Big mart Sales Prediction

by Manikya Sarashetti

Submission date: 01-Dec-2020 11:43AM (UTC+0530)
Submission ID: 1461154110
File name: PG26_Covariance-Final_Report.pdf (577.52K)
Word count: 2720
Character count: 13939

Big mart Sales Prediction

Manikya.Prakash Sarashetty
B.Tech ,CSE
Pes university, EC campus
manikyasarashetti@gmail.com

Ramyashree J.R
B.Tech , CSE
Pes university,EC campus
ramyarnayak0511@gmail.com

Nandini.R.Sonth
B.Tech, CSE
Pes university,EC campus
nandini.r.sonath@gmail.com

Abstract-In day today life malls and bigmarts collect their previous records in order to speculate the customer demand and not to get any shortage. In this paper, we started to experiment with different models to get the best fit model to forecast the sales of bigmart and gives highest accuracy. The final data will be useful to predict future sales with different machine learning techniques which will be useful for the retailers like Big Mart and other industries to increase their profit.

Keywords— Forecast Sales,Machine Learning Algorithms and techniques, Linear regression, random forest , Decision tree and XG booster.

I. INTRODUCTION

In this modern world all the organizations wants to increase their profit and revenue. At the end of the day, increase in the sales is the main goal of any organization. In this paper, we are trying to address the problem of big mart sales prediction to satisfy the consumer future demand in various big mart stores across various places and products based on the past records. In our work we used different machine learning techniques like XGBoost(proposed solution),Random forest,Decision Tree model and Linear regression Model to foresee the sales of bigmart.

ML Algorithms are capable of handling both non-linear data and huge dataset.We used Root Mean Squared Error (RMSE) as an evaluation metric to compare performance of models which we used . Here in this project where accuracy factor is most important to find the best model, we intend to use metrics parameter as accuracy measure for variables.

In this paper we worked on some of the above algorithms and tried to come up with a better model to predict sales.

The first and foremost technique used in predicting sale is the statistical methods and techniques, It is also called as traditional method. But the problem with these is it takes more time in predicting sales and it cannot handle non-linear data. So to overcome the problems in traditional methods machine learning techniques are introduced. Generally, in pre-processing of data raw data is converted into useful form of data to gain the better understanding of the data. Data preprocessing involves the following steps

- Data-cleaning.
- Data-transformation.
- Data-reduction.
- Feature Engineering

We can come to the conclusion that we can get to know about the strengths ,marketing plans,customer demands,which we can analyze from the previous records and situations.Sales prediction is all about the availability of resources from the past records. Indepth knowledge of history is essential for improving and enhancing the likelihood of marketplaces disregarding of any circumstances especially to the external circumstance, which is required to prepare for the upcoming needs for the business enhancement.

II.Literature survey

To solve our problem of bigmart sales prediction we Referred many research papers from the google scholar source.some of the insights are summarized below:

In paper[1] as we can see they have used Random forest regression approach and Xg booster approach to predict the sales of bigmart. They have also plotted graph to compare different models to predict the accuracy of the model.This boxplot mainly refers to the mean error produced by testing the algorithm of different models.
They have mainly focused on product level hypothesis.

In paper[2] they have used Xg boost technique to predict bigmart sales.The dataset is built on hypothesis

of item and store level. After visualizing and exploring the data ,data-set is divided into two parts, train dataset and test dataset in the ratio 80 : 20.
We got to know that the Xgboost has exclusive features like Sparse Aware (that is the missing data values are automatic handled) from this paper.

In paper[3] they have used random forest approach and linear regression model to solve the problem of predicting bigmart sales.We can infer accuracy is the key factor in predictive systems from this paper. There is a diagram which shows correlation between different factors which is very important to know about the variables to proceed with the building of model. The largest size store type did not produce the highest sales, in fact the location that produced largest sales is medium in size.

In this paper, we use comparison methodology in which raw data obtained at large mart will be pre-processed for missing data and outliers . Then machine learning algorithms 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 to predict the sales of each and every product in bigmart.

Other attempts-In this paper we are mainly trying to solve which method is best for forecasting sales of bigmart.Many others who have attempted to solve the same problem have tried to propose a model and then compare them with other models to get the correct model with the best accuracy.

