LEAD SCORE CASE STUDY

BY:

NIKHIL FRASER

DEEPJYOTI CHAKRABORTY

PROBLEM STATEMENT

- The company X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google and the people who land on the website might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%. Now the company wishes to identify the most potential leads, also known as ‘Hot 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.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%. To make this process effective let's build logistic regression model, the conversion predictions and evaluation metrics to achieve this.

ANALYSIS APPROACH

1. Reading and Understanding the data with the help of data dictionary.
2. Importing the required libraries and inspecting the data.
3. Data cleaning and manipulation:

- Replacing the 'Select' with NaN for not effecting the analysis.
- Dropped the columns with only one unique class and columns with index numbering which doesn't help for the analysis
- Dropped the columns which have missing values more than 35% but managed to hold few columns though having more missing values as they were seem to be important for the analysis.
- The columns still having missing values were imputed by necessary apt values.
- Outlier treatment was done for necessary continuous variables and few variables were bucketed into 'others' to control the data imbalance.
- Done Univariate analysis and Bivariate analysis wherever required.

4. Data Preparation:

- Converted few required variable into binary classification and created dummy variables for categorical data.
- Divided the data into train set and test set by 70:30 ratio and scaled the data using Standard Scaler.

5. Selected significant features using RFE technique.

6. Built models using logistic regression method and predicted.

7. With Confusion matrix got the Accuracy, Sensitivity, Specificity values and plotted required curves to get the optimal probability cut-off value.

8. With the Precision and Recall values plotted a curve to get a probability cut-off value.

9. With the obtained values predicted on the test set and attained good performance model. And marked the lead score for the leads accordingly.

10. Concluded the model insights and recommendations.

RESULTS:

The model built was with probability cut-off of 0.40 where we got the following for train set:

- Accuracy : 85.45%
- Recall : 82.54%
- Precision : 80.24%

And the following for the test :

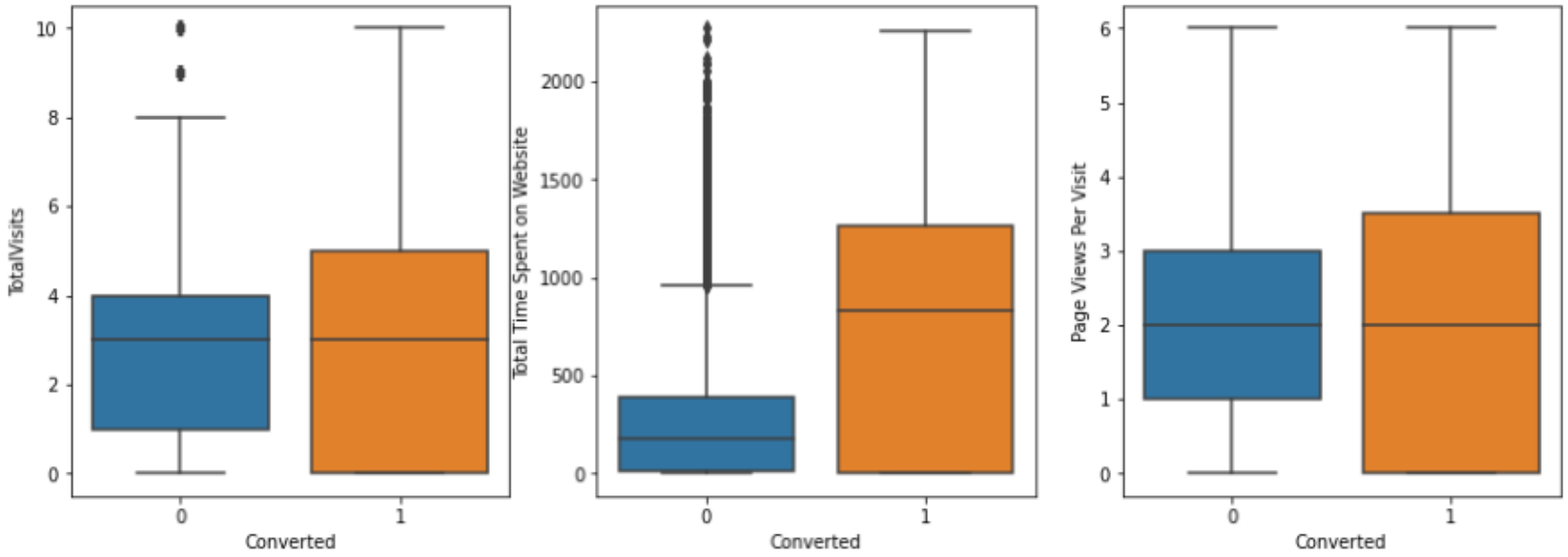
- Accuracy : 84.61%
- Recall : 80.58%
- Precision : 77.83%

