Lead Scoring Case Study

SUBMITTED BY :

1. Tanishq Samdarshi
2. Swati Ingole Nichat
3. Sanjana Tadem

CONTENTS

- ❏ Problem statement
- ❏ Problem approach
- ❏ EDA
- ❏ Correlations
- ❏ Model Evaluation
- ❏ Observations
- ❏ Conclusion

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. They have process of form filling on their website after which the company that individual as a lead.
- 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, this means 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

- Lead X wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full manpower and after achieving target what should be the approaches.

STRATEGY

- Importing Data.
- Data cleaning (handling null values and outliers) and removing columns with high % of null values.
- Exploratory Data Analysis to analyse behaviour of variables and select most contributing attributes.
- Train-Test Split and Scaling features.
- Using RFE (automated approach) to select best feature for model building.
- Building model and manual tuning to get p-value<0.05 and VIF <5.
- Predict model on train set and assign lead score to each lead based on threshold value by ROC Curve.
- Evaluate train model parameters.
- Test model on Test set.
- Evaluate test model parameters.

DATA CLEANING

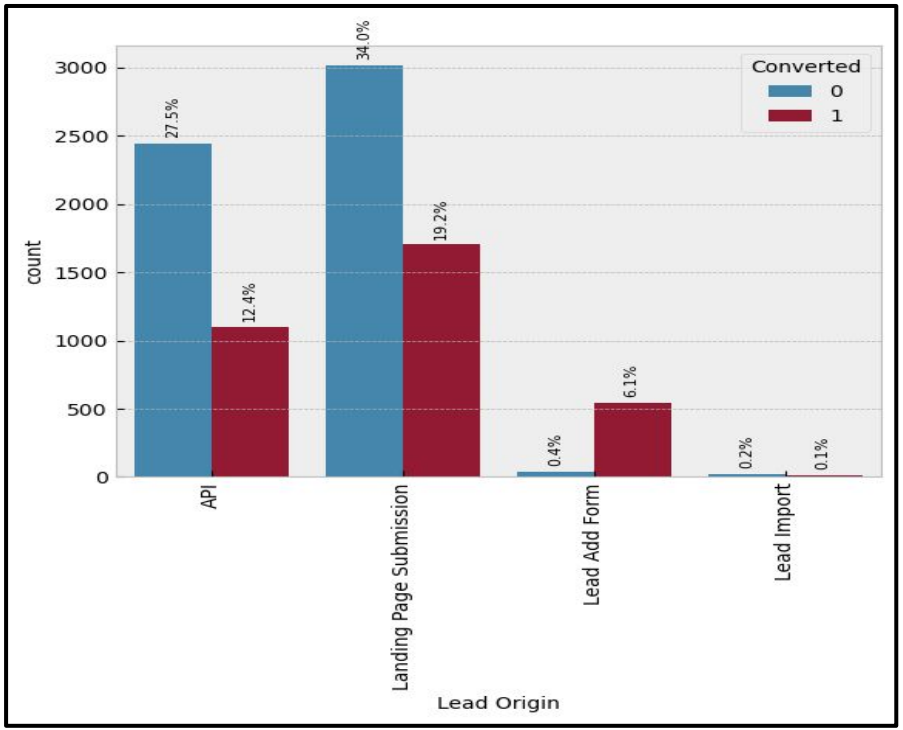
Assumptions and data cleaning performed on dataset:

- The "Select" level represents null values as customers did not choose any option from the list and were replaced by null values.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and their distribution.
- Dropped columns that didn't add any insight or value to the study objective ('City', 'Tags', 'Country', 'What matters most to you in choosing a course')
- Where the data was not skewed, imputation was used for categorical variables.
- Columns not required for further modeling ('Prospect ID', 'Lead Number') were dropped.
- Numerical data was imputed with mode after checking distribution.
- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers in 'TotalVisits', 'Total Time Spent on Website' and 'Page Views Per Visit' were treated and capped.
- Low frequency values were grouped together to “Others”. Binary mapping was performed for well distributed categorical column.
- Standardizing Data in columns by checking casing styles (“Lead Source” has Google and google)

