Lead Score Case Study
Logistic Regression

Peeyush Gera

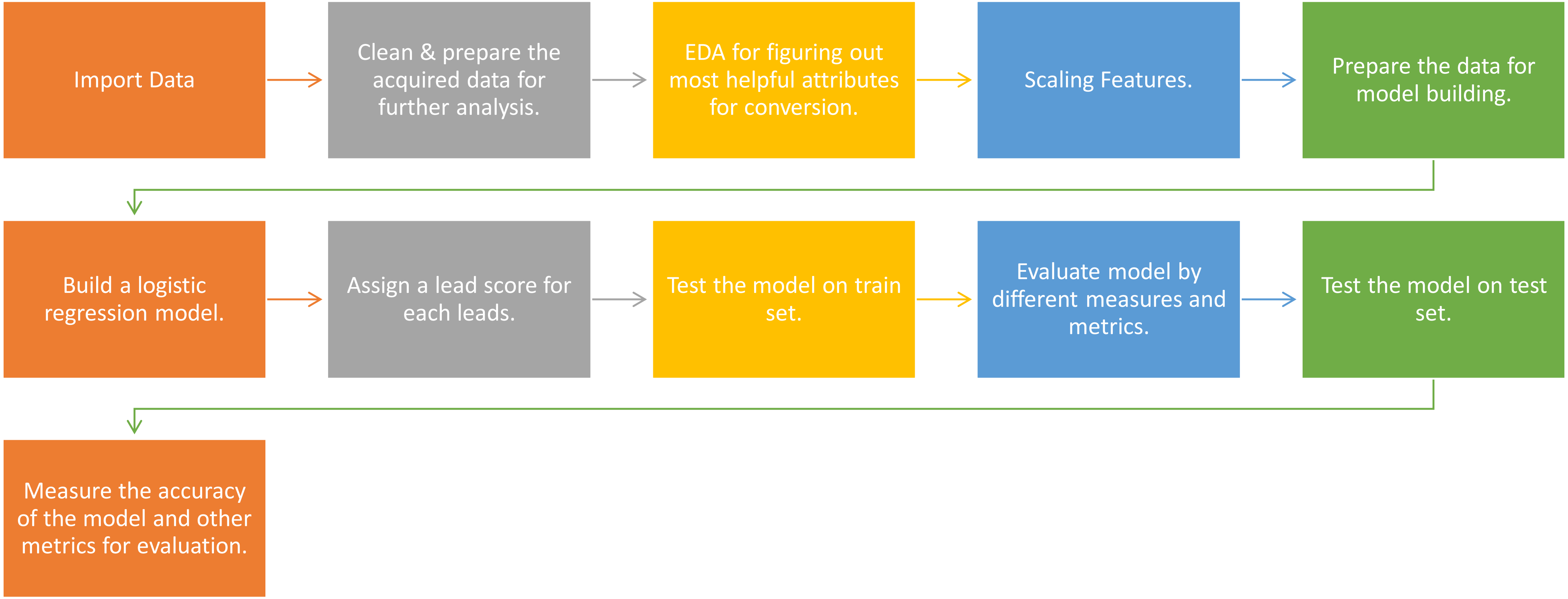
Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The company markets its courses on several websites and search engines like Google.
- 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'. 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.
- X Education has appointed you to help them select the most promising leads, i.e. 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%.

Business Goal

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e., is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Strategy



Import Data

Clean & prepare the acquired data for further analysis.

EDA for figuring out most helpful attributes for conversion.

Scaling Features.

Prepare the data for model building.

Build a logistic regression model.

Assign a lead score for each leads.

