Lead Scoring Case Study Summary Report

Objective

An education company named X Education needs a strategy to achieve higher lead conversion rate and use the manpower efficiently. They want to minimise the rate of useless calls made to the individuals. To analyse the data provided we followed the Logistic regression algorithm for classifying the leads as Converting (yes/no).

Following Data Analysis steps used :

Data Cleaning

- We first removed redundant columns such as “Prospect ID”, “Lead Number”, “Country”, “City” which were not going to help in predicting the conversion rate
- For categorical columns we merged the duplicate values. For “Lead Source” column data had “Google**”** & “google**”** as separate entities.
- 'Specialization', 'How did you hear about X Education', 'Lead Profile' columns were having “**Select**” as data in it. We replaced it with null.
- Null data analysis is performed and we removed columns which were having more than 30% null values. As they were not going to help in reaching any conclusion, we removed those columns.
- For remaining missing values, we imputed them with median & mode values as per the column’s use.

EDA

- **Univariate Analysis** was done for numeric, binary & categorical columns
- **Bivariate Analysis** was done to check the impact of the column with converted column.
- **Multivariate analysis** is performed by checking heatmap of the columns using Pearson Correlation.

Data Transformation

- For binary values, we replaced Yes with 1 & No with 0.
- Dummy variables are created for categorical columns such as Lead Source, Lead Origin, Last Activity, etc.
- Outlier Detection is performed for “TotalVisits” & “Page views per visit”. And bins were created to handle the outliers.

Data Preparation

- Data set was divided into Train & Test dataset
- We performed Standard Scaling on the numerical continuous columns to bring them to the same range with mean 0 and standard deviation of 1.
- Plotted the heatmap for finding the corelation among the columns and some high correlation columns were dropped.

Model Building

- We have built our model using Logistic Regression RFE with default 19 columns selected for feature selection and checked their p-values & VIF scores.
- We performed subsequent iterations and removed the columns which were having higher p-values and models were prepared. This process was iterated till we receive all columns with 0 p-value & VIF less than 5. Thus we have successfully removed all multicollinearity and have also removed insignificant columns from the model.
- In order to convert the probabilities returned by this model into the binary values, we need to find the optimum threshold instead of using the default threshold.
- For this we perform iterations on our model for several threshold values from 0.1 to 0.9. Then we compare the accuracy , sensitivity & specificity for each threshold and plot them. Visually from the plot we find that the converging point of the three scores is at 0.4. For the final model we use this as our optimal probability cut-off.
- We predict our binary values for the Converted column for the train dataset using this threshold. We evaluate the predicted values by finding the accuracy, sensitivity, specificity scores and plotting the Precision-Recall Trade-off. The model generated Recall value was 79% and Precision value was 73% for train dataset.
- Similarly, we predict our binary values for the Converted column for the test dataset using same threshold. We evaluate the predicted values by finding the accuracy, sensitivity, specificity scores and plotting the Precision-Recall Trade-off. The model generated Recall value as 75% and Precision value as 78% for test dataset.
- Using the generated probabilities, we calculated the lead score for test dataset, where anything above 40 was a good lead.

Conclusion

Top features for higher lead conversion rate :

- **Total Time Spent on Website**
- **Lead Origin_Lead Add Form**
- **What is your current occupation_Working Professional**

Leads with score above 80 can be called as Hot leads