EXPLORATORY DATA ANALYSIS

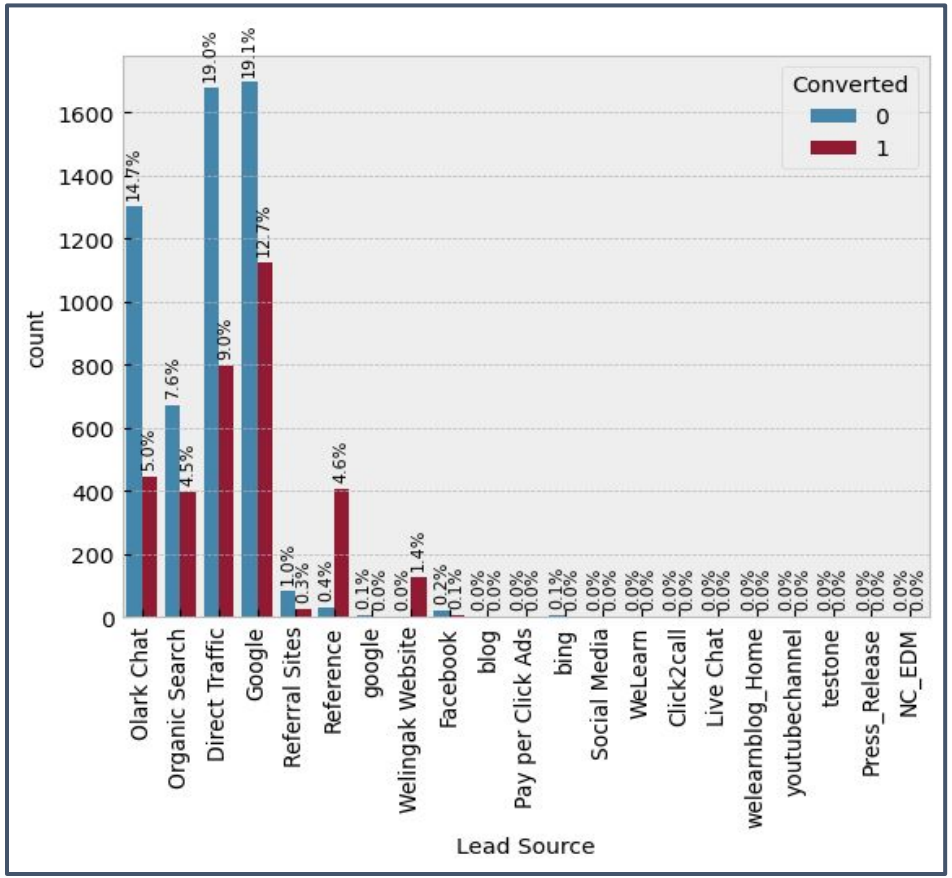
Univariate Analysis

1) Lead Origin Analysis



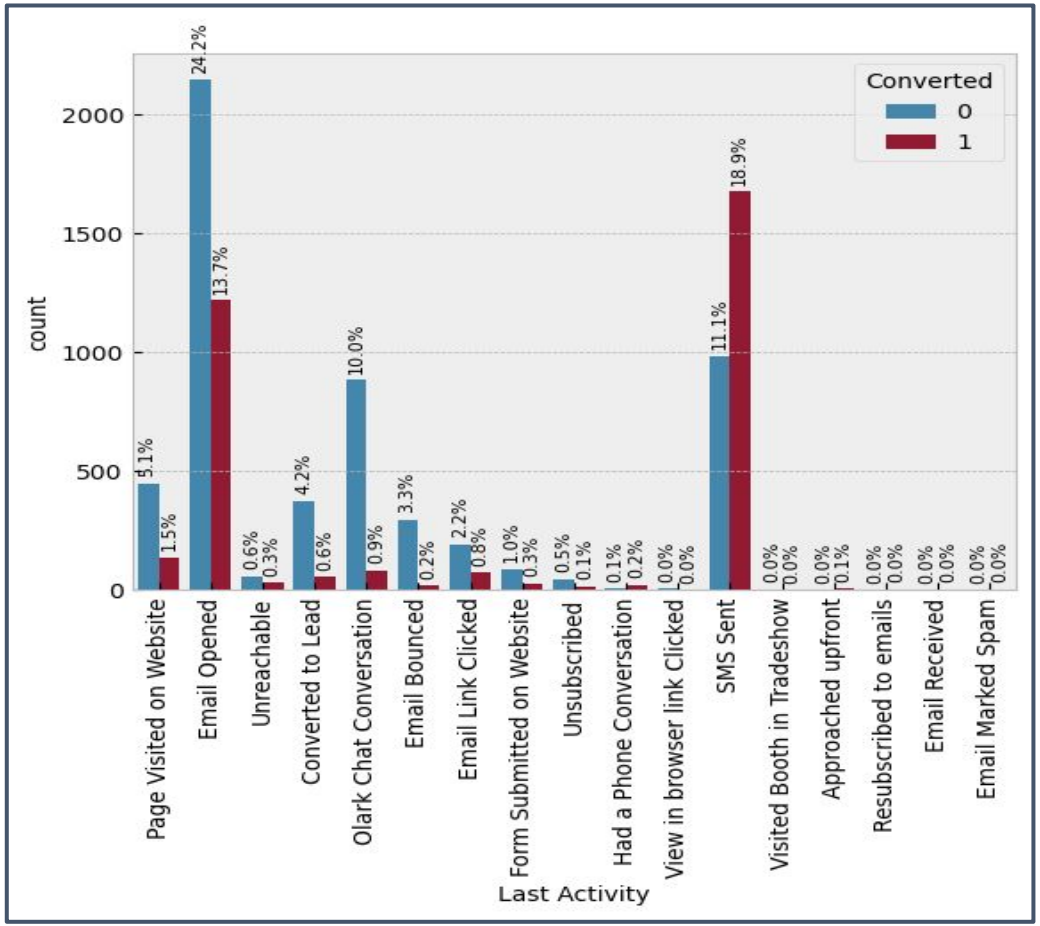
- API and Landing Page Submission have 12.4-19.2 % conversion rate but count of lead originated from them are considerable.
- Lead Add Form has good conversion rate but count of lead are not very high.
- Lead Import are very less in count.
- To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

