جامعة عجمان
AJMAN UNIVERSITY

Unlocking Business Value: Predicting Customer Profitability with Machine Learning

Lujain Adnan, Asmaa Luai Abdulla, Sara Al Najjar, Ahmad Jamal Shihad, Rayan Hassan, and Maitha Sultan Alremeithi

202110929, 202111148, 202110946, 202111564, 201920024, 202110248

Department of Information Technology, Ajman University

INS402: Business Intelligence and Data Warehousing

Prof. Ghazi Al- Naymat

Tuesday, 3 December 2024

Table of Contents

Table Of Figures

Abstract

With the competition in today's business environment, understanding and estimating customer profitability is an indispensable part of driving strategic decisions. Most businesses are struggling to develop an accurate prediction of who will be their most profitable customers in a scenario of limited or no clear feature-probability correlation. This gap has developed an urge for the implementation of high-power analytical methodologies that would model the intricate patterns of customer behavior. Machine learning involves the solution to the problem of the prediction of customers' profitability based on sales, discount, and product information among several other features. The respective customers shall, therefore, be segmented into various segments of values such as Low, Moderate, and High. The solution for such a classification task is designed as a Gradient Boosting Classifier, as it easily manages non-linear relationships and it is possible to increase the prediction accuracy by including several decision trees. It consists of data cleaning, preprocessing-encoding categorical variables and standardizing features-and model evaluation using accuracy, confusion matrix, and feature importance. The predictive model derived gives insights into customer segmentation to businesses for effective targeting of high-value customers, as well as offering strategic recommendations toward optimization of profitability. This model reaches an extremely high level of accuracy; it is well applicable in practical business scenarios.

I. Business Scenario

The primary goal of the business, centers on analyzing customer probability for Superstore, that is responsible for chain of retail stores in the US that offers for its customers products in Technology, Furniture, and Office Supplies. The business operates in four regions East, West, South, and Central of the US. And provides its goods for diversity of customers including Consumer, Corporate, and Home Office. The Superstore faces a-lot of challenges when it comes to identifying the reasons behind low profitability, as not all customers contribute in increasing the profit, causes can be because of excessive discounts or frequent product return these factors can lead to a huge loss for the company. Customers are categorized into High-Value, Low Value, and Unprofitable tiers based on their contribution. The purpose of this business analysis is to identify profitable segments, improve underperforming areas, and make data-driven decisions that enhance operational efficiency.

II. Business objectives

1.Maximizing Profitability:

Pinpoint customers, segments, regions, and products that maximizes the profit, to let the business focuses its resources to these main areas.

2.Improve Underperforming Areas:

Understanding and analysing the reasons behind low profitability in specific areas and then implement a strategy to solve these problems will increase the profitability and lower the chances of losing profits.

3.Enhancing Customer Retention:

By identifying top profitable customers, this will encourage the business to target them and offering them special offers, loyalty programs, this will lead in increasing CLV.

III. List of questions that can be answered by the analytical model

1. Customer-Level Profitability

- who are the most profitable customers?
- Who are the least profitable customers, and what is the reason behind it?

2. Sales vs. Profit Analysis

- does always high sales correlate with high profitability? if not, why is that?
- How does the discount rate impact the profits?

3. Segmented Profitability

- Which segment of customers is the most profitable?
- Does the profitability vary across the regions? And why?

4. Discount Optimization

- What is the optimal discount rate that does not decrease the profit?
- Customers that receives discounts consistently, erode the profit?

5. Churn Reduction

- Which customer segments show early signs of churn?
- What strategies can be implemented to prevent churn and maintain high profitability? be removed.

IV. Data understanding

a. Snapshot of the used data

The used dataset for analysis, contains multiple rows that represents customers transactions, with columns that captures key attributes that relates to customers, orders, products.

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ...	Customer ...	Segment	Country	City
1	CA-2016-152156	11/8/2016	11/11/2016	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson
2	CA-2016-152156	11/8/2016	11/11/2016	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson
3	CA-2016-138688	6/12/2016	6/16/2016	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles
4	US-2015-108966	10/11/2015	10/18/2015	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale
5	US-2015-108966	10/11/2015	10/18/2015	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale
6	CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles
7	CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles
8	CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles
9	CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles
10	CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles
11	CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles
12	CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles
13	CA-2017-114412	4/15/2017	4/20/2017	Standard Class	AA-10480	Andrew Allen	Consumer	United States	Concord
14	CA-2016-161389	12/5/2016	12/10/2016	Standard Class	IM-15070	Irene Maddox	Consumer	United States	Seattle
15	US-2015-118983	11/22/2015	11/26/2015	Standard Class	HP-14815	Harold Pawlan	Home Office	United States	Fort Worth
16	US-2015-118983	11/22/2015	11/26/2015	Standard Class	HP-14815	Harold Pawlan	Home Office	United States	Fort Worth
17	CA-2014-105893	11/11/2014	11/18/2014	Standard Class	PK-19075	Pete Kriz	Consumer	United States	Madison
18	CA-2014-167164	5/13/2014	5/15/2014	Second Class	AG-10270	Alejandro Grove	Consumer	United States	West Jordan
19	CA-2014-143336	8/27/2014	9/1/2014	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco
20	CA-2014-143336	8/27/2014	9/1/2014	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco
21	CA-2014-143336	8/27/2014	9/1/2014	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco
22	CA-2016-137330	12/9/2016	12/13/2016	Standard Class	KB-16585	Ken Black	Corporate	United States	Fremont

Figure 1 - Dataset Sample

-Time Period: the transactions span from 2014 to 2017.

