Lead Scoring Case Study using logistic regression

Submitted by :
Akash Mishra
DS C57

Problem Statement

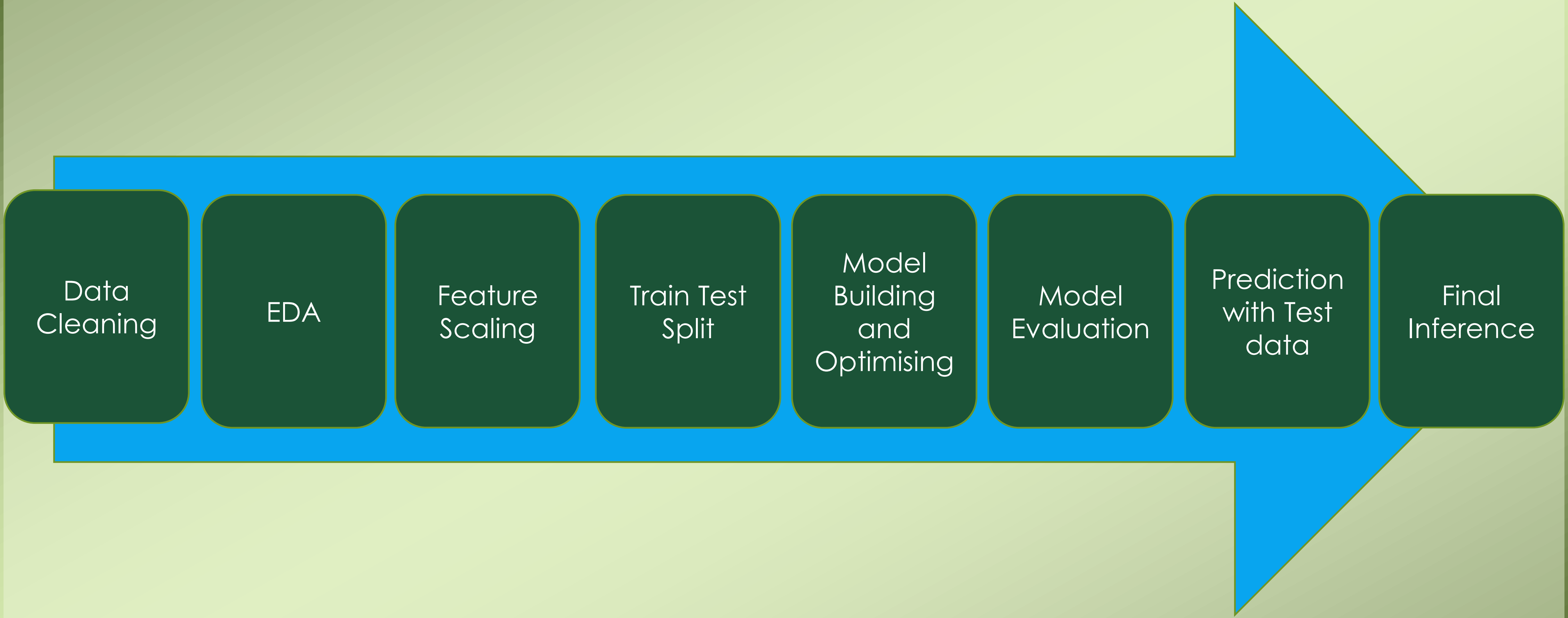
- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they 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%.

Business Goal

- 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.
- The company requires 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 CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Method

- Importing the data and inspecting the data frame
- Clean and prepare the data
- Exploratory Data Analysis.
- Feature Scaling
- Splitting the data into Test and Train dataset.
- Building a logistic Regression model and calculate Lead Score.
- Evaluating the model by using different metrics -Specificity and Sensitivity or Precision and Recall.
- Making predictions on test set



