Data Mining Project

Grisel Quispe Aramayo[1]

1. Introduction

The "Superstore Dataset" was made specifically for educational use in Tableau, a tool for visualizing data. Superstore Giant, the company behind it, wants to get better at what they do. They're interested in learning more about their products, where they sell them, what categories they fall into, and who their customers are. By studying this data, they hope to find ways to improve their business, like figuring out which products are most popular or which areas need more attention.

2. Dataset Understanding

The dataset consists of 21 features and 9994 instances. It contains both categorical and numerical variables. Among these, there are 13 categorical variables, including Ship Date, Ship Mode, Customer ID, Customer Name, Segment, Country, City, State, Region, Product ID, Category, Sub-Category, and Product Name. Additionally, there are 8 numerical variables, namely Row ID, Postal Code, Sales, Quantity, Discount, Profit, Order ID, Order Date, and Ship Date.

3. Dataset Preparation

The dataset underwent rigorous preparation to ensure its quality. We meticulously searched for missing or duplicate values, finding none. This meticulousness guarantees the integrity of the dataset for analysis.

To begin analyzing the dataset, is present a summary of its numerical data in Table 1. However, to maintain data cleanliness and avoid noise, certain columns such as Order ID, Order Date, Ship Date, and Postal Code were excluded from our analysis.

Upon scrutinizing the numerical data, it became evident that outliers exist, particularly in the Profit and Sales columns. For instance, in the Profit column, the minimum value of -6599.978 and the 1st quartile of 1.729 indicate significant deviation from the rest of the data. Similarly, the 3rd quartile of 29.364 and the maximum value of 8399.976 are also noticeably distant.

Likewise, the Sales column exhibits outliers, with values ranging from the 3rd quartile of 209.940 to a maximum of 22638.480. While these outliers may stem from various features within the dataset, their presence poses potential challenges during modeling. Addressing these outliers effectively is crucial to ensure the robustness and accuracy of any subsequent analyses or models derived from the dataset. To understood better the distribution of the data it would be use a boxplot to analyze the presence of the outliers.

	Profit	Sales	Quantity	Discount
Min.	-6599.978	0.444	1.00	0.0000
1st Qu.	1.729	17.280	2.00	0.0000
Median	8.666	54.490	3.00	0.2000
Mean	28.657	229.858	3.79	0.1562
3rd Qu.	29.364	209.940	5.00	0.2000
Max.	8399.976	22638.480	14.00	0.8000

Table 1: Summary statistics for Profit, Sales, Quantity, and Discount

3.1. Outliers

In Figure 1, it is employ a box plot to visually examine outliers within the Sales column. The plot illustrates a significant number of data points lying above the threshold of 200, indicating the presence of numerous outliers. These outliers, visually depicted as data points extending beyond the upper whisker of the box plot, suggest instances where sales values greatly exceed the typical range observed in the dataset

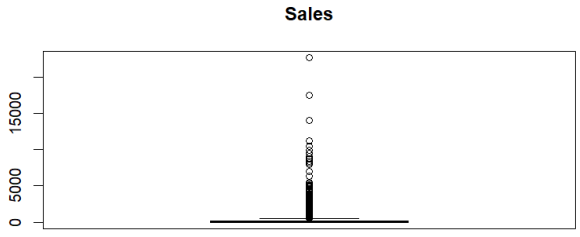

Figure 1: Outliers in sales

Figure 2 shows a box plot used to examine outliers in the

Profit column. The plot reveals numerous data points extending beyond the typical range, both above and below. These outliers represent significant deviations from the average profit values in the dataset, including both positive and negative extremes. This visual insight highlights the variability and potential irregularities within the profit data, prompting further investigation to understand their implications.

