Lead score case study

LOGISTIC REGRESSION

BY Uttam Kumar and Sunil Kishore

Problem statement:

The case study describes the steps in developing the lead scoring model. The lead score can be calculated by taking into account various factors such as the lead's likelihood of converting. X Education is a seller of online courses that are designed for industry professionals. It needs help identifying the most promising leads.

Business Goal:

A lead scoring model should be used by the company to assign a lead score to each prospect. It will help determine which leads have a higher chance of converting and which ones have a lower chance. The CEO of the company had stated that the company's objective is to have a lead conversion ratio of 80%.

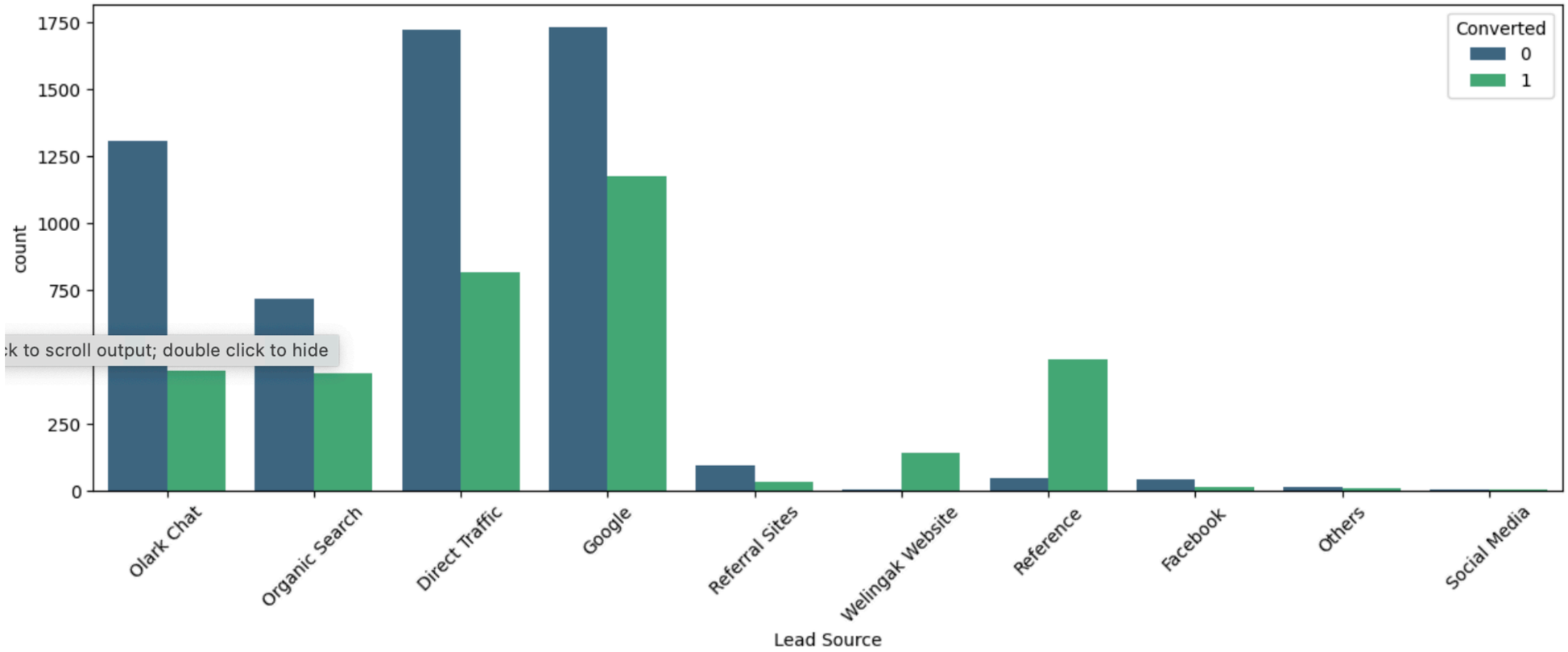
STRATEGY

1. Import data
2. Clean and Prepare the acquired data for EDA
3. Exploratory data analysis to find out important feature
4. Calling feature
5. Prepare the data for model building
6. Build a logistic regression model
7. Test the model
8. Evaluate the model
9. Measure the accuracy of the model