Data Cleaning

EDA

Feature Scaling

Train Test Split

Model Building and Optimising

Model Evaluation

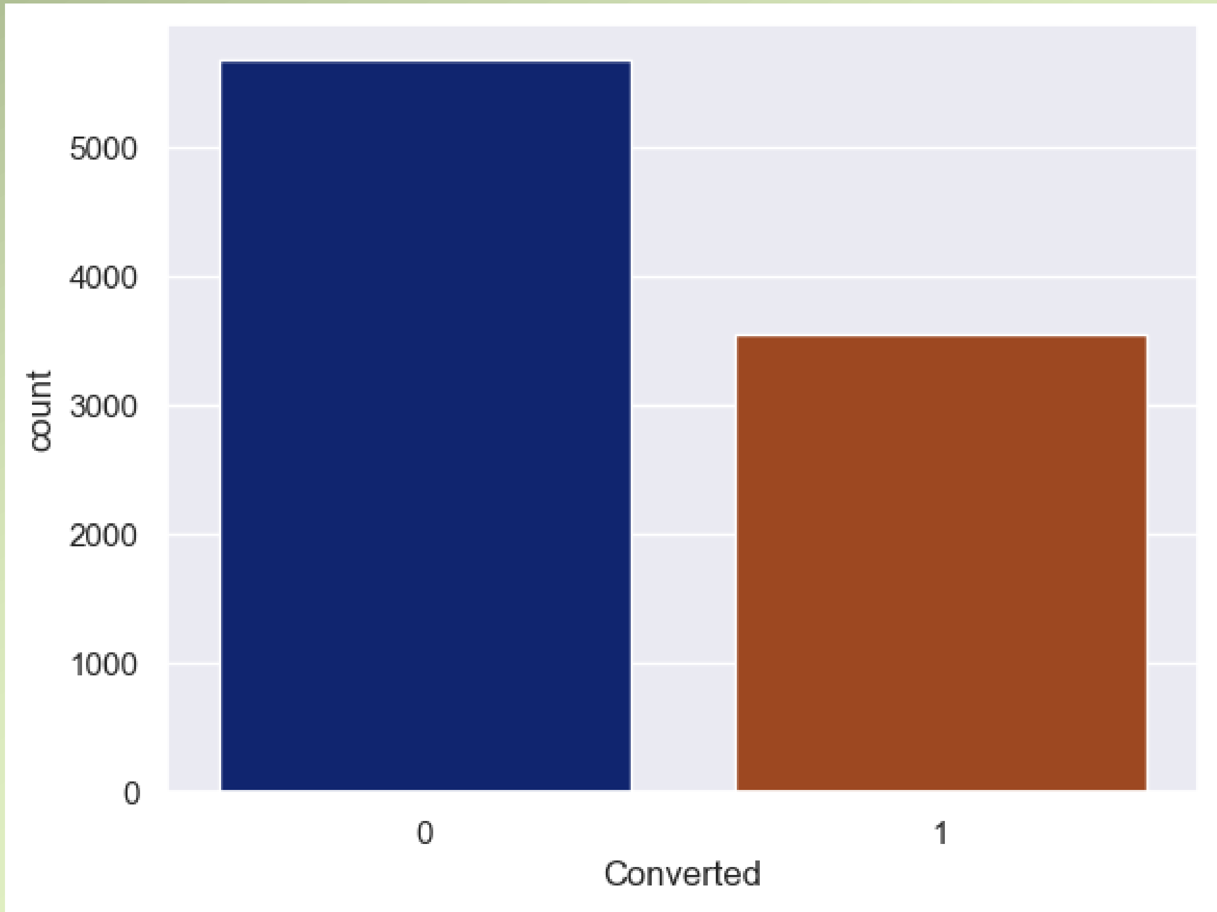
Prediction with Test data

Final Inference

Data Cleaning

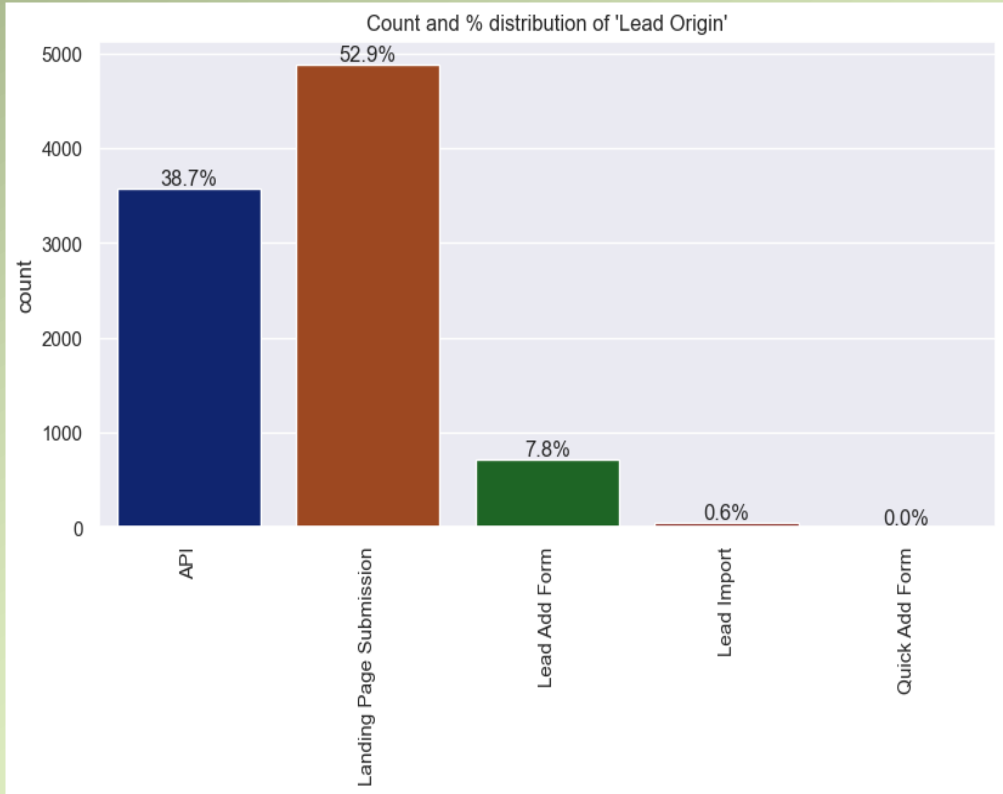
- Some of the categorical columns has a value “Select”. Since these values flow from a drop-down menu and “Select” is the default option. These can be deemed as missing or null and were replaced with same
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Columns which has one one unique value or were just row identifiers (eg Prospect ID) were dropped as they will not be useful in model creation
- Imputation was used for some categorical variables.
- Additional categories were created for some variables.
- Some of the numerical columns which contained missing values were imputed with median after checking distribution.
- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers were treated in numerical columns ‘TotalVisits’ and ‘Page Views Per Visit’ and were capped at 99 percentile.
- Invalid values were fixed and data was standardized for lead source.
- Low frequency values were grouped together to form new category “Others”