By the above results can tell that the model performs well as the recall is 80% and can help the business to not miss a lead though not a Hot Lead.

The precision and accuracy results are also good for the model. This model can predict the lead in way that no slippage of Hot leads for own and for the competitors.

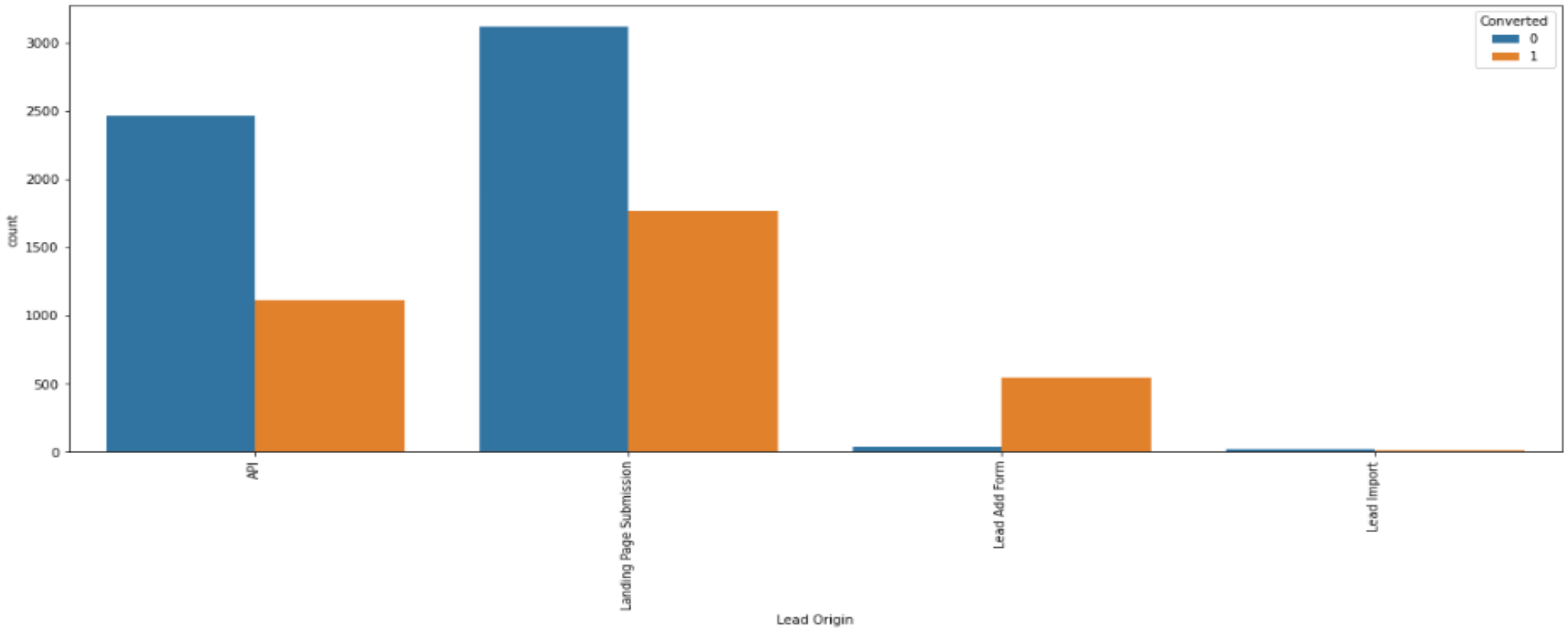
EDA ANALYSIS AND VISUALISATION:

