CS 5530-0002 Principles of Data Science
Spring 2023 Semester; University of Missouri - Kansas City

Prediction and Analysis of Sales of Products by Outlets

Based on the BigMart sales dataset of some stores during 2013

Ayushman Jeet Das
Graduate Student
School of Science and Engineering, UMKC
Kansas City, USA

Vishal Dhatrika
Graduate Student
School of Science and Engineering, UMKC
Kansas City, USA

Sahar Razzazi
Graduate Student
School of Science and Engineering, UMKC
Kansas City, USA

***Abstract*—This research paper aims to address the problem of inaccurate sales forecasting in the retail industry by proposing the development of predictive models using the 2013 BigMart sales dataset [1]. The objective of this project is to optimize inventory management and maximize profitability by accurately forecasting sales, identifying patterns and trends in sales data, and resolving risks associated with certain products. The methodology for this project involves data preprocessing, feature engineering, model development, and model evaluation, using Python/R scripts to wrangle, clean, and preprocess the data. Various machine learning algorithms will be trained and tested to develop accurate predictive models, and visualization techniques will be used to gain insights from the data. The results of this research may have implications for the wider retail industry in terms of improving inventory management practices and enhancing customer satisfaction.**

***Keywords*—Sales prediction, Retail industry, BigMart sales dataset, Inventory management, Profitability, Predictive modeling, Data preprocessing, Feature engineering, Machine learning algorithms, Data visualization.**

I. INTRODUCTION

The ability to accurately forecast sales is critical in the retail industry to optimize inventory management and maximize profitability. Inaccurate sales forecasting can lead to excess inventory, which incurs higher storage and handling costs, as well as increased risk of product spoilage and obsolescence. On the other hand, understocked inventory leads to lost sales and dissatisfied customers, who may seek out competitors with better stock availability. [2] Therefore, accurate sales forecasting is crucial for retailers to stay competitive and meet customer demands.

This research paper focuses on developing predictive models for sales forecasting and analysis using the 2013 BigMart sales dataset. The objective of this project is to identify patterns and trends in sales data, and use them to develop accurate predictive models that can be used to optimize inventory management and increase revenue. [3] The steps involved in achieving this objective include data preprocessing, feature engineering, model development, and model evaluation. By gaining insights from the data, such as item category by outlet size, total item sales by outlet type, and number of items per category, this study aims to provide retailers with the necessary tools to drive precision in sales forecasting and identify and resolve risks associated with certain products.

II. DATASET

The dataset used for this research paper is the 2013 BigMart sales dataset, which contains sales data for a fictional BigMart retail chain. The dataset consists of 8,523 records with 12 features, including 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, and Item_Outlet_Sales.

The dataset includes information about various products sold in BigMart stores, such as their weight, fat content, visibility, and type. It also contains information about the stores, including their location, size, establishment year, and type. The target variable in this dataset is Item_Outlet_Sales, which represents the sales of each item sold in the stores.

The dataset provides a comprehensive view of the sales data and allows for the analysis of sales trends across different outlets, items, and categories. With this dataset, retailers can identify the most profitable items and outlets, understand customer preferences, and optimize inventory management strategies. Furthermore, this dataset can be used to develop accurate predictive models to forecast sales, drive precision in inventory management, and increase revenue.

Column Name (Labels in Dataset)	**Description**
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The percentage of the total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which the store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in a particular store. This is the outcome variable to be predicted.

Table 1. Dataset Column names and description

III. Data Preprocessing

Data preprocessing is a crucial step in preparing a dataset for machine learning models. In this paper, we discuss three important aspects of data preprocessing: data cleaning, data encoding, and data scaling.

A. Data Cleaning

Data cleaning involves identifying and correcting or removing errors and inconsistencies in the data. In our methodology, we first identified missing values and filled them using appropriate values based on the column type. We then replaced zero values in the 'itemvisibility' column with the mean and merged repetitive values in the 'item_fatcontent' column. Additionally, we created a new feature, 'item_category,' by extracting the first two characters of the 'item_identifier' column and added a new category for 'item_fatcontent' called 'Non-Consumable' for non-consumable items. Finally, we created a new feature, 'outletyears,' to determine the age of each outlet.

B. Data Encoding

Data encoding is necessary because machine learning models require numerical data as input. We discussed three techniques for encoding categorical data: one-hot encoding, label encoding, and target encoding. One-hot encoding creates new binary columns for each category in a categorical variable, label encoding assigns a numerical value to each category, and target encoding assigns the mean of the target variable for each category in a categorical variable.

C. Data Scaling

Data scaling is important because some machine learning algorithms are sensitive to the scale of the input features. We discussed two techniques for scaling data: standard scaling and min-max scaling. Standard scaling transforms the data so that it has a mean of 0 and a standard deviation of 1, while min-max scaling transforms the data so that it has a minimum value of 0 and a maximum value of 1. The appropriate scaling technique depends on the distribution of the data.

