International Journal of Research in Engineering, Science and Management
Volume 6, Issue 4, April 2023
https://www.ijresm.com | ISSN (Online): 2581-5792

Global Superstores Sales Prediction and Data Visualization Using Power BI

Shruti Shivankar[1*], Shardul Mehetar[2], Neha Darade[3], Saachi Bhimanpalli[4], Dnyanada Dafale[5]

[1,2,3,4]*Student, Department of Computer Engineering, Terna Engineering College, Navi Mumbai, India*

[5]*Associate Professor, Department of Computer Engineering, Terna Engineering College, Navi Mumbai, India*

***Abstract*: Sales forecasting is an important aspect when it comes to companies who are engaged in retailing, logistics, manufacturing, marketing, and wholesaling. It allows companies to allocate resources efficiently, to estimate revenue of the sales and to plan strategies which are better for company's future. In this paper, predicting product sales from a particular store is done in a way that produces better performance compared to any machine learning algorithms. The dataset used for this project is Global superstore sales prediction from Kaggle. Nowadays shopping malls and Supermarkets keep track of the sales data of each individual item for predicting the future demand of the customer. It contains large amount of customer data and the item attributes. Further, the frequent patterns are detected by mining the data from the data warehouse. Then the data can be used for predicting the sales of the future with the help of several machine learning techniques(algorithms) for the companies like Big Mart. In this project, we propose a model using machine learning algorithm for predicting sales of companies like Big Mart and founded that it produces better performance compared to other existing models. An analysis of this model with other models in terms of their performance metrics is made in this project.**

***Keywords*: Data visualization, Business analytics, Machine Learning.**

1. Introduction

Supply and demand are two fundamental concepts of sellers and customers. Predicting the demand accurately is critical for the organizations to be able to formulate the plans. Big Mart is an online stop marketplace, where we can buy, sell, or advertise your merchandise at a very low cost. The major goal is to make Big Mart the shopping paradise for the buyers and the marketing solutions for the sellers. The goal is to multiply with the customers. The project "Global Superstore Sales Prediction and Data Visualization Using Power BI" aims to build a predictive model and find out the sales of each of the products at a particular store. The Big Mart can use this model to understand the properties of the products which plays a key role in increasing the sales. This can also be done on the basis hypothesis that should be done before looking at the data. The major aim of this machine learning project is to build a predictive model and to search out sales of each of the products at a particular selected store. Using this machine learning model, the Supermarket sales prediction tries to understand the properties of the products and stores which plays a key role in increasing the sales of products.

Sales Prediction is used to predict the sales of different products sold at various outlets in different cities of a Big Mart Company. Using this model, we will try to understand the properties of the products and stores which play a major role in increasing sales. Here python is used as programming language and Jupyter Notebook is used as tools. To build this application, machine learning aspects such as Supervised Learning task, Regression task are used. This is mainly done to predict the sales of a company stores in the future. The Various processes used are: Exploring the data and Data Pre-processing, Feature Engineering, Creating Model, Evaluation. Supervised learning helps you to understand the flow of the data and knowing the sale prices, etc. The regression task uses several different algorithms to predict the sales prices. It also includes task such as data visualization, cleaning, and transformation. Various Algorithms used are: Linear Regression, Multiple Linear Regression, Decision Tree Regression, XG Boost Regression, Random Forest Regression. The complete satisfaction of the customers.

The project "Global Superstore Sales Prediction and Data Visualization Using Power BI" builds a predictive model and finds out the sales of each of the product at a particular store. The Big Mart use this model to under the properties of the products which plays a major role in increasing the sales. This can also be done on the basis hypothesis that should be done before looking at the data.

2. Research Work

[1] In this paper, future sales of Big Mart Companies are predicted keeping in view the sales of previous years. For this, Machine Learning algorithms such as Linear Regression, K Nearest Neighbors algorithm, XGBoost algorithm and Random Forest algorithm have been used to predict the sales of various outlets of the Big Mart. Random Forest Algorithm is found to be the most suitable of all with an accuracy of 93.53%.

