ANALYZING BIG MARKET SALES USING MACHINE LEARNING ALGORITHMS

V.Ramyasri[1] SK.Manisha[2] V.Padma Priya[3] B. Mery Prasanna[4] Rajesh Yamparala[5]
Department of Computer Science and Engineering, Vignan'sNirula Institute of Technology and Science for Women, Pedapalakaluru, AP, India. Corresponding Author mail id: ramyavadlamudi057@gmail.com

Abstract
Supply and demand are two fundamental concepts of sellers and customers. Predicting demand is accurately is critical for organizations in order to be able to formulate plans. Sales Analysis is based on analysing sales for different outlets of Big mart companies, so that they can change the business model according to performance predicted. In this paper, we propose a new approach for demand prediction for Big Mart companies. The business model used by the Big Mart companies, for which the model is implemented, includes many outlets that sell the same product at the same time throughout the country where the company operates a market place model. The demand prediction for such a model should consider the price tag, outlet type, outlet location. In this study, we first applied linear regression for the specific set of outlets of one of the most popular Big Mart Companies. Then we used Random Forest regressor. Finally, all the approaches are evaluated on a real-world dataset obtained from the Big Mart Company. The experimental results show that the Random Forest regressor gives the pretty accurate sales results.
Keywords:-Data mining, emil Prediction, Naive Bayes Classifier, Decision tree Classifier

I.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. 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 onthe historical data of sales volume. According to the characteristics of the data, we can use the method of multiple linear regression analysis and random forest to forecast the sales volume.To improve the knowledge of marketplace A standard sales prediction study canhelp in deeply analysing the situations or the conditions previously occurred andthen, the inference can be applied about customer acquisition, funds inadequacyand strengths before setting a budget and marketing plans for the upcomingyear. In other words, sales prediction is based on the available resources fromthe past. In depth knowledge of past is required for enhancing and improvingthe likelihood of marketplace irrespective of any circumstances especially theexternal circumstance, which allows to prepare the upcoming needs for the business. Extensive research is going on in retailer's domain
for forecasting the futuresales demand. The basic and foremost technique used in predicting sale is thestatistical methods, which is also known as the traditional method, but thesemethods take much more time for predicting a sales also these methods couldnot handle non linear data so to over these problems in traditional methodsmachine learning techniques are deployed. Machine learning techniques can notonly handle non-linear data but also huge data-set anciently. To measure theperformance of the models, Root Mean Square Error (RMSE) and MeanAbsolute Error (MAE) are used as an evaluation metric as mentioned in theEquation 1 and2 respectively. Here Both metrics are used as the parameter foraccuracy measure of a continuous variable.Remaining part of this article is arranged as following: Section 1 briefly describesintroduction of sales prediction of Big Mart and also elaborate about the evaluation metric used in the model. Previous related work has been pointed in Section2. The detailed description and analysis of

proposed model is given in Section3. Whereas implementations and results are demonstrated in Section 4 and thepaper concludes with a conclusion in the last section.

In this paper, we propose a predictive model using random forest technique for predicting the sales of a company like Big Mart and we found our model produces better performance as compare to other model. Sales Prediction is used to predict sales of different products sold at various outlets in different cities of a Big Mart Company. As the volume of products, outlets are growing exponentially predicting them by hand becomes cumbersome. Predicting the right demand for a product is an important phenomenon in terms of space, time and money for the sellers.Now a days shopping malls and Big Marts keep 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. This data can be used for forecasting future sales volume with the help of random forests and multiple linear regression model. 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.

If sellers store much more product than the demand then this may lead to surplus. On the other hand, storing less product in order to save inventory costs when the product has a high demand will cause less revenue. Thus, Sales Prediction helps the companies to store products according with the predicted sales for products and different outlet locations helps companies to formulate a proper business model which helps them to organize and dispatch its product more efficiently thus cutting down on costs and increasing revenue.