III.Problem Statement

Our data analysis involves around predicting Bigmart Sales based on the characteristics or properties of all attributes. Our dataset contain the previous year's records using which we would be doing predictive analysis, to forecast the future sales.

"The next question in proceeding our project is to find out what are the characteristics of the items and how they will affect the sales of Bigmart which will be carried out by analyzing and understanding Big Mart sales.".To increase the sales of retailers we need to have a better understanding of the data about what attributes help in increasing sales.

Our next question definitely will be "What kind of variables should we use to predict sales and increase profit margin by precise model? " .Data visualization is helpful to find the useful information from the data through specific methodologies. Specifically, it looks for what kind of variables have impact on pastsales and if those influential variables change when the focus is shifted to sales of three specific regions.

Further we would like to model "The Sales Prediction based on how the attributes influence the sales". On analysis of the influence of each and every attribute we would be considering the most influential variables using data reduction and transformation techniques on the dataset.

A.Overview of data

Fig. 1. Steps followed for obtaining results

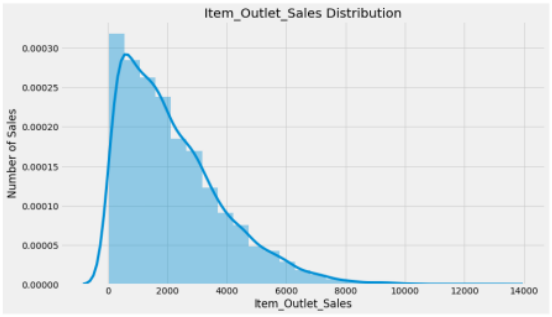
B.Data Description

We used bigmarket sales data as a dataset from the kaggle source in our work in which we have 13 attributes (when we combine both train and test dataset as one file). These 13 attributes define the basic features of the data which is being used for prediction. The dataset consists of 13 attributes .Among all the attributes target variable is the Item Outlet Sales attribute.

C. Data Exploration and Data VisualiZation

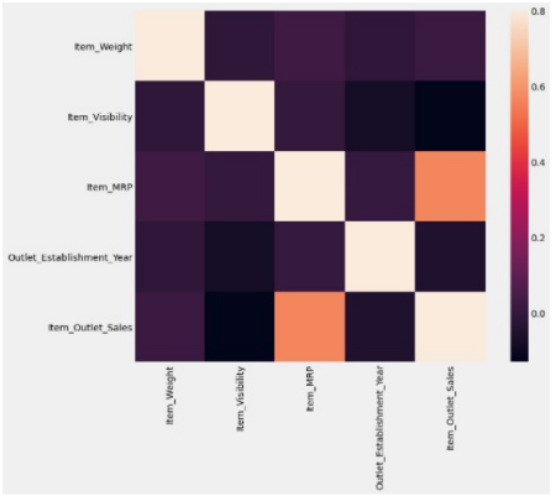
Data exploration is an important phase to collect useful information from the data which means to draw conclusions about the information that the data already has.

As you can infer from the diagram that Item outlet sales is a right skewd variable and it needs some data transformation to treat its skewness.



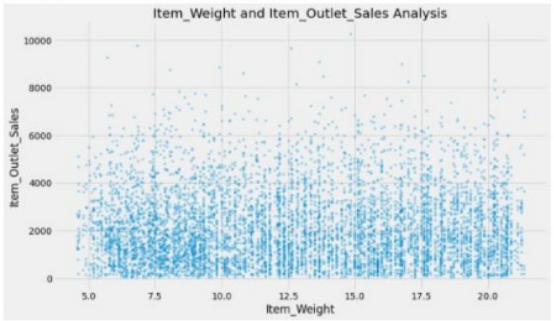
Data visualization is very important in data analysis As it is dependent on data that it can be done before or after data cleaning. In our case we visualized the data first to better understand the data and its characteristics. When we try to plot the correlation matrix we get to know that Item_Outlet_Sales is our

target variable.



Plot of Correlation matrix among the attributes of the data to see the which attributes have maximum and minimum correlation.

Next when we peek at the data we get to know about many insightful information that how other properties of the data affect the target variable.
Data visualization through graphs gives better knowledge about the data.



