Sales Forecasting on a Retail Dataset

Aleeza Inamdar[1], Harshitha M[2], Mihir Lakhani[3]

[1]Student of Computer Science , PES University, Bangalore, India, aleezainamdar2000@gmail.com

[2]Student of Computer Science , PES University, Bangalore, India, harshitha715@gmail.com

[3]Student of Computer Science , PES University, Bangalore, India, mihir.rajesh123@gmail.com

***Abstract*—**

Sales forecasting plays a very important role for a company/organization by providing insights on resource allocation, channelization of funds and workforce. Prediction of sales can be extremely beneficial to the company when making decisions about how and where to invest money. The focus of this study was to analyze a Retail Dataset and measure its performance with different Machine Learning Algorithms.

For analysis we have used two regression techniques- Linear Regression, Lasso Regression and two ensemble learning methods- XGBoost, Random Forest. In addition to evaluating the model with highest accuracy, the study also aims at examining whether the ensemble classifiers work better than single classifiers or vice-versa.

Keywords—Retail, linear regression, lasso regression, XGBoost, Random Forest, sales forecasting

I. INTRODUCTION

In today's competitive market, sales forecasting plays a predominant role in business planning because it can minimize risk in decision making to ensure smooth operation of the company.

Insufficiency in accurate financial data can often lead to lack of strategic planning to balance profit and expenditure in enterprises and organizations. Estimation of revenue and expenditure at the earliest is beneficial for the company which can further boost sales and marketing.

With the advancement in Big Data, there has been a major change in the way businesses use marketing strategies to attract their consumers. Recent studies and surveys have shown that utilizing Big data in the retail industry saw an 8 percent increase in profit and 10 percent reduction in the overall cost.

Monitoring and supervising the inventory proactively becomes as necessary and crucial as manufacturing products for a business. It can be seen that there is often inadequacy of information regarding demands for products which leads to inefficient management of inventory by enterprises and organizations. Situations like these can affect the company heavily as overflow of stocks and insufficient stocks can lead to wastage in funds and insufficient supply to consumers respectively. Sales forecasting at an early stage can help businesses to avoid these situations by providing insights for stipulation of products more accurately.

For this reason, consumer demand planning must start a few months prior because of the lead time for the procurement of products from suppliers.

Sales forecasting is not just a simple refill but an effective management and control to increase optimization of services to consumers, market penetration capacity and efficiency for companies.

Inaccurate sales forecasting can lead to inefficient investment for companies paving the way for substantial losses. To remedy this problem, this research proposes to find a method that is the best suitable for predicting the flow of sales out of the regression and ensemble techniques.

II. PREVIOUS WORK

As the volume and velocity of data is increasing day by day especially in the retail sector, Big Data has had a major breakthrough. Businesses and enterprises are trying to comprehend how the use of Big Data Analytics can help them in taking major decisions in their company. With the launch of tools like Hadoop, Hive and HBase volumes of data can be stored and processed easily. Technologies like Spark and Kafka can be used to process data in real-time which can be beneficial for retail in e-commerce sites [1]

Customer-behavioral analysis has also paved its way in the Retail Market that enables producers to analyze and track their consumers based on their purchase history and their average expenditure per se using concepts like RFM, Best-Worst Method and Clustering [3]

Traditional Regression Models like Linear Regression, Lasso Regression, Ridge Regression have been used to analyze the patterns in sales data [2] and these are used by various multi-category stores and e-commerce websites for the assessment or evaluation of their products. With the arrival of ARIMA and other Time-series models [4] analysis has become more convenient and effective since it considers external variables and all the additional factors affecting sales.

III. PROPOSED SOLUTION

In this study, we will be comparing some classic regression techniques to the ensemble methods of learning.

Relationship between independent attributes and the target attribute can be found using Simple linear Regression.

Lasso Regression is a regularized regression model which means that it makes slight modifications to the learning

algorithm such as shrinking its parameters so that the model generalizes better.

