International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER
SSRN

IJIEMR Transactions, online available on 09th Feb 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=ISSUE-02

DOI: 10.48047/IJIEMR/V12/ISSUE 02/43

Title **BIG MART SALES PREDICTION USING DATA ANALYSIS**

Volume 12, Issue 02, Pages: 277-289

Paper Authors
Dr.K.V. Sambasivarao, N.Harika, G.Kundana Sai, D.Ranjitha, Ch.Sathwika

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

BIG MART SALES PREDICTION USING DATA ANALYSIS

Dr.K.V. Sambasivarao[1]| N.Harika[2] | G.Kundana Sai[2]| D.Ranjitha[2]| Ch.Sathwika[2]

[1]Dean & Professor, Department of CSE, NRI Institute of Technology, Pothavarappadu(v)-521212, A.P., India. kvsrao@nriit.edu.in

[2,3,4,5] Final year students of Department of CSE, NRI Institute of Technology, Pothavarappadu(v)-521212, A.P., India.

namburuharika@gmail.com, sirichowdary.2401@gmail.com dasariranjita@gmail.com , gajula.kundanasai@gmail.com

ABSTRACT

Sales forecasting is an important aspect of different companies engaged in retailing, manufacturing, marketing and wholesaling. It allows companies to efficiently allocate resources, to estimate achievable sales revenue and to plan a better strategy for future growth of the company. Everybody wants to know how to buy goods cheaper or how to advertise them at low cost. Here is the answer. That is Big Mart. The goal is to make efficient marketing solutions for sellers. The ultimate idea is to prosper with the customers. This can be done based on hypothesis that should be done before looking at the data but its difficult time taking .The project "Big Mart Dataset" aims to build a predictive model and find out the sales of products at a particular store. Big mart will use this model to understand the properties of products and stores which play a key role in increasing sales.

1. INTRODUCTION

With the rapid development of global malls and stores chains and the increase in the number of electronic payment customers, the competition among the rival organizations is becoming more serious day by day. The growth of international malls and online shopping has led to an increase in the severity and acrimony of the competition between numerous shopping malls and massive supermarkets. Each organization is trying to attract more customers using personalized and short-time offers which makes the prediction of future volume of sales of every item an important asset in the planning and inventory management of every organization, transport service, etc.., in order to efficiently draw a big number of customers and determine the number of sales for each product, as well as for the business' logistics, distribution, and stock management requirements. The current machine learning is highly sophisticated and offers opportunities for forecasting or forecast demand for any type of organization in order to defeat low-cost prediction methods. For creating and enhancing market-specific marketing strategies, projections that are regularly updated are crucial.

Each Item is tracked for its shopping centers and Big Mart in order to anticipate a future demand for the customer and also improve the management of its inventory. Big Mart is an immense network of shops virtually across many places. Trends in Big Mart are very relevant and data

scientists evaluate those trends per product and store in order to create potential centers. Using the machine to forecast the transactions of Big Mart helps to test the patterns by store and product to achieve correct results. Many companies rely heavily on the knowledge base and need market patterns to be forecasted. Each shopping center and store endeavors to give the individual and present moment proprietor to draw in more clients relying upon the day, with the goal that the business volume for everything can be evaluated for organization stock administration, logistics and so forth.

Always better vaticination is helpful, both in developing and perfecting marketing strategies for the business, which is also particularly helpful. But not all machine-learning techniques are equal, and not all of them are equally accurate. As a result, a machine-learning algorithm may be extraordinarily effective when applied to a particular problem but ineffective when applied to another. Due to the cheap availability of computing and storage, it has become possible to use sophisticated machine learning algorithms for this purpose. So, Big Mart requires combining several machine-learning algorithms to produce a useful predictive model and projecting revenue with analytics. In order to find the most powerful predictive analytics .The working prototype of a machine learning-based sales forecasting system for Big Mart will be created. We must test the algorithm on Big Mart before launching this prototype. Genuine data from Mart has been collected by data scientists in the year 2013. The data is thereafter refined in order to get accurate predictions and gather new as well as interesting results with respect to the tasks data.

To address the issue of deals expectation of things dependent on client's future requests in various Big Mart across different areas diverse Machine Learning algorithms like Linear Regression, Random Forest, Decision Tree, XGBoost are utilized for gauging of deals volume. Deals foresee the outcome as deals rely upon the sort of store, populace around the store, a city wherein the store is located i.e., it is possible that it is in an urban zone or country and many more things should be considered. Because every business has strong demand, sales forecasts play a significant part in a retail centre. A stronger prediction is always helpful in developingand enhancing corporate market strategies, which also help to increase awareness of the market.

