Lead Scoring Case Study

Lead Scoring Case Study for X Education

- **Problem Statement** :

- X Education sells online courses to industry professionals. 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**:

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

- The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with 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%.

STEPS

➢ Following steps are performed as part of data cleaning and data manipulation.

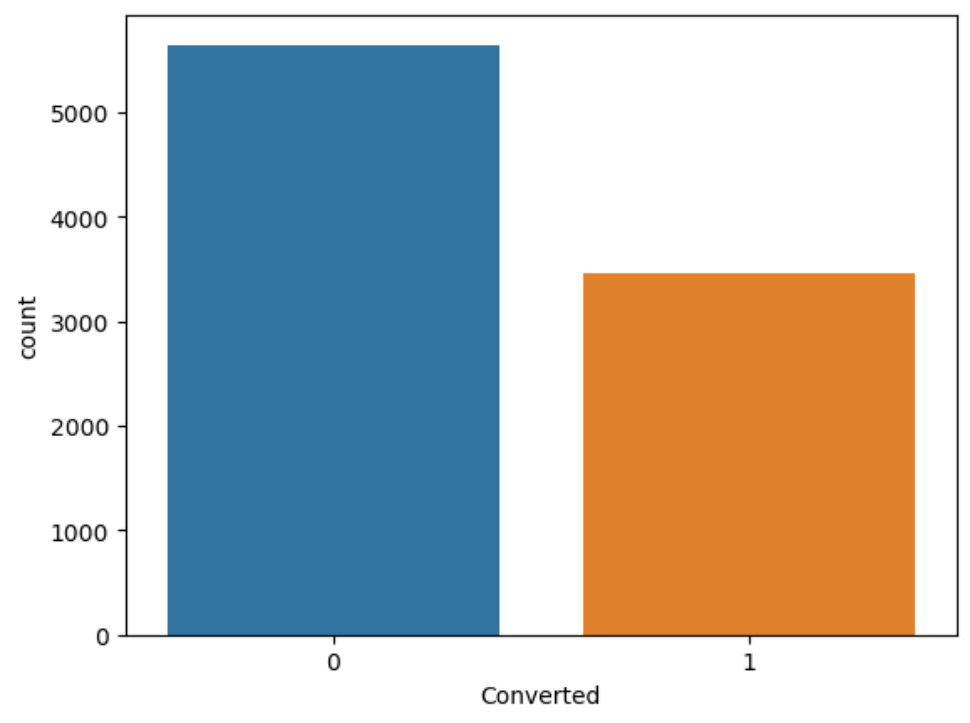
- Check and handle null values in columns.
- Drop columns, if it contains large amount of missing values and not useful for the analysis.
- Imputation of the values, if necessary.
- Check and handle outliers in data.

➢ EDA

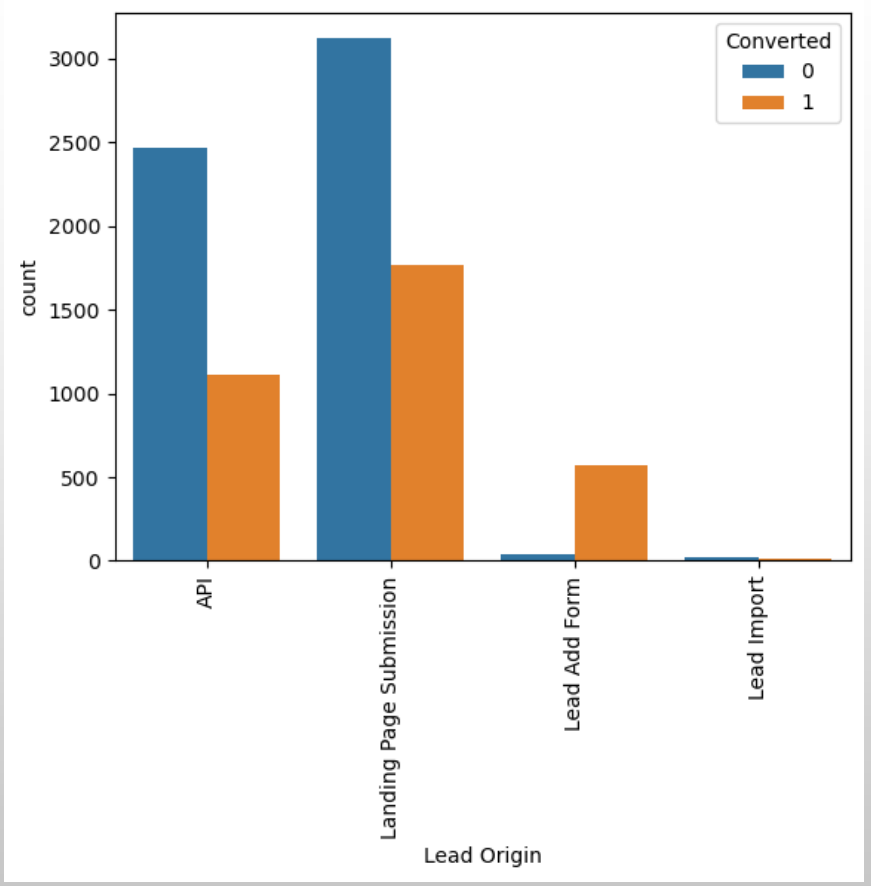
- Univariate data analysis: value count, distribution of variable etc.

