LEAD SCORING CASE STUDY

Submitted by :
Aravind R
Sanket Garg
Bose Karthik Y

PROBLEM STATEMENT

- X Education sells online courses to professionals who visit their website. They market courses on various platforms. Visitors can explore courses, fill out a form, or watch videos. Those who provide contact details become leads. Leads also come from referrals. The sales team contacts leads via calls and emails. The conversion rate is about 30%.
- Despite acquiring many leads, X Education wants to improve efficiency by identifying the most promising leads, called 'Hot Leads.' Focusing on these leads should increase the conversion rate and allow the sales team to prioritize communication with them.
- We are give a data set with over 9000 points of data along with data dictionary, we are to make a logistic regression model for the above case to filter the hot leads based on score of 0 to 100

DATA UNDERSTANDING

- We are given two files : Leads.csv and Leads data dictionary
- ‘Leads.csv’ contains around 9000 data points and the target variable is ‘Converted’ which is binary , where 0 indicates lead not converted and 1 indicates lead was converted
- ‘Leads Data Dictionary.xlsx’ file contains the data dictionary which explains each of the variables in ‘Lead.csv’

ANALYSIS APPROACH

- Data Preparation
  i)Cleaning dataset by removing null values, imputing missing values using 30% as threshold etc
- Exploratory Data Analysis
  i)Analysis using univariate bivariate and multivariate analysis
- Feature Selection
  i)Creating dummy variables
  ii)Making correlation Matrix and removing highly correlated variables using domain knowledge
- Model Development
  i)Applying logistic regression to predict the probability of leads being hot.
  ii)Train the logistic regression model
  iii)Removing variables using p and VIF values
  iv) Evaluate the model's performance using appropriate metrics like accuracy, precision, recall, and F1-score.
-

ANALYSIS APPROACH

- Model Validation

  i) Validate the suitable model on test data set

  ii) Check the model performance using evaluation metrics

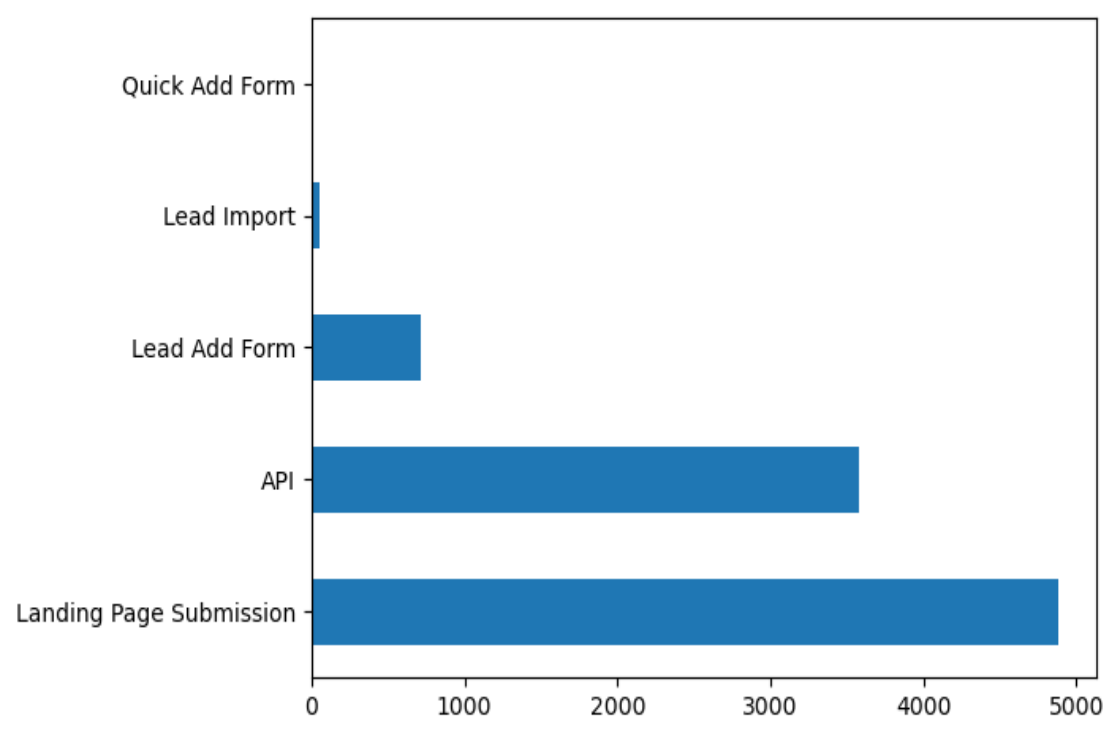
- Interpretation

  i) Analysis of coefficient of logistic regression how each contribute to predicting the hot leads

  ii) Identifying the top 3 variables that play a major role in predicting.

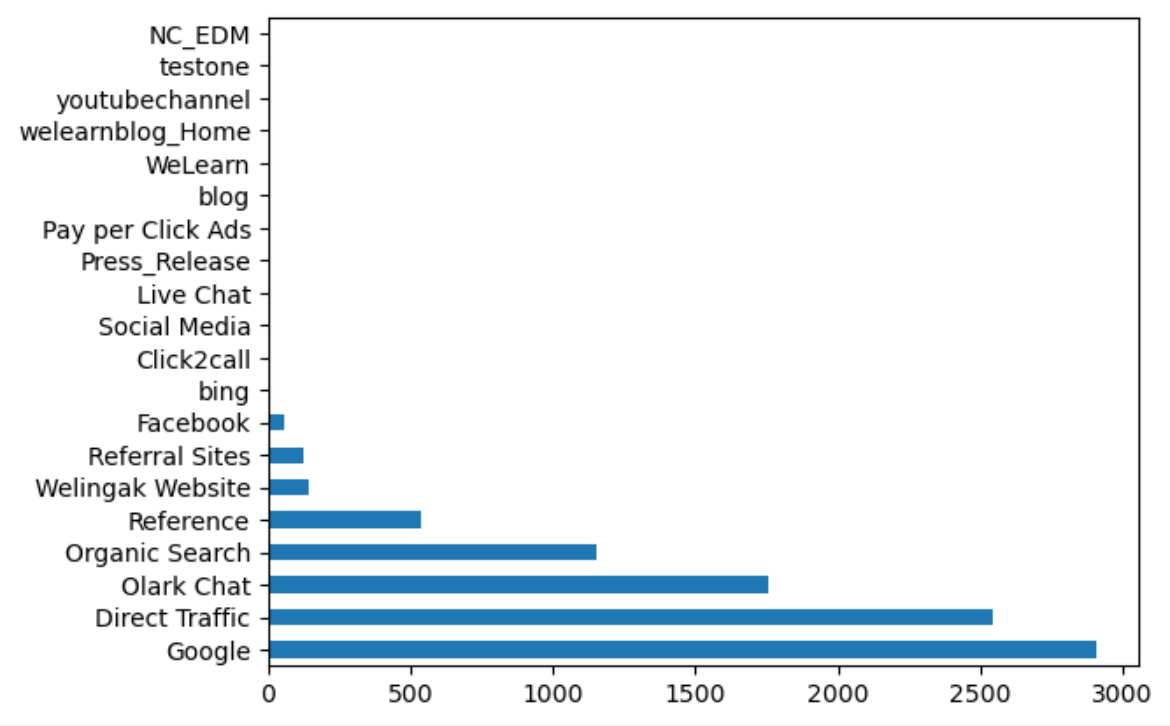
  iv) Provide recommendation based on insights