Visualising Total Visits, Total Time spent on Website, Page Views per Visit against Target variable



- Observation - People spending more time on the website are more likely to be converted and has major observation rather than the total vists and page views per vist

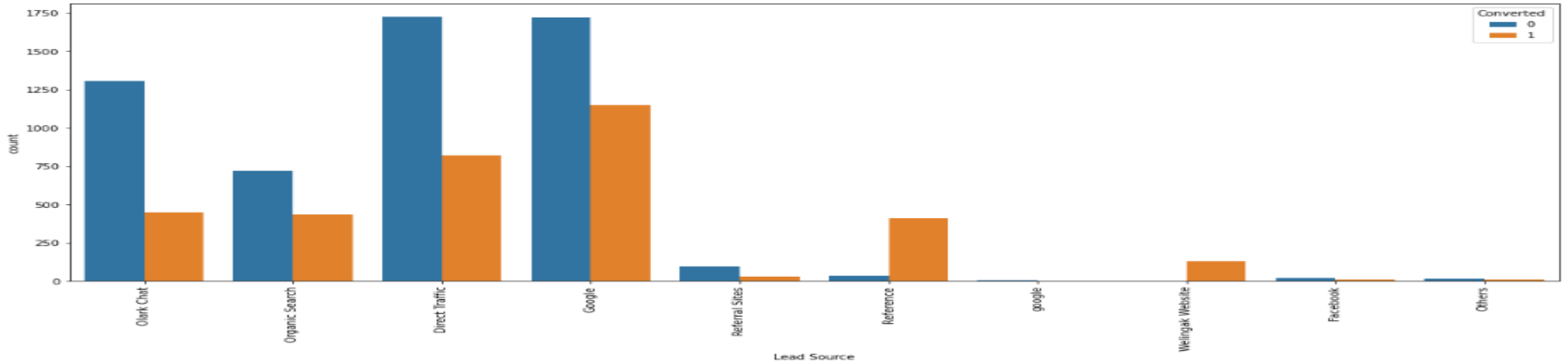
Plot of lead origin against target Variable



- **Observations:**'API' and 'Landing Page Submission' generate the most leads but have less conversion rates of around 30%. Whereas, 'Lead Add Form' generates less leads but conversion rate is great.

