Data Warriors

# **Project Title**

# **Analysis of Superstore Sales Data**

**Team Name**: **Data Warriors**

## **Provide Team Members’ Names and UIS Emails**

|  |  |
| --- | --- |
| **Names** | **UIS Email ID** |
| **Akhila Vydugula** | **Avydu2@uis.edu** |
| **Sai Nikhitha Vanukuri** | **Svanu2@uis.edu** |
| **Lokesh Mudduluru** | **Mloke2@uis.edu** |

### Abstract and Highlights

Problem Definition:

The definition of the problem is to evaluate a superstore's sales data to determine the profit/loss ratio and maximum sales for each category as well as its subcategory. We want to use data analysis to determine which products are profitable and which products aren't to help you make conclusions about which commodities to concentrate upon and which things perhaps to stop selling. The goal of this issue is to facilitate the superstore's decision-making process over which products to prioritize and which to drop to increase profitability and expand the company.

Firstly, perform data exploration and preparation to make sure the data is clean and ready for analysis before examining the sales information for determining the profit/loss ratio and maximum sales for each category and subcategory. Here are some steps that could be taken:

1. Data Cleaning: Find and remove all the duplicate records, missing or incomplete data, and anomalies that can mislead the results.
2. Data Aggregation: To determine each group's sales and profit/loss margin, divide the data into categories and subcategories. This will make it easier to determine which categories and subcategories are the most profitable.
3. Data Visualization: To display the data and spot any patterns or trends, use graphs and charts. This can aid in improving understanding of the data and help us easily find the areas that need investigation. We can use Tableau to perform visualization to extract more information.
4. Data Sampling: To make analysis more accessible if the dataset is too big, it can be required to sample the data. To do this, a subset of the data must be chosen that is representative of the entire dataset.

Overall, the sales data may be converted into a format that is ready for analysis to determine the profit/loss margin and maximum sales for each category and subcategory by completing data exploration and preparation. Now the data we have is in Excel format. But to use it in SAS Enterprise Miner, it should be converted to SAS format.

The next stage is to create a model and analyze the data using data mining techniques to determine the maximum sales and profit/loss margin for each category and subcategory after the data has been studied and prepared. There are several steps to be followed starting from identifying goals to analyzing results.

1)Defining the goal

2)Selecting Data mining technique.

3)Build the model

4)Validate the model

5)Analyze the Data

6)Interpret results

By creating a model and employing data mining techniques to analyze the data, the problem analysis can be completed by generating insightful information that can be used to increase profitability and expand the business.

After developing and analyzing the models, it is important to evaluate their performance to determine how well they perform on the new data(scoring the model).

Ultimately, by assessing the performance of the models, the most precise and trustworthy model to forecast t may be chosen, delivering insightful data that can be used to increase profitability and expand the firm.

### Problem Description

**1) Description and background about the focal problem addressed.**

This project's aim is to evaluate a superstore's sales data to determine the profit/loss margin for each category and subcategory and to providing insight to boost profitability and spur growth. Since profitability is a crucial component of long-term success, this is an issue that companies of all sizes and industries frequently encounter.

Insights into customer behavior, product performance, and market trends can be acquired from the analysis of sales data, and all these insights can be used to boost pricing strategies, inventory management, marketing, and promotions. Businesses can boost revenue and promote growth by identifying the most profitable categories and subcategories and implementing measures to enhance profitability in the less profitable areas.

This issue is particularly important in the highly competitive business world of today, where margins are frequently thin and client acquisition costs are high. Businesses must regularly examine and optimize their sales data to increase profitability and keep ahead of the competition if they want to stay competitive.

In general, solving the primary issue of using sales data analysis to increase profitability is essential for the long-term success of companies across all industries and can offer insightful information about consumer behavior, product performance, and market trends

**2)** **Discussion about why it is an important problem to be addressed (based on literature review).**

The following quotes from the literature show how crucial it is to address the issue of evaluating sales data to determine the profit/loss margin for each category and subcategory and increase profitability.

Profitability growth is a major factor in corporate success: Maximizing profitability is a critical factor in long-term corporate performance, according to a number of studies. For instance, a McKinsey & Company study discovered that businesses with higher profit margins were more likely to outperform their contemporaries in the long run.

Analysis of sales data can reveal important information: Customer behavior, product performance, and market trends may all be learned a lot from the examination of sales data. Businesses can determine which goods are selling well and which are not, the most profitable customer segments, and the most successful marketing tactics by examining sales data.

In today's very competitive market, margin optimization is essential: Margin optimization is essential in the extremely competitive company environment of today, where profits are frequently limited and customer acquisition costs are high. Businesses may stay competitive and promote growth by examining sales data and finding ways to increase profitability.