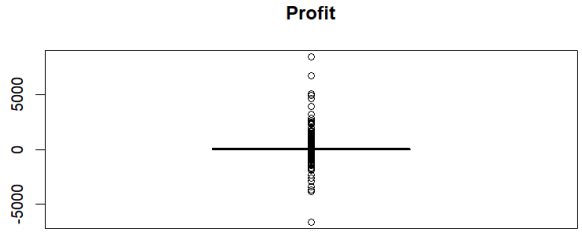

Figure 2: Outliers in profit

3.2. Correlation

There are two features on which is going to be focus: Segment and Profit. To analyze the correlation between the numerical and categorical features with the target variable (Segment or Profit), it's applied a transform to the numerical features using one-hot encoding. Additionally, also it deletes the present outliers using z-scores with a threshold of 20. Then it's applied random forest to find the most correlated features.

Feature	Importance
Profit	17.00868230
Sales	14.82370045
Discount	9.97082975
Quantity	6.55416291
CategoryOffice.Supplies	6.25956740
Ship.ModeStandard.Class	5.46829570
Ship.ModeSame.Day	5.15729157
RegionWest	4.87307861
CategoryTechnology	4.60910454
RegionEast	3.69329858
RegionSouth	0.70561880
Ship.ModeFirst.Class	0.49268585
Ship.ModeSecond.Class	-0.01779109

Table 2: Feature Importance for Segment

In the table 2 it can be seen the most important features for the target value "Segment".

For the profit In the table 3 it can be seen the most important features for the target value "Profit".

Feature	Importance
Sales	19.0072682
Discount	17.6217877
CategoryOffice.Supplies	3.2206995
Quantity	2.6589966
RegionWest	2.6033785
Ship.ModeSecond.Class	1.8902351
CategoryTechnology	1.6636187
RegionEast	0.1246425
RegionSouth	-0.3178857
Ship.ModeSame.Day	-0.4145870
Ship.ModeFirst.Class	-0.9433625
Segment	-1.7432202
Ship.ModeStandard.Class	-1.869202

Table 3: Feature Importance for the target Profit

The dataset preprocessing has been completed, with redundant data being cleaned. For instance, the "Country" column, containing only one value ("United States"), was removed. Additionally, to streamline the geographical data, the "State" and "City" columns, each containing over 100 different cases, were deleted, with only the broader geographic regions (East, West, North, and South) retained.

Furthermore, addressing numerical data, outliers were dealt with, and the data for "Sales" and "Profit" were standardized using the min-max technique, scaling values between 0 and 1 to ensure consistency and comparability across the dataset.

Moreover, a significant correlation between features and target variables has been identified. With these preparations completed, the dataset is now ready for modeling and analysis.

4. Modeling

4.1. Modeling for Segment

To find patterns and filter the data, the dataset includes customer IDs. By combining these with the Order IDs, analysts can discern how frequently customers make purchases. Upon examination of Table 4, it is noted that the most frequent occurrence is 37 times.

	Percentiles					
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Freq.	1.0	8.0	12.0	12.6	16.0	37.0

Table 4: Summary Statistics for Frequency

Candidates with a frequency of at least 10 were initially sought, resulting in 757 customers identified. Subsequently, a more stringent filter was applied, requiring a frequency of at least 30. This narrowed down the selection to 390 customers, who are considered the most important due to

their loyalty. Analyzing this subset of customers is deemed crucial for gaining insights into customer behavior and preferences.

To analyze their behavior, the dataset underwent clustering using the k-means algorithm with three clusters. Subsequently, the clusters were populated as follows: Cluster 1 comprised 347 customers, Cluster 2 contained 7 customers, and Cluster 3 included 36 customers. This segmentation enables the utilization of the Apriori algorithm to discover association rules among customer behaviors and preferences.

For the transactions it has been used: "Segment", "Region","Category", "Sub.Category", "Cluster", a support of 0.2 and a confidence of 0.8. In the table 5 it can be seen the

Num.	lhs
1	{}
2	{Category=Furniture}
3	{Region=East}
4	{Region=West}
5	{Category=Office Supplies}
6	{Segment=Consumer}
7	{Region=West, Category=Office Supplies}
8	{Segment=Consumer, Region=West}
9	{Segment=Consumer, Category=Office Supplies}

Table 5: lhs

Left Hand Side "LHS", the conditions which are observed in the dataset.

