LEAD SCORE ASSIGNMENT SUMMARY REPORT

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. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. 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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads in order to get a higher lead conversion.

The company requires us to build a model wherein we need 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. 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:

The effective way to attend this problem is by starting with the Hot Leads i.e., the leads that have the higher probability of getting converted. This will ensure we have a higher conversion rate in a less amount of time. We can focus more on nurturing the Hot leads to get in in converted and then focus on the low score (cold leads) if time permits.

We can determine the Hot leads and the cold leads using the Logistic Regression model. We will use the meta data provided to us and use it to build a logistic regression model and assign score to each lead.

The following are the steps followed in the process:

- Data Analysis
- Data Visualization
- Data Preparation
- Model Building and Evaluation

Data Analysis:

It involves importing the data to your workspace and inspecting the dataframe for the record counts, any null values, missing values, duplicate values.

There are few columns for which the value is populated as Select. We have converted the Select as Null for analysis purpose.

We have dropped the attributes that have the missing values >45 %.

We have imputed the columns having null values with median for the numerical columns and mode for Categorical column

Data Visualization:

We have plotted the distplot of the numerical variables to observe their spread.

Also, we have plotted the heatmap of the numerical variables to find if any columns have high correlation values. So that we can proceed removing one of the columns as they are insignificant during the model building.

Also, we have plotted the box plots of numerical variables to observe the outliers. In case of outliers, we cap the outlier value at 99 percentile or leave it as it is if it does not affect our analysis.

Also, we proceed with plotting the bivariate boxplots for numerical variable wrt target column to observe how these numerical variables affect the conversion.

We have plotted the count plot of Categorical columns wrt to Target column so as to observe how the categorical column affect the conversion.

Dropped the records which have null values now as it is insignificant compared to the total record count and will not affect our model.

Data Preparation:

Since Logistic regression uses numerical data, we will convert the Categorical columns using below technique:

- Converted the columns having the binary variables (Yes/No) to 0/1
- Created dummy variables for the Categorical columns using One hot Encoding(get_dummies)

We then proceeded with separating the independent variables and dependent variable. And also split the data into Training set (on which the model is build) and Test set (to make prediction).

We have then performed the scaling of numerical variable using StandardScaler to bring all features in the same standing, we need to do scaling so that one significant number doesn't impact the model just because of their large magnitude.

Model Building and Evaluation:

We have built a Binomial Logistic regression model using Generalized Linear Model Regression.

Then we have used the automatic feature selection Technique (RFE) to select the significant variables.

Then we manually proceed by removing the variables that have high p-value (>0.05) as they are insignificant during the model building. And also remove the variables with high VIF (>5) as they cause high multi collinearity between the variables. And again, we build a model.

All the p values of the predictor variables are <0.05 (predictors seems to be significant) and the corresponding VIF values (VERY LOW Multicollinearity between the predictors) are also less (< 5) after dropping the constant.

We evaluate the performance of the model based on:

- Accuracy
- Specificity
- Sensitivity
- ROC curve

We then use the below technique to find the optimal cut-off value

- Plot accuracy, sensitivity and specificity

Tradeoff between sensitivity and specificity is necessary to find the optimal cut off. And based on this cut off we find predicted value.

We will assign a lead score to each lead using probability predicted by model (Lead Score=Predicted probability *100).

We used the model build on Training set to predict the Target values of the Test set. And also evaluated the same using the metrics.

Conclusion:

The model we have built has the following metrics for the Training set and Test set:

Train Data:

- Accuracy: 92.49%
- Sensitivity: 91.71%
- Specificity: 92.96%

Test Data:

- Accuracy: 91.94%
- Sensitivity: 91.98%
- Specificity: 91.91%

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model.

Recommendations:

After completing the Model building the following are my recommendations:

From the Coeff values of the model it is clearly evident that the following variables should be focussed the most in order to increase the probability of lead conversion

- a) Tags Closed by Horizon
- b) Tags Lost to EINS
- c) Tags Will revert after reading the email

And the 3 variables that influence the lead conversion are-Tags, Lead Source, Last Activity. Then followed by Lead Origin, Last Notable Activity, Specialization and Total Time Spent on Website

Apart from this from the EDA we conclude that:

- The customer who has wished to not receive any emails tend to be not converted. So, it’s better to target only people who opt to receive emails from the organization
- The customer had seen the ad in any of the listed items: ‘Search', 'Newspaper Articles’ Education Forums', 'Newspaper', 'Digital Advertisement' are very low.
- The converted count is high for Lead Source of Google and Direct Traffic. Hence it is suggested the company promotes their ads on the Google page.
- The Landing Page Submission has the high conversion value. So, when the people have submitted the forms their chance of converting is high. Hence, we can target those clients through calls, emails and SMS.
- Also, when the customer sent us SMS or when they have the email is opened their chance of turning to lead is high.
- The unemployed, working professionals, student have higher chance of converting to leads. Hence, we have to target these categories of clients.