➢ Scaling, dummification and encoding of the data.
➢ Classification technique: logistic regression used for the model making and prediction.
➢ Validation of the model.
➢ Model presentation.
➢ Conclusions and recommendations.

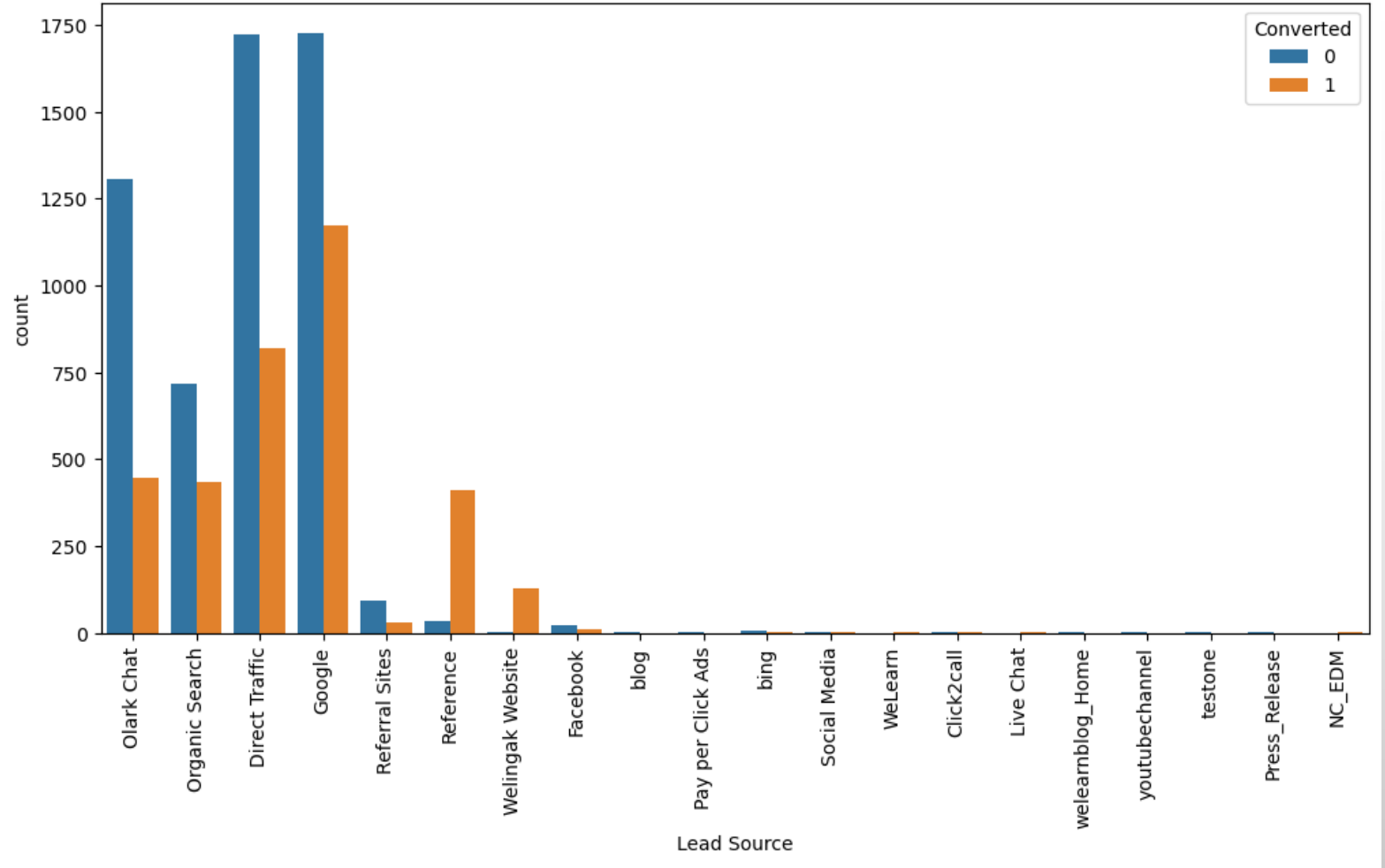
Analysis For Categorical Columns



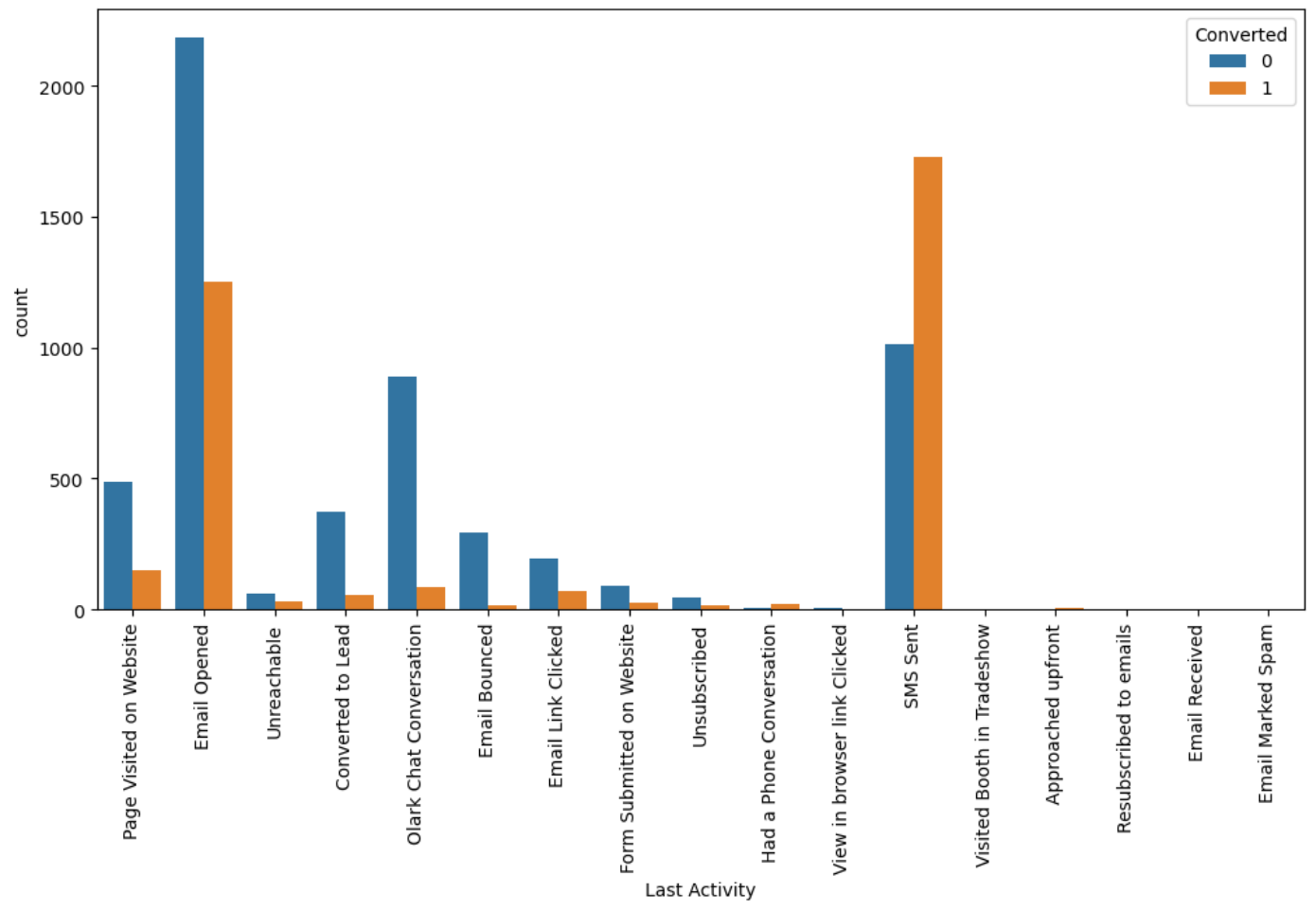
