Lead Scoring Case Study

Presented by:-
Deeksha Shrivastava
DCS59 EPGP

Problem Statement

- X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google. When people land on the website, browse the courses and fill up a form providing their email address or phone number, they are classified to be a lead.
- 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

- X Education wants to know the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein we 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 target lead conversion rate is around 80%.

Analysis Approach

- Loading and Inspecting the Data
- Data Quality Check and Data Cleaning
  - > checking duplicates,
  - > checking null values,
  - > dropping unique columns,
  - > imputing missing values
- Data Understanding and Imbalance percentage
- Exploratory Data Analysis (EDA)
  - > Univariate Data Analysis – checking value count,vizualise data distribution of variable, analysis of categorical and numerical variables
  - > Bivariate Data Analysis - correlation coefficient, pattern between the variables

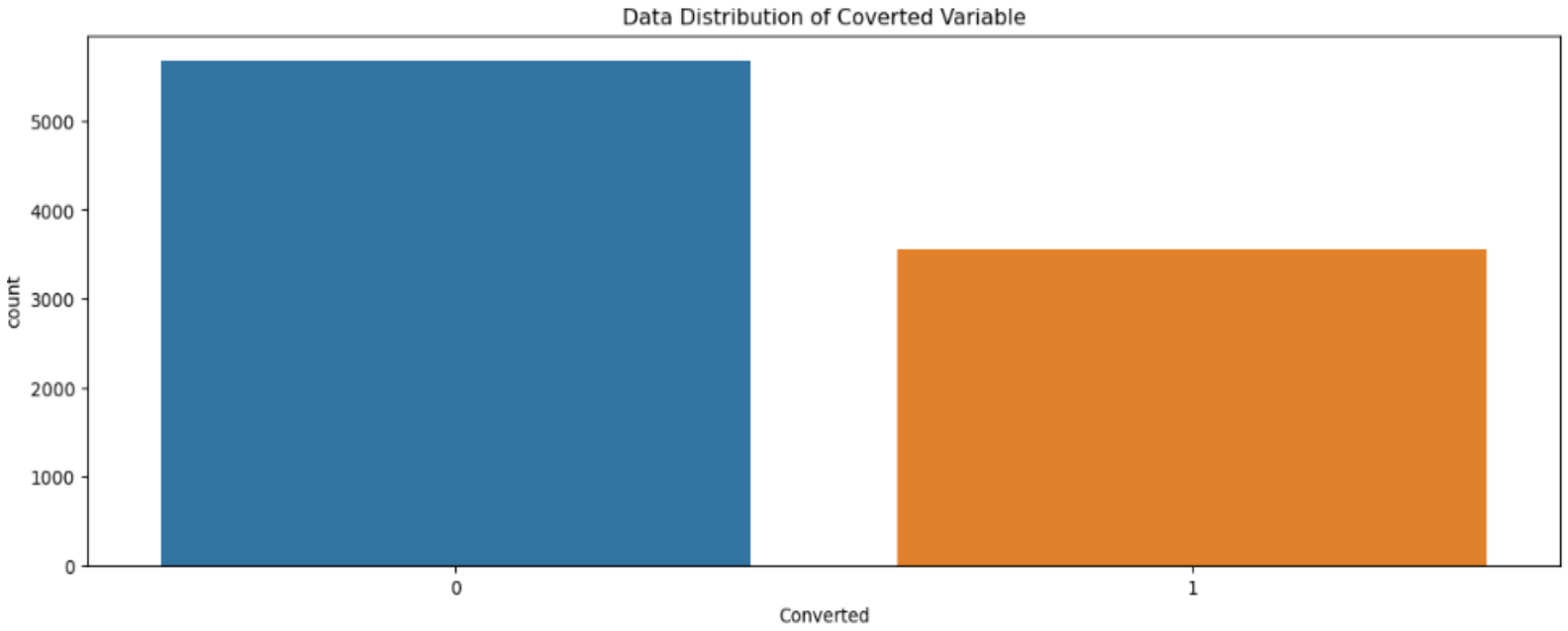
Analysis Approach

- Data Preparation –Dummy variables
- Test-Train Split
- Feature Scaling-checking correlation matrix
- Model Building- using stats model, feature selection using RFE , using manual feature selection
- Check VIF and p-value,
- Making predictions on train set
- Model Evaluation – confusion matrix, accuracy, sensitivity, specificity, recall, precision, precision and recall tradeoff
- Plotting the ROC Curve
- Finding Optimal Cutoff Point
- Making predictions on the test set
- Final Observation