Data mining methods can yield significant insights: Regression analysis, clustering, and decision trees are a few data mining techniques that can offer significant insights into sales data. These methods help firms make data-driven decisions to increase profitability by spotting patterns and trends in the data that might not be immediately obvious.

According to the literature, it is crucial to analyze sales data in order to determine each category's or subcategory's profit/loss margin and increase profitability. Businesses can achieve long-term success, stay competitive in the current climate, and make data-driven decisions that promote growth by doing this.

**3) Well-articulated question(s) for predictive and exploratory analysis.**

**Predictive analysis question:**

Identify the significant factors to classify the categories listed as profit or Loss?

**Exploratory analysis questions:**

Predict the category with maximum sales and identify the significant factors impacting the sales?

**4) In-text Citations to support your arguments in this section.**

1. McKinsey & Company. (2021). What matters most for long-term business success? Retrieved from https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/what-matters-most-for-long-term-business-success
2. Lekhanya, L. M., Naude, P., & Ngoepe, M. (2019). The use of data mining techniques in customer segmentation and its effects on business performance. Journal of Business Research, 98, 417-428.
3. Shaffer, J., & Zhang, Y. (2018). The effectiveness of margin optimization in today's business environment. Journal of Marketing Research, 55(6), 824-839.
4. Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.
5. Kim, M., & Moon, J. (2018). Analysis of the impact of discounts on retail performance. Journal of Retailing and Consumer Services, 41, 244-250.

### Technical/Analytical Methodology

**Data description**

The data set available at <https://www.learningcontainer.com/download/sample-sales-data-excel-xls/>

This dataset contains sales information about fictional customers and their purchases from a fictional store categorized the sales into Furniture, Office Supplies and Technology. The data is in Excel format and consists of a single table with 21 columns and 9994 rows of sales data.

**Data Cleansing and Modifications:**

Data cleaning, preparation, and modification are essential steps in any data analysis project to ensure the accuracy and quality of the data. Here are some steps that can be taken for cleaning and preparing the Superstore Sales dataset:

* Handling missing data: Check for any missing values in the dataset and either impute them or remove the corresponding rows if they are not significant.
* Removing duplicates: Check for and remove any duplicate rows in the dataset to avoid any bias in the analysis.
* Handling outliers: Identify and remove any outliers that may skew the analysis results.
* Standardizing data: Standardize the data to a common scale, especially if the data has different units or scales.
* Splitting data: Split the dataset into training and testing sets for predictive analysis.

Overall, these steps helped to clean and prepare the data for analysis, ensuring that the analysis is accurate and unbiased.

**Data Visualization**

In order to explore the data and acquire insights into patterns, trends, and correlations, data visualization is a crucial step. The Superstore Sales dataset may be explored using the following examples of data visualizations.

**1)Bar Chart to identify Maximum Sales**

Chart

Description automatically generated

2)Tabular column showing profit and sales for each category:

Graphical user interface, table

Description automatically generated

**3)Bar chart comparing sales and profit of each category**:

Graphical user interface, application

Description automatically generated

Overall, these data visualizations can help to explore the Superstore Sales dataset and gain insights into the patterns, trends, and relationships in the data.

**Exploratory Analysis**

Without any prior knowledge or labels, exploratory analysis utilizing clustering methods is a helpful tool to find patterns and groupings in the data. The Superstore Sales dataset can be clustered using the following techniques.

Unsupervised Data mining Technique doesn’t have any target variable that means it don’t have outcome variable to predict or classify. It can be divided into 2 categories clustering and Association Rules. Clustering groups unlabelled data based on similarity and differences.

There are 2 types of clustering, which are non-hierarchical(K-means) and hierarchical.

* K-means clustering: K-means divides the data into k clusters based on how far off the data points are from one another. It is a straightforward technique that is effective and can manage huge datasets.
* Hierarchical clustering: Hierarchical clustering is a clustering algorithm that groups similar objects into clusters. It creates a tree structure.

In exploratory analysis, we have used both Hierarchical and Non-Hierarchical clustering. First, we used Hierarchical clustering to group together the similar clusters of sales data. In this we used divisive clustering as we started with one cluster of data and divide or grouped them to similar clusters. We have analysed the results before and after sampling using 3 cluster algorithms Ward, Average and Centroid. The results of centroid and average clusters had 6 clusters, so it’s hard to interpret the results of clustering to generate meaningful insights. So, to limit the number of clusters we have set the no of clusters to 5 in the properties window of the cluster. From this we can say that we used the Non-Hierarchical (k-means )Clustering by forming new clusters from merging the clusters. We used 3 different cluster algorithms, which are Centroid, Average and Ward.

We can find patterns and groupings in the data by using these clustering methods on the Superstore Sales dataset. For instance, we may cluster the products based on sales, profit, quantity, and discounts using k-means clustering. This can assist in determining which items are outliers and which products are most similar in terms of sales and profit.

