CRM Data Analysis Case – Technical Overview

1. Data Exploration

In the initial phase of the CRM data analysis, our goal was to understand the dataset, identify key variables, and uncover preliminary insights. We followed a systematic approach, which included:

- **Data Loading & Cleaning:** We used *panda* to load the data and inspect its structure using *info(), head(),* and *describe()* methods. This allowed us to get a sense of the data types, missing values, and basic statistics.
- **Missing Value Handling:** Upon checking for missing values with *isnull(),* we confirmed that there were no missing entries, which streamlined our analysis process.
- **Outlier Detection & Data Consistency:** We visualized key features like *Income* using histograms to detect outliers and anomalies. This helped us understand the spread of income levels and spending behavior among the respondents.
- **Feature Exploration:** We used *seaborn* and *matplotlib* for data visualization, creating histograms, bar plots, and box plots to explore features like *Income, Recency,* and various spending categories (e.g., *MntWines, MntMeatProducts).* This provided insights into respondent demographics and their preferences.
- **Descriptive Statistics:** Through calculating metrics such as means, medians, and quartiles, we developed an understanding of income distribution, spending habits, and engagement levels across different segments.

Tools Used: *pandas* for data manipulation, *seaborn* and *matplotlib* for visualization, *numpy* for numerical calculations.

Outcome: The exploration phase allowed us to identify key variables for further analysis, such as income, spending patterns, and campaign responses. It provided the foundation for the segmentation and modelling phases by highlighting which variables are most relevant for customer analysis.

2. Segmentation

To better understand our customer base, we implemented a segmentation analysis using clustering techniques. The objective was to group customers based on similarities in their behavior, enabling targeted marketing strategies. Here's how we approached this:

- **Feature Selection:** We selected key variables related to customer behavior, including *Income, Recency,* and spending on various product categories. The features were chosen based on their relevance to customer engagement and potential to differentiate between segments.
- **Standardization:** Since clustering algorithms are distance-based, we standardized the data using *StandardScaler* from *sklearn* to ensure that each variable contributed equally to the distance calculations.

- **K-Means Clustering:** We applied the K-Means clustering algorithm to the scaled data. Using the **Elbow Method,** we identified the optimal number of clusters, choosing *K=4* where the within-cluster sum of squares (WCSS) began to stabilize.
- **Cluster Interpretation:** After assigning each customer to a cluster, we analyzed the mean values of key features within each cluster. This allowed us to understand the defining characteristics of each segment, such as income levels, spending habits, and shopping preferences.
- **Visualization with PCA:** For better interpretability, we used **Principal Component Analysis (PCA)** to reduce the data to two dimensions and visualized the clusters in a 2D plot. This made it easier to see how customers were grouped and how distinct each segment was.

Tools Used: *sklearn* for *StandardScaler, KMeans,* and *PCA, pandas* for data manipulation, *seaborn* and *matplotlib* for visualization.

Outcome: The segmentation provided actionable insights into different customer groups, such as high-income customers preferring premium products and more price-sensitive segments. It allowed the marketing team to tailor strategies based on specific cluster needs, maximizing engagement and profitability.

3. Predictive Model

The final phase involved building a predictive model to forecast customer responses to marketing campaigns. This model aimed to identify the customers most likely to engage with future campaigns, thereby optimizing resource allocation. The process included:

- **Feature Selection:** For predictive modeling, we focused on features that were likely to influence campaign response, such as *Income, Recency, MntWines, NumWebVisitsMonth,* and household composition (*Kidhome, Teenhome).* The target variable was *Response,* which indicates whether a customer accepted the campaign offer.
- **Handling Class Imbalance with SMOTE:** We used **Synthetic Minority Over-sampling Technique (SMOTE)** to address the imbalance in the target variable, ensuring that both responder and non-responder classes were well represented in the training data. This helped in avoiding biased predictions towards the majority class.
- **Model Selection & Training:** A **Random Forest Classifier** was chosen for its ability to handle complex relationships between variables and provide feature importance scores. The model was trained using an 80/20 train-test split and evaluated on accuracy, **ROC AUC,** and **classification report** to assess performance.
- **Model Evaluation:** We plotted the **ROC Curve** to visualize the trade-off between true positive rate and false positive rate, providing insights into the model's ability to distinguish between responders and non-responders. We also used a **confusion matrix** to analyze the model's precision and recall.
- **Feature Importance & Recommendations:** After training, we analyzed which features contributed most to predicting customer response. This helped us identify key factors influencing engagement, such as recent purchases, income, and discount sensitivity. We used this information to refine marketing strategies.

Tools Used: *imblearn* for *SMOTE*, *sklearn* for *RandomForestClassifier, train_test_split,* and evaluation metrics, *pandas* for data management, *seaborn* and *matplotlib* for ROC and confusion matrix plots.

Outcome: The predictive model achieved high accuracy and a strong ROC AUC score, indicating that it could reliably identify potential responders. By understanding the key drivers of response, the marketing team can better target high-probability customers, improving the efficiency and impact of future campaigns.