# **Data Mining for Giant Superstore**

**Dataviz team**

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

|  |  |
| --- | --- |
| **Names** | **UIS Email ID** |
| Tejasri , Bollu | [tboll4@uis.edu](mailto:tboll4@uis.edu) |
| Alekya , Kurapati | [akura3@uis.edu](mailto:akura3@uis.edu) |
| Sumanth , Battu | [sbatt9@uis.edu](mailto:sbatt9@uis.edu) |
| Satya Sairam Kumar, Katakamsetti | [skata5@uis.edu](mailto:skata5@uis.edu) |

May 14, 2023

**Abstract and Highlights**

The business strategy for identifying advantageous customer categories and analyzing profits is one of the most prominent variables for the Superstore-related sector. It is significantly advised, and the Superstores commonly utilize it to identify client groupings and raise profit margins (Mahfuza et al., LRFMV: An efficient customer segmentation model for superstores, 2022). Dataset is from Kaggle where it includes a U.S. retail store's sales and profits from 2014 to 2017 (Chowdhury, *Superstore dataset* 2022). The goal of this dataset is to discover variables that affect profit and create a model that can precisely forecast profit for upcoming sales. The goal of this issue is to offer the retail store suggestions on how to increase its profitability. Several data mining approaches, such as data preparation, data cleansing, and data visualization, were used to examine and prepare the data. The SAS Enterprise Workstation Miner model creation and evaluation process then initially involved splitting the data into training and testing sets. Unsupervised learning techniques for cluster analysis and supervised learning models, such as decision trees, linear regression, and neural networks, were created to forecast profit. Various measures, including mean absolute error and root mean squared error, were used to assess each model's performance. Decision tree ("RegTree B3D6") was shown to be the best-performing model, with the lowest root mean squared error and mean absolute error. Sales, discounts, and the region or category of the product were shown to be the most crucial variables influencing profit. According to the analysis results, the retail outlet should concentrate on boosting sales while controlling discounts and enhancing the profitability of specific product categories as a practical recommendation. For stores aiming to maximize their profitability through data-driven decision-making, this dataset offers useful insights.

**Problem Description**

A Superstore Giant wants to know your expertise to determine what will work best for them given the expanding needs and fierce competition in the market. They want to know which goods, areas, categories, and consumer groups they should focus on or ignore (Superstore Dataset, n.d.).

Making a decision regarding how to prioritize the product after giving it some thought. to boost profitability and increase sales. Additionally, with the help of the time series analysis researchers have done the forecasting on sales and profit (Sasakitetsuya, *Data Analysis for marketing strategy* 2023).

To stay competitive and succeed in the market, the Superstore Giant should focus on identifying high-growth categories, targeting specific customer segments, optimizing inventory management, expanding to new markets, monitoring competitors, embracing technology. By concentrating on those areas, they can enhance their operations, increase sales and maintain a competitive edge. Supportive visualizations are required for the observations from the available dataset (Ashwinshetgaonkar, *Super Store Analysis,* 2022).

The major problems can be with the store’s sales, profit along with customer satisfaction and demand in the market as per the research. To elaborate on this:

1. To better target the company's sales and marketing, it is important to understand the traits and behaviors of various client groups. To deal with segment customers, we can use clustering or classification techniques that divide them into groups according to things like the region, past purchases, and product preferences.
2. Predicting future revenue generation can assist the business in making plans for employees, inventory, and supplies. This Sales Forecasting can be done using regression models that consider various factors impacting on sales and this is done using the sales data available from the existing dataset.

In our project, data mining plays a crucial role as it would unfold the typical insights and patterns that cannot be observed manually from the given data. By using Supervised and Unsupervised learning data mining can help businesses to make more effective decisions and optimization to stay competitive in a market that is changing quickly.

Furthermore, data mining automates a variety of tasks that would take a long time or become impossible for a human to complete manually, like processing enormous data sets or figuring out intricate relationships between different variables.

**Technical/Analytical Methodology**

DataSet: The dataset is called Giant Superstore, and it was downloaded from kaggle.com.

