Lead Score Case Study

- Submitted by :
  Prashik Bansod
  Ratna Singh

Problem Statement

An X Education company need help to select 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 you need to assign a lead score 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%.

Goals of the Case Study

•Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
•There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

Business Objective

- X Education wants to know most promising leads
- Build model which identifies the hot leads. (More likely to buy course)
- Deployment of model for future use.

Steps:

1. Reading and Understanding data
2. Data Cleaning
    - *Handling Null Values*
    - *Dropping insignificant columns*
    - *Data Imbalance*
3. Exploratory Data Analysis (EDA)
    - *Univariant Analysis (Categorical Columns)*
    - *Univariant Analysis (Numerical Columns)*
    - *Bivariant Analysis*
4. Prepare data for Model Building
    - *Converting binary fields Yes/No to 1/0*
    - *Dummy Variable method for other categorical columns*
5. Model Building
    - *Train Test Split*
    - *Feature Scaling*
    - *RFE (Recursive Feature Elimination)*
    - *VIF (For Manual Feature Elimination)*
    - *Predictions*
6. Model Evaluation
    - *Confusion Matrix*
    - *ROC Curve*
    - *Precision and Recall*
7. Prediction and Evaluation on test/unseen dataset
8. Conclusion

Data Understanding and Cleaning

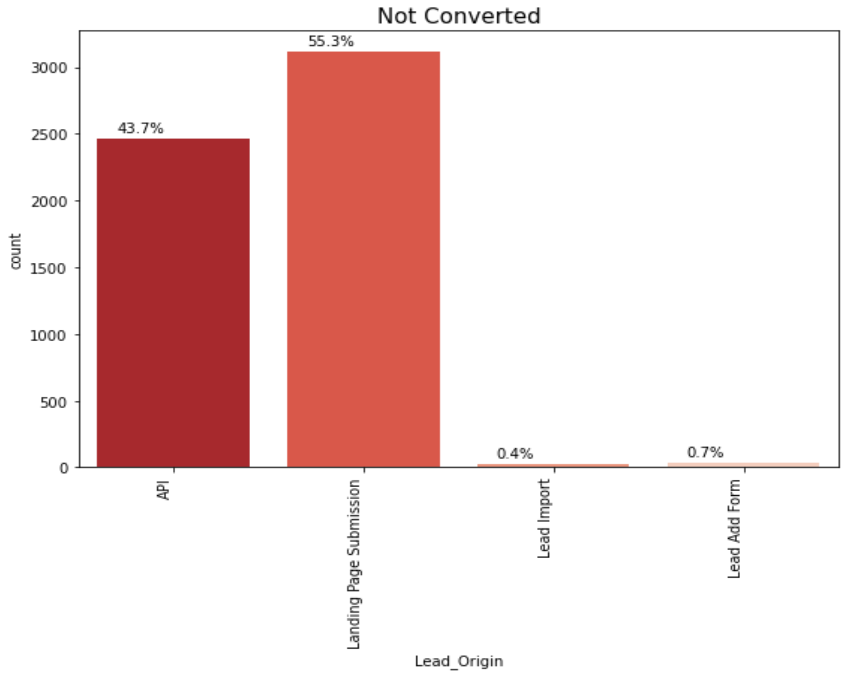
- Provided dataset **'leads.csv'** has 9240 rows and 37 columns
- For simplicity, renaming column names of data frame. By replacing ' ' with '_'
- The are many values as 'Select' in dataset, which is equivalent to null [As mentioned in problem statement]
  - Reason: In the UI, for the fields which has drop down for an option to select, if user doesn't select anything, the default value is 'Select'
  - Hence, replacing all such values as null
- Further cleaning
  - There are many columns with most of the values as NULL. Hence, dropping columns with null values more than 40%.
  - 'City' Column is also approx. to 40%, we can consider 'City' column also to drop.
  - Also, column 'Lead_Number' seems to be an identifier of each record, hence dropping.
  - Four columns with higher percentage of missing values:
    - Considering the problem statement, these columns seems to be important and hence, its not good to drop those columns.
    - Therefore, replacing the null values of with 'NA' (Additional category)
    - For other columns with minimal percentage of null values: Choosing rows with sum less than 1
- Data Imbalance
  - 38% which means data isn't imbalanced