UNIVARIATE ANALYSIS



Lead Origin count

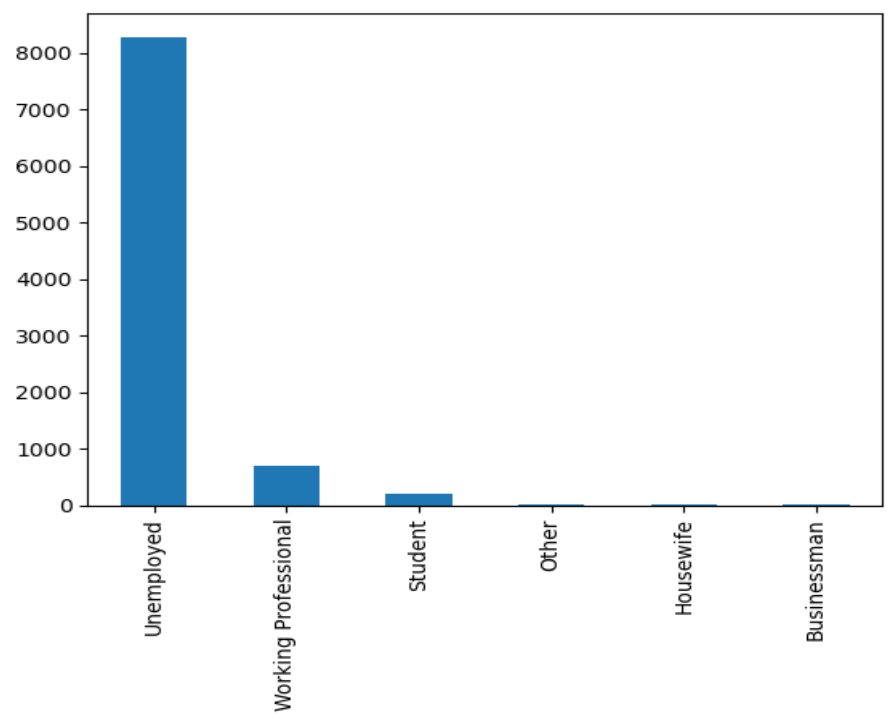
- This shows that origin identifier with which the customer was identified to be a lead mostly done using API, Landing Page Submission and lead add form



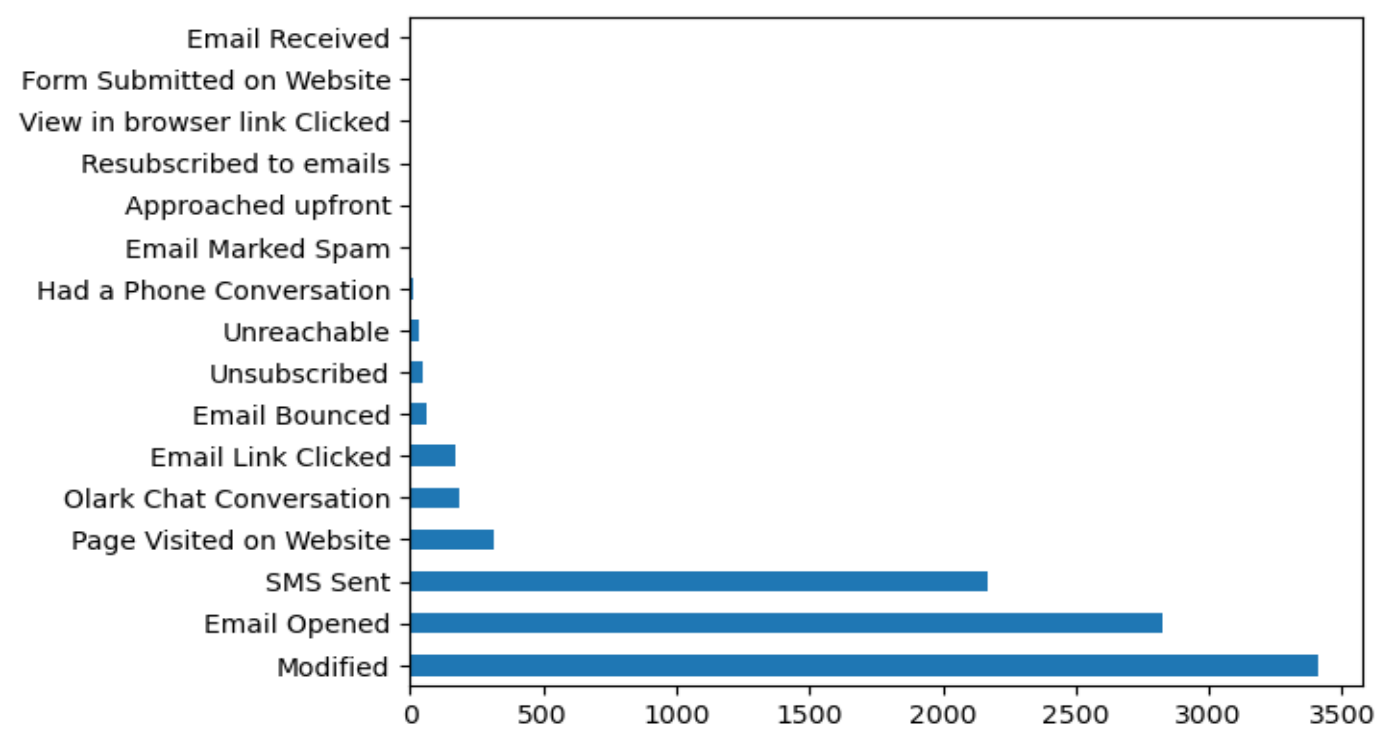
Lead Source Count

- The source of the lead mostly consist of Google, Direct traffic, Olark chat and Organic Search

