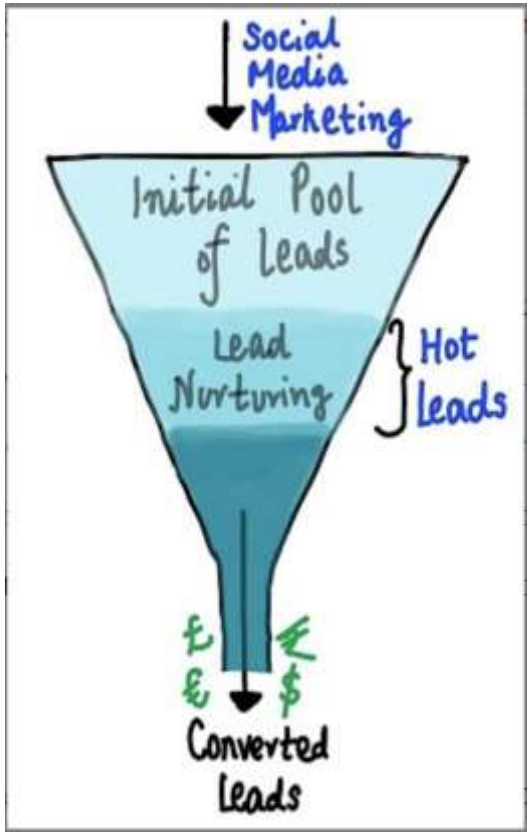
LEAD SCORING CASE STUDY

BY- SHRUTI JAIN
VAIBHAV KOHLI
KIRAN K

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 typical lead conversion rate at X education is around 30%.
- Now, 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.



BUSINESS OBJECTIVE:

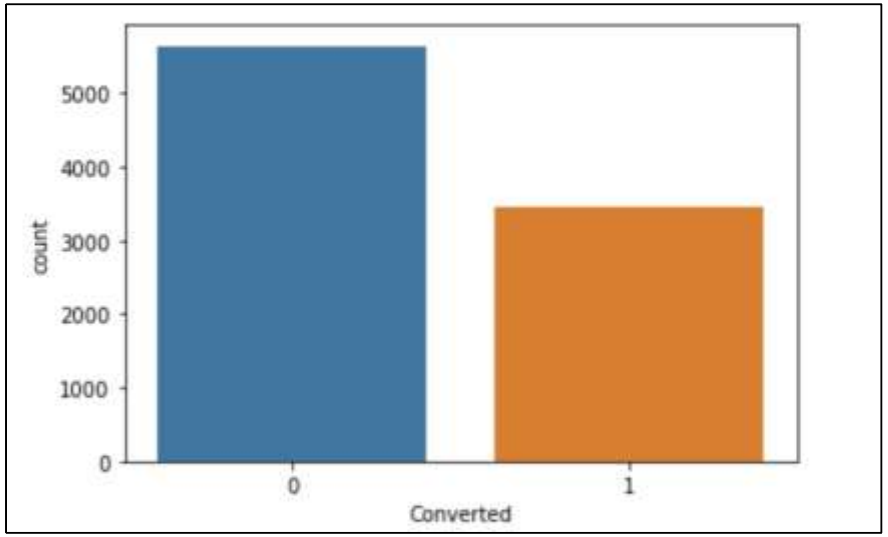
- To build a model in selecting the most potential leads(‘Hot Leads’) whose lead conversion rate is around 80%.
- In this model, a lead score is assigned to each of the leads in such a way that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

APPROACH:

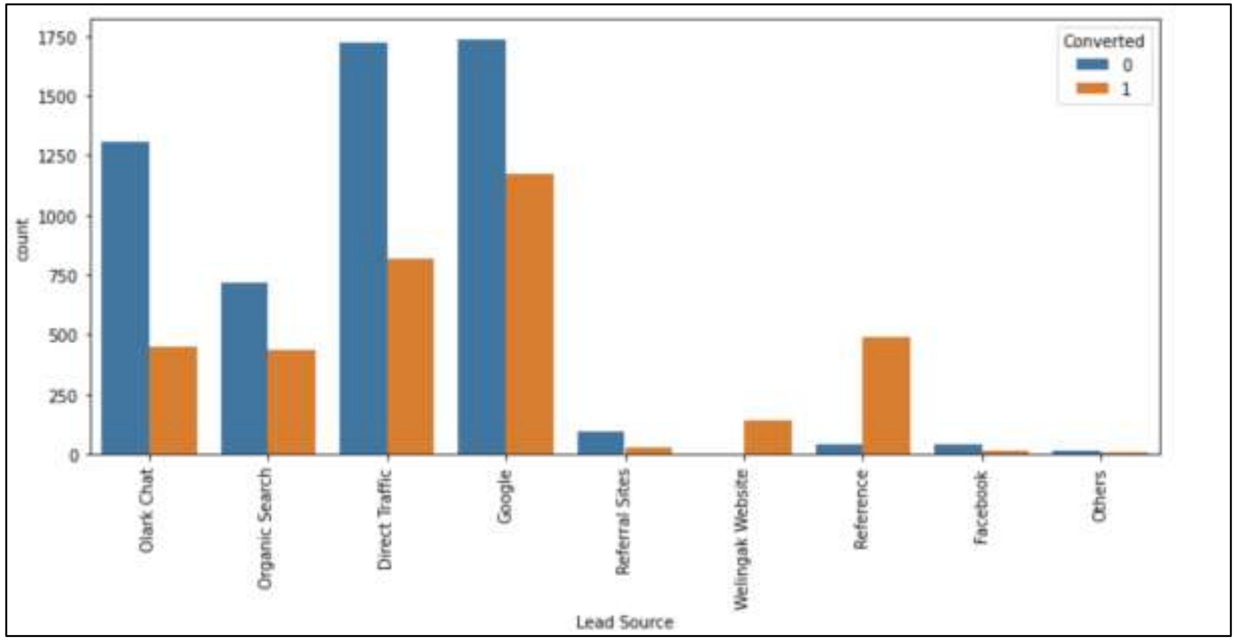
- Data cleaning and manipulation
  - ➢ Handling missing, null values
  - ➢ Handling ‘Select’ values
  - ➢ Drop columns which aren’t useful
  - ➢ Imputation of the data
  - ➢ Handling Outliers
- EDA
- Dummy Variables
- Feature Scaling
- Building model using Logistic Regression
- Model Evaluation
- Conclusion

DATA INSIGHTS:

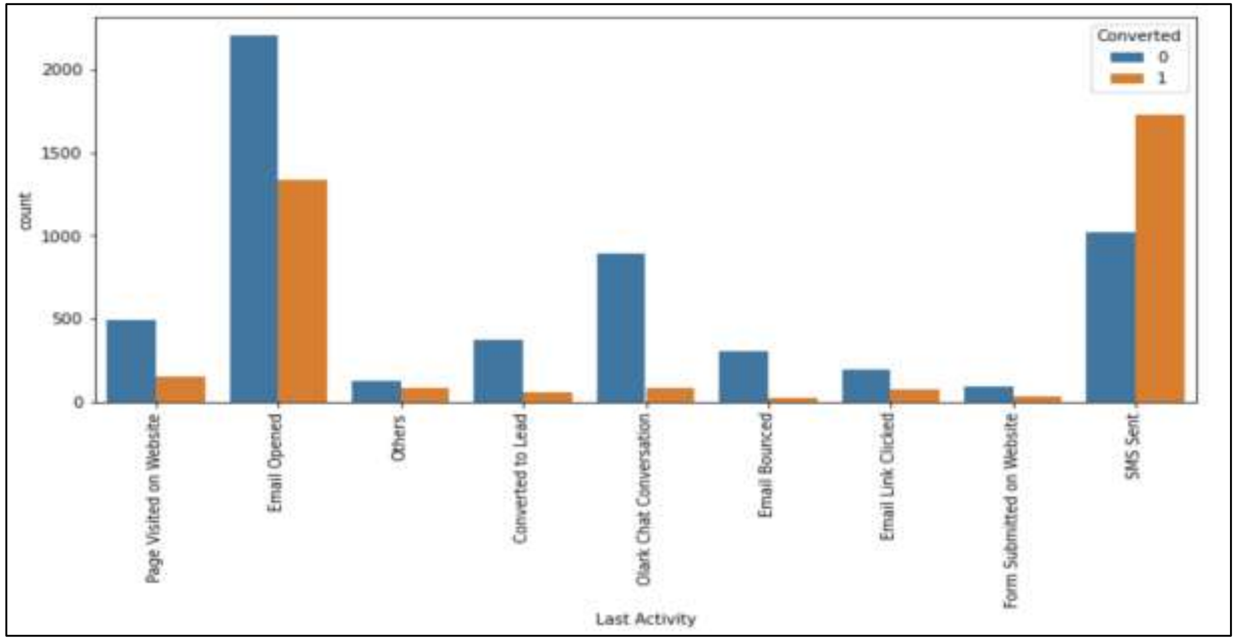
- Converted: Potential Leads can be identified on the basis of Leads Score (which is probability of leads getting converted). Out of 9240 entries we see that around 37% of leads are converted and 73% of leads are not converted.



