IMPROVIZING BIG MARKET SALES PREDICTION

Meghana N*
Department of Computer Science and Engineering
GITAM School of Technology, Bengaluru campus, Karnataka, India

Pavan Chatradi
Department of Computer Science and Engineering
GITAM School of Technology, Bengaluru campus, Karnataka, India

Avinash Chakravarthy V
Department of Computer Science and Engineering
GITAM School of Technology, Bengaluru campus, Karnataka, India

Sai Mythri Kalavala
Department of Computer Science and Engineering
GITAM School of Technology, Bengaluru campus, Karnataka, India

Mrs.Neetha K S
Assistant Professor
Department of Computer Science and Engineering
GITAM School of Technology, Bengaluru campus, Karnataka, India

Abstract- Estimating future sales is the major aspect of the numerous distributions, manufacturing, marketing and wholesaling companies involved. This helps businesses to allocate capital effectively, to forecast realistic sales revenues as well as to prepare a better plan for potentially increasing the business. In this paper, estimating product sale from a single outlet is carried out using a random forest regression approach, XG booster approach which provides better predictive results compared to a linear regression model. This approach is carried out on data from Big-Mart Sales where data discovery, processed and sufficient relevant data is extracted which play a vital role in predicting accurate outcome.

Keywords – *Sales* Forecast, Machine Learning, Linear regression, random forest , XG booster.

I. INTRODUCTION

Estimating future sales is an important aspect of any business. Accurate prediction of future sales help companies to develop and improve business strategies as well as to gain proper market knowledge. Standard sales forecast helps companies to analyze the situation which has occurred before and then apply customer purchase inferences to identify inadequacies and weaknesses before budgeting as well as to prepare good plan for the next year. A detailed knowledge of past opportunities permits one to plan for future market needs and increase the possibility of success Regardless of external factors, firms which see sales modeling as its first step towards improved performance compared to those who don't.

In this paper, we use random forest regressor and XG-booster approach to predict sales where data mining techniques such as discovery, data transformation, feature development, model creation and testing are used. In this technique raw data collected by a big mart will be pre-processed for missing data, anomalies and outlier. An algorithm will then be trained to construct a model on that data. We will use this model to forecast the end results. It is a system in which three functions are combined. It is used to extract and transform the data from one database into an appropriate format. Generally, in pre-processing of data raw data is converted into useful form of data. Data pre-processing involves the following steps

- Data-cleaning.
- Data-transformation.
- Data-reduction.

In this proposed system we have used Random Forest Algorithm to incorporate predictions from multiple decision trees into a single model.

This paper consists of following sections: In Section 1 we discuss briefly about the Introduction to sales forecasting. In Section 2 the research done so far is presented. Section 3 contains a detailed information about the methodology used by the proposed method. Section 4 presents the system architecture and algorithms used, Section 5 presents final results and Section 6 ends with the conclusion.

.

II. LITERATURE SURVEY

Machine Learning is defined as the computer program which learns by itself from its experience without any human interference. Research on sales prediction has been done and some of them has been discussed below:

In paper[1], general linear approach, decision tree approach and good gradient approach were used to predict sales. The initial data set considered included many entries, but the final data set which is used for analyzing was much smaller than the original as it consists of non-usable data, redundant entries and insignificant sales data.

In paper[2], linear regression method has been organized into structured data. Then it involves modeling data for predictions using machine learning techniques where the expected accuracy was 84%.

In paper[3], they used linear regression and XG booster algorithm to forecast sales that included data collection and translation into processed data. Ultimately, they predicted which model would produce the better outcome.

In paper[4],sales were predicted using three modules that are hive, R programming and tableau. By analysing the stores history which helps get an understanding of the store's revenue to make some improvements to the target so it can be more successful. Within the diagram, key values are obtained to reduce all intermediate values by reducing the intermediate key feature to obtain the results.

Mohit Gurnani in his research proves that composite models achieve good results in comparison to individual models. He also stated that decomposition mechanisms are far better than hybrid mechanisms [5]. J. Scott Armstrong in his research discussed about predicting solutions to interesting and difficult sales forecasting problems [6].