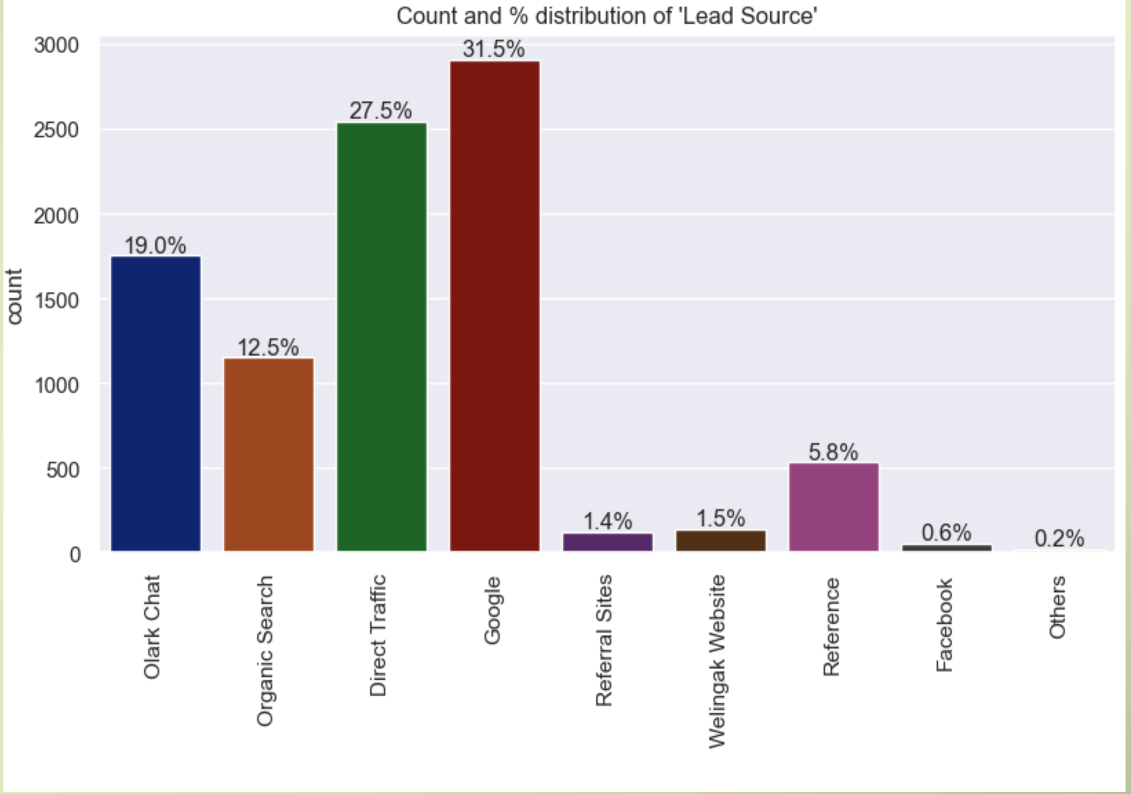
Exploratory Data Analysis



- There is a data imbalance with Target variable “Converted”
- Plot shows that only 38.5% of Leads were converted
- Around 61.5% of leads did not convert

