Summary

Leads Scoring Case Study

Problem Statement:

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.

Solution Approach:

1.Data Cleaning:

- First step to clean the dataset we choose to remove the redundant variables/features
- The data set was partially clean except for a few null values and the option 'Select' has to replace with a null value since it did not give us much information.
- Dropped the high percentage of Null values more than 30%.
- Checked for number of unique Categories for all Categorical columns.
- From that Identified the Highly skewed columns and dropped them.
- Treated the missing values by imputing the favourable aggregate function like (Mean, Median, and Mode).

2. Exploratory Data Analysis:

- A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good but found the outliers
- Performed Univariate Analysis for both Numerical and Categorical variables.
- Performed Bivariate Analysis with respect to Target variable.
- Changed the multicategory labels into dummy variables and binary variables into '0' and '1'.

3. Dummy Variables:

- The dummy variables are created for all the categorical columns.
  1) For Binary variables, we convert them to 0/1 from Yes/no Column examples: 'Do Not Email'.
  2) For multivariable we create dummy variables and drop original variable. Column examples: Lead Origin', 'Lead Source', 'What is your current occupation', 'What matters most to you in choosing a course'.

4. Train-Test Split:

- The Spit was done at 70% and 30% for train and test the data, respectively.

5. Scaling:

- Now there are a few numeric variables present in the dataset has different scales. It is extremely important to rescale the variables so that they have a comparable scale. If we do not have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. So Standard Scaling is used to scale ['TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'] these columns.

6. Model Building:

- By using RFE with provided 20 variables. It gives top 20 relevant variables. Later the irrelevant features was removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value 0.06 were kept).
- Remove those high p values and then HIGH VIF values one by one for each iteration then perform modelling it until we get all features with low p and low VIF.

8.Model Evaluation on Train Set:

- Train model will used for predicting the values of output variables in set.
- We will plot RoC Curve and Precision recall trade off
- We will find optimal cut-off point (from curve we see cutoff as 0.28)
- Now we choose 0.28 as cut off to convert predicated values 0 and 1
- Check Accuracy, sensitivity, specificity, precision, and confusion matrix.
- Precision should be achieved around 80% and our model achieved 77.13%
-

9.Prediction on Test Data:

- Apply model on test data and check the confusion matrices to find the accuracy, sensitivity, specificity and Precision which came to be around 80%.

10.Observation:

- We can see that our model has accuracy up to 79% on train data set and 77% on test data set and attain 77% Precision. We can now say that the lead conversion will be around 80% as required by the CEO of the company.
- Column final_predicted , converted and Conversion_Probability with corresponding cut off used as 0.28. probability score less than 0.28 is cold lead meaning low chance of getting converted. Greater than is hot lead meaning higher chances of getting converted.

- Top features for good conversion rate:
  1) Lead Origin_Lead Add Form
  2) Current occupation_Working Professional
  3) Lead Source_Welingak Website

- Consider Lead Score > 75 as a high score. Leads with a score of 75-85 should be given to more experienced and senior sales employees