2) Lead Source



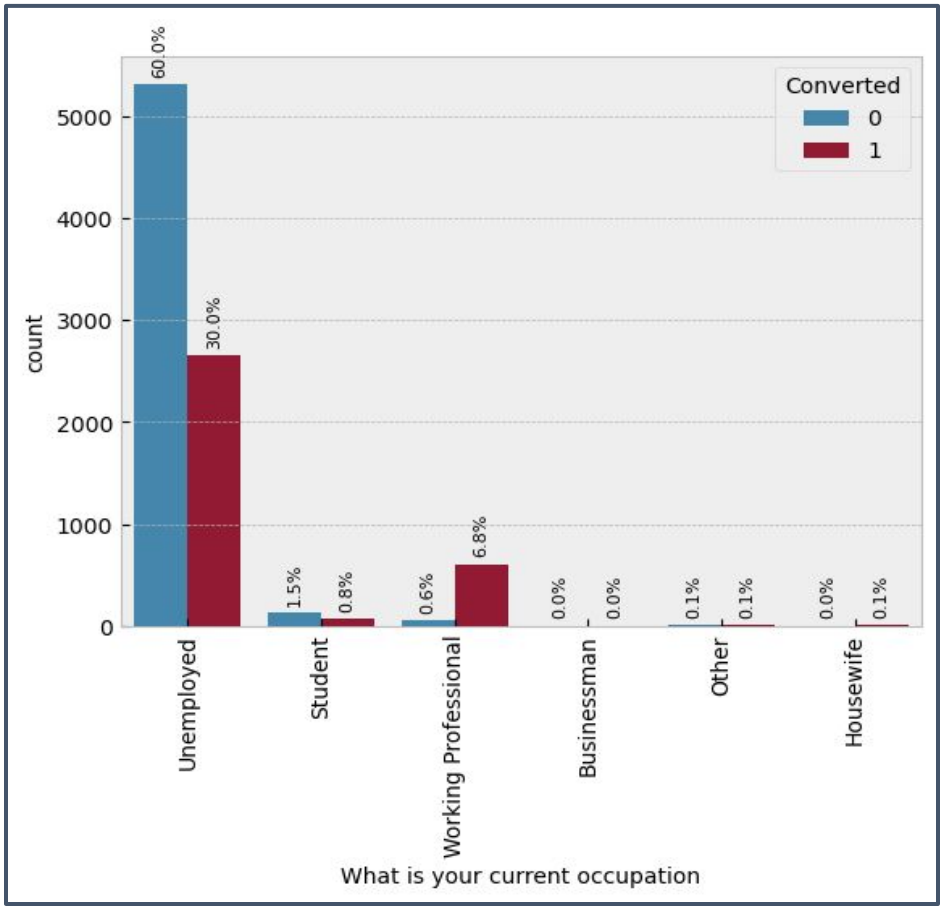
- Google and Direct traffic generates maximum number of leads having 30-35 % of conversion rate.
- Conversion Rate of reference leads and leads through welingak website is highest.
- Apart from 5-6 categories , the contribution of others is negligible.
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