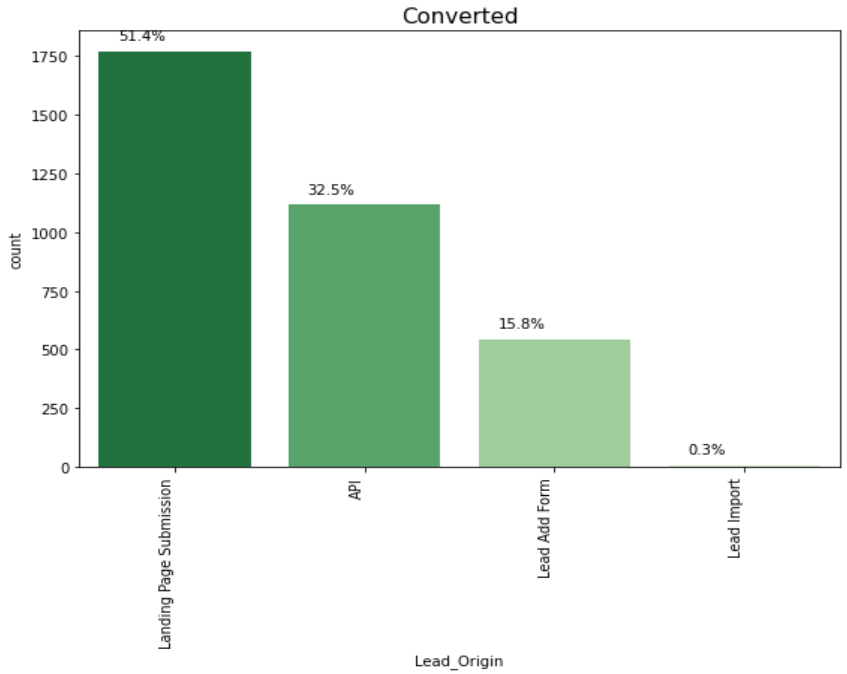
Exploratory Data Analysis (EDA)

- Create two different datasets based on target feature viz. "Converted"
  - df_0 -> records for which converted = 0
  - df_1 -> records for which converted = 1

Univariant Analysis (Categorical Column)

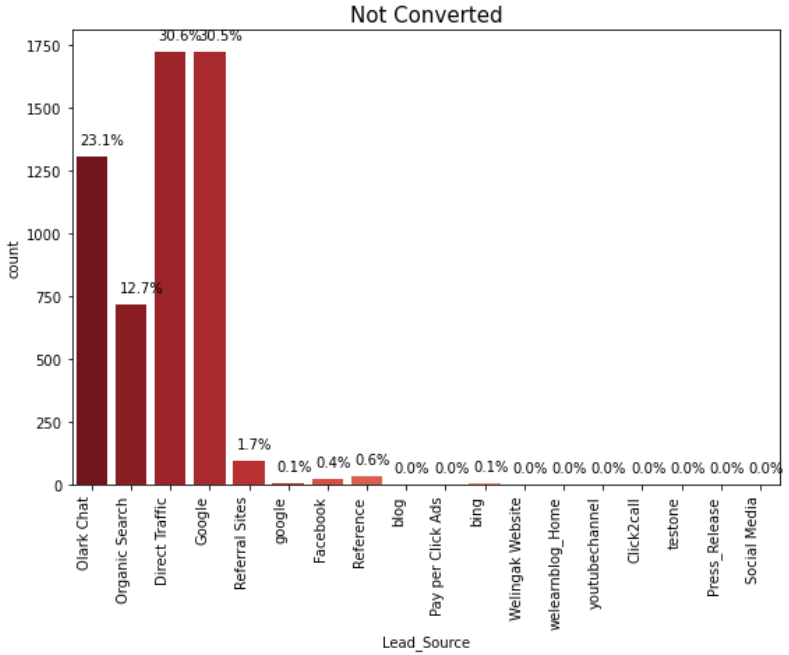
- New user defined function named as “plot_cat_cols” which will plot categorical columns based on target value
  - Eg: For Column “Lead Origin”

