LEAD SCORING CASE STUDY

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. When these people fill up a form providing their email address or phone number, they are classified to be a lead. 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'. 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 . The company requires 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.

STEPS FOLLOWED FOR SOLVING THE PROBLEM

Data understanding and EDA

- Reading the dataset , Inspecting , cleaning the null values & handling outliers , EDA
- The data is converted to a clean format suitable for analysis in Python.

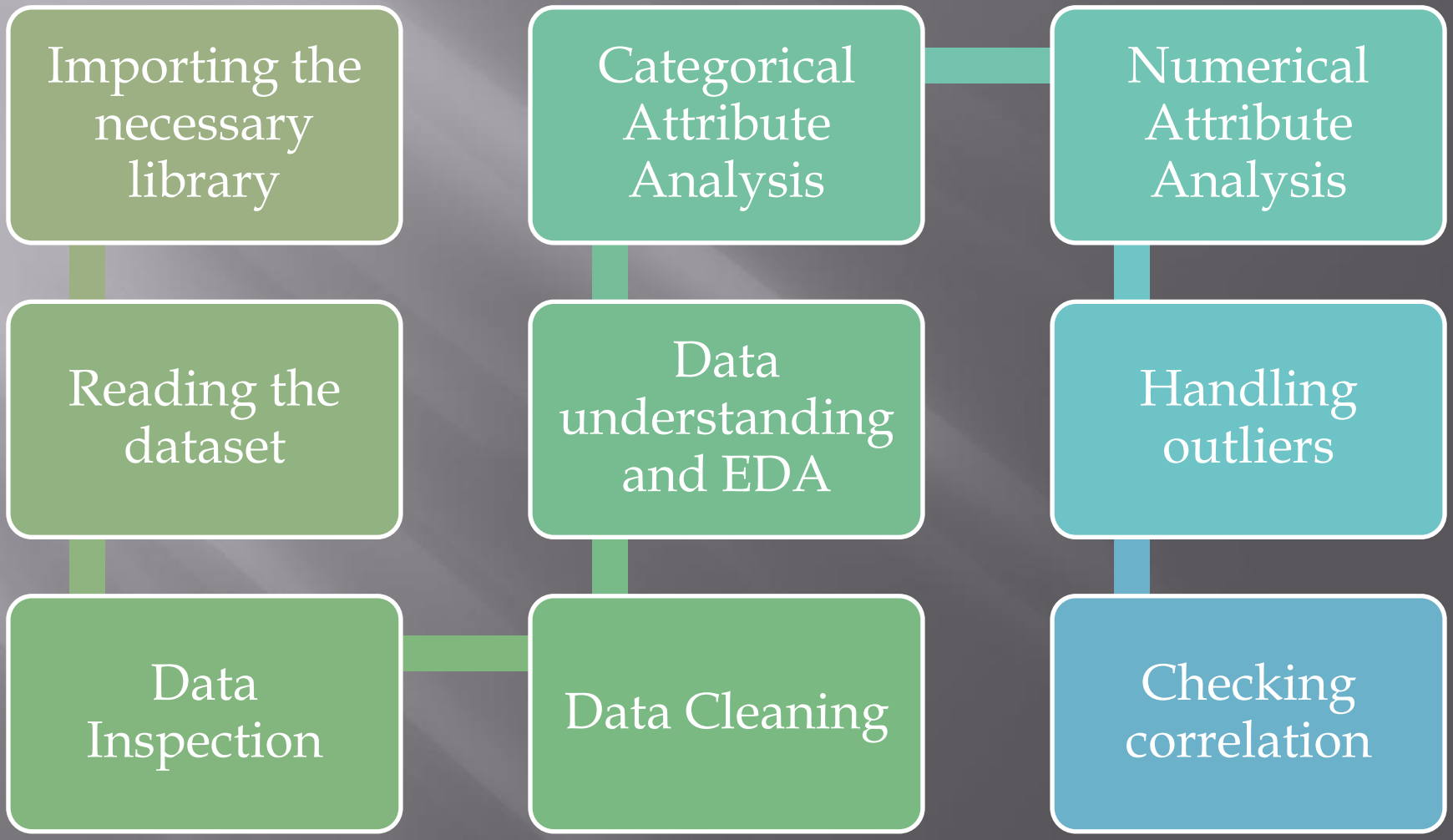
Data preparation

- Transformation(Dummy variables creation/label encoding)
- Train test split

Model building and evaluation

- Using Logistic regression A reasonable number of different models are attempted and the best one is chosen based on key performance metrics
- The results are at par with the best possible model on the dataset.

Data understanding and EDA



Data understanding and EDA

- Imported necessary library
- Read the data
- Found 9240 Rows and 37 columns
- There are no duplicates under Prospect ID and Lead Number and since they just indicate only the identification number, dropped these 2 columns
- There are some columns where the value has mentioned as 'select' it probably because the leads do not select any options we replaced the 'select' to null values .
- Dropped those columns whose null values percentage more than 45 % .
- Visualized the categorical variable through countplot
- Found some variable having same values more than 99% also some values equal to 100% , these are not useful so we dropped these columns .
- Also replaced some values of similar type but different name for better analysis

Data understanding and EDA

Numerical Attribute Analysis

Displayed correlation using Heatmap

Data Preparation

- Dummy Variable Creation
- Changing some yes/no values to binary values 1/0
- Dropping the original columns after dummy variable creation
- Splitting data and building logistic regression model
  - Used ‘train_test_split’ from ‘sklearn.model_selection’
  - Taking 70/30 cutoff for spliting the data
- Scaling numeric columns by ‘StandardScaler’