EDA – Univariate Analysis

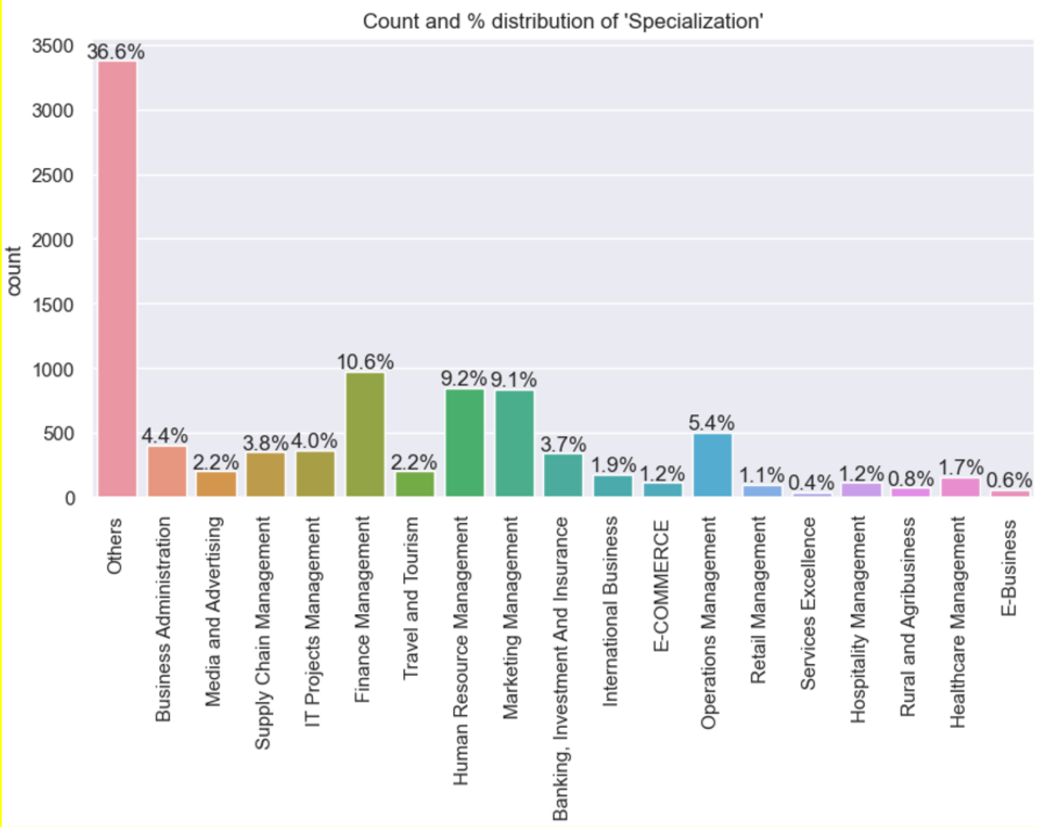


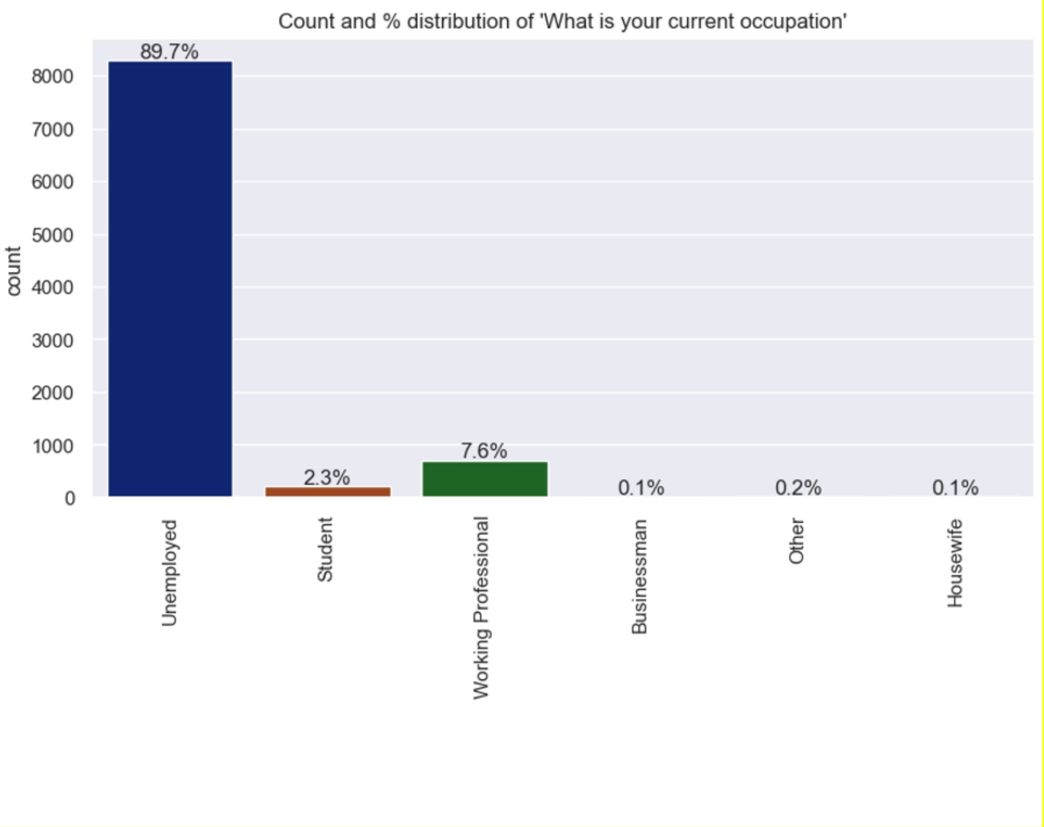


- ‘Lead Origin’ – Most of the Leads came from ‘Landing Page Submission’ followed by API.
- These two alone accounts for ~91.6% of all leads

- ‘Lead Source’ – Majority of the Leads (31.5%) are from ‘Google’ followed by ‘Direct Traffic’ (27.5%)

EDA – Univariate Analysis

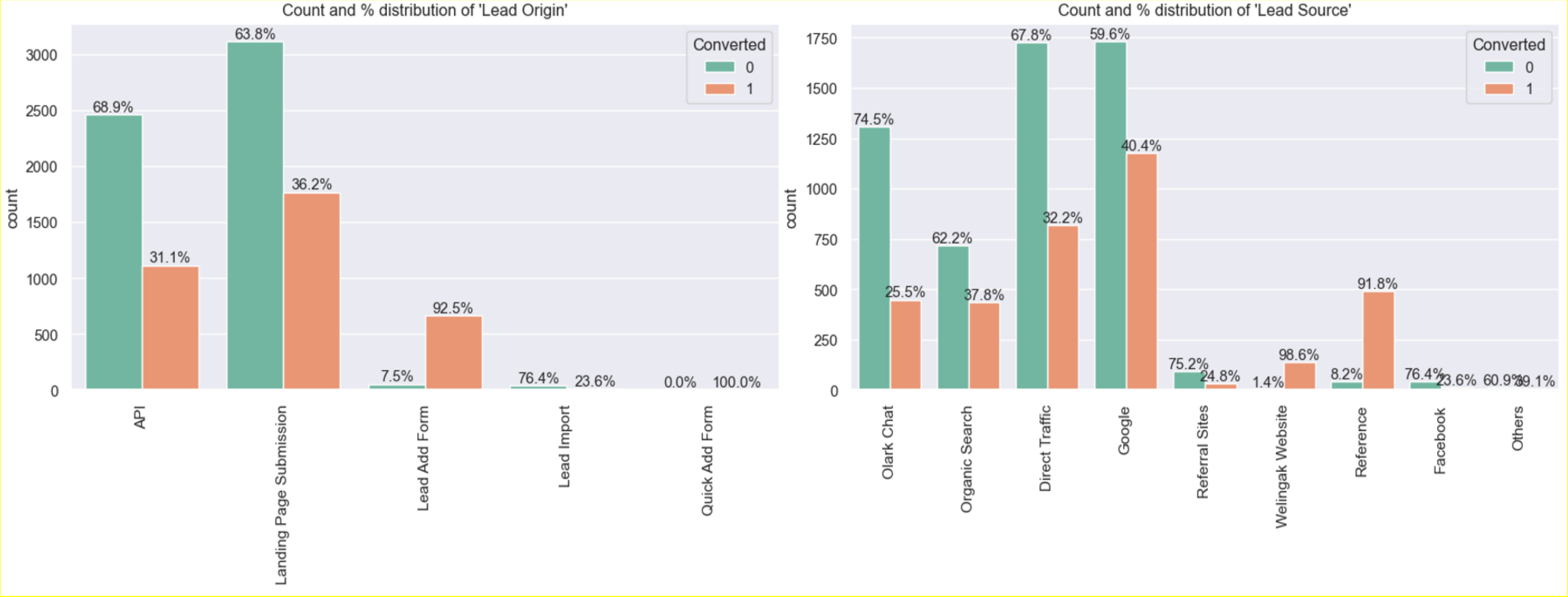




- ‘Specialization’ – Most of the leads are from different flavors of Management.
- Moreover, for 36.6% of all leads there are no information regarding specialization and are tagged as ‘Others’

- ‘Current Occupation’ – Most of leads (89.7%) were unemployed (or had missing value which was imputed as Unemployed).