EXISTING SYSTEM

With the rapid development of global malls and stores chains and the increase in the number of electronic payment customers, the competition among the rival organizations is becoming more serious day by day. Each organization is trying to attract more customers using personalized and short-time offers which makes the prediction of future volume of sales of every item an important asset in the planning and inventory management of every organization, transport service, etc. Due to the cheap availability of computing and storage, it has become possible to use sophisticated machine learning algorithms for this purpose.

The existing system is built using Support Vector Machine and Linear Regression. But building the system using SVM and linear regression does not give much accuracy. Though there are various ways to build a model, building it with optimistic method helps prevent the issues. As this system can be useful in real time, confining it with these models cannot help always in all circumstances. Thus these techniques are to be replaced with the one that gives best results. The existing system

may become slower when the data size becomes larger. It also doesn't work well with noisy data and missing values.

Thus the basic idea behind boosting algorithms is building a weak model, making conclusions about the various feature importance and parameters, and then using those conclusions to build a new, stronger model and capitalize on the misclassification error of the previous model and try to reduce it. The model is effective with large dataset and capable of handling missing values. The default base learners of XGBoost are tree ensembles. The tree ensemble model is a set of classification and regression trees (CART). Trees are grown one after another, and attempts to reduce the misclassification rate are made in subsequent iterations

PROPOSED SYSTEM

The data scientists at Big Mart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store. Using this model, Big Mart will try to understand the properties of products and stores which play a key role in increasing sales.

We make use of XGboost for the sake of improvement in accuracy. XGboost is a Machine Learning algorithm that deals with structured data, and uses the gradient boosting framework at its core. Boosting is a sequential technique which works on the principle of an ensemble. It combines a set of weak learners and delivers improved prediction accuracy.

3.3. METHODOLOGY

3.3.1. Linear Regression:

Build a fragmented plot as a linear or non-linear pattern of data and a variance (outliers). Consider a transformation if the marking isn't linear. If this is the case, outsiders, it can suggest only eliminating them if there is a non-statistical justification. Link the data to the least squares line and confirm the model assumptions using the residual plot and the normal probability plot .A transformation might be necessary if the assumptions made do not appear to be met.

3.3.2. Random Forest:

The ability of the Random Forest Algorithm to handle data sets containing continuous variables, as in the case of regression, as well as categorical variables, as in the case of classification, is one of the most essential characteristics of this algorithm. When used to classification issues, it provides superior performance results.

3.3.3. XGBoost:

Extreme Gradient Boosting is same but much more effective to the gradient boosting system. It has both a linear model solver and a tree algorithm. Which permits xgboost in any event multiple times quicker than current slope boosting executions. It underpins various target capacities, including relapse, order and rating. As "xgboost" is extremely high in prescient force however generally delayed with organization, it is appropriate for some rivalries.

3.4. IMPLEMENTATION

For building a model to predict accurate results the dataset of Big Mart sales undergoes several sequence of steps and in this work we propose a model using Xgboost technique. Every step plays a vital role for building the proposed model. After preprocessing and filling missing values, we used ensemble classifier using Decision trees, Linear regression, Random forest and Xgboost. RSME is used as accuracy metrics for predicting the sales in Big Mart. From the accuracy metrics it

was found that the model will predict best using minimum RSME.

3.5. ADVANTAGES:

- This is an easily scalable model to provide detailed information and accurate predictions for sales volume for different types of products as there is a lot of data out there.
- It is the percentage of display space in a store given to that particular item. Looking at the average visibility of items given in each store type and outlet.

3.6. GOALS:

- Replacing the Nans, identifying outliers, feature selection and normalization – for both training and testing data.
- Building the regression models: linear, decision tree, random forest and XGboost. Predicting the sales, cross validating the scores, calculating the Root Mean Square Error (RMSE).
- Classifying the training data with a decision tree and a random forest and calculating the accuracy score.

INPUT DESIGN AND OUTPUT DESIGN

INPUT DESIGN:

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

10.2. OBJECTIVES:

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.
2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.
3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

10.3. OUTPUT DESIGN:

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
2. Select methods for presenting information.
3. Create document, report, or other formats that contain information produced by the system.

