Big Mart Sales Prediction Using Machine Learning Techniques

Kaushal Dhingra[1], Dr.Anu Jaglan Rathee [2]

[1] Student,Department of Information and Technology,Maharaja Agrasen Institute of Technology,Delhi

[2] Faculty, Department of Information and Technology,Maharaja Agrasen Institute of Technology,Delhi

---*********************--

ABSTRACT

In this paper detailed analysis is done on each and every parameter of the data and predictions are made for the future sales using the Exploratory techniques of Machine Learning and Data Science.As we know forecasting of sales plays very crucial role for creating different strategies, marketing ,stocking of the products,Customer Lifetime Value(CLV),revenue generated and profits made.This research and exploratory analysis will help the company to plan a perfect strategy and help achieve good revenue sales and will provide a boom in the growth of the company in near future.Machine Learning along with Data Analysis gives better result than other methods.In the whole process Data Exploration,Data preprocessing,Data Cleaning,feature selection,feature transformation will play important role and will help give us effective output in terms of accuracy.

Keywords: Data Exploration,Data analysis,Machine Learning, Big Mart Sales data, Linear Regression model,Ridge Regression model,XGBoost

1. INTRODUCTION

In today's competitive world where every big company ,huge shopping malls,big marts like Walmart are recording our data related to the sales they made or the product we bought and at what time we bought,where we bought and many other things.All this data collected is of the great use because so many results can be derived and predictions can be made using this data.This data will help in predictions of future demands and resources and stock management.The dataset used has many dependent and independent variables and is a mixture of item attributes,data gathered through customers and also data related to stocks or inventory management.The data received is well defined in a form of sheet but is still raw for making the predictions so the data is therefore refined for making the predictions and also to get some interesting insights and results that can increase our knowledge.This will further be used for forecasting future sales by means of using machine learning algorithms such as the linear regression , ridge regression , random forests and XGBoost.

1.1 Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.
The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly.

1.2 Data Analysis or Exploratory Analysis

Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.Data exploration is an approach similar to initial data analysis, whereby a data analyst uses visual exploration to understand what is in a dataset and the characteristics of the data, rather than through traditional data management systems.Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data. ... An essential component of ensuring data integrity is the accurate and appropriate analysis of research findings.

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate.

1.3 Machine Learning Algorithms

Various Machine Algorithms that are used are for the predictions are:-

Linear Regression :- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting

Ridge Regression :- Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.Ridge regression is a technique used to eliminate multicollinearity in data models. In a case where observations are fewer than predictor variables, ridge regression is the most appropriate technique.

Decision Tree :- A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules.

XGBoost :- XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.XGBoost is a scalable and accurate implementation of gradient boosting machines and it has proven to push the limits of computing power for boosted trees algorithms as it was built and developed for the sole purpose of model performance and computational speed.

Contribution

Contribution of the paper can be summarized as follows:-

(1)Exploratory analysis of data will help see the factors which are more crucial for any store sales predictions and will help see tangible and intangible things of data.
(2) Feature engineering will help turn data into the format in which data is required.
(3) Machine Learning algorithms are used for the predictions.

