SUMMARY – LEAD SCORING CASE STUDY

Submitted By: Aditi Marwaha and Smriti Pradhan (Batch: DS C57)

Problem Statement:

X Education generates a significant number of leads, but its lead conversion rate is relatively low, standing at approximately 30%. The company has tasked us with developing a model to assign a lead score to each lead, aiming to prioritize leads with higher scores that indicate a greater likelihood of conversion. The CEO's target for the lead conversion rate is set at around 80%.

Analysis Approach:

In our analysis of the provided dataset, we followed a systematic approach to ensure a thorough examination of the data. Firstly, we identified relevant columns based on the Data Dictionary and proceeded to eliminate any invalid or redundant columns that may not contribute meaningfully to our analysis. To ensure data integrity, records with over 40% missing data were removed, and we strategically imputed missing values in selected columns.

A critical step in our analysis involved identifying potential data columns that could significantly impact accurate predictions. To gain insights into the distribution and relationships within the dataset, we employed visualizations, such as graphs, and removed outliers in numerical variables to enhance the quality of our analysis.

To prepare the data for modeling, we encoded categorical data into dummy variables, facilitating their integration into a predictive model.

The dataset was then divided into training and test sets in a 70:30 ratio, with the training data being scaled to address any disparities in data magnitude that could impact model predictions. The Generalized Linear Model (GLM) was applied to the training data, and ineffective variables were further refined using Recursive Feature Elimination (RFE) and Variance Inflation Factor (VIF) techniques.

The resulting model, consisting of 12 variables and one constant, demonstrated an impressive performance with over 80% accuracy and precision in predicting the training dataset. This well-tuned model was subsequently applied to the test dataset, which had also been appropriately scaled, revealing consistent results with accuracy and precision exceeding 80%.

The formula from the final LogReg model is:

10.9851**Tags_Lost to EINS* + 9.2047**Tags_Closed by Horizzon* + 5.3743**Tags_Will revert after reading the email* + 2.6362**Tags_Busy* + 1.9729**Last Activity_SMS Sent* + 1.6948**Lead Origin_Lead Add Form* + 1.2544**What is your current occupation_Working Professional* + 0.9371**Total Time Spent on Website* - 1.1699**Tags_Ringing* - 1.3787**Last Notable Activity_Modified* - 1.8118**Last Activity_Email Bounced* - 3.7292**What is your current occupation_Other*

Conclusion:

The insights derived from the analysis highlight several key points:

1. Customers or leads who actively fill out forms represent potential leads for conversion.

2. Targeting working professionals is recommended, as they exhibit a higher probability of conversion and are likely to possess better financial capabilities for service payments compared to those who haven't specified their occupation.

3. Leads with 'Last Activity' recorded as 'SMS Sent' demonstrate a higher conversion rate, suggesting that they should be prioritized in targeted marketing efforts.

4. Analyzing the behavior of customers who spend more time on the website can enhance user experience and increase conversion rates. Therefore, the company should focus on creating compelling content and ensuring user-friendly navigation to encourage extended website engagement.

5. Understanding the popularity of different specializations enables tailored course offerings and marketing campaigns. Providing targeted content and resources, particularly for popular specializations like Management, can attract and retain customers in those specific fields. These insights contribute to a more effective and targeted approach in customer engagement and business strategies.