[2] In this paper, the three models k nearest neighbor Regression Model, Multinomial Regression Model and Decision tree Regression- Ada Boost Model were implemented. Multinomial and Decision Tree with Ada Boost regression models results in with an accuracy of 100%.

*Corresponding author: shrutishivankar10@gmail.com

[3] In this paper, three machine learning algorithms were implemented. Based on the performance accuracy the best algorithm is chosen for the prediction. Gradient Boost Algorithm is showing 98% overall accuracy and the second stands Decision Tree Algorithms with nearly 71% overall accuracy and followed by Generalized Linear Model with 64% accuracy.

[4] In this research paper, the four algorithms XGBoost Regression, ANN, Random Forest, SVR have been implemented out of which Random Forest regressor performs well compared to the other three algorithms with RMSE as 1171.429 and R2_score as 0.55.

[5] This paper outlines the sales forecasting by implementing three algorithms where Random Forest Algorithm has 89 percent accuracy, Decision Tree Algorithms is second with approximately 78 percent accuracy, and Linear Regression Model is third with 70 percent accuracy, according to the data.

[6] In this paper, different types of Machine Learning techniques such as Support Vector Regression, Gradient Boosting Regression, Simple Linear Regression, and Random Forest Regression have been evaluated on food sales data to find the critical factors that influence sales to provide a solution for forecasting sales. After performing metrics such as accuracy, mean absolute error, and max error, the Random Forest Regression is found to be the appropriate algorithm.

[7] In this paper, sales of a product is predicted using a two-level statistical model that reduces the mean entire error value. The two-level statistical model outperformed the other single model predictive techniques and donated better predictions to the big mart dataset.

[8] In this paper, Random Forest classifiers and regression have been implemented and the result shows that when the dataset is passed to both models their accuracy and other metrics vary a lot. Random Forest Regression works best for the dataset chosen.

[9] In this paper, dataset of Rossmann drug store's income statistics which is the second greatest drug save in Germany is considered. Data mining techniques like ARIMA models and XG Boost algorithm were used. Also, Linear Regression, Random Forest Regression were also considered. XGBoost algorithm excelled at prediction.

[10] In this paper, various algorithms like Decision tree, Deep learning Artificial Neural Networks (ANN), Naive Bayes, Random Forest. The Random Forest classifier had an accuracy score of 98% with a precision of 97%. The recall for the model is at 98%, with an F1 score of 98% and roc of 100%

[11] In this paper, the authors have implemented four different algorithms out of which XGBoost has the highest accuracy of 61.14%. Performance metrices like RMSE, Cross validation score, STD were also calculated.

[12] This paper provides an opportunity for accuracy of findings by implementing data exploration and analysis algorithms. Based on the performance, the XGBoost algorithm has showed 82% accuracy, RMSE value 5023 and therefore considered as the best fit comparing to other algorithms.

[13] This paper anticipates the purchases of a retail shop on Diwali sales and predicts the sales of various products based on their predictor factors by evaluation of consumer data. The algorithm used here is Random Forest Regression which has shown the most accuracy.

[14] In this paper, the authors have implemented a sales prediction model on the Indonesian footwear industry using real-life data. Their approach is based on the classification decision tree. The decision trees depict the results of data classification. View, price, and type are the significant variables to form the model.

[15] This paper provides insights forecasting sales of a big mart data. The authors have implemented a few algorithms out of which Gradient Boosted Tree has shown the most accuracy. It has shown 95.84% accuracy with an error rate of 41.6.

3. Proposed Methodology

In the proposed system, we consider multiple products from the multiple stores where stores are spread across the global. The sales of products are being analyzed in terms of the market, segments, country basis, state basis, city basis. So, after providing necessary inputs, the machine learning model gives the approximate sales value. This could be achieved by providing the desired dataset to ML model which will in turn help us in accurate predictions.