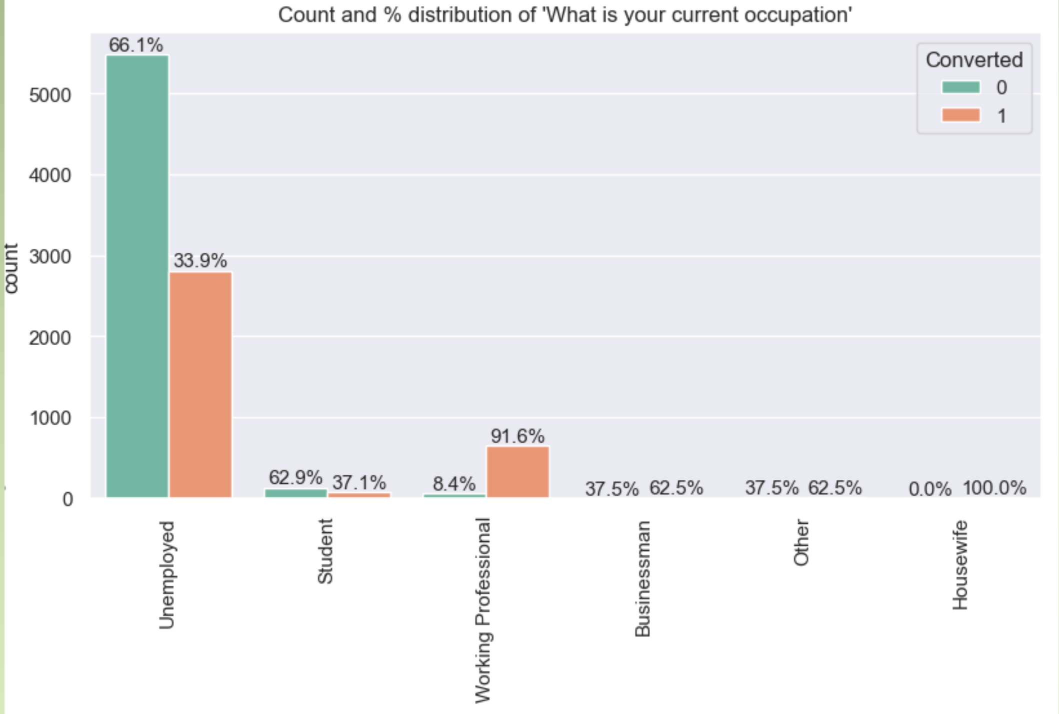
EDA – Bivariate Analysis

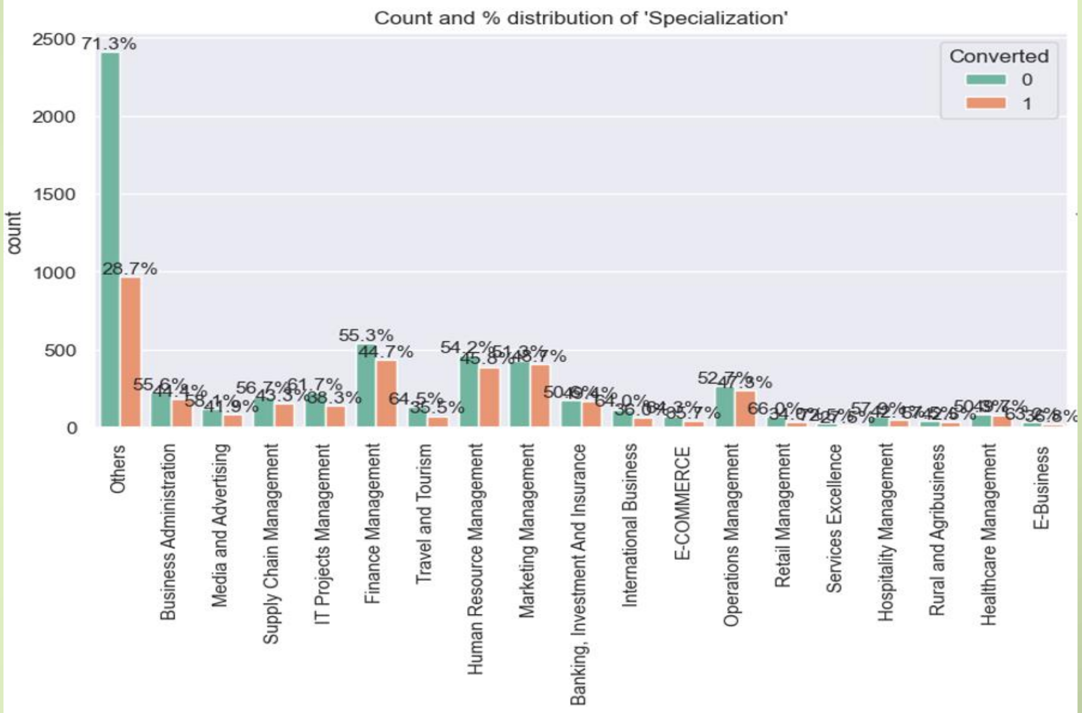


- Lead Origin – We see most lead conversions rate(LCR) for origin ‘Lead Add Form’ (92.5%).
- Though there are more leads coming in from other origins, their LCR is between 30-40%