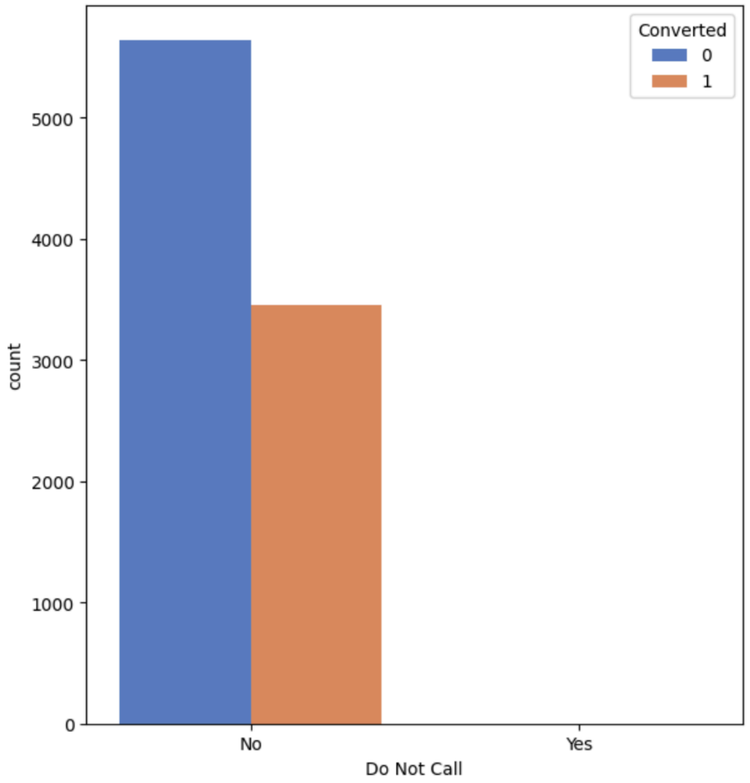
Exploratory Data Analysis

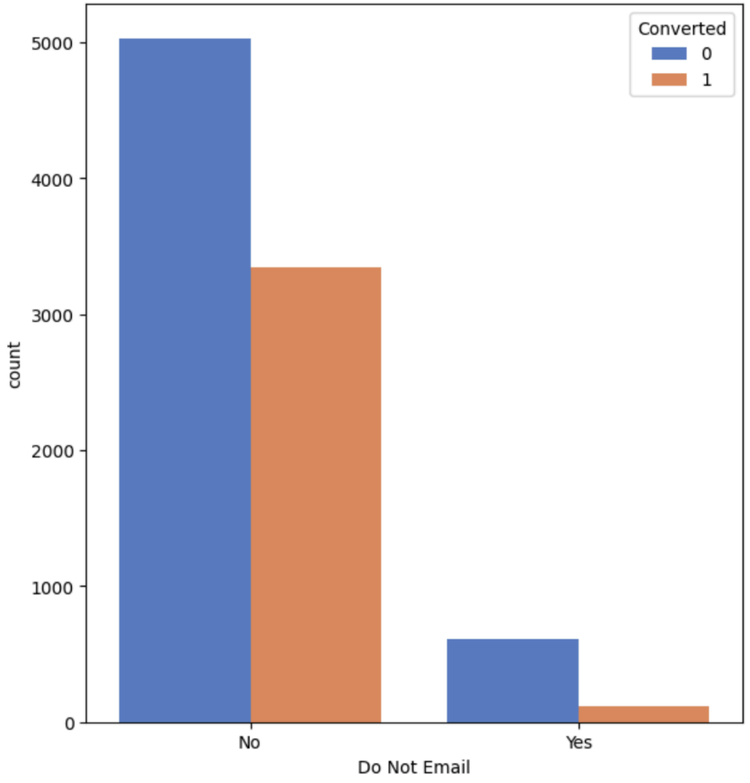