A. Description of Data

The dataset entitled Global Superstore Sales Prediction is used in our project. It consists of tables which include columns like:

- Order ID
- Order Date
- Ship Date
- Ship Mode
- Customer ID
- Customer Name
- Segment
- City
- State
- Country
- Postal Code
- Market
- Region
- Product ID
- Category
- Sub-Category
- Product Name
- Sales
- Quantity
- Discount
- Profit
- Shipping Cost
- Order Priority
- Sales

B. Project Modules

1) Machine Learning Module

Collecting Data: As you know, machines initially learn from the data that you give them. It is of the utmost importance to collect reliable data so that your machine learning model can find the correct patterns. The quality of the data that you feed to the machine will determine how accurate your model is. If you have incorrect or outdated data, you will have wrong outcomes or predictions which are not relevant.

Preparing the Data: After you have your data, you must prepare it by:

- Putting together all the data you have and randomizing it. This helps make sure that data is evenly distributed, and the ordering does not affect the learning process.
- Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc. You might even have to restructure the dataset and change the rows and columns or index of rows and columns.
- Visualize the data to understand how it is structured and understand the relationship between various variables and classes present.
- Splitting the cleaned data into two sets - a training set and a testing set. The training set is the set your model learns from. A testing set is used to check the accuracy of your model after training.

Choosing a Model: A machine learning model determines the output you get after running a machine learning algorithm on the collected data. It is important to choose a model which is relevant to the task at hand. Over the years, scientists and engineers developed various models suited for different tasks like speech recognition, image recognition, prediction, etc. Apart from this, you also must see if your model is suited for numerical or categorical data and choose accordingly.

Training the Model: Training is the most important step in machine learning. In training, you pass the prepared data to your machine learning model to find patterns and make predictions. It results in the model learning from the data so that it can accomplish the task set. Over time, with training, the model gets better at predicting.

Evaluating the Model: After training your model, you must check to see how it’s performing. This is done by testing the performance of the model on previously unseen data. The unseen data used is the testing set that you split our data into earlier. If testing was done on the same data which is used for training, you will not get an accurate measure, as the model is already used to the data, and finds the same patterns in it, as it previously did. This will give you disproportionately high accuracy. When used on testing data, you get an accurate measure of how your model will perform and its speed.

Parameter Tuning: Once you have created and evaluated your model, see if its accuracy can be improved in any way. This is done by tuning the parameters present in your model. Parameters are the variables in the model that the programmer generally decides. At a particular value of your parameter, the accuracy will be the maximum. Parameter tuning refers to finding these values.

Making Predictions: In the end, you can use your model on unseen data to make predictions accurately.

2) Data Visualization Module

Connect with Data Sources: Power BI Desktop data sources include Excel spreadsheets, CSV files, Q-data feed, online services, and data on the cloud. This data could be structured, semi-structured, or unstructured, depending on the business type. The first step in the process is connecting with the above data sources to import data. After importing the data from different sources, you can move forward with transforming and filtering the data based on various constraints.

Transform Data & Create Models: Using Power Query Editor, you can extract valuable information, remove anomalies, and add some conditions for a better understanding of the data. It is like sculpting a block of wood by cutting the edges, removing extra wood, shaving off the projections, and adding other ingredients to make it look as intended. You can also change columns and data types and add default values into the columns with null values.

Create Visuals: Visuals are the graphical representations of the data you stored in a model. Microsoft Power BI Desktop provides drag-n-drop features through which you can visualize the raw business data in the form of charts, graphs, maps, and KPIs. After creating the visuals, they can be attached to the dashboards or live reports in the form of tiles. Custom visuals also help you identify the problems in various departments and the market behavior and make better decisions based on them.

C. Evaluation Metrices

The machine learning regression algorithms are used in this project. So, the evaluation metrics are as follows:

1) *Mean Absolute Error (MAE):* MAE is a very simple metric which calculates the absolute difference between actual and predicted values.
2) *Mean Squared Error (MSE):* MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value.
3) *Root Mean Squared Error (RMSE):* As RMSE is clear by the name itself, that it is a simple square root of mean squared error.
4) *R Squared (R2):* R2 squared is also known as Coefficient of Determination or sometimes also known as Goodness of fit. R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform. In contrast, MAE and MSE depend on the context as we have seen whereas the R2 score is independent of context. So, with help of R squared we have a baseline model to compare a model which none of the other metrics provides.
5) *Adjusted R Squared:* The disadvantage of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decreases because it assumes that while adding more data

variance of data increases. But the problem is when we add an irrelevant feature in the dataset then at that time R2 sometimes starts increasing which is incorrect. Hence, to control this situation Adjusted R Squared came into existence.