Samaneh Beheshti-Kashi in his research reviewed different Various approaches on the predictive potential of consumer-generated content and search queries [7]. Gopal Behera has done effective study on Big mart sales prediction and has given prediction metrics for various existing models [8].

In this paper, we use random forest and XG booster methodology in which raw data obtained at large mart will be pre-processed for missing data, anomalies and outliers. Then an algorithm will be used to predict the final results. ETL stands for Extract, Transform and Load and finally we compare all the models and predict which model gives accurate result.

III. RESEARCH METHODOLOGY

1. Dataset Collection

We used wide market sales data as a dataset in our work where the dataset consists of 12 attributes. These 12 attributes define the basic features of the data which is being forecasted. These attributes are divided into Answer Variable and predictors. Here we use dataset which contains 8523 items spanning various locations as well as cities. Store-and product-level hypotheses are the main factors on which our dataset focuses on. Attributes such as area, population density, capability of the store, location etc have been included in store level. At last the dataset is divided as training and test dataset.

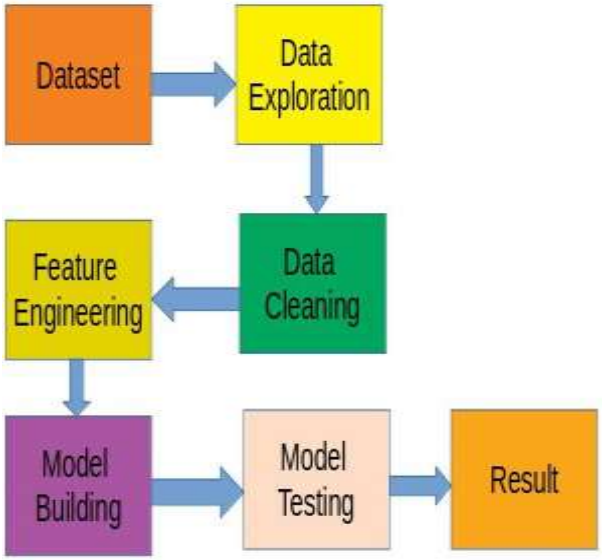

Fig 1: Steps Involved In Forecasting.

2. Data Exploration

Valuable data information is drawn-out from the dataset in this step. Outlet year of establishment ranges from 1985 to 2009. These Values in this form may not be sufficient. There are 1559 different items present in the dataset and 10 different outlets Here we classify the data from the hypothesis vs available evidence which indicates that the size of the outlet attribute and the weight of the object faces the question of missing values, as well as the least value of Object view is Zero which is not feasible. The Item type attribute contains 16 specific values. Variable outlet sales have been skewed positively. So, a log is applied on Answer Variable to skewedness.

3. Data Cleaning

In previous section it has been found that attributes Outlet size and Element weight lack values. Here in place of missing value for outlet size, we replace with mode value of that attribute and in place of missing values of that particular attribute of object weight, we substitute by mean value. The missing attributes are numerical, where correlation between the imputed attributes decreases as well as the mean and mode replacement decreases. we believe that there is no relation among the attributes calculated and the attribute imputed in our model.

4. Feature Engineering

Feature engineering is all about converting cleaned data into predictive models to present the available problem in a better way. During data exploration, some noise was observed. In this phase, this noise is resolved and the data is used for building appropriate model. New features are created to make the model work precisely and effectively. A few created features can be combined for the model to work better. Feature engineering phase converts data into a form understandable by the algorithms.

5.Model building

After Feature engineering, the processed data is used to give accurate results by applying multiple algorithms. A model is a set of algorithms that facilitate the process of finding relation between multiple dataset. An effective model can predict accurate results by finding exact insights of data.

