Lead Scoring Analysis

Problem statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

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%.

To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.

The company requires us 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%**.

Goal :

1. 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.
2. There are some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

Data Analysis:

1. There are no duplicates in the dataset
2. From the above analysis we understood that Prospect ID and lead Number potrays the same customer and they do not hold any relevance for further analyis hence they can be dropped
3. Through EDA we analysed India is the most occurring country in the Dataset, it might not be suitable in terms of analysis as it may lead to classification issue. Hence the 'Country' variable can be dropped.
4. Here we can see that Mumbai is indeed the most frequent occurring city in the Dataset.
5. Management is having the highest number of leads converted. This is important and cannot be removed.
6. Management has the highest number of leads as compared to other specializations
7. Here Working Professional can go for the course as the chances are high
8. Unemployed leads are the highest over here.
9. Better Career Prospects has higher number of leads and influences the column hence we can remove this variable.
10. Highest number of the leads are generated from Google and Direct Traffic.
11. While the least being the Live chat attribute through which leads can be generated

Leads can be maximized from References and other Welingak Website

12. API and Landing page is giving away the most leads for conversion.
13. Lead Add form has high number leads for conversion but the count of the leads is less.
14. Higher count of leads in Lead Add form can lead to higher conversion
15. Here we can see 'Do not Call' and 'Do not Email' has higher number of conversion as compared to said being 'Yes'
16. SMS Sent and Email opened is having higher number of leads as compared to others.
17. Modified has the least number of leads conversion ration

Model Preparation and Analysis:

1. Created dummy variables and removed original for analysis purpose.

2. Split the data to train and test set.

3. Scaling the data using standard scalar.

4. Differentiate RFE supported columns and non-supported columns.

5. Using Variance Inflation Factor to see any correlation between Variables.

6. Check for sensitivity and specificity.

7. Plotted ROC curve, found the optimal cut off point.

8. Predict using the test data set.

Output of the Data Modelling and Analysis:

After running the model on the Test Data these are the figures we obtain:

- Sensitivity : 91.98%
- Specificity : 93.26%
- Accuracy : 92.78%

Final Observation:

Now we will be comparing the values obtained from Train & Test Set:

Train Data:

- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%

Test Data:

- Sensitivity : 91.98%
- Specificity : 93.26%
- Accuracy : 92.78%