- ‘Lead Source’ – Highest conversion are for direct ‘Reference’ source (91.8%).
- While ‘Google’ is accounting for a LCR of 40.4%

EDA – Bivariate Analysis

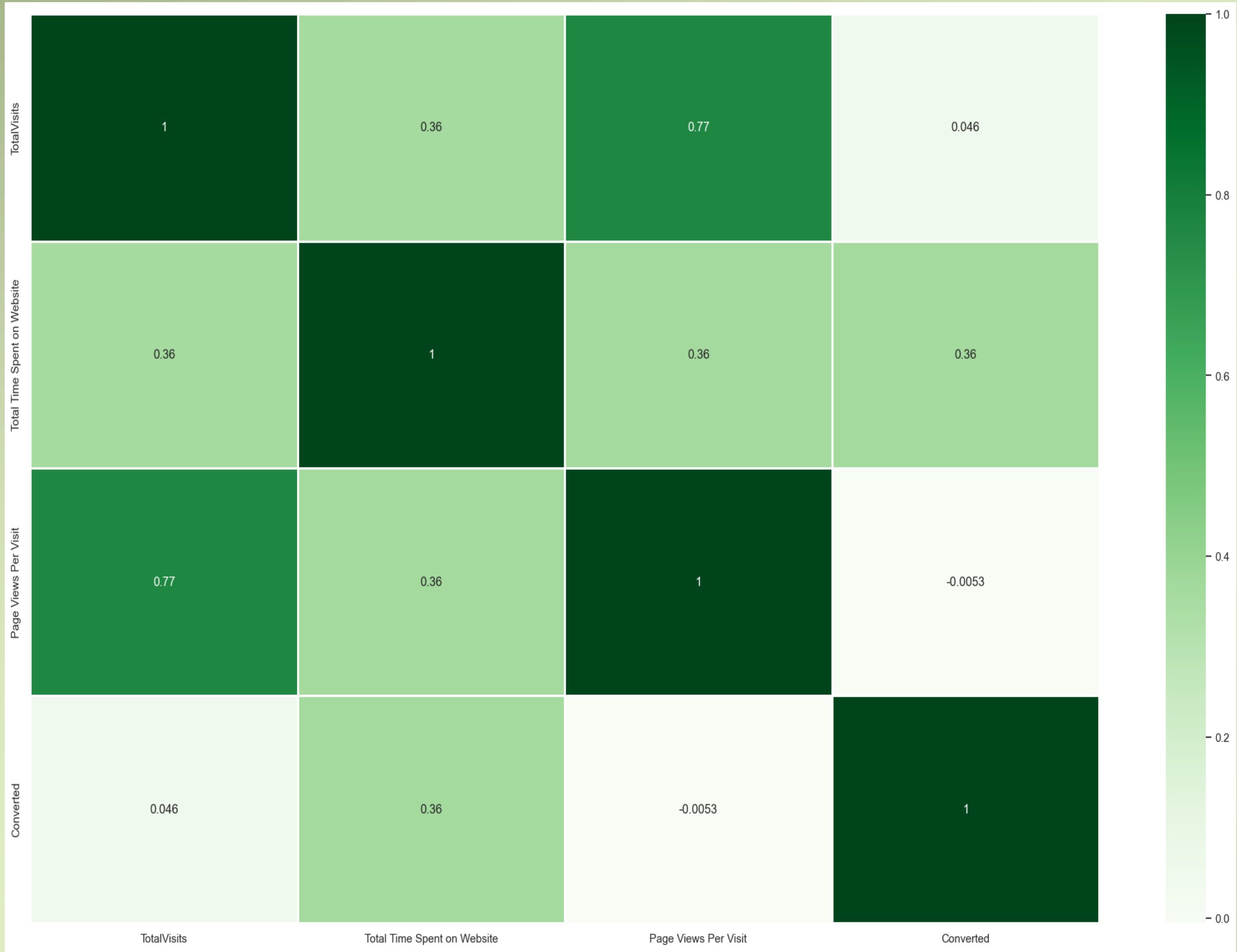




- Current Occupation - We see highest LCR for 'Working Professional' (91.6%) followed by Unemployed (33.9%)

- 'Specialization' : We see a good LCR for the 'Management' related specialization

EDA – Bivariate Analysis – Numerical Columns



- We see a correlation between ‘Total time spent on Website’ with converted variable.
- Indicating more the time spent by a Lead higher are their chances of converting

Data Preparation for Model Building

- Dummy Variables were created for categorical features ('Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity', 'Specialization', 'What is your current occupation', 'A free copy of Mastering The Interview’ and 'Last Notable Activity').
- Data set was split into two sets, train and test, at a ratio of 70:30
- Numerical Columns were scaled using MinMax scaling to convert them into normalized form
- Numerical columns are 'TotalVisits', 'Total Time Spent on Website’ and 'Page Views Per Visit'

