Big Mart Sales Prediction

A project report submitted in partial fulfillment of the requirements for the award of the degree of

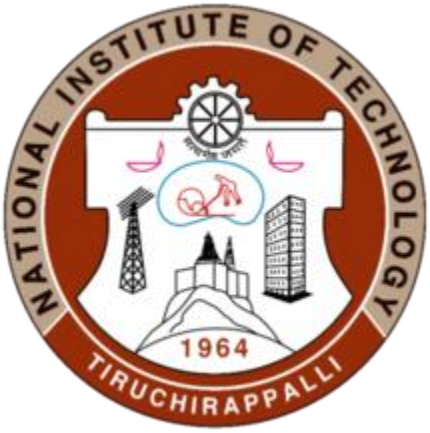
Master of Computer Applications

in

Computer Applications

By

Rajnish Kumar Verma (205121078)



DEPARTMENT OF COMPUTER APPLICATIONS

NATIONAL INSTITUTE OF TECHNOLOGY,

TIRUCHIRAPPALLI 620015

DECEMBER 2023

BONAFIDE CERTIFICATE

This is to certify that the project **"Big Mart Sales Prediction"** is a project work successfully done by

Rajnish Kumar Verma (205121078)

in partial fulfillment of the requirements for the award of the degree of Master of Computer Applications from the National Institute of Technology, Tiruchirappalli, during the academic year 2022-2023 (5th Semester – CA749 Mini Project Work).

Dr. S. Nickolas	Prof. Dr. Michael Arock
Project Guide	Head of the Department

Project viva-voce held on 11/12/2023.

Acknowledgment

Every project, big or small, is successful largely due to the effort of several wonderful people who have always given their valuable advice or lent a helping hand. I sincerely appreciate the inspiration, support, and guidance of all those people who have been instrumental in making this project successful.

I express my deep sense of gratitude to **Dr. G. Aghila**, Director, National Institute of Technology, Tiruchirappalli for giving me an opportunity to do this project.

I am grateful to **Dr. Michael Arock**, Professor and Head, Department of Computer Applications, National Institute of Technology, Tiruchirappalli, for providing the infrastructure and facilities to carry out the project.

I express my gratitude to my Project Guide, **Dr. S. Nickolas**, Professor, Department of Computer Applications, National Institute of Technology, Tiruchirappalli, for his support and arranging the project on a good schedule, and who assisted me in completing the project. I would like to thank him for duly evaluating my progress and evaluating me.

I express my sincere and heartfelt gratitude to **Project Evaluation Committee**, Department of Computer Applications, National Institute of Technology, Tiruchirappalli. I am sincerely thankful for its constant support, care, guidance, and regular interaction throughout my project.

I express my sincere thanks to all the faculty members and scholars of NIT Trichy for their critical advice and guidance to develop this project directly or indirectly.

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 objective of machine learning is to build models and apply 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 trained to meet management expectations and used in a variety of contexts, enabling precise action to be done in order to meet the organization's goal. In this report, the case of Big Mart, a one-stop-shopping- center, has been discussed to predict the sales of different types of items and to understand the impact of several variables on the sales of the products. High accuracy results are achieved by considering multiple elements of a dataset gathered for Big Mart and the process of developing a predictive model. These observations may then be used to inform decisions aimed at increasing sales.

The proposed solution will be based on the dataset:

https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data

“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.

Keywords: Sales prediction, Regression Task

Table of Contents

Title of the Project

Big Mart Sales Prediction

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.

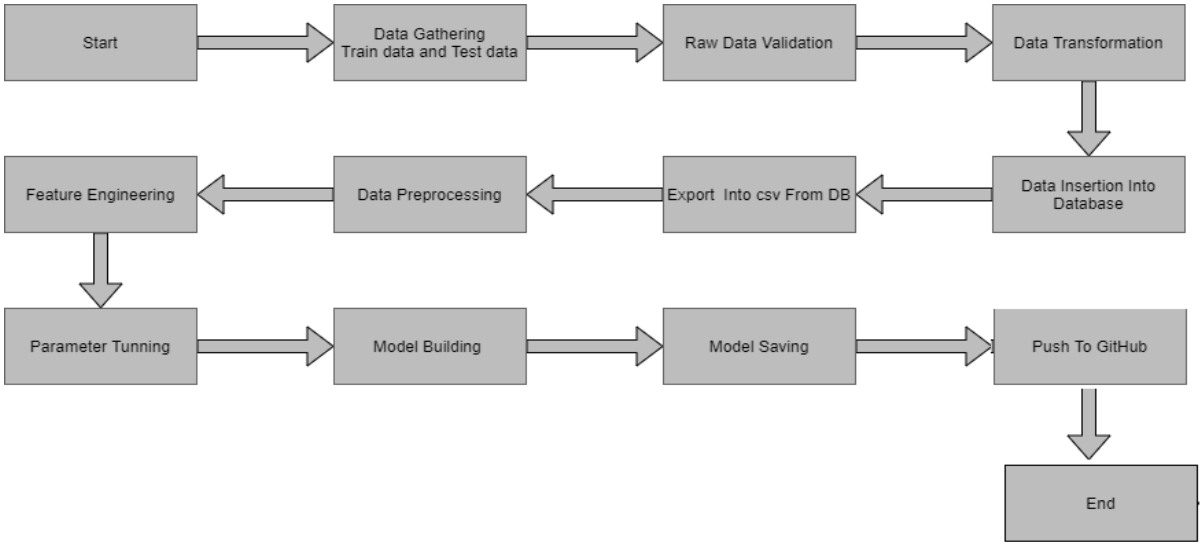
1.3 Problem Statement

"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.