-Size of the Dataset contains over 9,000 rows and 21 attributes.

Data Collection Sources:

- Point-of-Sale Systems: Capturing transactions at the same time as the purchase time.
- Order Management Systems: By tracking shipping details and order processing.

b. Key Data Attributes

2.Order Attributes:

- Order ID: Unique Order ID for each Customer.

- Order Date: Order Date of the product.

- Ship Date: Shipping Date of the Product.

- Ship Mode: Shipping Mode specified by the Customer.

3.Customer Attributes:

- Customer ID: Unique ID to identify each Customer.

- Customer Name: Name of the Customer.

- Segment: The segment where the Customer belongs.

4.Geographical Attributes:

- Country: Country of residence of the Customer.

- City: City of residence of the Customer.

- State: State of residence of the Customer.

- Postal Code: Postal Code of every Customer.

- Region: Region where the Customer belongs.

5.Products Attributes:

- Product ID: Unique ID of the Product.

- Category: Category of the product ordered.

- Sub-Category: Sub-Category of the product ordered.

- Product Name: Name of the Product.

6.Sales and Financial Metrics:

- Sales: Sales of the Product.
- Discount: Discount provided.
- Profit: Profit/Loss incurred.
- Quantity: Quantity of the Product.

V. Data Modeling

a. Preprocessing

- Five excel files were uploaded in Power Bi.
- Using the Transform Data option some cleaning happened.
- Applying grouping by Customer name to avoid duplicates.
- Ensuring no missing values.
- Checking and removing outliers to avoid measurement errors.
- Sales and Profit were Normalized.
- Confirm the changes and uploading the new cleaned datasets to start the modeling part.

b. Dimension Tables

- *Orders_Dimension:* Includes all the attributes that is related to Order.
- *Customers_Dimension:* Includes all the attributes that is related to the Customer Attributes.
- *Products_Dimension:* Includes all the attributes that is related to the Products.
- *Time_Dimension:* Order_Date, Ship_Date

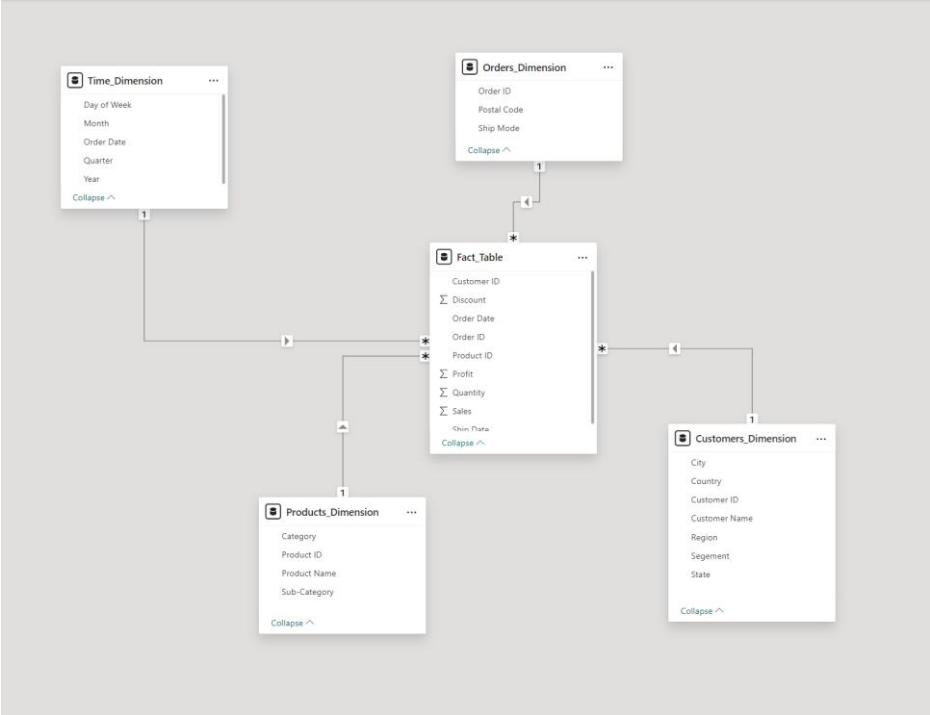

Figure 2 - Data Model - Customer Profitability Analysis

VI. What is the suitable analytical model chosen

a. Model to be used

The present model takes an inspiration from the principle of one of the most powerful ensembles learning methods for classification problems known as the Gradient Boosting Classifier. It has seen wide applications in real-world scenarios where relations among features and the target can be complex, nonlinear, or difficult to model using any other traditional method.

b. Describe the model

Gradient Boosting Classifier is a popular ensemble learning algorithm for classification. It's strong for combining a number of weak learners in view of achieving a very strong predictive model, usually a decision tree. It works effectively in handling a complex data set with interaction effects due to non-linear relationships between features and the target variable.

c. Gradient Boosting Classifier Key Features

1. Ensemble Method: It's a sequence or series of decision trees in which each succeeding tree tries to correct errors created by the predecessor. This, in turn, increases predictive accuracy by learning from the mistakes of previous models.

2.Boosting: It is a form of boosting, which means that it is an iterative process where the new models focus on those instances which have been misclassified by previous models. It forms a strong predictive model through different predictions generated from various weak models.

3. Classification: Gradient Boosting Classifier is only applied to classification problems where one tries to predict categorical outcomes. In this case, the model predicts customer profitability categories such as Low value, Moderate value, and High value.