Variables contributing in Lead Conversion

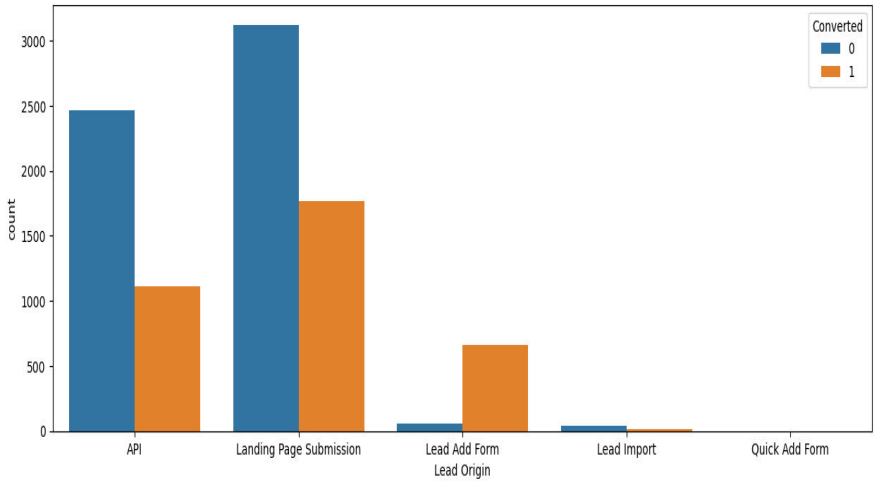
- Specialization_Others
- Lead Source_Olark Chat
- Lead Origin_Landing Page Submission
- What is your current occupation_Other
- Last Activity_SMS Sent
- Lead Origin_Lead Add Form
- Last Activity_Olark Chat Conversation
- Lead Source_Welingak Website
- Total Time Spent on Website
- Do Not Email
- What is your current occupation_Working Professional
- Last Activity_Unsubscribed
- Last Activity_Converted to Lead

Exploratory Data Analysis



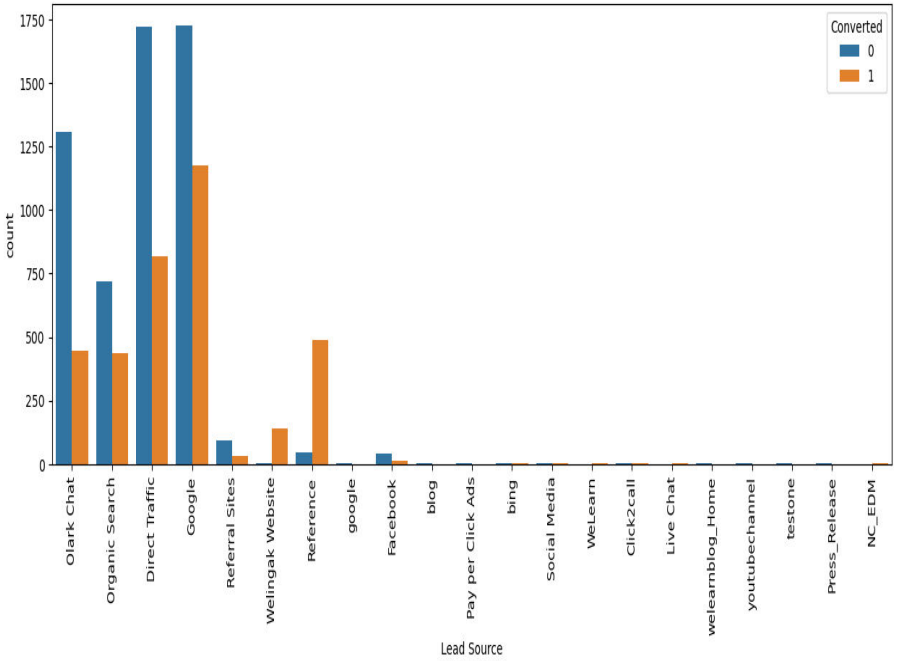
Target Variable - 'Converted'
O - 'Not Converted' – 5679 leads – 61.47%
1 - 'Converted‘ - 3561 leads – 38.53%
This shows we need to increase lead conversion rate.

