Lead Scoring Case Study Summary

Problem Statement:

- X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead scores have a higher conversion chance and the customers with lower lead scores have a lower conversion chance.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary:

Step 1: Importing Libraries and Data

1. Import required libraries like NumPy, Pandas, Matplotlib, Seaborn, Statsmodel, and Scikit-learn.
2. Reading the dataset and understanding the dataset using pandas info and describing functions.

Step 2: Data cleaning

1. There were a few columns with the value 'Select' which means the leads did not choose any given option. We changed those values to Null values.
2. Dropping the column having null values greater than 40%.
3. Value counts within categorical columns were checked to decide the appropriate action: if imputation causes skew, then the column was dropped, created a new category (others), impute high-frequency value, drop columns that don't add any value.
4. Numerical categorical data were imputed with mode.
5. columns with only one unique response from a customer were dropped.
6. We checked for outliers and used the Capping/Winsorization method to treat them.
7. We also fixed invalid data and grouped low-frequency values.

Step 3: Exploratory Data Analysis

1. Data imbalance checked: only 38.5% of leads converted.
2. Performed univariate and bivariate analysis for categorical and numerical variables. 'Lead Origin', 'Current occupation', 'Lead Source', etc. provide valuable insight into the effect on the target variable.
3. Time spent on the website shows a positive impact on lead conversion.

Step 4: Data Preparation

1. Changed the binary variables into '0' and '1'.
2. We created dummy variables for the categorical variables.

3. The next step was to divide the data set into test and train sections with a proportion of 70-30% values.
4. We used the Min Max Scaling to scale the original numerical variables.

Step 5: Model Building

1. Used RFE to reduce variables to 15. This will make the data frame more manageable.
2. A manual Feature Reduction process was used to build models by dropping variables with p-values> 0.05.
3. A total of 2 models were built before reaching the final Model 3 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.
4. log_m_2 was selected as a final model with 13 variables, we used it for predicting train and test sets.

Step 5: Model Evaluation:

1. A confusion matrix was made, and a cut-off point of 0.35 was selected based on the accuracy, sensitivity, and specificity plot. This cut-off gave accuracy, specificity, and recall all around 81%. Whereas the precision-recall view gave fewer performance metrics around 72%.
2. As to solving business problems, the CEO asked to boost the conversion rate to 80%, but metrics dropped when we took a precision-recall view. So, we will choose a sensitivity-specificity view for our optimal cut-off for final predictions.
3. The lead score was assigned to train data using 0.35 as a cut-off.

Step 6: Making Predictions on Test Data:

1. Scaling test dataset and predicting using the final model.
2. Evaluation metrics for train and testing are very close to around 81%.
3. A lead score was assigned.
4. We have determined the following features that have positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.
    a. Total Time Spent on Website
    b. Lead Origin in Lead Add Form
    c. Current Occupation in Working Professional
    d. Lead Source in Welingak Website
    e. Last Activity in SMS Sent
    f. Last Activity in Others
    g. Lead Source in Olark Chat
    h. The last Activity in the Email Opened
    i. Total Visits

5. We have also identified features with negative coefficients that may indicate potential areas for improvement. These include:
    a. Specialization in Hospitality
    b. Management Do Not Email
    c. Lead Origin in Landing Page Submission
    d. Specialization in Others

Recommendation

To increase our Lead Conversion Rates:

1. Focus on features with positive coefficients for targeted marketing strategies
2. The sales team should prioritize calling leads who have spent a significant amount of time on the website, as Total Time Spent on the Website is a good indicator of interest in X Education's services, with a coefficient of 4.416148.
3. Engage working professionals with tailored messaging. Working professionals to be aggressively targeted as they have a high conversion rate and will have better financial situations to pay higher fees too.
4. We should focus on more budget/spending on Welingak Website in terms of advertising, etc. to attract more leads.
5. The coefficients for Last Activity in SMS Sent and Last Activity in Email Opened are 2.140983 and 0.975803 respectively, indicating that leads who have been sent SMS messages or have opened emails are more likely to convert.
6. Optimize communication channels based on lead engagement impact.

To identify areas of improvement:

1. Analyse negative coefficients in specialization offerings.
2. Review the landing page submission process for areas of improvement.