4. Non-linear Relationships: A major strength of Gradient Boosting is the capability for modeling nonlinear relationships, that always exist in data. This becomes important in business contexts, since customers behave nonlinearly.

While an ensemble learning-based classification model, Gradient Boosting Classifier stands out for its wide acceptability due to its flexibility, coupled with high predictive performances against complex patterns in the data.

d. State-of-the-art about

Gradient Boosting Classifier

Gradient Boosting Classifier has, in the recent context, developed into one of the robust and most powerful machine learning algorithms, being one of the most sought-after in classification tasks owing to their excellent performance in predictions. The recently performed changes in this area contributed much to enhancing efficiency, scalability, and real-world data handling complexity. Key developments were the development of optimized variants: XGBoost, LightGBM, and CatBoost are the most important ones; these introduce crucial enhancements, including parallelization, tree pruning, and advanced handling of categorical features that make GBC faster and more scalable when applied to big datasets.These have enabled Gradient Boosting to reach even higher accuracy with the same or improved computational efficiency. This allows GBC to intrinsically calculate feature importance for further business insights, especially in such areas as customer segmentation and profitability prediction, where the main drivers of customer behavior have to be identified. It has also been extended by various regularization techniques, such as shrinkage and subsampling, to avoid overfitting, hence generalizing better and giving a better performance on unseen data. Other variabilities in development include hybridizing GBC with other machine learning techniques, such as Random Forest and neural networks, which can capture intricate relationships and high-level abstractions within the data, hence boosting their predictiveness. Introduction of methods like SHAP has also given more interpretability to GBC, presenting transparent insight into model predictions for a business, which is called for in actionable decision-making.But most interesting is GBC's adaptation to real-time predictions through streaming data, meaning one can continuously update the models as new data arrives; this way, predictions should be relevant and accurate at any moment in time.These combined improvements have made Gradient Boosting, in its turn, a really all-purpose, robust, state-of-the-art tool for various predictive tasks, making this technique especially valuable in these industries that are focused on the precise modeling of customers' behavior, such as e-commerce, finance, and marketing

e. How the Model Works

The GBC is an ensemble learning algorithm, which is very powerful. It works by iteratively building a sequence of weak learners, usually decision trees, in which each new tree attempts to correct the errors of the previously built trees. This is how trees are built in a stage-wise manner, leading to a strong predictive model.

GBC works from the very notion of boosting, basically meaning that a sequence of models combines their predictions with the best result. The following steps explain how this model really works:

1. Base Model: This is usually the most basic model in the chain, often a simple decision tree, which is considered a "weak learner." It might have low accuracy because it is a very simple model. It provides a simple prediction, but due to simplicity, the prediction will probably have high bias, meaning errors in the predictions.

2. Computation of Error: Subsequent to creating the basic tree, it calculates residual errors. It will represent the differences of predicted versus real target values. In other words, the residuals would depict regions that the model is lagging behind and would need an amendment.

3. Subsequent Trees: Every new tree was a consecutive one, where each looked towards the error residuals of the previous. And so, every new tree tried to set right the errors of the one prior to it in succession and gave more weights towards such data points. By so doing, iteratively the model is improved in terms of reductions both the bias and variance inside the prediction.

4. Weighted Voting: For every additional addition of a tree, all the predictions made by them are combined to arrive at a final prediction. It is usually an average weighted in nature. The weight of each tree depends on how well it corrected the errors from the previous trees. That is, the better it has done at reducing the errors, the more weight it gets for the final decision.

5. Learning Rate: The learning rate in Gradient Boosting is a hyperparameter that controls the contribution of each tree to the final prediction. A lower learning rate will result in each tree making smaller and smaller contributions toward the final prediction, which may make the model more robust but less liable to overfit. This might require a higher number of trees for the same level of performance.

6. Stopping Criterion: It is stopped when the desired number of trees has been constructed, or the model has reached a point where the addition of further trees does not make substantial improvements. The regularization technique can be enforced to avoid overfitting, which may include restricting the depth of trees or imposing early stopping.

Thus, the algorithm builds a very accurate, robust model that can handle even very complex patterns in the data by iteratively improving the model's predictions, with a focus on those instances previously misclassified. Such sequential error correction is an important feature of GBC, making it one of the most effective approaches when there are either linear or non-linear relationships in a dataset.

It follows from the process of enhancing the predictions in steps, starting with weak learners, toward a final strong model, which can then classify examples with high accuracy-for example, predicting customer categories using features such as sales, profit, and discounts.

f. Model Evaluation

Several performance evaluation metrics and visualizations were done for the performance of the Gradient Boosting Classifier. These include accuracy, classification report, confusion matrix, learning curve, and feature importance. Each of these tools gives unique insights into how well the model is predicting customer categories and points toward areas where the model can be improved.

1. Accuracy:

The Gradient Boosting Classifier returned an impressive test accuracy of 92.9%, which assured that the model performance was effective in categorizing customers based on available features. Accuracy is one of the important metrics for a classification model, which determines the overall proportion of correct predictions a model makes, which, as in this case, turned out to be high. That proved the model was reliably distinguishing the classes of High value, Moderate value, and Low value customers.

2.Classification Report:

The classification report summarizes the classes based on a number of important metrics. Therefore, precisions, recalls, and F1-scores are provided for each customer category:

1.High value: Precision 0.91 and Recall 0.92. These mean that the model attains high precision and recall in classifying the customers as High value, showing that most of the true positive customers have been captured by this class with fewer false positive classifications.

2. Low value: At the same time, regarding Low Value customers, their precision was 0.90 when their recall was a little bit worse at 0.85, meaning it gets a majority of those being correct, low value but has sometimes misinterpreted and graded them into the middle section or moderation value customers.