3) Last Activity



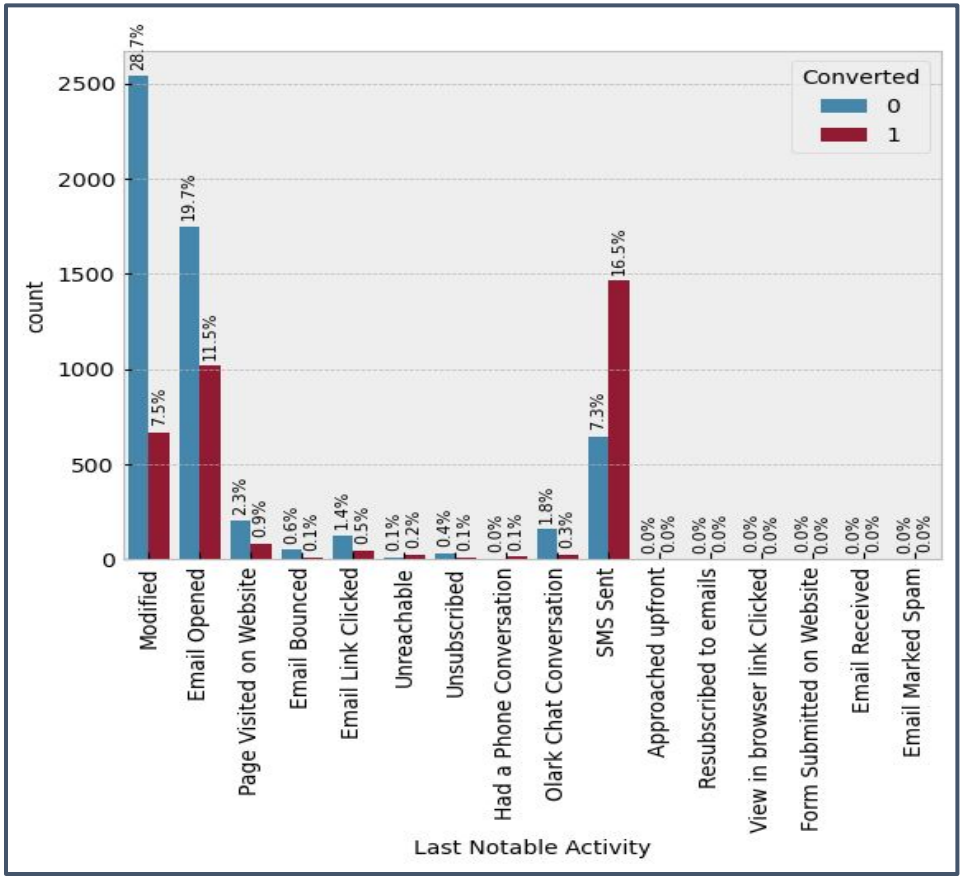
- Most of the lead have their Email opened as their last activity and has approx 35% conversion rate.
- Email Opened and SMS sent have majority of the leads
- Conversion rate for leads with last activity as SMS Sent is almost 60%.
- Focus should be on these two categories to improve lead conversion % as other categories lead conversion percentage is poor

4) What is your current occupation



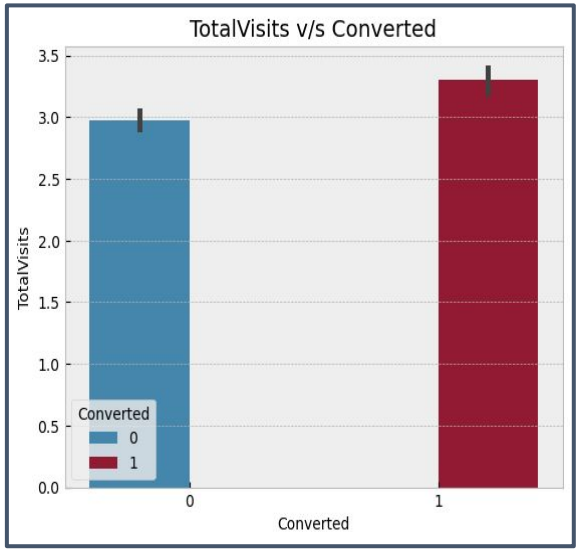
- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in numbers but has around 30-35% conversion rate and are maximum in number of leads.
- To improve conversion rate preference should be given to Unemployed customer and after that to Working professional .
- This will increase efficiency of work and maximize productivity.

