Lead Scoring Case Study

Submitted By:
Devika Aggarwal
Soumita Das

Problem Statement:

X Education sells online courses to industry professionals.

- **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.**

Business Objective:

1. **X education wants to know most promising leads.**
2. **For that they want to build a Model which identifies the hot leads.**
3. **Deployment of the model for the future use.**

Methodology

- Data cleaning and data manipulation.

1. Check and handle duplicate data.
2. Check and handle NA values and missing values in Rows and Columns
3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

- EDA

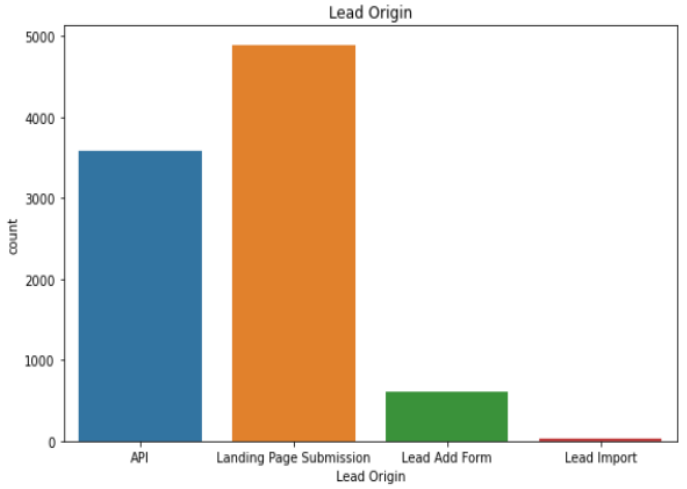
1. Univariate data analysis: value count, distribution of variable etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

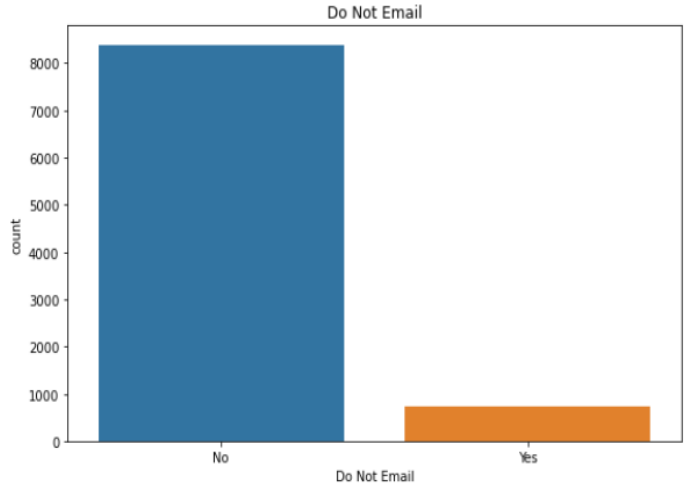
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