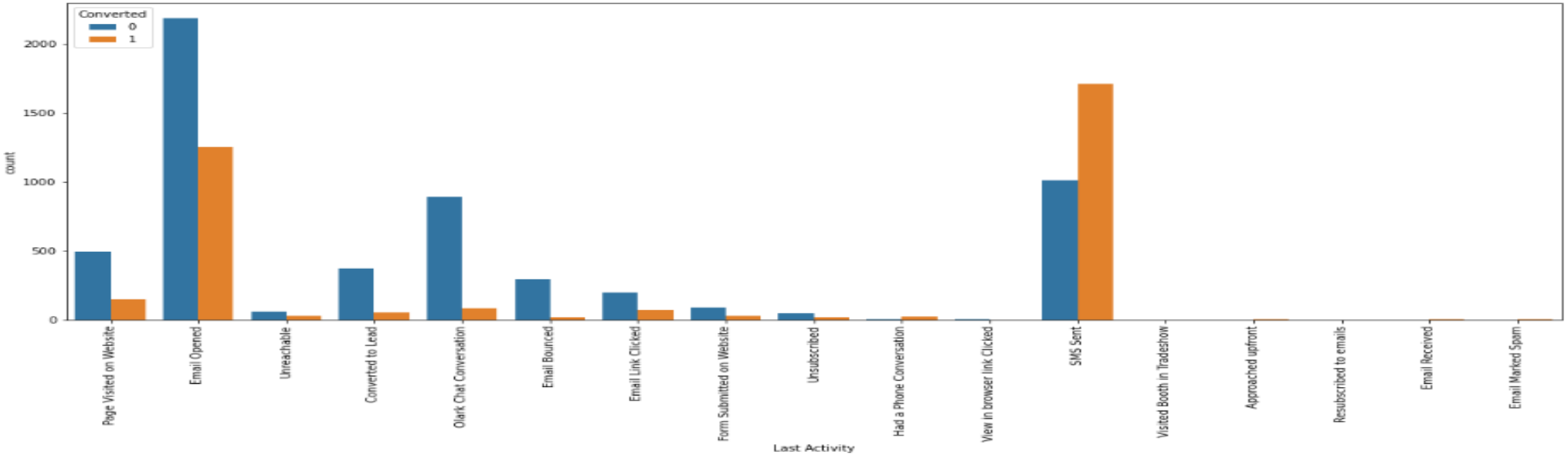
- **Analysis:**
  1. We should try to increase conversion rate for 'API' and 'Landing Page Submission'
  2. Increase leads generation using 'Lead Add Form'. 'Lead Import' does not seem very significant.

Plot of Lead Source against target Variable



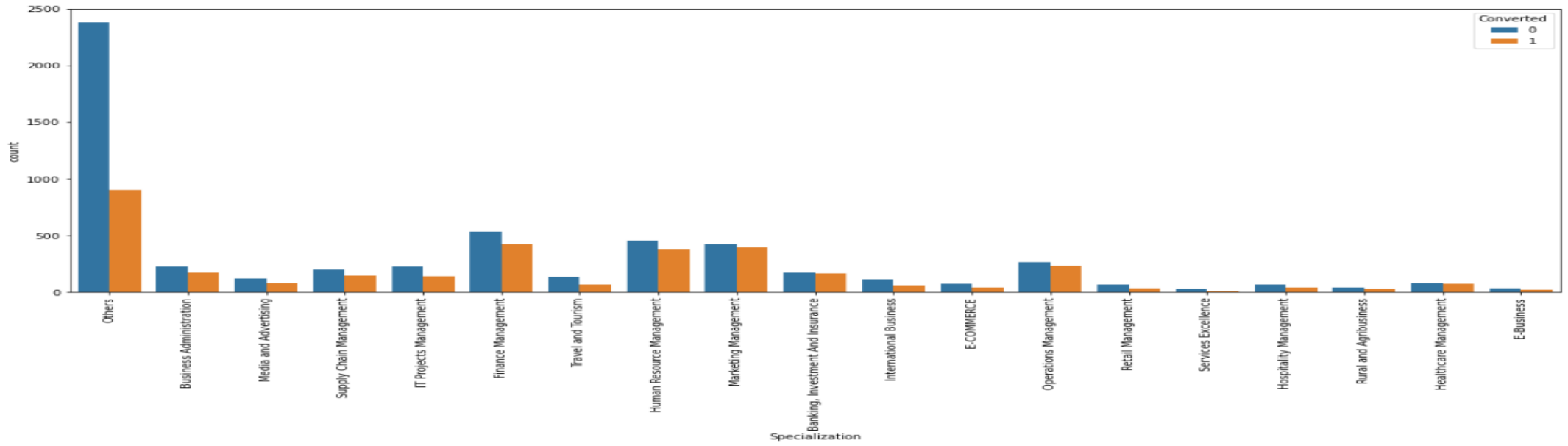
- Observation: 'Direct Traffic' and 'Google' generate maximum number of leads while maximum conversion rate is achieved through 'Reference' and 'Welingak Website'