II.Literature Survey

Sales forecasting as well as analysis of sale forecasting has been conducted bymany authors as summarized: The statistical and computational methods arestudied in also this paper elaborates the automated process of knowledgeacquisition. Machine learning [1] is the process where a machine will learn fromdata in the form of statistically or computationally method and process knowledge acquisition from experiences. Various machine learning(ML) [6,3] techniqueswith their applications in deferent sectors has been presented in [2]. Pointed out most widely used data mining technique inA Comparative Study of Big Mart Sales Prediction[7]the field of business is the Rule Induction (RI) technique as compared to otherdata mining techniques. Whereas sale prediction of a pharmaceutical distribution company has been described in [11]. Also this paper focuses on twoissues: (i) stock state should not undergo out of stock, and (ii) it avoids thecustomer dissatisfaction by predicting the sales that manages the stock level ofmedicines. Handling of footwear sale punctuation in a period of time has beenaddressed in [5]. Also this paper focuses on using neural network for predictingof weekly retail sales, which decrease the uncertainty present in the short termplanning of sales. Linear and non-linear [12] a comparative analysis model for salesforecasting is proposed for the retailing sector [13]performed sales prediction in fashion market. A two level statistical method [10]is elaborated for forecasting [9] the big mart sales prediction. Xia and Wong proposed the divergences between classical methods (based on mathematical andstatistical models) and modern heuristic methods and also named exponentialsmoothing, regression [14,15], auto regressive integrated moving average (ARIMA), generalized auto regressive conditionally heteroskedastic (GARCH) methods. Mostof these models are linear and are not able to deal with the asymmetric behaviourin most real-world sales data [8]. Some of the challenging factors like lack ofhistorical data, consumer-oriented markets face uncertain demands, and shortlife cycles of prediction methods results in inaccurate forecast.

III. Proposed Model

We are proposing an approach to use Random Forest Regressor algorithm to increase the accuracy of the project. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of

predictions, and it predicts final output. Greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Working of Random Forest Algorithm

Random Forest works in two-phase first is to create the random forest by combining N decision tress, and second is to make predictions for each tree created in the first phase. Working process can be explained in the below steps and diagram.

Step 1- Select random K data points from the training set.

Step 2- Build the decision trees associated with The selected data points(Subsets).

Step 3- Choose the number N for decision trees to build.

Step 4- Repeat Step 1 & 2

Step 5- For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

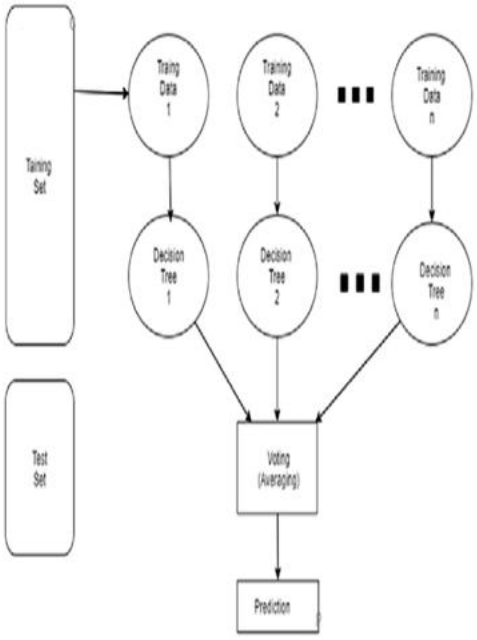

Fig: Random Forest Algorithm

The above diagram explains the working of the Random Forest algorithm. Random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts final output.

Advantages of Random Forest

- Random Forest is capable of performing both classification and Regression tasks.
- It is capable of handling large datasets.
- It enhances the accuracy of the model and prevents the overfitting issue.

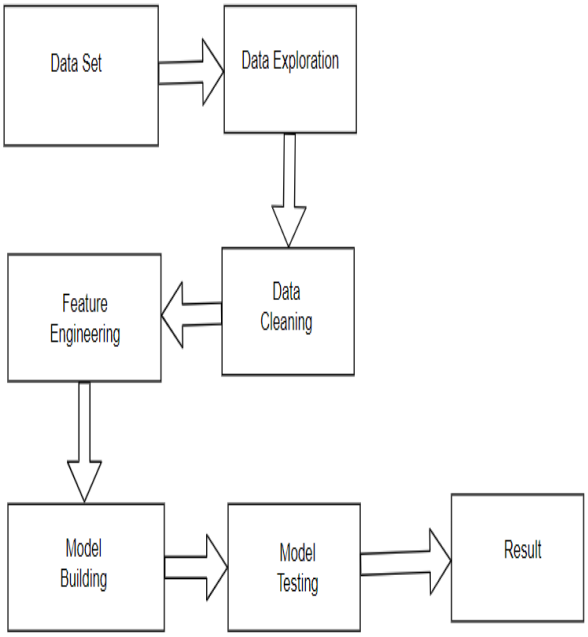

Fig: System Architecture

A **data set** is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the dataset in question.