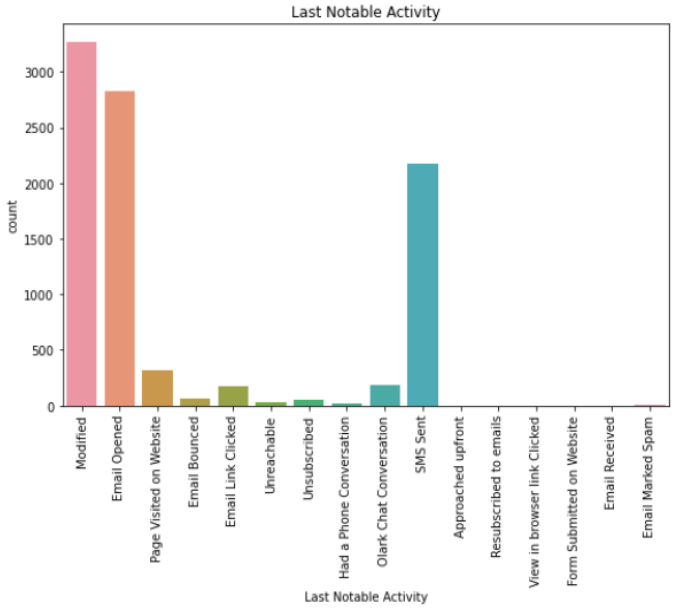
Data Manipulation

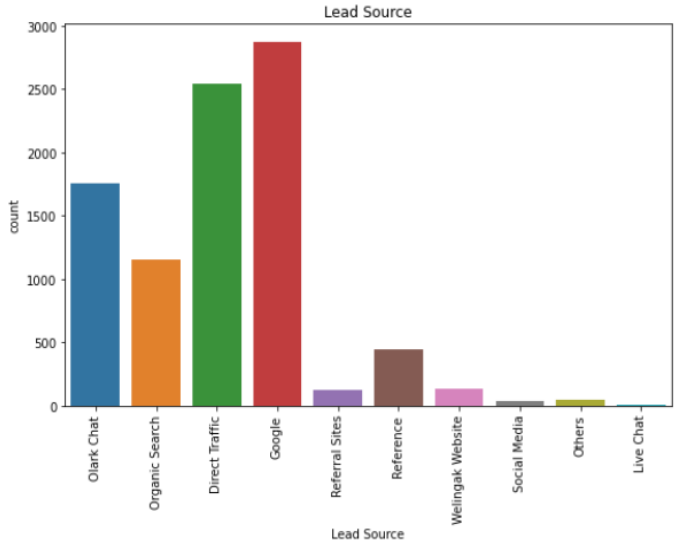
- Number of columns and rows: 37 and 9240 respectively
- Removing and imputing if missing values as per data and there are no duplicate values present.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis. After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 40% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’
- Dropping the rows as per the data as if wont have affected a data as it was less than 1.50%

