Lead Scoring Summary

In this assignment, our objective was to build a logistic regression model to assign a lead score to each lead, helping X Education improve their lead conversion rate by focusing on the most promising leads. Here's how the assignment was approached, and the key learnings gathered throughout the process:

**1. Data Exploration and Preparation**

Our first step was to thoroughly explore the dataset, which comprised 9,240 data points and 37 attributes related to leads. We conducted a detailed Exploratory Data Analysis (EDA) to understand the distribution, relationships, and characteristics of each feature. This process also included identifying and addressing missing values and null categories, such as 'Select', which did not provide any meaningful information.

We discovered that several columns had varying levels of missing data, ranging from 1% to as high as 50%. Additionally, some columns had multiple categories, while others contained outliers, such as ‘TotalVisits’. Columns with more than 30% missing data were deemed unfit for analysis and were removed. Furthermore, any columns with constant values were dropped, as they added no variability or predictive power to the model.

For the remaining columns, we ensured proper imputation: missing numerical values were filled using mean or median imputation, and categorical values were replaced with a new category, 'Unknown', to preserve the dataset's integrity. This thorough data cleaning process set a solid foundation for accurate and reliable model building.

**2. Model Building and Evaluation**

After cleaning and preparing the dataset, we split it into training and testing sets (70-30 split) to build the logistic regression model. To ensure consistency, we scaled all the features before beginning the model building. We used Recursive Feature Elimination (RFE) to remove less relevant features systematically.

During RFE, we eliminated features with high p-values that did not significantly influence the target variable ("Converted"). After several iterations, we refined the model, achieving a Variance Inflation Factor (VIF) below 5 and a p-value threshold of 0.5, which indicated multicollinearity and statistical significance were within acceptable limits. We then proceeded to make predictions.

Our first model performed well, with an Area Under the Curve (AUC) of 89%. Despite attempting further improvements, the second iteration showed a decline in performance. Therefore, we reverted to the first model to fine-tune the cut-off threshold. We identified an optimal cut-off value of 0.3, where sensitivity (true positives / (true positives + false negatives)) was maximized without compromising accuracy and specificity.

Finally, we applied the model to the test data and evaluated the metrics, which confirmed the model’s robustness. The evaluation metrics met our expectations, and we finalized the model based on these results.

**3. Insights and Learnings**

Analyzing the model coefficients, we identified that the most influential variables are:

A screenshot of a computer

Description automatically generated

These variables are strong indicators of lead engagement and intent, revealing that engaged behaviour is crucial for conversion.

**4. Conclusion**

Based on the model results,

1. Although the initial dataset contained numerous features that could potentially influence lead conversion, multiple iterations of model development helped identify the most impactful ones.
2. Leads with a high VIF score are more likely to convert.
3. Improving customer engagement through targeted emails and calls, especially focusing on Lead Origin\_Lead Add Form, can significantly enhance conversion rates. Sending SMS messages also proves beneficial in engaging leads.