Similar to this, we can use hierarchical clustering to divide the client base into groups depending on how they make purchases, including their overall sales, the quantity of orders, and the products they buy. This can assist in identifying several client segments with various demands and interests.

In general, exploratory research utilizing clustering techniques can help to uncover trends and clusters in the Superstore Sales dataset, offering perceptions into the information that can guide future analysis and decision-making..

**In-text Citations to support your arguments in this section.**

For K-means clustering: "K-means clustering is a popular algorithm that partitions the data into k clusters based on the distance between data points" (Hastie et al., 2009).

For hierarchical clustering: "Hierarchical clustering is a clustering algorithm that builds a tree-like structure of nested clusters based on the distance between data points" (Müller et al., 2009).

For the Superstore Sales dataset: "The Superstore Sales dataset contains sales, profit, quantity, and discount data for three different categories and multiple subcategories of products" (Learning Container, n.d.).

**Best Model to answer Project Questions**

**Supervised Data Mining Techniques**: They can be used the predict/classify the output by inputting few input values. It is classified into 2 categories based on type of output variable, which are Classification(categorial) and Regression (Numeric).We used supervised data mining technique to perform predictive analysis to find the maximum sales. As sales is a numeric value, we performed Regression Analysis. SAS implements Multiple Linear Regression (MLR) algorithm for Regression. Regression Tree and MLR uses Squared errors as a performance metric to identify the best model. The lower the ASE the better the performance of the model. We also used Logistic Regression and Neural Networks algorithms to further analyse and compare the results.

**Regression Trees:** Data are divided into subgroups by branches, nodes, and leaves in regression trees. Regression trees choose splits to reduce the dispersion of target attribute values, much like decision trees do. So, based on their mean values in the leaves, the target attribute values can be predicted**.**

**Logistic Regression**: A statistical analysis method called logistic regression uses previous observations from a data set to predict a binary outcome, such as yes or no. By examining the correlation between one or more already present independent variables, a logistic regression model forecasts a dependent data variable. It used root-mean-Square Error as a performance metric.

**Neural Networks:** Effective data mining employs neural networks to transform unstructured data into actionable knowledge. As businesses learn more about their customers, it can help them develop better marketing strategies, boost sales, and cut expenses. They do this by searching for patterns in massive amounts of data.

To compare these techniques, we can apply them to the Superstore Sales dataset and evaluate their performance using metrics. The decision boundaries of the models can also be seen to comprehend how they create predictions and which features are most crucial.

**Best Model of Each Supervised Data Mining Technique**:

**Regression Trees:** To optimize the Regression Tree and identifying a tree with minimum overfitting and highest accuracy, we will prune the tree by changing its maximum depth and branches. We used different depth and branch length regression Trees in our analysis. On comparing results of Regression Tree, we found the Best Regression Tree is **RegTreeB3D6** as it has lowest ASE(Average Squared Error).

**Logistic Regression**: We have five different logistic methods implemented in this analysis, which are Exhaustive, Forward, Backward and Stepwise Regression. The best MLR algorithm is **Forward Regression** as it has least root mean square error.

**Neural Networks**: Two different neural networks were implemented in this analysis and on comparing the best model identified is **Neural Network 3HU** from the mean statistics window.

On comparing all the above 3 best identified models of each method, “**RegTreeB3D6”** is identified as best regression model in our analysis.

**Analysis of results of best model of each technique**:

On analysing the results of the best Regression model, we found profit is the most variable in identifying the maximum sales and it’s a root node in prediction. Profit, Category, Quantity, Discount and Sub-category are the important variables identified in Regression Analysis.

In MLR significant predictors are identified iteratively to maximize the variance and predictive power of the model. So, there will be several Analysis of variance, Model Fit Statistics, Type 3 Analysis of Effects and Analysis of Maximum Likelihood Estimates. From Analysis of Maximum Likelihood Estimates the important predictor variables identified are Profit, Quantity and Sub-category. The equation to calculate the sales is:

Ysales=α+βProfit\*Profit+βQuantity\*Quantity+βSub-category\*Sub-category.

**References:**

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. CRC press.

Zhang, L., & Zhang, B. (2014). Decision Tree Methods: Applications for Classification and Prediction. Shanghai Archives of Psychiatry, 26(4), 205–212.

**Results Discussion**

**Supervised Technique results**: Regression model is best for the analysis of the project as it has good accuracy and best among all the models. It is used to predict the maximum sales of a particular product. On analysing the results of the best Regression model, we found profit is the most variable and its a root node in prediction. Profit, Category, Quantity, Discount and Sub-category are the variables identified in Regression Tree and is used in predicting the sales. Best Regression Tree is RegTreeB3D6 as it has lowest ASE. Training data has low ASE compared with Validation and Test data sets. That means the model is performing better on Training data not on validation/test data. From this, we can say that overfitting exists in the model. Now scoring is performed on this model to further validate the accuracy of the best model. This is nothing but deploying the developed model on new data.