Data exploration is an approach similar to initial data analysis, whereby a data analyst uses visual exploration to understand what is in a dataset and the characteristics of the data, rather than through traditional data management systems.

Data cleaning is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Feature engineering is the process of using domain knowledge to extract features from raw data via data mining techniques. These features can be used to improve the performance of machine learning algorithms. Feature engineering can be considered as applied machine learning itself.

Model building is the process of developing a probabilistic model that best describes the relationship between the dependent and independent variables.

Model-based testing is an application of model-based design for designing and optionally also executing artifacts to perform software testing or system testing. Models can be used to represent the desired behaviour of a system under test or to represent testing strategies and a test environment.After performing all the above tasks, then we get final result.

IV.UML Diagrams

Sequence Diagram

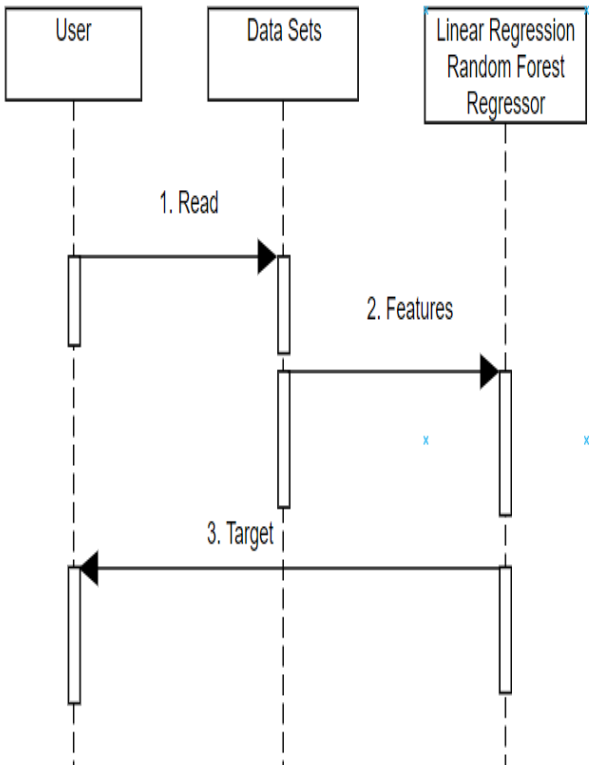

Fig: Sequence Diagram

The above figure shows that User can read the datasets and apply features of Linear Regression and Random Forest Algorithms. Then Find Target variable and send to user.

V: Result

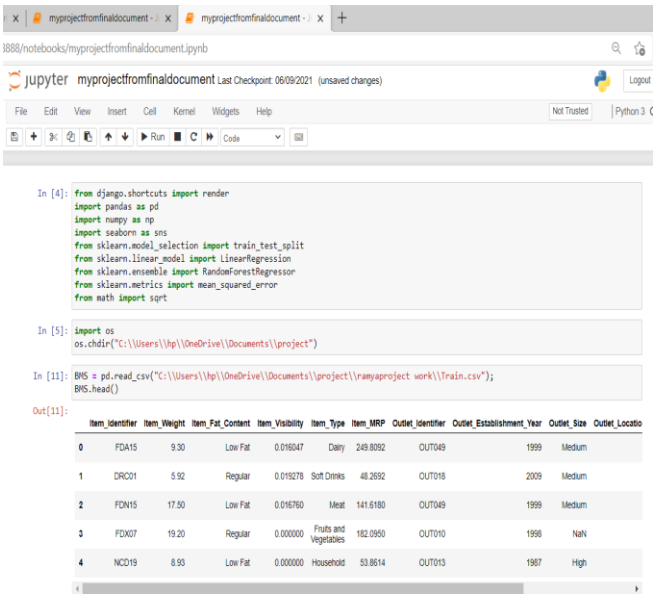

Fig: Representing first five input data

The above shows the Sample dataset representing First five Input data contains some parameters like item-size, item-fat-content, item-weight, outlet-location, outlet-type etc.

