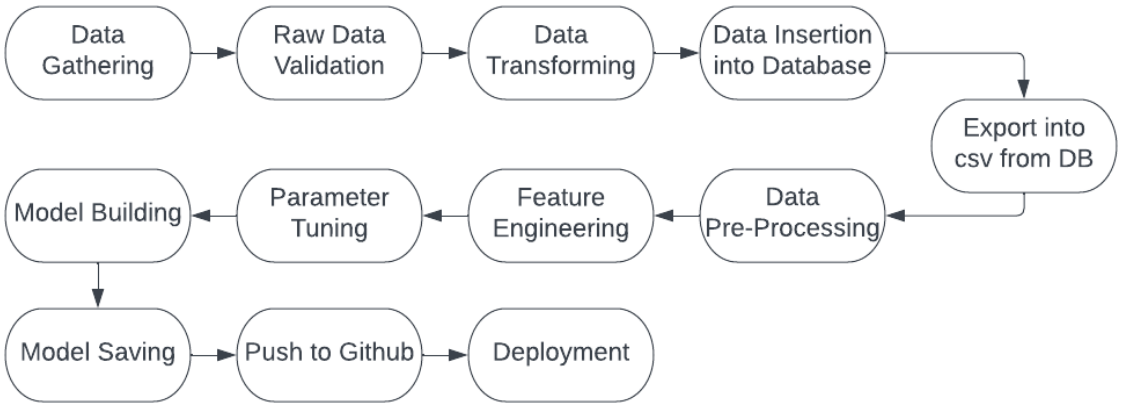
ML can take many different forms. First, numerous ML applications and the kinds of data they work with are explored in this study. The problem statement that is the focus of this work is then codified.

1.3 Problem Statement

Nowadays, shopping malls and Big Marts keep track of individual item sales data in order to forecast future client demand and adjust inventory management. In a data warehouse, these data stores hold a significant amount of consumer information and particular item details. By mining the data store from the data warehouse, more anomalies and common patterns can be discovered.

2. Architecture:

Following workflow was followed during the entire project.



2.2 Data Gathering

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

2.3 Raw Data Validation

Before moving on with any operation, multiple sorts of validation must be performed on the loaded data. Validations include ensuring that all of the columns have a standard deviation of zero and ensuring that no columns have any complete missing data. They are necessary because the qualities that include them are useless. It won't have any impact on how much a product sells at the relevant stores.

For example, if an attribute has zero standard deviation, all of its values are the same and the attribute's mean is zero. This suggests that regardless of whether sales are up or down, the attribute will remain the same. According to this, it serves no purpose to take any property into account when operating if all of its values are missing. Increasing the likelihood of the dimensionality curse is needless.

2.4 Data Transformation

Data transformation is necessary before transferring the data to the database so that it can be transformed into a format that makes database insertion simple. The two properties in this case, "Item Weight" and "Outlet Type," have the missing data. So, they are filled out with supported relevant data types in both the train set and the test set.

2.6 Data Pre-processing

All of the steps necessary before transmitting the data for model development are completed in data pre-processing. For instance, some of the "Item Visibility" characteristics in this instance have values of 0, which is inappropriate given that if an item is available on the market, how can its

visibility be 0? In its place, the average value of item visibility for the relevant "Item Identifier" category has been used. A new characteristic called "Outlet years" was added, which subtracts the current year from the supplied establishment year.

2.7 Feature Engineering

It was discovered after Pre-processing that certain of the attributes are not crucial to the item sales for the specific retailer. Hence, their qualities are dropped. To turn the categorical data into numerical features, even one hot encoding is carried out.

2.8 Parameter Tuning

Randomized search CV is used to fine-tune the parameters. In order to solve this issue, Linear Regression and Random Forest are utilised. These two algorithms' parameters are adjusted and sent to the model.

2.9 Model Building

After executing all kinds of Pre-processing operations indicated above and doing scaling and hyperparameter tweaking, the data set is passed into the models, Lasso Regression and Random Forest Regressor and Gradient Boosting Regressor. It was determined that Gradient Boosting Regressor performs best with the maximum Accuracy value 65%. As a result, the " Gradient Boosting Regressor " did well in this problem.

2.10 Model Saving

Model is saved in ".sav" format using the pickle library. So as to use the model in further applications

3. Data Description

The data scientists at Big Mart gathered sales information from 10 stores spread across various locales, each of which sold 1559 distinct products. It is deduced what function specific attributes of an item play and how they impact sales using all the observations. The following describes the dataset:

```
[6]  1 train.head(5)
```

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

Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
Medium	Tier 1	Supermarket Type1	3735.1380
Medium	Tier 3	Supermarket Type2	443.4228
Medium	Tier 1	Supermarket Type1	2097.2700
NaN	Tier 3	Grocery Store	732.3800
High	Tier 3	Supermarket Type1	994.7052

The data set consists of various data types from integer to float to object as shown in Figure:

```
1 train.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
```

There may be several kinds of underlying patterns in the raw data, which can potentially provide insights into the issue and in-depth knowledge about the subject of interest. However, attention should be exercised when dealing with data because it could include null values, redundant values, or different sorts of ambiguity, which necessitates pre-processing of the data. So, a dataset should be investigated as thoroughly as feasible.