Considering the changes in the business environment, it has become difficult for companies to raise their standards in a competitive world by identifying and targeting profitable products, regions, categories, and customer segments. The Superstore Giant understands how important this is and is looking for professional assistance to help them reach their objectives. To succeed in the competitive market, it is crucial for the superstore to understand the customer's needs, preferences, and buying behavior. By analyzing the customers’ data, they identify the patterns and trends of which type are used for buying different products which can be used to improve the marketing strategies. To identify which goods and categories are generating the most money and which ones are failing, it is also crucial to study sales data. The business can maximize profitability and improve inventory management by concentrating on the most lucrative goods and categories. The most profitable regions can be identified through geographic analysis, as well as those that could need more funding or a different strategy. With the use of this data, targeted marketing campaigns can be created, and pricing and promotion methods can be modified as necessary. In general, Superstore Giant can decide where to focus their resources and which methods will be most successful in attaining their business objectives by utilizing data analytics and professional insights.

DataSource:<https://www.kaggle.com/datasets/vivek468/superstore-dataset-final?resource=download> (Chowdhury, *Superstore dataset* 2022).

From <https://www.kaggle.com/datasets> website.

Data Cleaning was done to remove all the null values, and check if the data has some duplicate values, we also used a filter to remove all the special characters from the dataset. We checked all the data types and found no errors in it. We also performed correlation in our data to check which variables are highly correlated, that means the variables having correlation 0.8 or more than 0.8. By this in our dataset the variables are not highly correlated. Later, we did dimension reduction in SAS EM as our data is clean and good to go.

The below graphs represent the Data Visualization in different aspects.

The selection of variables for the analysis as well as the derivation of those variables can both benefit from the usage of visualization techniques. If binding of numerical variables is required (for instance, if a yes/no choice is needed, a numerical result variable may need to be changed to a binary variable), they can also assist in finding the proper bin sizes. As a part of the data reduction process, they can also contribute to the combination of categories (Shmueli et al., *Data mining for Business Analytics: Concepts, techniques, and applications in R* 2019, p.56).

The first Graph shown is between the Sales and Profit by Region and Category

A screenshot of a computer

Description automatically generated with medium confidence

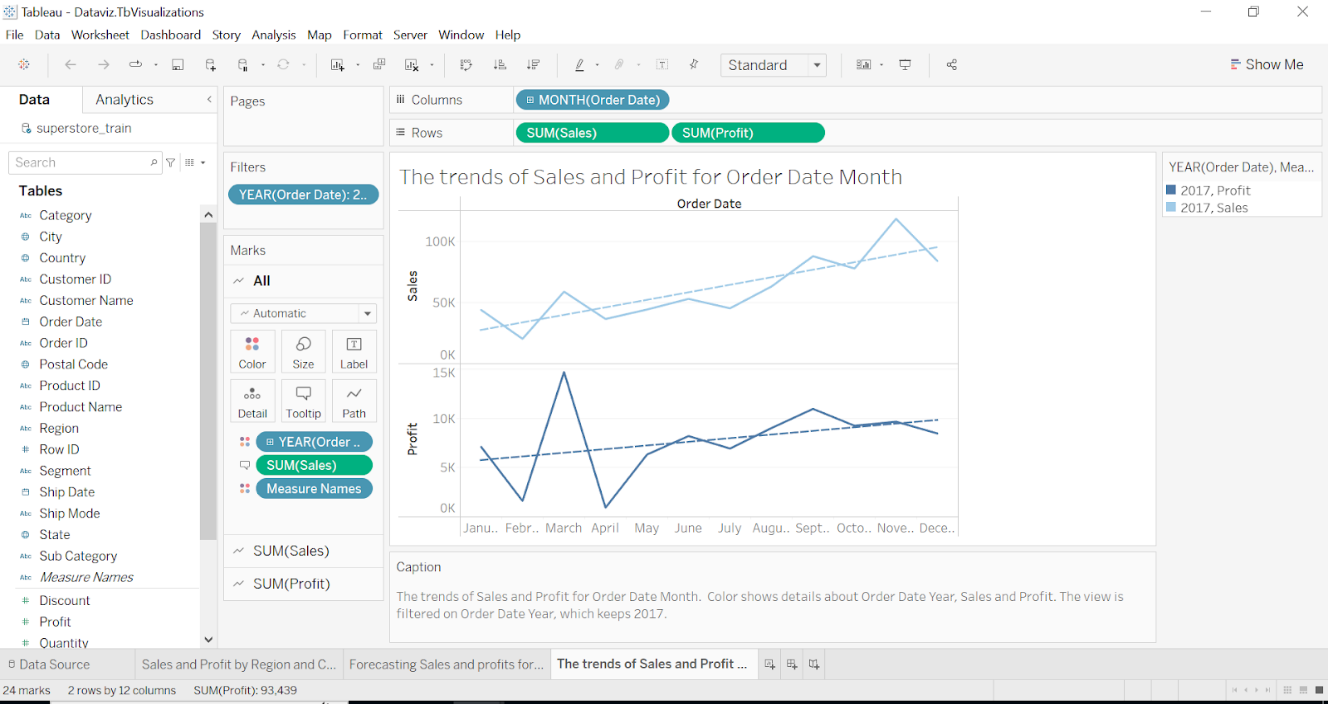
The horizontal bar graph represents the Region and Category and the vertical bar in the graph represents the Sales. We also took sub-categories and profits into consideration so all the products are represented in the graph to show their sales of those products in the regions. We specifically used a bar graph so that we can get a relation between the highly demanded products and their sales when we consider all the categories like furniture, office, and technology. When we look at the graph, we can see that technology has the highest demand and the sub-category which has highest sales from 2014 - 2017 is Phones. There is another reason that in Exploratory Analysis we can get the accurate values of how much the product has demand which can be used to recommend to the customers.