Ensemble learning is a widely used technique that seems to yield results with a high level of accuracy. The core crux of ensemble learning techniques is that several machine learning models are trained sequentially and the combination of results of each model is used to make predictions. Using bagging and boosting techniques have shown extremely good results.[5]

In this study, we will be using XG boost which is an advanced implementation of the Gradient Boosting Algorithm which has shown almost 10 times better results compared to other gradient boosting techniques.

We will also be using a Random Forest Regressor that uses the bagging technique and implements a decision tree model for its predictions.

IV. REGRESSION

This is the most widely used modelling in the fiel of machine learning. Regression is used when we want to determine the association or relation between a set of independent attributes and a target attribute.

Regression can be used for a variety of purposes. It can be used to identify patterns between a set of variables and a target class or it can also be used to make predictions and render valuable information regarding the future of a business. Regression can be of various types depending on the outcome one wants to achieve.
In our study we will be using four different types of regression.

A. Model 1- Simple Linear Regression

This is the simplest form of regression. An association between the target variable and one or more independent variables can be found using the formula:

Y= b0 + b1X

where, X is the independent feature, b1 is the slope, b0 is the y-intercept and Y is the dependent/target variable.
A linear relationship can be found between X and Y as depicted in the figure below by the straight line.

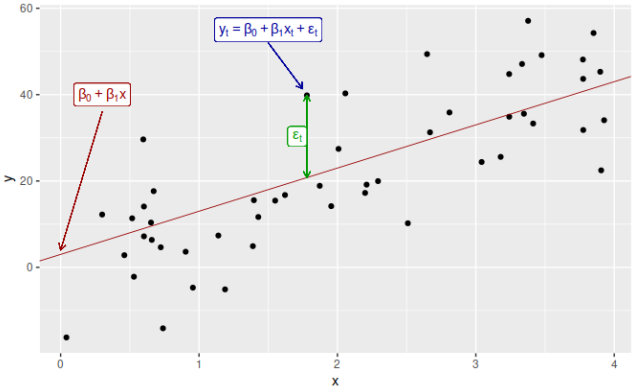

Figure 1. Simple Linear Regression

Our goal here is to find the best-fit line so that the least amount of miscalculations are made. The metric we're using to evaluate its performance can be a type of squared error such as RMSE or MSE. In order to minimize the error, b0 and b1 must be selected appropriately.
.

$$b_0 = \bar{y} - b_1\bar{x}$$

Figure 2. Calculation of intercept

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Figure 3. Co-efficient formula

A positive relationship exists between x and y if b1>0
A negative relationship exists between x and y if b1<0

B. Model 2- Lasso Regression

Also called Penalized Regression, this is another form of linear regression that is used when a subset of features must be selected from a whole dataset. It also uses the method of shrinkage of coefficients in order to minimize the variation in data samples.

This type of regression performs the L1 regularization which is a method where a penalty or a certain cost is added to the magnitude of coefficient values. This leads to obtaining a sparse matrix as a result where most of the values are zero and can henceforth be eliminated. The ideal solution to yield simpler models is to bring most of the coefficients closer to zero which can be incurred from larger penalties. This algorithm in used to minimize the coefficients:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Figure 3. L1 Regularization

Where λ is the tuning parameter.

1. No parameters are eliminated when $\lambda = 0$.
2. The values of coefficients tend to become zero as λ increases.
3. The bias increases as λ increases.
4. The variance increases as λ decreases.

C. Model 3- XGBoost

This algorithm uses a technique called as boosting where each model tries to learn from the errors of the previous model and is performed sequentially. It is an ensemble learning method that performs regression on our data. It is a decision-tree based ensemble.

It is also called as Regularized boosting technique because it performs regularization as seen in Lasso Regression and tends to reduce overfitting thereby increasing the performance of the model.

Boosting is implemented through three simple steps:

1. We define a preliminary model F0 which is used to predict the target class 'y'. (y-F0) is the residual for this model.