The following table illustrates various statistically significant factors for numerical properties, including mean, standard deviation, median, count of values, maximum value, etc.

To ensure that analysis and model fitting are accurate, preprocessing of this dataset entails performing analysis on the independent variables, such as checking for null values in each column and then replacing or filling them with supported relevant data types. Some of the representations created using Pandas tools, which provide information on model values for categorical columns and variable counts for numerical columns, are displayed above. Deciding which value to priorities for further investigation activities and analysis depends on the maximum and minimum values in numerical columns as well as their percentile values for the median. During the model building process, data types from various columns are also employed for label processing and one-hot encoding.

```
1 train.describe()
```

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	7060.000000	8523.000000	8523.000000	8523.000000	8523.000000
mean	12.857645	0.066132	140.992782	1997.831867	2181.288914
std	4.643456	0.051598	62.275067	8.371760	1706.499616
min	4.555000	0.000000	31.290000	1985.000000	33.290000
25%	8.773750	0.026989	93.826500	1987.000000	834.247400
50%	12.600000	0.053931	143.012800	1999.000000	1794.331000
75%	16.850000	0.094585	185.643700	2004.000000	3101.296400
max	21.350000	0.328391	266.888400	2009.000000	13086.964800

4. Implementation and Results

4.1 Implementation Platform and Language

Python is a general-purpose, interpreted-high level language that is frequently used in modern times to solve domain problems rather than handle system complexity. For programming, it is also known as the "batteries included language." It has a number of libraries utilized for scientific research and inquiries as well as a number of libraries from other parties to facilitate effective issue solutions.

The Python libraries Numpy and Matplotlib have been utilized in this work for scientific computation and 2D visualization, respectively. Moreover, the Python Pandas tool has been used to conduct data analysis. To complete jobs by assembling the random forest approach, utilize the random forest regressor. Jupyter Notebook has been utilized as a development environment because it excels at "literate programming," where human-friendly code is interspersed within code blocks.

4.2 Correlation

```
1 train.corr()
```

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
Item_Weight	1.000000	-0.014048	0.027141	-0.011588	0.014123
Item_Visibility	-0.014048	1.000000	-0.001315	-0.074834	-0.128625
Item_MRP	0.027141	-0.001315	1.000000	0.005020	0.567574
Outlet_Establishment_Year	-0.011588	-0.074834	0.005020	1.000000	-0.049135
Item_Outlet_Sales	0.014123	-0.128625	0.567574	-0.049135	1.000000

- There is almost no association between the dependent variable, Item Outlet Sales, and the type of grocery store outlet. This suggests that product visibility has no impact on sales, which runs counter to the common belief that "greater visibility equals more sales."
- Sales at an outlet have a positive correlation with item MRP (maximum retail price), suggesting that the price that an outlet quote has a significant impact on sales.
- Depending on each outlet's particular sales, different outlets will quote different MRPs.

4.3 Metrics for Data Modelling

The coefficient of determination R2 (R-squared) is a statistic that assesses how well a model fits the data, or how closely the predictions of regression come close to the actual data points. The value 1 of R2 indicates that regression predictions fully match the real data points, and higher values of R2 show stronger model successes in terms of prediction coupled with accuracy. The application of improved R2 measurements produces even better outcomes. The dataset's target column's logarithmic values turn out to be important for the prediction procedure. So, it may be argued that improved results can be determined by adjusting the columns used in the prediction.

Taking the square root of the column could have also been a way to incorporate correction. Also, because the target variable's square root tends to have a normal distribution, it makes the dataset and target variable easier to visualize.

An essential metric during the estimating phase is the measuring of error. For measuring the correctness of continuous variables, the terms Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are frequently employed. It may be claimed that both MAE and RMSE can be used to express the average model prediction error in terms of the variable of interest. The MAE, which gives equal weight to each individual difference, is the average of the absolute disparities between forecast and actual observation over the test sample.

The term "RMSE" refers to the square root of the average of the squared discrepancies between the prediction and the actual observation. R2 is a relative measure of fit, whereas RMSE is an absolute value. RMSE is a quadratic scoring rule that aids in determining the average error of the variable. Better model fitting results from low RMSE values acquired for linear or multiple regression.

As the metric RMSE ratio is estimated to be equal to the ratio between the train and test sample, it can be concluded from the results of this work that there isn't much of a difference between our train and test sample. RMSE is a good measure together with measuring precision and other necessary characteristics, and it may be used to infer findings about how accurately responses are predicted by our model.

