LEAD SCORE CASE STUDY SUMMARY

- G.P. Shiva Prasad
- Chennibabu Dogiparthi

Problem Statement

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned to each of the leads such that thecustomers 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%.

Solution Summary

1.Reading and Cleaning Data

- There are no duplicate entries in the dataframe

- There are some columns with "Select" as a value which is the default for those columns. Hence, should be treated as "null" values. So, replaced "Select" value with "null" value

- There are some columns with null percentage

- Dropped the columns having null percentage more than 60%

- Imputed the null values with the mode value of those respective columns for some columns having significant null values and are categorical

- For Column "Specialization", the null values are replaced by string "Others" because there is no obvious entry with high occurrences and also with a understanding that null values are there because the specialization which they want to choose is not present. Thus, replacing with "Others".

- For numerical columns, null value is imputed with median value because of outliers presence

- Dropped the rows having null values as the remaining columns' null percentage is less than 1.5%

- There are some numerical columns with outliers

- The outliers are capped with 95 percentile or 5 percentile depending on the range of the value, i.e. highest value is capped using 95 percentile value and lowest value is capped using 5 percentile

- Dropped 7 columns because they are score variables. Thus, these 7 variables are not useful in prediction because of their dependence on Target/dependent variable

- Dropped ID columns as well because they don't have any influence on dependent variable

2. Data Analysis

- Did univariate analysis with respect to dependent variable (Converted) and dropped the columns which have only one value or highly skewed. These columns don’t help in analysis. So, dropped them.

- Checked for correlation between the numerical columns

- Some values of “Last Activity” are covered as values under “Last Notable Activity”. So, dropped “Last Notable Activity”

3.Data Preparation

- Converted binary variables(having only two unique values(Yes/No))

  - Replaced “Yes” with 1 and “No” with 0 in those binary columns

- Created dummy variables for categorical columns and dropping the first column

- Performed Train-Test set split with 80% train size

- Performed Feature Scaling using Standardization

4.Model Building

- Feature Selection is done in Mixed approach. First used RFE(Recursive Feature Elimination) and then basing on the P-Value and VIF, tuned the model and finally came up with the model with 12 predictors for the dependent variable

- Created a dataframe containing actual “Converted”(dependent variable) values and the predicted variables. Initially, the threshold for probability is kept at “0.5”. The metrics are as below

  - Accuracy = 0.8049318087890894

  - Sensitivity = 0.6607402031930334

  - Specificity = 0.8931823228958472

- Plotted ROC Curve to determine how good the model is and got the area under curve to be 0.88 with is in acceptable range. Thus, the model is good

- Plotted graph to know the optimal cut-off threshold where we get balanced sensitivity and specificity and came up with the threshold of value “0.34”

- Then assigned Lead Score to the train dataset based on the converted probability.

  - Accuracy = 0.8065849290535887

  - Sensitivity = 0.8156748911465893

  - Specificity = 0.8010215411947591

  - Precision = 0.7910512597741095

  - Recall = 0.6607402031930334

  - F1 Score = 0.7200474495848161

- Made Predictions on the Test Dataset with threshold cutoff value of “0.34”

