# 

ANALYSIS ON GLOBAL SUPER STORE DATA

BY

PHANIDHAR KASUBA VENKATA NAGA

DATE: 10/09/2023

# **INTRODUCTION**

Our shopping habits have significantly changed in a world that is facing challenges. How we buy goods and services has altered significantly as a result of the COVID-19 epidemic. As a result, online buying has evolved into more than just a preference. Understanding the complexity of e-commerce and retail operations is essential in such a circumstance. This research investigates the Global Superstore dataset in order to shed light on this changing environment. This collection includes data from years pertaining to different products, customers, and transactions.

The dataset being examined here is from the area of retail sales and management. It contains a lot of knowledge obtained from the activities of Global Superstore, a hypothetical yet representative retail organization. This dataset is crucial for comprehending how retail organizations operate at a time when conventional buying habits have undergone considerable disruption.

**SOURCE:** Vikas Kumar Appani (2023). Global Super Store [Dataset]. <https://data.world/vikas-0731/global-super-store>

The Global Super Store dataset, collected from Vikas Kumar Appani's repository and found online, contains approximately 49,000 observations and is a sizable collection of retail data. For this report, I have focused on a subset of the data that is only relevant to the United States of America, yielding roughly 9,996 records. Here’s the sample of the data.

Sample data:

A screen shot of a computer

Description automatically generated

Here in our data set, there are different types of data types such as Character, Numeric, POSIXct(date-time) below are the detail description of the data types involved in our dataset.

Columns Descriptions:

* Row ID (Numeric): Unique identifier for each row in the dataset.
* Order ID (Character): Unique identifier for each sales order.
* Order Date (Date Time): Date when the order was placed.
* Ship Date (Date Time): Date when the order was shipped.
* Ship Mode (Character): Shipping mode used for the order (e.g., Standard Class, Express).
* Customer ID (Character): Unique identifier for each customer.
* Customer Name (Character): Name of the customer.
* Segment (Character): Customer segment (e.g., Corporate, Consumer).
* Country (Character): Country where the order was placed.
* City (Character): City where the order was placed.
* State (Character): State where the order was placed.
* Postal Code (Integer): Postal code of the location where the order was placed.
* Region (Character): Geographical region of the order (e.g., West, East).
* Product ID (Character): Unique identifier for each product.
* Category (Character): Product category (e.g., Furniture, Office Supplies).
* Sub-Category (Character): Sub-category of the product (e.g., Chairs, Paper).
* Product Name (Character): Name of the product.
* Sales (Numeric): Sales amount for the order.
* Quantity (Numeric): Number of items sold in the order.
* Discount (Numeric): Discount applied to the order.
* Profit (Numeric): Profit generated from the order.

Our dataset comprises a diverse array of variables, each revealing a facet of retail operations. These variables include:

A screenshot of a computer code

Description automatically generated

As a crucial indicator of the Global Superstore’s financial performance, I choose "Profit" as the analysis's main outcome variable. However, a large number of predictor variables, including "Category," "Region," and "Ship Mode," are likely to have an impact on this result. This report embarks on a journey of hypothesis generation and testing. Our primary aim is to unearth valuable insights into the factors that affect the profit margins of the Global Superstore. While I will formulate hypotheses during the course of our exploration, one overarching hypothesis to test is whether certain categorical predictors, like "Category" or "Region," significantly impact the profitability of the store's operations. In the analysis of the goal variable "Profit" in the Global Superstore dataset, categorical factors (like "State" and "Category") and continuous covariates (like "Quantity" and "Discount") are used as predictor variables. Because they stand for particular categories that are anticipated to have a consistent impact on earnings, "State" and "Category" are referred to as fixed factors. Because I’m are not utilizing mixed-effect models, which require random effects, there are no random components in our analysis.

I tested our hypotheses in a linear regression analysis of the Global Superstore dataset with "Profit" as the target variable. I want to see if there is a significant linear relationship between predictor variables, such as categorical factors like "State" and "Category" and continuous covariates like "Quantity" and "Discount", and "Profit" which we predict as a product category of different countries may affect profitability, while sales volume may have a positive effect and discounts may have a negative effect The statistical significance of these correlations is checked with the help of linear regression to account for variable profitability accounts.

A screenshot of a computer screen

Description automatically generated

# **PRELIMINARY EXPLORATION OF THE DATA**

## Missing Values

Checking if there are any missing values:

A screenshot of a computer

Description automatically generated

## Exploratory Data Analysis

A graph of a bar chart

Description automatically generated

🡪 From the above bar chart, we can see that which of the Sub-Categories are most selling in the United States of America i.e., Phones and chairs. From the below graph we can see how the sales are over the years 2014-2018. We can observe that sales achieved their peak between 2014 and 2015, though it managed to achieve 17,500 sales during 2016-2017 sales are reducing with year-on-year basis.