3.Moderate value: In the category of Moderate value, the highest performance is realized with a precision of 0.94 and recall of 0.95. This evidences that the model identifies moderate-value customers highly accurately.
Because of this, weighted precision and recall also are really well-balanced across classes, along with the F1 score, which may indicate that the model was relatively well tuned and performing in all customer value categories effectively.

Accuracy: 0.9279639819909955
Classification Report:

	precision	recall	f1-score	support
High value	0.91	0.92	0.91	383
Low value	0.90	0.85	0.88	392
Moderate value	0.94	0.95	0.95	1224
accuracy			0.93	1999
macro avg	0.92	0.91	0.91	1999
weighted avg	0.93	0.93	0.93	1999

Figure 3 – Classification Report

3.Confusion Matrix:

The confusion matrix allows a better view of the model classification for each category of customer. It shows, graphically, the number of correct and incorrect predictions made for each category:

In the High value class, there is a high number of correct predictions, 355, with only some misclassifications as medium or low-value customers.

-Low Value class is misclassified at a higher rate; 47 Low-Value customers were misclassified as Moderate Value customers, showing that the model slightly got confused between these classes.

-Moderate value customers were mostly classified correctly, though there were minor misclassifications to other classes.

This matrix serves to highlight the fact that while the model is strong at differentiating between high- and moderate-value customers, it has some overlap in distinguishing the low and moderate customer value predictions.

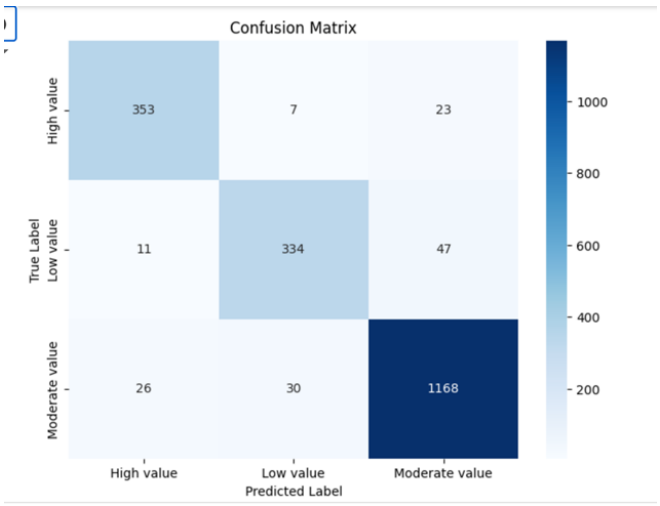

Figure 4 – Confusion Matrix

4.LearningCurve:

The learning curve was generated to evaluate how the model's performance evolves as the number of training samples increases. It illustrates both the training and validation accuracy as a function of the size of the training data:

The curve is one that shows a steady increase in accuracy as the amount of training data increases, which implies that the model benefits from more data. The difference between training and validation curves is little, meaning the model generalizes well to data on which it has not been trained. This further confirms that this model does not suffer from a high variance problem; high variance would be depicted as a wide gap between the two curves.

Figure 5 - Training vs Test Accuracy Across Increasing Estimators

5.Feature Importance:

This feature importance plot after the model showcases the most relevant variables for model predictions. Feature importance becomes explicitly clear from this plot that Sales and Discount have a larger share. That means Sales and Discount hold the key position in selecting the customer categories. Let's see here:

Sales- The higher number of sales normally indicates high-valued customers and enforces revenue generation as vital in defining the category of the customer.

Discount: Similarly, Discount is very important to identify the customers, which implies that the amount of discount provided to a customer is an essential predictor of their value to the business.

Other features like Sub Category, Quantity, and category are less important but still very informative for the model. Based on these feature importance values; businesses can develop strategies focused on sales and discount optimization for maximizing customer profitability.

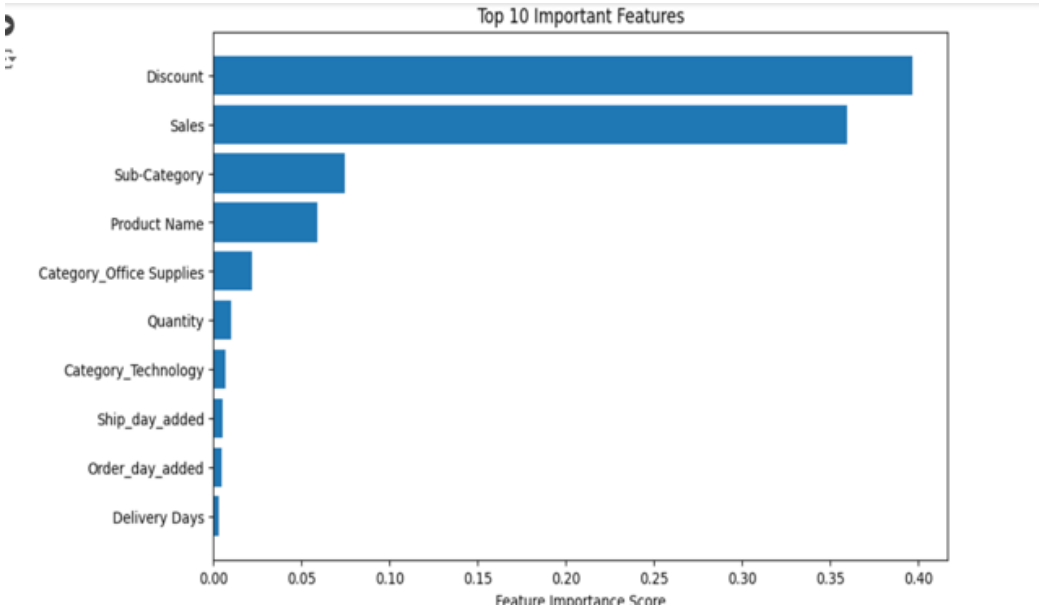

Figure 6 - Feature Importance

g. Advantages and Disadvantages of the Chosen Model

-Advantages:

1.GBC has high accuracy with flexible applicability, especially over datasets on the high side and at moments where no relationship between either other features and the target feature exists, or none are linear. In this work, as it could be seen above, GBC gave a good accuracy of 92.9%, indicating that even under worse circumstances, GBC could do some reliable predictions.