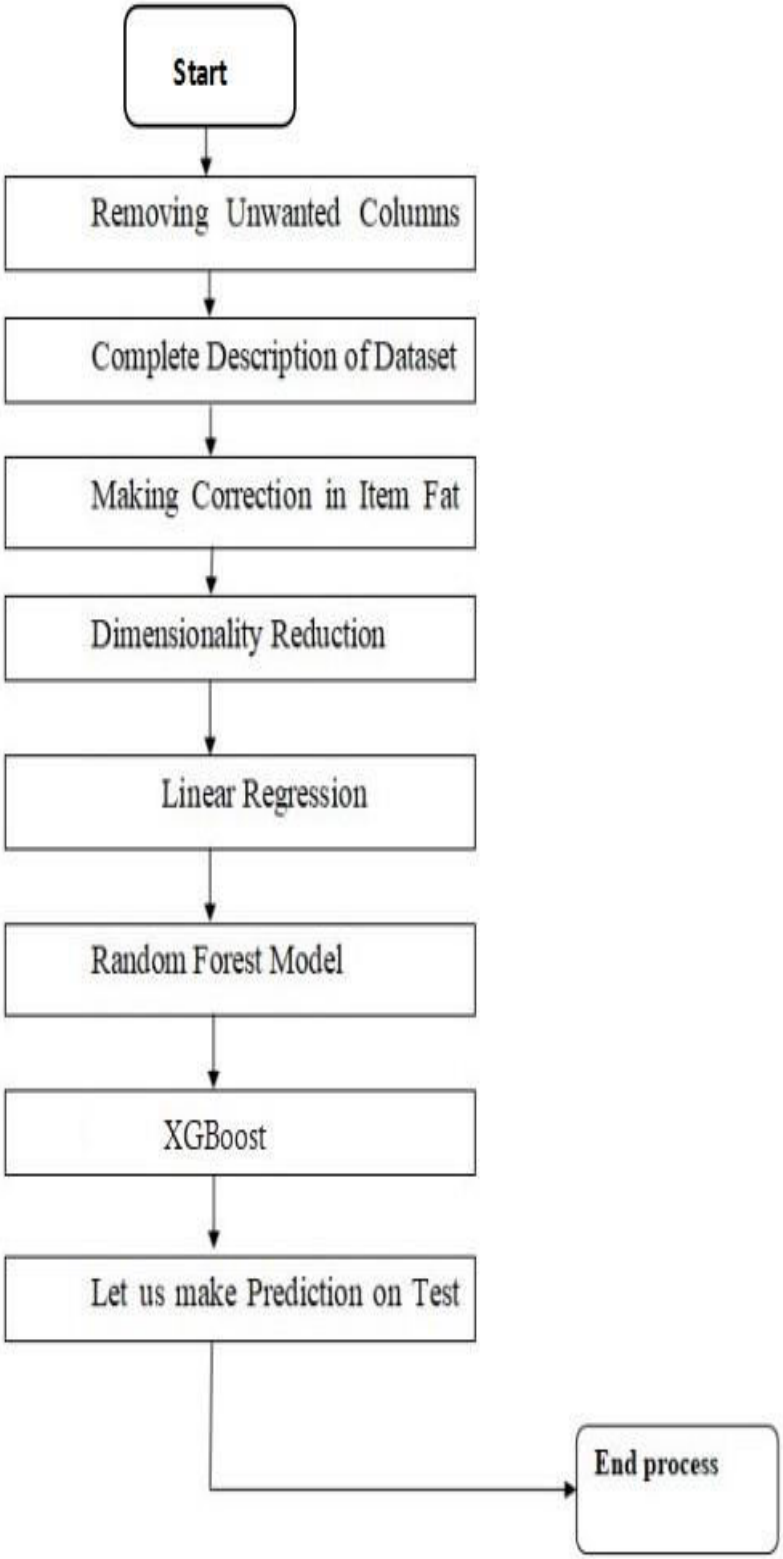
The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

DATA FLOW DIAGRAM:

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modelling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org