UNIVARIATE ANALYSIS



Occupation and their count

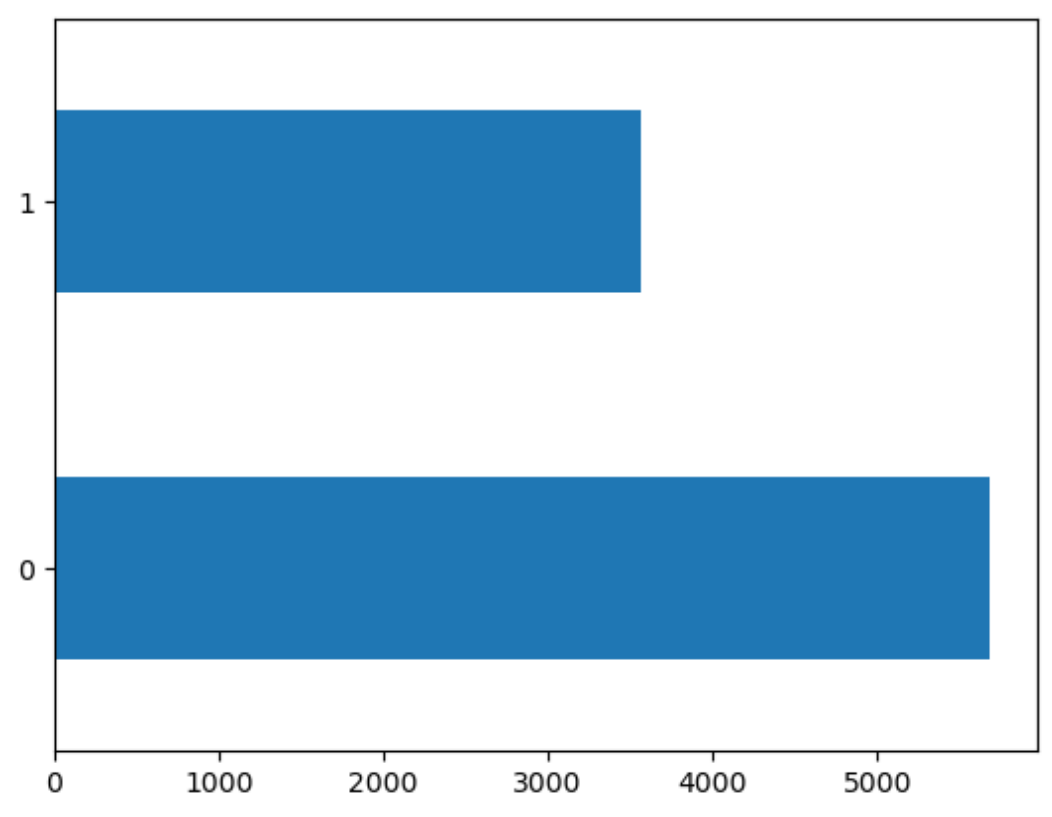


Last Notable Activity count

- More than 8000 of the leads are unemployed and other significant are working professional and student

- The last Notable activity among the leads are given where most of them are engaged in modification of details, viewing email or using SMS