Plot of Last Activity against target Variable



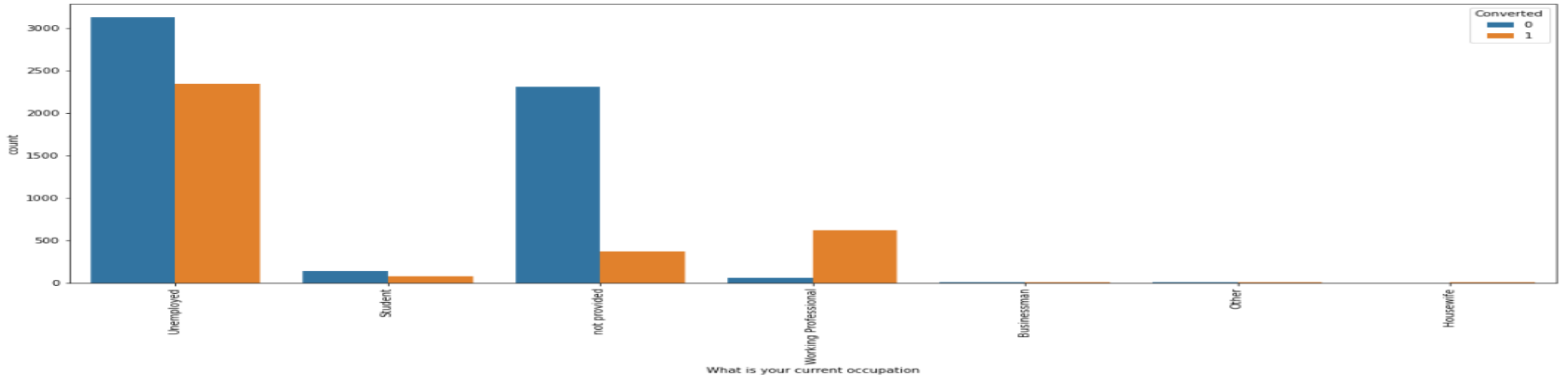
- Observation - Highest number of lead are generated where the last activity is 'Email Opened' while maximum conversion rate is for the activity of 'SMS Sent'. Its conversion rate is significantly high.

Plot of Specialization against target Variable



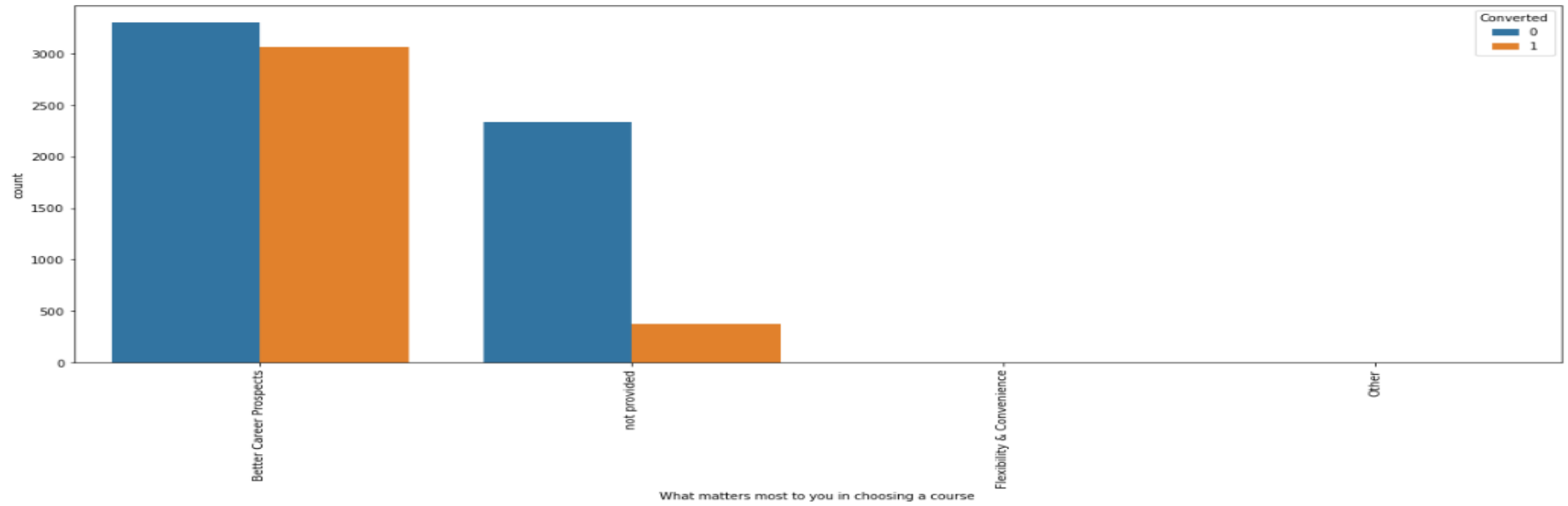
Observation : Got no relevent observation. Conversation rate is almost same in the ratio of total leads generated for different specialization streams.

