Lead Scoring Case Study Summary

Objective

The company requires you to build a model wherein you 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%.

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 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%.

Approach Used

As a start of the case study, firstly we needed to understand the data we had. For that we inspected the whole data set. The number of rows and columns in the dataframe were identified to know the scale of the set we have. The number of null values in each of the columns were identified as part of the data inspection. Also a check was done to remove the calls with only 1 unique value as it won't impact our analysis and model at all. Then few fields like `Asymmetrique Activity Score' & 'Asymmetrique Profile Score'` were dropped, as they can be represented through their index value. Then the imputing of null values was done with the mode value for categorical variables. For some of the variables the null values were imputed with the 'Unknown' field and later while creating dummies the Unknown field column was removed. Various variables were taken care of during this phase of the building and analysis using the plots and other analysis.

The next phase was the EDA(Exploratory Data Analysis) of the data we had in our hand. This phase helped us to know about the present conversion rate , the relation among data, what does our data represent in context to our objective and how can we use it to reach our objective. We figured out the current conversion rate to be 65%. Also the major conversion for lead source was from Google. Similarly Last Activity of SMS Sent had the most number of conversions. Also

the conversions were more for Unemployed people. We could see that 50% businessmen were conversions while housewives had 100% conversion as per the data.
X Education Forms and Newspaper Article were dropped since they had a single value.
Also we could infer that Conversion 0 or 1 has similar visualization irrespective of page views per visit & Total time spent on the website where as for total visits count is more for conversion 0. This phase gave us the insight of how to approach the dataset for our objective and what kind of variables we have and how many dummy variables we needed.
After the EDA, outlier treatment was done to remove the data that can impact our model. This treatment was done on the numeric variables. The variables identified were
`'TotalVisits','Total Time Spent on Website','Page Views Per Visit'`
The outlier treatment was done on these numeric variables. Outliers were identified using box plots and interquartile percentile ranges.

After the outlier treatment, the data was prepared. In this step the categorical variables were then converted to have dummy columns in the data set, with drop_first true to optimise the number of columns in the data set. But as earlier mentioned we had fields with value as Unknown, those columns were dropped from dummy columns list. This will help to identify the dependency of categorical variables in the model we will build.

Finally after all the EDA and data preparation, the data was split into test-train(70-30).
Once this was done, Feature scaling was needed to be done to bring all the numeric variables to a common scale, otherwise relatively higher scale variables will impact our model and thus the whole of our prediction.
After the above step, the model was built using GLM (Generalized Linear models) from statsmodels. After the model is built, we select features in the model using feature selection technique. For this we used RFE (recursive feature elimination). We took top 20 features and built and evaluated the model on those.
Once all the feature selection was done, the model was evaluated and VIFs values were computed to know the multicollinearity in the model to remove if any. The features having high values were removed. The model stats were looked into then to identify variables with higher p-value as they can lead to insignificant models and thus not reliable results. Those features were dropped one by one and the model building iteration was performed one by one.
The evaluation was done by keeping a threshold of churn probability as >= 0.5 and comparing with the original values of churn. The accuracy, sensitivity and specificity values were computed using the existing values of churn.
The VIF values were computed every time, also the confusion metrics were taken into account to see we are staying within limits while we drop the columns. Once all the values including accuracy, sensitivity and specificity were taken care of, we moved to plotting the ROC curve.
The ROC curve helped us to identify the right threshold value as we can see False Positive Rate and True Positive Rate.The area under the curve represents the aggregated performance value across all the thresholds. This comes out to be approximately 0.96, which is a good value.
The area under the curve of ROC is generally calculated using the integral calculus in the specified region. The accuracy, sensitivity and specificity values were plotted to identify the optimal threshold which comes out to be ~0.35 in our case. The precision and recall values

were again drawn to identify the trade offs. From that we could figure out an optimal threshold of 0.42, but we could get 80% lead conversion with 0.35 as well as we did earlier. So we went with that particular threshold value.
Finally the predictions were made after the model is built on the test set. We got the following results -

a. Accuracy - 89%
b. Sensitivity - 95%
c. Specificity - 85%

The F1 score was computed to be 0.8714.
After all this, the final features were identified that impact the lead conversion a lot. The following features were identified -

1. Tags_lost to eins
2. Tags_closed by horizzon
3. Tags_will revert after reading the email
4. What is your current occupation_working professional
5. What is your current occupation_unemployed
6. Last Activity_sms sent
7. Lead Quality_worst
8. Asymmetrique Activity Index_03.low
9. Tags_opp hangup
10. Tags_interested in full time mba