Analysis of Categorical Variables



Lead Origin Vs Converted

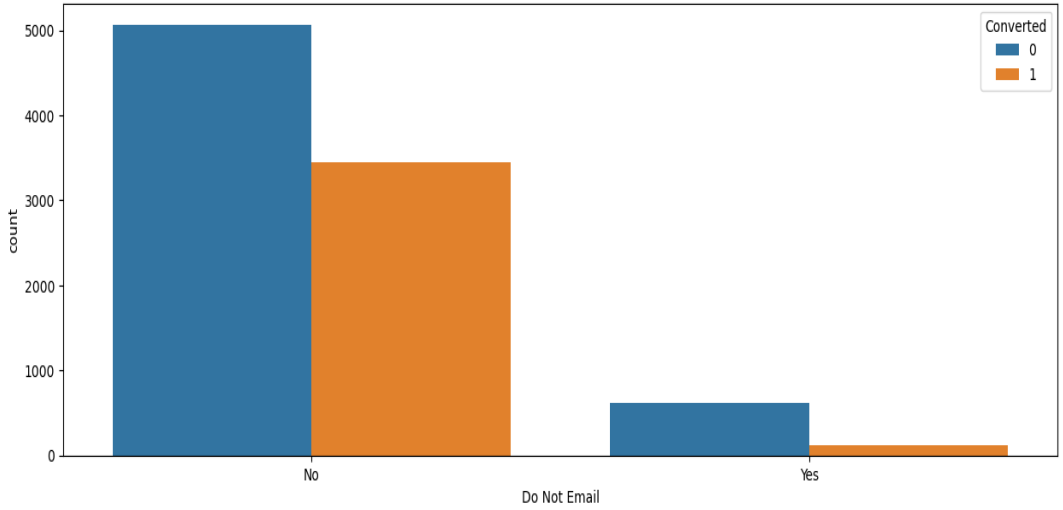
Landing Page Submission has highest lead conversions = 1768



Lead Source Vs Converted

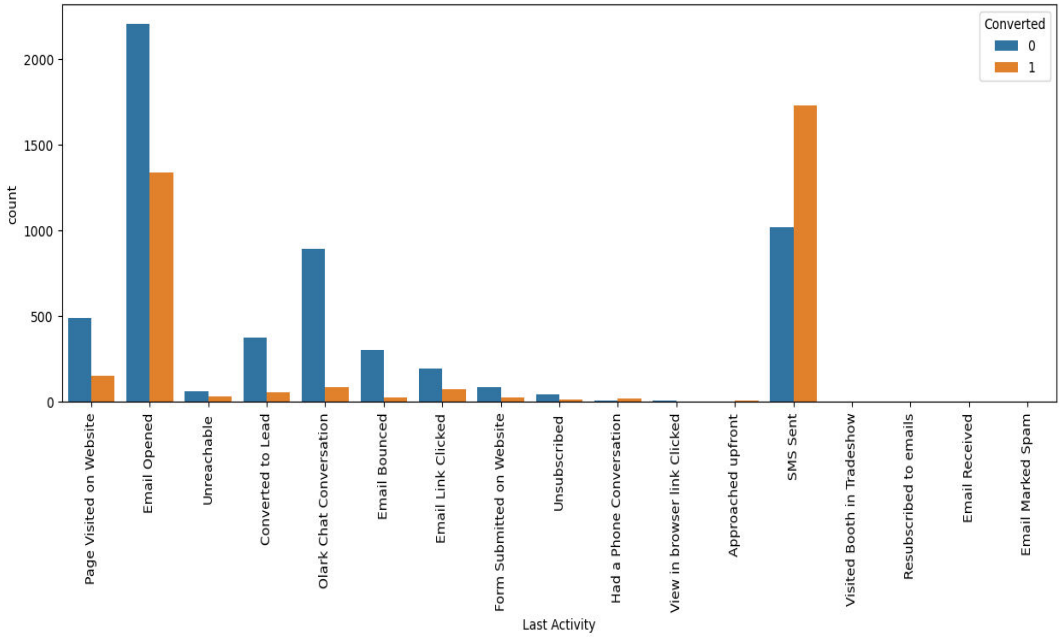
Google has highest lead conversions = 1176 .Welingak Website and Reference have lower leads but very good conversion rate 98% and 91% respectively.

Analysis of Categorical Variables



Do not email Vs Converted

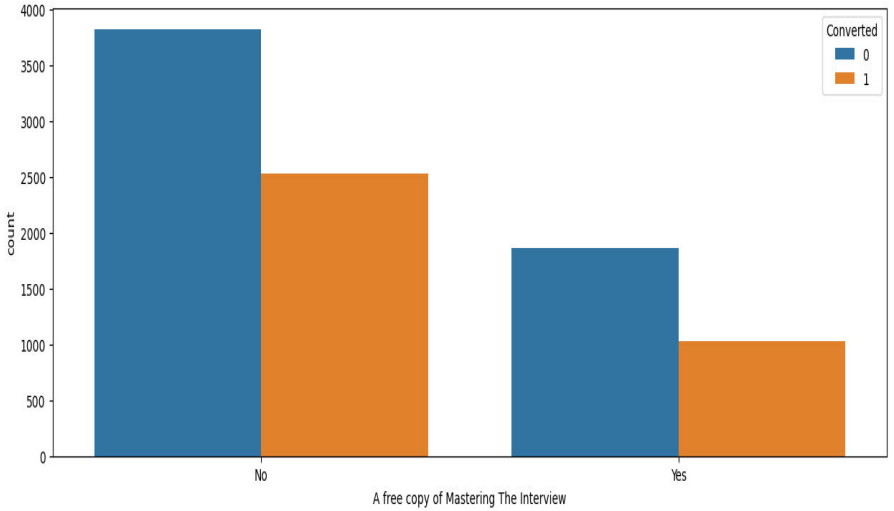
Leads opted for email has higher lead conversions (40%=3443)