IV. ARCHITECTURE OF PROPOSED SYSTEM

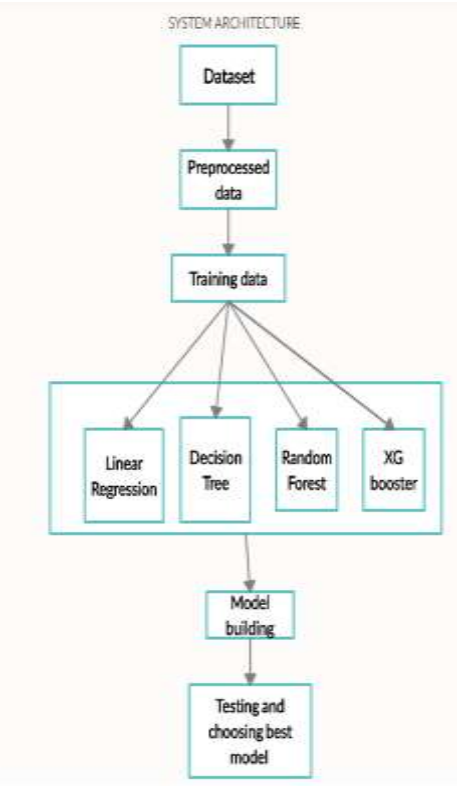

Fig 2: Diagram of Proposed System.

The algorithms used in this framework are,

- **LINEAR REGRESSION**:

  Linear regression algorithm tries to predict the results by plotting the graph between an independent variable and a dependent variable that are derived from the dataset. It is a general statistical analysis mechanism used to build machine learning models. The general equation for linear regression is

$$Z = a + bE$$

  Where, Z is the dependent variable and E is independent variable.

- **RANDOM FOREST**:

  Random Forest Algorithm is used to incorporate predictions from multiple decision trees into a single model. This algorithm uses bagging mechanism to create a forest of decision trees. It the incorporates the predictions from multiple decision trees to give very accurate predictions. The Random Forest algorithm has two steps involved,

  - Random forest formation.

- Predict by Random forest classifier generated.

- **XG BOOSTER APPROACH**:

  The XG Boost algorithm is developed using Decision trees and Gradient boosting. This algorithm stands on the principle of boosting other weaker algorithms placed in a gradient decent boosting framework. This approach works very accurately beating almost all other algorithms in providing accurate prediction. It can be defined as an extension to Gradient Boosting algorithm. Features of XG Boost are,

  - Parallelized tree building.
  - Efficient handling of missing data.
  - In built cross validation capability.
  - Tree pruning.
  - Cache Awareness.

V. RESULTS AND DISCUSSION

In the below graph, we can observe that the feature with the lowest correlation with our target variable is the Item Visibility. So, the less available the commodity is the higher the price would be in the shop. The most positive finding is from Item MRP.

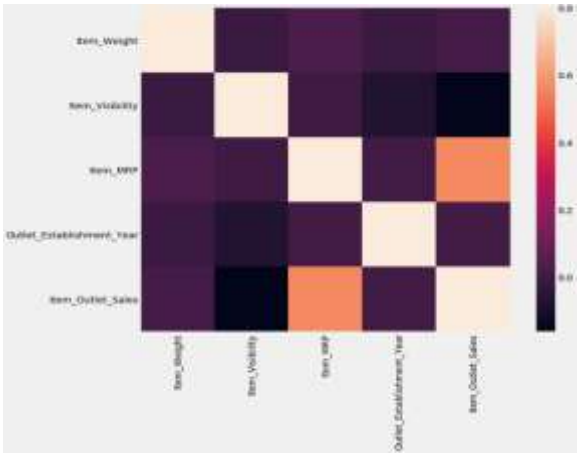

Fig 3: Graph to predict Correlation of variables with target variable.

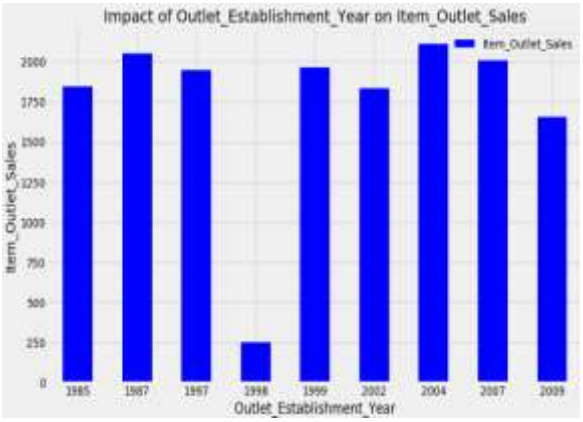

Fig 4: Yearly sales visualization.