- Conversion rate is around 38.02%



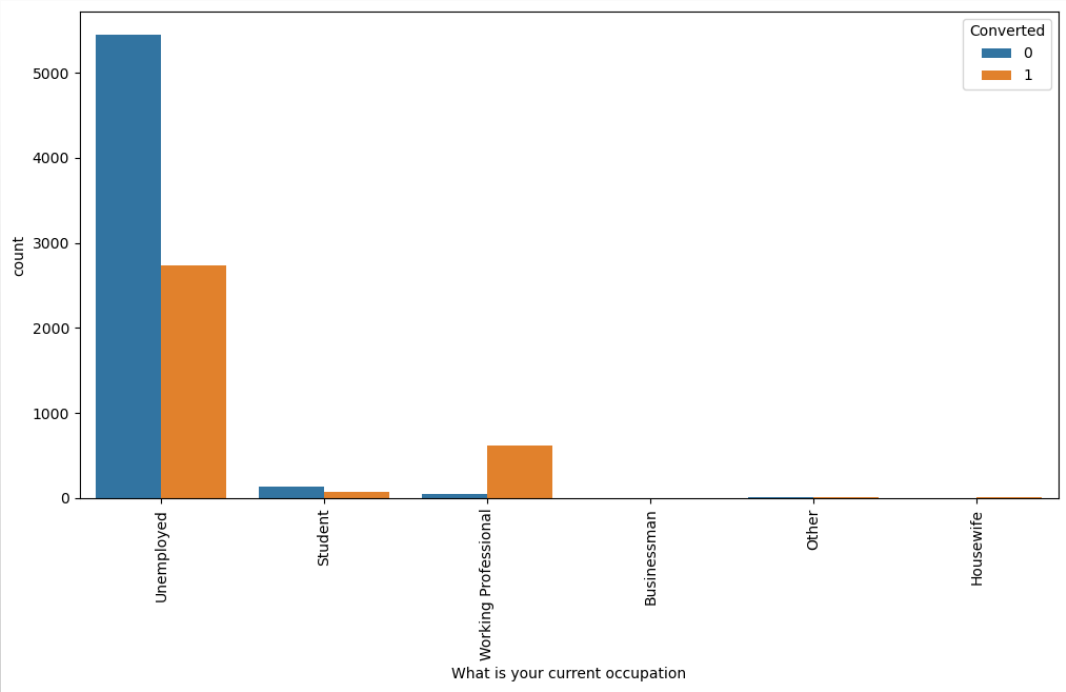
- API and Landing Page Submission have high count of lead
- Lead Add Form have more conversion rate than other