Last Activity Vs Converted

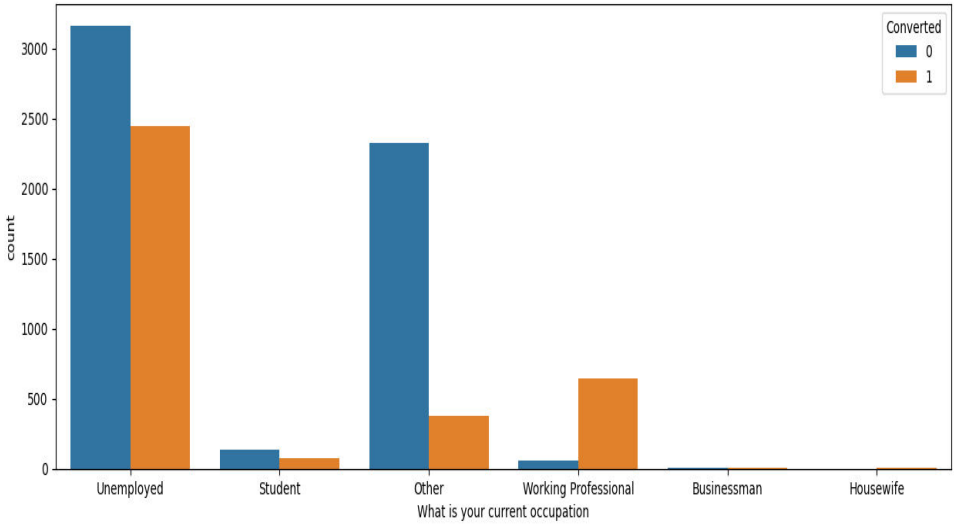
SMS has highest lead conversions(1727),followed by emails(1334)

Analysis of Categorical Variables



Free copy of mastering the interview Vs Converted

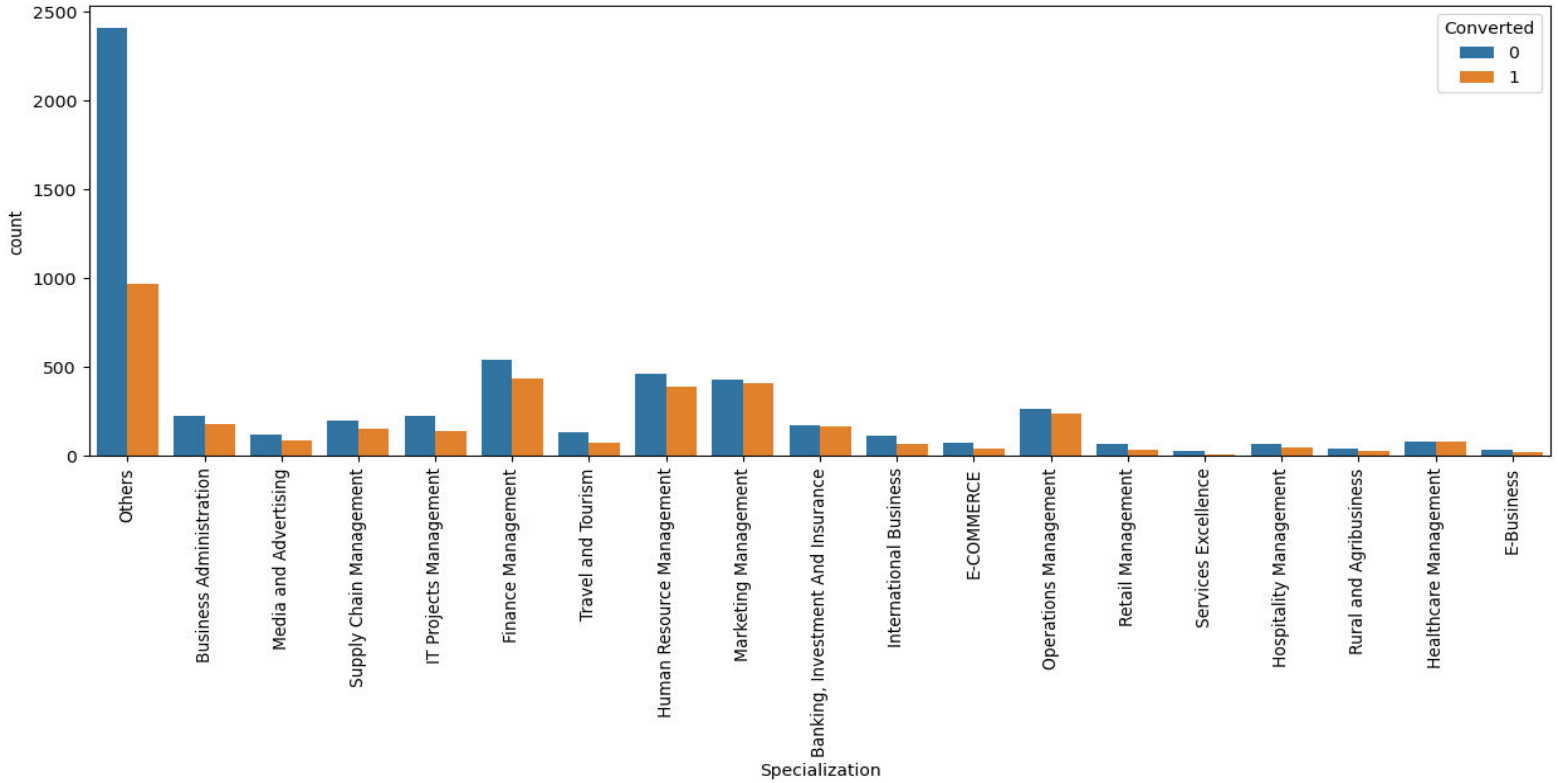
Leads prefer less copies .