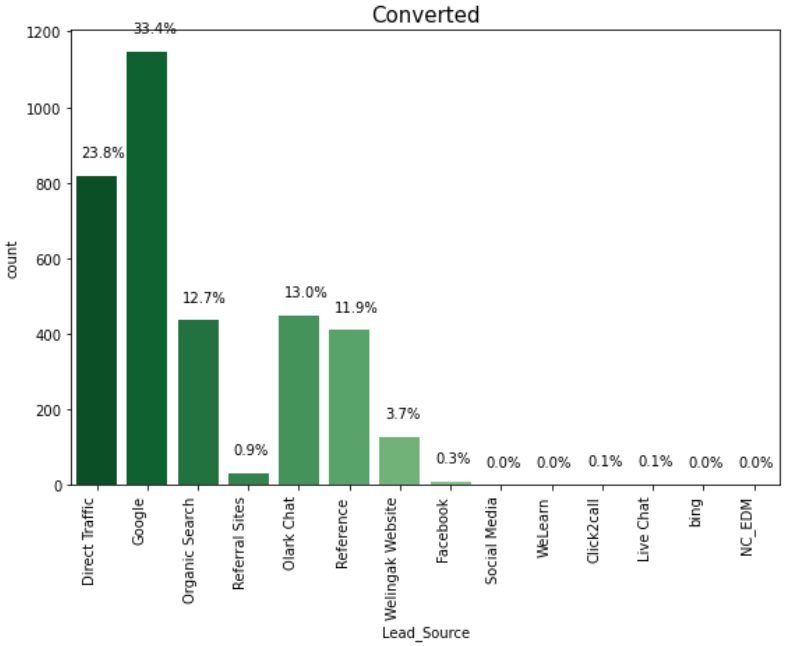




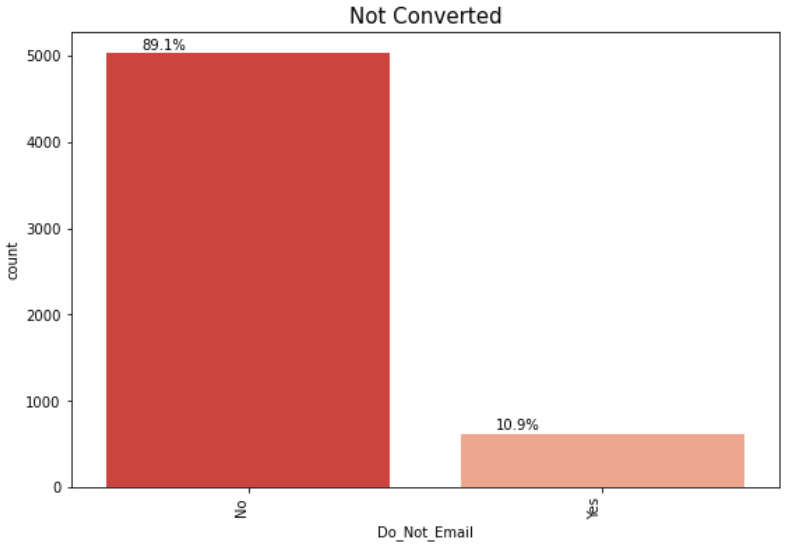
Exploratory Data Analysis (EDA)