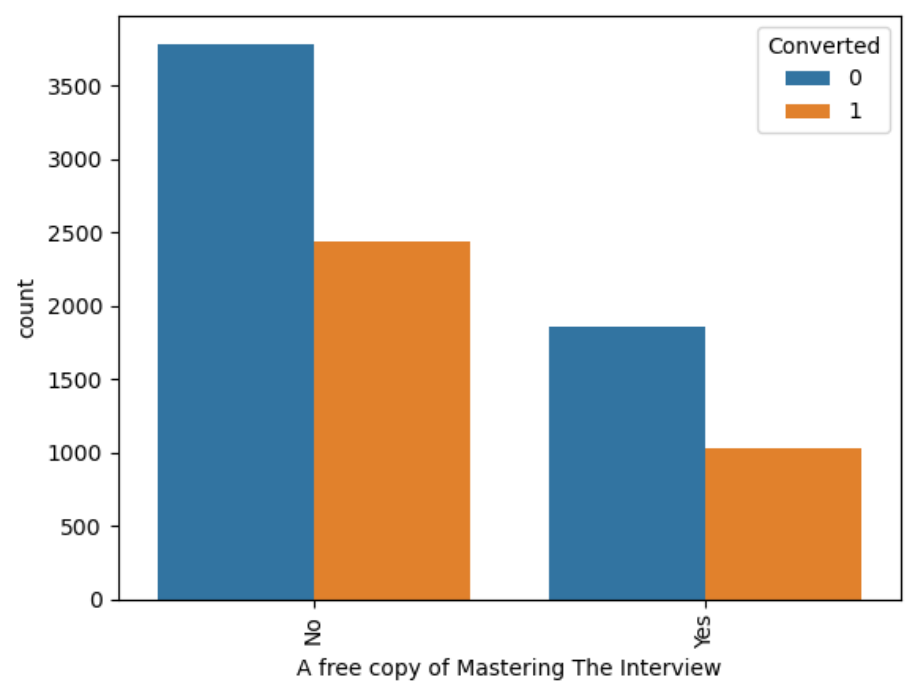


- Google and Direct Traffic cause most number of leads
- conversion rate is high in Reference and Welingak Website



- Most people's has last activity as email opened
- Conversion rate for last activity as SMS Sent is highest
- We can avoid leads from Olark chat conversion and email bounced as last activity

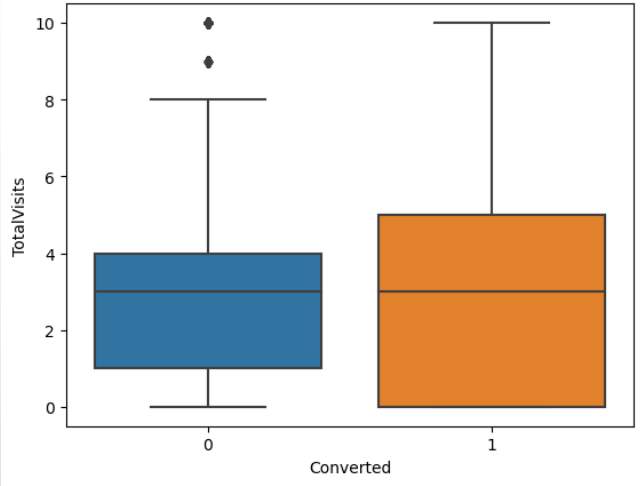




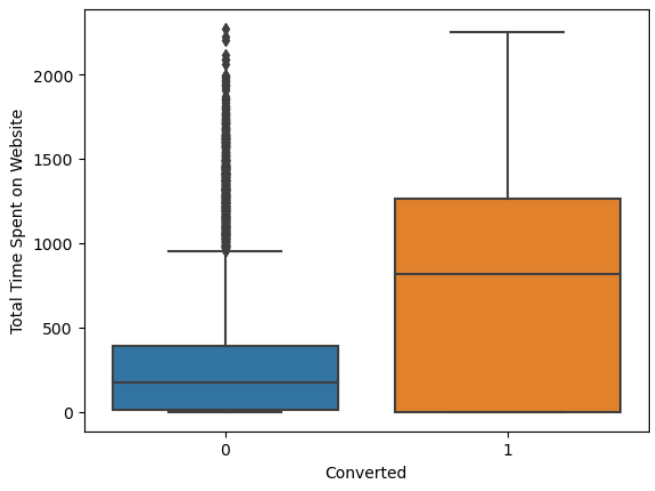
- Working Professional have higher conversion rate
- Unemployed category have high count of conversion
- Even

- conversion rate is high on leads who do not want a free copy of Mastering Interviews

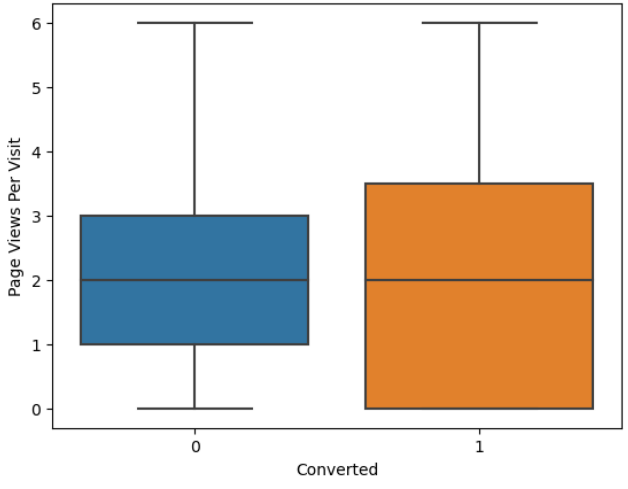
Analysis For Numerical Columns



- Median is same for both so we cant get any inference by knowing that total visits made by customer affect the conversion rate



- Customer who spend more on website are more likely to converted



- Average number of pages on the website viewed during the visits doesn’t add any value to conversion rate .

