Summary

The model building and prediction is being done for company X Education and to find ways to convert potential users. We will further understand and validate the data to reach a conclusion to target the correct group and increase conversion rate. Let us discuss steps followed:

1. Data Cleaning

- Dropping columns that have only one unique values for all the leads.
- Handling 'Select' variable that is present in many categorical variables.
- Droping all the columns with more than 40% missing values
- Dropping columns with high data imbalance
- Using imputation technique for columns having less % of missing values
- Combining categories having low percentages into one single category.

2. EDA:

- Univariate Analysis - Numerical values:
- The max probability for TotalVisits is found to be around 15-20 visits. It increases initially but decreases further.
- The max probability for PageViewsPerVisit is found to be around to be 3-5 pages.
- The probability of time spent is found to be high for time between 0-300 seconds and Decreases further.
- 'TotalVisits' and 'Page Views per Visit' columns have outliers.
- Univariate & Bivariate Analysis – categorical Variables
- Max leads source is from Google and Direct Traffic.
- Maximum leads generated are unemployed
- People usually do not subscribe for a free copy of mastering the interview.
- If Lead source is Add Form, the ratio of lead conversion is very high.
- We need to target people via Emails and SMS as it is found that the probability of response in Case Converted leads is found to be higher.
- Google is found to be the important source for Lead Conversion
- It is clearly visible from the graph that we need to target the Unemployed and Working Professional to get a higher conversion rate. The ratio of conversion rate is higher than not Converted people for working professionals.
- We also worked on numerical variable outlier's treatment.

3. Data Preparation

Following steps are done as part of data preparation:

- Converted binary variable (Yes/No) to 1/0.
- Created dummies for categorical columns
- Performed train-test split: The split was done at 70% and 30% for train and test data Respectively.

- Performed Scaling: We have done the standard scaling on the variables ['Total Visits', 'Page Views per Visit', 'Total Time Spent on Website']
- Finding correlations between different variables using heat map.

4. Model Building

- RFE was used for feature selection.
- Then RFE was done to attain the top 15 relevant variables.
- Later the rest of the variables were removed manually depending on the VIF values and p-value.
- A confusion matrix was created, and overall accuracy was checked which came out to be 81.77%.

5. Model Evaluation

- **Sensitivity – Specificity**

If we go with Sensitivity- Specificity Evaluation. We will get:

- **On Training Data**

1) The optimum cut off value was found using ROC curve. The area under ROC curve was 0.88.

2) After Plotting we found that optimum cutoff was 0.35 which gave
   ACCURACY 78 %
   SENSITIVITY – 81 %
   SPECIFICITY – 76 %

- **Prediction on Test Data**

We got:-
ACCURACY 78 %
SENSITIVITY – 81 %
SPECIFICITY – 76 %

- **Precision – Recall**:

If we go with Precision – Recall Evaluation

1) With the cutoff of 0.42 we get the Precision & Recall of 72.92% & 80.05% Respectively.

2) So to increase the above percentage we need to change the cut off value. After Plotting we found the optimum cut off value of 0.4

3) If we go with 0.4 it'll decrease the sensitivity which is not desired in this problem Statement.

So if we go with Sensitivity-Specificity Evaluation the optimal cut off value would be 0.35

&

If we go with Precision – Recall Evaluation the optimal cut off value would be 0.44

CONCLUSION :-

• For Lead Conversion Rate to be high (more than 80% as per problem statement) we need sensitivity/Recall/TPR to be above 80%.

• If we want to further increase sensitivity we can decrease the cut off by 0.01-0.02. But we can't decrease further because specificity will decrease.

• There is not much difference between train and test data's performance metrics. This implies that our

Final model didn't over fit training data and is performing well. Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact less no of people and the conversion rate would be high.

From our model, we can conclude following points:

• The customer/leads who fills the form are the potential leads. (Add Form)

• We must majorly focus on working professionals.

• It's always good to focus on customers, who have spent significant time on our website.