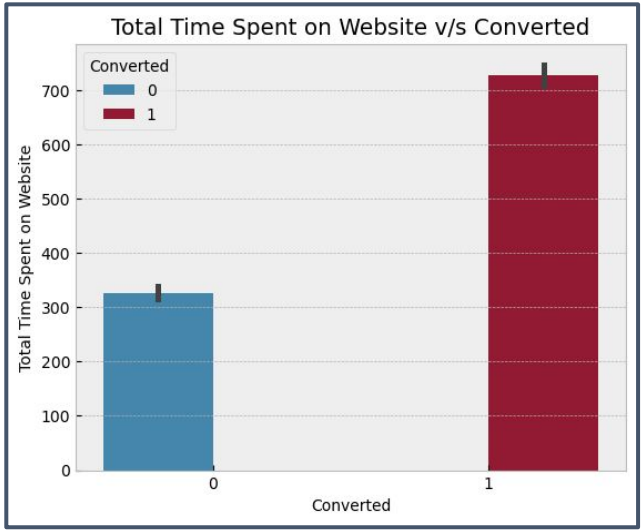
5) Last Notable Activity

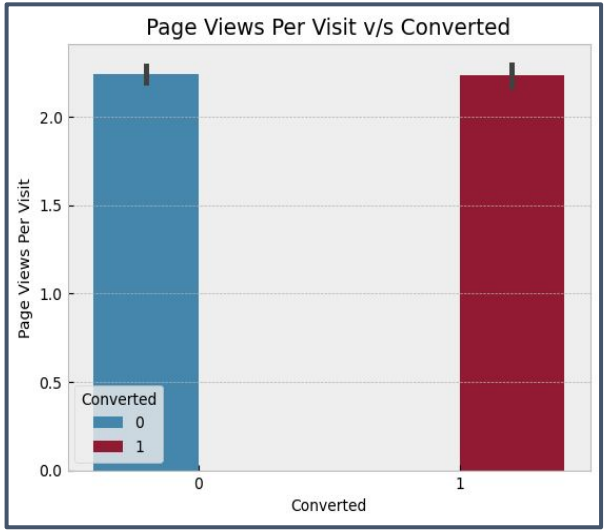


- Modified , email sent and email opened holds maximum share of leads.
- SMS sent has highest conversion rate followed by Email Opened and Modified
- For increasing conversion rate SMS sent leads must be given preference..

Bivariate Analysis







- TotalVisit v/s Converted and Page View per Visit v/s Converted has almost comparable number of converted and unconverted leads
- Total Time Spent On Website v/s Converted has more number of converted leads as compared to not converted
- For getting conversion rate higher focus should be on Total Time Spent attribute , higher time spent on website, higher the chances of conversion of leads

Note: In above graphs: 0 indicates not-converted and 1 indicates - converted

Multivariate analysis

In support with bi-variate analysis as depicted in multivariate analysis graph:

- There is a strong positive correlation between 'Total Visits' and 'Page Views per Visit', indicating that customers who visit the website more frequently tend to view more pages per visit.

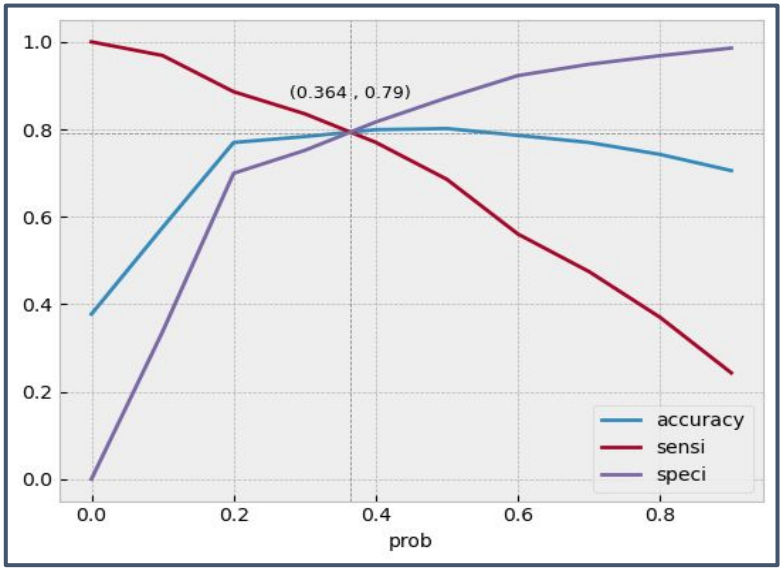
- Customers who spend more time on the website have a higher LCR, indicating that increasing the time spent on the website can lead to higher conversion rates.