2. Handles Non-Linearity Well: GBC works well in cases when the relationship between features and the target is not linear, which in fact often is the case for real data. Since it uses decision trees as base learners, complex boundaries of decisions can be modeled; therefore, it's good for complex datasets where the interactions of the variables are varying.

3. Robust to Overfitting: It has regularization techniques such as shrinkage and tree pruning in Gradient Boosting; thus, this will avoid overfitting, especially if the model is trained with care by tuning the hyperparameters like the learning rate and tree depth. This will help it generalize well on unseen data.

4. Diversity: GBC also handles diversified data, both numerical and categorical variables. It is a most versatile model working in various financial, marketing, healthcare, and e-commerce arenas.

-Disadvantages

1.Greedy and Computationally Heavy: Generally speaking, GBC is computationally heavy, especially if there are a lot of estimators or when working on huge dataset problems. That would introduce more computational complexity at train time compared to simple model baselines without any optimized hyperparameters.

2. Sensitive to Noisy Data: Although resistant to overfitting, GBC can be sensitive to noise in the data. If there are a lot of irrelevant features or outliers in the data, the model will not work that well. Proper data preprocessing, such as handling missing values and outliers, is very important to enhance the performance of the model.

3.Hyperparameter Tuning: GBC necessitates serious tuning of hyperparameters. The learning rate, number of estimators, and even the depth of trees are examples of parameters that can severely affect the performance of a model. This will surely take more time and expertise to optimize.

4. Interpretable: GBC allows extracting features, but in general, such systems are still "black box" since the understanding of why decisions were made in each given forecast is not as transparent compared with simpler models, like a linear regression or decision tree. That may complicate the reasoning behind certain forecasts, for example, during the explanation of model outcomes for non-technical stakeholders.

VII. Dashboards and Visualizations

Dashboards were designed to provide a deep visual insight into the datasets and its outcomes, this will help managers to take the reasonable actions for the business. To add, it also helps in monitoring key metrics such as customer segments, profit in a specific region, sales during a specific time. And a good dashboard indicates a good business.
The creation of the below dashboards created by Power bi a great visualization tool, by Microsoft. That is interactive and easy to use.

The link for interactive visualizations:
https://app.powerbi.com/groups/me/reports/8ac7e6d7-9700-4582-bd8c-00a2d4eed197?ctid=a0f09f69-3caa-4b21-8efd-4ef01f250d10&pbi_source=linkShare