Categorical Nominal Analysis

Google search has the highest conversion compared to other platform

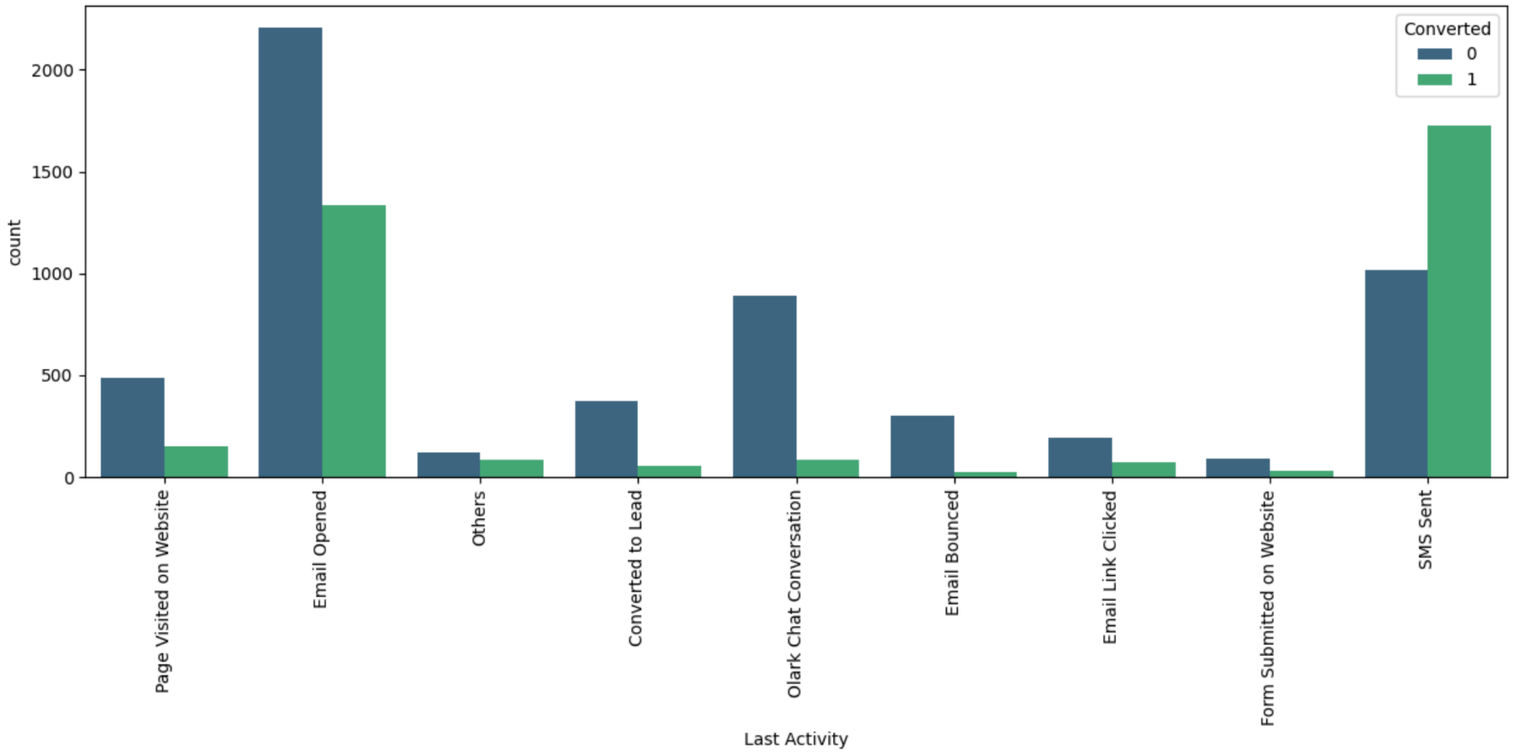


