Summary: Lead Scoring Case Study

Problem Statement

An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google which creates their leads when these people fill up a form providing their email address or phone number. Various sources are:

- Search engine
- Social platforms
- Direct website
- Referrals

Current conversion for these leads is 30% and the company wants to assign a lead score for each lead so that the conversion cut-off improves to 80%.

Objective

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

2. Provide recommendations on the basis of model for future situation.

Solution Approach

The process involved multiple stages, each designed to refine the data and enhance the predictive model for assigning lead score to improve conversion rate.

1. **Data Understanding:** After the data is imported functions like info(), describe(), unique(), head(), shape helps to understand the structure of the dataset, the different variables and their datatypes, missing values, etc.
2. **Data Cleaning:** The initial step involved cleaning the dataset, which was generally well-organized but contained some imperfections. Missing values were addressed by imputing the mode for categorical variables and median for numerical columns. In some cases, the null values were dropped. Some categorical datapoints like Last Activity, Tags were categorized into one bucket to simplify analysis and improve clarity.
3. **Exploratory Data Analysis (EDA):** Exploratory Data Analysis was performed to assess the quality and relevance of the data. It was noted that some categorical variables included irrelevant or redundant information, which could distort the results. However, numeric values were accurate although contained outliers that need to be handled, which facilitated further analysis.
4. **Creation of Dummy Variables:** Categorical variables were converted into dummy variables to prepare the data for machine learning algorithms. Additionally, MinMaxScaler was applied to normalize the numeric data, ensuring all variables were on a comparable scale.
5. **Train-Test Data Split:** The data was divided into training and testing subsets, with 70% allocated for training the model and 30% for testing its performance. This approach allowed

the model to be trained on one subset and evaluated on another, ensuring its effectiveness in predicting outcomes.

6. **Model Development:** For model development, Recursive Feature Elimination (RFE) was used to identify the top 15 relevant variables from the dataset. RFE systematically removed less significant variables to enhance model accuracy. Subsequently, additional filtering was performed based on Variance Inflation Factor (VIF) and p-values, retaining variables with VIF values below 5 and p-values below 0.05.
7. **Model Evaluation:** The model's performance was assessed using a confusion matrix to determine its classification based on Recall value. The optimal cutoff value was determined using the ROC (Receiver Operating Characteristic) curve, resulting in accuracy, recall, sensitivity, and specificity rates of over 84%. These metrics indicated that the model was effective in predicting potential leads.
8. **Prediction Analysis:** Predictions were conducted using the test data with an optimal cutoff value of 0.42. This cutoff was chosen to balance accuracy and recall, which were maintained around 80%, demonstrating the model's reliability.
9. **Precision-Recall Assessment:** To further validate the model, a precision-recall analysis was performed. Adjusting the cutoff to 0.42 yielded a recall rate of 84%. Precision measures the proportion of true positive predictions among all positive predictions, while recall assesses the proportion of actual positives identified by the model.

Key Insights

1) Top three variables in your model which contribute most towards the probability of a lead getting converted
    - Tags_Closed by Horizzon
    - Tags_Lost to EINS
    - Total Time Spent on Website

2) Top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion
    - What is your current occupation_Unemployed",
    - Tags_Ringing
    - Lead Origin_Landing Page submission

3) **Total Time Spent on the Website:** Longer website engagement is strongly correlated with higher conversion likelihood.

4) **Total Number of Visits:** Increased visit frequency is indicative of greater interest.

5) **Lead Source:** Leads originating from Google, direct traffic, and organic search were more valuable compared to other sources.

6) **Last Activity:** Engagement through SMS and chat conversations was associated with higher potential for conversion.

7) **Lead Origin and Occupation:** Leads from specific formats and those from working professionals showed higher conversion potential.

By leveraging these insights, X Education can tailor its marketing strategies to better target the potential leads.