Summary

Problem statement

An education company named X Education sells online courses to industry professionals. The company wishes to find the "Hot leads".

X Education focuses on to select the most promising leads, 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 a higher lead score have a higher conversion chance and the customers with a 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

Understanding the Data

Reading and understanding the data. And analyzing their data description and information.

Data cleaning

- Columns with greater than 35% of null values has been dropped, which includes Imputing the missing values as and where required.
- The outliers were identified and removed using box plot.
- Numerical categorical data were imputed with mode and columns with only one unique response from customer were dropped.

Data Analysis

- Data imbalance checked with 38%.
- Performed univariate and bivariate analysis for categorical and numerical variables. Provided the valuable insight on effect on target variable.
- Maximum Leads are generated by Google and Direct Traffic.
  Leads from Reference and Welingak Website are high.
- Most of the data are from India, we cannot draw any inference from this.
- Maximum leads generated are unemployed and their conversion rate is more than 50%. Conversion rate of working professionals is very high.

- Maximum leads are generated having “Last activity” as “Email opened” but the conversion rate is not too good. “SMS sent” as the last activity has a high conversion rate.

Data Preparation

- Created dummy features (one-hot encoded) for categorical variables.
- Splitting Train & Test Sets.
- Feature Scaling using Standardization.
- Dropped few columns, they were highly correlated with each other.
- Low frequency columns are combined into one single column.
- Imputations are done where ever required in the columns.

Model Building

- Used RFE to reduce features from 22 to 15. This will make data frame more manageable.
- Manual Feature Reduction process was used to build models by dropping features with p – value.
- Iteratively, 5 models were built up achieving p-values <0.05 and VIF <5 and no sign of multicollineratity.
- Logm5 was selected as final model with 11 features; we used it for making prediction on train and test set.

Model Evaluation

- Confusion matrix was used to evaluate the model and cut off point of 0.3 was arrived based on accuracy, sensitivity and specificity plot.
- The Train Data: Accuracy: 77.05%, Sensitivity: 64.9, Specificity: 89.8% Precision: 66.5% , Recall: 88.2%.
- The Test Data: Accuracy: 77.52%, Sensitivity: 64.9%, Specificity: 89.8%, Precision: 66.4% , Recall: 83%.
- Based on the above recall and precision values we have concluded our fifth model as our final model.

Making predictions on test data

- Making Predictions on Test: Scaling and predicting using final model.
- Evaluation metrics for train & test are very close to around 83%. Lead score was assigned.
- The below are the 11 features were selected for the better conversion rate:

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2020	0.094	-12.723	0.000	-1.387	-1.017
Do Not Email	-0.3600	0.043	-8.348	0.000	-0.445	-0.276
Total Time Spent on Website	1.1023	0.038	28.710	0.000	1.027	1.178
Lead Origin_Lead Add Form	4.6119	0.523	8.816	0.000	3.587	5.637
Lead Source_Direct Traffic	-1.0496	0.107	-9.783	0.000	-1.260	-0.839
Lead Source_Google	-0.7804	0.102	-7.615	0.000	-0.981	-0.580
Lead Source_Organic Search	-0.8639	0.124	-6.987	0.000	-1.106	-0.622
Lead Source_Reference	-1.7425	0.564	-3.089	0.002	-2.848	-0.637
Lead Source_Referral Sites	-1.3749	0.336	-4.094	0.000	-2.033	-0.717
What is your current occupation_Student	1.1342	0.224	5.057	0.000	0.695	1.574
What is your current occupation_Unemployed	1.2613	0.082	15.384	0.000	1.101	1.422
What is your current occupation_Working Professional	3.7575	0.189	19.919	0.000	3.388	4.127