Univariate Analysis





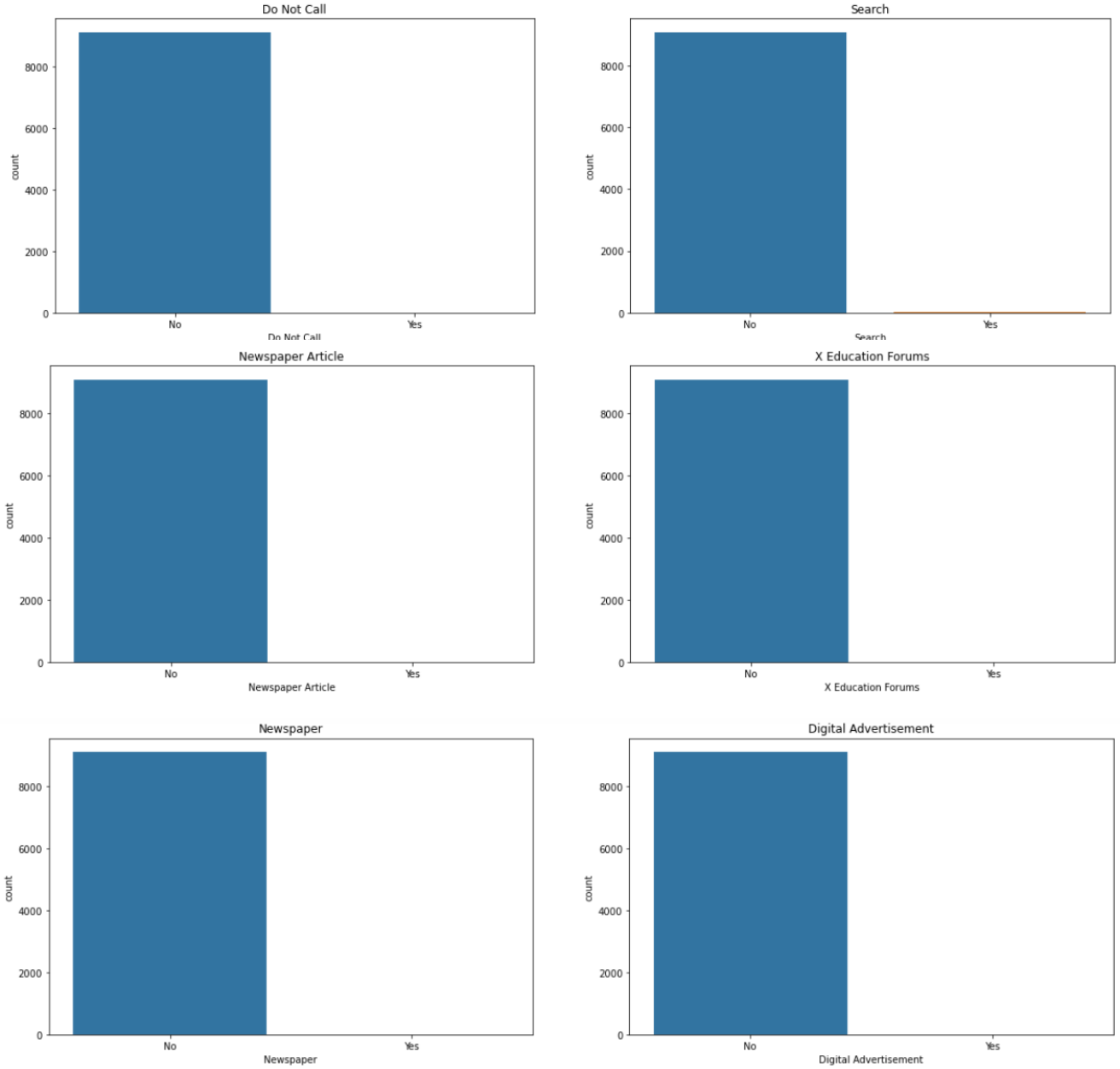




Inferences:

1. Through Landing Page Submission followed by Api we can heavily identify the lead.
2. Most of the leads want the course details to be mailed.
3. Google is the most effective website to generate the leads.
4. Modified is the most Last Notable Activity.

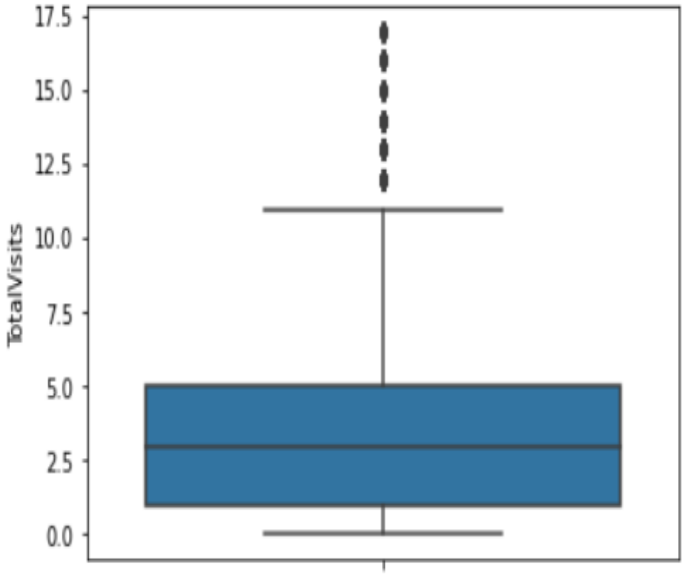
Univariate Analysis

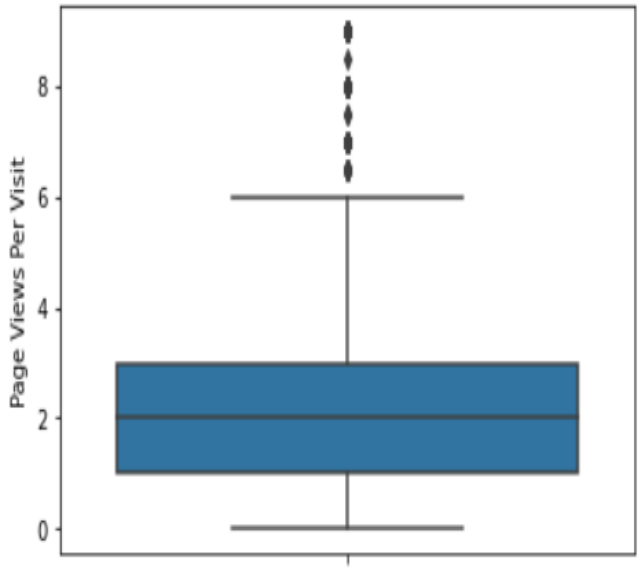


Inference:
Imbalance of data is clearly visible and can be dropped as we cant gather Any information from this