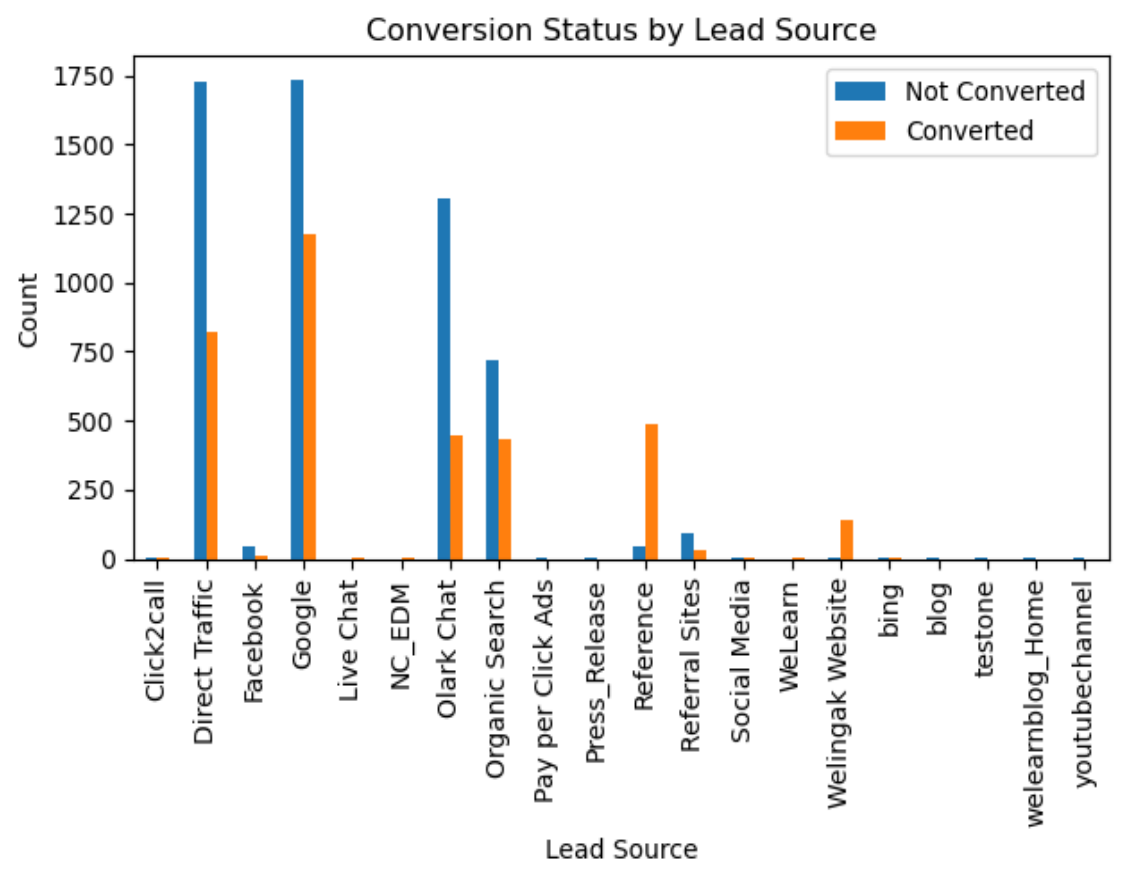
CHECKING DATA IMBALANCE

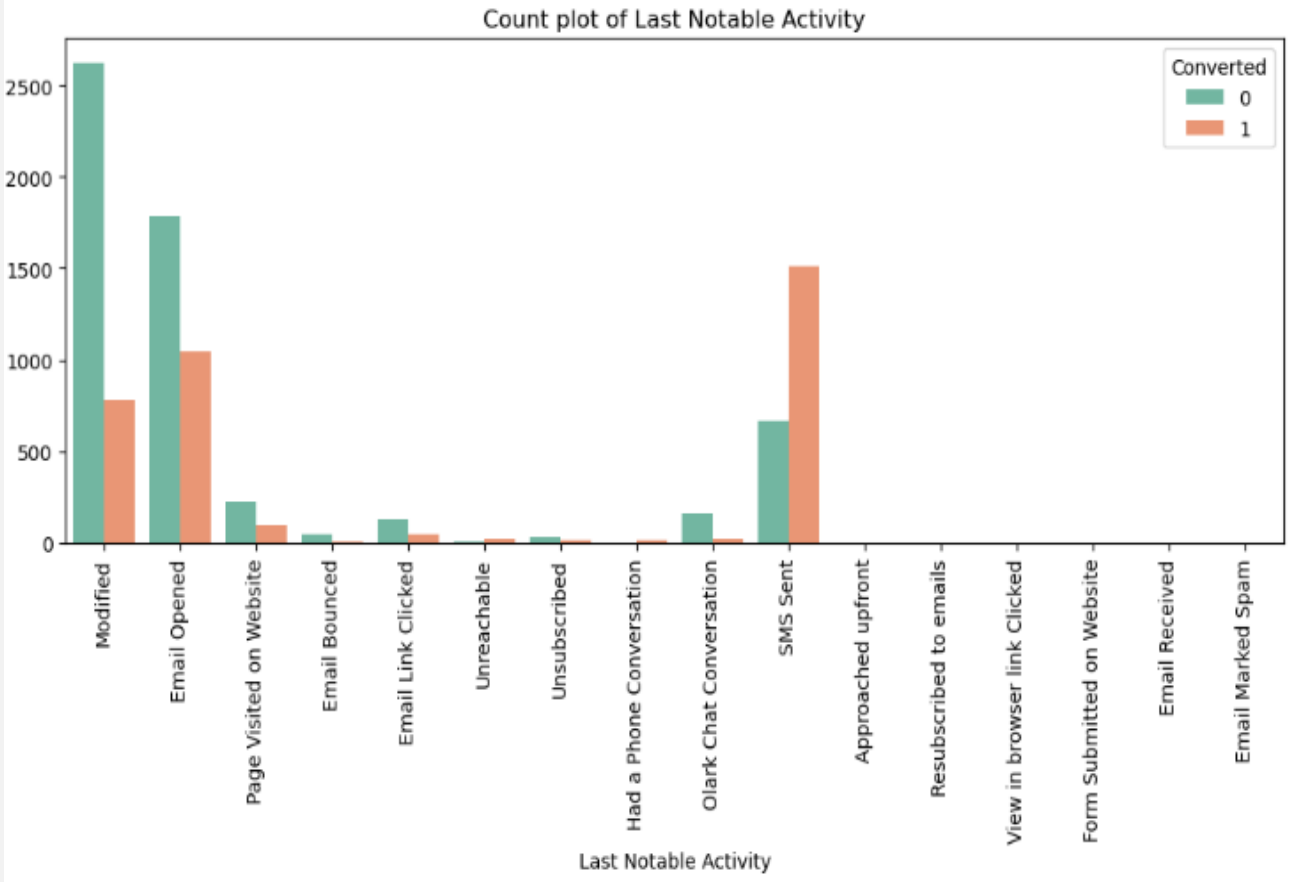


Count of converted and not converted

- There is small amount of imbalance , but this should be fine for making model
- Zero implied the lead is converted and One not converted
- The number of converted is around 3500 and not converted is above 5000