International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

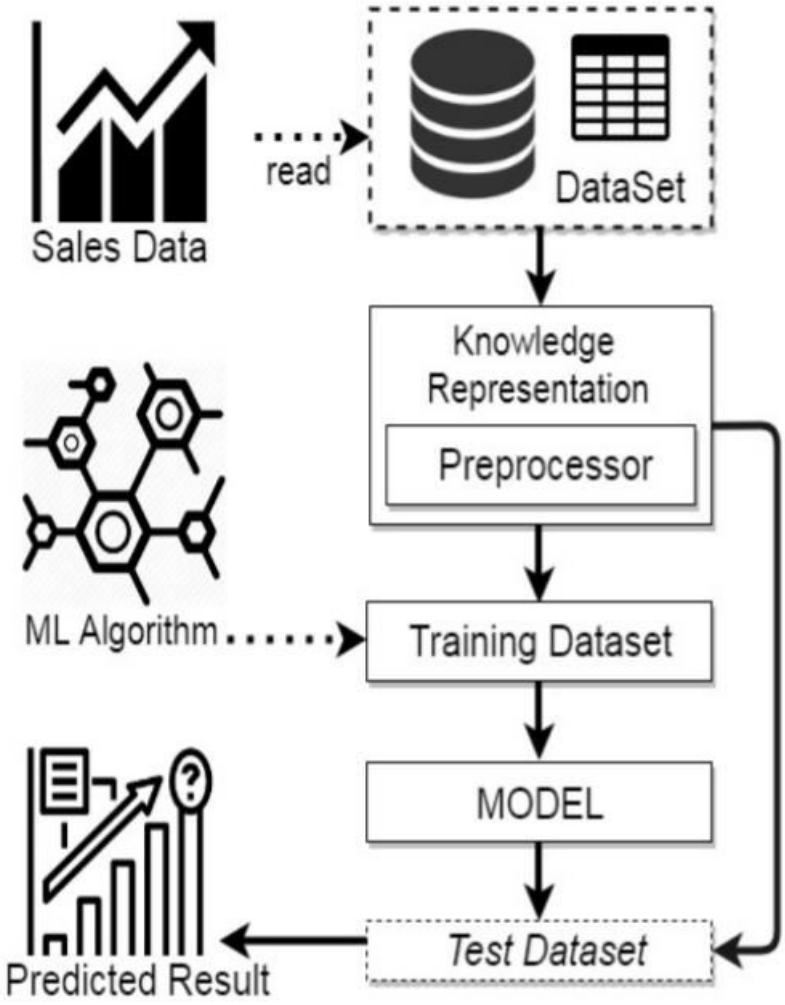


Figure :1 System architecture

RESULTS:

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

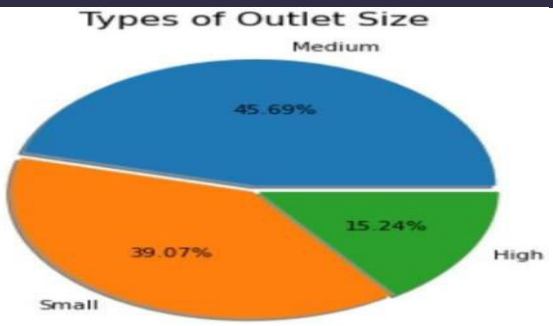


Fig- 2. Types of Outlet Size.

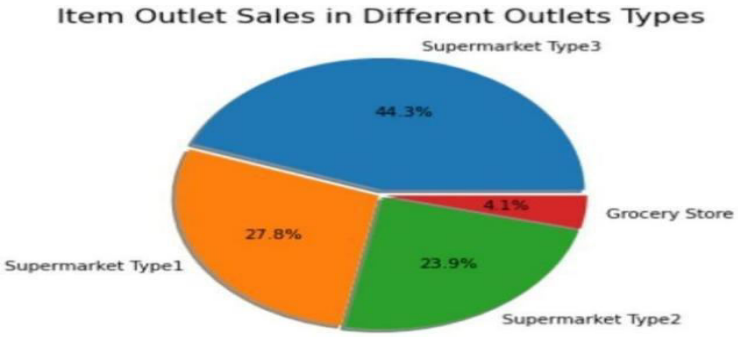


Fig-2. Sales in Different Outlet Types.

