International Journal of Scientific Research and Engineering Development-– Volume2 Issue 2, Mar –Apr 2019

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Sales Prediction System using Machine Learning

Archisha Chandel[*], Akanksha Dubey[**], Saurabh Dhawale[***], Madhuri Ghuge[****]

[*] [**] [***] (Computer Department, Mumbai University/Bharati Vidyapeeth College of Engineering, Navi Mumbai)

**** (Assistant Professor of Computer Department, Mumbai University/Bharati Vidyapeeth College of Engineering, Navi Mumbai)

Abstract:

Supply and demand are two fundamental concepts of sellers and customers. Predicting demand accurately is critical for organizations in order to be able to formulate plans. Sales Prediction is based on predicting the 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 and decision tree algorithm for the specific set of outlets of one of the most popular Big Mart Companies in the USA. Then we used XGBoost regressor, a gradient-based algorithm to predict sales [1]. Finally, all the approaches are evaluated on a real-world data set obtained from the Big Mart Company. The experimental results show that the XGBoost regressor gives pretty accurate sales results.

Keywords **—machine learning, B2B sales forecasting, sales prediction, XGBoost regressor**

I. INTRODUCTION

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. Sellers may have limited time or need to sell their products as soon as possible due to the storage and money restrictions. Therefore, the demand of a product depends on many factors such as price, popularity, time, outlet type, outlet locationetc. Forecasting sales become hard manually when the number of factors increases. Demand prediction is also closely related to Sales revenue. 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 to expected sales for the region and outlet type [2].Thus, providing companies with the predicted sales for products and different outlet locations helps companies to formulate a proper business model which helps them toorganize and dispatch its

product more efficiently thus cuttingdown on costs and increasing revenue.

In Section 2 we have discussed the system overview and in section 3 we have described feature extraction. Section 4 describes our system algorithm and data definitions. And in section 5we conclude this work and future work.

II. SYSTEM OVERVIEW

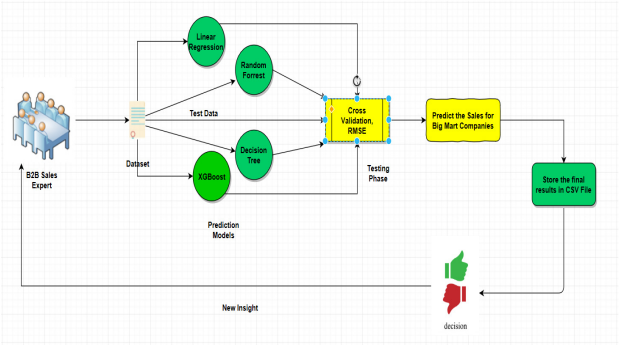

Fig. 1: High-level overview of the presented intelligent system.

In this model, a five-step procedure is used to solve the problem of predicting the Sales revenue for different products at different outlet locations for Big Mart Companies. First, the data is acquired, collected and divided into training and test label. This data undergoes a preliminary analysis which includes univariate and bivariate analysis. In the second stage, data pre-processing is performed which takes care of missing and erroneous values in the dataset. In the third stage, the features are selected and modified to get the best results. In the fourth stage, feature transformation is used to convert categorical features to numerical features. In the fifth stage, using various algorithm techniques models are built and the results are evaluated. These results are communicated to the firm and finally, after approval the results are applied by the firm to generate a business model for next year. Using this method guarantees more accurate and better results.

III. FEATURE EXTRACTION

Correlation and Data mining is used for feature selection over here. Feature extraction is the technique of extracting features of the data like its mean, variance, standard deviation, entropy etc.

A. Motivation

After collecting and integrating the dataset that is acquired, the nature of data needs to be understood. For doing so correlations, general trends and outliers need to be identified by calculating the mathematical statistics (mean, median, range, standard deviation, etc.) for each feature. By doing so the features that affect the sales of a product can be shortlisted. Visualization techniques are also used to support the needed. After the preliminary analysis missing, duplicate and inconsistent values are checked for and corrected. The features that can be grouped together or those that are not needed are filtered. Dimension analysis further enhances the feature selection approach. If this step is not performed when a lot of unrequired features would be analysed in the further steps and might produce a major difference in the result obtained. This makes working with the dataset easier and faults tolerant.

