IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

Best Selling Product and Category Prediction Using Sales Analysis

Ms. Archana Nikose[1], Tejal Mungale[2], Minal Shelke[3], Rohini Shelote[4], Priyal Solanke[5]
Assistant Professor, Department of Computer Science & Engineering[1]
UG Students, Department of Computer Science & Engineering[2,3,4,5]
Priyadarshini Bhagwati College of Engineering, Nagpur, Maharashtra, India

Abstract: *A sales analysis is a detailed report that tells about more profound understanding of a business's sales performance, customer data, and the revenue. This tells you which deals are worth chasing and which are better left behind. Also, for the deals your sales team does decide to pursue, they'll have a good approach ready to make the lead or customer more receptive to the sale. Using Sales Analysis helps to take retailers towards profit in this world of competition. Nowadays shopping malls keep the track of their sales data of each and every individual item for predicting future demand of the customer and update the inventory management as well. These data stores basically contain a large number of customer data and individual item attributes in a data warehouse. Further, anomalies and frequent patterns are detected by mining the data store from the data warehouse. The resultant data can be used for predicting future sales volume with the help of different machine learning techniques for the retailers like Big Mart. A predictive model is build using different algorithms. In this paper, we investigate forecasting sales for a Big Mart, with four machine learning algorithms (Random Forest, Linear Regression, Decision Tree and XG Booster Algorithms). The results show that the Random Forest algorithm performs better than the other two models.*

Keywords: Sales Analysis, Machine Learning, Random Forest, Linear Regression, Decision Tree Regression, XG Booster etc.

I. INTRODUCTION

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. In this project, we are providing forecast for the sales data of big mart in a number of big mart stores across various location types which is based on the historical data of sales volume.

Sales forecasting has always been a very significant area to concentrate upon. An efficient and optimal way of forecasting has become essential for all the vendors in order to sustain the efficacy of the marketing organizations. Manual infestation of this task could lead to drastic errors leading to poor management of the organization, and most importantly would be time consuming, which is something not desirable in this expedited world. A major part of the global economy relies upon the business sectors, which are literally expected to produce appropriate quantities of products to meet the overall needs. Targeting the market audience is the major focus of business sectors. It is therefore important that the company has been able to achieve this objective by employing a system of forecasting. The process of forecasting involves analyzing data from various sources such as market trends, consumer behavior and other factors. This analysis would also help the companies to manage the financial resources effectively. The forecasting process can be used for many purposes, including: predicting the future demand of the products or service, predicting how much of the product will be sold in a given period. This is where machine learning can be exploited in a great way. Machine learning is the domain where the machines gain the ability to outperform humans in specific tasks. They are used to do some specialized task in a logical way and gain better results for the progress of the current society. The base of machine learning is the art

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

of mathematics, with the help of which various paradigms can be formulated to approach the optimum output. In case of sales also forecasting machine learning has proved to be a boon. It is helpful in predicting the future sales more accurately .There are several ways of forecasting sales in which companies have previously focused on various statistical models such as time series and linear regression, feature engineering and random forest models to obtain future sales and demand prediction. Time series contains data points that are stored over a fixed period and are used to forecast the future. Time series is a collection of data points which are collected in period at sequential, evenly spaced points. The most important components to analyse are patterns, seasonality, irregularity, cyclicity. Linear regression is a mathematical tool used to forecast past values. It can help to determine the underlying trends and address cases involving overstated rates . Feature engineering is the use of data on domain knowledge and the development of features to make predictive Machine Learning models more accurate. It makes for deeper data analysis and a more useful perspective. A decision tree is a fundamental principle behind a model of random forests. The decision tree approach is a technique used in data mining to forecast and classify data. The decision tree approach does not provide any conceptual understanding of the issue itself. Random forest is the more sophisticated method that allows and merges many trees to make decisions. The random forest model results in more accurate forecasts by taking out an average of all individual tree decision predictions. The entire data set is usually divided into two parts, namely the training data and the test data. Training data is a data that is used to train the model, and test data is the data used to evaluate the trained model. A classical approach is 80-20 split, stating that 80 percent of the data is used to train the model, and the remaining 20 percent of the data is used to test the model.