Model building and evaluation

Model Building using Stats Model & RFE

Build total 3 Model

Finalize Model 3 because

- All the p values are less
- The VIF values seem to be in order

Checked the overall accuracy which is 0.92

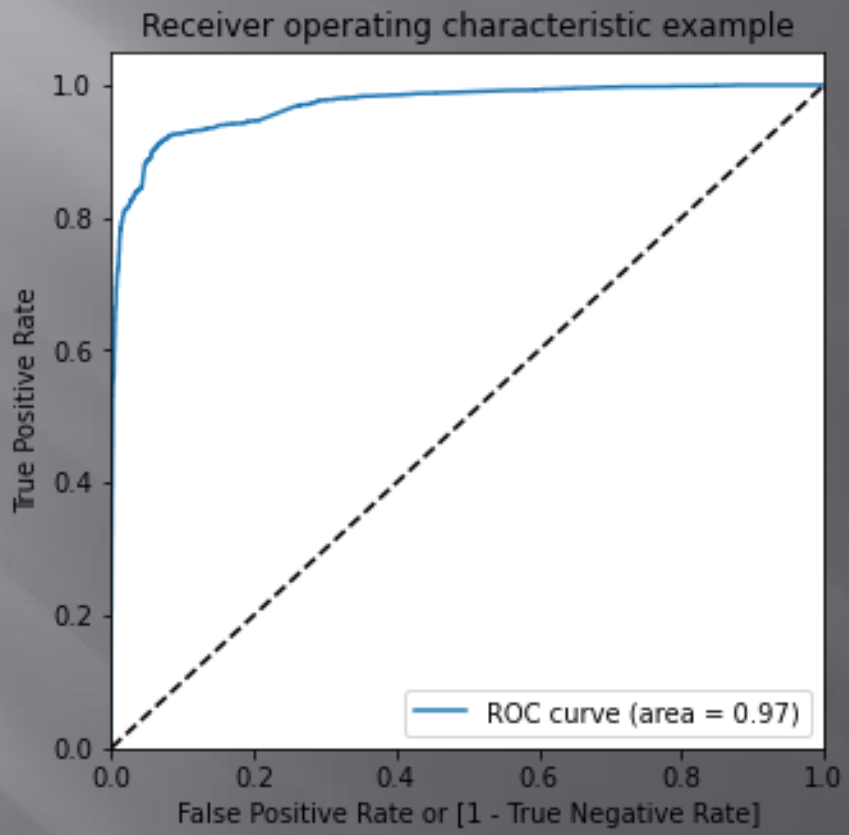
Sensitivity of our logistic regression model is 0.88

Specificity of our logistic regression model is 0.95

Positive predictive value is 0.91

Model building and evaluation

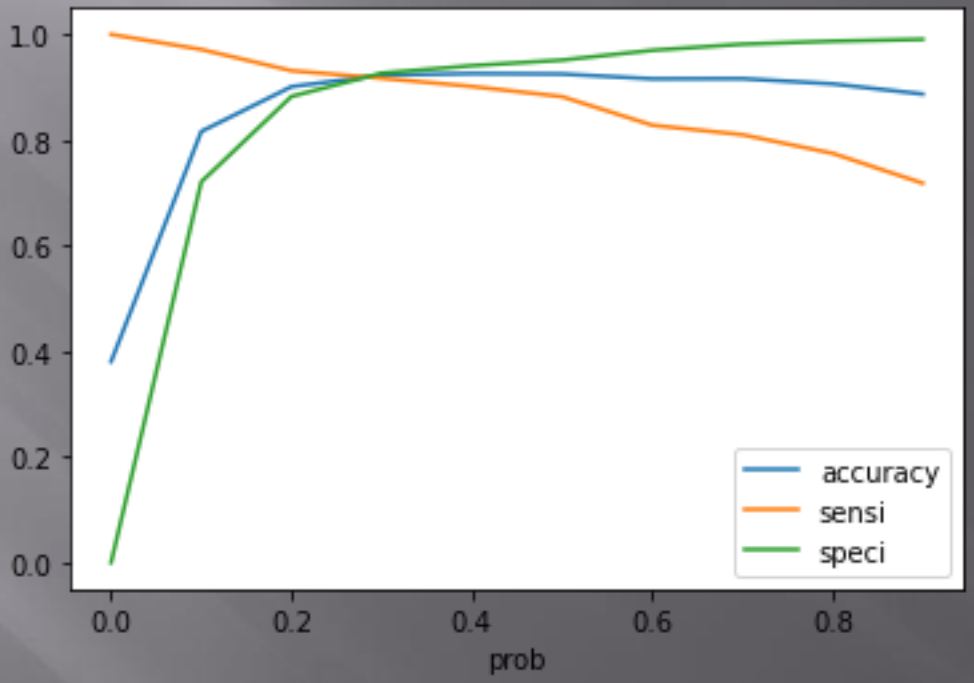
ROC Curve



We are getting a good value of 0.97 indicating a good predictive model.

Model building and evaluation

plotted accuracy sensitivity and specificity for various probabilities.



Model building and evaluation

- we can see that the final model seems to be performing well.
- Some observations are:
- The ROC curve has a value of 0.97, which is very good.
- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%
- Positive predictive value = 0.88
- Precision = 0.88
- Recall = 0.91
- Also do Predictions on Test data
- Putting CustID to index

Model building and evaluation

Observations:
Observations on the Test Data after the model is executed

Accuracy : 92.78%

Sensitivity : 91.98%

Specificity : 93.26%

Final Observation:

Train data observations vs Test data observations

	Accuracy	Specificity	Sensitivity
Train Data	92.29%	92.66 %	91.70%
Test Data	92.78%	93.26%	91.98%

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model