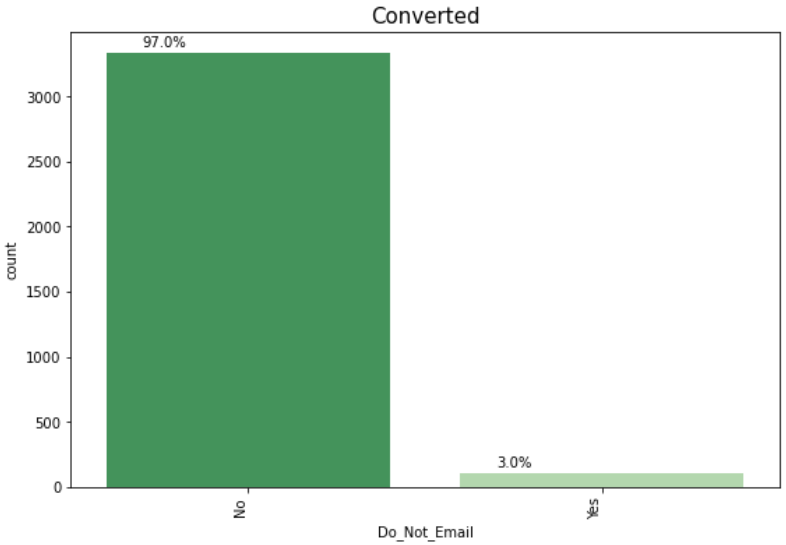
Column : Lead Score





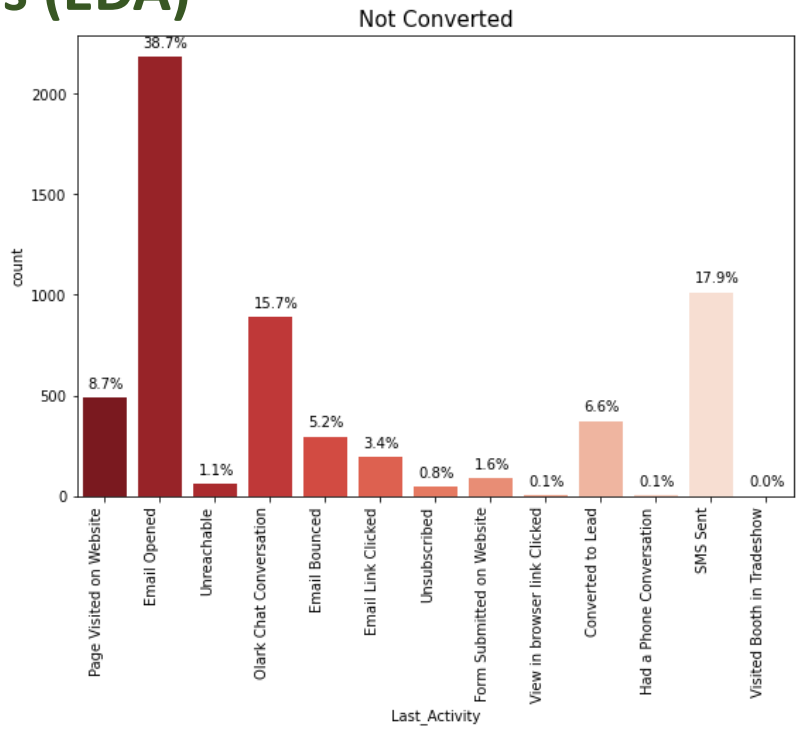
Column : Do Not Email

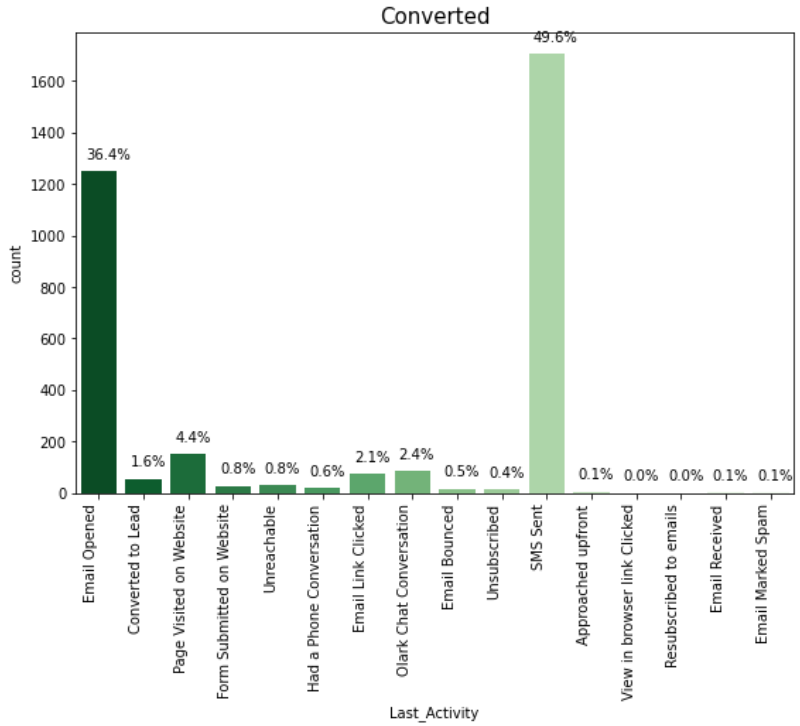




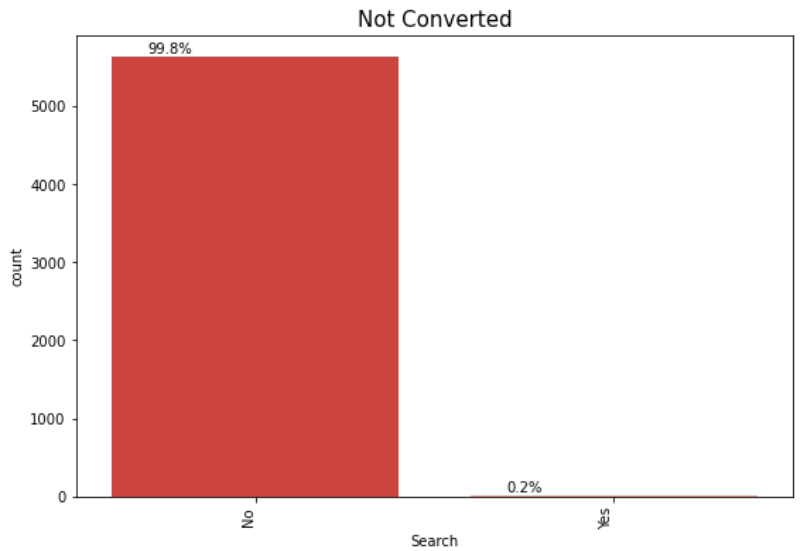
Exploratory Data Analysis (EDA)