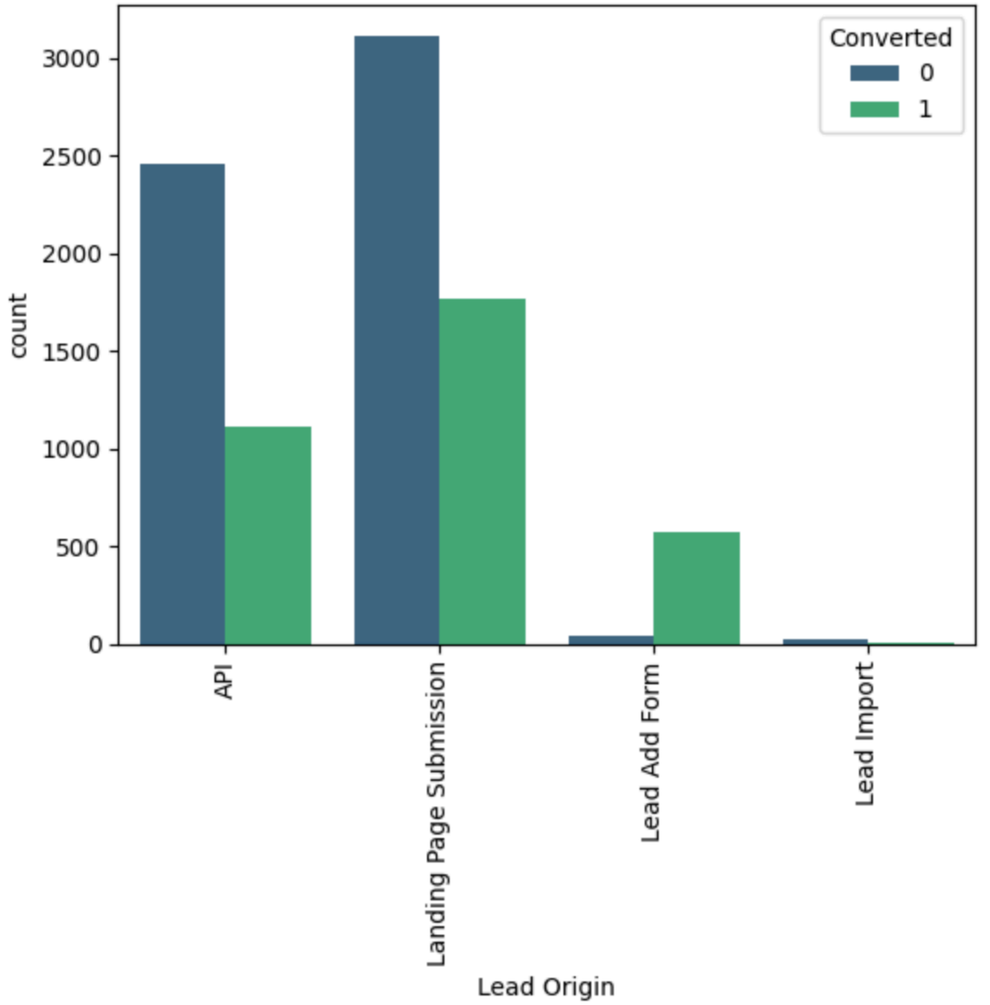
Most leads prefer not to informed through phone call





