**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%.

 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. A typical lead conversion process can be represented using the following funnel:

**Lead Conversion Process - Demonstrated as a funnel**

As you can see, 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 well (i.e. educating the leads about the product, constantly communicating etc. ) in order to get a higher lead conversion.

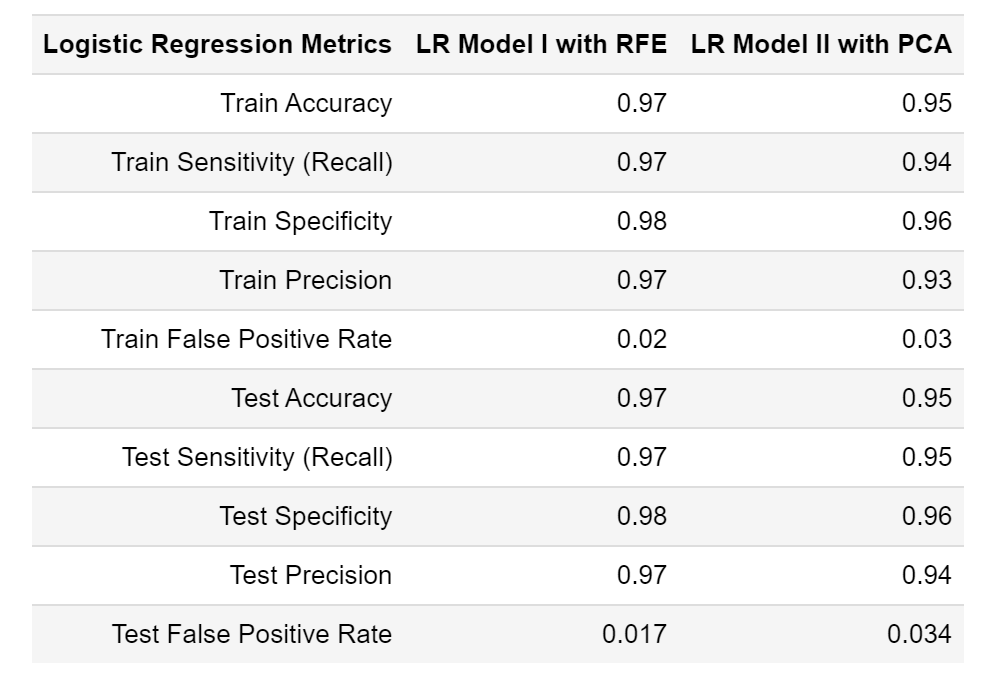
X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have 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%.

**Summary of Logistic Regression – Lead Scoring Case Study**

The given data has 9000+ data points with nearly 37 columns. Of total there are 8 numerical columns and 29 categorical columns. The target column is **Conversion** and have 2 binary outputs. Since the data is labeled and having 2 possible outputs, we have used Logistical Regression Model in this case study. We have created 2 models of Logistical Regression model one using RFE and other with PCA. We have compared both models and then evaluated the better model based the Logistical Regression Model metrics.

Following are the detailed steps used in the assignment.

1. Import Python Libraries for data analysis and ML
2. Sourcing the Data
3. Read the data
4. Inspect the data
5. Visualizing and Exploring the data
6. Perform Univariate Analysis
7. Perform Segmented Univariate Analysis wrt Target variable
8. Perform Bivariate Analysis
9. Inspect and Clean the Data
10. During the initial inspection, we observe that the data is imbalanced, high number of missing values across several columns.
11. Hence, we thought it makes sense to explore the data and derive imputation strategy.
12. Outlier Detection & Treatment
13. We had observed that there were couple of columns with outliers. Based on business judgement, we have treated outliers.
14. Impute Missing values
15. There were many columns with missing values
16. Apart from missing values, as given in the problem statement, we have seen that users had not selected any option and the given data has “SELECT” which should be considered as null. This was imputed with NAN.
17. Our strategy was to drop the columns with 40% of missing data. For the columns having missing data below 40%, we have inspected each column individually to see if we can impute based on mode or with any other strategy. Couple of columns have been dropped and couple of columns have been imputed.
18. Other than that, we are dropping all the columns for which there is no or very low variance in data values.
19. We are also dropping the columns for which we do not see any major impact on the Conversion. For example, City, Specialization etc as identified above post EDA.
20. Preparing the data for modelling (train-test split, rescaling etc)
21. Create Dummy Variables
22. We have used binary encoding and pd.get\_dummies technique to create dummy variables
23. Scaling of the Data
24. We have used standard scaling technique across numerical variables to bring them to one scale
25. Perform Fit and Transform on Train data
26. Perform Transform on Test data
27. Analyzing Correlation
28. **Model I using RFE**
29. There were nearly 105 features to represent the data.
30. On training data, using RFE, we selected the 22 features
31. During this process, we have evaluated VIF to check the collinearity and significance to the model with p-value.
32. We have dropped the columns having p-values above 0.05 and VIF higher than 5.
33. After some iterations, we had final set of features where there was no collinearity and p-value being very low.
34. Now using Logistic Regression model we have predicted the probability of each record.
35. To derive the final cutoff we have derived the labels at different probability cutoffs. We have calculated Sensitivity, Specificity, Accuracy, Precision, FPR across each probability cutoff.
36. The above metrics, we have following plots
37. ROC Curve
38. Accuracy, Sensitivity and Specificity
39. Recall and Precision Curve
40. We have derived our final cut off based on the Recall and Precision Curve. As in the business requirement, we have decided that Recall and Precision curve is right metric to decide the cut off. Based on the convergence of Recall and Precision, we have decided the cutoff probability as **0.5** to derive our final model.
41. We ran our model with test data and derive the final probability predictions.
42. With **0.5** cutoff we have derived our evaluation metrics and lead scores. We took the probabilities \* 100 as measure of our lead scores.
43. With this model we got lead conversion rate of 97% (Precision of .97)
44. **Model II using PCA**
45. There were nearly 105 features to represent the data.
46. On training data, using PCA we have observed that nearly 22 features explain 95% variance in the data and 50 features explain 99% of the data.
47. Beyond 95% we see that each feature was adding less variance and we selected 95% variance to derive principal components.
48. We have derived PC’s for 95% variance both on training data and test data.
49. We have used Scikit learn LR model to make predictions on data.
50. To derive the final cutoff we have derived the labels at different cutoffs. We have calculated Sensitivity, Specificity, Accuracy, Precision, FPR across each probability cutoff.
51. The above metrics, we have following plots
52. ROC Curve
53. Accuracy, Sensitivity and Specificity
54. Recall and Precision Curve
55. We have derived our final cut off based on the Recall and Precision Curve. As in the business requirement, we have decided that Recall and Precision curve is right metric to decide the cut off. Based on the convergence of Recall and Precision, we have decided the cutoff probability as **0.42** to derive our final model.
56. We ran our model with test data and derive the final probability predictions.
57. With **0.42** cutoff we have derived our evaluation metrics.
58. With this model we got lead conversion rate of 93% (Precision of .93)
59. Model evaluation, Comparisons and Conclusions and final summary



* Precision is around 0.97 and Lead Conversion rate is around 97% both on train and test set with **Model 1**.
* Precision is around 0.93 and Lead Conversion rate is around 93% both on train and test set with **Model 2**.
* **All the given metrics are better with Model 1 and hence we recommend LR Model 1 using RFE for final analysis.**
* Based on Model 1, we have derived our lead scores.
* Following is the summary