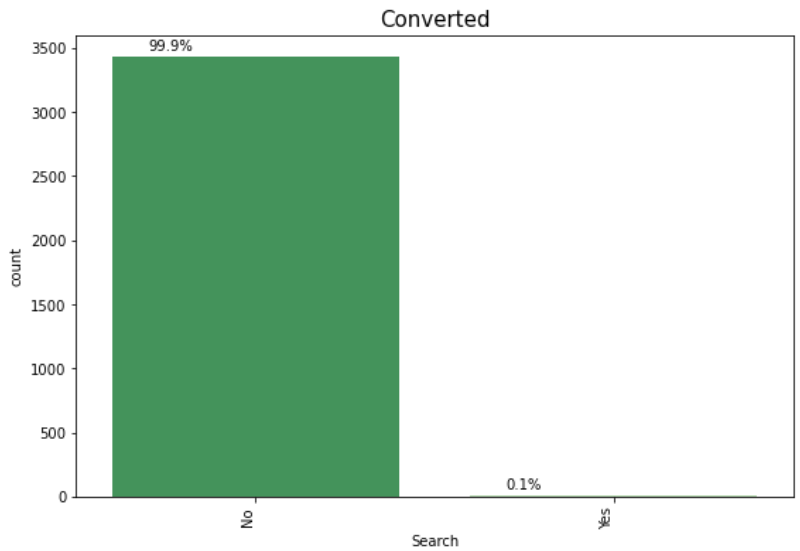
Column : Last Activity





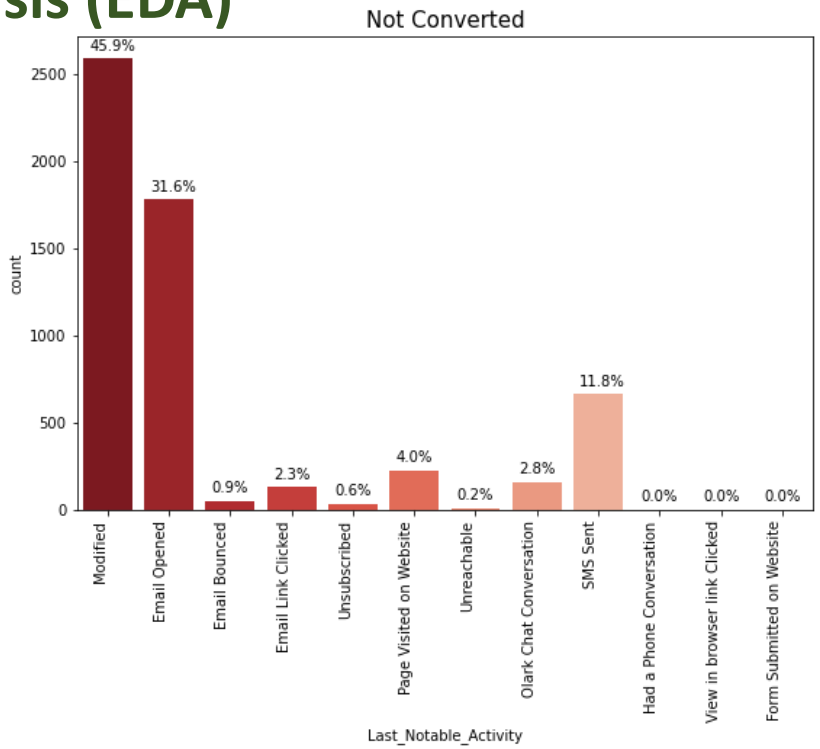
Column : Search