Plot of Current Occupation against target Variable



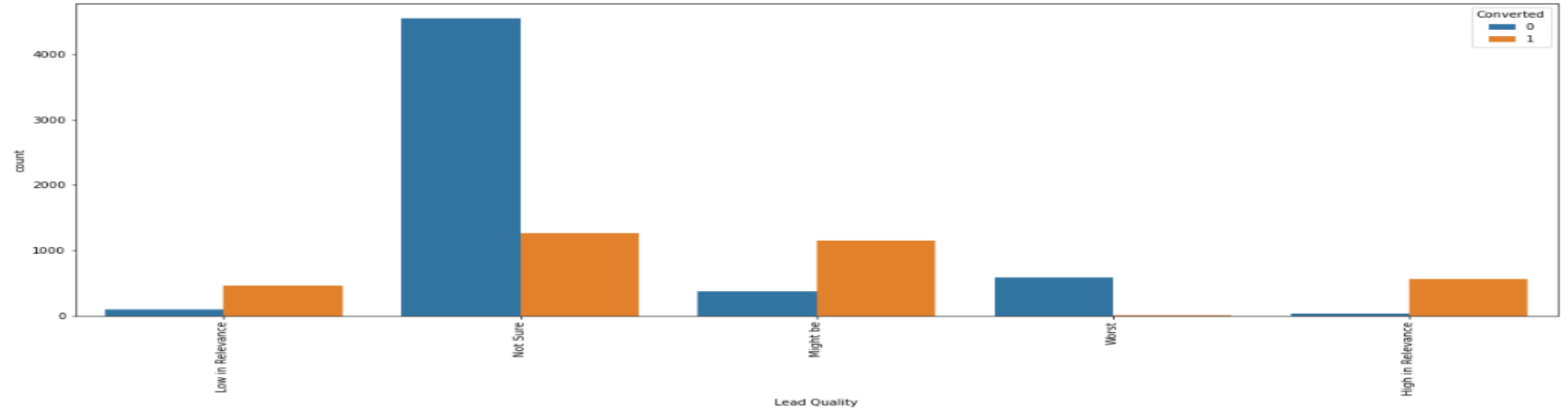
Observation: Convertion rate is very high for the working professionals

Plot of what matters most to you choosing a carrer against target Variable



Observation: Leads generated and Convertion rate is very high for 'Better career Prospects'