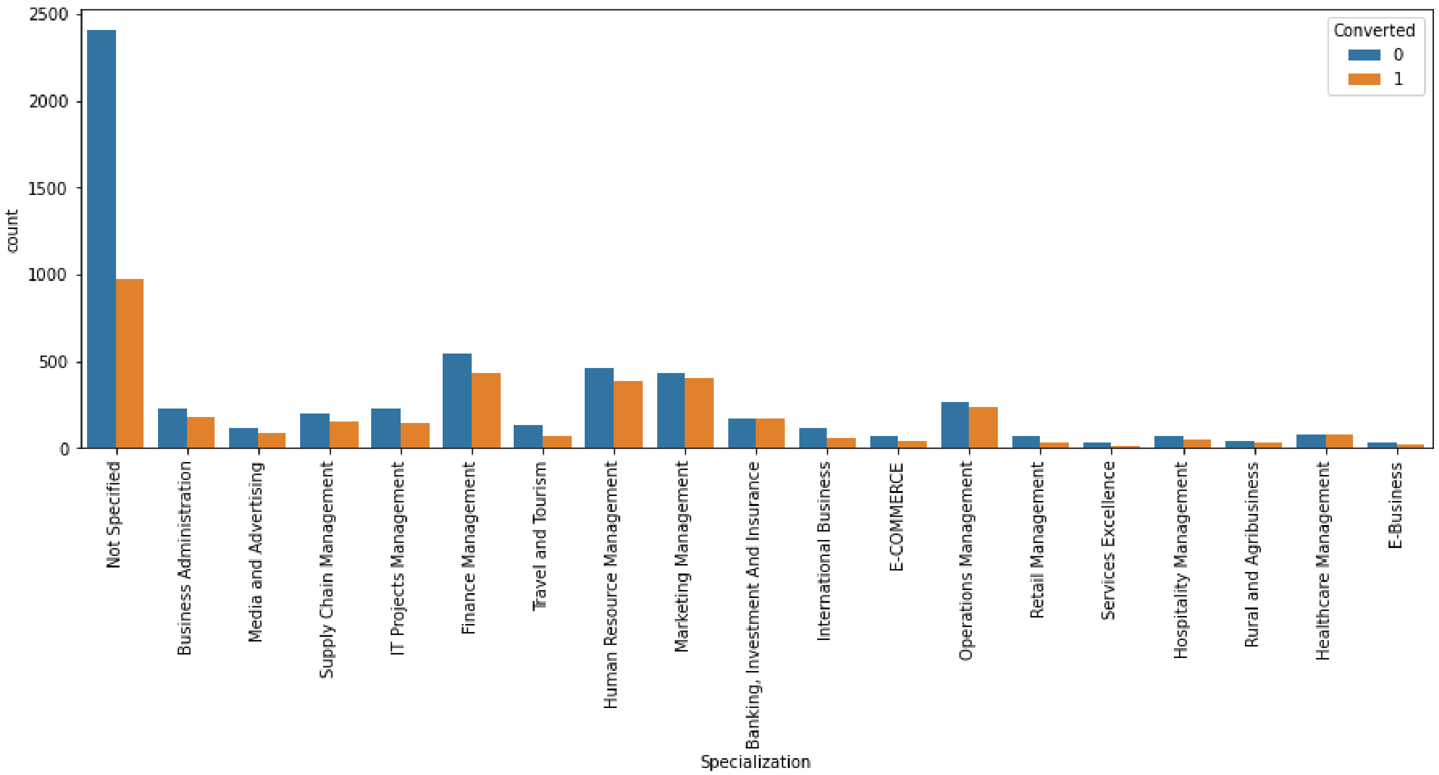
Test the model on train set.

Evaluate model by different measures and metrics.

Test the model on test set.

Measure the accuracy of the model and other metrics for evaluation.