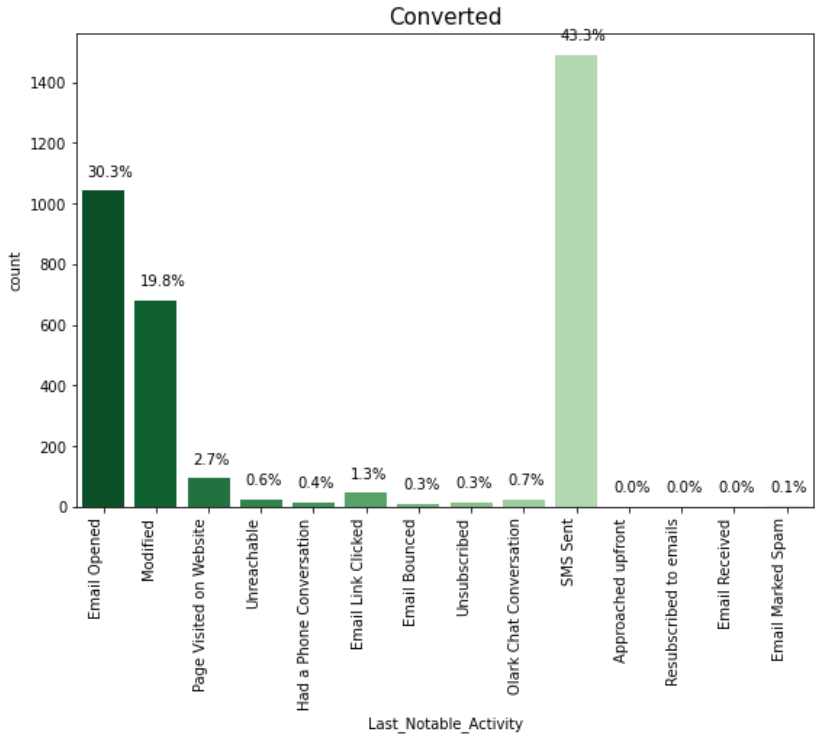




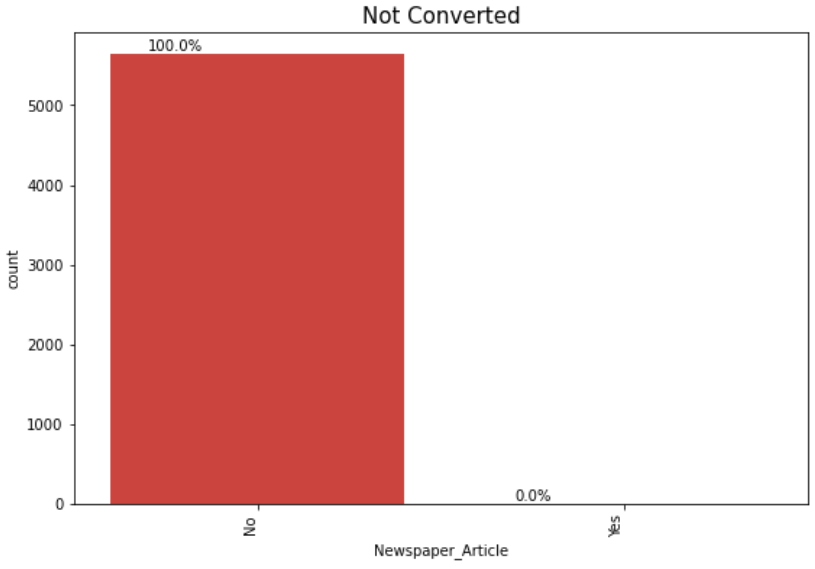
Exploratory Data Analysis (EDA)