The aim is to style a model that gains from the market information utilizing machine learning strategies and gauge the long run patterns available value development. The Support Vector Machine (SVM) may be used for both classification and regression. It's been observed that SVMs are more employed in classification based issue like ours. The SVM technique, we plot every single data component as some extent in n-dimensional space (where n is that the number of features of the dataset available) with the worth of feature being the worth of a specific coordinate and, hence classification is performed by finding the hyper-plane that differentiates. Predictive methods like Random forest technique are used for the identical. The random forest algorithm follows an ensemble learning strategy for classification and regression. The random forest takes the common of the varied subsamples of the dataset, this increases the predictive accuracy and reduces the over-fitting of the dataset.

II. LITERATURE SURVEY

A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression

Random Forest and Linear Regression used for prediction analysis which gives less accuracy. To overcome this they use XG boost Algorithm which will give more accuracy and will be more efficient.[1]

Sales Prediction System Using Machine Learning

The objective is to get proper results for predicting the future sales or demands of a firm by applying techniques like Clustering Models and measures for sales predictions. The potential of the algorithmic methods are estimated and accordingly used in further research.[2]

Intelligent Sales Prediction Using Machine Learning Techniques

Multiple Instance Learning Accurately predicting the exchange could be a challenging task, but the fashionable web has proved to be a really useful gizmo in making this task easier.[3]

Big Mart Sales Prediction Using Machine Learning

Predicting the accuracy for XG Boost Regressor. Big Mart Sales Prediction Using Machine Learning**.** The results predicted will be very useful for the executives of the company to know about their sales and profits. This will also give them the idea for their new locations or Centre's of Big Mart.[4]

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

Researching Sales Forecasting Practice

Clarifying the audit function is particularly important since sales forecasting often has a low organisational profile until events turn sour with damaging consequences to organisational viability.[5]

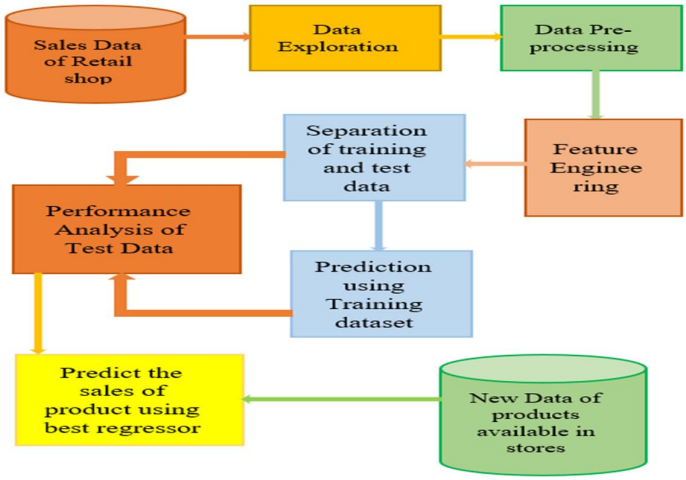
A Study of Demand and Sales Forecasting Model using Machine Learning Algorithm

The fields and attributes used in the study were inadequate for further analysis. Big data can be used as a method for predictive analytics in sales forecasts to speed up the latest studies.[6]

III. Problem Statement

The aim of this project is to help retailers make a healthy relationship with customers by knowing their needs and changing trends. While opening a new store retailers or Big Marts must know about the products they have to focus on. To know the future sales and best selling products in the locality, we are going to use this predictive model. By collecting the dataset of stores and marts in the specific region, it is easy to build a model predicting best selling as well as less selling product in the market.

IV. System Architecture



Collecting Data from various shops to explore needs of customers. Performing different processes on the collected data like, Data exploration, Data pre-processing, Separation of training and testing data. To explore data in any data science competition, it is advisable to append test data to the train data. So, we will combine both train and test to carry out data visualization, feature engineering, one-hot encoding, and label encoding. Later we would split this combined data back to train and test datasets. Most of the times the given features in a dataset are not enough to give satisfactory predictions. In such cases, we have to create new features which might help in improving the model's performance. Before feeding our data into any model, it is a good practice to pre-process the data. We will do pre-processing on both independent variables and target variable. After all this, we will perform prediction on test dataset.