BIVARIATE ANALYSIS

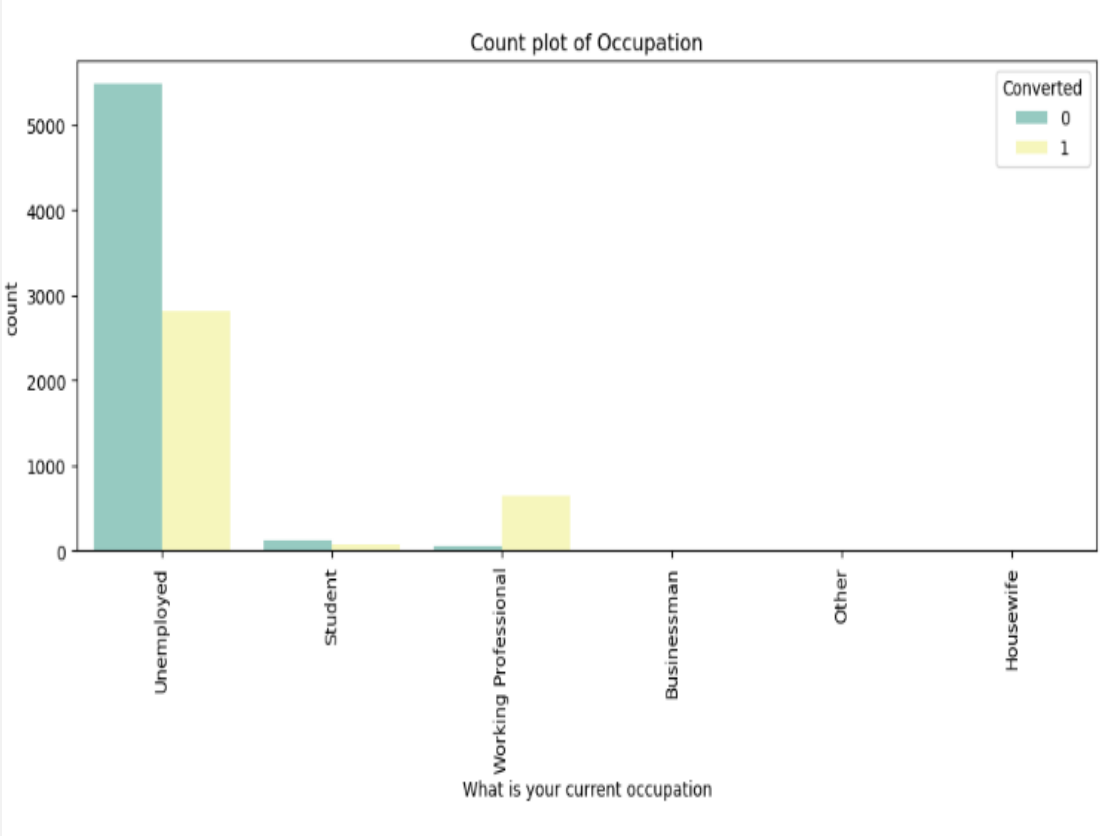




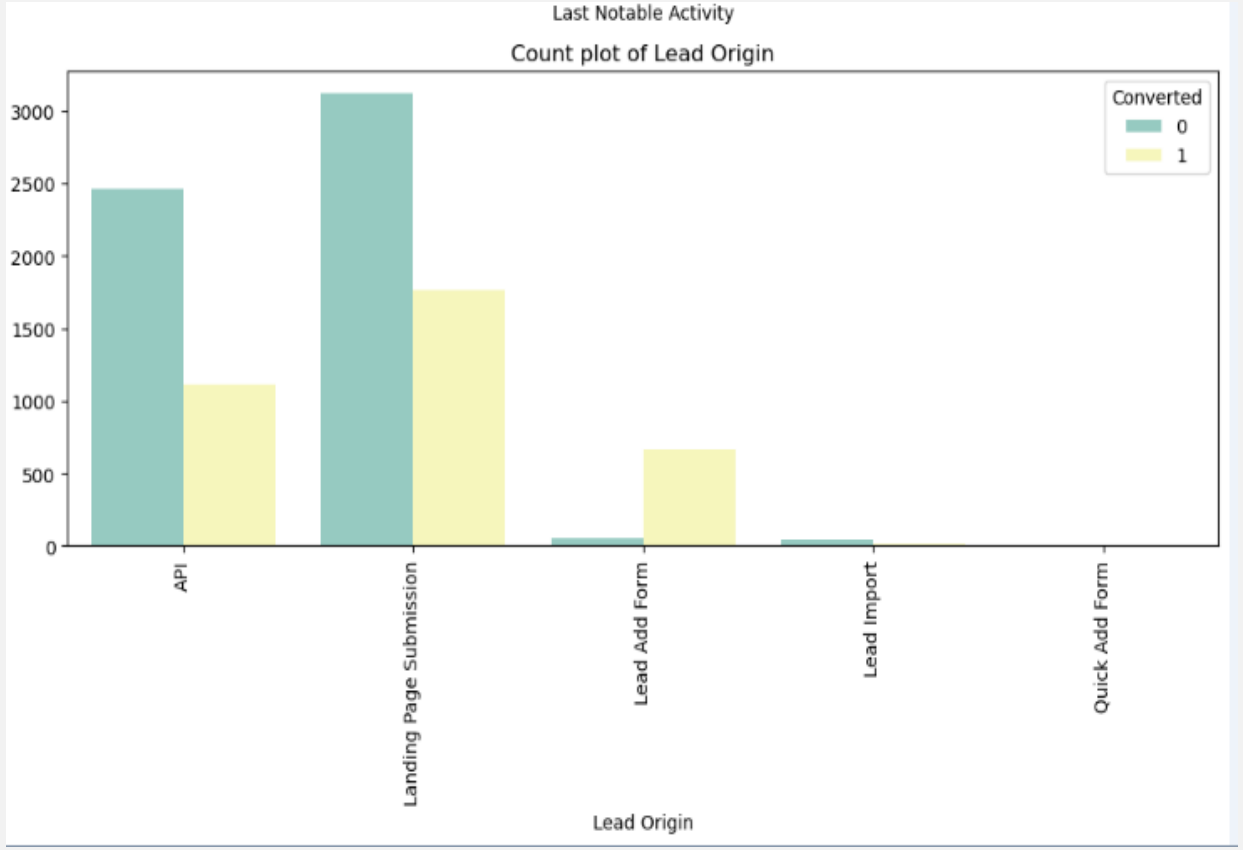
- Lead source is which has most conversion ratio by significant amount is shown by reference and Welingak website .

- Conversion ratio for a lead whose SMS was sent recently seem to be higher.
- Modified and email opened also show significant amount of converted however ratio remains low.