rhs	supp	conf	coverage	lift	count
{Cluster=[1,3]}	1.00	1	1.00	1	390
{Cluster=[1,3]}	0.22	1	0.22	1	87
{Cluster=[1,3]}	0.32	1	0.32	1	126
{Cluster=[1,3]}	0.34	1	0.34	1	135
{Cluster=[1,3]}	0.60	1	0.60	1	235
{Cluster=[1,3]}	0.66	1	0.66	1	258
{Cluster=[1,3]}	0.22	1	0.22	1	86
{Cluster=[1,3]}	0.27	1	0.27	1	106
{Cluster=[1,3]}	0.37	1	0.37	1	147

Table 6: rhs, support, confidence, coverage, lift, count

In Table 6, you'll find the RHS (Right Hand Side), representing the conditions—in this case, clusters 1 and 3. These results unveil intriguing insights into customer behavior, highlighting the most significant categories of items, such as Furniture, the regions where customers are located, and the Customer segment that prevails within this filtered dataset.

With the clusters, it can be also analyze the Sales and Profit. Figure 3 illustrates that while clusters 1 and 2 do not have the highest values, they exhibit similar patterns. Conversely, cluster 2 depicts significantly higher sales and profit. However, it's noteworthy that there are fewer instances of such high sales, indicating their infrequency in the store.

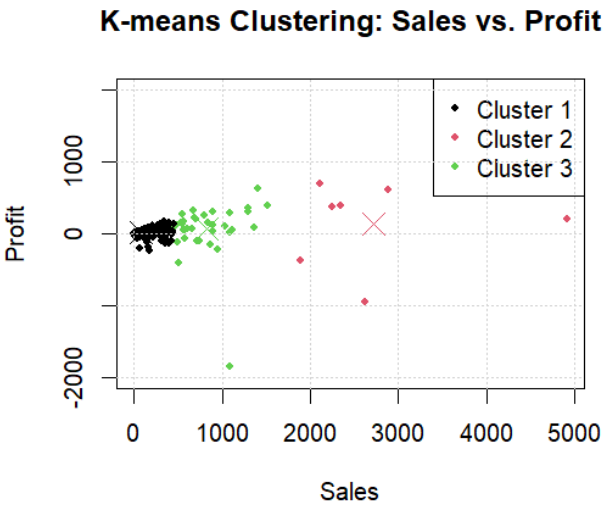

Figure 3: K-means

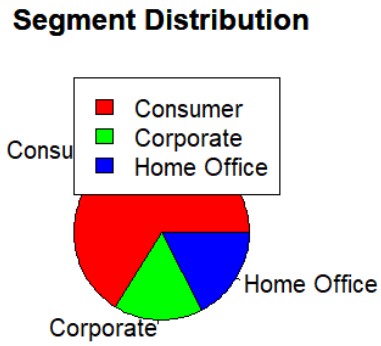

Figure 4: Segment for the cluster 1

Figure 4 reveals that the most common segment within cluster 1 is Consumer.

4.2. Modeling for Profit

To analyze the target variable "Profit," the selected features include Sales, Quantity, Discount, Ship Mode Second Class, Region East, Region West, Category Office, and Category Technology. These features, identified through random for-

est analysis, exhibit the highest correlation with Profit. Additionally, Sales and Profit have been normalized for consistency.

Initially, regression models such as Linear Regression, Lasso Regression, Ridge Regression, and Elastic Net Regression were employed to predict Profit. Despite tuning and cross-validation efforts, the results were not deemed satisfactory.

After experimenting with various regression models, the decision was made to implement a decision tree for regression (Diachkov, 2024). The model's structure is depicted in Figure 5. The plot illustrates that the most influential features in predicting Profit are Discount and Sales. The overall dataset was splitting in training 80 % and testing 20%. To evaluate the model's performance, predictions are