More data exploration combined with outlier detection and high leverage points could significantly improve results. Using ensemble learning, which combines multiple low-dimensional sub-models that are simple to verify by subject matter experts, is another strategy that is conceptually simpler.

4.4 Prediction results

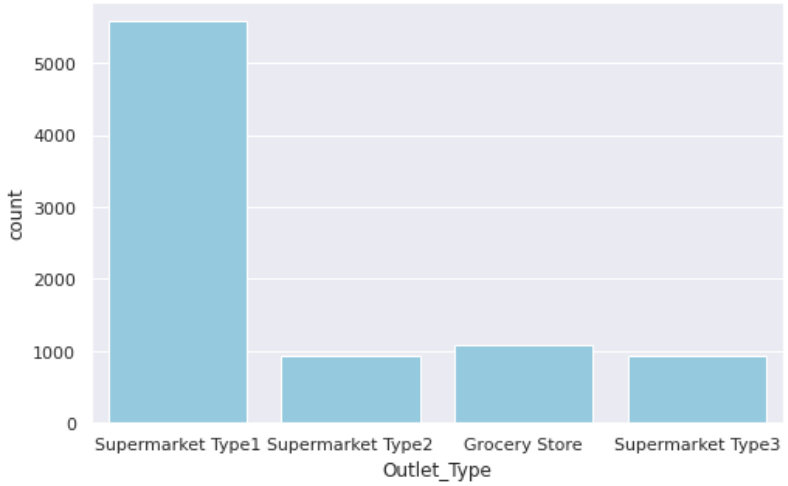
- Not the biggest site generated the most revenue. The location at OUT027, a Supermarket Type3 with a size categorized as medium in our dataset, generated the largest number of sales. It can be said that the performance of this outlet was significantly superior than that of any other outlet site with any size offered in the taken dataset.
- Item Outlet Sales, the target variable, has a median value of 3364.95 for the OUT027 location. A median value of 2109.25 was found at the site (OUT035) with the second-highest median score.
- For the Gradient boost model, adjusted R-squared and R-squared values are higher than usual. Also, compared to other models with the greatest CV score, its RMSE value is low. As a result,

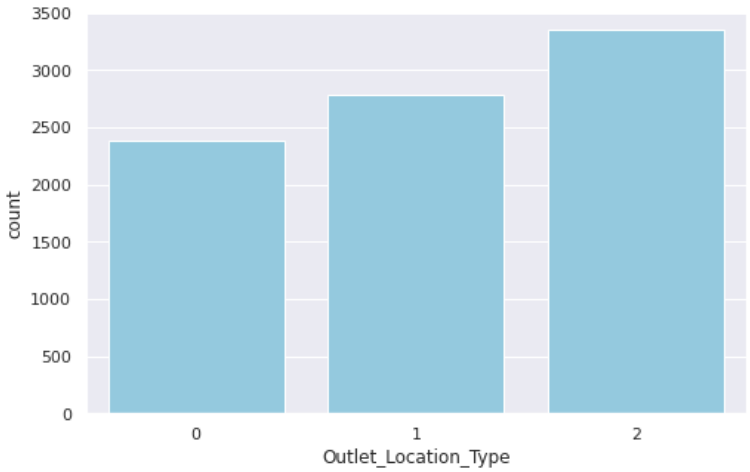
the gradient boost model is more accurate and fits the data better.

5. Conclusion

The fundamentals of machine learning and the related data processing and modelling techniques have been covered in this project, followed by an application of these concepts to the problem of predicting sales in several Big Mart shopping complexes. The predicted results after implementation indicate the relationship between the many factors taken into account and how a specific site of a medium size recorded the best sales, implying that additional retail locations should adopt a similar strategy for increased sales.

It may be determined that more sites should be upgraded to Tier 3 under the "Supermarket Type 3" outlet type in order to boost product sales at Big Mart. This approach can help any one-stop-shopping mall, like Big Mart, by predicting how many of its products will sell in the future at various locations. (Tier1: 0, Tier2: 1, Tier3: 2)





6. Future Scope

- To increase the originality and success of this sales prediction, many instances parameters and other elements can be used. With more factors being employed, accuracy which is crucial in prediction-based systems can be considerably improved. Also, understanding the operation of the sub-models can boost system productivity.
- For increased usability, the project can be further developed in a web-based application or

in any device supported by an in-built intelligence thanks to the Internet of Things (IoT).

- In order to develop more exact results that are closer to actual world circumstances, many stakeholders involved with sales information can also contribute more inputs to aid in hypothesis formulation.
- The old approaches could be observed to have a greater and more beneficial impact on the overall development of a corporation's tasks when paired with efficient data mining methods and features.
- One of the key benefits is that the regression outputs are more expressive and, to a certain extent, more intelligible. Moreover, variations can be added to the suggested strategy to boost its adaptability at a crucial point in the regression model-building process.
- Further experiments are required for accurate resource efficiency measures in order to properly assess and optimize.