BIVARIATE ANALYSIS

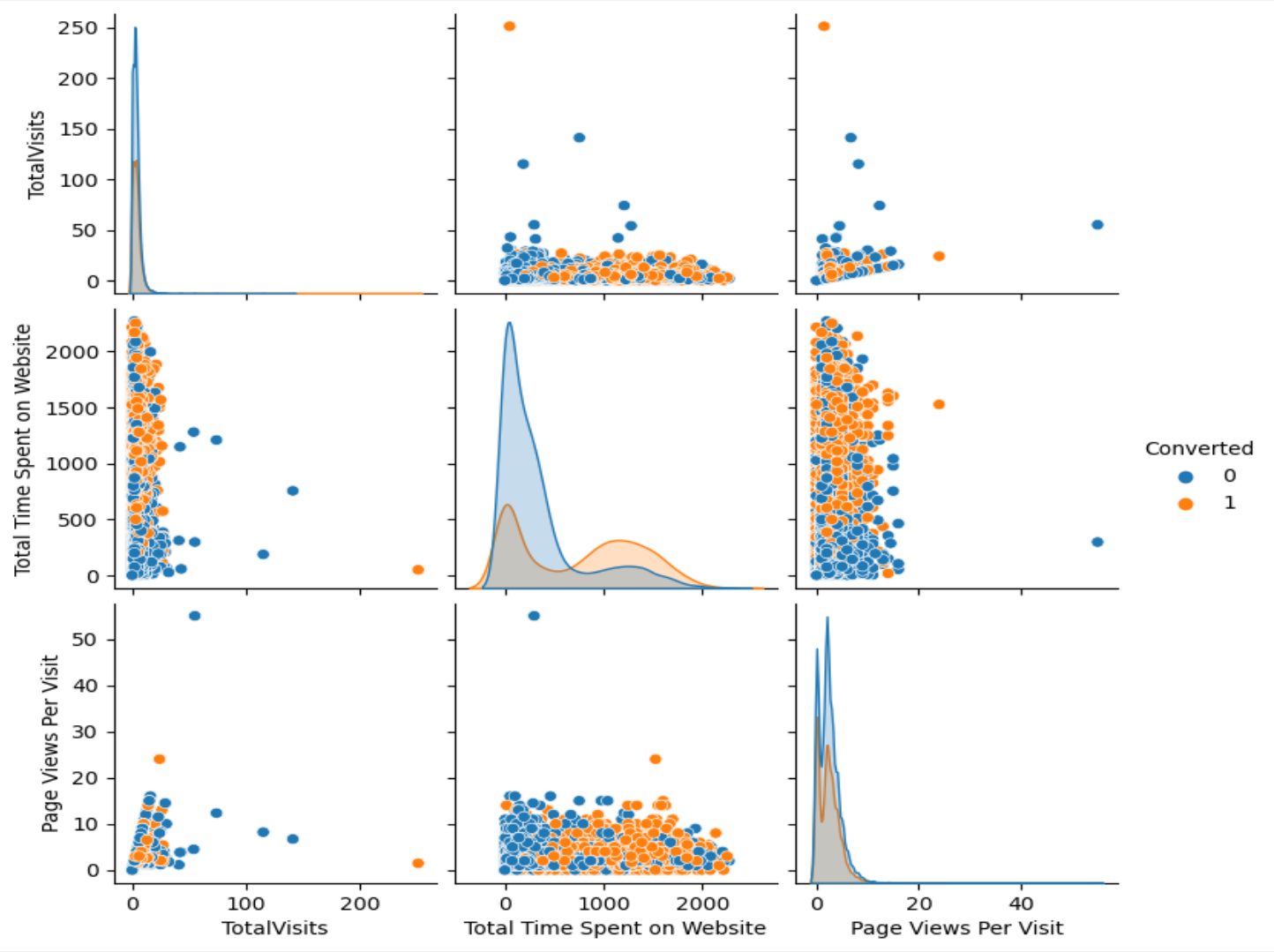


- Conversion rate of working professionals to hot leads is higher
- Unemployed leads however contribute to the significant amount of both converted and not converted leads



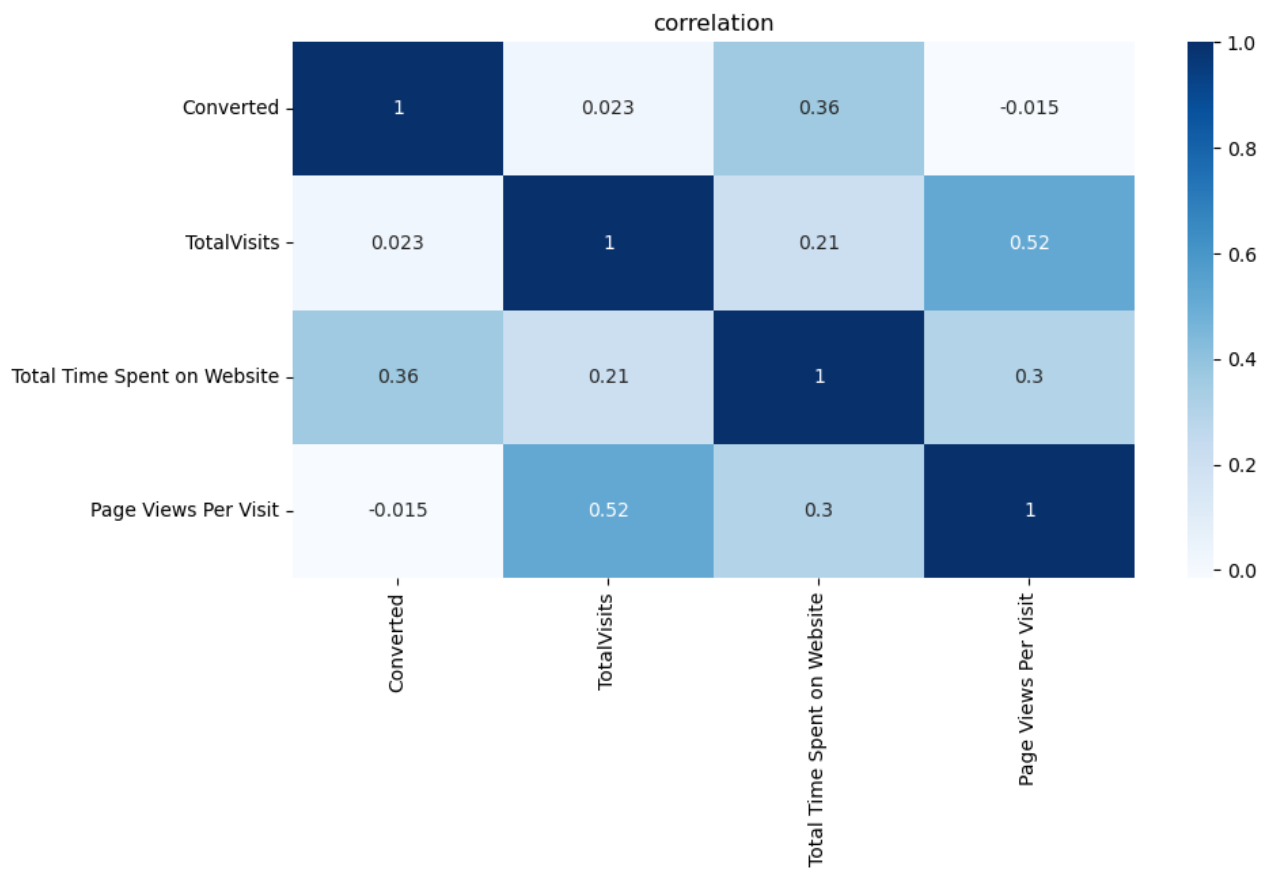
- Lead ADD form has the highest conversion ratio
- API and landing page contribute a large amount of conversion but they also contribute to significantly larger count of not converted.