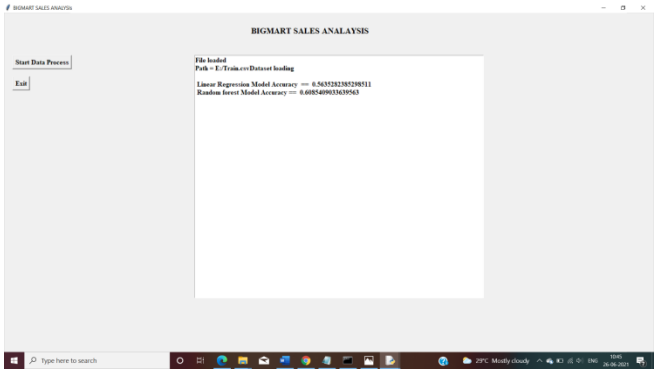

Fig: Final Result after applying algorithms

The above figure shows final result after applying both linear regression and random forest algorithms. It shows the accuracy of both the algorithms and we see random forest gives more accurate results compared to linear regression.

VI. Conclusion

We examine the problem of demand forecasting on an e-commerce web site. We proposed stacked generalization method consists of sub-level regressors. We have also tested results of single classifiers separately together with the general model. Experiments have shown that our approach predicts demand at least as good as single classifiers do, even better using much less training data (only %20 of the dataset). We think that our approach will predict much better when more data is used. Because the difference is not statistically significant between the proposed model and random forest, the proposed method can be used to forecast demand due to its accuracy with fewer data.

VII. Future Enhancement

In future, we will use the output of this project as part of the price optimization problem which we are planning to work on. Here we focus only on Food items. In future, consider other parameters like price and also focus on other things like Grocery Items.

VIII.References

[1]. Alpaydin, Ethem. Introduction to machine learning MIT press, 2020.

[2] Marsland, Stephen. Machine learning: an Algorithm perspective. CRC press, 2015.

[3]. Mohri, Mehryar, AfshinRostamizadeh, and Ameet Talwalkar. Foundations of machine Learning. MIT press, 2018.

[4]. Yamparala, R., &Perumal, B. (2019). Secure Data Transmission with Effective Routing Method Using Group Key Management Techniques-A Survey. *International Information and Engineering Technology Association*, *52*(3), 253-256.

[5]. Zhang, G., Eddy Patuwo, B., Hu, M.Y.: Forecasting with artificial neural networks: The state of the art. Int. J. Forecast. 14(1), 35–62 (1998)

[6]. Yamparala, R., &Perumal, B. (2020). EFFICIENT MALICIOUS NODE IDENTIFICATION METHOD FOR IMPROVING PACKET DELIVERY RATE IN MOBILE AD HOC NETWORKS WITH SECURED ROUTE. *Journal of Critical Reviews*, *7*(7), 1011-1017.

[7].Behera, Gopal, and Neeta Nain. " A Comparative Study of Big Mart Sales Prediction.

[8]. Chandel, Archisha, et al. "Sales Prediction System using Machine Learning."

[9]. CHALLA, R., YAMPARALA, R., KANUMALLI, S. S., & KUMAR, K. S. (2020, November). Advanced Patient's Medication Monitoring System with Ardunio UNO and NODEMCU. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 942-945). IEEE.

[10]. Punam, Kumari, RajendraPamula, and Praphula Kumar Jain. "A Two-Level Statistical Model for Big Mart Sales Prediction." 2018 International Conference on Computing, Power and Communication Technologies (GUCON). IEEE, 2018.

[11]. Krishna, K. V. S. S. R., Chaitanya, K., Subhashini, P. P. S., Yamparala, R., &Kanumalli, S. S. (2021). Classification of Glaucoma Optical Coherence Tomography (OCT) Images Based on Blood Vessel Identification Using CNN and Firefly Optimization. *Traitement du Signal*, *38*(1).

[12]. Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. Vol. 821. John Wiley & Sons, 2012.

[13]. Yamparala, R., Challa, R., Kantharao, V., & Krishna, P. S. R. (2020, July). Computerized Classification of Fruits using Convolution Neural Network. In *2020 7th International Conference on Smart Structures and Systems (ICSSS)* (pp. 1-4). IEEE

[14]. Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.

[15]. Segal, Mark R. "Machine learning benchmarks and random forest regression." (2004).