IV. Module

The various modules of the project would be divided into the segments as described.

4.1 Data Collection

Data collection may be a very basic module and also the initial step towards the project. It generally deals with the gathering of the proper dataset. The dataset that's to be employed in the market prediction must be wont to be filtered supported various aspects. Data collection also complements to boost the dataset by adding more data that are external. Our data mainly consists of the previous year information of various shops. Initially, we are going to use Kaggle Dataset, we'll be using the model with the information to predict accurately.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

4.2 Data Cleaning

It was observed from the previous section that the attributes Outlet Size and Item Weight has missing values. In our work in case of Outlet Size missing value we replace it by the mode of that attribute and for the Item Weight missing values we replace by mean of that particular attribute. The missing attributes are numerical where the replacement by mean and mode diminishes the correlation among imputed attributes. For our model we are assuming that there is no relationship between the measured attribute and imputed attribute.

4.3 Feature Engineering

Some nuances were observed in the data-set during data exploration phase. So this phase is used in resolving all nuances found from the dataset and make them ready for building the appropriate model. During this phase it was noticed that the Item visibility attribute had a zero value, practically which has no sense. So the mean value item visibility of that product will be used for zero values attribute. This makes all products likely to sell. All categorical attributes discrepancies are resolved by modifying all categorical attributes into appropriate ones. Finally, for determining how old a particular outlet is, we add an additional attribute year to the dataset.

4.4 Data Visualization

Data visualization is defined as a graphical representation that contains the information and the data .By using visual elements like charts, graphs, and maps, data visualization techniques provide an accessible way to see and understand trends**,** outliers**,** and patterns in data .In modern days we have a lot of data in our hands that is in the world of Big Data, data visualization tools, and technologies are crucial to analyze massive amounts of information and make data-driven decisions. The graphs are shown below.

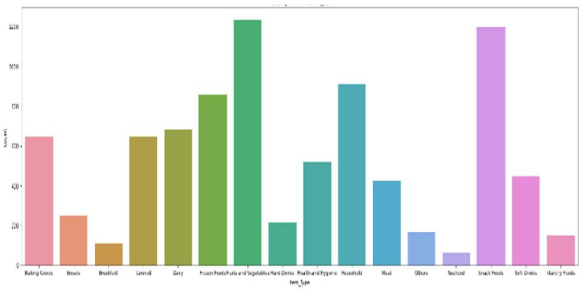


Fig. Count plot for Item Type

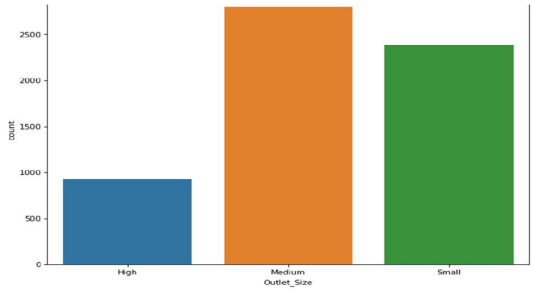


Fig. Count plot for Outlet Size

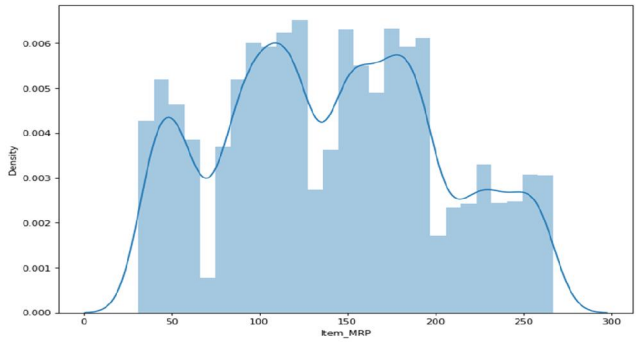


Fig. Data Distribution of Item Weight

V. METHODOLOGY

5.1 Random Forest Algorithm

Random forest algorithm is being used for the stock market prediction. Since it has been termed as one of the easiest to use and flexible machine learning algorithm, it gives good accuracy in the prediction. This is usually used in the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

classification tasks. Because of the high volatility in the stock market, the task of predicting is quite challenging. In stock market prediction we are using random forest classifier which has the same hyper parameters as of a decision tree.