MULTIVARIATE ANALYSIS



- People who have lower number Total visits but higher time spend on the website are showing higher conversion rate.

CORRELATION MATRIX



- The correlation matrix is showing slightly higher correlation of 0.52 between page views per visit and total visits
- Total time spend on Website is showing moderate 0.36 correlation between converted

DATA PREPARATION

- All the ‘Yes’ and ‘No’ are converted into 0s and 1s respectively
- The redundant factors are removed by checking correlation matrix and using domain knowledge
- Dummy variables are created for categorical variables
- The data set is split into test and train data sets 70% train and 30% test data
- The data set is scaled using Standard Scaler

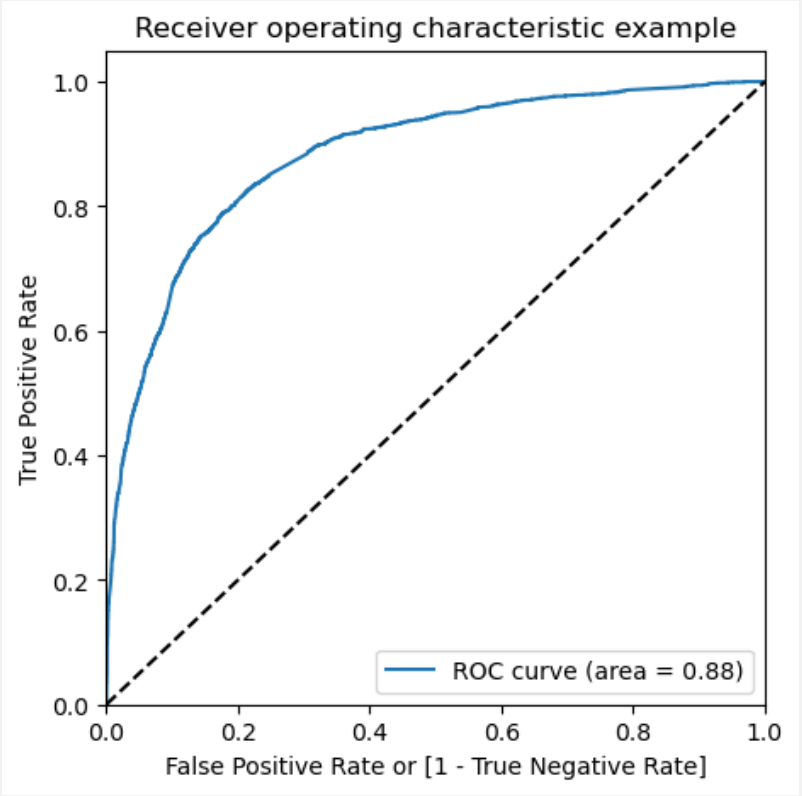
MODELLING

- The prepared data is taken for logistic regression
- Initial steps of RFE is done and the feature limit is set to 15
- After RFE the data is a 1st model is trained using logistic regression with the 15 columns
- VIF And P values are checked for the model and feature which is having high VIF values(>5) or p values (>0.05) is removed one by one
- The model is again trained on the data set with the removed feature and this process is repeated till all the all the feature's VIF and P-values are within acceptable range

MODEL EVALUATION

- Based on default of 0.5 cut off we are getting the following results

Confusion Matrix 1		
Actual/Predicted	Not Converted	Converted
Not Converted	3550	452
Converted	744	1722
Accuracy of our train model		82%
Sensitivity of our train model		70%
Specificity of our train model		89%