Correlations b/w different variables

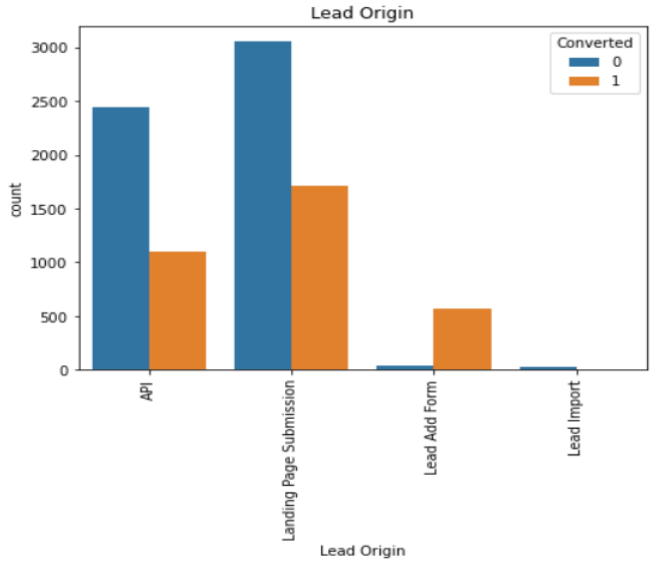
Outliers Check

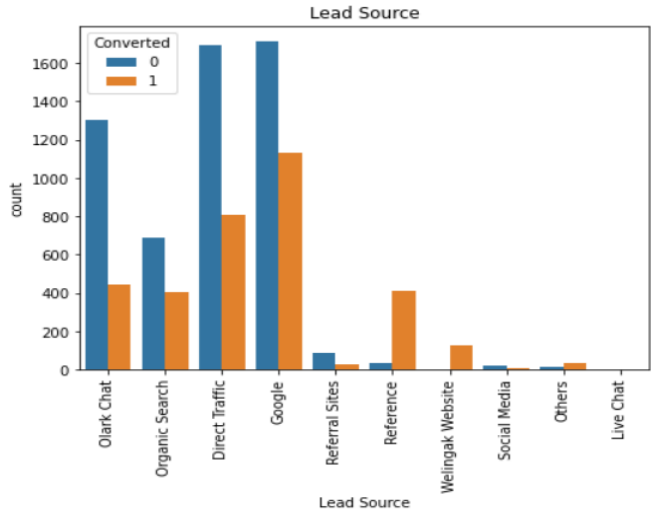


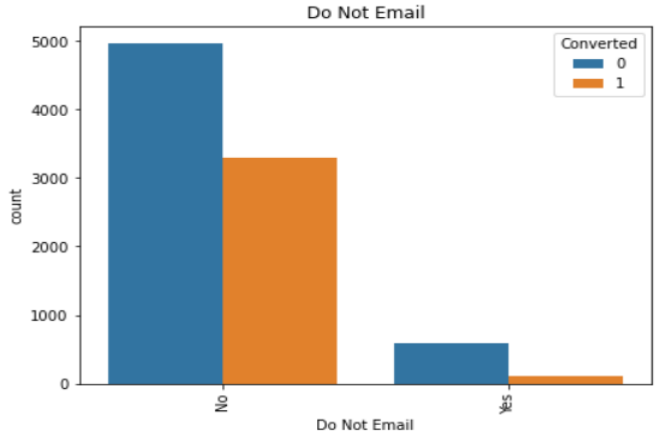


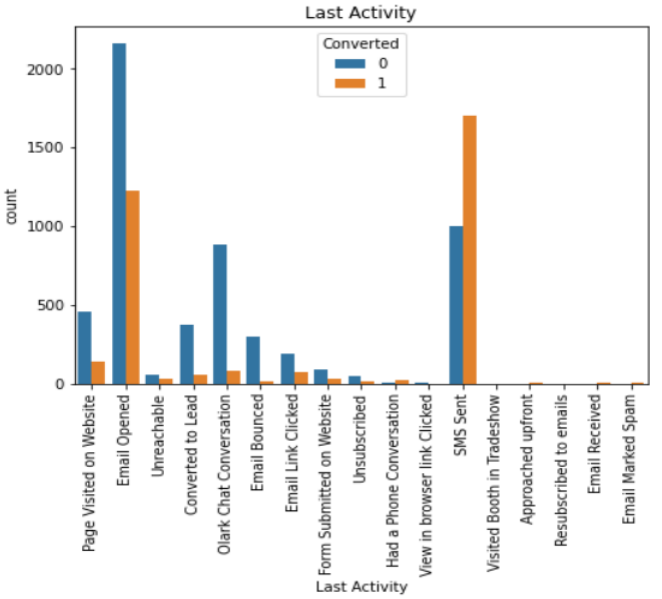
Inference:
Outliers were present in the Total Visits and Page per Visit which were handled.

Bivariate Analysis









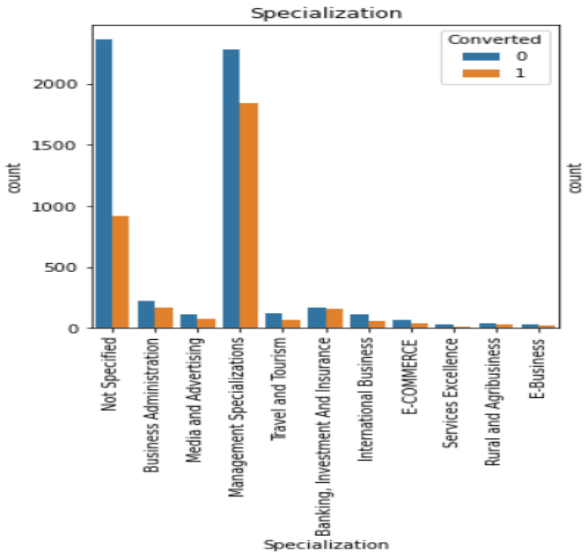
Inferences:

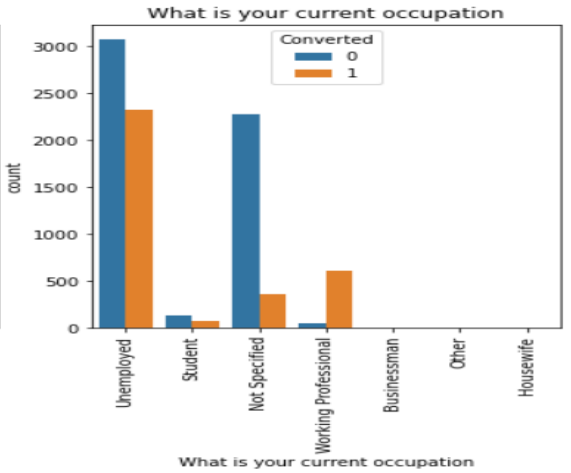
Maximum leads are generated through Google and Direct traffic.

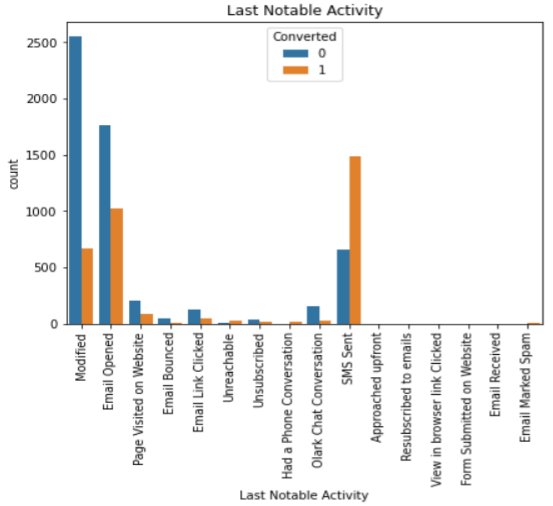
Conversion Rate of reference leads and leads through welingak website is high.

API and Landing Page Submission bring higher number of leads as well as conversion.

Bivariate Analysis







Inferences:
Working Professionals have high chances of joining the course.
Unemployed have high chance of Lead conversion.
Management Specialization have high chances of getting leads converted

Dummy Variables

Train-Test Split

Scaling Of Data

RFE

Model Building

Dummy Variables encodes all of the independent variables as dummy variables allows easy interpretation of the odds ratios, and increases the significance of the coefficients.Dummy Variables are created for object type variables.Numerical Variables are Normalised

Train-Test Split helps to compare our own machine learning model results to machine results. The split was done at 70% and 30% for train and test data respectively.