2. The residual (y-F0) from the previous step is fit on a new model named h1

3. F1 is the new model that is obtained from the combination of models F0 and h1 . The mean squared error and variance of the boosted model F1 is much lower than that of F0. The formula for obtaining F1 is as follows:

$$F_1(x) <- F_0(x) + h_1(x)$$

We can perform the above-mentioned steps up to 'm' iterations until the accuracy of the final model is improved. Here F_m denotes the latest model, F_{m-1} denotes the previous model and h_m denotes the residual obtained for the previous model.

$$F_m(x) <- F_{m-1}(x) + h_m(x)$$

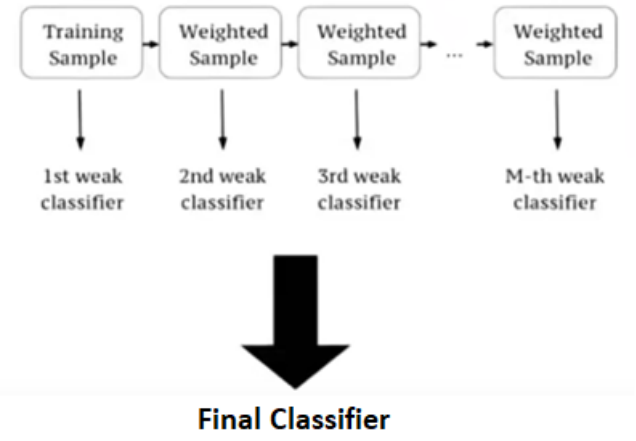

Figure 4. Boosting

D. Random Forest

Random forest is an ensemble learning method that makes predictions based on the combined results of multiple decision trees. Each decision tree stump runs on a subset of data from the original dataset obtained by random sampling. The results obtained from each decision tree is then added and a mean value is obtained. This technique is also called as bagging. The majority vote of each model is taken using the formula:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Each node can be split into multiple features or variables. A limit or a threshold is set on this number and is called as the hyperparameter.

Since each decision tree stump obtains a small part of the original dataset through the means of random sampling, we can sure that this element of randomness can help reduce overfitting.

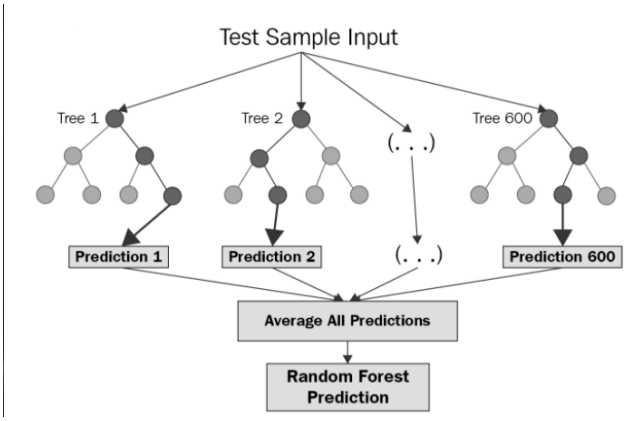

Figure 5. Bootstrap Aggregation/Bagging used in Random Forest

V. METHODOLOGY

The dataset is divided into two parts- a training set that is used to train the model and a testing set that is used to evaluate the model. All our models were fitted to the same training and testing data.

A. DATASET

In order to demonstrate the forecast of sales, we sourced a dataset from Kaggle.
The dataset has variables related to the sales of products.

Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
0 FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.1380
1 DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
2 FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.2700
3 FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	Tier 3	Grocery Store	732.3800
4 NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052

The parameters within the dataset are Item Identifier, Item Weight, Item Fat Content, Item Visibility, Item Type, Item

MRP, Outlet Identifier, Outlet establishment Year, Outlet Size, Outlet Location Type, Outlet Type.

The parameter 'Outlet Sales' is the Target variable and we will use four different models in order to predict its outcome.
The data contains 12 columns and 8524 rows

The data has been thoroughly cleaned before beginning the analysis. There were a few missing values which have been replaced by the mean of that particular column.
Similarly, categorical data and numeric data have been encoded successfully.

The vast dataset provides for a good range of data to split between train and test in order to obtain good results. We have ensured a 70-30 split for good analysis.