Exploratory Data Analysis

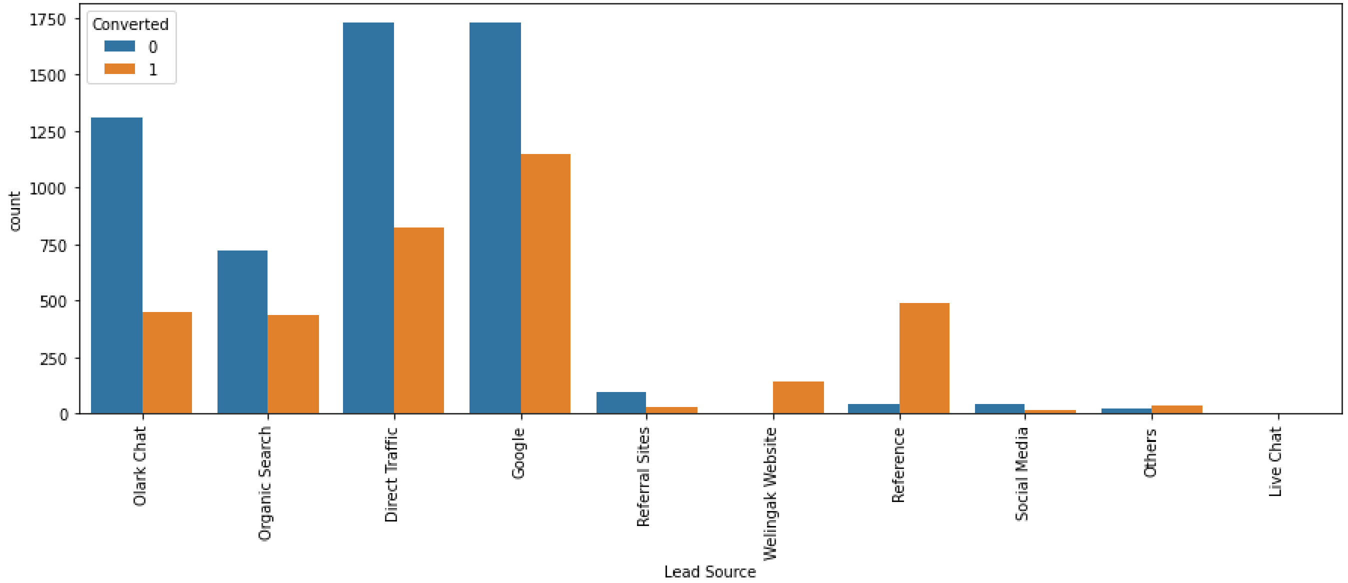


- Spread of Specialization column

We see that specialization with Management in them have higher number of leads as well as leads converted. So, this is definitely a significant variable and should not be dropped.