2.4 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'.

2.5 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.

2.6 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.

2.7 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.

2.8 Github

Whole project directory will be pushed into GitHub repository.

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:

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

```
In [5]: df_train.head()
```

Out[5]:

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Loc
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	

```
In [5]: df_train.head()
```

Out[5]:

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 [6]: df_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
```

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.

```
In [7]: df_train.describe()
```

Out[7]:

	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

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 Result

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 Correlations

- 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

4.5 Implementation Details

Software

To implement experimentation, we've used the following technologies. A brief description of used technologies along with versions is represented below.

- Python - Python is an interpreted, high level and object-oriented programming language. It is an open source.
- Jupyter notebook - It is a web-based interactive computing platform. It helps in developing, documenting, and executing code.
- Pandas - It is a free open-source python library. It is mainly used for data analysis. It helps to perform various data manipulation operations.
- Numpy - It is a python library that is used for scientific computing in python. It is used to perform wide mathematical operations on data.
- Matplotlib - It is a Python package used to create static, animated, and interactive visualizations.
- Seaborn - It is a python library that is used for data visualization. It is based on the matplotlib library. It helps to make statistical graphics using python.
- Scikit-learn - It is a python library that is used for machine learning. It consists of many machine learning algorithms, best suitable for predictive analysis.

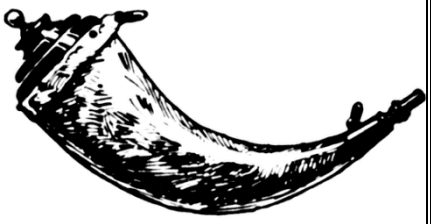
Hardware

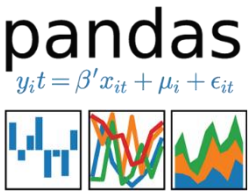
- Hard Disk 1 TB
- Processor Intel i5 8th gen
- RAM 8 GB

Tool Used

The programming language is Python that is used here, also we will use some other python-based libraries like, for ml, we will use Scikit-Learn library, for data manipulation we will use pandas, for numerical computation Numpy, for custom APIs creation Flask web frameworks. Visual Studio Code is used as python IDE for

all modular coding and custom APIs creation. And storing all code files for publically available we will use GitHub.

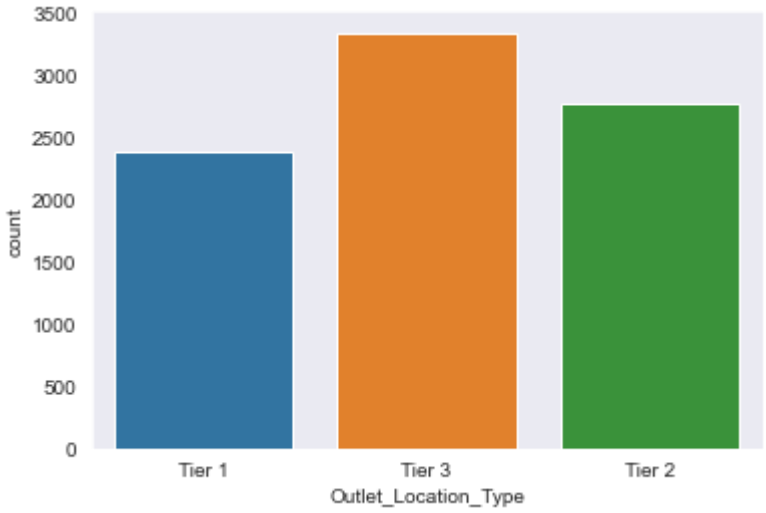


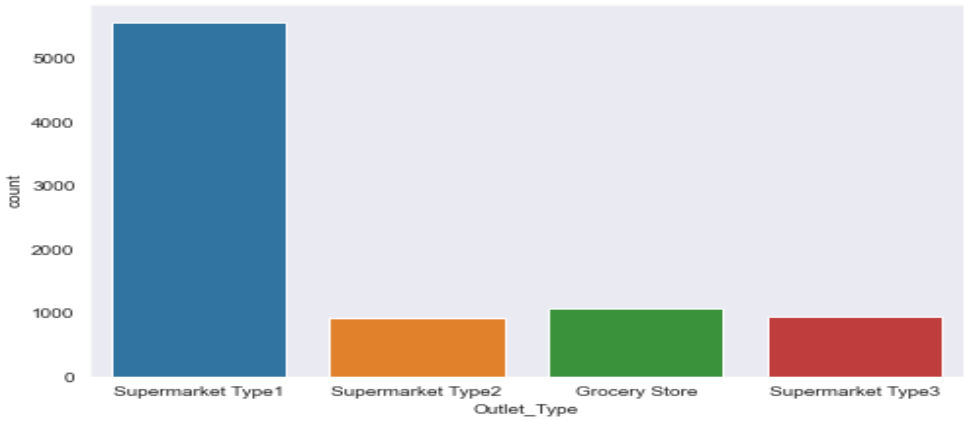


5. CONCLUSION

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 Work

1. 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
2. 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. References

1. Book: Hands on Machine learning with with scikit Learn Keras and Tensor Flow 2nd Edition – Aurelien Geron.

2. Abstract: https://ijarcce.com/papers/big-mart-sales-prediction-using-machine-learning/

Github Link : https://github.com/erajneeshverma/Store-Sales-Prediction