current occupation Vs Converted

Unemployed has highest lead conversions(2441). Working Professional has good conversion rate(91%)

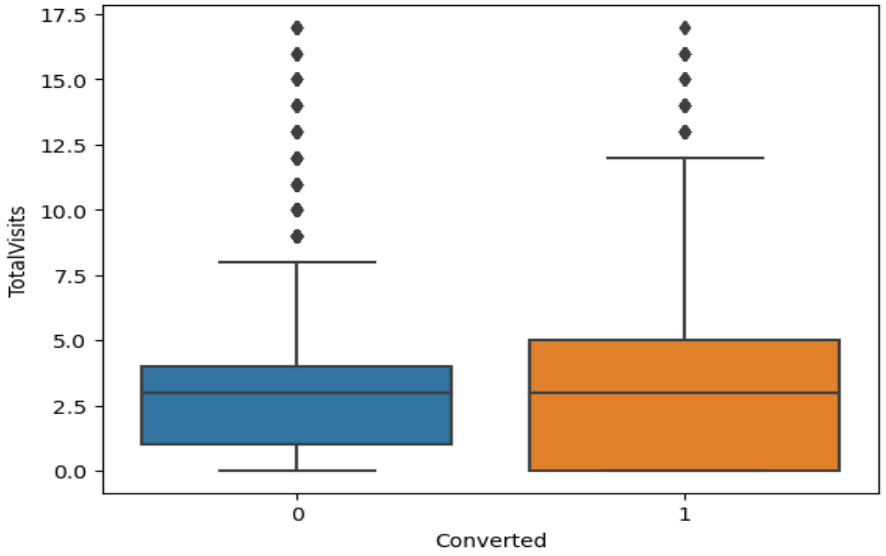
Analysis of Categorical Variables



Specialization Vs Converted

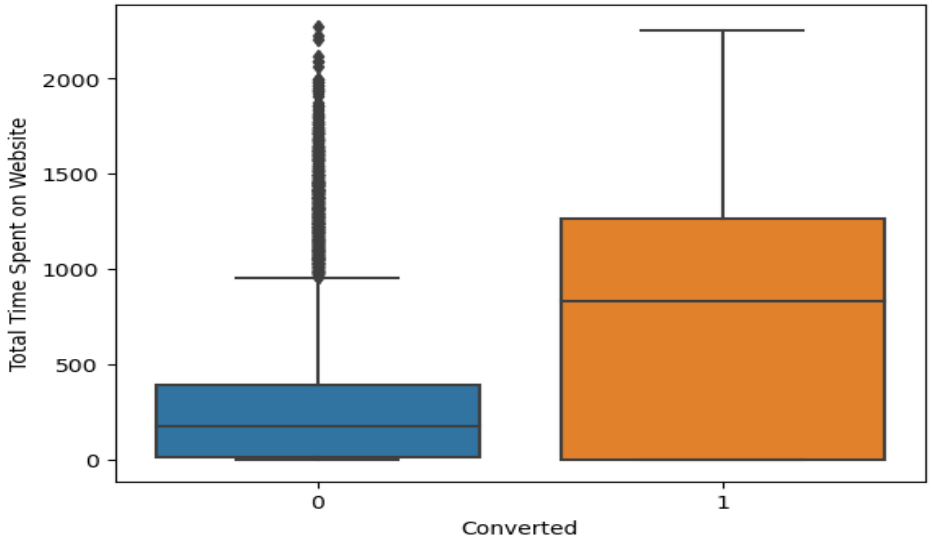
Most leads have no information about specialization have highest conversion(3380)
Finance, Marketing and Human Resource can also be potential leads.