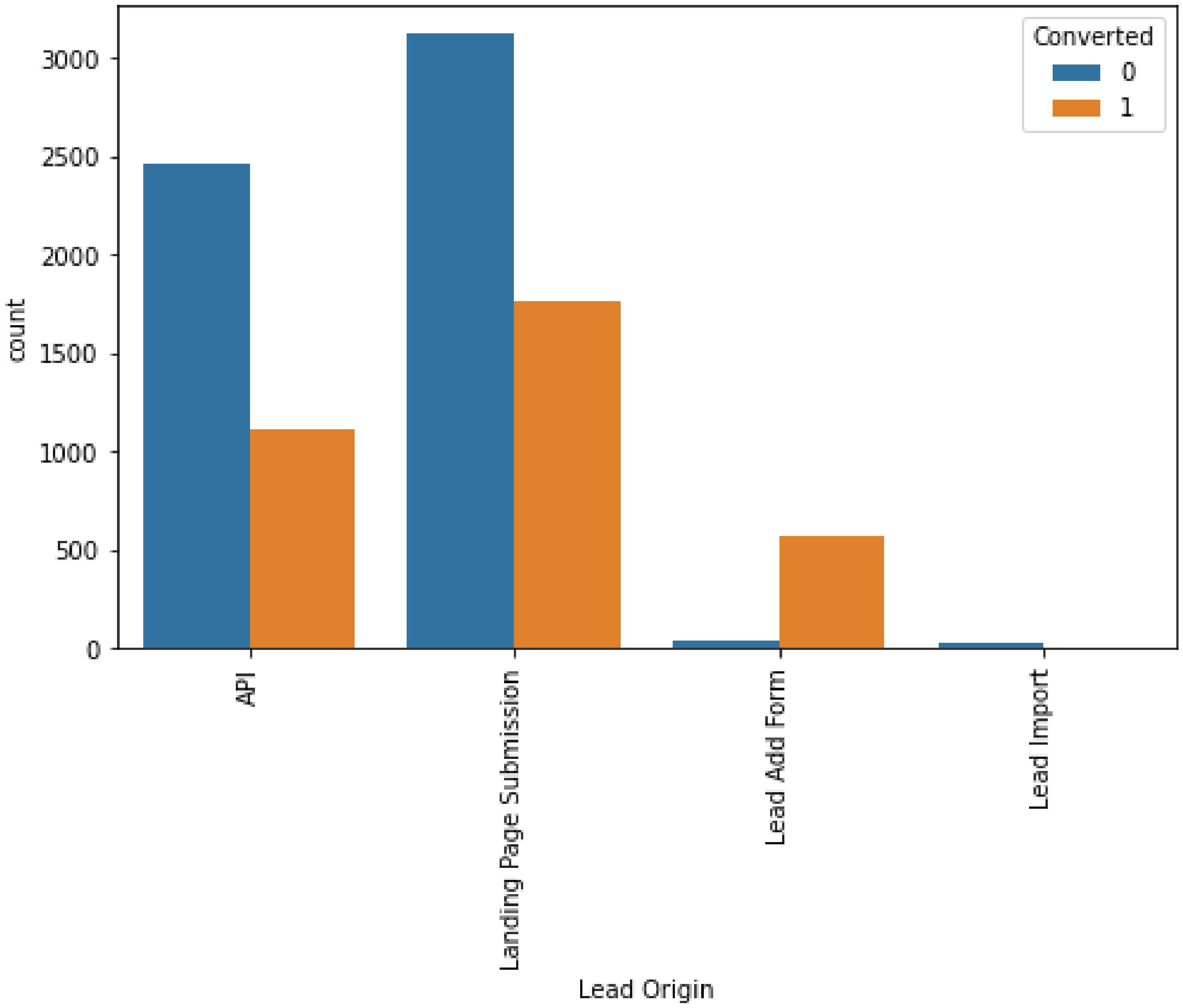
EDA

- Visualizing count of Variable based on Converted value
- Inference Maximum number of leads are generated by Google and Direct traffic. Conversion Rate of reference leads and leads through welingak website is high. To improve overall lead conversion rate, focus should be on improving lead conversion of Olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.