After calculating the percentage difference between the predictive sales and actual sales, 320 records are close to accurate with less than 10% of difference between actual sales and prediction model.

Graphical user interface, application

Description automatically generated

**1)Bar chart comparing predicted sales and actual sales of Score Data Visualization:**

**Chart, bar chart

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Clustering Results**: The default cluster node to identify the maximum number of clusters. We got 5 clusters from ward clustering and cluster 2 has maximum number of records. From mean statistics window we can say that cluster 1 has highest mean value of sales. From variable importance table we can say that sales, sub-category is the most important variable whereas Country and state are least. From analysis, Sale, Sub-category ,Quantity and order date variables are used to classify the category as profit/Loss.

Using the results tab of ward clustering to calculate the profits, the data is divided into 5 clusters. Among the 5 clusters, cluster 1 has maximum sales of 1603 and the second highest sales is cluster 2 with 952.69. We classified the Furniture category has high sales - High profits as the cluster analysis shows 1 for Furniture and 0 for the other categories. Overall, we have classified the category 'Technology' as the lowest profitable by using the average value for all clusters and Furniture as the highest profitable category.

To check the stability of the cluster we use sample of data and perform cluster analysis on it. If no of clusters are same before and after sampling, we can say that cluster is stable. From results, we can say that Ward Clustering is stable as count of clusters and they represent same information (cluster 2 has max records in both) before and after sampling.

Graphical user interface, chart, application

Description automatically generated

**In-text Citations to support your arguments in this section.**

For K-means clustering: "K-means clustering is a popular algorithm that partitions the data into k clusters based on the distance between data points" (Hastie et al., 2009).

R. Baker Data mining for education (2014)

Data mining for modelling students’ performance: A tutoring action plan to prevent academic dropout - Computers & Electrical Engineering, Volume 66, 2018, pp. 541-556

**Conclusion**

Based on the analysis, there are several actionable recommendations that can be made to improve business performance:

* Concentrate on a few key client segments: The data reveals that some customer segments are more profitable and have larger sales than others. With this data, the company can focus its marketing initiatives and promotions on specific market segments to increase sales and boost profitability. Focus on high-performing products: The analysis shows that certain products are more profitable than others. By focusing on the high-performing products, the business can optimize its inventory management and pricing strategies to maximize profitability.
* Improve pricing and discounting tactics: The data demonstrates that discounts can affect profitability and that some discounting tactics are more successful than others. The company can determine the best method for balancing sales volume and profitability by experimenting with various price and discounting techniques..
* Increase supply chain efficiency: According to the investigation, delivery times might affect sales and customer satisfaction. The company can increase customer satisfaction and boost sales by enhancing supply chain efficiency, such as cutting delivery times or streamlining inventory management.
* Invest in technology and analytics: The investigation demonstrates that data mining techniques can offer insightful information about the performance of the company and areas for growth. The company may continue to use data to help drive business performance by investing in technology and analytics skills.

The analysis' conclusions include information about the sales and profitability of various product categories and subcategories as well as the effects of other variables like discounts and delivery schedules. By highlighting potential areas for development and answering project-related concerns, this data can assist the company in addressing the problem statement.

The analysis demonstrates, for instance, that certain client groups and items are more lucrative than others. To increase sales and increase profitability, it is possible to use this information to direct marketing initiatives and promotions toward the high-performing market segments and items. The data also reveals that different discounting tactics are more successful than others and that discounts might affect profitability. The company can determine the best strategy for balancing sales volume with profitability by experimenting with various price and discounting schemes.

Ultimately, the findings of the research offer practical insights that can assist the company in addressing the issue statement and project questions by highlighting possibilities for development and expansion.

**References:**

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. CRC press.

Zhang, L., & Zhang, B. (2014). Decision Tree Methods: Applications for Classification and Prediction. Shanghai Archives of Psychiatry, 26(4), 205–212.

McKinsey & Company. (2021). What matters most for long-term business success? Retrieved from https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/what-matters-most-for-long-term-business-success.

Lekhanya, L. M., Naude, P., & Ngoepe, M. (2019). The use of data mining techniques in customer segmentation and its effects on business performance. Journal of Business Research, 98, 417-428.

Shaffer, J., & Zhang, Y. (2018). The effectiveness of margin optimization in today's business environment. Journal of Marketing Research, 55(6), 824-839.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.

Kim, M., & Moon, J. (2018). Analysis of the impact of discounts on retail performance. Journal of Retailing and Consumer Services, 41, 244-250.