**DSC 550 Final Term Project**

**Topic: Sales Analysis for Customer Retention**

**Author: Era Ebhodaghe**

**Introduction**

XYZ Company is currently facing challenges in accelerating its sales revenue growth. Over the years, the company has experienced slower-than-expected revenue increases, which has raised concerns about its overall profitability and long-term sustainability. As the company prepares its annual budget for the upcoming fiscal year, it is crucial to identify key areas for improvement and investment.

The primary objective of this analysis is to assess various aspects of the business by evaluating the company’s sales reports. By identifying patterns and trends, the goal is to develop strategic recommendations that can drive revenue growth, enhance customer retention, and improve overall customer satisfaction. Addressing these challenges is essential for boosting profitability and ensuring the company remains competitive in the market.

**Stakeholders and Project Impact**

Conducting this analysis will require data collection and in-depth evaluation, which may lead to additional expenses for the company and its stakeholders. However, the insights gained will provide valuable information to guide investment decisions.

To support this initiative, I will be presenting a customer retention strategy to key stakeholders. This strategy aims to increase customer subscriptions, a critical factor in driving sustained revenue growth. The analysis will identify the top five variables that most significantly influence a customer’s likelihood to subscribe to XYZ Company's services.

By understanding these key drivers, the company’s board will be able to allocate resources more effectively, focusing on areas that have the highest potential for increasing subscription rates and overall sales revenue. This targeted approach will ensure that investments are made strategically, maximizing returns while improving customer engagement.

**Data Source**

For this analysis, the dataset was obtained from Kaggle, a well-known open-source platform for data science and analytics. The selected dataset is relatively clean, minimizing the need for extensive preprocessing and allowing for a more efficient analysis process.

The dataset can be accessed via the following link:  
[Customer Shopping Latest Trends Dataset](https://www.kaggle.com/datasets/bhadramohit/customer-shopping-latest-trends-dataset)

By leveraging this dataset, I aim to extract meaningful insights that will help XYZ Company refine its sales and marketing strategies, ultimately leading to stronger revenue growth and improved business performance.

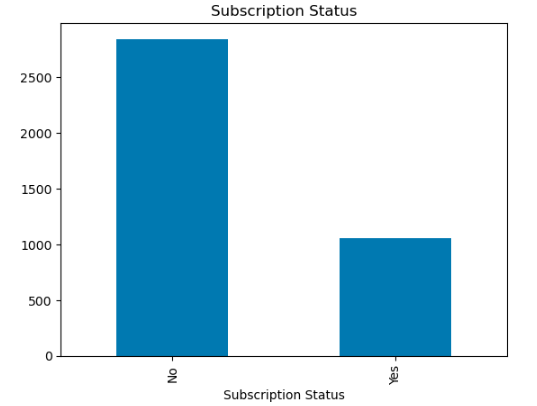
**Summary of Project Milestones**

The project began by loading the dataset and ensuring that it was properly imported without errors. After confirming the integrity of the data, I proceeded with an initial exploration to understand its structure and key characteristics. This involved examining the statistical summary of non-categorical variables, allowing me to gain insights into the distribution, range, and central tendencies of numerical features.

To enhance my understanding of the data, I selected specific columns for visualization. Visual representation is a powerful tool in data analysis, as it helps identify patterns, trends, and potential outliers that may not be immediately apparent from raw data. For this project, I initially created five visualizations, with the flexibility to generate additional ones if needed. I anticipated that once the key factors influencing customer subscriptions were identified, supplementary visuals might be required to support budget allocation and strategic decision-making.

Given that the primary focus of this analysis is subscription status, I have included a critical visualization of the sales data. This visualization highlights a significant imbalance between customers with active subscriptions and those without. The data reveals that a substantially larger proportion of customers do not have a subscription compared to those who do. This disparity underscores the importance of identifying factors that drive customer conversions and devising strategies to increase the subscription rate.

Having established a clear understanding of the current subscription distribution, the next step involves conducting a deeper analysis to predict subscription status based on other available variables. By leveraging statistical techniques and machine learning models, the objective is to uncover key drivers that influence a customer's likelihood of subscribing. These insights will be instrumental in guiding the company’s marketing strategies, customer engagement efforts, and resource allocation to optimize revenue growth.

Fig 1: Subscription Status Count

Data Exploration

In order to analyze the data, I proceeded to examine the number of rows in the dataset with missing values. Out of a total of 3900 rows, there were no missing values in the dataset. Then I proceeded to explore the categorical columns with visuals to better understand the data spread.