Item_Outlet_Sales and Item_Fat_content is spread across all regions.

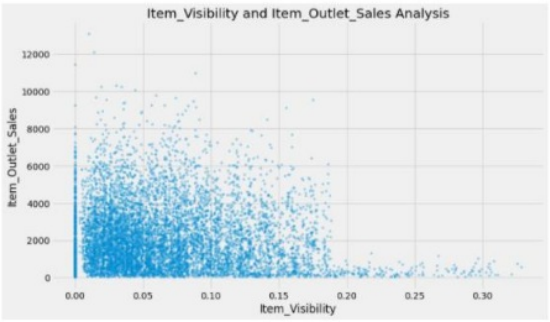


Fig. From the above plot we can infer that there are more purchases of less visible items.

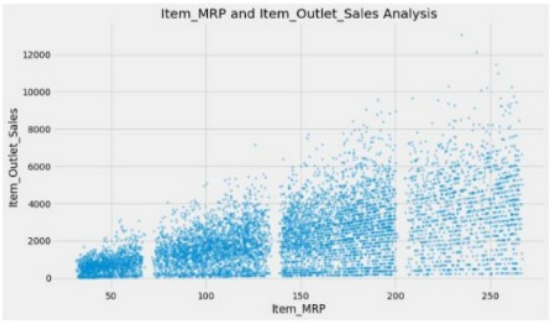
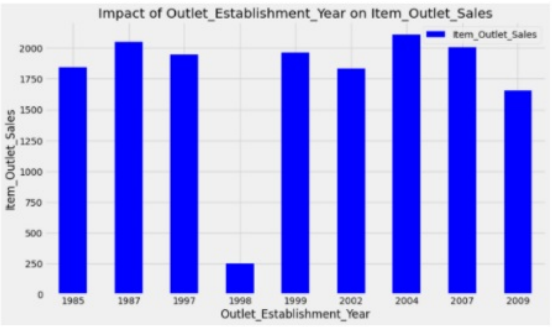
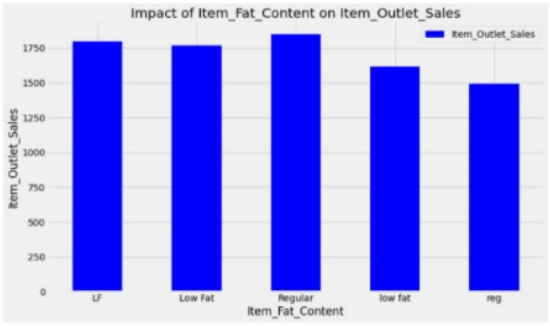
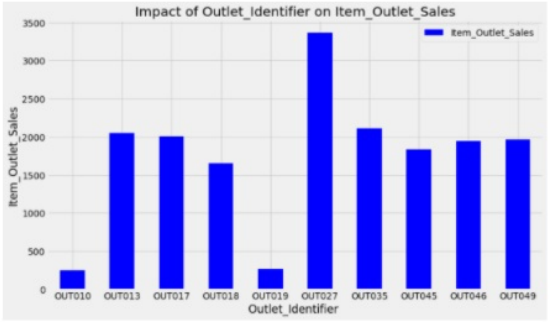


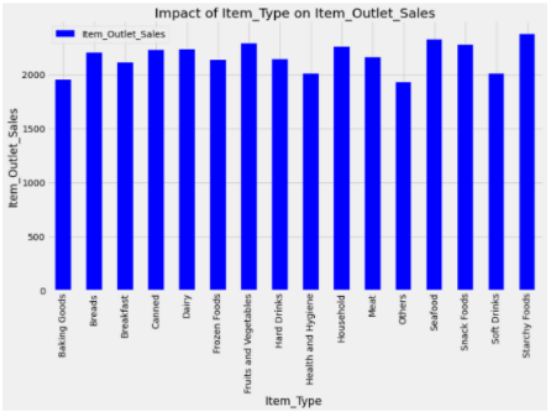
Fig. In the above plotwe can clearly see 4 segments of prices.