The second graph shows the forecasting in between sales and profit for 3 years.

A screenshot of a computer

Description automatically generated with medium confidence

The years and months are taken to predict the future of the sales, From the above graph we can see that in 2014 the sales were around 200k and then decreased to almost 10k then rose to 60k in March and from there were ups and downs, but the highest sales were recorded in between September to November and the year ended with almost 75k in December. But from the next coming the sales were pretty much the same as increments in the year beginning and then decreased and finally ended with values much greater than the beginning. While coming to the forecast, the years have almost the same for the next 3 years with minimal changes in the sales.



The third above graph represents the trends of sales and profit for order date month, which means this graph shows the pattern of how the customers buy products that impacts on the sales. The superstore mainly records this information to identify trend patterns to improve their sales.

In terms of the time and effort dedicated to various variables, the use of basic charts will depend on the nature of the data mining work and domain expertise about the data. The result variable will receive more attention in supervised learning. The outcome variable is often connected to the y-axis in scatter plots. It is preferable to use simple plots that show associations in unsupervised learning (for the purpose of data reduction or clustering), such as scatter plots (Shmueli et al., *Data mining for Business Analytics: Concepts, techniques, and applications in R* 2019, p.59).

Exploratory Analysis:

In this analysis we take the dataset and connect them to the clusters, there are three types of clusters they are:

1. Ward Clustering
2. Centroid Clustering and
3. Average Clustering

In these clusters we will check the numbers of clusters formed after running the results. The below results show the best model in clustering for the dataset, which is Ward Clustering. We can see that the number of clusters in Segment Size is 3 before and after the Ward clustering of 85% sampling.

Initially we took the 90% sampling to check the stability of 3 different clusters before and after the sampling. After the dataset is executed, we observe the results. Further, we sampled it to 85% and executed to check which cluster is better stable than any other cluster analysis.

We have used just the hierarchical cluster analysis in our project as we got the clustering results before and after different percentage of samplings resulted in 6 or below 6 clusters. The result of one of our clustering results is as shown below which has before and after sampling results. Further we verify the mean statistics window to analyze which model is stable.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Best chosen model - EACH Supervised technique**

The following three supervised data mining methods, such as decision trees, regression, and neural network models are used to analyze data and create models for prediction. There are two different types of diagrams performed, they are:  
1. A diagram in which all the models directly connect from the data partition node from the dataset.  
2. A diagram in which all the models connect from the principal component’s node after the data partitioning.

We have done 3 different models in each diagram such as Decision tree, Regression and Neural Network models.

From these three techniques of diagram 1: In Decision tree models the best model is “RegTree B3D6”.  In Regression models “Exhaustive Regression” is the best model. In neural networks “NeuralNet 5HU” is the best model (Shmueli et al., *Data mining for Business Analytics: Concepts, techniques, and applications in R,* 2019).