The numerical columns were further explored with visuals as well to identify outliers. There were no apparent outliers in the dataset with the data evenly distributed as shown below:

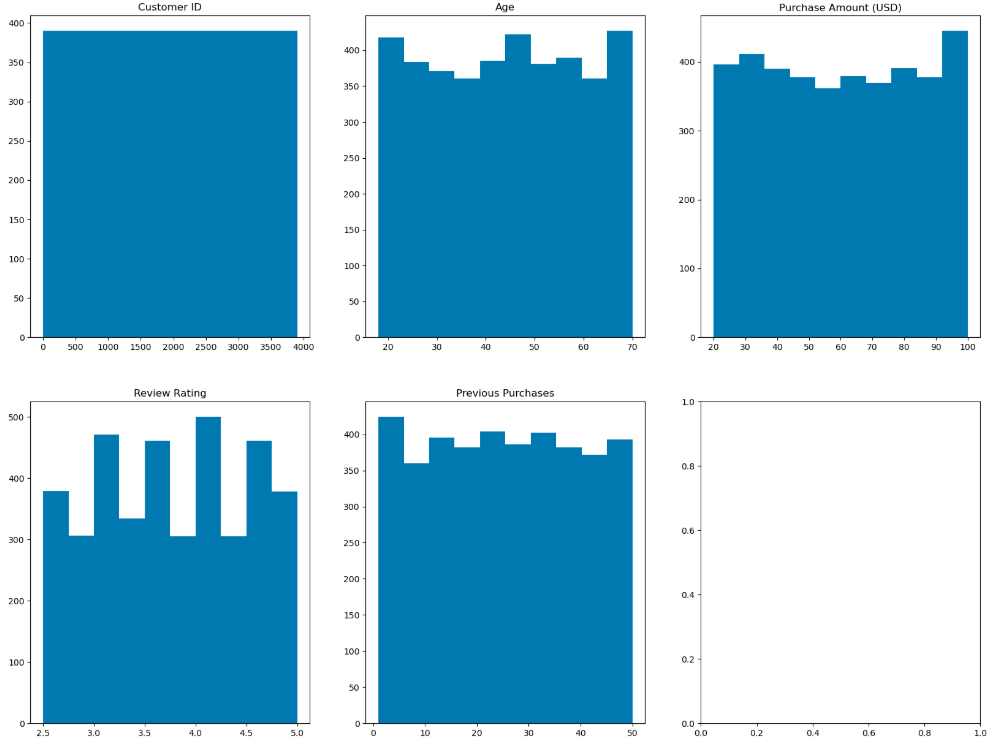


Fig 2: Plot for numerical columns

The next focus was the analysis of the subscription column. There are 1053 customers with subscriptions and 2847 customers without subscriptions. The goal is to identify ways to increase the number of subscriptions.

Data Cleaning

Although the dataset was fairly well-structured, data cleaning was essential before processing. I converted categorical columns with "yes" and "no" values into Boolean expressions and transformed other categorical features into dummy variables. These modifications are crucial for preparing the data for machine learning models.

Additionally, unnecessary columns such as Promo Code Used and Customer ID were removed, as they provided no meaningful value to the analysis and could potentially impact results negatively.

Model Building and Evaluation

Once the data was cleaned, I prepared it for modeling. Using sklearn’s train\_test\_split, I split the dataset into training and test sets. Since this is a classification problem, I implemented a Python dictionary for parameter tuning and utilized a for-loop for grid search. This approach allowed me to test multiple algorithms simultaneously to determine the model with the highest accuracy and optimal parameters.

To further refine the model, I leveraged sklearn’s feature selection to identify the five most influential factors affecting subscription likelihood. The top five features were **Age, Gender, Shipping Type, Discount Applied,** and **Previous Purchases**.

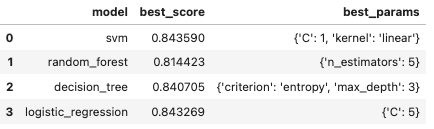


Fig 3: Model accuracy and parameters

Following feedback from the last project milestone, I evaluated Precision, Recall, F1 scores, and the confusion matrix. While my model's overall accuracy was high, the F1-score for the "subscription = yes" class was significantly lower. This is primarily due to class imbalance, as shown in the support column. Out of 780 examples, 222 belong to class 1 (subscription = "yes"), while 558 falls under class 0 (subscription = "no"). As a result, the model performs well in predicting class 0 but struggles with class 1 due to the imbalance.

Additionally, the confusion matrix in Fig. 5 highlights a notable number of false positives, which can also be attributed to the class imbalance.

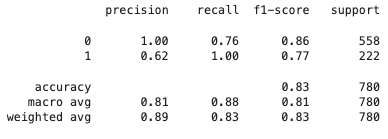
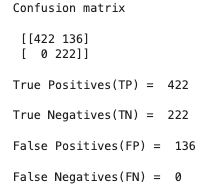


Fig 4: Precision, recall, f1-score and support

Fig 5: Confusion matrix

**Conclusion**

The model is not yet ready for deployment, as class imbalance needs to be addressed before rerunning the model. Due to time constraints, I was unable to implement class balancing in this project.

Once an optimized model is developed, my recommendation to the company would be to invest in targeted marketing, leveraging the five key features identified. Additionally, operational adjustments could be made by refining shipping options, discount structures, and analyzing previous purchase patterns.

This process should be implemented iteratively, with continuous monitoring of subscription levels following these changes. Regular reviews and refinements will be essential to maximize customer conversions and further increase subscriptions over time. By adopting a data-driven approach, XYZ Company can improve customer retention, boost revenue, and maintain a competitive edge in the market.