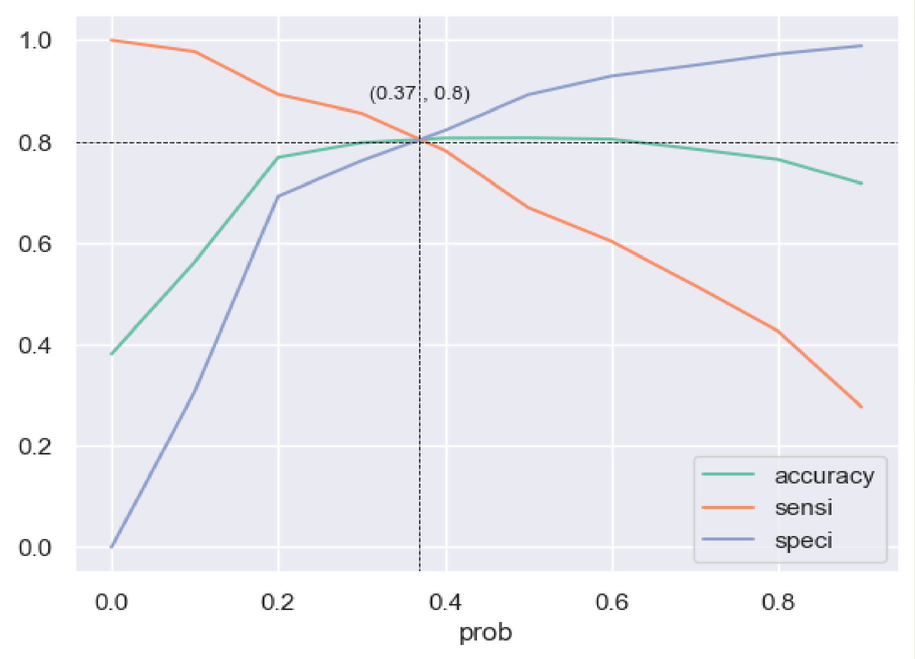
Model Building

- First model was build using all of the features
- This gave an accuracy of 82.4%
- However, p-value for some of the features were >0.05 and VIF was infinite for few
- Recursive Feature Elimination (RFE) technique was used to select only the top 15 columns
- This was further optimized to give a better model based on statsmodel parameter (p-value) and VIF
- Features like ‘What is your current occupation_Housewife’ were removed to fine tune the model

Model Building

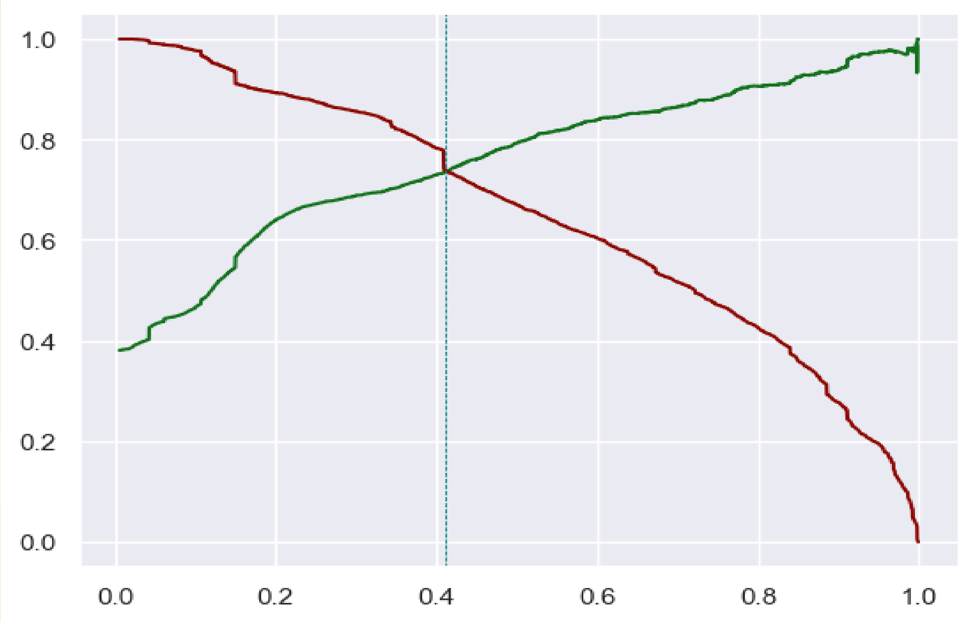
- Model 3 looks stable after three iteration with:
  - significant p-values within the threshold (p-values < 0.05) and
  - No sign of multicollinearity with VIFs less than 5
- This gave an accuracy of 80.72%
- All other metrices were calculated for this model
- An arbitrary probability of 0.5 was selected which gave a poor balance between specificity and sensitivity
- Accuracy, Specificity and sensitivity were plotted and their intersection at 0.37 was used as probability threshold
- Precision and Recall tradeoff were also calculated which gave a threshold of 0.41.
- However, metrics were better balanced with Accuracy, Specificity and sensitivity intersection, so 0.37 was selected
- ROC curve was also plotted which gave area under curve as 0.88 out of 1.

Model Evaluation - Train Data



Evaluation Metrics with 0.37 as cutoff

- Accuracy : 80.38%
- Sensitivity : 0.8077
- Specificity : 0.8013
- Precision : 0.7147
- Recall : 0.8077



Evaluation Metrics with 0.41 as cutoff

- Accuracy : 80.38%
- Sensitivity : 0.7404
- Specificity : 0.8358
- Precision : 0.7354
- Recall : 0.7404

Model Evaluation - Train vs Test Data

Model has following metrices on Train dataset :

- Accuracy : 80.38%
- Sensitivity : 80.78%
- Specificity : 80.18%
- Precision : 71.5%
- Recall : 80.78%
- Area under ROC : 0.88
- Confusion Metrics :[[3207 795]