4. Results and Dimensions

We have used K-Fold method for evaluating our model with the help of the following dimensions. Below are the observations from the various implemented models.

Dimensions: Segment, Market, Region, Category, Sub-Category, Product Name, Quantity.

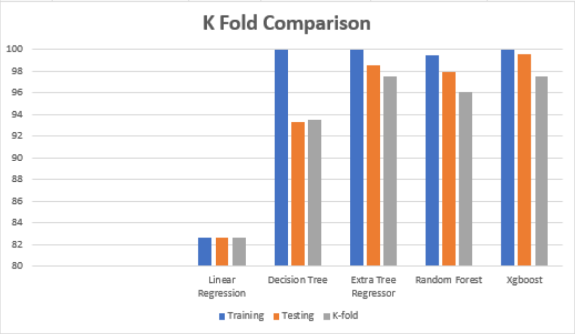

Fig. 1. K-Fold comparison

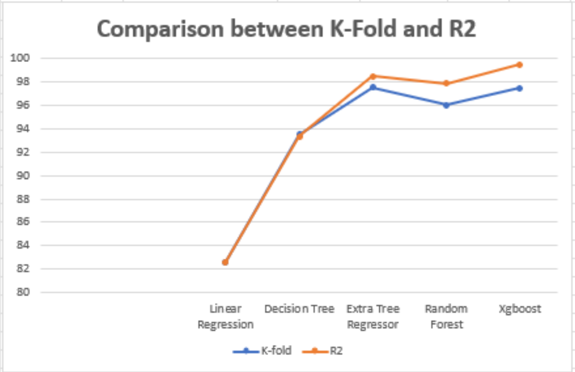

Fig. 2. Comparison between K-Fold and R2

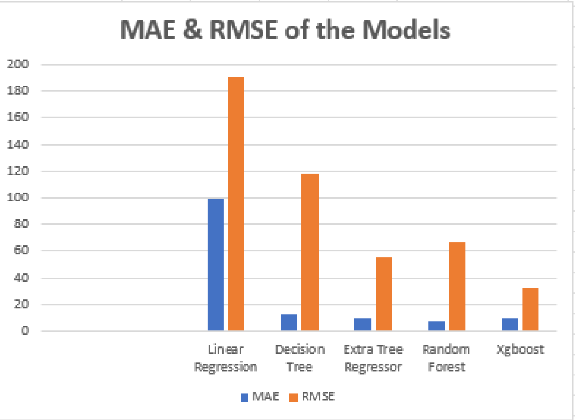

Fig. 3. MAE & RMSE of the models

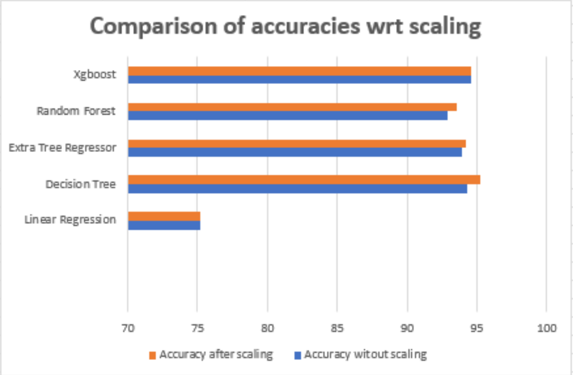

Fig. 4. Comparison of accuracies wrt scaling

5. Conclusion

In this project, basics of machine learning and the associated data processing and modelling algorithms are described, and their application in predicting sales of different Big Mart shopping outlets. The implementation, show the correlation among different attributes considered and how a particular location of medium size recorded the highest sales, suggesting that other shopping locations should follow similar patterns for improved sales. Multiple instances parameters and various other factors can also be used for predicting the sales more innovatively and successfully. Accuracy plays a major role in prediction systems, can be significantly increased when the parameters used are increased. Also, how the sub-models work also can lead to improving the productivity of the system. As the profit made is directly proportional to the sales predictions made accurately, the big marts aim accurate predictions so that the company will not suffer any losses. In this project, we have designed a model by Xgboost technique, linear regression, random forest etc. and experimented it on the Big Mart 2013 dataset for sales prediction of the product of a particular outlet. Experiments support that our technique produces more accurate prediction compared to than other available techniques like decision trees, ridge regression etc. Also, the visualization of chosen data has been done in Power BI to gain the insights. To conclude, our system predicts the near to correct sales for the products globally.

References

[1] Prajwal Amrutkar, Shubhangi Mahadik, “Sales Prediction Using Machine Learning Techniques”, International Journal of Research Publication & Reviews, Volume 3, 2022.

[2] Varshini S., D. Preethi, “An Analysis of Machine Learning Algorithms to Predict Sales”, International Journal of Science and Research, 2022.

