**Sales Prediction System using Machine Learning**

Md.Ataur Rahman

Id: 16155032

Dept of it

University of information Technology and science

- - - - - - - - - - - - -- - - - - - - - - - - - - - - - --\*\*\*\*\*\*\*\*\* - - - - - - - - -- - - - - -- - - - - - - - - - -- - - - - - - - -- - -

**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 RESEARCH Data

The data has been taken from the Kaggle data analytics competition, it contains data of 45 Walmart

stores and its various departments. The original data files used for our analysis were stores.csv,

train.csv and features.csv which contained the below mentioned fields:

stores.csv: This file contains anonymized information about the 45 stores, indicating the type and

size of store.

train.csv: This is the historical training data, which covers to 2010-02-05 to 2012-11-01. Within

this file you will find the following fields:

* Store - the store number
* Dept - the department number
* Date - the week
* Weekly\_Sales - sales for the given department in the given store
* IsHoliday - whether the week is a special holiday week

features.csv: This file contains additional data related to the store, department, and regional

activity for the given dates. It contains the following fields:

* Store - the store number
* Date - the week
* Temperature - average temperature in the region
* Fuel\_Price - cost of fuel in the region
* MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is

running. MarkDown data is only available after Nov 2011, and is not available for all stores

all the time. Any missing value is marked with an NA.

* CPI - the consumer price index
* Unemployment - the unemployment rate
* IsHoliday - whether the week is a special holiday week

We merged the data files to our convenience for analysis which have been uploaded in our github

page. Methodology Selection

The methodologies which we used in this project are:

1. Merge different data sets: We had three data files – training, features and stores. We merged

all the three data files to see effect of different variables on sale. Since this data was already

cleaned, we didn’t do any data cleaning work.

2. Study summary descriptive statistics: We have studied how different factors like week,

month, store size, temperature effect sales by using ggplot function in R. An interesting

observation during descriptive statistics came up where we realized markdowns were also

impacting sales. But since markdown data wasn’t big enough for linear regression model,

we restricted ourselves with just descriptive statistics work.

3. Build linear regression models: We built linear regression in R to predict sales using week of the year, store size and temperature. We have used backward selection model to analyze the effects of various predictors on the sales.

Model Building

We have created separate dashboards to analyze variation of sales for departments with Week,

Month, Temperature, Store Size, Markdowns. Next, we tried to get insights by comparing the

factors like fuel prices, temperature, unemployment rate, CPI etc., with sales of the stores. We

found that there were not much insights by comparing the sales with unemployment, CPI and that

resulted in not much variation in the graphs. Store size, Temperature and Week of the year have

shown some interesting findings and it affected the sales by month significantly.

**SYSTEM OVERVIEW**

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

SYSTEM ALGORITHM:

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

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 ith observation.

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

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 Lecturer Md Mynouddin sir, who has been a great mentor and has always helped and encouraged us with extreme sincerity and affection. We are much obliged to our honorable Head of Department Al Shayokh whose support and co-operation was always helpful and motivating. As we give expression to our love and appreciation our heartis fill. And we in sincere appreciation of your valuable help.

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

**REFERENCE**

Shttps://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://www.kaggle.com/ashydv/sales-prediction-simple-linear-regression%20%5b3)]

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

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