5.2 Linear Regression Algorithm

Linear Regression is the most commonly and widely used algorithm Machine Learning algorithm. It is used for establishing a linear relation between the target or dependent variable and the response or independent variables. The main aim of this algorithm is to find the best fit line to the target variable and the independent variables of the data. It is achieved by finding the most optimal values for all θ. With best fit it is meant that the predicted value should be very close to the actual values and have minimum error.

5.3 Decision Tree Regression

A Decision tree is a machine learning algorithm that can be used for both classification and regression . Decision trees are predictive models that use a set of binary rules to calculate a target value. Each individual tree is a fairly simple model that has branches, nodes and leaves. A decision tree is arriving at an estimate by asking a series of questions to the data, each question narrowing our possible values until the model get confident enough to make a single prediction. The order of the question as well as their content are being determined by the model. In addition, the questions asked are all in a True/False form. This is a little tough to grasp because it is not how humans naturally think, and perhaps the best way to show this difference is to create a real decision tree from. In the above problem x1, x2 are two features which allow us to make predictions for the target variable y by asking True/False questions.

5.4 XG Boost Algorithm

XG Boost also known as Extreme Gradient Boosting has been used in order to get an efficient model with high computational speed and efficacy. The formula makes predictions using the ensemble method that models the anticipated errors of some decision trees to optimize last predictions. Production of this model also reports the value of each feature's effects in determining the last building performance score prediction. This feature value indicates that outcome in absolute measures – each characteristic has on predicting school performance. XG Boost supports parallelization by creating decision trees in a parallel fashion. Distributed computing is another major property held by this algorithm as it can evaluate any large and complex model. It is an out-core-computation as it analyses huge and varied datasets. Handling of utilization of resources is done quite well by this calculative model. An extra model needs to be implemented at each step in order to reduce the error.

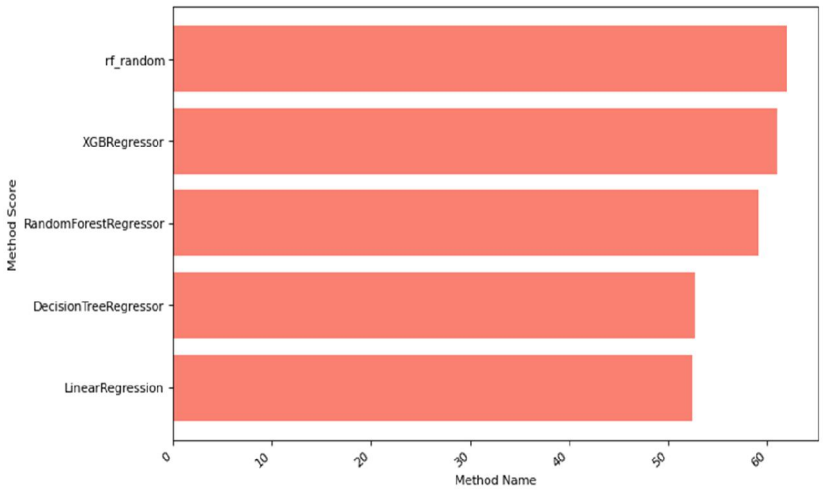


Fig. Accuracy of every algorithm.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

VI. Software Requirement

- Python
- IDE(Jupyter / Colab)
- Numpy
- Seaborn
- Pandas
- Matplotlib
- Streamlit
- Sklearn

VII. Result And Discussion

The records set divided into two corporations, one used for training and different for trying out. The training set consists of 70% of the aggregate records and remaining 30% are used as checking out. We also perform experiments on equal (30% or 70%) dataset that is training in addition to testing for KNN classifier.

In first output, we are predicting the outlet sales, on the basis of parameters like Item Weight, Item Fat Content, Item Visibility, Item Type, Item MRP, Outlet Establishment Year, Outlet Size, Location Type, and Outlet Type. Shown in Figure.