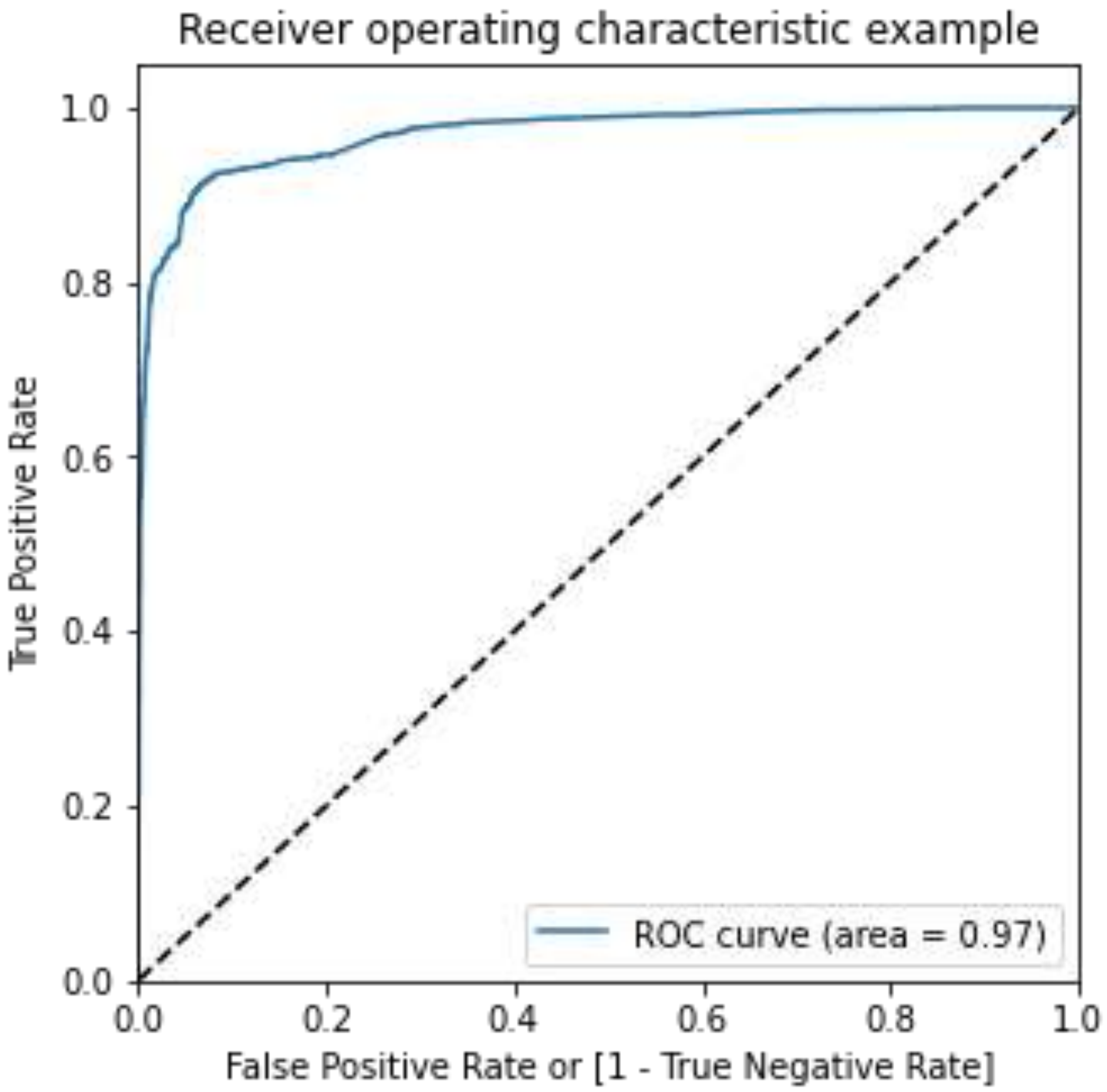
Lead Origin Vs Count

- Inference API and Landing Page Submission bring higher number of leads as well as conversion. Lead Add Form has a very high conversion rate but count of leads are not very high. Lead Import and Quick Add Form get very few leads. In order to improve overall lead conversion rate, we have to improve lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.



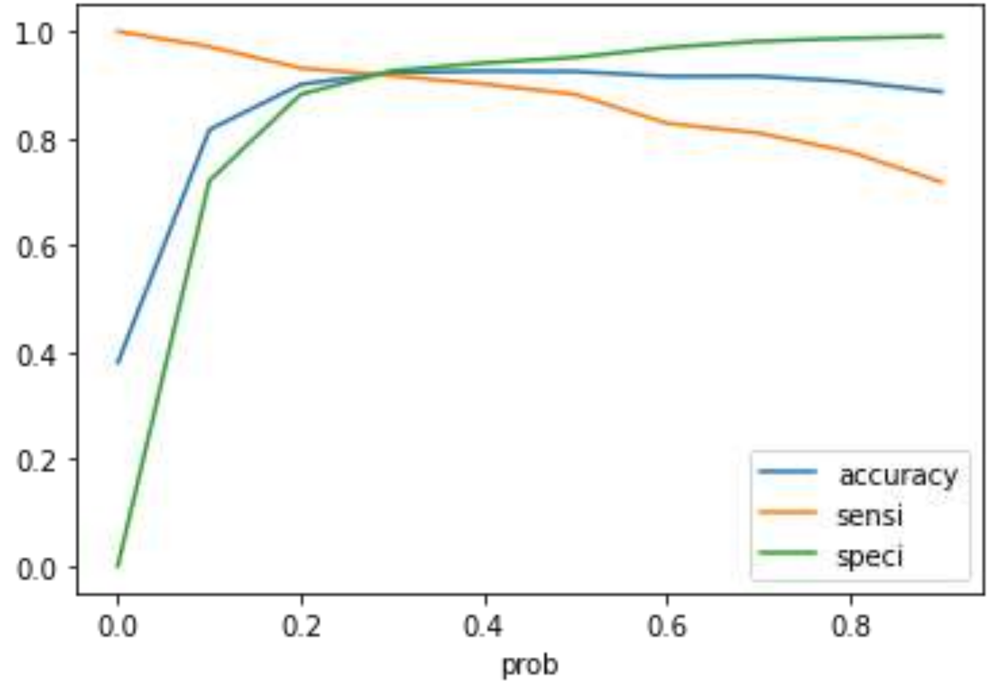
PLOTTING ROC CURVE

- The ROC Curve should be a value close to 1. We are getting a good value of 0.97 indicating a good predictive model.
- Finding Optimal Cutoff Point
- Above we had chosen an arbitrary cut-off value of 0.5. We need to determine the best cut-off value and the below section deals with that:



Plot accuracy sensitivity and specificity for various probabilities.

- Observation:
- So as we can see above the model seems to be performing well. The ROC curve has a value of 0.97, which is very good. We have the following values for the Train Data:
- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%
- Some of the other Stats are derived below, indicating the False Positive Rate, Positive Predictive Value,Negative Predictive Values, Precision & Recall.



PREDICTIONS ON TEST SET

- Observation:
- After running the model on the Test Data these are the figures we obtain:
- Accuracy : 92.78%
- Sensitivity : 91.98%
- Specificity : 93.26%
- Final Observation: Let us compare the values obtained for Train & Test:
- Train Data:
- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%
- Test Data:
- Accuracy : 92.78%
- Sensitivity : 91.98%
- Specificity : 93.26%
- 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