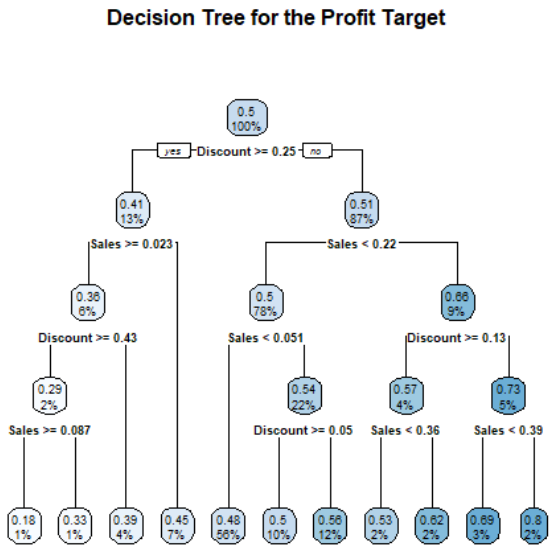

Figure 5: Decision tree for profit

compared with actual values in Figure 6. As the result was not improving, another approach has been decided, changing the profit from numerical to categorical splitting into 3 categories, and creating a new feature Profit.category. Also the distribution change: for training 70% and for testing 30%. The model can be seen in the figure 7.

In the new scenario where profit is treated as a class variable, it's noteworthy that discount, sales, and the category of technology emerge as crucial features for the decision tree.

Discount and sales are likely significant because they directly impact the profit margin. Higher discounts or increased sales could potentially lead to lower profits, while lower discounts or decreased sales might result in higher

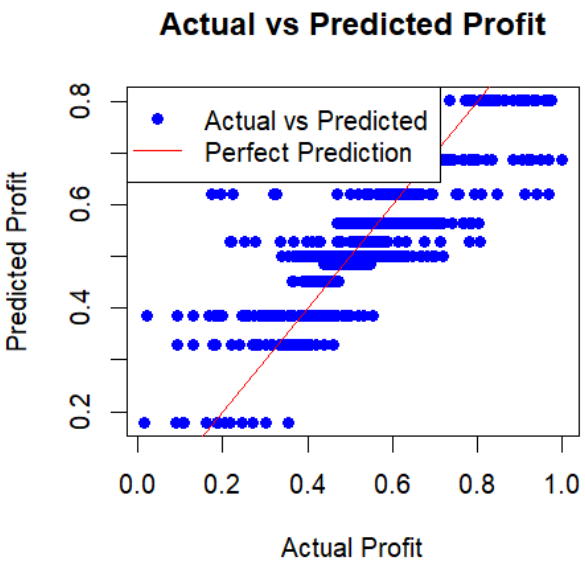

Figure 6: Real vs Prediction

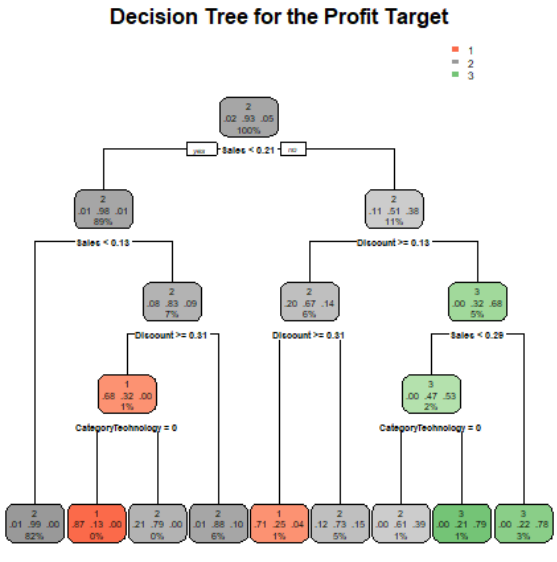

Figure 7: Decision tree for profit category

profits. Thus, the decision tree might prioritize these features to understand the relationship between these variables and the profit class.

Additionally, the category of technology might play a crucial role because it could represent a high-profit margin category. Certain product categories, such as technology products, may inherently yield higher profits compared to others due to factors like pricing strategies, demand elasticity, or production costs. Therefore, the decision tree might identify the category of technology as an important feature for classifying profit levels.