In conclusion, data preprocessing is an essential step in preparing a dataset for machine learning models. It involves data cleaning, data encoding, and data scaling. By following a systematic and thoughtful approach to these steps, we can improve the quality and accuracy of machine learning models.

IV. Framework

We predicted outlet sales using various regression techniques. In particular, four different algorithms were used. Linear regression, random forest regression, decision tree regression, support vector regression, and artificial neural network.

According to the regression models, the independent variables predict the dependent variables [4]. Regression analysis estimates dependent 'y' variable values due to the range of independent variable values 'x' [5]. Outlet sales is the Y variable in this dataset. The model estimates the values of the regression coefficients that best fit the data, and then uses these coefficients to predict the values of the dependent variable for new data. We used linear regression as the base model.

Decision tree regression is a type of tree-based structure used to predict the numeric outcomes of the dependent variable. [6]

Decision Trees are great for obtaining non-linear relationships between input features and the target variable and is a type of regression analysis in which the predictor variable is split into two or more branches using a tree-like model. It works by recursively partitioning the data into subsets based on the most important features, and creating a tree-like structure of decisions that lead to a final prediction.

In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node [7].

Random forest is an ensemble of decision trees. This is to say that many trees, constructed in a certain "random" way form a Random Forest. Each tree is created from a different sample of rows and at each node, a different sample of features is selected for splitting. Each of the trees makes its own individual prediction. These predictions are then averaged to produce a single result.

Support vector regression: The algorithm works by finding the best hyperplane that fits the training data within a certain margin of error, and then using that hyperplane to make predictions on new data points. The final prediction is obtained by estimating the function that best fits the data points.

Artificial neural network is a technique that uses a network of interconnected nodes to make predictions. In this technique, the data is passed through multiple layers of nodes, with each layer processing the data in a different way. The final prediction is obtained by combining the outputs of all the nodes in the last layer.

V. Experiment

We first split the dataset into a 80% training set and a 20% test set in order to evaluate the performance of our models on unseen data and avoid overfitting. Following metrics used to compare regression models and evaluate them, MAE (Mean Absolute Error), R-squared, and RMSE (Root Mean Squared Error). Each of these metrics provides a different perspective on how well the model is performing. To determine which model is better, we compared the values of these metrics for each model.

R-squared measures the proportion of variance in the target variable that is explained by the model and indicates how well the model fits the data. R-squared ranges from 0 to 1, where 0 indicates that the model does not explain any of the variance in the target variable, and 1 indicates that the model explains all of the variance. R-squared can have negative values, which mean that the regression performed poorly [8]. A higher R-squared indicates that the model is performing better.

MAE measures the average difference between the predicted and actual values of the target variable. It represents the average absolute distance between the predicted values and the true values. It measures the normal size of the errors in a lot of forecasts, without thinking about their heading [9]. A lower MAE indicates that the model is performing better.

RMSE measures the difference between the predicted and actual values of the target variable. It is similar to MAE, but it takes the square root of the average squared difference between the predicted values and the true values and provides an estimation of how well the model is able to predict the target value. A lower RMSE indicates that the model is performing better.

In general, a model with a lower MAE and RMSE, and higher R-squared is considered to be better.