a. Overview Dashboard

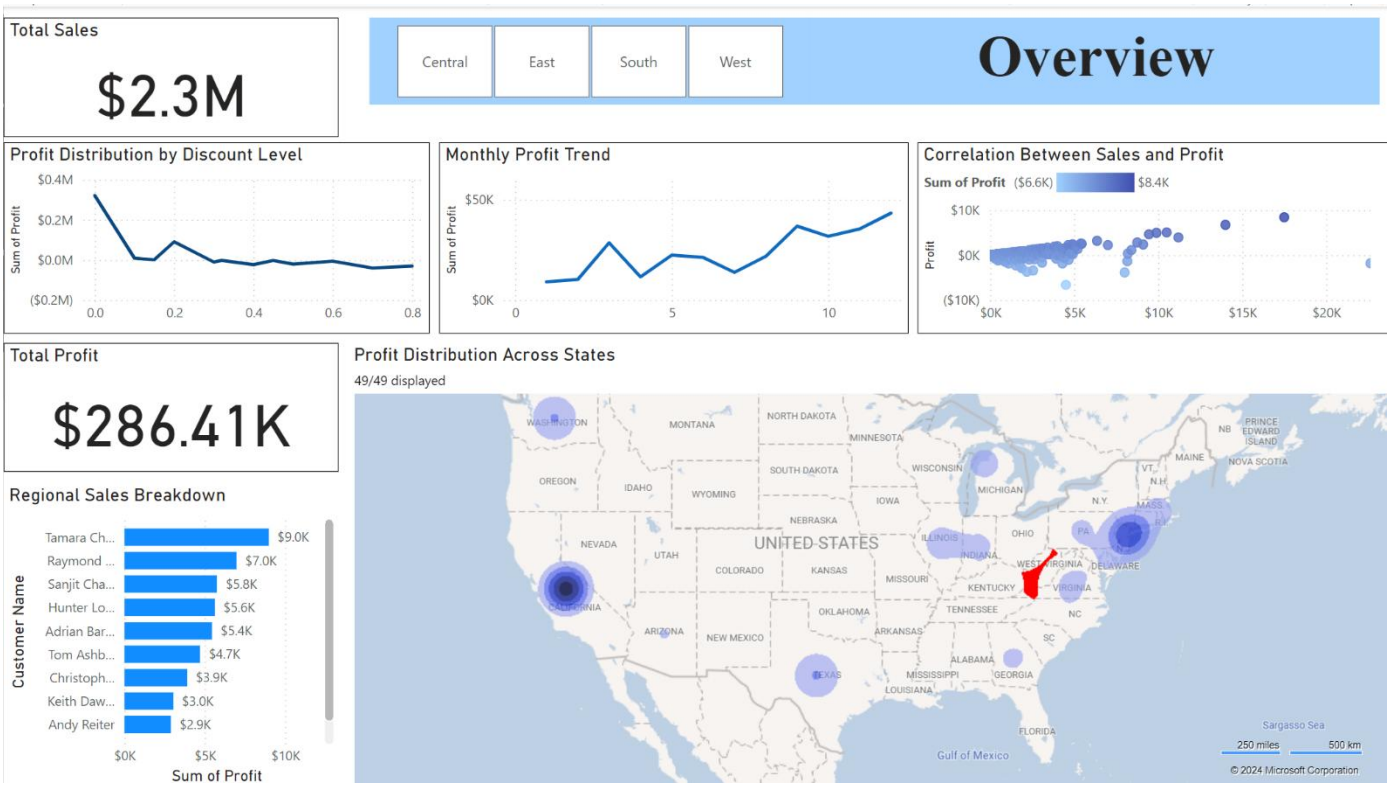

Figure 7 – Overview Dashboard

The above *Figure 7* shows a dashboard that provides the manager with all the needed information of the Superstore, especially of sales and profit data. it highlights that the overall total sale during 2014 to 2017 in US is ($2.3M) and profit($286.41K), along with a graph that describes the profit distribution by different levels of discounts, and it shows that the higher the discount the less the profit. A scatter plot that describes the relationship between profit and sales and it has now obvious relationship. Finally, the bar chart plots the top 10 customers that provided the Superstore with the highest profit, also there this US heat map that shows the profit distribution across the 49 states and we got California the most profitable state.

b. Customer Insights Dashboard

The dashboard in *Figure 8* provides a customer insights based on their profitability. The pie chart describes the customer distribution by probability, it categorizes Superstore customers into 3 categories, high-value customer (19.55%), low-value (49.94%), and unprofitable customers (30.52%) this indicates that 49.94 of the Superstore customers come with benefit to the store and that is a good indicator. The Bar chart shows the profit by customer segments, which type of customers that contributes the most to profit and it is the Consumer with ($134K), followed by Corporate ($92K) and Home Office ($60K). Finally, the Pareto Chart defines the top contributors to profitability, shows that some customers contribute only the top 20% customers contribute 80% of the profits, adhering to the 80/20 principle. Together, these charts give better customer targeting and decisions.