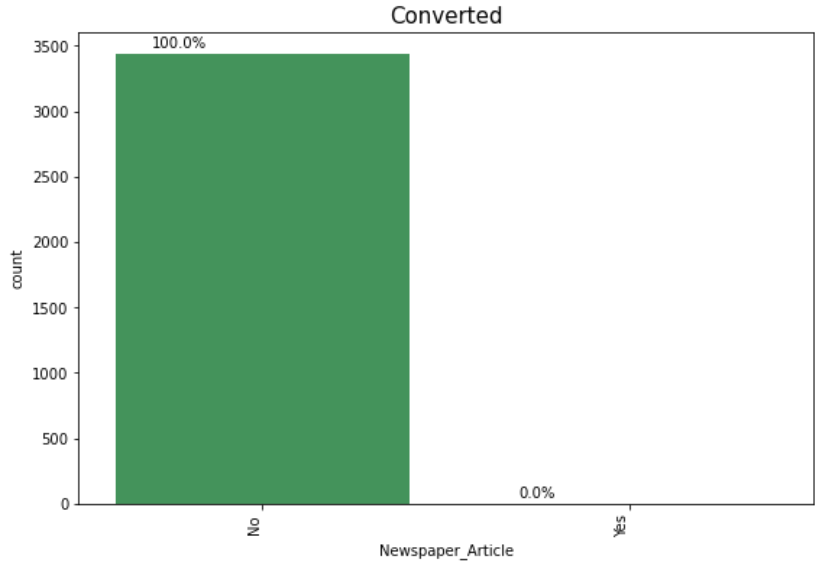
Column : Last Notable Activity





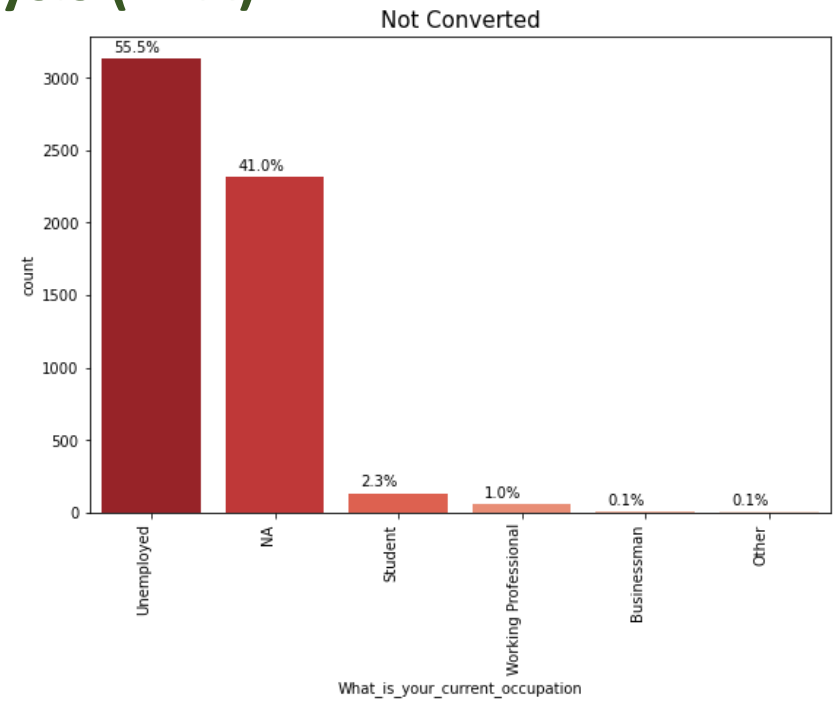
Column : Newspaper Article

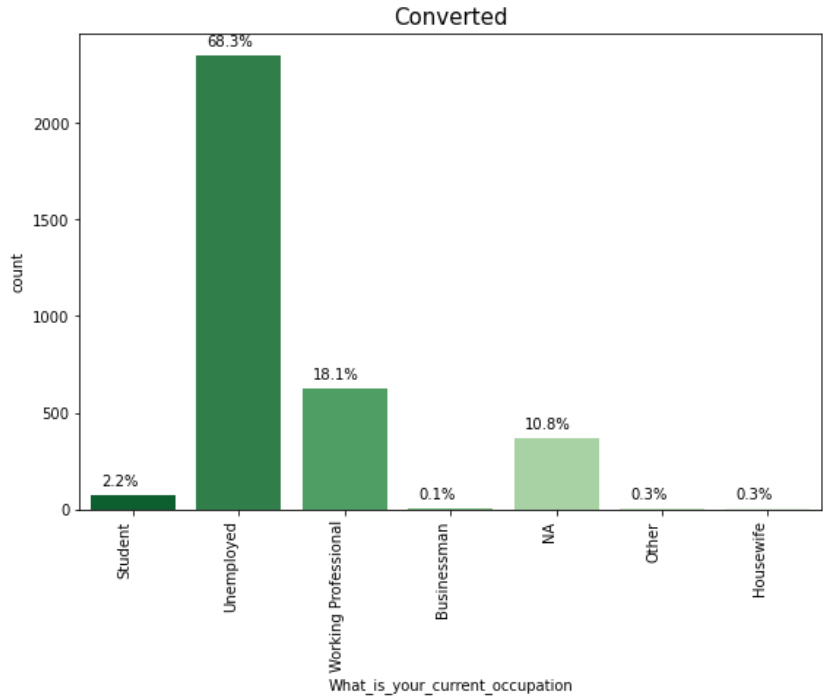




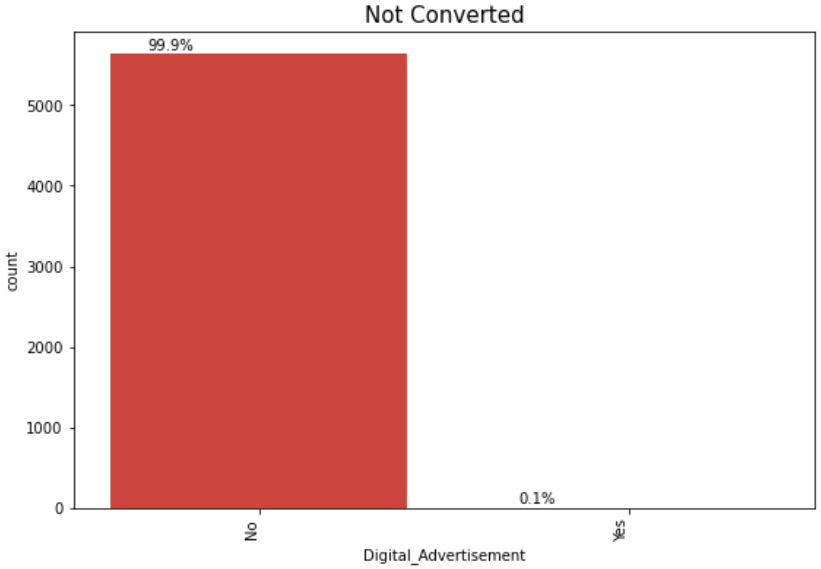
Exploratory Data Analysis (EDA)