Data Preparation before Model building

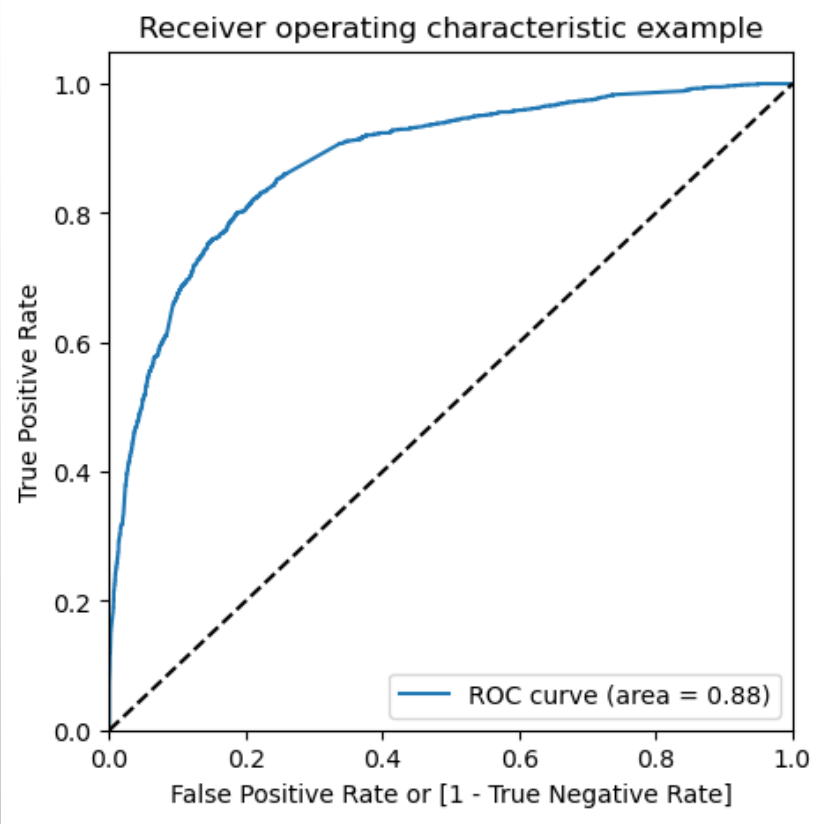
- Created dummy features (one-hot encoded) for categorical variables
- Added dummy variable and dropped the main categorical columns for which we created dummies
- Split the data into 70:30 Train ,Test set with 100 random state .
- Feature scaling - MinMaxScaler used to scale the features
- Checking the correlations

Model Building

Feature Selection

- The data has lots of dimension and large number of features. This will reduce model performance and might take high computation time.
- Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns and eliminate highly correlated columns
- Then we can manually fine tune the model with the RFE outcome
- Pre RFE – 64 columns & Post RFE – 20 columns
- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Model 5 looks stable with significant p-values within the threshold (p-values < 0.05) and No sign of multicollinearity with VIFs less than 5
- Hence, it will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