[3] Aneesh Tony, Pradeep Kumar, Rohith Jefferson, Subramanian, “A Study of Demand and Sales Forecasting Model Using Machine Learning”, Psychology and Education, 2021.

[4] Bandaru Srinivasa Rao, Kamepalli Sujatha, Nannpaneni Chandra Sekhara Rao, T. Nagendra Kumar, “Retail Sales Prediction Using Machine Learning Algorithm”, Turkish Online Journal of Qualitative Inquiry (TOJQ), Volume 12, 2021.

[5] Purvika Bajaj, Renesa Ray, Shivani Shedge, Shravani Vidhate, “Sales Prediction Using Machine Learning Algorithms”, International Research Journal of Engineering and Technology, Volume 7, 2020.

[6] Akshay Godse, Poonam Pawar, Sairaj Sawant, Shirin Mujawar, “Intelligent Sales Prediction Using Machine Learning Techniques”, IRJECE, Volume 7, 2019.

[7] Sai Nikhil Boyapati Ramesh Mummidi, "Predicting Sales Using Machine Learning Techniques", Blekinge Tekniska Hogskola, 2020.

[8] A. Bhuvaneswaria, T.A. Venetiaa, "Predicting Periodical Sales of Products Using a Machine Learning Algorithm", International J. Nonlinear Anal. Appl., Volume 12, 2021.

[9] B. Sri Sai Ramya, K. Vedavathi, "An Advance Sales Forecasting Using Machine Learning Algorithm", International Journal of Innovative Science and Research Technology, 2020.

[10] Kenneth Ofoegbu, "A Comparative Analysis of Four Machine Learning Algorithms to Predict Product Sale for A Retail Store", Dublin Business School, 2019.

[11] Vidya Chitre, Shruti Mahishi, Sharvari Mhatre, Shreya Bhagwat "Big Mart Sales Analysis", International Journal of Innovative Technology and Exploring Engineering, Volume 11, Issue 5, April 2022.

[12] Naveen Kumar R, Jegan J, Yogesh V, Kavitha S, "Sales Prediction Analysis", International Research Journal of Engineering and Technology, Volume 8 Issue 5, May 2021.

[13] Swapna G, Adarsh K, Aniketh H, Latha V, M. Sreelakshmi, "Diwali Sales Prediction using machine Learning", Journal of Emerging Technologies and Innovative Research, Volume 9, Issue 3, March 2022.

[14] Raden Johannes, Andry Alamsyah, "Sales Prediction Model Using Classification Decision Tree Approach for Small Medium Enterprise Based on Indonesian E-Commerce Data, School of Economic and Business, Telkom University, 2015.

[15] Sanjay N. Gunjal, D. B. Kshirsagar, B. J. Dange, H. E. Khodke, C.S. Kulkarni, "Machine Learning Approach for Big-Mart Sales Prediction Framework", International Journal of Innovative Technology and Exploring Engineering, Volume 11, Issue 6, May 2022.