MODEL BUILDING

1. Splitting dataset into train and test dataset.
2. Scale Variable using MinMaxScaler in the dataset.
3. Use RFE to select variables (automated approach) which effectively contributes to model building.
4. Build the first model.
5. Using manual approach remove feature whose p-value or VIF is not under acceptance value
6. Repeat above step till p-value and VIF is under acceptable range.
7. Predict using train set.
8. Evaluate accuracy and other metrics.
9. Predict using test set.
10. Evaluate accuracy and other metrics.
11. Compare evaluation parameter of train and test set.

MODEL EVALUATION

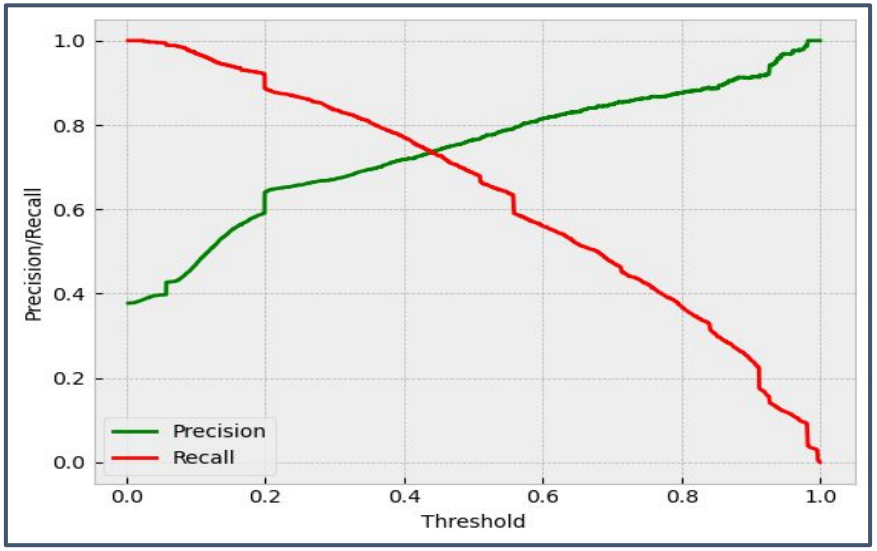
ROC curve



Model evaluation was carried out using Confusion matrix with a cutoff point of 0.36 suggested by ROC curve. It lead to following evaluation parameters for test set:

1. Accuracy : 78.3%
2. Sensitivity : 83.6 %
3. Specificity : 75.2 %

Precision-recall trade-off

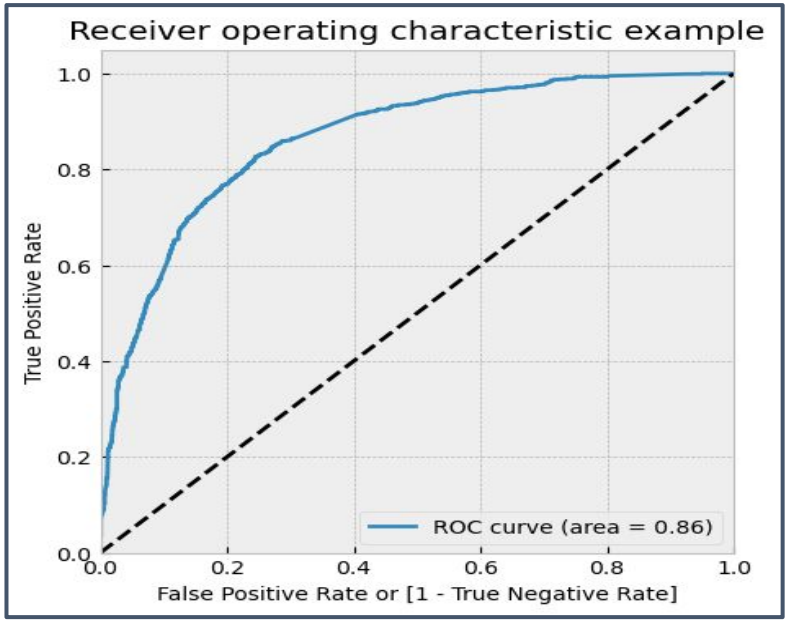


The precision-recall yielded the following with a cut-off of 0.41:

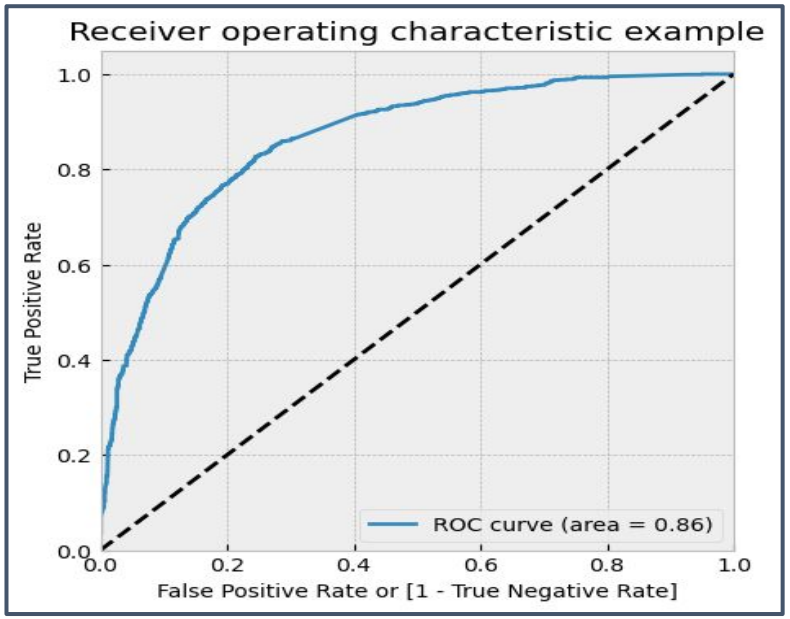
- Precision score = 0.7
- Recall score = 0.7962

Roc curve -for train and test

Test



Train



ROC Curve findings:

1. The Area under ROC for both train and test data set is 0.86 which indicates that the model is a good predictor.
2. The curve plotted as close as to the top left corner of the plot indicates that the model has a high true positive rate and a low false positive rate at all threshold values.

MODEL EVALUATION COMPARISON

Train set

1:Accuracy : 78.3%

2:Sensitivity : 83.6 %

3:Specificity : 75.2 %

Test set

1:Accuracy : 78.3%

2:Sensitivity : 83.6 %

3:Specificity : 75.2 %

CONCLUSION

- There are total 434 hot lead which needs to be contacted first .
- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should give priority to the leads whose lead origin is "Lead add form" as they highly likely to get converted.
- The company should avoid making calls to the leads whose lead origin is "Landing page submission" as they are not likely to get converted.
- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Olark chat" as these are more likely to get converted.
- The company should make calls to the leads whose last notable activity was "SMS Sent" and "Others" as they are more likely to get converted.
- The company should make calls to the leads whose total visit is higher in number as they are more likely to get converted.
- The company should avoid making calls to the leads whose last activity is "Olark chat conversation" and "Email bounced" as they are not likely to get converted.
- The company should not make calls to the leads whose Specialization was "Others" as they are not likely to get converted.
- The company can give preference to leads having Specialization in “Finance Management " , “HR Management" , “Marketing Management" and “Operation Management" as their lead conversion rate is very good despite having less number of leads

SOME USEFUL BUSINESS RECOMMENDATIONS:

- Advertisements on “Welingak Websites" and "Olark chat" can be made more interesting and engaging as more converters are for these two sources
- An automated response form or genAI support to to be given customers who are likely to or usually spend more time on website to know their interest areas and provide information accordingly. As those spending more time on website have higher conversion rate.
- Other strategies like chatbots to increase user friendly communication may lead to attract more customers.
- Further focus to be given on specialization to acquire more data, so that tailored information and courses can be offered since we have good conversions from “Finance Management ” , “HR Management” , “Marketing Management” and “Operation Management”.
- More incentives and offers for referral benefits can be provided to attract more customers