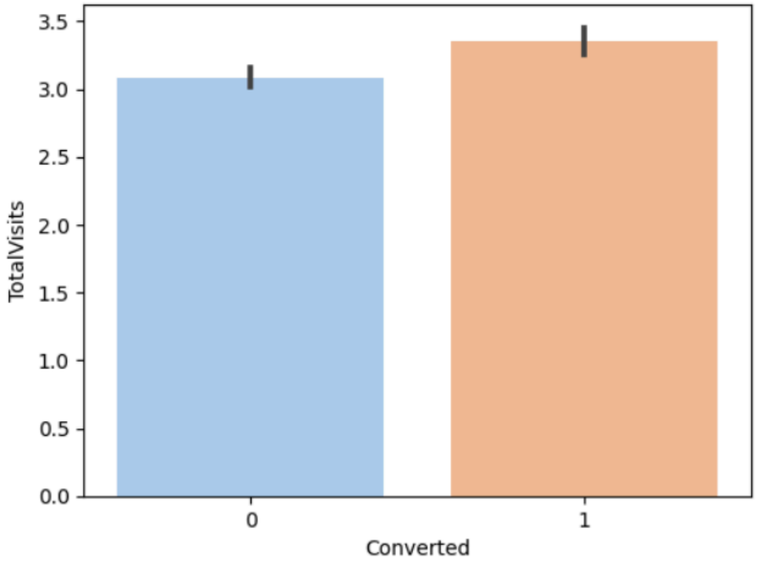
SMS and Email has shown to be a promised method for getting higher confirmed leads

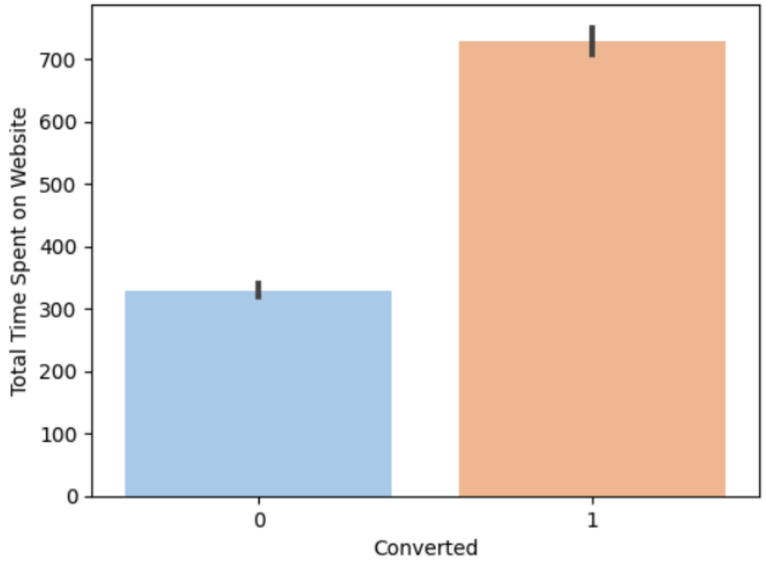


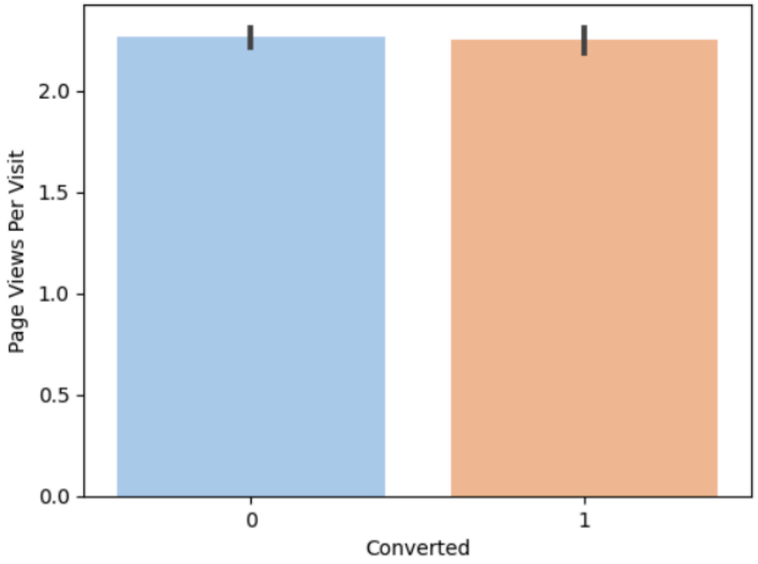


Landing page submission has high lead conversion.