In second output, we are predicting the Best Selling and Less Selling Products. The prediction is based on Dataset collected from the shops, and Big Marts in the locality. Shown in Fig.6.2.

```
hp = bigdata.groupby('Item_Type').agg({'sales': 'max'})['sales'].to_frame('Sales').reset_index()
hp.sort_values(by=['Sales'], inplace=True, ascending=False)
hp.set_index('Item_Type', inplace=True)
hp
```

Item_Type	Sales
Breakfast	6182.808594
Frozen Foods	6182.808594
Fruits and Vegetables	6182.808594
Health and Hygiene	6182.808594
Household	6182.808594
Snack Foods	6182.808594
Dairy	6176.418457
Starchy Foods	6131.549316
Meat	6127.083496
Canned	6120.693359
Soft Drinks	6086.665527
Others	5966.435059
Baking Goods	5732.297363
Hard Drinks	5216.457520
Seafood	5216.457520
Breads	5163.319336

Figure 1: Best Selling Product

```
lp = bigdata.groupby('Item_Type').agg({'sales': 'min'})['sales'].to_frame('Sales').reset_index()
lp.sort_values(by=['Sales'], inplace=True)
lp.set_index('Item_Type', inplace=True)
lp
```

Item_Type	Sales
Household	28.915073
Meat	44.982426
Health and Hygiene	46.031727
Baking Goods	51.065739
Frozen Foods	53.264042
Hard Drinks	53.264042
Seafood	59.272663
Snack Foods	59.272663
Canned	66.504974
Dairy	142.655670
Fruits and Vegetables	142.655670
Breads	147.633179
Soft Drinks	206.865067
Others	225.083542
Starchy Foods	305.182770
Breakfast	336.011871

Figure 2: Less Selling Product

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

VIII. CONCLUSION

In this project, by using Machine Learning algorithm, we have developed a system that is capable of finding the best selling product and predicting future sales. This will help the owner of Big Marts or other retailers who are planning to open a new shop in specific locality. By using Streamlit, we have created a web application capable of predicting best selling products.

REFERENCES

[1]. Heramb Kadam, Rahul Shevade and Prof. Deven Ketkar: "A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression". In: 2013 International Journal Of Engineering Development and Research .

[2]. M Panjwani , R Ramrakhiani and H Jumnani: **"**Sales Prediction System Using Machine Learning" April 23, 2020.

[3]. Sunitha Cheriyan and Shaniba Ibrahim: "Intelligent Sales Prediction Using Machine Learning Techniques" 2018 International Conference on Computing, Electronics & Communications Engineering (ICCECE) 07 March 2019.

[4]. Rohit Sav, Pratiksha Shinde, Saurabh Gaikwad : "Big Mart Sales Prediction Using Machine Learning" International Journal of Creative Research Thoughts (IJCRT)

[5]. Robert Fildes, Stuart Bretschneider and Fred Collopy "Researching Sales Forecasting Practice", International Journal of Forecasting, 2003.

[6]. Aneesh Tony, Pradeep Kumar, Rohith Jefferson, Subramanian "A Study of Demand and Sales Forecasting Model using Machine Learning Algorithm",2021.

[7]. Singh Manpreet, Bhawick Ghutla, Reuben Lilo Jnr, Aesaan FS Mohammed, and Mahmood A. Rashid. "Walmart's Sales Data Analysis-A Big Data Analytics Perspective." In 2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), pp. 114-119. IEEE, 2017.

[8]. Panjwani, Mansi, Rahul Ramrakhiani, Hitesh Jumnani, Krishna Zanwar, and Rupali Hande. Sales Prediction System Using Machine Learning. No. 3243. EasyChair, 2020.

[9]. Cheriyan, Sunitha, Shaniba Ibrahim, Saju Mohanan, and Susan Treesa. "Intelligent Sales Prediction Using Machine Learning Techniques." In 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), pp. 53-58. IEEE, 2018.

[10]. Saltz, J. S., & Stanton, J. M. (2017). An introduction to data science. Sage Publications.

[11]. T. Alexander and D. Christopher, quot;An Ensemble Based Predictive Modeling in Forecasting Sales of Big Martquot;, International Journal of Scientific Research, vol. 5, no. 5, pp. 1- 4, 2016.