Prepared for: Head of analytics and Sales Team

This report provides an analysis of the dataset concerning sales approaches for the new product line. The objective of this analysis is to validate and clean the data, conduct exploratory analysis to understand the characteristics and relationships of key features, define a business metric for monitoring, and offer initial insights and recommendations.

Data Validation

In this section, I describe the steps taken to validate and clean the dataset, focusing on each column. The goal of this process was to ensure that the data is accurate, complete, and usable for analysis.

- The column of the sales_method contained several inputs for only three sales methods, so the labels were unified to only contain three sales methods: "Email", "Email + Call", and "Call".
- Checking whether each column was assigned the appropriate data type or not. Each column was assigned the appropriate data type. In addition to a description for the data to view whether it contains nulls or not. Only the revenue column contained nulls and those records represented a 7.16% of the data which is considered a very small portion so no actions were taken for the analysis in hand.
- Since the company was established in 1984 and we are in 2024 then this means that it has been 40 years since the company's establishment, thus any year more than 40 years doesn't make sense accordingly two rows were deleted from the data.
- Checking for duplicates. There were no duplicates in the data.
- Checking for Outliers, and 1978 rows in the data were detected to be outliers, and the outliers in the revenue column were detected to be 634, so the revenue log-transformation were executed that will help us in further analysis like predictive models and no outliers were detected in the revenue column after the log-transformation (fig 1, 2 in the code).
- The two revenue columns were plotted to see the distribution of the revenue before and after the transformation (fig 3 in the code). It became less skewed and more stable however since our goal is to understand the raw distribution of revenue or how much each customer is contributing to overall sales, the original revenue will be used in the rest of the analysis to solve the business problem in hand.
- Descriptive statistics were calculated to view the log-transformed revenue against the original one and it was found that the mean is much reduced in the log-transformed revenue column and standard deviation is much lower indicating a more stable normalized revenue.
- There is a positive correlation between no. of new products sold and weeks and a positive correlation between revenue and no. of new products sold, which implies that when one of those variables increase the other variable will increase as well.

Explanatory Analysis

To better understand the dataset and uncover patterns, I conducted exploratory data analysis (EDA) using both univariate and multivariate analysis techniques.

- In the donut chart (fig 5 in the code) the revenue from each sales method is calculated. It is concluded from the graph that Email method has the highest revenue (672220.61) in comparison to the Call + Email (408256.69) and Calls (227513.02) method.
- In the bar chart (fig 6 in the code) the number of customers in each sales method is calculated. It is concluded from the graph that Email method has the highest number of customers (7465 customers) in comparison to the Call + Email (2572 customers) and Calls (4961 customers) method.
- In the boxplot and histogram (fig 7, 8 in the code) the spread of overall revenue is plotted. As viewed in the graph the spread of the revenue is quite wide, which means there is a lot of variations in the revenue. The revenue is somehow unpredictable, this could mean a mix of high-risk high-reward opportunities, as well as some volatility.
- In the 3 histograms and boxplots (fig 9, 10 in the code) the spread of the revenue in the three sales approaches is plotted.
- In the Email histogram most of the revenue values fall between 75 and 150, and the density plot peaks around 100, indicating this is the most common revenue for this method.
- In Email + Call histogram revenue values are more spread out, ranging from 150 to 250, Multiple peaks in the density plot, especially around 150 and 200, suggesting variability in revenue.
- In the Call histogram the majority of revenue values also fall between 25 and 75, A peak around 50 in the density plot, but with a slightly higher density.
- The Email + Call is the most stable method that doesn't contain outliers.
- If you're aiming to maximize revenue, the Email + Call method looks promising due to its wider range and multiple high peaks. And for consistency, the Email or Call methods are more stable with revenue concentrated around 100.
- In the following line chart (fig 11 in the code) the trends of the three sales method is plotted against the number of sales weeks since the product launch. As shown the three sales method is increasing overall by time.
- In the following heat map (fig 12 in the code) the states are visualized with the number of new products sold. Showing that California State is the highest in the number of products sold which implies that most of our customers are located in California.

Definition of a metric for the business to monitor

Based on the exploratory analysis and the business problem at hand, I recommend defining the following metric for monitoring the business's progress:

- To monitor the success of the new sales methods, the team should focus on key metrics such as Revenue per Customer, Conversion Rate, and Sales Method Efficiency. By comparing these metrics across the three sales methods (Email, Call, and Email & Call), we can determine which method delivers the best results. Additionally, tracking Sales Trends over time will provide insights into the sustainability and long-term effectiveness of each sales approach.

- The revenue per customer was calculated per sales method and it was found that the Email+Call method has the highest revenue per customer (158.7). The conversion rate needs more data to be included as it is calculated by dividing the number of purchases by the number of customers contacted, so if these data was included the conversion rate could be calculated and it helps assess how effective each sales method is at converting potential leads into actual sales. In addition to tracking revenue growth over time for each sales method will help identify whether the method continues to be effective as time passes. This will help assess if a method that works initially starts losing effectiveness or if another method becomes more successful over time.

Given the results, I recommend the company continue using the Email & Call method, as it offers a good balance of efficiency and revenue generation. However, further testing over a longer period is suggested to determine if the benefits of the hybrid method are sustained over time. We could also streamline the call duration to maximize the time-to-revenue ratio without sacrificing too much in terms of conversion rate.