Lead Score Case Study Summary Report

I. Problem Statement

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- The company requires to build a model wherein it needs to assign a lead score 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%.

Objective:

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.

Solution Approach

II. Data Preparation:

- Dropped all the scoring variables which we cannot take into model building since these are a part of Lead Profiling activity – Tags, Lead Profile, Lead Quality and all Asymmetrique scores.
- Dropped all the Highly Skewed variables which are having more than 90% entries related to one category.
- *Page Views Per Visit, Total Visits & Total Time Spent on Website* are only the continuous columns after dropping unnecessary columns. The first two had missing values which are dropped as they are very less in numbers & outliers are treated with mode value.
- In all the below features, there are certain categories which are very less in count. All those are grouped into one single category 'Others'.
    - Last Activity
    - Lead Source
    - Lead Source
    - Specialization
    - City
- In some of the above features 'Select' category is been converted as missing values.
- Below are the final set of features that will be used for modelling. All the categorical variables are converted to binary variables by creating respective dummy variables.

Binary Features	Continuous Features	Categorical Features
Converted	TotalVisits	Lead Origin
Total Time Spent on Website	Page Views Per Visit	Lead Source
A free copy of Mastering The Interview	Lead Number	Last Activity
		Specialization
		What is your current occupation
		City

- The data imbalance ration of about 38% is identified on the ‘Converted’ column.

III. EDA:

- With these bivariate analysis of categorical variables, the following observations are evident:
  - **SMS Sent** activity is potential lead candidate
  - **Lead Add Form origin** is as well a potential lead candidate
  - **Olark Chat source** is not a reliable lead candidate.
- When the bivariate analysis performed on the continuous features, it is observed that the Total Time Spent on Website influences the lead conversion.

IV. Model Building:

- To begin with, identified 20 variables with the help of RFE method.
- After about 9 iterations of eliminating variables with high p-value and high VIF, a final model is achieved with 12 variables and the variables are listed in the order of highest coefficients.

Feature	Coefficient
Lead Origin_Lead Add Form	4.16
Last Activity_SMS Sent	2.22
Lead Source_Welingak Website	1.69
Last Activity_Other	1.62
Lead Origin_Lead Import	1.52
Lead Source_Olark Chat	1.40
Last Activity_Email Opened	1.17

Total Time Spent on Website	1.15
Last Activity_Website Activity	0.54
Specialization_Travel and Tourism	-0.43
Last Activity_Olark Chat Conversation	-0.40
Specialization_Finance Management	-0.38

V. Identify the Probability Cut-off

- With the ROC area of 0.86, the model seems to be a good one.
- The Precision_Recall_Curve and Sensitivity_Specificity tradeoff curves are plotted to identify the optimal cut-off probability so as to keep all the metrics good.
- 0.41 is found to be the optimal cut-off probability from Precision-Recall Trade-off and 0.35 from Sensitivity-Specificity trade-off.
- By the objective of the case study, we need to identify about 80% potential lead candidates. Hence, we should target high Sensitivity and should be around 80%. We will stick to the probability cut-off of 0.35.

VI. Predictions & Metrics on Test Data

- The final model is applied on the test data with a probability cut-off of 0.35 and the metrics seems to good, not much of drop compared to train data.
- The probability is also converted as lead score ranging from 0-100 for all the leads of train and test data.

VII. Final Metrics

- Train Set: ***Accuracy: 78.62%, Sensitivity: 78.99, Precision: 69.60***
- Test Set: ***Accuracy: 78.63%, Sensitivity: 77.76, Precision: 67.99***