LEAD SCORE CASE STUDY

GROUP MEMBERS:

Himanshu Chaudhary

Sandeep Poduval

Problem Statement

- X education sells online courses to industry professionals
- Although, X Education gets many leads but the Lead conversion is poor at only 30%
- Goal is to identify Hot leads that have a higher chance of conversion
- Once identified, the lead conversion rate should go up as the Sales team can concentrate on converting the Hot leads

Business Objective:

To build a Logistic Regression model to identify the “Hot Leads” and achieve a lead conversion rate up to 80%

Solution Methodology

- Data reading
- Cleaning the data
  - Checking and removing redundant columns
  - Converting the label “select” to null values
  - Removing columns with >35% null values
  - Other columns with missing values are imputed with the mode
- Data Transforming
  - Changing the labels to dummy variables and “Yes”:1 and “No”:0
  - Removing the duplicate columns
- Test Train Split
  - Splitting the data in Test and Train split and scaling the data
  - Plotting the Heat Map
- Model Building
  - Running ref on 15 variable and checking the linear model regression results.
  - Removing the columns with high P value one by one to and checking the VIF value.
  - Predicting on the train model.
  - Checking the ROC curve and checking the accuracy, specificity and senility.
  - Checking and calculating precision and recall with cut-off of 0.35 and 0.41 on train and test data.

Understanding the data

Total Number of Rows = 9240

Total Number of Columns = 37

Removing Single value features – Magazines, Receive more updates about our Courses, update me on supply

Removing Prospect ID and Lead Number

Dropping columns having more than 35% missing values

Data Conversion

Numerical variables were scaled using MinMax Scaler

Dummy variables were created for object type variables

Model Building

Splitting the data into Training and Test slit with 70:30 ratio

Using RFE for feature selection

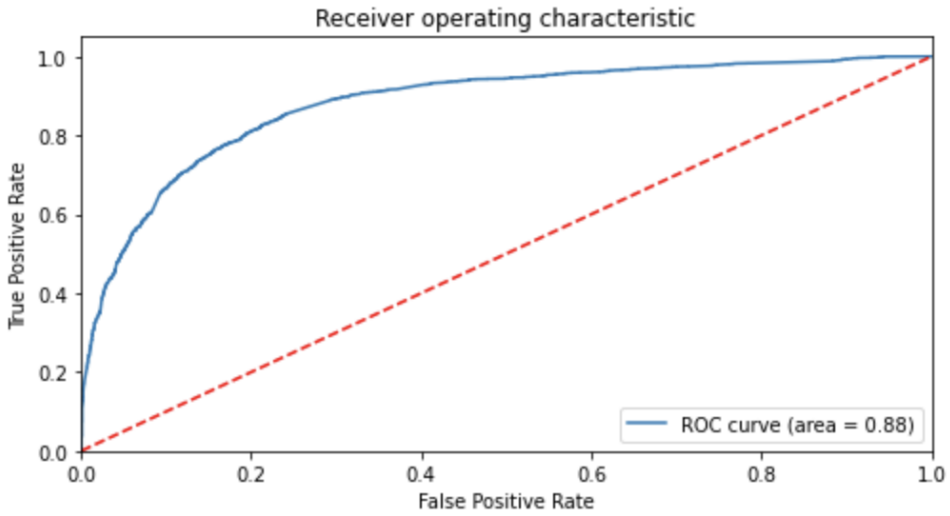
Running RFE with 15 variables

Building Model by removing the variable whose p-value is greater than 0.05 and VIF is greater than 5

Predicting the Test Data

Overall accuracy is 81%

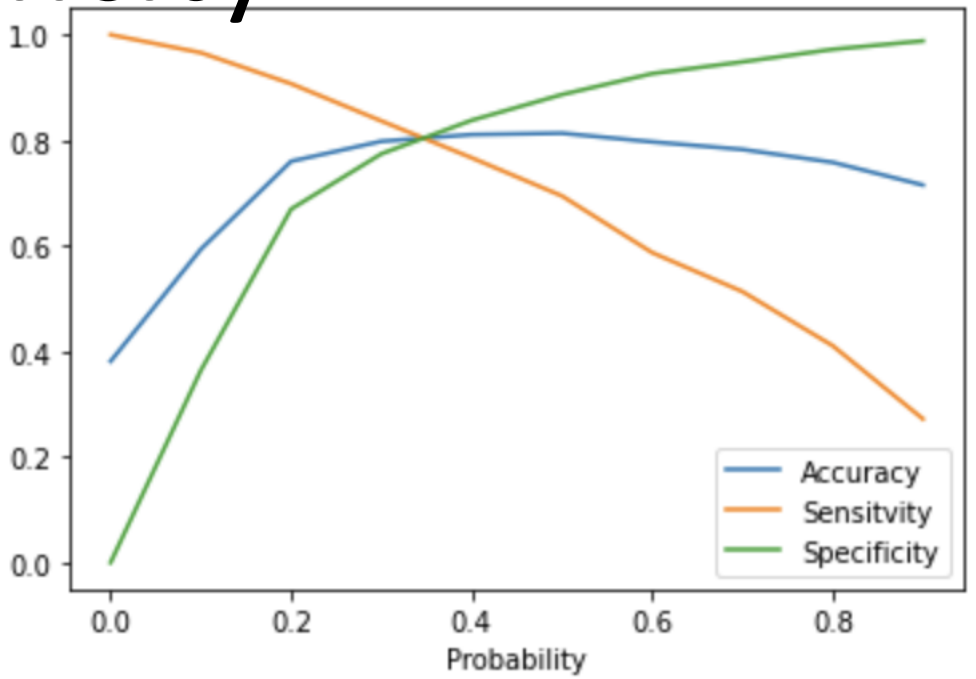
ROC Curve



Finding the Optimal Cut off point where we have balanced Sensitivity and Specificity.

Optimal Cut off is **0.35**

Accuracy, Sensitivity, Specificity



Conclusion

The variables that mattered the most in Potential buyers:

For a good conversion ratio:

- Lead Origin is through a Lead Add form that is the lead has entered his details in the form provided.
- Last Notable Activity is Had a phone conversation, the customer has had a phone conversation with the Sales team.
- And the Lead is a working professional