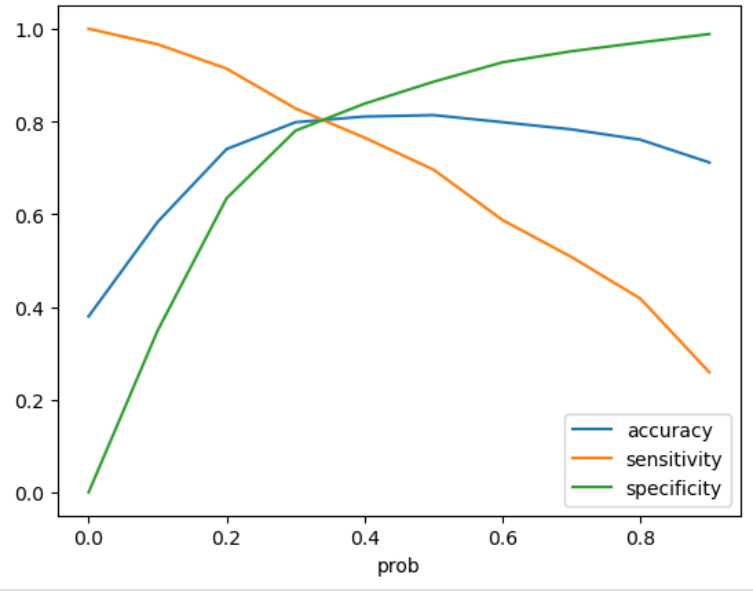
Model Evaluation



- With 0.5 probability the accuracy is above 81%
- Sensitivity -69.61%
- Specificity-88.56%

- Area under curve is 0.88 out of 1 which shows a good predictive model For train data set

Optimal Cutoff point

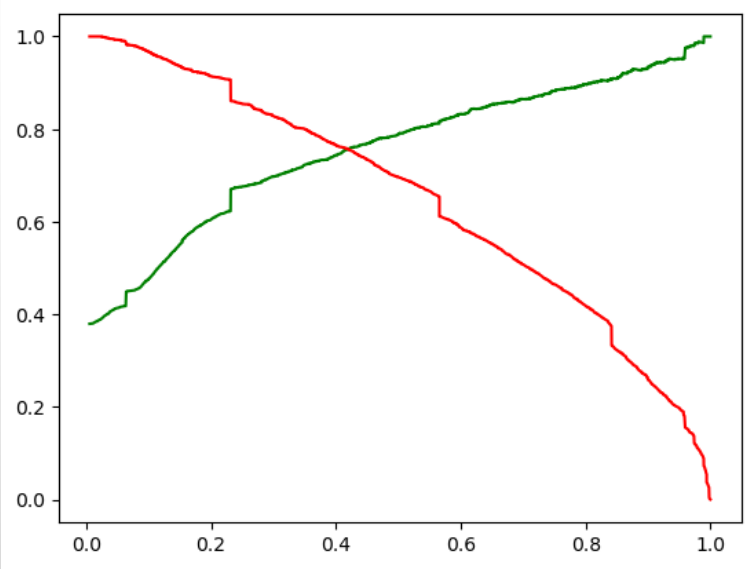


From the curve around 0.3 is seems the optimum point to take it as a cutoff probability.

	prob	accuracy	sensitivity	specificity
0.0	0.0	0.379630	1.000000	0.000000
0.1	0.1	0.583333	0.966515	0.348849
0.2	0.2	0.740427	0.914014	0.634202
0.3	0.3	0.798336	0.828028	0.780167
0.4	0.4	0.810578	0.765192	0.838351
0.5	0.5	0.813716	0.696155	0.885656
0.6	0.6	0.798493	0.587433	0.927650
0.7	0.7	0.782957	0.507648	0.951429
0.8	0.8	0.760672	0.417941	0.970402
0.9	0.9	0.711551	0.258785	0.988616

- With 0.3 As probability we get the train set accuracy 79.83% so we can say it around 80% which is not bad for further analysis
- Sensitivity -82.80%
- Specificity-78.01%
- Precision – 69.74%
- Recall- 882.80%

Prediction On Test Set



After prediction On test data set with 0.3 prob and assigning Lead score We got the metrices

- Accuracy- 80.22%
- Sensitivity -83.20%
- Specificity-78.39%
- Precision – 70.37%
- Recall- 83.20%

Recommendation based on Final Model

- As per the problem statement, increasing lead conversion is crucial for the growth and success of Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.
- We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.
    - Total Time Spent on Website
    - Lead Origin Lead Add Form
    - Current occupation Working Professional
    - Last Notable Activity Had a Phone Conversation
    - Lead Source Welingak website
    - Last Notable Activity SMS sent
    - Last Notable Activity Unreachable
    - Lead Source Olark Chat

Recommendation based on Final Model

To increase our Lead Conversion Rates

- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- More budget/spend can be done on Welingak Website in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage providing more references.
- The "Working Professionals" should receive calls from the company because they are more likely to convert.
- target people based on their most recent SMS and email opening activity.
- Last Notable activity Had a phone conversation as the most recent noteworthy activity.
- have better financial situation to pay higher fees too.
- To identify areas of improvement