2. LITERATURE REVIEW

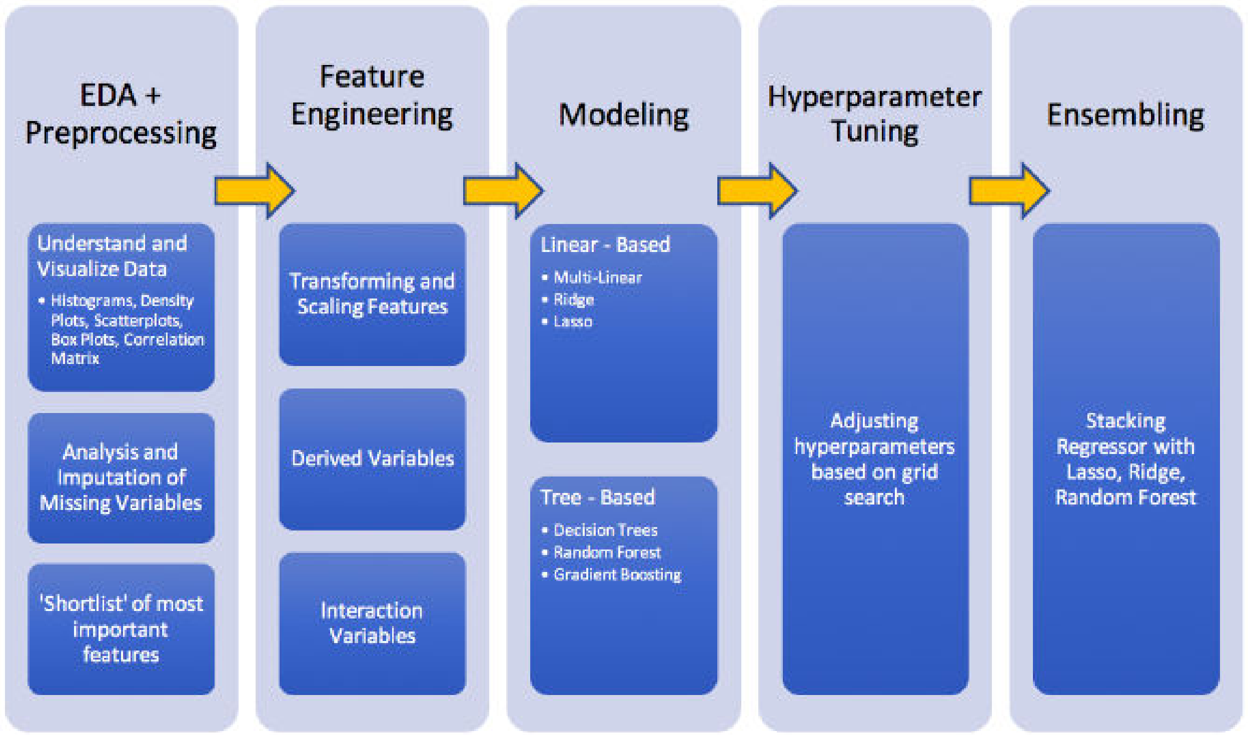
The data available is increasing day by day and such a huge amount of unprocessed data is needed to be analysed precisely, as it can give very informative and finely pure gradient results as per current standard requirements. It is not wrong to say as with the evolution of Artificial Intelligence (AI) over the past two decades, Machine Learning (ML) is also on a fast pace for its evolution. ML is an important mainstay of the IT sector and with that, a rather central, albeit usually hidden, part of our life [1].Sales forecasts provide insight into how a firm should manage its workforce, cash flow, and the means. This is an important precondition for the planning and decision-making of enterprises. It allows businesses to formulate their business plans effectively[2].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[3].In paper[4],general linear approach, decision tree approach and good gradient approach were used to predict sales. 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.Using the Random Forest, prediction of the sales is made easier and care is taken in fixing the optimum number of trees[5].Samaneh Beheshti-Kashi in his research reviewed different Various approaches on the predictive potential of consumer-generated content and search queries [6]. Gopal Behera has done effective study on Big mart sales prediction and has given prediction metrics for various existing models [7]. 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 [8].

3. METHODOLOGY FOR BIG MART SALES PREDICTION

Our goal is to identify the most important variables and to define the best regression model for predicting our target variable. Hence, this analysis will be divided into five stages:

1. Exploratory data analysis (EDA);
2. Data Pre-processing;
3. Feature engineering;
4. Feature Transformation;
5. Modeling;
6. Hyperparameter tuning
7. Ensembling.

Chart below illustrates the workflow divided into 5 stages :


EDA + Preprocessing
Understand and Visualize Data
• Histograms, Density Plots, Scatterplots, Box Plots, Correlation Matrix
Analysis and Imputation of Missing Variables
'Shortlist' of most important features
Feature Engineering
Transforming and Scaling Features
Derived Variables
Interaction Variables
Modeling
Linear - Based
• Multi-Linear
• Ridge
• Lasso
Tree - Based
• Decision Trees
• Random Forest
• Gradient Boosting
Hyperparameter Tuning
Adjusting hyperparameters based on grid search
Ensembling
Stacking Regressor with Lasso, Ridge, Random Forest

1.Exploratory Data Analysis :-The goal for this section is to take a glimpse on the data as well as any irregularities so that we can correct aht in the next section, **Data Pre-Processing.** In this we can check for duplicate values.Univariate analysis,Numerical predictors,correlation between numerical predictors and target variable,distribution of variable,bivariate analysis were also seen.

