LEAD SCORING CASE STUDY SUMMARY

Problem Statement:

X Education is an online education company that offers courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once people land on the website, they may browse the courses, fill out a form for a course, or watch some videos. When people provide their email address or phone number, they are classified as a lead. The leads are then contacted by the sales team to convert them into customers. The typical lead conversion rate at X Education is around 30%.

X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. A model is required to be built wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution:

Step 1: Importing and Understanding Data.

- Importing the Data
- Checking first 5 rows to see the data.
- Checking Non-Null values and data types of each column.
- Checking the quantitative spread of the data.
- Checking the shape (No of rows and Columns in the data).

Step 2: Data Cleaning

- Checking for duplicates if any
- Checking for null values and imputing them with appropriate methods
    - We dropped columns with more than 40% null values.
    - We used mode imputation for categorical columns, if there is skewness in the data.
    - We used median imputation for numerical columns, if there is skewness in the data.
- Detecting outliers and Handling them
- Drop columns that are not useful for the analysis.

Step 3: Data Visualization

- Performed Univariate analysis on categorical column to see the count of each value in the each column.
- Performed bivariate analysis on categorical columns to see how they vary W.R.T Converted column.
- Performed Univariate analysis on numerical columns by plotting count plot.
- Performed bivariate analysis on numerical columns with Converted column to see how the leads are related to these columns.
- In this step we also plotted the correlation matrix to identify the columns which are correlated.

Step 4: Data Preparation for Model Building, Test-Train Split and Feature Scaling.

- Converted all binary variables to 0 and 1.
- Created dummy variables for all remaining categorical variables as logistic regression needs the input as numerical values.
- Concated the dummy variables to the cleaned data set and dropped the initial columns.
- Rescaled all the required columns using Standard Scaler.

Step 5: Model Building

- Create a Logistic regression model using the prepared data.
- Use RFE to select the best 15 variables.
- Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05.
- Total 3 models were built with the 3rd model having P-Value < 0.05 and VIF < 5.
- Created a confusion matrix and find the overall accuracy with a cut-off value of 0.5.

Step 6: Model Evaluation using different matrices.

- ROC curve was plotted for the features and the curve came out be pretty decent with an area coverage of 89% which further solidified the of the model.
- Plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The intersecting point of the graphs was considered as the optimal probability cutoff point. The cutoff point was found out to be 0.35

- Based on the new value we could observe that close to 80% values were rightly predicted by the model. We could also observe the new values of the 'accuracy=80.76%, 'sensitivity=80.55%', 'specificity=80.88%'.
- Found out the Precision and Recall metrics and the values came out to be 78.41% and 68.33% respectively on the train data set.
- Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.4.

Step 6: Making Predictions on Test Set.

- After finalizing the optimum cutoff and calculating the metrics on train set, we predicted the data on test data set. Below are the observations:

Train Data:

- Accuracy: 80.76%
- Sensitivity: 80.55%
- Specificity: 80.88%

Test Data:

- Accuracy: 81.93%
- Sensitivity: 80.55%
- Specificity: 80.88%

Step 7: Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 82%, 81% and 82% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 78%
- There are 511 Hot Leads. They should be targeted as they have a high chance of getting converted
- Hence overall this model seems to be good.

Step 8: Recommendations

- As per the problem statement, increasing lead conversion is crucial for

the growth and success of X Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.

- We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.
    - Lead Origin_Lead Add Form: 3.49
    - Lead Source_Welingak Website: 2.49
    - Current_occupation_Working Professional: 2.31
- We have also identified features with negative coefficients that may indicate potential areas for improvement. These include:
    - Last Activity_Olark Chat Conversation : -1.52
    - Last Activity_Email Bounced : -1.42