B. Data Cleaning

After understanding the nature of the data and finding a correlation between different features and target variable i.e. sales. The erroneous values in the dataset need to be replaced with values that make sense, the missing values need to be replaced with appropriate numerical or categorical value depending on the type of feature [5]. The redundant information in the dataset is to be removed. This fills the gaps within the dataset and makes it wholesome, which enables better results.

C. Feature Transformation

Data cleansing gives us a wholesome error-free dataset to work with, Feature Transformation is the family of algorithms used to create new features from existing features, in this we use a

linear combination of two or more features to make a new feature, this new feature gives better results with respect to target variable i.e. sales.

This system also uses categorical feature transformation to numerical feature transformation. The redundant features are dropped from the dataset for the new ones.

IV. SYSTEM ALGORITHM

Various algorithms are used to predict highly accurate results. In the following section all the algorithms used are described:

1. Linear Regression:

The basic idea of this algorithm is to fit a straight line between the selected features in training dataset and a continuous target variable i.e. sales. This algorithm finds a line that best fits the data.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x) [3]. So, this regression technique finds out a linear relationship between x (input) and y(output).

The equation of the regression line is represented by:

$h(Xi)=B0+B1Xi$
$h(Xi)$ is the expected value for i^{th} observation.

A. Algorithm

Input: Dataset with proper input and output labels
Output: Predict sales value and store in csv file

begin
i. Calculate mean, variance for the list of values
Def Mean(values):
Sum(values) divided by Len(values)
Calculate Mean_x, Mean_y
Def Variance (values, Mean):
sum([(values-Mean) ^2])
Calculate Variance_x, Variance_y

ii. Calculate covariance

Covar<=0
Def Covariance (x, Mean_x, y, Mean_y):
For length of x do:
Covar<= Covar + (x[i] - Mean_x) * (y[i] - Mean_y)
End

iii. Estimate coefficients

B0 <= covariance (x, Mean_x, y, Mean_y) / variance (x, Mean_x)
B1 <=Mean_y – B0*Mean_x

iv. Predict

For every X in the test set do:
Y <=B0+B1X
End
Store predicted values in CSV file

2. XGBoost Regressor:

XGBoost stands for e**X**treme**G**radient **B**oosting. The implementation of the algorithm was engineered for the efficiency of computing time and memory resources [4]. Boosting is a sequential technique which works on the principle of an ensemble. It combines a set of weak learners and improves prediction accuracy. At any instant t, the model outcomes are weighed based on the outcomes of previous instant t-1. The outcomes predicted correctly are given a lower weight and the ones miss-classified are weighted higher.

A. Algorithm

XGBoost's split finding a greedy algorithm

Input: Dataset with proper input and output labels, I instance a set of the current node.
Output: Predict sales value and store in csv file

Begin
i. Gain <= 0

ii. For every i belongs to I do:
G <= sum (g i), H <= sum (h i)

iii. For k=1 to m do:
GL <=0, HL <=0
For j in sorted (I, by Xjk) do:
GL <= GL + gj, HL <= HL + hl
GR <= G - GL, HR <= H – HL
Score <= [Score, GL^2/ (HL + λ) + GR^2/ (HR + λ) - G^2/ (H + λ)]
End
End

V. CONCLUSION

In this paper, 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. In the future, we will use the output of this project as part of the price optimization problem which we are planning to work on.

ACKNOWLEDGEMENT

We express esteemed gratitude and sincere thanks to our project guide Prof.Madhuri Ghuge, who has been a great mentor and has always helpedand encouraged us with extreme sincerity and affection. We are much obliged to ourhonorable Head of Department Dr. D. R. Ingle whose support and co-operation was always helpful and motivating. And also thank you to Project Coordinator Prof. ShivsagarGondil who supported us in numerous ways. Our parents to whom we are forever grateful for constant support and encouragement. As we give expression to our love and appreciation our heartis fill. And we in sincere appreciation of your valuable help.

REFERENCES

https://ieeexplore.ieee.org/document/8073492 [1]

https://www.leadingindia.ai/downloads/projects/BS/bs_7.pdf [2]

https://www.kaggle.com/ashydv/sales-prediction-simple-linear-regression [3]

https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/ [4]

https://bigdata-madesimple.com/machine-learning-sales-forecasting-tackle-insufficient-data-issue/ [5]