Analysis of Numerical Variables



Total Visits Vs Converted

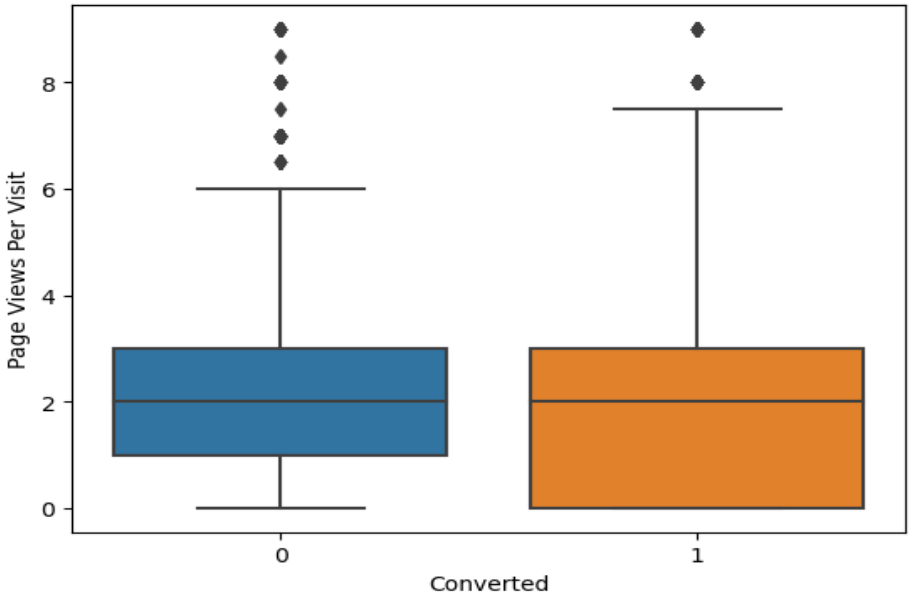
Total visits = 0 has highest number of leads = 2189 and lead conversions = 916. Higher total visits have a slighter higher chance of being a promising lead.



Total Time Spent on Website Vs Converted

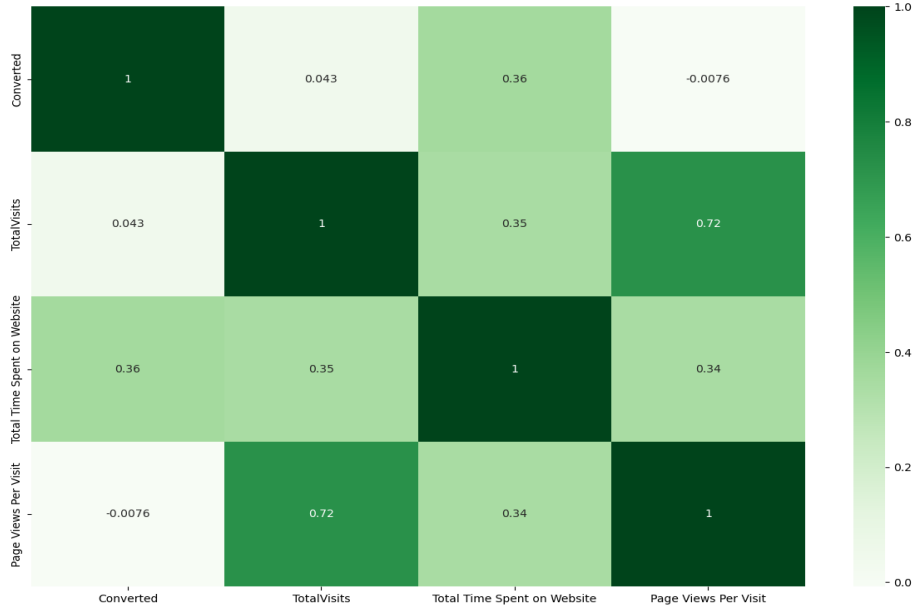
The conversion rate of leads spending time on webiste is really high.