A graph of a number of numbers

Description automatically generated

When analyzing the Global Superstore dataset for your report, scatter plots can be a useful visualization tool, especially when looking at correlations between different numerical variables.

A screenshot of a graph

Description automatically generated

Box plots are helpful for summarizing, visualizing, and pinpointing probable outliers in the distribution of numerical data. But here

A screenshot of a graph

Description automatically generated

Tree map according to profit generated by each state in United states. By this we can see the top profit-making states in United States of America are New York, California, Washington, Michigan, Virginia, Indiana, and Georgia.

A screenshot of a color palette

Description automatically generated

A pie chart of a company

Description automatically generated

* The above Pie-chart shows the profit accumulated by three categories namely Furniture, Office Supplies and Technology. We can clearly see that most of the profits are earned from the Technology and Office Supplies categories.

## Correlation:

Now I have dropped all the un-relevant columns and plotted a correlation Heatmap for the rest of the data which is as follows: A screenshot of a graph

Description automatically generated

Insights from the above correlation heat map:

* Profit and sales have a highly positive correlation, suggesting that as sales increase, profit tends to increase.
* Profits have a moderate positive correlation against Ship mode, Region, Day, indicating that as they increase the profits might increase accordingly.
* Profit may be mild positively correlated to the category Discount, Year, indicating that high discounts might be associated with higher quantities sold.
* Segment, Day Of Week and sub-category are negatively correlated with our target variable (Profit).

## 2.4 Data Pre-Processing

* To streamline the dataset and focus on relevant variables for our analysis, I have removed the following unnecessary columns: 'Row ID,' 'Order ID,' 'Customer ID,' 'Product ID,' 'Product Name,' 'Postal Code,' 'Ship Date’, 'Customer Name', and 'Country.' As they are in no way affecting our profit.
* Linear regression and random forest can be considered to model profit as the outcome variable.
* Potential predictor variables could include sales, ship mode, day, category and Discount.
* Data transformations might be needed to address skewness in certain variables.
* Interaction terms between variables, if relevant, could be explored.
* The categorical variables (e.g., product categories, regions, segments), are encoded using factor functions to train the machine learning algorithm.

A computer screen shot of numbers

Description automatically generated

All the data in our data frame is now either int or Numeric data. Now the data is all set to get trained by machine learning algorithm i.e., Linear Regression Model and Random Forest for predicting profit as the target variable.

# **ANALYSIS METHODS**

The primary goal in analyzing the Global Superstore dataset is to predict and explain variations in profit, which is a crucial performance metric for the business. I aim to build a predictive model that helps us understand how various factors within the dataset, such as sales, quantity sold, discounts, product categories, customer segments, and regions, influence profit. Specifically, we want to uncover:

* The impact of sales, quantity sold, and discounts on profit.
* The significance of other variables like product categories, customer segments, and regions in relation to profit.
* The nature and strength of the relationships between profit and these predictor variables.

By achieving this understanding, I can create a model that provides valuable insights to optimize business strategies, improve profitability, and make informed decisions for the Superstore's operations.

Our analysis suggests that linear regression and Random forest, with appropriate data preprocessing and consideration of interactions and categorical variables, can be a suitable approach for modeling profit based on the observed relationships in the data.

To train our model and assess the metrics for the machine learning model, load the necessary libraries "dplry" and "caret," scale the data suitably, and then divide the data into "X\_train," "X\_test," "Y\_train," and "Y\_test." using 80% of the data to train our model and the remaining 20% to test it.

A screenshot of a computer program

Description automatically generated

## 3.1 LINEAR REGRESSION MODEL:



A screenshot of a computer

Description automatically generated

Results of various metrics on our linear regression model are as follows.

A computer code with numbers and symbols

Description automatically generated with medium confidence

|  |  |  |
| --- | --- | --- |
| S.No | METRIC | VALUE OBTAINED |
| 1. | Mean Absolute Error (MAE): | 58.5029 |
| 2. | Mean Squared Error (MSE): | 38709.1726 |
| 3. | Root Mean Squared Error (RMSE): | 196.7465 |
| 4. | Adjusted R-squared (R²): | 0.3142 |

**LINEAR REGRESSION PLOTS**:

A graph of a graph with black dots and a red line

Description automatically generated

Both the linearity and homoscedasticity assumptions of linear regression should be verified by this plot. You expect the residuals to be randomly scattered along empty horizontal lines in a normal sample. The pattern in this plot may indicate linearity or heterogeneity issues. It helps in your assessment of the model’s ability to represent the underlying relationships in the data.

A graph with a line

Description automatically generated

The residuals distribution is compared to a normal distribution using the Q-Q plot. The residuals in a sound model ought to be linear. Departures from normality can be shown by deviations from a straight line. Extreme variances imply that either data changes are required or that the model may not be adequate.