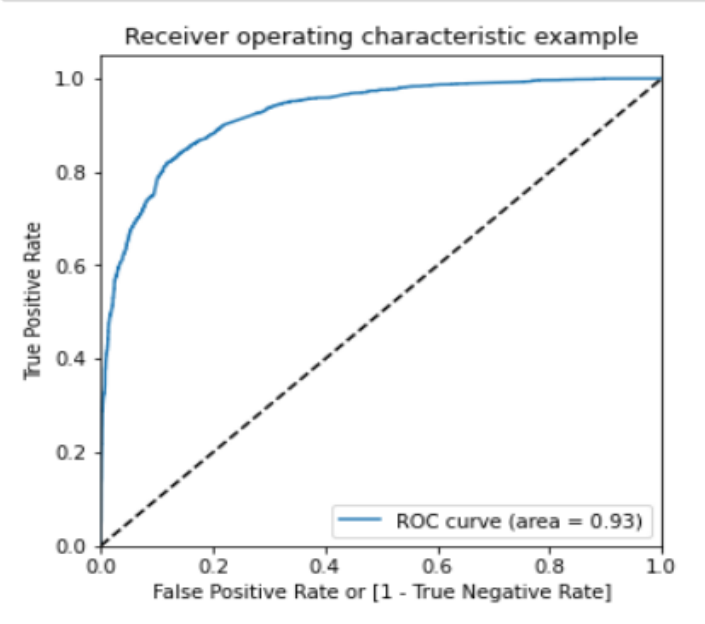
Plot of Lead quality against target Variable

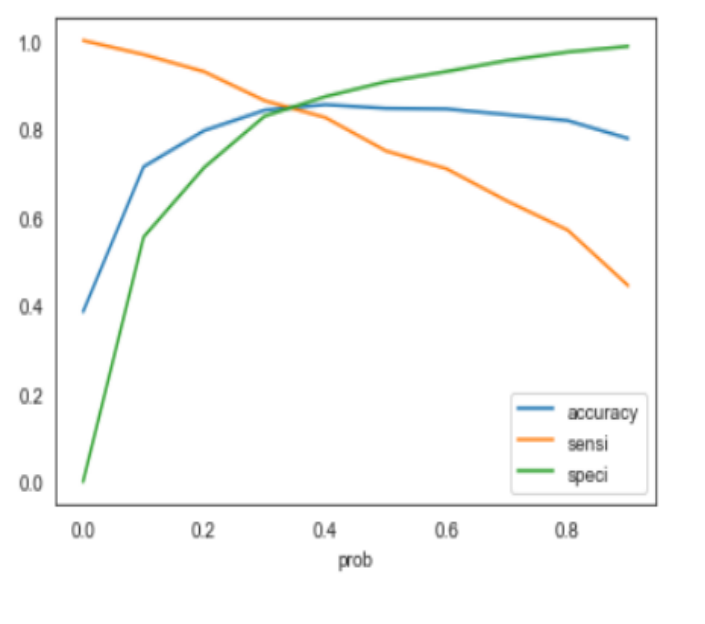


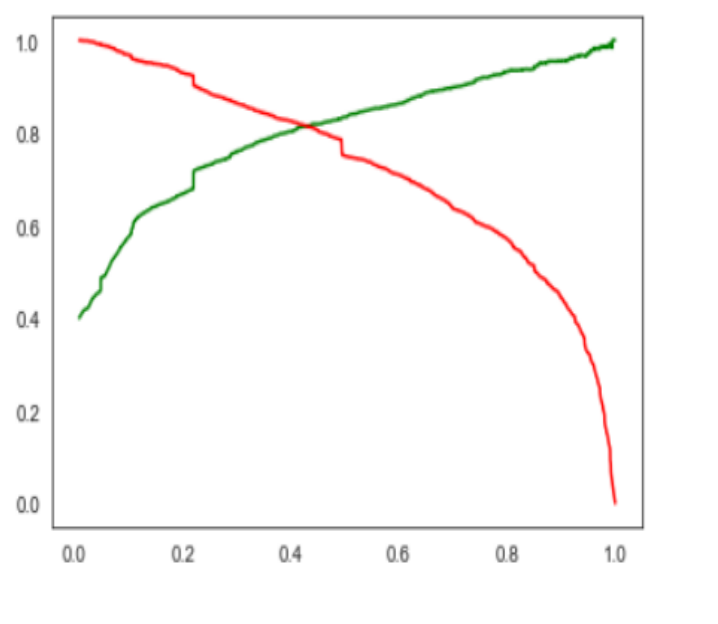
Observation:

- As expected a huge amount of convertion rate can be seen from people marked as 'Might be' & 'High relevence'. And low convertion rate can be noticed in 'Worst' marked data.
- Suprisingly we can see a good convertion rate for the 'Low Relevance' marked people.