People spending higher than average time are more promising lead.

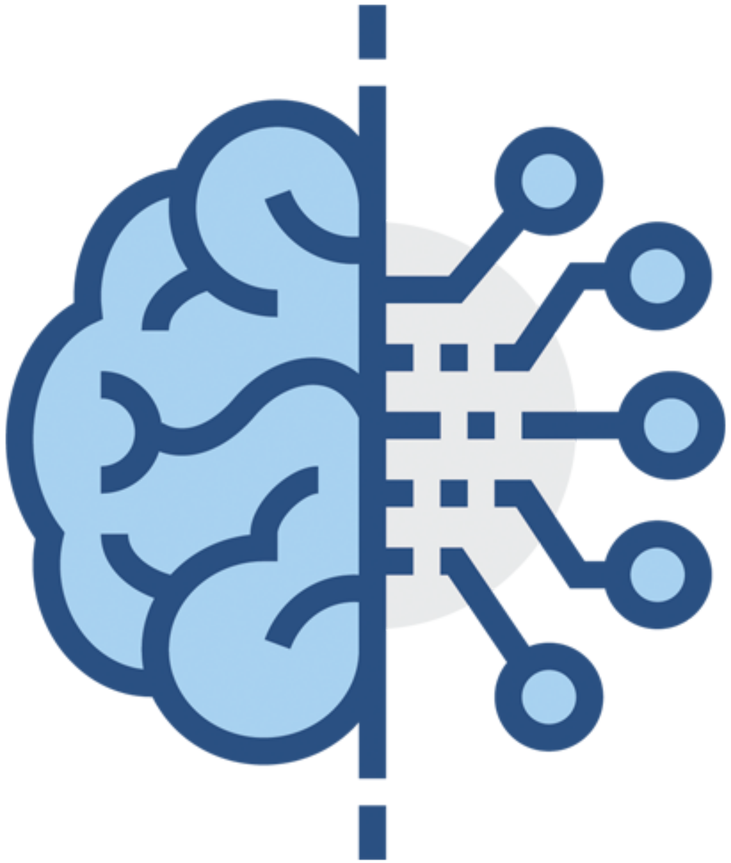




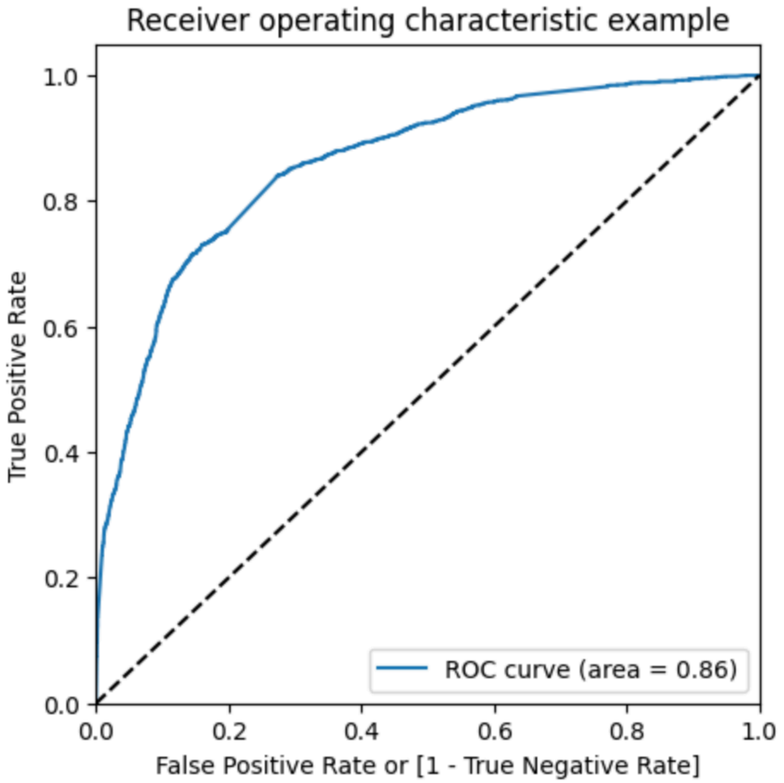


BUILDING LOGISTIC REGRESSION MACHINE LEARNING MODEL

- Splitting into train and test set
- Scale variables in train set
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate variable based on high p-values
- Check for VIF value
- Predict using train set
- Evaluate accuracy and other metric
- Predict using test set
- Precision and recall analysis on test prediction



ROC is 0.86. Which indicate good predictive model.

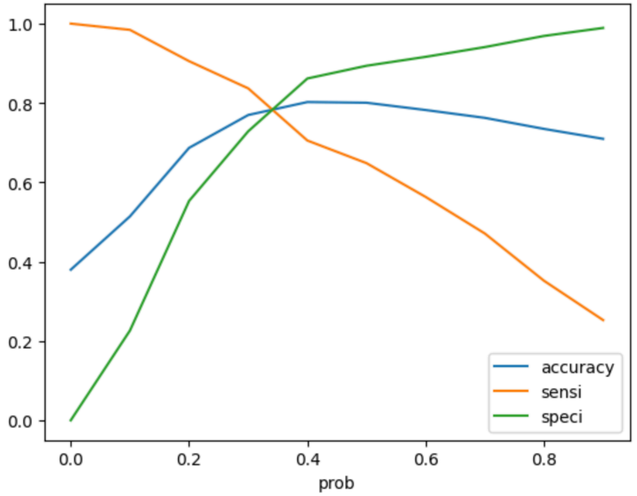


The ROC curve is a statistical tool used in machine learning to analyze the classification models' performance. It shows the difference between true and false positive rates depending on the classification threshold.