5. Evaluation

To evaluate the best model, (the decision with Profit as category) it has been used the method confusion matrix. In the table 7

Table 7: Confusion Matrix and Statistics

Reference	Prediction		
	1	**2**	**3**
1	24	6	1
2	46	2666	55
3	0	18	85

The model performs relatively well in predicting class 2, with a high number of true positives (2666) and a relatively low number of false positives (46) and false negatives (55). Class 1 has a relatively low number of true positives (24), indicating that the model struggles to correctly classify instances of this class. However, the false negatives for class 1 (6) are relatively low compared to the false positives (6), suggesting that while the model misses some instances of class 1, it also misclassifies other instances as class 1 when they belong to different classes. Class 3 has a moderate number of true positives (85) and relatively low false positives (18) and false negatives (0), indicating reasonable performance in classification. The model require further optimization to improve the performance, but it's a better result than other models that has been tried before.

In the table 8 there are metrics about the model performance.

- Accuracy: The accuracy of 95.66% proves a correctly predict class fir the 95.66% of the instances.
- Confidence Interval: The Confidence Interval between [94.85% and 96.37%] means that we can be 95% confident that the true accuracy falls within this range.
- No information Rate: In this case, the no information rate is 92.73%, suggesting that the model's accuracy of 95.66% is significantly better than random guessing.
- P-Value [Acc NIR]: The p-value is 4.464e-11, is very close to zero which means that the model's accuracy is significantly better than the no information rate.
- Kappa: A kappa value of 0.6176 indicates substantial agreement between the predicted and actual class labels beyond what would be expected by chance.

Table 8: Overall Statistics

Metric	Value
Accuracy	0.9566
95% CI	(0.9485, 0.9637)
No Information Rate	0.9273
P-Value [Acc NIR]	4.464e-11
Kappa	0.6176

5.1. Time performance

The time for the model compilation was 30.59 secs.

5.2. Repository

The code can be found in: https://github.com/GriselQ23/Data-Mining-project

6. Conclusions

This project posed significant challenges, with numerous outputs warranting discussion. Given the nature of this business dataset, it's common to encounter outliers that deviate significantly from the distribution of features. Removing these outliers could impact a significant segment of buyers. Even though these outliers may not occur frequently, if the store aims to enhance this sector, these variables should be analyzed separately.

The order ID feature played a crucial role in uncovering transaction patterns. It facilitated the segmentation of the most frequent clients, allowing for an analysis of the segment from which the majority of profits originate, namely "Consumer". Armed with this insight, the store can optimize its inventory management to meet consumer demands more effectively and make informed business decisions.

Using the Apriori algorithm, several intriguing transaction patterns were uncovered. For instance, it was observed that the Furniture category is frequently in demand. Additionally, attention should be given to the West region and the Consumer segment. Another noteworthy finding is the association between the Office Supplies category and the Consumer segment. While it might be assumed that the

Corporate segment would be the primary consumer of office supplies, the data suggests otherwise. This insight can inform business decisions regarding product placement and marketing strategies.

Building the profit model proved to be challenging due to the nature of the data. In this project, a novel approach was undertaken by employing decision trees and transforming profit from continuous to discrete. Despite the availability of this public dataset and numerous prior attempts using regression, which yielded poor results (Vivek468, 2024), this project achieved remarkable success. The exceptionally high accuracy of 95% demonstrates that alternative methodologies can outperform traditional approaches when applied thoughtfully.

Something very intriguing is that the dataset is provided by Tableau, a visualization tool. Most of the previous work has focused on EDA (Exploratory Data Analysis), indicating that many answers can be found using visualization tools alone. However, within these tools lie rules, candidates, and data mining engineering capabilities. Hence, machine learning may not always be the most appropriate tool for finding solutions. Additionally, complex models such as SVM, Hidden Markov Models, or Deep Learning are not always necessary when answers can be derived using simpler methods or just through rules

References

Diachkov, D. How to start using decision tree classification in r. https://medium.com/data-and-beyond/how-to-start-using-decision-tree-classification-in-r, 2024.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Vivek468. Superstore dataset final. https://www.kaggle.com/datasets/vivek468/superstore-dataset-final/code, 2024.