Plots of (ROC curve), (accuracy sensitivity specificity) & (precision and recall)

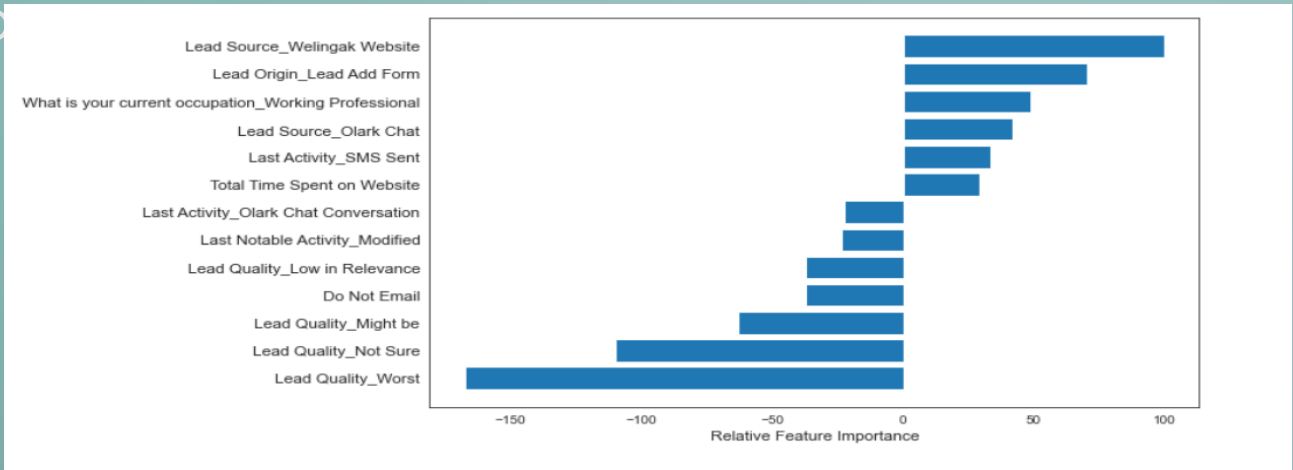






- The Area under curve (auc) is approximately 0.925 which is very close to ideal auc of 1.
- From the accuracy sensitivity specificity plot the optimum point is 0.35.
- From the Precision & Recall plot the optimum point for the probability cut-off is 0.4

CONCLUSION AND RECOMMENDATIONS:



•The positive features that a lead converts into hot leads are the following:
Leads sourced from Welingak Website.
Leads originated from Lead Add Form.
Leads whose current occupation is Working Professional.
Leads sourced from Olark Chat.
Whose last activity is SMS sent.
Leads who spend most time on the website.

•The negative features that have impact on conversion rate:
Leads marked by the assigned employee as Worst Quality.
Leads marked by the assigned employee as Not Sure.
Leads marked by the assigned employee as Might be.
Customers opted not to be emailed about the course (Do Not Email).
Leads marked by the assigned employee as Low in Relevance.
If the Last Notable Activity by the student is Modified.
If the Last activity performed by the customer is Olark Chat Conversation.

The X Education can increase the lead conversion rate by focusing the above factors and improving the negative impacting factors towards the positive approach. The company can increase the target lead conversion rate to be around 80% by using the model with threshold of 0.4 as per the CEO's inclination.