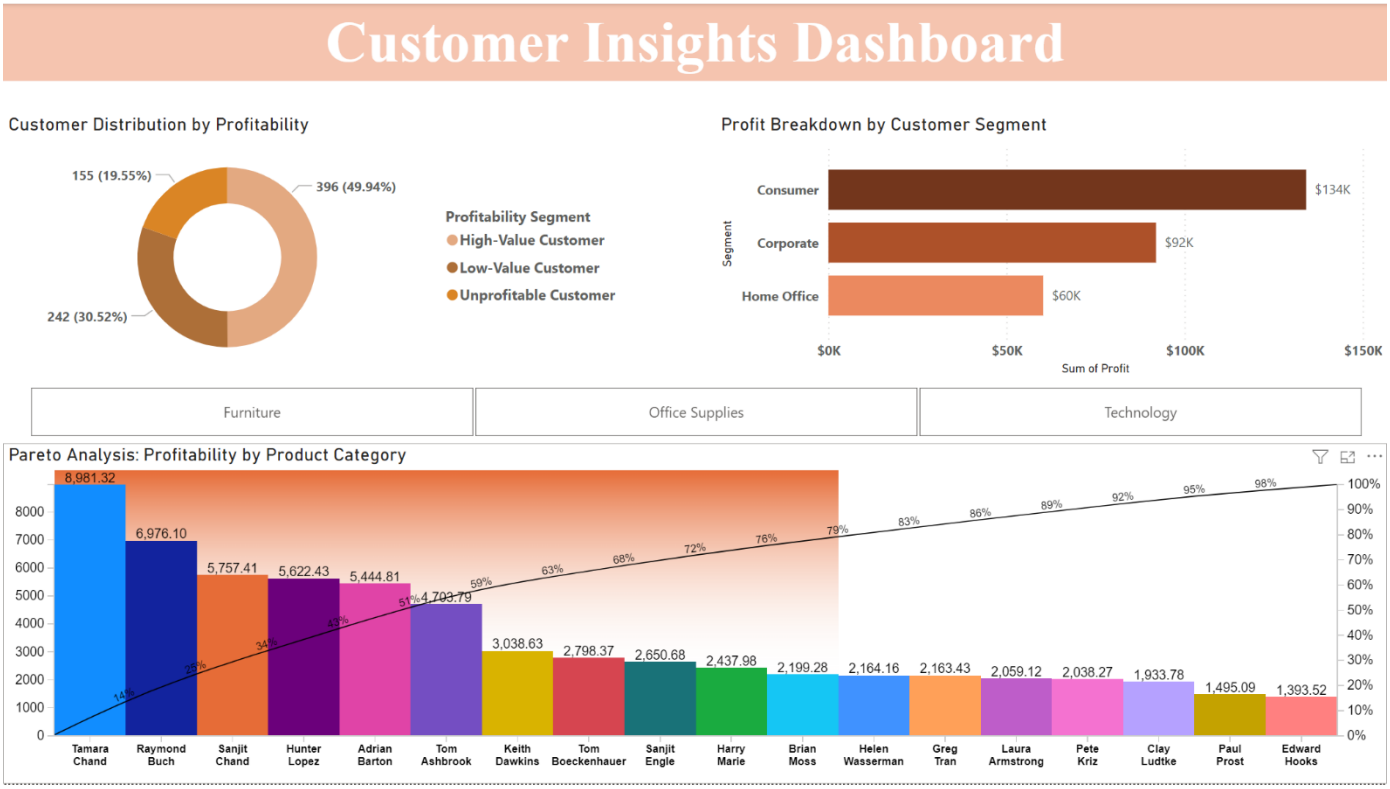

Figure 8 - Customer Insights Dashboard

c. Product & Category Analysis

Figure 9 - Product & Category Analysis

The *Figure 9* dashboard will provide an overview of the products and category performance, specifically in sales and profits, horizontal bar chart that represent profit by sub-category, highlights the most and the least sub-categories highlighting their profit and here we have Copiers, Phones, and Accessories with the most profitable sub-category. The second horizontal bar chart that shows the distribution of the category profit, indicates that Technology as the most profitable category and Furniture is the least. The third and last pinpoint the total sales of each product and in our situation Canon Image Class got the most sales. Furthermore, the dashboards provide us with the important key metrics including 3 types of Categories, 17 type of sub-category, 3 segments, and 1850 total of products. Lastly, a timeline slicer that helps the manager to follow up with trends for each product and category.

d. Regional Profitability Dashboard

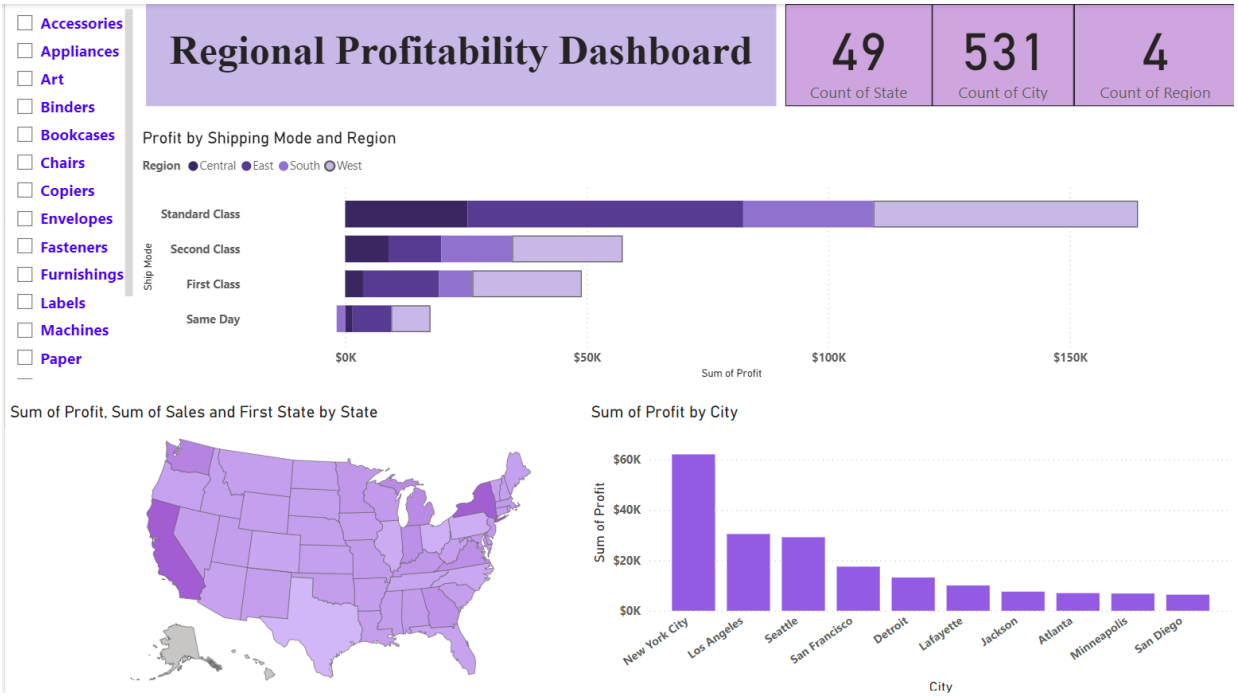

Figure 10 – Regional Profitability Dashboard

The *Figure 10* dashboard provides a detailed view profit distribution across the US, it highlights how can the shipping mode affect the profit and which of the three mode is the most profitable across Regions (Standard Class, Same Day, Second Class, First Class). The US map visualize the profit and sales distribution among states, showcasing the states that contribute the most in profits, that helps in locating more stores in that specific state, like California. to add on Cities like New York and Los Angeles stands out with the most profitable cities that what the bar chart draws. The dashboard makes it easy for mangers to understand the geographical factors that impact the profit.