We have also performed Principal Component Analysis (PCA) on our dataset to reduce the dimensionality.
Since our data had no strong correlations in the beginning, PCA was essential in order to interpret the data in a more concise manner.

VI. PERFORMANCE OF THE ALGORITHMS

The metric we are using to evaluate the performance of our algorithms is Root Mean Squared error (RMSE).
The standard deviation of the error in the prediction value can be measured through RMSE. RMSE measures the standard deviation of the prediction errors. It gives us an information about how concentrated or scattered the data points are around the line of best-fit.

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N]^{1/2}$$

This is the formula that is used to measure RMSE where,
Zfi= Predicted output
Zoi=Actual output
N= Number of samples in data.

The model that gives the least value for RMSE will be considered as the most effective for forecasting sales.

A. LINEAR REGRESSION MODEL

RMSE: 1365.157887

Since we are dealing with a lot of parameters in our dataset, The RMSE turns out pretty high. We can conclude that Considering that Linear Regression models have high error trade-offs, using them is not that effective in case of our dataset.

B. Lasso Regression Model

RMSE: 1287.426210

This model produces results better than Linear regression Model.
Lasso Regression has the tendency to increase bias and lower the variance. Since it performs regularization and converges the coefficient of irrelevant variables to 0, it seems to be more effective than Linear regression.

C. XGBoost Regressor

RMSE: 1276.355176

As we know that ensembling techniques are well known for Their tendency to reduce variance, this model performs the best out of all the models.
Since boosting runs sequentially, and repetition of errors do not occur, we can say that the combined variance of Multiple weak regressors < variance of a single strong Regressor

SL NO.	MODEL	RMSE
1	XG Boost regressor	1276.355176
2	Lasso Regression	1287.426210
3	Random forest regressor	1332.699190
4	Linear regression	1365.157887

D. Random forest regressor

RMSE: 1332.699190

Since Random forest is a bagging technique, there is no Method to handle the repetition of errors because training takes place simultaneously. Random forest can be thought as an extension of decision trees.
In spite of having low-bias, we can observe that it does not do a good job on lowering the variance significantly and performs slightly worse than the Lasso Regression Model.

E. PERFORMANCE COMPARISON OF MODELS

We can observe that XG Boost Regressor performs the best compared to the other three models.

VII. CONCLUSION

In this paper, we aimed at forecasting sales of different items in a retail store. We used four different models- Linear regression, Lasso Regression, XGBoost regression and Random forest regression, out of which two of them were

classic regression models and two of them were ensemble learning techniques.
XGBoost regressor performed the best and we can conclude that ensemble learning techniques are very useful in predictive analysis and can lead to more precise and accurate estimation of data.

The dataset we used in this study is small. Results can be more clearly seen with a vaster and more usable dataset.

VIII.CONTRIBUTIONS

Aleeza -Data Pre-processing and Exploratory data analysis

Mihir- Principal component analysis and Regression Models

Harshitha- Ensemble learning techniques for the models and Model Evaluation

REFERENCES

[1] Hamza Belarbi, Abdelali Tajmouati, Hamid Bennis, Mohammed El Haj Tirari, “Predictive Analysis of Big Data in Retail Industry”, May 2020

[2] Mr. Faraz Hariyani 1, Mrs. Haripriya V 2 1MSc IT Student, Dept of MSc IT, JAIN (Deemed-to-be University), INDIA 2Assistant Professor. Department of MSc IT, JAIN (Deemed-to-be University)

[3] Rendra Gustriansyah, Ermatita, Dian Palupi Rini, Reza Firsandaya Malik, “Integration of Decision-Making method and Data Mining Method as a preliminary study of Novel Sales Forecasting method”, International Journal of Advanced Trends in Computer Science (IJATCSE), Volume 9, no. 4, July-August 2020

[4] Dr. Ravi Mahendra Gor, “Forecasting Techniques”, July 2011

[5] P. Subhashini, Yash Kimtani, Yuvraj Talukdar, Ajit Kumar Shah, “Sales forecasting using ensemble methods”, International Research Journal of Computer Science (IRJCS), Issue 04, volume 6, April 2019