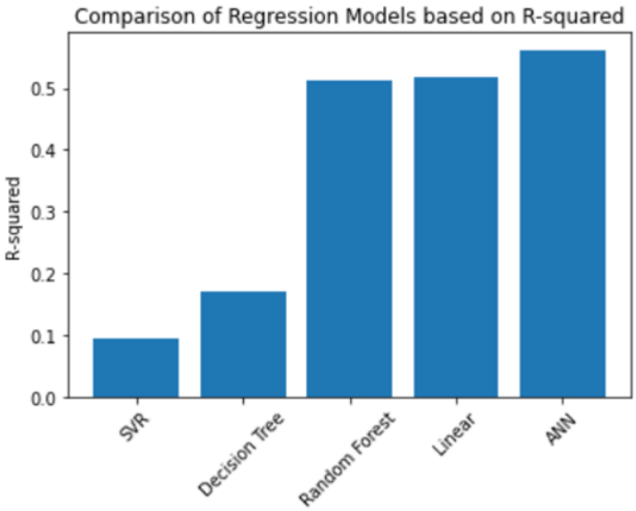

Figure 1. Comparison of R-squared values

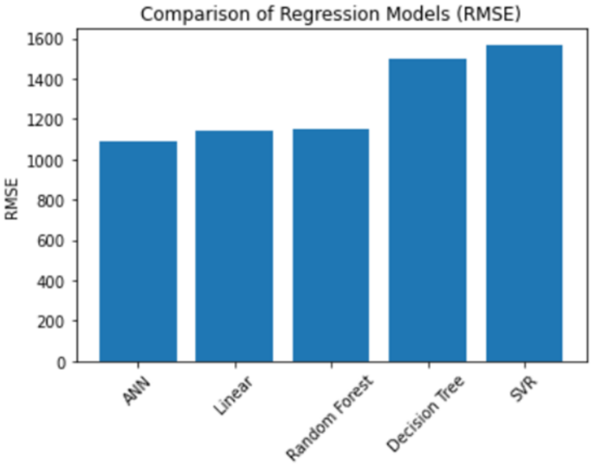

Figure 2. Comparison of RMSE values

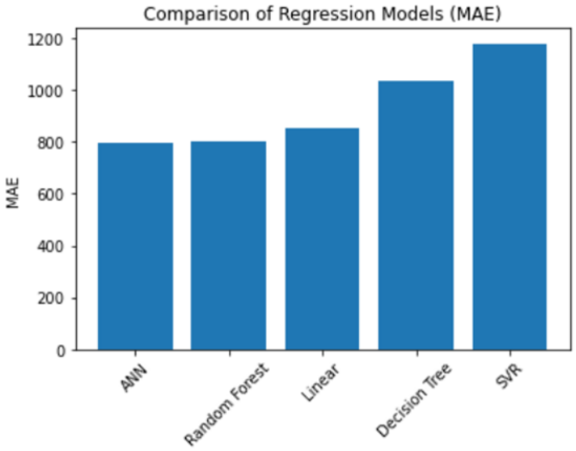

Figure 3. Comparison of MAE values

Based on the R-squared values of these models, the Artificial Neural Network has the highest R-squared value of 0.56, indicating that the model explains about 56% of the variance in the target variable. The Support Vector Regressor has the lowest R-squared value and explains about 9% of the variance in the target variable.

The RMSE metric measures the average difference between the predicted values and the actual values, squared. It is similar to MAE, but it penalizes larger errors more heavily. Based on the RMSE values of these models, we can see that the Artificial Neural Network has the lowest RMSE value, and the Support Vector Regressor has the highest RMSE of 1567.648.

For The MAE, A lower MAE value indicates that the model has better predictive accuracy. Based on the MAE values of these models, we can see that the Artificial Neural Network has the lowest MAE value and the Support Vector Regressor has the highest MAE.

Overall, based on those performance metrics, the Artificial Neural Network appears to be the best performing model, with the highest R-squared value and the lowest MAE and RMSE values. Also, lowest RMSE of 1091 in artificial neural networks suggesting that the predicted values are closer to the actual values compared to other models. The Linear Regression and Random Forest Regressor also performed relatively well, while the Decision Tree Regressor and Support Vector Regressor had lower performance scores.

VI. Conclusion

Overall, based on those performance metrics, the Artificial Neural Network appears to be the best performing model, with the highest R-squared value and the lowest MAE and RMSE values. The Linear Regression and Random Forest Regressor also performed relatively well, while the Decision Tree Regressor and Support Vector Regressor had lower performance scores. [10] These findings emphasize the importance of selecting appropriate models and performance metrics to optimize the accuracy of predictions. Further research is needed to investigate the effectiveness of these models in other scenarios and domains.

VII. References

[1] Shivan Kumar, "Big Mart Sales Prediction Datasets" Kaggle. https://www.kaggle.com/datasets/shivan118/big-mart-sales-prediction-datasets (Accessed May 13, 2023).

[2] R. Berman and A. Israeli, "The Value of Descriptive Analytics: Evidence from Online Retailers," SSRN Electronic Journal. Elsevier BV, 2020. doi: 10.2139/ssrn.3745748.

[3] KPMG, "Clarity on Data and Analytics" https://assets.kpmg.com/content/dam/kpmg/cl/pdf/2015-10-kpmg-chile-advisory-data-analytics-clarity.pdf (Accessed May 13, 2023).

[4] H. Roopa and T. Asha, "A linear model based on principal component analysis for disease prediction," IEEE Access, vol. 7, pp. 105314-105318, 2019.

[5] G. A. Seber and A. J. Lee, Linear regression analysis vol. 329: John Wiley & Sons, 2012.

[6] Rathore, Santosh Singh, and Sandeep Kumar. "A decision tree regression based approach for the number of software faults prediction." ACM SIGSOFT Software Engineering Notes 41.1 (2016): 1-6.

[7] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.

[8] Cameron, A. Colin, and Frank AG Windmeijer. "An R-squared measure of goodness of fit for some common nonlinear regression models." Journal of econometrics 77.2 (1997): 329-342.

[9] Arnaud de Myttenaere, Boris Golden, Bénédicte Le Grand and Fabrice Rossi Neurocomputing, 192 (2016), pp. 38-48

[10] Link to Google Colab IPYNB notebook "DS_Project.ipynb" https://colab.research.google.com/drive/1IP-DtImitJuZMYH8BqN8I94Vl6o77Vne?usp=sharing (Accessed May 13, 2023).