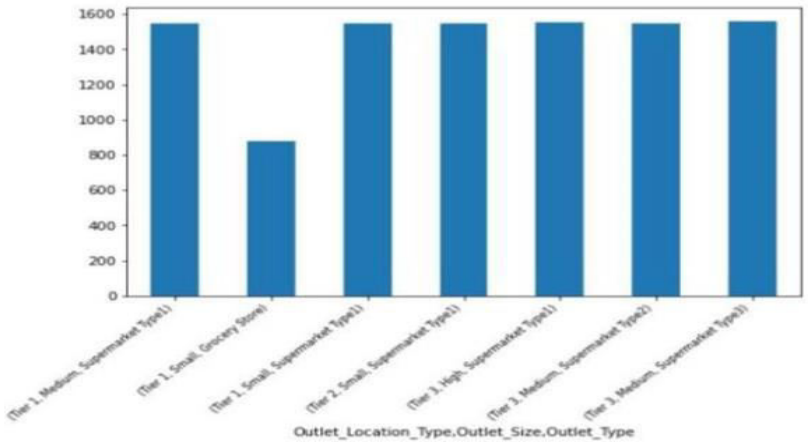


Fig-3. Grouping attributes.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

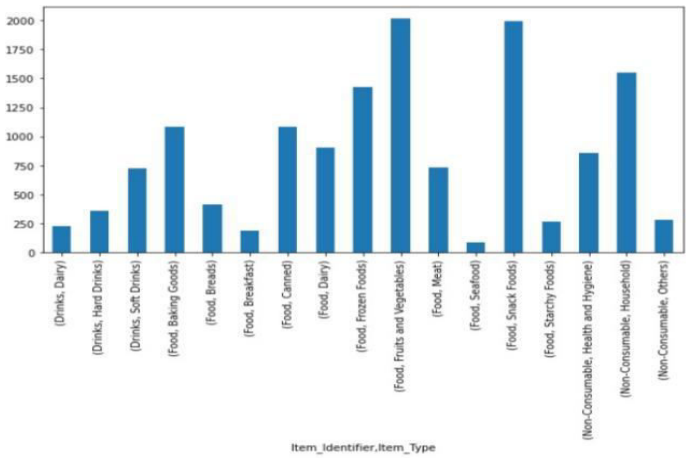


Fig-4. Market Basket Analysis.