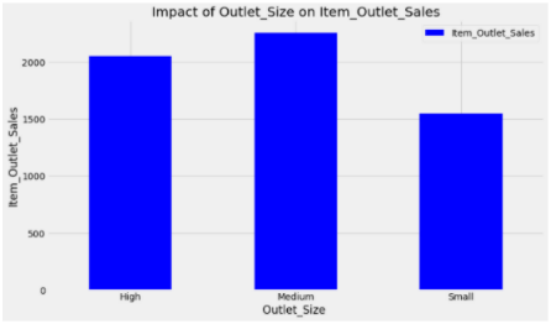
There is lesser number of sales in the year 1998.

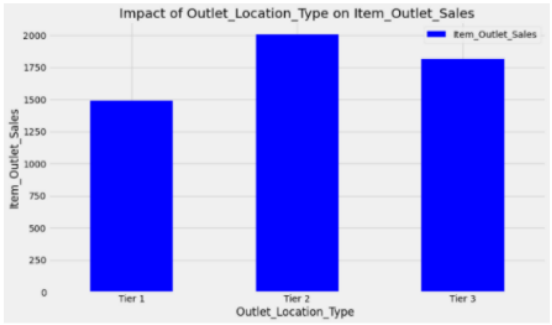


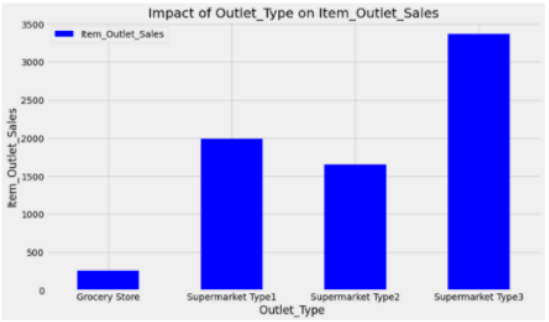




Scattering of Item_Outlet_Sales variable across the categories of Item_Type and Item_Fat_Content is not very distinct. The distribution for Outlet identifier 010 and 019 are quite alike and they are different from other outlet identifiers.







From the above plot we can tell that Grocery store has less sales compared to other outlet types.

D.Data Cleaning

Data cleaning is the vital step in preprocessing.In our dataset we had missing values of 3 attributes- item outlet sales,item weight and outlet size.We filled missing values of outlet size with the mode of that particular attribute.For the other two we fill missing values with the mean of the corresponding attributes.
The least value of item visibility is zero which is not possible practically so it should be taken care while cleaning the data.

E. 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 features can be combined for the better working of model. Feature engineering phase converts data into a form understandable by the algorithms.We drop the unnecessary columns and divide the data into train and test data.To continue our analysis we export our files to modified versions.

IV.Implementation and Result

A. Model Building

As mentioned earlier we use a comparative methodology like we try to build most of the models and compare them. We tried to use models like Linear Regression, Decision Tree model ,Random Forest and Xg booster approach to predict the sales of bigmart.By calculating RMSE(root mean squared error) we will come to conclusion about the best model to predict the sales.

1.Linear Regression model

Regression can be termed as a parametric technique which is used to predict a continuous or dependent variable on basis of a provided set of independent variables. This technique is said to be parametric as different assumptions are made on basis of data set.

In simple terms 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.

After modifying data we filt multiple linear regression to the training set.After predicting the test set results we perform cross validation and predict accuracy.With the linear regression model we got accuracy of 56.36% and RMSE of 1127.

2. Decision Tree Model

Decision tree model as the supervised learning algorithm can solve regression and classification problems.The main aim of the decision tree model is that it creates a training model to forecast the value of response variable by learning simple desion rules deduced from the data.With the decision tree model we got accuracy of 59% and RMSE of 1095.

3.Random Forest Model

Random forest also belongs to the group of supervised learning algorithms. The "forest" it builds, is a group of decision trees, usually taught with the "bagging" method. The conclusion of the bagging method is that a combination of learning models will increase the overall effect.
To put it in simple words : random forest assemble Different decision trees and group them to get an exact and invariable prediction.
With the Random Forest model we got accuracy of 61% and RMSE of 1062.

4.Xgboost approach

The gradient boosting decision tree algorithm is taken from XGBoost library.
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. algorithm. Features of XG Boost are,

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

With the Xgboost model we got RMSE of 1031.