e. Time Analysis Dashboard

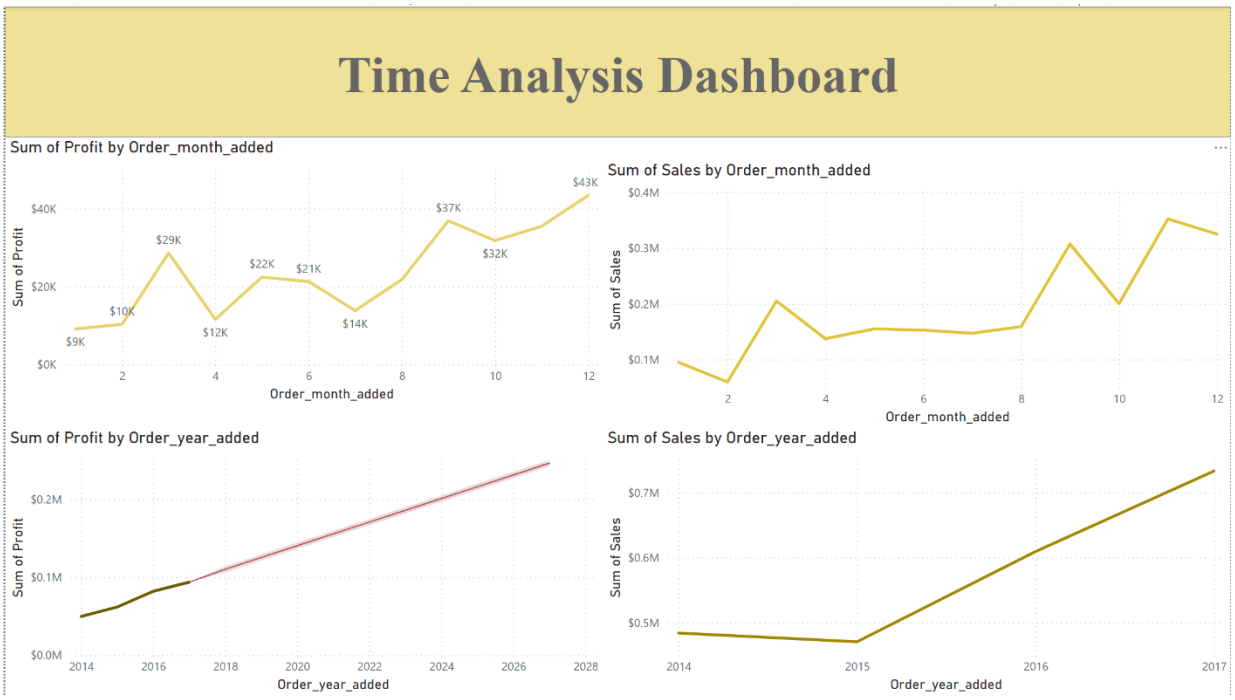

Figure 11 – Time Analysis Dashboard

Figure 11, The final dashboard presents time-based trends in profits and sales. Its shows annual and monthly trends that highlight sales and profitability how do they change over time. Monthly trends help in discovering seasonal increase, on other hand yearly trends indicate long-term performance growth. For example, through the years sales and profits have been increasing consistently, with some months reaching higher than usual levels. This dashboard will help business to plan, scheduling activities at times of excellent performance, and monitor their future growth.

VIII. Conclusion

In conclusion, the implementation of a Gradient Boosting Classifier for customer profitability analysis enables businesses to accurately identify and segment their customers based on profitability, thereby facilitating targeted strategies that enhance operational efficiency and maximize profitability. This analytical approach not only addresses the complexities of customer behaviour but also provides actionable insights for improving underperforming areas and fostering customer retention.

IX. References

- ashishtomar99. (2020). *GitHub - ashishtomar99/ColgatePalmolive-Amazon_Profitability_IndustryPracticum: Amazon profitability prediction and trade profit optimization using SKU bundling strategy for a New York based Fortune 200 CPG company*. GitHub. https://github.com/ashishtomar99/ColgatePalmolive-Amazon_Profitability_IndustryPracticum
- Bhadoria, A. (2020, October 28). *Customer Profitability Analysis: Definition, Formula, Benefits - SmartKarrot*. SmartKarrot L Comprehensive Customer Success. https://www.smartkarrot.com/resources/blog/customer-profitability-analysis/
- Chowdhury, V. (2021). *Superstore Dataset*. Www.kaggle.com. https://www.kaggle.com/datasets/vivek468/superstore-dataset-final
- *Customer Profitability Analysis*. (2024, June 13). DealHub. https://dealhub.io/glossary/customer-profitability-analysis/
- Gupta, A. (2020, December 24). *Regression Analysis In Excel With Example - Simplilearn*. Simplilearn.com. https://www.simplilearn.com/tutorials/excel-tutorial/regression-analysis
- Kenton, W. (2023). *What is pareto analysis? How to create a pareto chart and example*. Investopedia. https://www.investopedia.com/terms/p/pareto-analysis.asp
- PAGE. (2022, September 29). *Creating ONE PAGE with MULTIPLE TABS in Power BI Report using BOOKMARK NAVIGATOR BUTTONS*. YouTube. https://youtu.be/Ng1EylK8KXU?si=W8NrAzMsG5aDmMJ2
- *SALES ANALYSIS ON SUPERSTORE DATASET*. (2023). https://doi.org/10.56726/irjmets36572
- *Sign in | Microsoft Power BI*. (n.d.). Powerbi.microsoft.com. https://app.powerbi.com/

X. Appendices

Appendix A: Tools Used

- Kaggle: platform for data scientist and machine learners to achieve their goals in designing and building predictive models or even descriptive analysis for penalty of datasets

https://www.kaggle.com/datasets/vivek468/superstore-dataset-final

- Google Colab: platform by google used to implement python code in your own browser. It is well suited for data scientists to make their ideas work.

https://colab.google/

Microsoft Power BI: Microsoft developed software where simple visualization became interactive, easy to use, simple, and got many options. Mainly used for business intelligence where you share your latest insights and ideas.

https://www.microsoft.com/en-us/power-platform/products/power-bi