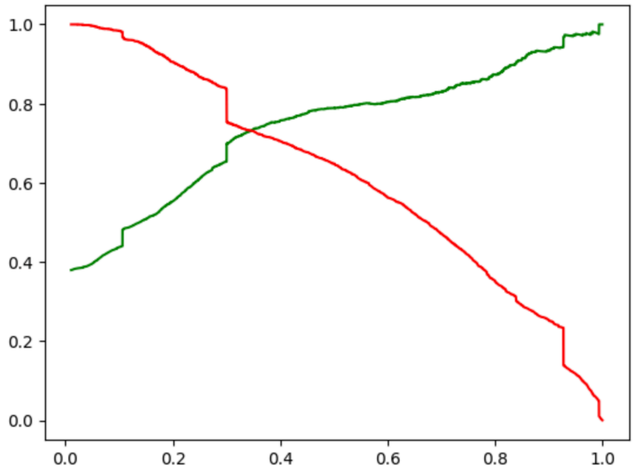
The ROC curve is a measure of the ability of a model to differentiate between classes. It can help determine the model's overall performance and provide insight into how well it can handle diverse datasets.

The ROC Curve should be around one. We are currently getting an excellent value of 0.86. This indicates that the model has a good predictive mode.

Model Evolution Training data set



Precision : 65%
Recall: 83%



Confusion Matrix

2880	1073
395	2024

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

Inference:

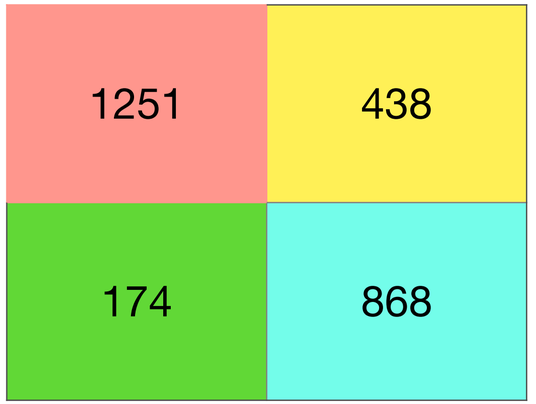
The model is performing well according to the inference shown above. The value of the ROC curve is 0.86. For the Train Data, we have the following values.