Thus we conclude that Xgboost approach is the best model to predict the sales with the least root mean squared error.The RMSE is used as accuracy measure to find the best model.Lower values of **RMSE** indicate better fit.

V.Conclusion

We started **with** collecting appropriate dataset and made some hypothesis.Then we thought that data visualizing is vital so we started to look at the data and understand the properties of the data through graphs and charts which we plotted.Then comes the data cleaning step and in feature engineering we try to transform the data by dropping some columns and made one-hot encoding of variables which was required for the model. Then we started building different models and calculated their RMSE value.

As moving further towards data analysis we come to the point that understanding about the data is really vital to draw the conclusions from the information that the data contain.We got to know about the different models to forecast sales. As boosting is a grouping model that comes with an easy to read and interpret algorithm, making its predictions and interpretations easy to handle. The prediction or forecast ability is efficient through the use of its clone methods, such as random forest, and decision trees. Boosting is a flexible model and method that restricts over-fitting easily.
One disadvantage of boosting that it makes the model to fail is that it is sensitive to outliers since every classifier is required to fix the errors in the predecessors. Thus, the method is too dependent on outliers. Another disadvantage of the boosting approach is that the method is almost impossible to scale up. This is because every estimator is based on its correctness on the previous predictor sand forecast, thus making the course of action
difficult to streamline.

In these days where every part of the world is digitally connected and all the retailers want to know about consumers need ahead of time to not disappoint the customer due to shortage of the products in the store.Increasing competition between retailers day by day they are demanding for the accurate model to satisfy user demands and boost their store's revenue.The research on finding the accurate model is going on because the profit made by store is proportional to foreseeing the future sales so that the organizations doesn't suffer any loss. Our predictions help big marts to refine their methodologies and strategies which in turn helps them to increase their profit.We can use the output to help retailers like Bigmart.

References

[1] IMPROVIZING BIG MARKET SALES

PREDICTION, N Meghana, P Chatradi, A Chakravarthy, SM Kalavala, GITAM School of Technology, Bengaluru campus, Karnataka, India (URL: http://xajzkjdx.cn/gallery/423-april2020.pdf)

[2] A Comparative Study of Big Mart Sales Prediction, Gopal Behera, N Nain ,September 2019 ,Conference: 4th International Conference on Computer Vision and Image Processing, At: MNIT Jaipur(url : https://www.researchgate.net/publication/336530068_A_Comparative_Study_of_Big_Mart_Sales_Prediction)

[3] Parichay: Maharaja Surajmal Institute Journal of Applied Research Volume 3, Issue 1; January-June 2020, Sales Prediction Model for Big Mart 22 Nikita Malik, Karan Singh (Url: https://msi-ggsip.org/msijr/pdf/MSIJAR_VOL03_ISSUE01.pdf#page=26)

Authors

Manikya.Prakash.Sarashetty,
Student ,SRN:PES2201800301 Department of CSE PES University ,EC Campus

Ramyashree.J.R,
Student ,SRN:PES2201800728, Department of CSE PES University ,EC Campus

Nandini.R.Sonth,
Student ,SRN:PES2201800401 Department of CSE PES University ,EC Campus

Contribution of each Team member

1.Manikya Prakash Sarashetty – Data Exploration and Model Building

2.Ramyashree.J.R – Data visualization and Model Building

3.Nandini.R.Sonth- Data cleaning and Model Building

By working on this project we got to know about many things like how to handle the data, the insights we see as consumers are the result of a great deal of work,scraping data and problem solving skills.We will try to take our project to the advanced level to make our proposed solution to be best fit.

Big mart Sales Prediction

ORIGINALITY REPORT

4%	0%	3%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Gopal Behera, Neeta Nain. "Chapter 37 A Comparative Study of Big Mart Sales Prediction", Springer Science and Business Media LLC, 2020 Publication	2%
2	Reshma Boggavarapu, Pooja Agarwal, Rohith Kumar D.H. "Aviation Delay Estimation using Deep Learning", 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019 Publication	1%
3	Submitted to Universiti Sains Malaysia Student Paper	<1%
4	Kumari Punam, Rajendra Pamula, Praphula Kumar Jain. "A Two-Level Statistical Model for Big Mart Sales Prediction", 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 2018 Publication	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		