**Diagram 1 output of predictive analysis:**

The results window shows the best decision trees, that is “RegTree B3D6”. This tree has three branches, and the depth is 6.

A screenshot of a computer

Description automatically generated with medium confidence

The results window shows the best Neural Network model, that is “NeuralNet 5HU”.

A screenshot of a computer

Description automatically generated with medium confidence

The results window shows the best Regression model, that is “Exhaustive Regression”.

A screenshot of a computer

Description automatically generated with medium confidence

**Diagram 2 output of predictive analysis:**

From the three techniques of diagram 2: In Decision tree models the best model is “RegTree B2D6”.  In Regression models “Exhaustive Regression” is the best model. In neural networks “NeuralNet 5HU” is the best model (Shmueli et al., *Data mining for Business Analytics: Concepts, techniques, and applications in R,* 2019).

The results window shows the best Neural Network model, that is “NeuralNet 5HU”.

A screenshot of a computer

Description automatically generated with medium confidence

The results window shows the best decision trees, that is “RegTree B2D6”. This tree has two branches, and the depth is 6.

A screenshot of a computer

Description automatically generated with medium confidence

The results window shows the best Regression model, that is “Exhaustive Regression”.

A screenshot of a computer

Description automatically generated with medium confidence

Comparison of different models in diagram 1 according to their performance metrics:

A screenshot of a computer

Description automatically generated with medium confidence

Comparison of different models in diagram 2 according to their performance metrics:

A screenshot of a computer

Description automatically generated with medium confidence

**Overall, best chosen Model to answer both predictive model and exploratory Research Question(s)**

Exploratory analysis: What are the primary elements influencing the profitability of the highest profit-margin products? Which product has the highest profit over a particular region or segment? Do any regions or categorical products have the least margin and constant non-profit status?

Predictive analysis: Predict which product is on demand and could have the most profit in future depending on the region, history of sales and price of the product?

Supervised learning: By comparing three different models, according to their performance metrics for diagrams 1 and 2. The decision tree from diagram 1 that is, “RegTree B3D6” is considered as the best model. This is considered as there is no barrier as, Principal component node which affects the model with its variables as it is in diagram 2, with the best Neural Network model “NeuralNet 5HU”. Also, the “RegTree B3D6” has the ASE Average square error of 24358.44 for train, 13503.58 for validation and 28097.93 for test and its maximum absolute error (MAX) for training is higher than the validation and test that is 5935.324, 2200.052 and 5670.72 respectively (Shmueli et al., *Data mining for Business Analytics: Concepts, techniques, and applications in R* 2019, p.121).

A screenshot of a computer

Description automatically generated with medium confidence

Unsupervised learning: When the objective is to produce clusters with nearly similar sizes and variance and the data set contains a manageable number of observations, ward clustering is a reasonable option. However, when working with huge data sets, it might be computationally expensive.

Outliers can pull the centroid away from the other observations in the cluster, making centroid clustering vulnerable to them. When there are many observations in the data set and the objective is to identify groups of observations that are related in some way, it might be helpful.

When trying to find groups of observations that are related in some manner and the data set comprises a lot of observations, average clustering is a useful option because it is less susceptible to outliers than centroid clustering.

However, from the results of before and after the sampling to 85%, we have observed that the Ward clustering is the best stable unsupervised technique in this case.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Scoring results of predictive analysis for diagram 1:**

A screenshot of a computer

Description automatically generated

**Scoring results of predictive analysis for diagram 2:**

A screenshot of a computer

Description automatically generated

From the above 2 scoring results of diagrams 1 and 2 of the predictive model. As we mentioned before diagram 1 predictive model is having the higher priority. Another reason for this is the adj R square value is 38% in the case of diagram 1(when decision tree “RegTree B3D6” is the best model). Whereas it dropped down to 18% in the case of diagram 2(when NeuralNet 5UH is the best model). And the P-value for both the predictive diagram results in <0.0001 which is less than +/- 0.005.

Generally, the higher the value of adjusted R squares the better the model is. Hence, diagram 1 with decision tree B3D6 is the best model and more accurate than any other models of predictive analysis. Further performed the residual analysis for the scored data.