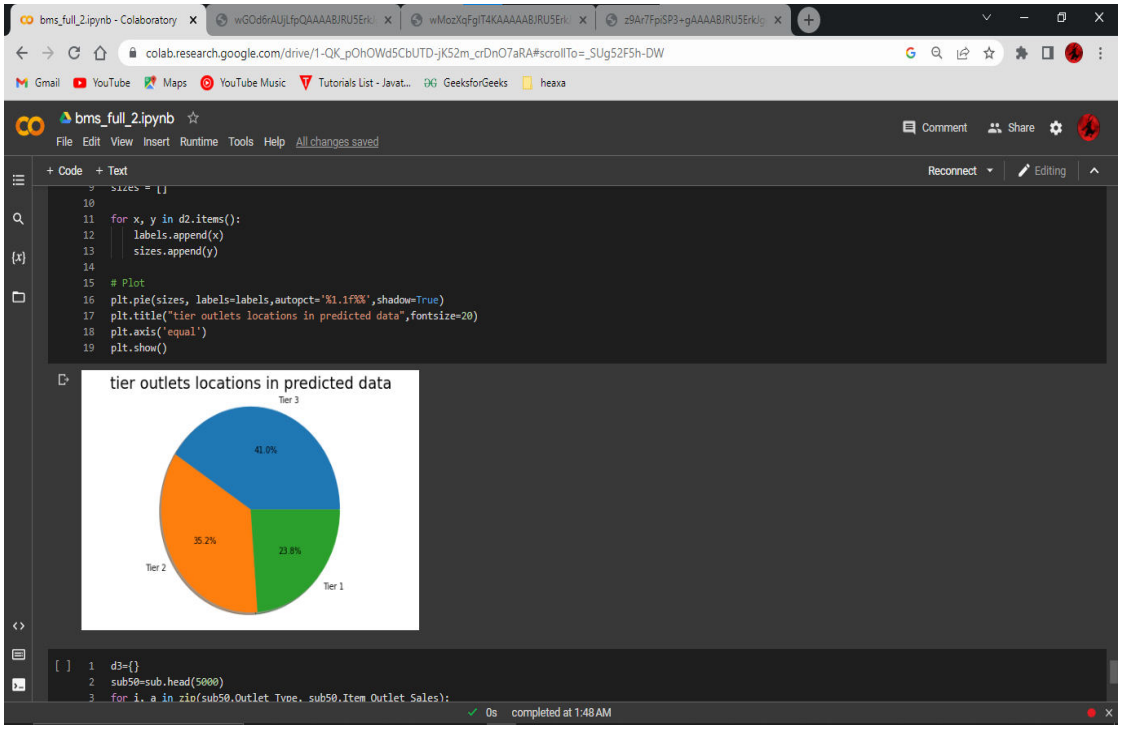


Fig- 5. Tier outlet location in predicted data-1

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

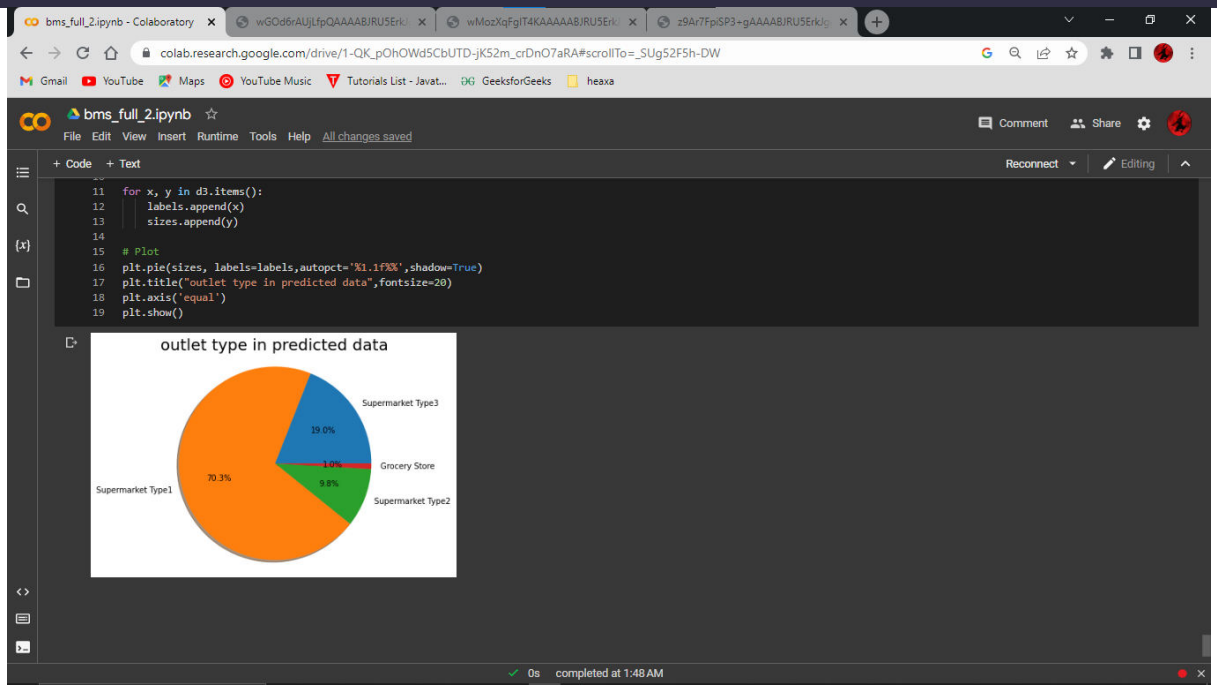


Fig- 6. Outlet Type in Predicted data.

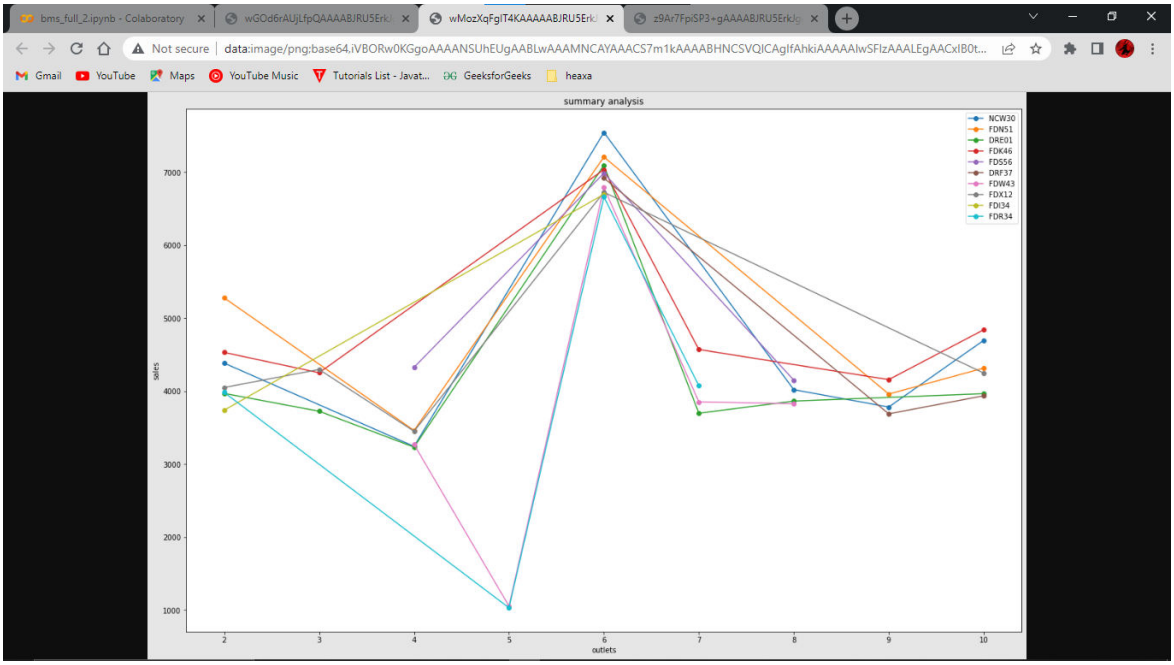


Fig-7. Top 10 Products (Sales Vs Outlet)