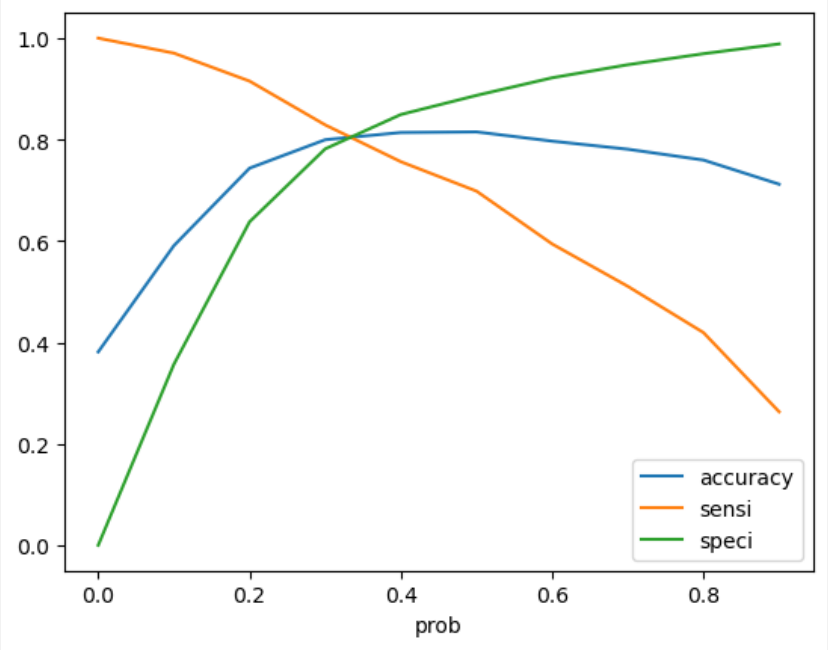
ROC CURVE

ROC curve area is 0.88 which indicates it is a good model

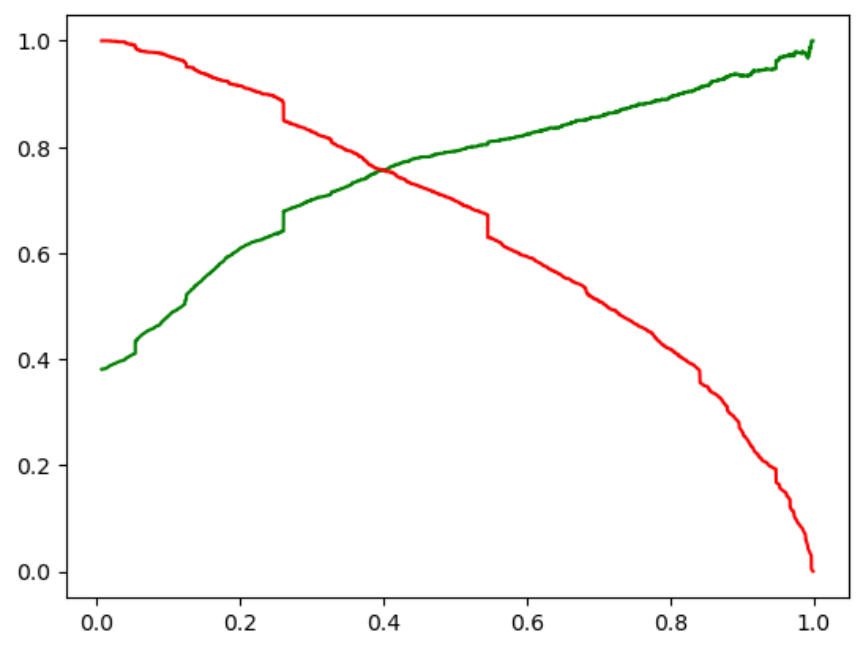
MODEL EVALUATION

- Model scores based on the cutoff of 0.35 the graph of accuracy sensitivity and specificity

Confusion Matrix 2		
Actual/Predicted	Not Converted	Converted
Not Converted	3264	738
Converted	510	1956
Accuracy of our train model		81%
Sensitivity of our train model		80%
Specificity of our train model		81%
precision_score:	72%	
recall_score:	79%	



PRECISION RECALL CURVE

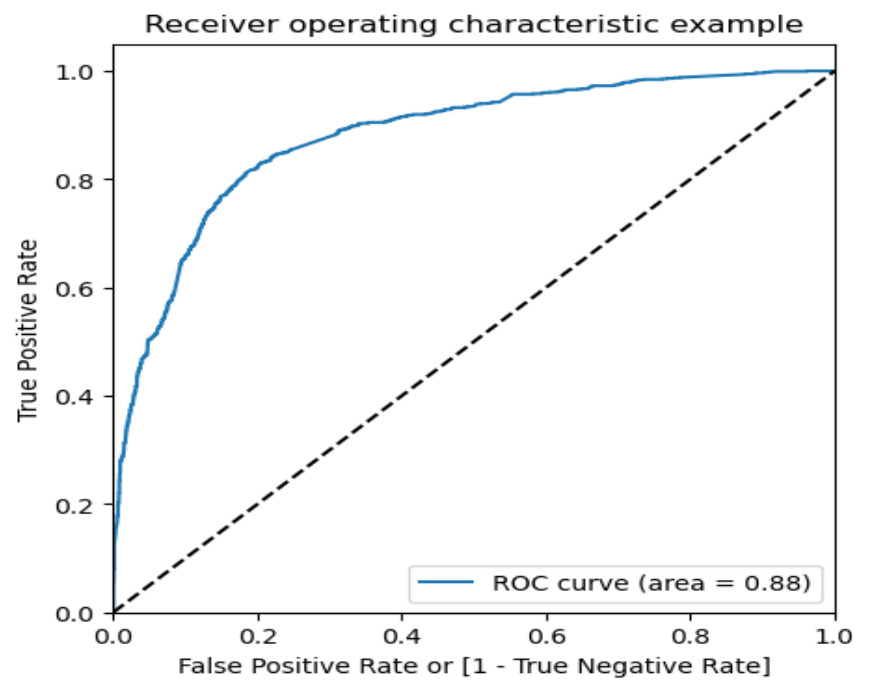


Precision Recall Curve

- From the precision-recall curve we are getting a value of 0.4 for cut-off

PREDICTIONS ON TEST SET

- Based on decision on 0.35 as optimal cutoff the model score are shown



- Area under ROC curve is 0.88 indicating a good model

Confusion Matrix 3		
Actual/Predicted	Not Converted	Converted
Not Converted	1392	285
Converted	230	865
Accuracy of our test model		81%
Sensitivity of our test model		79%
Specificity of our test model		83%

precision_score:	75%
recall_score:	79%

INTERPRETATIONS

- We are getting a model with the following

  On train data :

  - Accuracy : 81%
  - Sensitivity : 80%
  - Specificity :81%
  - Precision Score:72%
  - Recall Score:79%

- On test data:
  - Accuracy : 81%
  - Sensitivity:79%
  - Specificity:83%
  - Precision Score:75%
  - Recall Score 79%

- The model consistently performs well across different evaluation metrics for both training and test datasets. This means that the model is reliable, avoids overfitting, and generalizes effectively to new data. The absence of significant biases in its predictions further supports its positive performance. Overall, this indicates that the model is consistently accurate and trustworthy.

INSIGHTS AND RECOMMENDATION

- Google has contributed most for getting the leads to convert which is then proceeded by direct traffic.
- Most of the leads belongs to the unemployed occupation, working professional and student stood at 2 and 3 position respectively
- India is the country with most leads and it stood at about 96% of the total leads from any country.
- Leads from the google and direct traffic are converted the most
- Leads who have lower number Total visits but higher time spend on the website are showing higher conversion rate.

INSIGHTS AND RECOMMENDATION

- Leads who Origin is from Lead Add Form has a higher probability to become a hot lead
- Working professionals can be considered as a hot lead, so it is recommended for the sales team to pursue working professionals in getting a convert
- Leads whose source is from Olark Chat has a higher probability to become a hot lead. So any leads from Olark chat must be focused by the sales team.