2.Data Pre-Processing :- Data preprocessing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects.Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, and missing values, etc...In data pre-processing missing values are being looked,then missing values are filled.

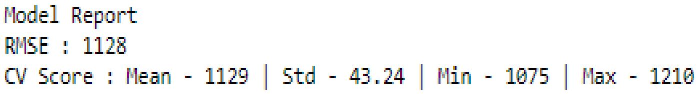
3.Feature Engineering :- Extracting of extra features from data is called feature engineering and various ideas were considered and implemented like should we combine the outlet_types of supermarket_1,2,3 into 1,does item_visibility 0 make sense?, creating a broad category of item_type,modifying categories of item_fat_content.

4.Feature Transformation :- It is modifying the data but keeping the information intact,this is done so that machine learning algorithms can understand it easier and give us better results.One-hot encoding is one of the step done in feature transformation

5.Model Building :- In this instead of repeating the codes again and again, a generic function which takes the algorithm and data as input and makes the model, performs cross-validation and generates submission is created.Models used for predictions on the dataset are 1. Linear Regression Model , 2.Ridge Regression Model ,3. Decision Tree Model,4.Random Forest Model and 5.XGBoost Model

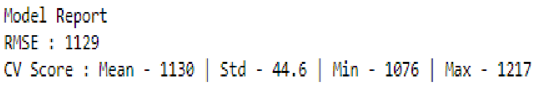
4. RESULTS

Linear Regression Model Result



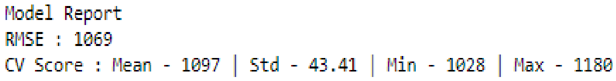
```
Model Report
RMSE : 1128
CV Score : Mean - 1129 | Std - 43.24 | Min - 1075 | Max - 1210
```

Ridge Regression Model Result



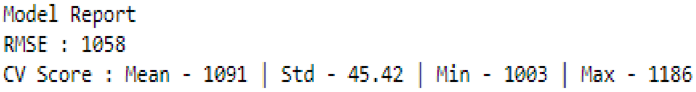
```
Model Report
RMSE : 1129
CV Score : Mean - 1130 | Std - 44.6 | Min - 1076 | Max - 1217
```

Decision Tree Model Result



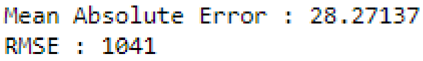
```
Model Report
RMSE : 1069
CV Score : Mean - 1097 | Std - 43.41 | Min - 1028 | Max - 1180
```

Random Forest Model Result



```
Model Report
RMSE : 1058
CV Score : Mean - 1091 | Std - 45.42 | Min - 1003 | Max - 1186
```

XGBoost Model Result



```
Mean Absolute Error : 28.27137
RMSE : 1041
```

5. CONCLUSION

Machine Learning algorithm that performed the best was XGBoost with RMSE = 1041.Hence in this paper sales prediction using machine learning algorithms are implemented.The outcome of the research done and algorithms applied will increase the efficiency and accuracy of the dataset and will help the company in a big way.

REFERENCES

[1] Smola, A., & Vishwanathan, S. V. N. (2008). Introduction to machine learning. Cambridge University, UK, 32, 34.

[2] S. Cheriyan, S. Ibrahim, S. Mohanan and S. Treesa, quote;Intelligent Sales Prediction Using Machine Learning Techniques,quote; 2018 International Conference on Computing,Electronics amp; Communications Engineering (iCCECE), Southend, United Kingdom, 2018, pp. 53-58

[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-DecisionMaking).

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

[5] T. Alexander and D. Christopher, quot;An Ensemble Based Predictive Modeling in Forecasting Sales of Big Martquot;, International Journal of Scientific Research, vol. 5, no. 5, pp. 1-4, 2016. [Accessed 10 October 2019]

[6] 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.

[7] 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.

[8] 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.