  - Accuracy = 0.8

- Sensitivity = 0.8156748911465893

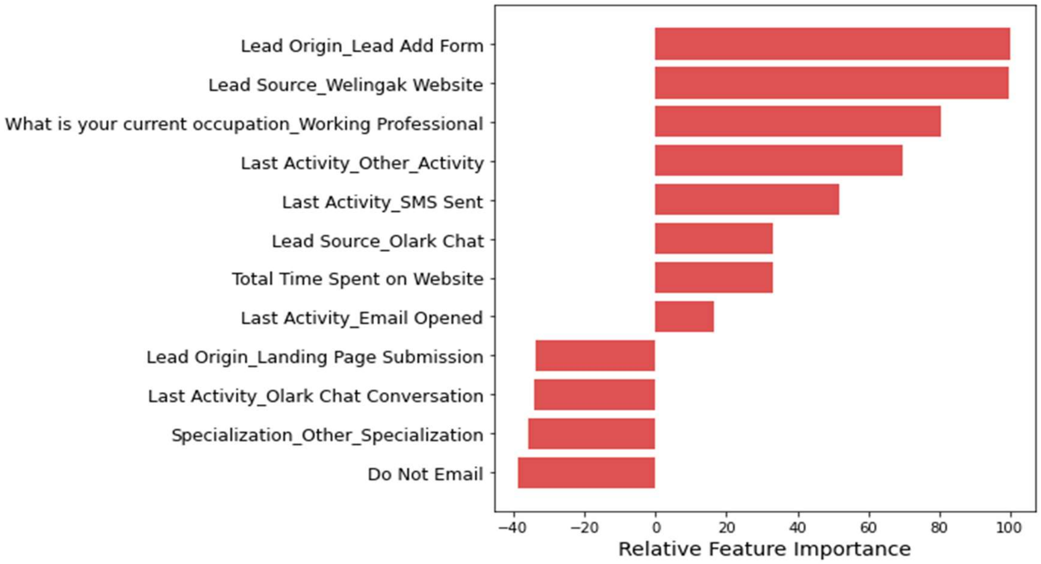
- Specificity = 0.8010215411947591

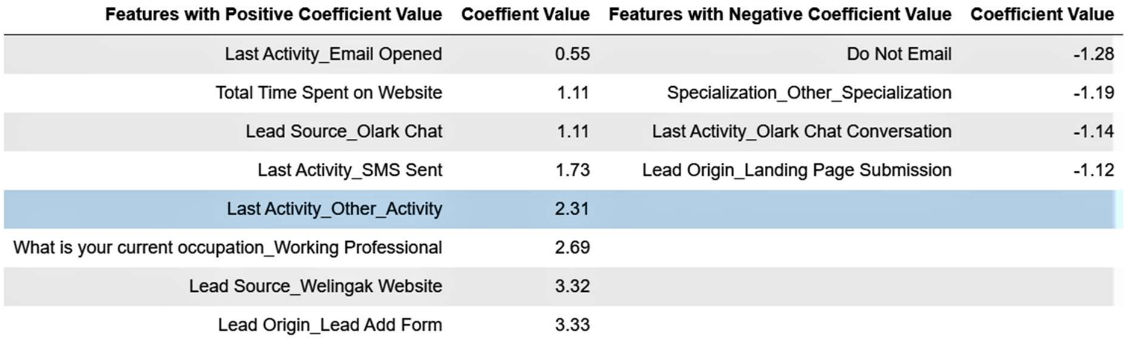
- Precision = 0.7158469945355191

- Recall = 0.7717231222385862

- F1 Score = 0.7427356484762581

Finally built a model with 12 predictors with various Importance. They are





Features with Positive Coefficient Value	Coeffient Value	Features with Negative Coefficient Value	Coefficient Value
Last Activity_Email Opened	0.55	Do Not Email	-1.28
Total Time Spent on Website	1.11	Specialization_Other_Specialization	-1.19
Lead Source_Olark Chat	1.11	Last Activity_Olark Chat Conversation	-1.14
Last Activity_SMS Sent	1.73	Lead Origin_Landing Page Submission	-1.12
Last Activity_Other_Activity	2.31		
What is your current occupation_Working Professional	2.69		
Lead Source_Welingak Website	3.32		
Lead Origin_Lead Add Form	3.33		

Above are the 12 predictors with their coefficient values. Positive Coefficient means that it affects the dependent variable in the positive way(if predictor changes, the probability of lead conversion increases) and vice versa.

RECOMMENDATION:

- Website Should be made more intuitive and engaging with more information added because it's a strong predictor with a strong positive coefficient.
- “Lead Add Form” in “Lead Origin” should be given high importance.
- “Welingak Website” in Lead Source is a reliable one in attracting new leads.