Analysis of Numerical Variables



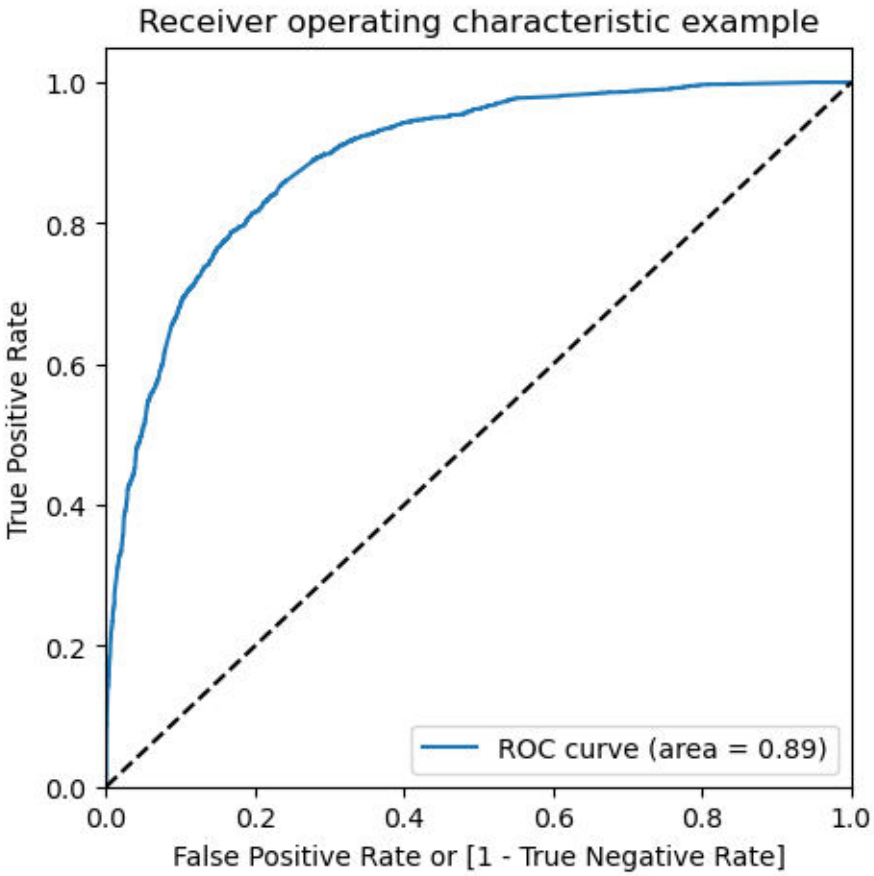
Page views per visit Vs Converted

The median of both converted and non converted is close. Higher page views can be more promising leads