- Accuracy : 76.96%
- Sensitivity :83.67%
- Specificity : 72%

Model Evolution Test data set

- **Precision : 65%**
- **Recall: 83%**

Confusion Matrix



Inference:

Finding After running the model on the Test Data:

- Accuracy : 77.59%
- Sensitivity :83.30%
- Specificity : 74.06%

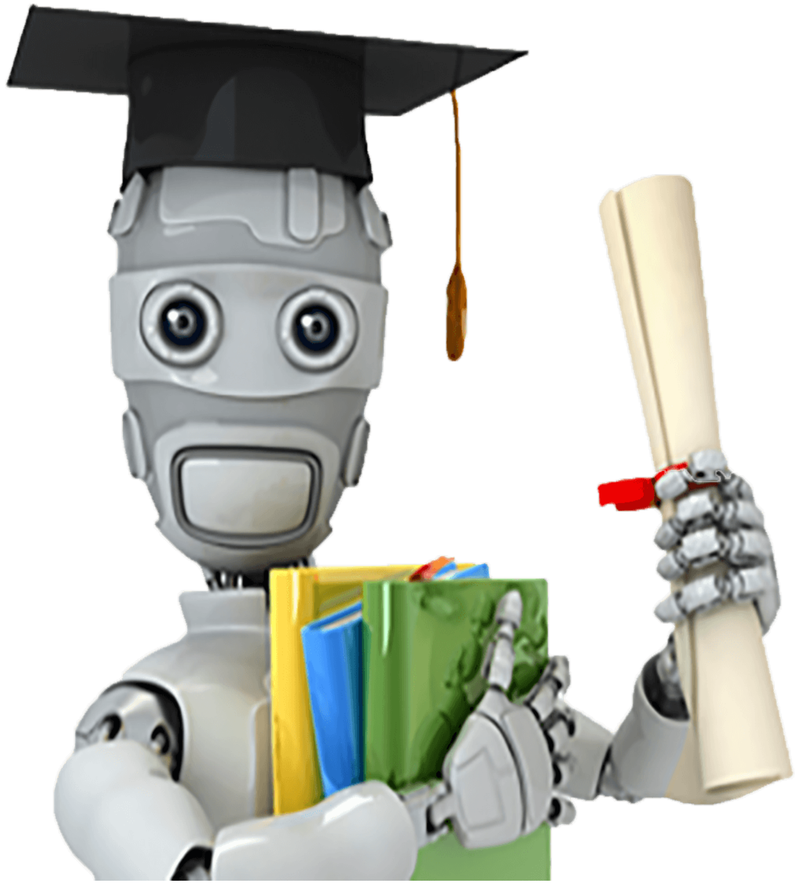
Case Summary

Conclusion:

- We have also performed various other tests such as recall metrics and sensitivity-specificity. For the final prediction, we have decided to use the optimal cutoff based on these two factors.
- The sensitivity, specificity, and accuracy values of the test set are as follows: 77%, 83%, and 74%. These are close to the values that are derived from the trained set.
- The lead score that was obtained from the trained data showed that the predicted model had a conversion rate of around 80%.
- Overall, the model seems to be good.

- The lead score that was obtained from the test set revealed that the model can predict a lead's conversion rate at 83%. This is in line with the expectations of the CEO, who has estimated that the lead conversion rate should be around 80%. The model's sensitivity to the data will help identify the most promising leads.
- The top 3 variables that contribute for lead getting converted in the model are :

1. **Total time spent on website**
2. **Lead Add Form from Lead Origin**
3. **Had a Phone Conversation from Last Notable Activity**



THANK YOU