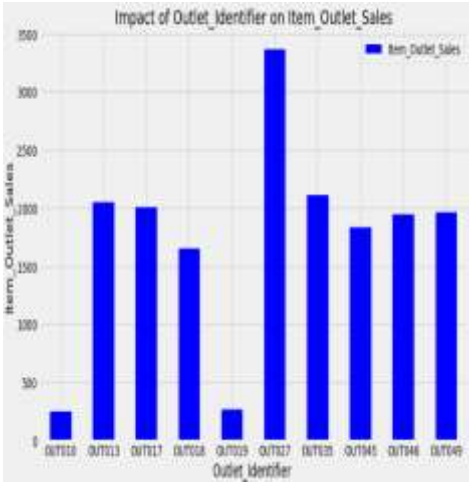

Fig 5: Comparison of sales at different outlets.

algorithm	mean error
Random Forest	13
Xgboost	13
Decision Tree	19
Linear	33

Fig 6: Mean error of different models.

Mean error is the important factor based on which the accuracy of the model relay on lesser the meanerror higher is the accuracy of the predicted outcome.

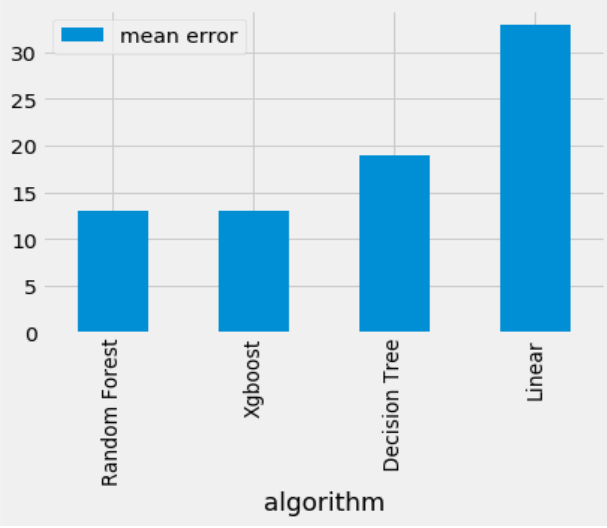

Fig 7: Comparison between the models.

VI. CONCLUSION

The objective of this framework is to predict the future sales from given data of the previous year's using machine Learning techniques. In this paper, we have discussed how different machine learning models are built using different algorithms like Linear regression, Random forest regressor, and XG booster algorithms. These algorithms have been applied to predict the final result of sales. We have addressed in detail about how the noisy data is been removed and the algorithms used to predict the result. Based on the accuracy predicted by different models we conclude that the random forest approach and XG Booster approach are best models. Our predictions help big marts to refine their methodologies and strategies which in turn helps them to increase their profit.

VII. REFERENCES

[1]**"Applied Linear Statistical Models",** Fifth Edition by Kutner, Nachtsheim, Neter and L, Mc Graw Hill India, 2013.

[2]Demchenko, Yuri & de Laat, Cees & MembreyPeter,**"Defining architecture components of the Big Data Ecosystem"**, 2014.

[3] Blog: Big Sky, **"The Data Analysis Process: 5 Steps To Better Decision Making"**, (URL: https://www.bigskyassociates.com/blog/bid/372186/The-Data- Analysis-Process-5-Steps-To-Better-Decision-Making).

[4] Blog: Dataaspirant, **"HOW THE RANDOM FOREST ALGORITHM WORKS IN MACHINE LEARNING"**, (URL: https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/.).

[5] Mohit Gurnani, Yogesh Korke, Prachi Shah, Sandeep Udmale, Vijay Sambhe, Sunil Bhirud, **"Forecasting of sales by using fusion of machine learning techniques",**2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), IEEE, October 2017.

[6] Armstrong J,**"Sales Forecasting"**, SSRN Electronic Journal, July 2008.

[7] Samaneh Beheshti-Kashi, Hamid Reza Karimi, Klaus-Dieter Thoben, Michael Lütjen, **"A survey on retail sales forecasting and prediction in fashion markets"**, Systems Science & Control Engineering: An Open Access Journal. 3. 154-161. 10.1080/21642583.2014.999389.

[8]Gopal Behera, Neeta Nain, **"A Comparative Study of Big Mart Sales Prediction"**, 4th International Conference on Computer Vision and Image Processing, At MNIT Jaipur, September 2019.