Bigmart Sales Prediction Using Machine Learning

Narmadha A[1], Sushmitha B[2], A S Hepsi Ajibah[3]
[1, 2, 3] Dept of EEE
[1, 2, 3] GKM College of engineering and technology

Abstract- *In today's world big malls and marts record sales data of individual items for predicting future demand and inventory management. This data Stores a large number of attributes of the item as well as individual customer data together in a data warehouse. This data is mined for detecting frequent patterns as well as anomalies. This data can be used for forecasting future sales volume with the help of random forests and a multiple linear regression model*

Keywords- Xg boost, Machinelearning, linear Regression, Random Forests

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 paper, 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.

II. LITERATURE REVIEW

Tanu Jain, AK Sharma [1] interprets that the algorithms which are frequently used in the field of association rule mining are Eclat and Apriori (market basket analysis) algorithms. Both of these algorithms are mainly used for mining of primarily data sets and to find fraternity(associations) between these regular data sets using R which is a domain based language for data exploration, analysis and analytics. Several packages and libraries of R has been used by the authors to examine the performance of Eclat and Apriori algorithms on different item sets on the basis of execution time taken by both of the algorithms..

Author [2] basically interprets that what is data analysis and how we can do it efficiently?. In this paper author recommends R for data analysis because of its tremendous capability of data exploration, several inbuilt packages, easy to implement several machine learning algorithms etc. As we know that R is a statistical language as well as programming language which helps in effective model prediction and better visualization techniques. So after survey authors found that the with R data analysis is much more efficient.

Author of this paper [3] uses the list of top 10 machine learning algorithms to observe the influence of these algorithms which was published by IEEE in ICDM(International Conference on Data Mining) in the month of December 2006.List contains top 10 data mining algorithms which are mostly used by the research community are as follows: k-means algorithm, support vector machine, Apriorist algorithm, Naive Bayes, Nearest neighbour(kNN), Decision tree, Expectation Maximization(EM)algorithm, PageRank, AdaBoost, Eclat algorithm.

Hilda,Jurgen and joseph [4] analyzes and compares three famous mechanisms or tools of data mining known as – Rapid Miner,Weka,R respectively to use their expertise in the area of structural health monitoring. This paper interprets several functionalities of R,Weka, Rapid miner in time series analysis for structural health monitoring like visualizing, filtering, applying statistical models etc.

Author's of this paper [5] analyzes a free and emerging statiscal language known as R for mining big data . This paper provides information about implementing a clustering technique known as K-means with the help of Ron a huge data set. In today's date data coming from several sources is immense and to mine such a huge data in not a simple task,but this paper shows us that we can do it efficiently using R.

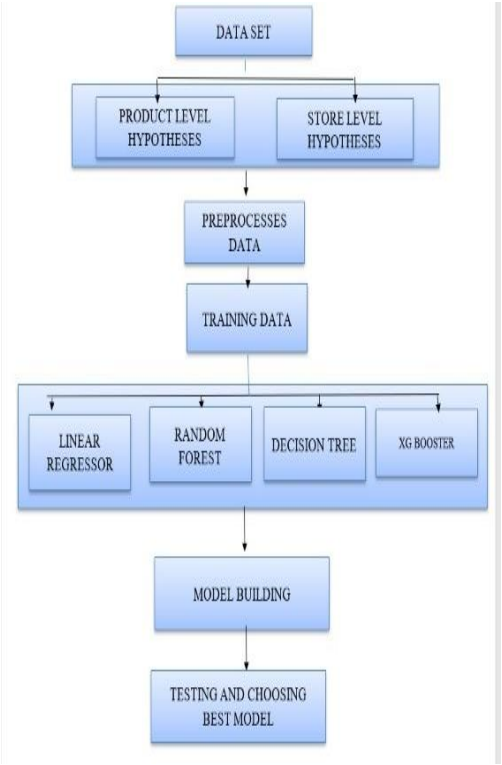
III. OBJECTIVE

The goal of businesses is to make profits, Which can be achieved by making more sales than expenses. Sales are the lifeblood of each, and every organization and sales forecasting play a critical role in conducting every business.

Better forecasting helps to build and enhance business plans by growing awareness about the marketplace

Big mart sales of an item on the potential demand of customers in various large mart stores across different locations and items based on the previous record. The main objective is to provide a more accurate predictive model in the prediction of outlet item sales in Bigmart to increase its growth Now it is the time to articulate the research work with ideas gathered in the above steps by adopting any of below suitable approaches

IV. SYSTEM ARCHITECTURE



V. MODULES

Sales prediction is preferably a regression problem than a time series problem Practice shows that the use of regression procedures can often supply us better results comparing with time series techniques. Machine learning algorithms make it possible to find patterns in the time series. Bigmart sales dataset consists of 2013 sales data for 1559 products throughout 10 special stores in unique towns.