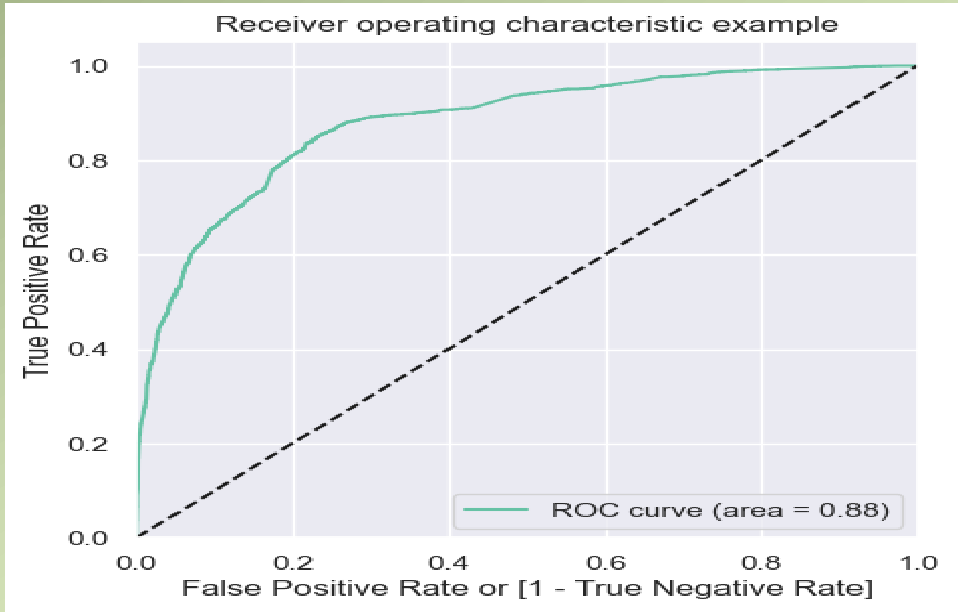
  [474 1992]]

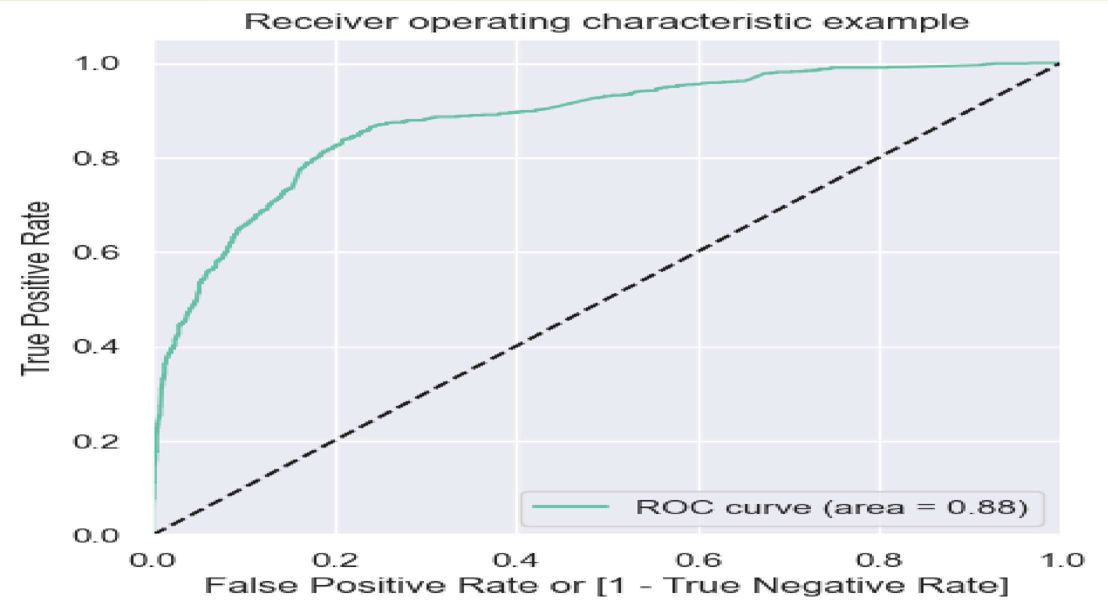
Model has following metrices on Test dataset :

- Accuracy : 81.20%
- Sensitivity : 81.27%
- Specificity : 81.15%
- Precision : 73.79%
- Recall : 81.27%
- Area under ROC : 0.88
- Confusion Metrics : [[1361, 316],

  [205, 890]]

Model Evaluation - Train vs Test Data





ROC curve for Train and Test data set:

- Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.
- Top three features that contribute to the model are `Total Time Spent on Website`, `Lead Origin_Lead Add Form` and `What is your current occupation_Working Professional`
- Probability cutoff of 0.37 was selected. Any probability above this is predicted as converted (1) and below this are not-converted
- There probability values when scaled to the range 0-100 can be deemed as Lead Score. Higher the Lead score value more is the change of conversion and vice versa.

Summary And Conclusion

Summary And Conclusion

- Top 4 features which contributes to Conversion are :
  - Total Time Spent on Website: 3.978878
  - Lead Origin_Lead Add Form : 2.690811
  - What is your current occupation_Working Professional : 2.611332
  - Lead Source_Welingak Website : 2.516484
- There are few negative coefficients as well :
  - Do Not Email_Yes : -1.617348
  - Last Notable Activity_Page Visited on Website : -1.623070
  - Last Notable Activity_Modified : -1.754393
  - Last Notable Activity_Email Link Clicked : -1.825428
- Leads who spent more time on website, more likely to convert.

Summary And Conclusion

- For the Lead Origin, 'Lead Add Form' has a very high conversion rate of 92.5%. More focus can be given on this
- In Lead Source category those through Reference has a conversion rate of 91.8%. X Education should focus more on references and probably device an incentive for same. This will boost count in this area more. For the most frequency category 'Google' Conversion rate is decently placed at 40.4%, which is above average.
- Amongst Occupation category 'Working Professional' has a high conversion rate of 91.6%. X Education should focus more on working professionals and tailor executive level courses which will attract attention of more 'Working professionals'
- More budget/spend can be done on Welingak Website in terms of advertising, etc.