**Conclusions and Practical (actionable) recommendations**

The study of the Superstore dataset offered insightful information for increasing profitability and locating high-profit goods for upcoming sales. Based on the study, the following helpful recommendations may be made.

According to the exploratory analysis, the product category, region, and sales were the main factors affecting the profitability of the items with the greatest profit margins. The top three categories with the largest profit margins were office supplies, technology, and furniture. The southern area has the lowest profit margin and the western region the greatest. When compared to other items, binders and paper had low profit margins while bookcases and tables had large profit margins. These findings suggest that firms should concentrate on marketing high-profit items in lucrative markets and seek to increase the profitability of low-margin products by cutting costs or raising pricing.

Based on previous sales, price, and area data, it was discovered that a model could be trained to anticipate a product's demand and profitability. To determine which product will be in demand and likely to provide the most profit in the future, decision tree, regression, and neural network models were applied. Businesses may plan their inventory, pricing, and marketing tactics based on these projections to optimize earnings.

The analysis's findings addressed the issue statement by outlining the key variables that affect profitability and highlighting highly profitable markets and products. Finding the most and least lucrative categories, locations, and items was aided by the exploratory analysis. With the help of predictive analysis, it was possible to forecast a product's demand and profitability, allowing for the creation of well-informed future sales plans.

Overall, the analysis of the Superstore dataset produced useful suggestions for enhancing profitability and locating high-profit items for future sales. These insights may be used by businesses to make data-driven choices and boost profitability.

**Team Member Contributions**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Name** | **Tasks** |
| **1** | **Tejasri Bollu** | **Project Proposal, Data Acquisition and Exploration,Variable Selection and Dimension Reduction, Model Development, finalizing model, Interpreting and Disseminating Results, Project Report** |
| **2** | **Sumanth Battu** | **Project Proposal, Data Acquisition and Exploration, Data Cleaning and Preparation, Model Development, Performance evaluation, Interpreting and Disseminating Results, Project Report** |
| **3** | **Alekya Kurapati** | **Project Proposal, Data Acquisition and Exploration, Data Cleaning and Preparation, Model Comparison, Performance evaluation, Interpreting and Disseminating Results, Project Report** |
| **4** | **Satya Sairam Katakamsetty** | **Project Proposal, Data Acquisition and Exploration, Variable Selection and Dimension Reduction, Model Comparison, finalizing model, Interpreting and Disseminating Results, Project Report** |

**REFERENCES:**

* **Shmueli, Galit, et al. Data Mining for Business Analytics : Concepts, Techniques, and Applications in R, John Wiley & Sons, Incorporated, 2017. ProQuest Ebook Central,** [**http://ebookcentral.proquest.com/lib/uis/detail.action?docID=5014630**](http://ebookcentral.proquest.com/lib/uis/detail.action?docID=5014630)
* *Find open datasets and Machine Learning Projects*. Kaggle. (n.d.). <https://www.kaggle.com/datasets>
* **Chowdhury, V. (2022, February 17). *Superstore dataset*. Kaggle.** [**https://www.kaggle.com/datasets/vivek468/superstore-dataset-final?resource=download**](https://www.kaggle.com/datasets/vivek468/superstore-dataset-final?resource=download)
* **Mahfuza, R., Islam, N., Toyeb, Md., Emon, M. A. F., Chowdhury, S. A., & Alam, Md. G. R. (n.d.). *LRFMV: An efficient customer segmentation model for superstores*. PLOS ONE.** [**https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0279262#sec004**](https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0279262#sec004)
* **Sasakitetsuya. (2023, January 5). *Data Analysis for marketing strategy*. Kaggle.** [**https://www.kaggle.com/code/sasakitetsuya/data-analysis-for-marketing-strategy**](https://www.kaggle.com/code/sasakitetsuya/data-analysis-for-marketing-strategy)
* **Ashwinshetgaonkar. (2022, May 27). *Super Store Analysis 🏪: Data Visual.: Seaborn🔴*. Kaggle.** [**https://www.kaggle.com/code/ashwinshetgaonkar/super-store-analysis-data-visual-seaborn/notebook**](https://www.kaggle.com/code/ashwinshetgaonkar/super-store-analysis-data-visual-seaborn/notebook)