Summary

The model building and prediction is being done for a EdTech company X Education. The main aim of the model building is to find potential leads ('Hot Leads') or customers who will bring in revenue and in turn help in proliferating the business.

Below is the stepwise approach used for the analysis:

Step-1: Reading and Understanding Data:

- In this step, first we import the necessary library modules that are required for the analysis.
- Next part of this step involves understanding the structure of the data. So, for that data quality checks are done and data dictionary is referred.
- Upon inspection of the data, we found 36 features with 1 target variable. The data consists of 30 categorical features and rest 6 are numerical features.

Step-2: Data Pre-processing

- First inspection is done on the categorical columns and we found many entries as 'Select'. These entries were then addressed as missing values and replaced by 'NaN'.
- Then we did a through inspection of all the missing values and dropped the values at a threshold of 70% for columns and 45% for rows and rest were imputed by mean and median.
- Then we dealt with the duplicate values and merged such entries into same columns.
- Outliers were also identified and were dropped wherever large variations were found.

Step-3: Exploratory Data Analysis

In this step Univariate, Bivariate and Multivariate analysis were done and insights were collected upon the data.

Step-4: Dealing with Categorical Variables/ Creation of Dummy Variables

In this step dummy variables are created for the identified categorical columns. Also, another dataframe with the cleaned data is now introduced where 98.2% of the original data is retained.

Step-5: Splitting the Data into Training and Testing Sets

In this step the cleaned data is split into a ratio of 70% to 30% for training and testing.

Step-6: Rescaling the features

Here the training data and testing data are scaled using the standard scaler.

Step-7: Finalising with the training model

Here various models are built and checked which is best for the analysis. The models are evaluated with the help of statsmodels library. Finally, the best model is selected for further analysis.

Step-8: Predicting Values based on final model

Here predictors are evaluated based upon the final model.

Positive Predictors for HOT LEADS: "Closed by Horizzon", "Lost to EINS", "Will revert after reading the email", "Welingak Website, "Working Professional" or "Unemployed"

This summary report is made by Sudeshna Mahapatra and Dip Ghosh

Negative Predictors for HOT LEADS: "Already a student", "switched off", "Not doing further education", "Diploma holder (Not Eligible)", "Ringing", "Interested in other courses", "Interested in full time MBA", "Worst"

Step-9: Metrics Evaluation

Here various metrics using confusion matrix are derived where accuracy was found to be 91.7%.

Step-10: VIF Evaluation

This step involves checking of VIF factor for finding the multicollinearity in the final model.

Step-11: Metrics beyond simply accuracy

In this step the other metrics i.e., Precision, Accuracy, Sensitivity, Speficity, Positive and negative leads are evaluated**.**

Step-12: Plotting the ROC Curve

This step involves plotting of roc curve for further analysis

Step-13: Finding Optimal Cutoff Point

This step is used to find the optimal cutoff point which is crossed by the three parameters Accuracy, Sensitivity and Specificity. In our model we found it to be 0.25.

Step-14: Making predictions on the test set

Final step involves testing the model against the test dataset and comparing it.

Observations:

After running the model on the Train Dataset these are the figures we obtain:

Accuracy: 91.7%

Sensitivity: 87.1%

Specificity: 94.6%

After running the model on the Test Dataset these are the figures we obtain:

Accuracy: 89.75%

Sensitivity: 91.1%

Specificity: 88.9%

Conclusion:

1. 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.

2. Accuracy, Sensitivity and Specificity values of test set are around 89.75, 91.1 and 88.9% which are approximately closer to the respective values calculated using trained set.

3. Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is above 80%.

4. Hence overall this model seems to be good.

This summary report is made by Sudeshna Mahapatra and Dip Ghosh

🏵❀Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:

1. Lead Origin_Lead Add Form
2. What is your current occupation_Working Professional
3. Total Time Spent on Website

Recommendations:

From our model, we can make following recommendations:

- The customer/leads who fills the form are the potential leads.
- We must majorly focus on working professionals.
- We must majorly focus on leads whose last activity is SMS sent or Email opened.
- It's always good to focus on customers, who have spent significant time on our website.
- It's better to focus least on customers to whom the sent mail is bounced back.
- If the lead source is referral, he/she may not be the potential lead.
- If the lead didn't fill specialization, he/she may not know what to study and are not right people to target. So, it's better to focus less on such cases.

This summary report is made by Sudeshna Mahapatra and Dip Ghosh