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

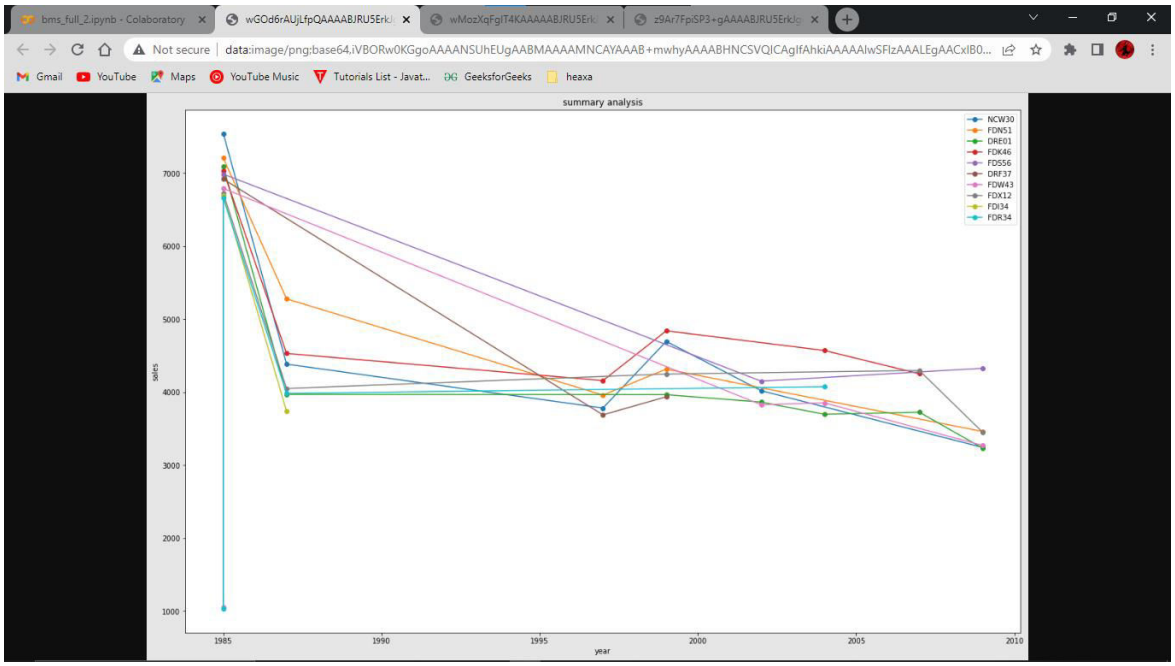


Fig-8. Top 10 Products yearly Sales (Sales Vs Year).