We have 2 dataset the train dataset which has 8523 rows and 12 features and the test dataset which has 5681 rows and 11 columns. The train dataset has 1 extra column which is the target variable. We will predict this target variable for the test dataset. Calculations done in the Python environment using the main packages pandas, sklearn, numpy, matplotlib, seaborn etc. To conduct the analysis, we will be using Jupyter Notebook. The goal of the Bigmart sales prediction ML challenge is to build a regression model for expecting the sales of every of 1559 products for the following year in every of the 10 specific Bigmart stores. The Bigmart sales dataset additionally includes certain attributes for each product and store. This model allows Bigmart to know the properties of products and stores that play an essential position in growing their universal sales. We divided the entire analysis process to following five stages:

1. Exploratory data analysis (EDA)
2. Data Pre-processing
3. Feature engineering & Feature Transformation
4. Modeling
5. Hyper parameter tuning and Evaluation

Each step is explained below in details.

EXPLORATORY DATA ANALYSIS (EDA)

In this phase useful information about the data has been extracted from the dataset. That is trying to identify the information from hypotheses vs available data. Which shows that the attributes Outlet size and Item weight face the problem of missing values, also the minimum value of Item. Visibility is zero which is not actually practically possible. Establishment year of Outlet varies from 1985 to 2009. These values may not be appropriate in this form. So, we need to convert them into how old a particular outlet is. There are 1559 unique products, as well as 10 unique outlets, present in the dataset. The attribute Item type contains 16 unique values. Where as two types of Item Fat Content are there but some of them are misspelled as regular instead of Regular' and low fat, LF instead of Low Fat.

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

FEATURE ENGINEERING & FEATURE TRANSFORMATION

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. In some cases, it was noticed that non-consumables and fat content property are not specified. To avoid this, we create a third category of Item fat content i.e. none. In the Item Identifier attribute, it was found that the unique ID starts with either DR or FD or NC. So, we create a new attribute Item Type New with three categories like Foods, Drinks and Non-consumables. Finally,for determining how old a particular outlet is, we add an additional attribute Year to the dataset

MODEL BUILDING

After completing the previous phases, the dataset is now ready to build proposed model. Once the model is built it is used as predictive model to forecast sales of Big Mart. In our work, we make model based on different algorithms such as Random Forest algorithm, Linear regression, Lasso Regression, Ridge regression, Decision tree etc. and compare it with other machine learning techniques. All models received features as input, which are then segregated into training and test set. The test dataset is used for sales prediction.

HYPERPARAMETER TUNING AND EVALUATION

The next and final step in our project is the tuning of different parameters in every model and saw improvement in model performance. While this is an important step in modeling, it is by no means the only way to improve performance.

VI. CONCLUSION

In this paper, Making a prediction model for predicting sales and to select the best model that can be reliable i.e it's accuracy should be highest .We applied three typical forecasting models and several attributes to the trigger model through training and testing the classification model with present and past sales data.Finally we obtained more accurate forecasting results than could be obtained by comparing the models. We have also tested results of single classifiers separately with the general model

REFERENCES

[1] Research on sales information prediction system of e-commerce enterprises based on time series model Authors: Jian Liu, Chunlin Liu, Lanping Zhang& Yi Xu
[2] a completely unique Trigger Model for Sales Prediction with data processing Techniques Authors: Wenjie Huang , Qing Zhang, Wei Xu,Hongjiao Fu, Mingming Wang, Xun Liang
[3] Sales Prediction Model Using Classification Decision Tree Approach For Small Medium Enterprise Based on Indonesian E – Commerce Data Authors: Raden Johannes H. P., Andy Alamsyah
[4] Sales Demand Forecast in E-commerce Using a Long Short-Term Memory Neural Network Methodology Author:Kasun Bandara, PeibeiShi,Christoph, Bergmeir, Hansika,Hewamanage,Quoc Tran,Brian Seaman.
[5] Sales Prediction System using Machine Learning Authors: Archisha Chandel, Akanksha Dubey, Saurabh Dhawale, Mathura Ghuge
[6] A survey of machine learning techniques forfood sales prediction. Author: Grigorios Tsoumakas
[7] Schroeder, G., Klim, A., Heinz, G., et al.(2010) System for predicting sales lift and profitof a product supported historical sales information: U.S. Patent 7,689,456.
[8] Yuan, H., Xu, W & Wang, M. (2014) Canonline user behavior improve the performance of sales prediction in E-commerce? IEEE International Conference on Systems, Man, and Cybernetics, Patent 2377–2382.
[9] Pavlyuchenko, B.M. Linear, machine learning and probabilistic approaches for statistical analysis. In Proceedings of the IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine,