Lead Score Case Study

Students –

1. Venugopalan MR
2. Vaasanthi Polu
3. Dipali Gurnule

Business Understanding

- An education company named X Education sells online courses to industry professionals.
- People land on the website through different sources, sometimes reference and fill in the form. Once the form is filled they are converted to lead.The typical lead conversion rate at X education is around 30%.
- 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', as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective

To select the leads that are most likely to convert into paying customers.

- The company requires a model wherein a lead score is assigned 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.
- A ballpark of the target lead conversion rate to be around 80%.

Solution Process -

- Data Cleaning and Manipulation –

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

- EDA –

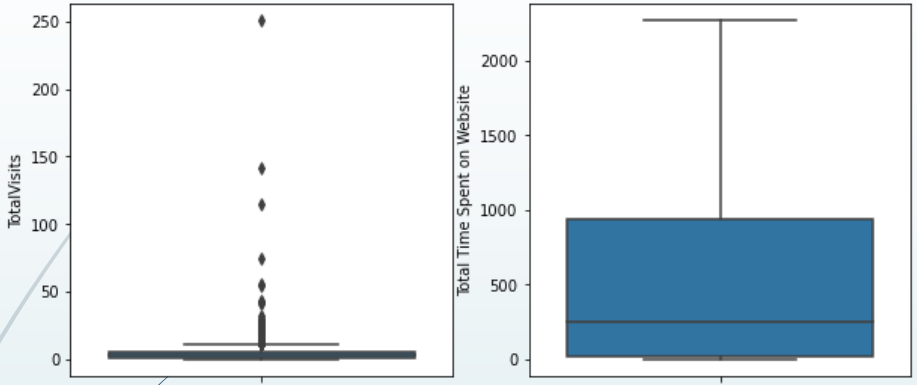
1. Univariate data analysis: value count, distribution of variable
2. Bivariate data analysis: correlation coefficients and pattern between the variables

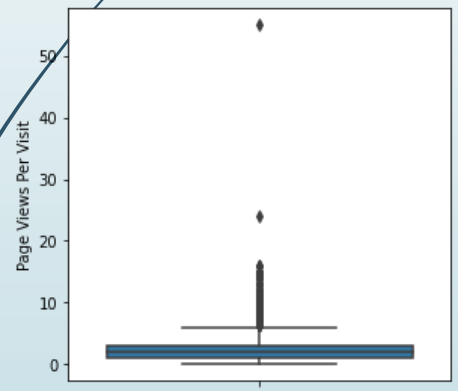
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction
- Validation of the model
- Model presentation
- Conclusions and recommendations.

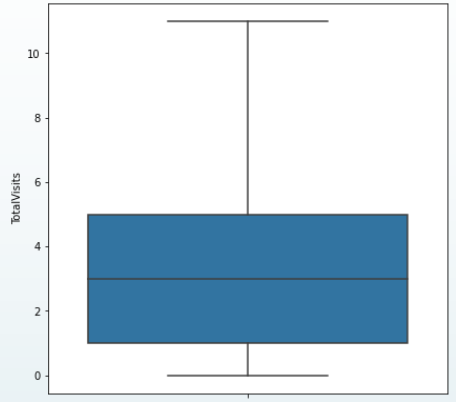
Data Manipulation

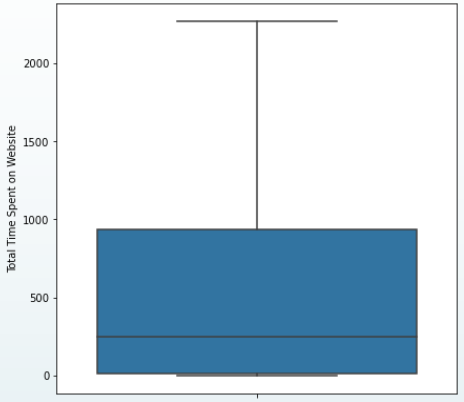
- There is a value select in few features which is almost equal to null, change all the select values to ‘np.nan’
- Drop ‘Prospect ID’, ‘lead number’ columns as they are index columns
- Single value features like 'I agree to pay the amount through cheque’, 'Magazine', 'Receive More Updates About Our Courses’, ‘ Update me on Supply Chain Content', 'Get updates on DM Content‘ are dropped as they show no variance on the data.
- Dropping columns with no variance after univariate analysis 'Do Not Call', 'Search', 'Newspaper Article',‘ X Education Forums', 'Newspaper', 'Digital Advertisement', ‘Through Recommendations‘ , 'Country',‘ What matters most to you in choosing a course'
- Dropping the columns having more than 35% as missing value such as How did you hear about X Education and Lead Profile.

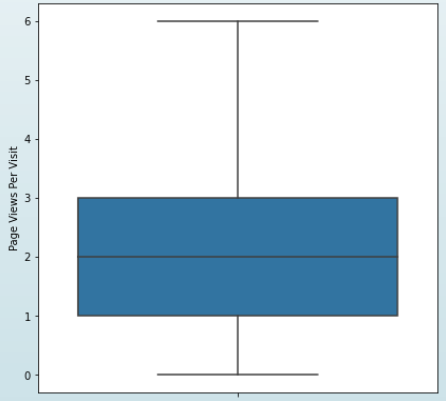
Outliers are identified and handled using capping .



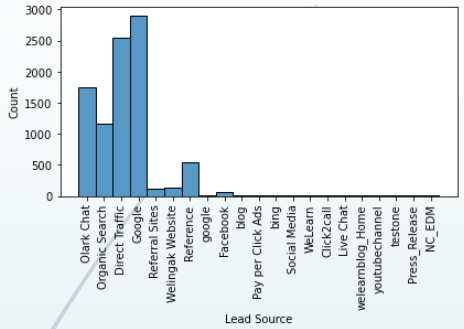


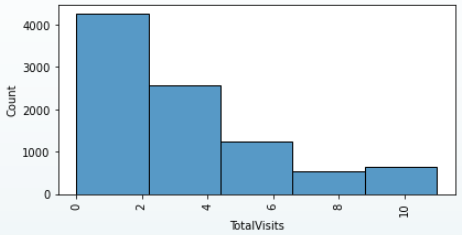


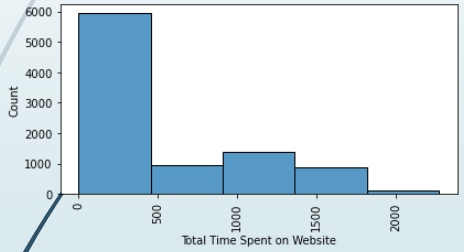


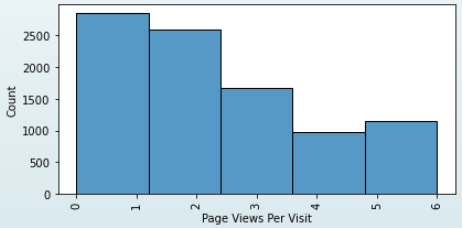


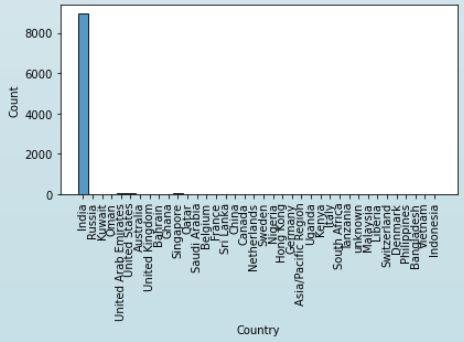
Univariate Analysis





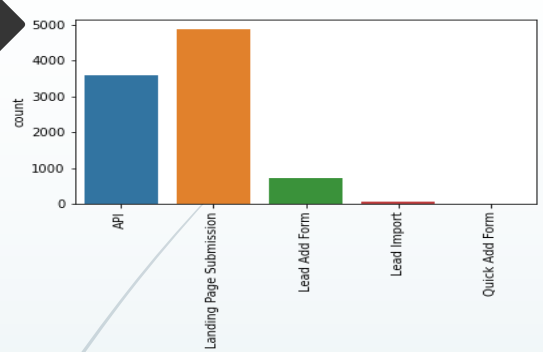


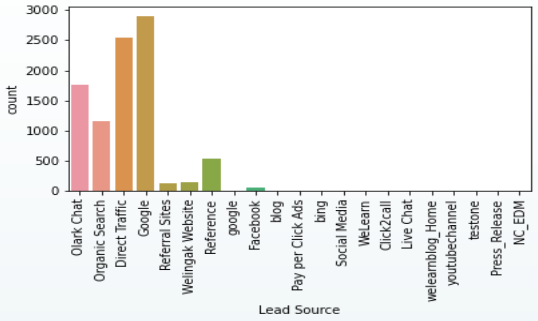


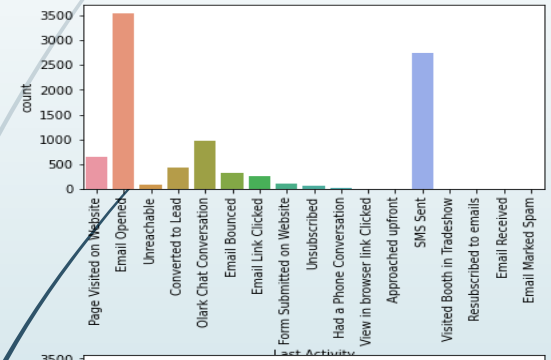


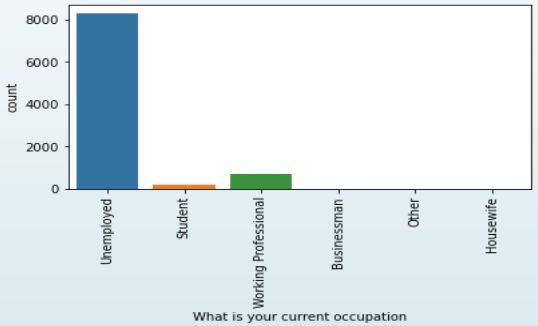
Using histplot for numerical varaibles:

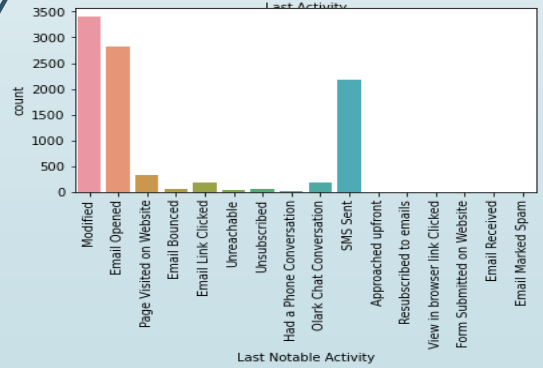
1. We see the lead source is mostly from google
2. Only India has more leads
3. Total visits are graddually decreasing







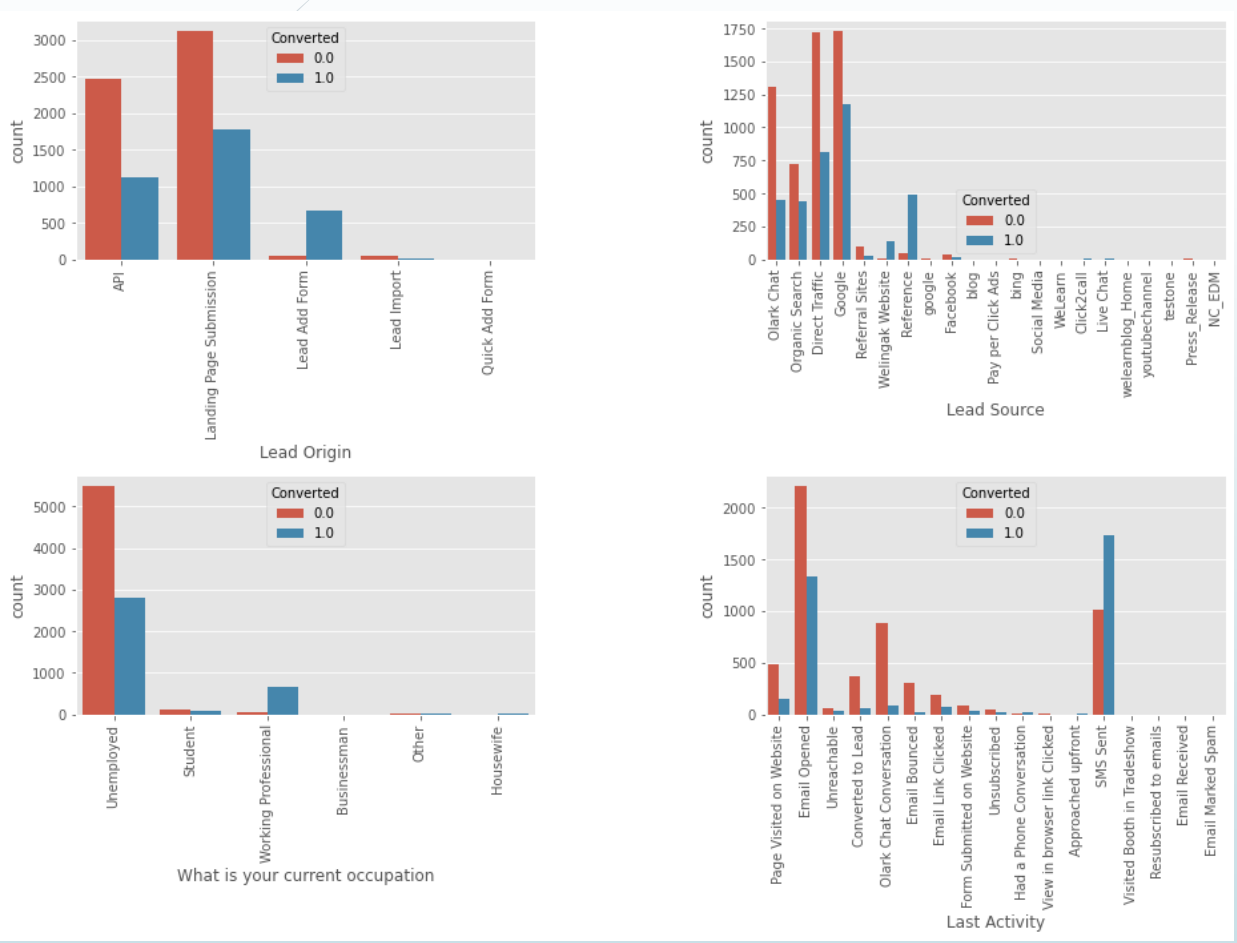




Using Countplot for categorical variables

1. Unemployed preferred to chose the company more
2. lead source is more from google followed by direct traffic

Bivariate Analysis



1. People who landed on page submission are converted
2. people from lead source as google are considered for conversion
3. Working professionals are considered for conversion
4. People who receive an sms are considered for conversion

Data Preparation

- Binary Variables are mapped.
- Categorical Variables with multiple levels are one-hot encoded(dummy variables are created)
- Highly Correlated dummy varibles are dropped

Model Building

- Train-Test split
  - Splitting the dataset in 70,30 ratio for train,test
- Using StandardScaler for scaling the numerical variables.
- Using logistic regression() for building the model
- Using RFE algorithm for feature selection using 20 variables as output
- Using GLM to fit the model
- Eliminating the insignificant variables whose p value is greater than 0.05 and VIF value greater than 5

Out[83]:

The final model obtained is:

	Feature	VIF
0	Do Not Email	1.806070
1	Total Time Spent on Website	1.203181
2	Lead Origin_Lead Add Form	1.420101
3	Lead Source_Olark Chat	1.668478
4	Lead Source_Welingak Website	1.234910
5	Last Activity_Converted to Lead	1.240665
6	Last Activity_Email Bounced	1.794549
7	Last Activity_Olark Chat Conversation	1.988498
8	What is your current occupation_Working Profes...	1.136381
9	Last Notable Activity_Email Link Clicked	1.018204
10	Last Notable Activity_Email Opened	1.103369
11	Last Notable Activity_Had a Phone Conversation	1.001363
12	Last Notable Activity_Modified	1.915776
13	Last Notable Activity_Olark Chat Conversation	1.322950
14	Last Notable Activity_Page Visited on Website	1.024364

Evaluating the model

Using confusion matrix on train data to find accuracy, specificity, sensitivity

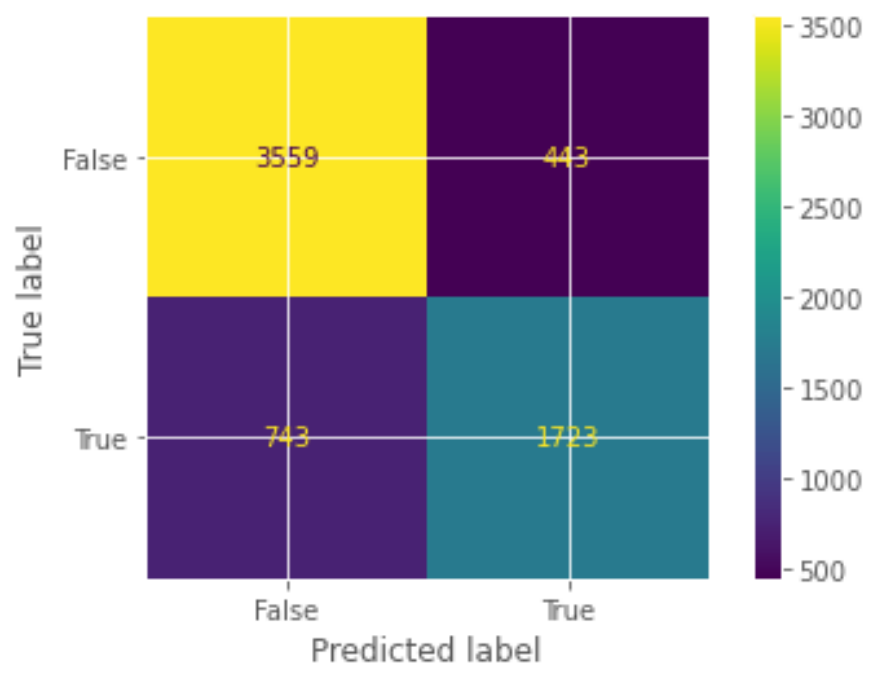
Using confusion matrix on test data to find accuracy, specificity, sensitivity

Train set:

accuracy=81.6%

sensitivity=69.9%

specificity=88.9%

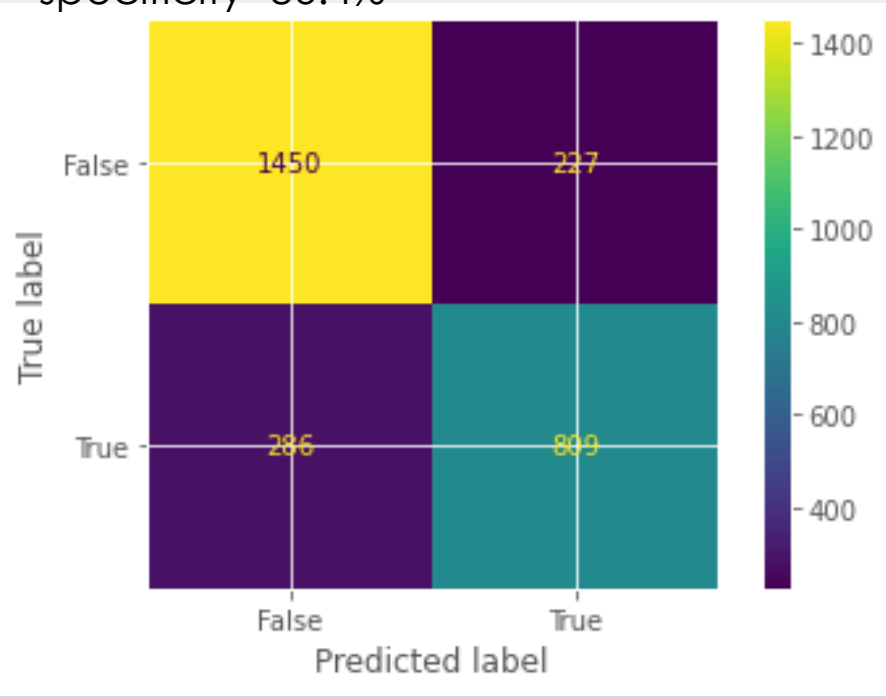


Test set:

accuracy=81.7%

sensitivity=73.8%

specificity=86.4%

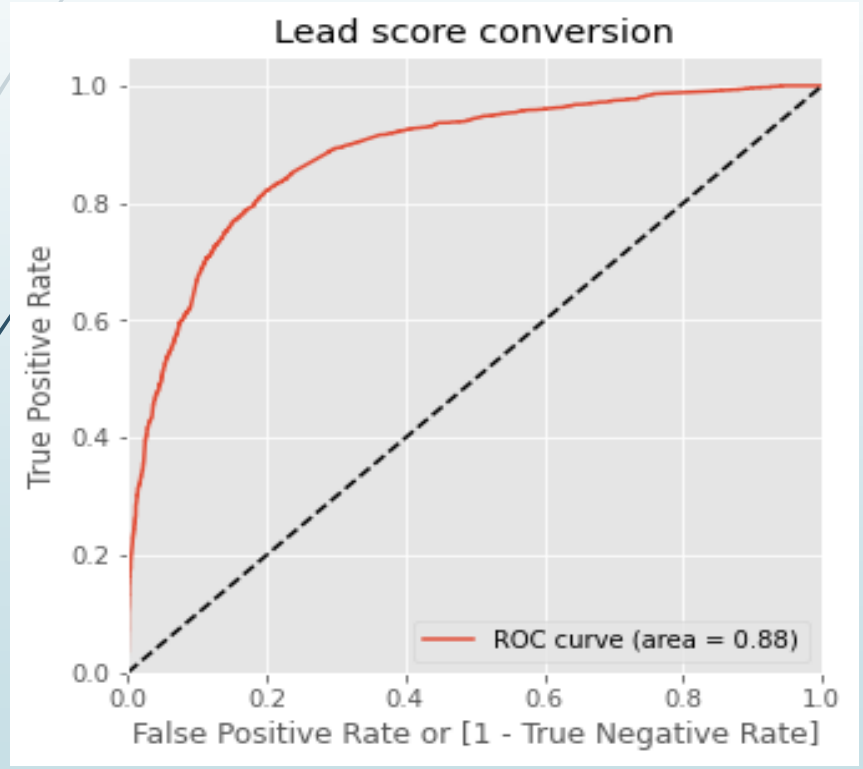


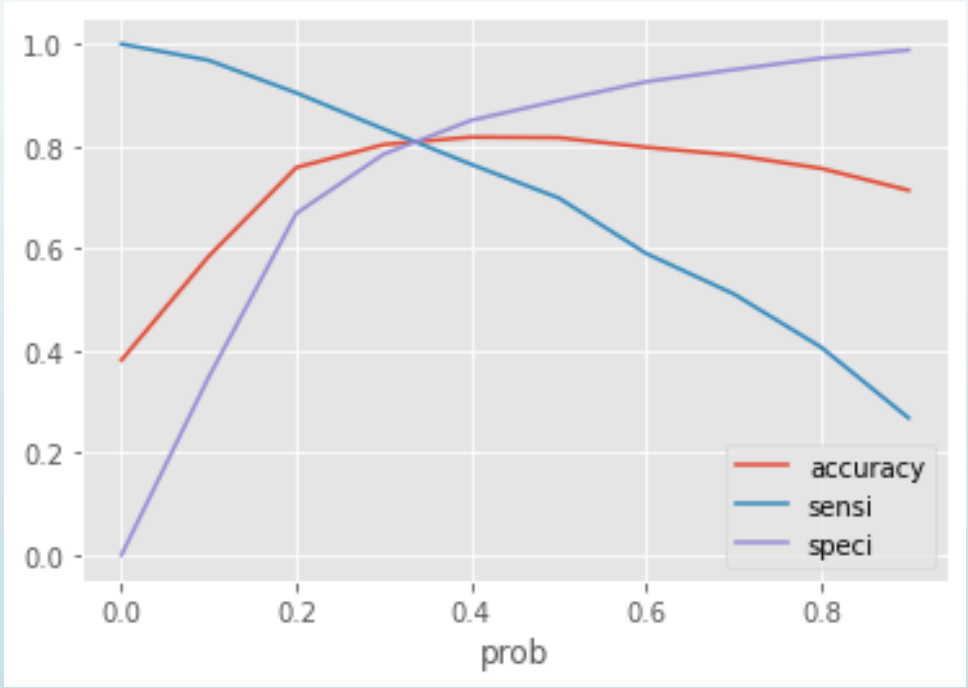
ROC Curve

It shows the tradeoff between sensitivity and specificity

Optimal cutoff probability is that probability where we get balanced sensitivity and specificity

From the curve 0.3 is the optimum point to take it as a cutoff probability.





Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order)

- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional
- Last Notable Activity_Had a Phone Conversation
- Lead Source
  - Welingak Website
  - Olark Chat
- Total Time Spent on Website

The Model has achieved 80% accuracy required by the company.

The top leads are listed which can be contacted are more than 100