LEAD SCORING CASE STUDY SUMMARY

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.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. 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%.

Solution summary -

1) **Reading and understanding the data:**
   The first step to solving the problem is to read the data and understand its various aspects such as - how many columns are present, how each feature(numerical and categorical) impacts the solution, etc.
2) **Cleaning the data:**
   i) Dropping the columns which have more than 40% values missing
   ii) Creating new category NA for columns with missing values less than 40% and dropping rows whose features are less than 2% missing data
   iii) Identified few features that were inputs provided by the sales team, rather than prospect data, and dropped those features
   iv) Performed analysis for outliers to identify if any numerical columns had a large amount of outliers
   v) Dropped features than had data imbalance
3) **Data transformation:**
   Modifying the binary variables to have values ‘0’ and ‘1’ and creating dummy variables for the categorical columns
4) **Train-Test split:**
   Performed a random split of the data with a proportion on 70-30
5) **Feature scaling:**
   i) Scaled the train data using the Standard Scaler
   ii) Plotted heat map of the categorical variables to drop those that have a high correlation
6) **Model building:**
   i) Used RFE to generate a base model with 18 best features
   ii) Dropped columns with high p-values and VIF(>5) in the subsequent models till the a model with all features with low p-values and VIF are obtained
   iii) Evaluated final model and got a 79% accuracy
   iv) Performed evaluation on metrics - Sensitivity(64%), Specificity(88.5%), Recall(63.9%), Precision(77.7%)

v) To improve model used different probability cutoffs to find the optimal threshold and obtained it as 0.3

7) **Deploying model on test data:**
   i) The test data showed a conversion rate of 81.7%
   ii) Test data metrics showed the following values – Sensitivity (81.7%), Specificity(73.5%), Recall(81.7%), Precision(63.8%)

Solution summary -

The following conclusion can be derived from the model that we have created.

- The sales team can concentrate on the Unemployed candidates and the working professionals as they have the highest conversion rates among all other occupations
- Those who have spent more time on the website are more likely to be converted as they would be intrigued to find about the courses. The marketing team could make their website more accessible and interesting and more informative.
- Most of the leads are from the city of Mumbai. Metropolitan cities have more probability to become a lead.