A graph of a graph with numbers and a line

Description automatically generated with medium confidence

To evaluate homoscedasticity (constant residual variance), this graph is utilized. A horizontal line with points dispersed randomly is what you would anticipate from a well-fitted model. When the spread of the points changes consistently with the fitted values, heteroscedasticity is likely. The scale-location graphic sheds light on the residual variation over the gamut of fitted values.

A graph of a number of lines

Description automatically generated with medium confidence

With the use of this plot, outliers that have a big impact on the regression coefficients can be found to be important data points (or influential data points). The upper-right or lower-right corner points on a residual vs. leverage graphic could be outliers. It is important to look at high leverage points further because they might significantly impact the model's fit.

## 3.2 RANDOM FOREST MODEL:

A screen shot of a computer

Description automatically generated

Results of various metrics on our Random Forest model are as follows.

A computer code with blue text

Description automatically generated

|  |  |  |
| --- | --- | --- |
| S.No | METRIC | VALUE OBTAINED |
| 1. | Mean Absolute Error (MAE): | 24.58867 |
| 2. | Mean Squared Error (MSE): | 13075.28 |
| 3. | Root Mean Squared Error (RMSE): | 114.3472 |
| 4. | R-squared (R²): | 0.698877 |

**RANDOM FOREST PLOT:**

**A graph of trees and text

Description automatically generated**

In the context of Random Forest (RF), an "error vs. trees" plot, also known as a "out-of-bag (OOB) error vs. number of trees" plot, is a graphical representation that illustrates how the performance of a Random Forest model changes as you modify the number of decision trees in the ensemble. Showing the correlation between the number of trees and the error rate (classification or regression) performed on OOB data assists you in choosing the ideal number of trees. This diagram usually shows that increasing the number of trees initially lowers the error rate, but eventually the improvement plateaus, signifying the point of diminishing returns. It's essential for choosing the right level of model complexity while preventing overfitting. When I attempted to take 150 or 200 trees in this situation, the error grew after 100 trees. Thus, we restricted the number of trees to 100.

## Models’ summary:

The Linear Regression model exhibits limitations in terms of predictive accuracy and explanatory power, as indicated by a relatively high Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and a low R-squared (R²) value. Its generalization ability is limited to data resembling the training set.

In contrast, the Random Forest model outperforms the Linear Regression model with lower MAE, MSE, RMSE, and a higher R² value. It provides more accurate predictions and explains a larger portion of the variance in the dependent variable. The Random Forest model can be confidently generalized to make accurate predictions for new data within the domain similar to the training set.

Overall, the Random Forest model is a more suitable choice for predictive modeling due to its superior performance in terms of accuracy and explanatory ability.

# **CONCLUSION**

This study aimed to investigate the relationship between predictor variables such as sales, quantity sold, and discounts, and the outcome variable, profit, within the context of the Global Superstore dataset. To analyze how these factors impact profit and assess model performance, both Linear Regression and Random Forest models were constructed and evaluated.

**Study Approach:**

* The study commenced by collecting and exploring the Global Superstore dataset, identifying potential predictor variables and designating profit as the outcome variable.
* Two models, Linear Regression and Random Forest, were developed to predict profit based on the selected predictor variables.
* Model performance was assessed using various metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²).

**Key Findings:**

* The Linear Regression model exhibited limited performance, indicated by high MAE, MSE, RMSE, and a low R² value. These metrics imply significant prediction errors and a weak ability to explain profit variance.
* In contrast, the Random Forest model outperformed Linear Regression, demonstrating lower MAE, MSE, RMSE, and a higher R² value, signifying improved predictive accuracy and explanatory capability.

**Implications and Meaning of Results:**

* The Linear Regression model may not generalize well to new, unseen data, and its predictions may be error-prone.
* While it remains usable for prediction, decision-makers should consider its limitations when relying on its output.

**Future Research Directions:**

Given the model limitations, further research avenues include:

* Feature Engineering: Exploring additional predictor variables or transformations to enhance model performance.
* Alternative Modeling Techniques: Considering non-linear models, ensemble methods, or advanced regression techniques for better capturing data relationships.
* Data Collection: Expanding and diversifying data sources to improve the model's generalization ability.
* Domain-Specific Insights: Incorporating domain knowledge to refine the model and identify crucial profit-affecting factors.
* Business Strategy Optimization: Using the improved model for optimizing pricing, discount strategies, and resource allocation across different product categories, customer segments, or regions.

In conclusion, this study delved into the impact of predictor variables on profit using both Linear Regression and Random Forest models. While the Linear Regression model exhibited limitations, the Random Forest model showcased superior predictive performance. Further research avenues aim to enhance the analysis, ultimately supporting more effective business decision-making in the context of the Global Superstore dataset.

References:

Vikas Kumar Appani. (2023). Global Super Store Dataset. Kaggle.

URL: <https://www.kaggle.com/datasets/apoorvaappz/global-super-store-dataset>.

Accessed [01-10-2023].