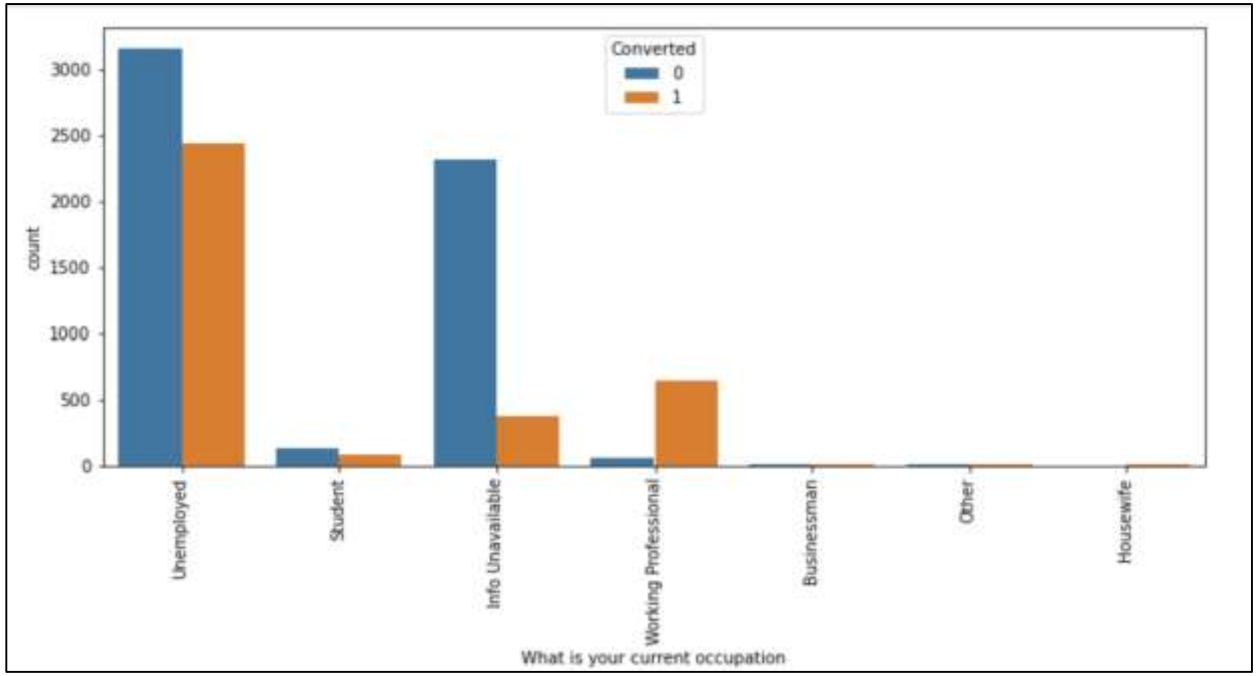
- Lead Source: Majority source of the lead is Direct Traffic & Google. Leads from Google and Reference have more probability of conversion.