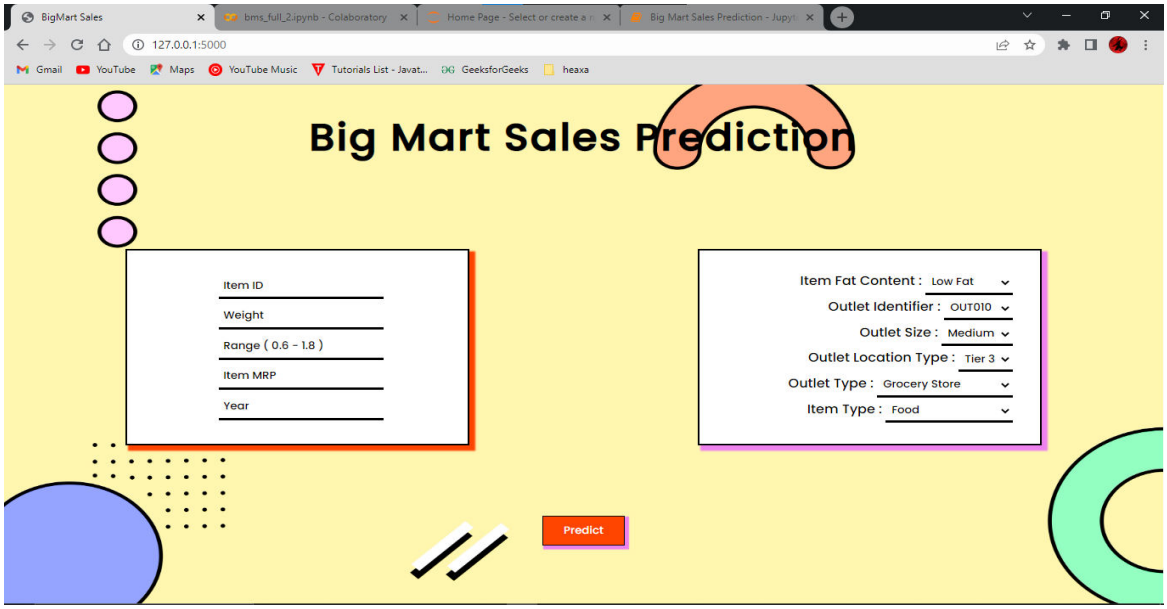


Fig-9. Home Page

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

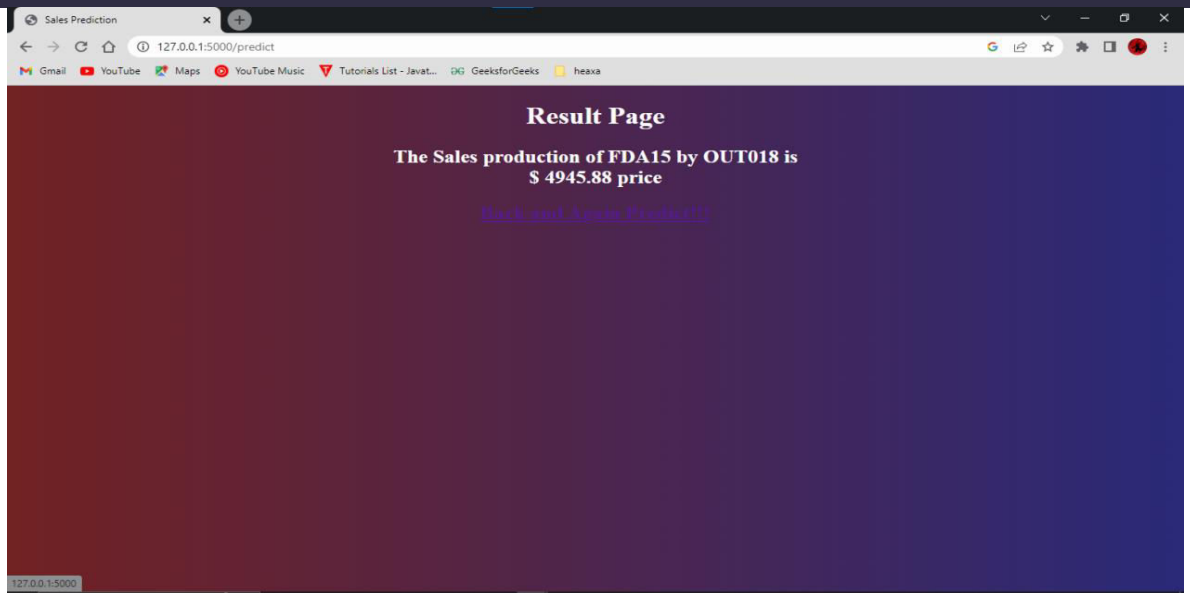


Fig- 10. Result Page

Conclusion

In this paper, basics of machine learning and the associated data processing and modelling algorithms have been described, followed by their application for the task of sales prediction in Big Mart shopping centres 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. 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.

REFERENCES

1) Smola, A., & Vishwanathan, S. V. N. (2008). Introduction to machine learning. Cambridge University, UK, 32, 34.

2) Saltz, J. S., & Stanton, J. M. (2017). An introduction to data science. Sage Publications.
3) Shashua, A. (2009). Introduction to machine learning: Class notes 67577. arXiv preprint arxiv:0904.3664.
4) MacKay, D. J., & Mac Kay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge university press.
5) Daumé III, H. (2012). A course in machine learning. Publisher, ciml. info, 5, 69.
6) Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.
7) Cerrada, M., & Aguilar, J. (2008). Reinforcement learning in system identification. In Reinforcement Learning. IntechOpen.
8) Welling, M. (2011). A first encounter with Machine Learning. Irvine, CA.: University of California, 12.
9) Learning, M. (1994). Neural and Statistical Classification. Editors D. Mitchie et. al, 350.
10) Mitchell, T. M. (1999). Machine learning and data mining. Communications of the ACM, 42(11), 30-36.
11) Downey, A. B. (2011). Think stats. " O'Reilly Media, Inc.".
12) Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.
13) Nikita Malik, & Karan Singh (2020), Sales Prediction Model for Big Mart, Vol 3. Issue 1.
14) Aditi Narkhede , Mitali Awari , Suvarna Gawali & Amrapal Mhaisgawali , Big Mart Sales Prediction Using Machine Learning Techniques.
15) Kumari Punam, Rajendra Pamula, Praphula Kumar Jain (2018), A Two-Leval Statistical Model for Big Mart Sales Prediction.
16) Sunitha Cheriyan, Shaniba Ibrahim, Saju Mohanan, Susan Treesa (2019), Intelligent Sales Prediction Using Machine Learning Techniques .