Checking correlations of numerical variables using heatmap

Model Evaluation

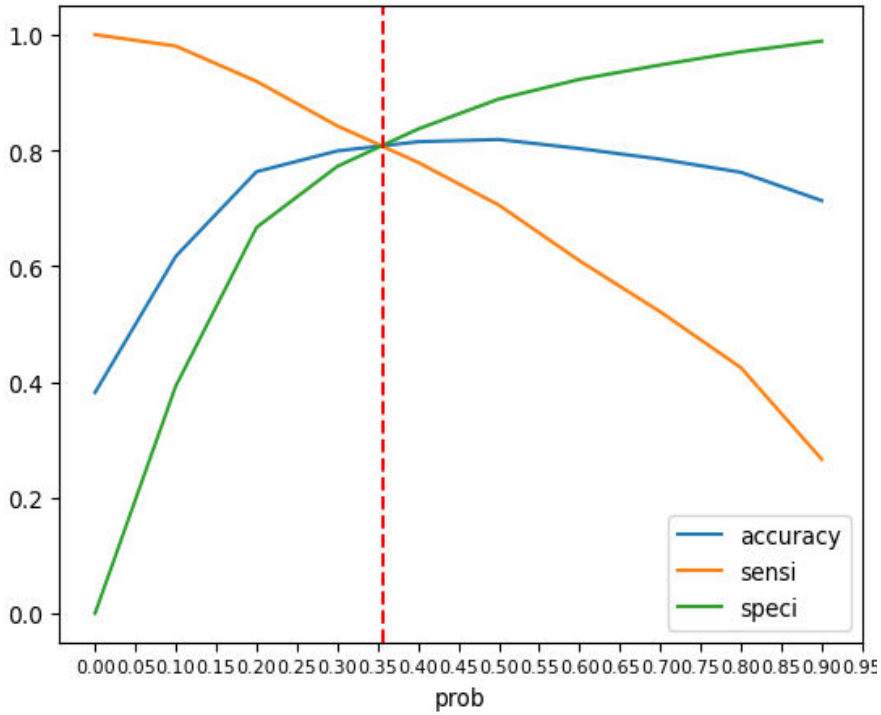


ROC Curve

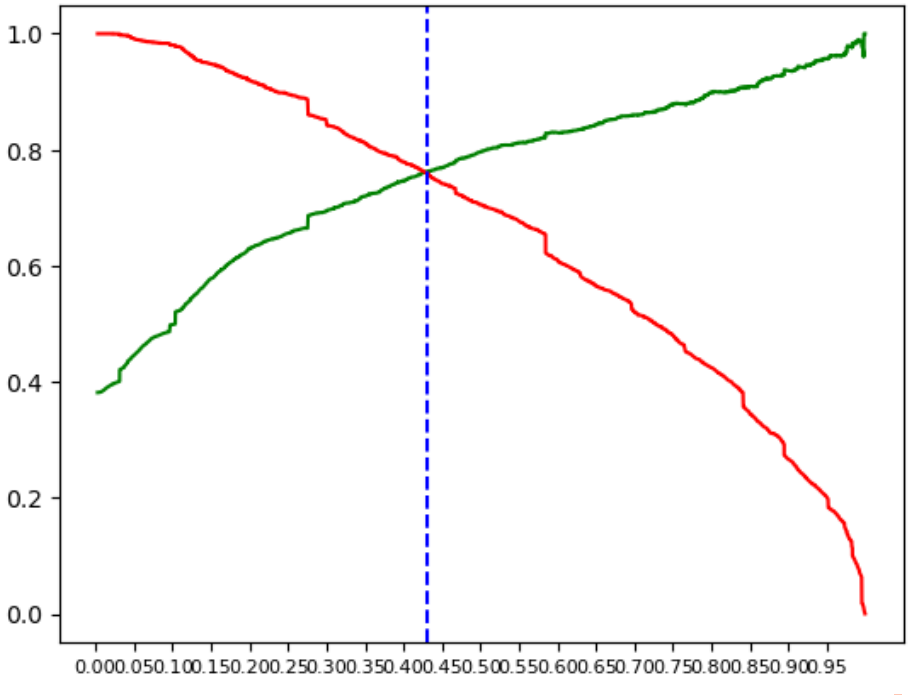
ROC curve value should be close to 1 and we are getting value as 0.89. This shows it is a good predictive model.

Model Evaluation

Optimal Cut off Point



Precision Recall Trade off Curve

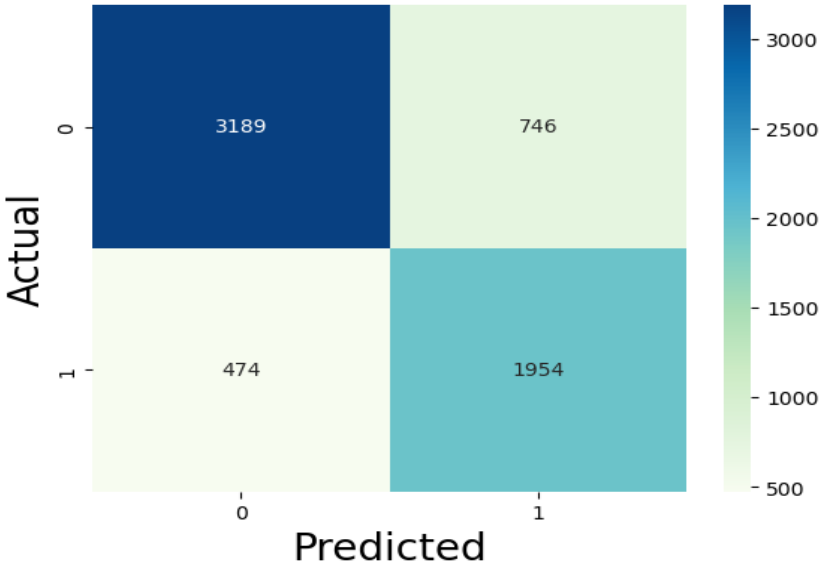


- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity.
- From the plotted curve(left), we have found 0.357 as the best optimal cutoff point.
- From the plotted curve(right), we have found 0.43 as the best optimal cutoff point for precision-recall curve.

Observations

- Train Data Set
- ***Accuracy = 80%***
- ***Sensitivity = 80%***
- ***Specificity = 81%***
- ***Precison = 72%***
- ***Recall = 80%***

- Test Data Set
- ***Accuracy = 80%***
- ***Sensitivity = 80%***
- ***Specificity = 80%***
- ***Precison = 72%***
- ***Recall = 80%***



Confusion Matrix

Actual/Predicted	0	1
Not converted 0	3189	746
Converted 1	474	1954

Conclusion

- The top variables in my lead scoring model which contribute most towards the probability of a lead getting converted are:

  >Lead source_welingak website

  >Lead origin_Lead Add Form

  >What is your current occupation_working professional

- We need to focus on improving lead conversion of:

  >Landing Page Submission and API. The conversion rate is 92% in lead add form so we can generate more leads in this field.

  >Direct traffic and Google.The conversion rates in Welingak Website and Reference are 98% and 91% respectively we can generate more leads in these fields.

  > Unemployed and working professionals will be more willing to take a course

- We need to increase reach of SMS and emails to maximum aspirants.
- We should focus on each specialization as it is one of the important aspect of lead conversion.
- Model is performing well with 80% accuracy,sensitivity and specificity.

Thank You