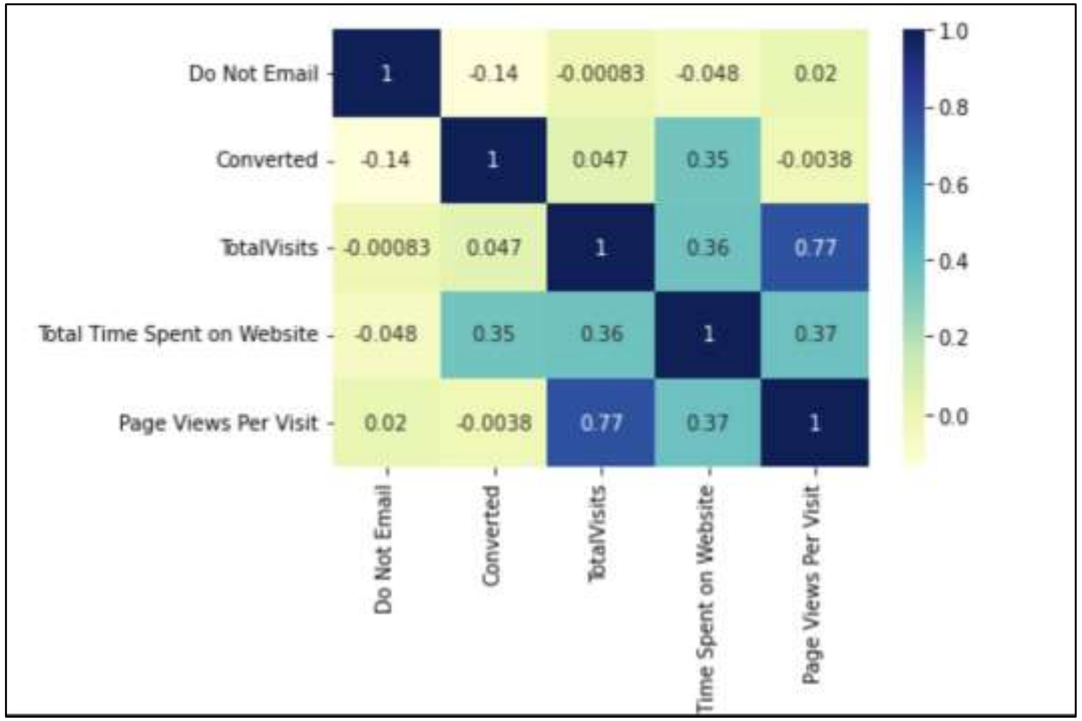
- Last Activity: Customers whose last activity was SMS Sent have higher conversion rate. Customers who last activity was Email Opened constitute majority of the customers and have less conversion rate.



- What is your current occupation: Majority of the leads are from Unemployed and they have more probability of getting converted. However, Working Professionals have the highest probability of getting converted.



- Correlation among the Numeric columns: We can observe that the variables are not highly correlated with each other. But there is multicollinearity among some features



MODEL BUILDING:

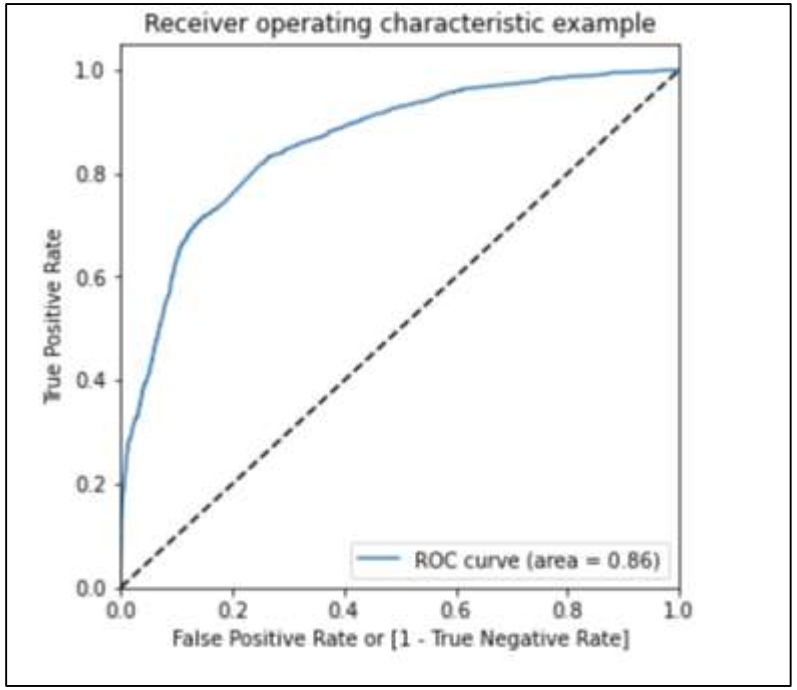
- Splitting the data into Train and Test sets(70:30) to check model stability.
- Use RFE for feature selection with top 17 variables as the required output.
- Building Model by removing the variable whose p- value is greater than 0.05 and VIF is greater than 5
- Prediction on the Test set.

