Summary

Problem Description:

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%.

X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. A model is required to be built wherein 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 :

1. Data Cleaning:

The first step is to clean the data and remove unwanted variables.After removing we found that columns are having ‘Select’ as their labels and we need to replace those with the null values. Removed columns having more than 35% of null values. For remaining missing values, we have replaced the maximum number of occurrences of the column. We have some data with All Capital or All small values by replacing it with their correct format.

2. Data Transformation:

Changed the multicategory labels into binary variables in the form of ‘0’and ‘1’. Created dummy variables for some variables. Checked the outliers and removed some of the numbers using 0.99-0.1% analysis.

3. Data Preparation:

Splitting the dataset into train and test dataset. Scaled the dataset using the StandardScaler(). Plotted heatmap for finding the correlations and dropping them.

4. Model Building:

We build our model with the help of RFE with 19 variables. Checked the VIF Score for each variable, as all of the variables are having VIF Score < 5.0, we proceed to our next step. We then removed the insignificant variables using the P-Value Score. For our final model we checked the optimal probability cutoff by finding points and checking the accuracy, sensitivity and specificity. We found one convergent point and we chose that point for cutoff and predicted our final outcomes.

We checked the precision and recall with accuracy, sensitivity and specificity for our final model and the tradeoffs. Prediction was made now in the test set and predicted value was recoded. We did model evaluation on the test set like checking the accuracy, recall/sensitivity to find how the model is. We found the score of accuracy and sensitivity from our final test model is in acceptable range. We have given the lead score to the test dataset for indication that high lead scores are hot leads and low lead scores are not hot leads.