Scaling Of Data transforms data as it fits within a specific scale. MinMax Scalar is performed.

RFE was performed to attain the top 15 relevant variables

Model was build checking VIF values and p-value (Keeping the variables VIF < 5 and p-value < 0.05).

Predictions

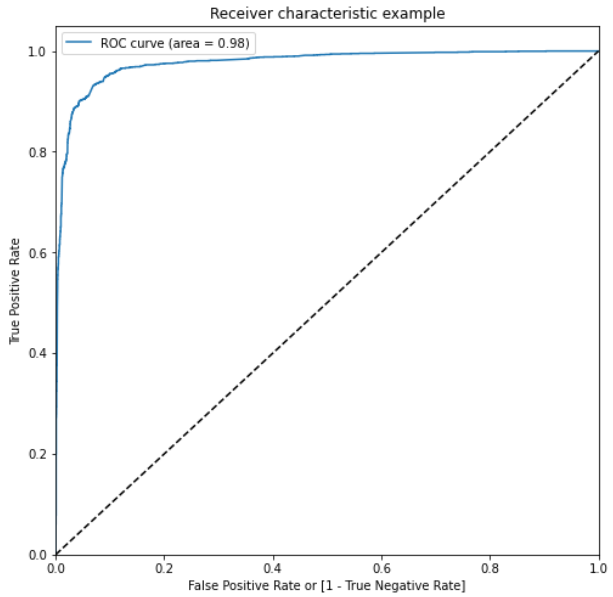
ROC

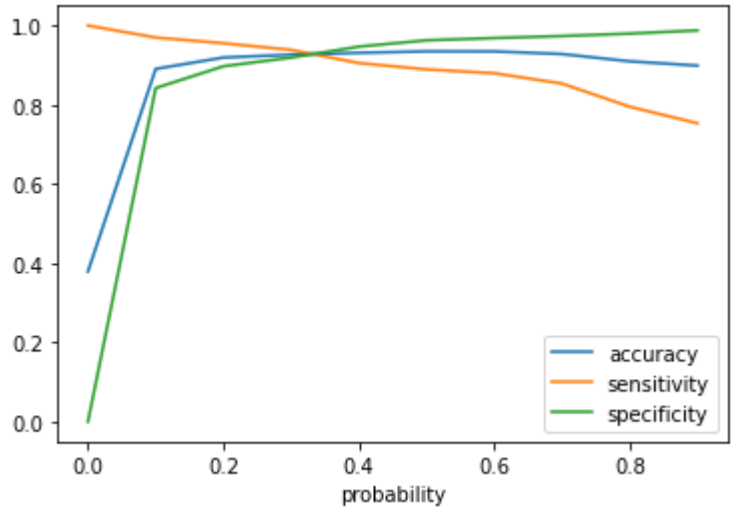
With cutoff 0.5 we have

Accuracy = 93.48% **Sensitivity = 88.93%**

Specificity = 96.27% **Precession = 93.57%**

We found out that specificity is 96.27% which is good and Sensitivity is 88.93% which needs to be handled.





ROC should be closer to 1 and we are Getting ROC = 0.98 which is very good value which indicates that it is a great model.Its clear from Probability Graph that Optimal CutOff is **The optimal cutoff is at 0.35**.

Results And Conclusion

With cutoff at 0.35 following are the results:

a. Train Data

Accuracy = 93.01% Sensitivity/Recall = 93.18%

Specificity = 92.90% Precession = 88.91%

b. Test Data

Accuracy = 93.18% Sensitivity/Recall = 93.13%

Specificity = 93.22% Precession = 89.35%

It seems that Model will predict the Conversion Rate very well.

Recommendations

X Education have a high chance of getting all the potential buyers to buy their courses by focusing on :

1. The customers are visiting the website more often/ is eager to know about the course by visiting the website several times and downloading the program sheet.

2. There last modified activity falls under the category ‘Modified’.

3. The customers who have opened the Email to have a look over the program.