- Below are the features responsible for Lead Conversion.
    - ➢ *Lead Source_Olark Chat*
    - ➢ *Lead Origin_Lead Add Form*
    - ➢ *Lead Source_Welingak Website*
    - ➢ *What is your current occupation_Unemployed*
    - ➢ *What is your current occupation_Info Unavailable*
    - ➢ *Total Time Spent on Website*
    - ➢ *Do Not Email*
    - ➢ *What is your current occupation_Student*
    - ➢ *Lead Origin_Lead Import*
    - ➢ *What is your current occupation_Other*

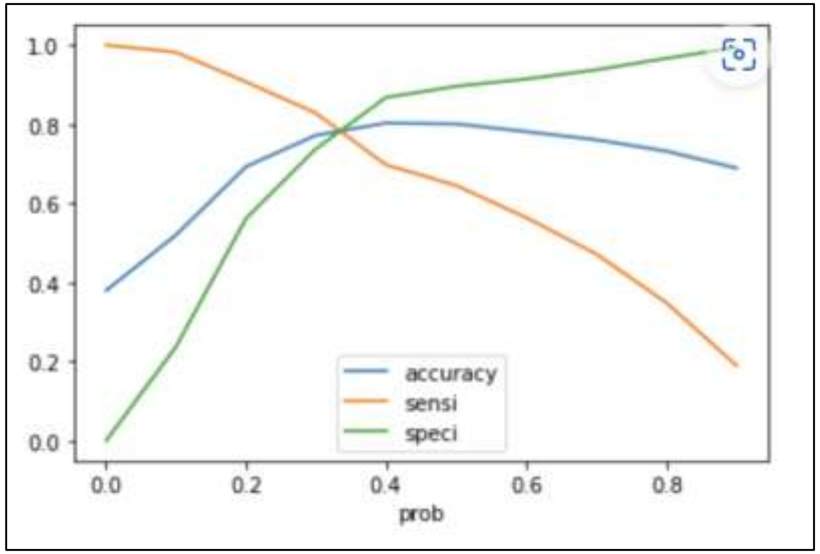
ROC CURVE & OPTIMAL CUT-OFF POINT:

- ROC Curve represents how much the model is able to distinguish between the classes.

- We can see that the Area under Curve is 0.86 which means almost 86% of the times, the model is able to distinguish the 1’s and 0’s



- Upon plotting the Accuracy, Sensitivity, Specificity of the model together, we can find out the optimal cut-off point which is around 0.3.
- With this proboablity of 0.3, we will predict the y-values from the X train data such that any conversion probability of greater than 0.3 will be converted to the lead



MODEL EVALUATION:

We get the below metric upon running the model on the feature selected.

1. **Train Dataset:**

   Accuracy: 77%

   Sensitivity: 84%

   Specificity: 73%

   Precision: 65%

   Recall: 84%

2. **Test Dataset:**

   Accuracy: 77%

   Sensitivity: 83%

   Specificity: 73%

   Precision: 66%

   Recall: 83%

The Model seems to predict the conversion rate well and should be helpful for the education company to identify the Hot Leads.

CONCLUSION:

- The customer who fills the form are the potential leads.
- We must majorly focus on working professionals.
- If the lead source is referral, they may not be the potential lead.
- We must majorly focus on leads whose last activity is Email opened or SMS sent.
- It’s better to focus least on customers to whom the sent mail is bounced back.
- It’s always good to focus on customers, who have spent significant time on our website.
- If the customer didn’t fill specialization, they may not know what to study and are not right people to target. So, it’s better to focus less on such cases.