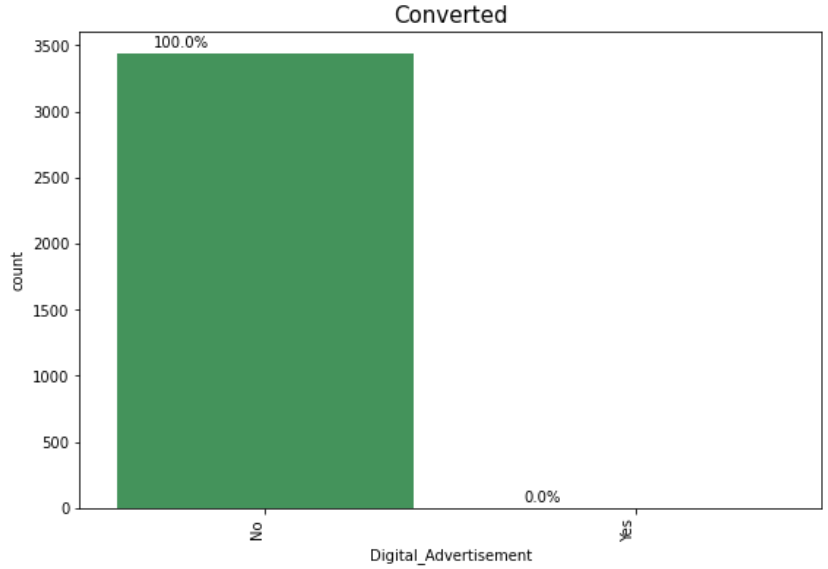
Column : Current Occupation





Column : Digital Advertisement



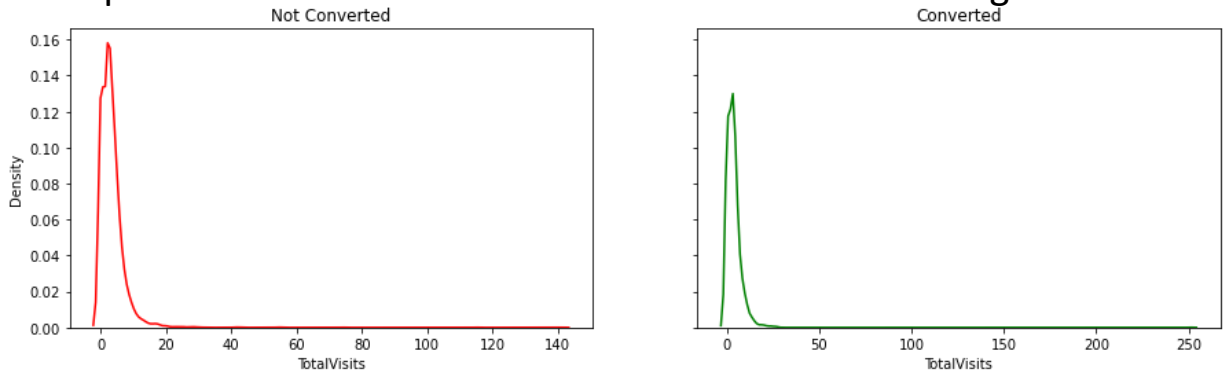


Exploratory Data Analysis (EDA)

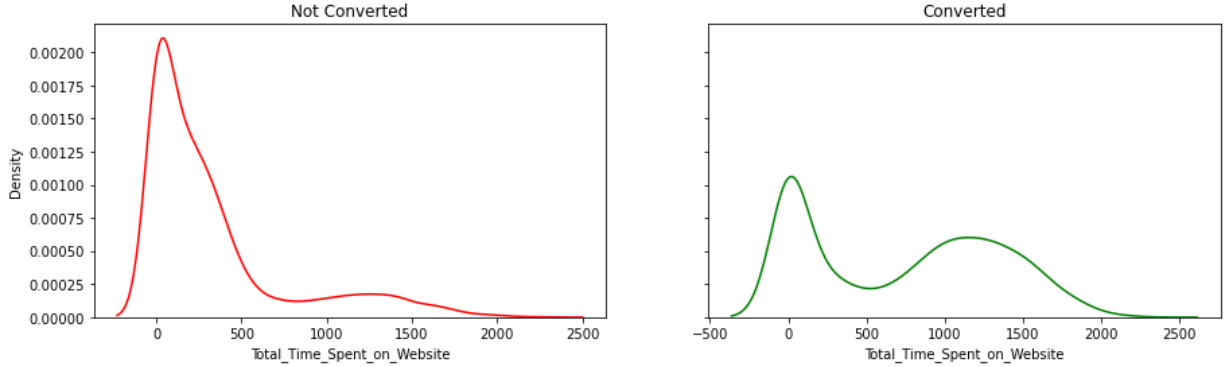
Univariant Analysis (Numerical Columns)

- New user defined function named as “plot_dist” which will plot distribution of numerical columns based on target variable.

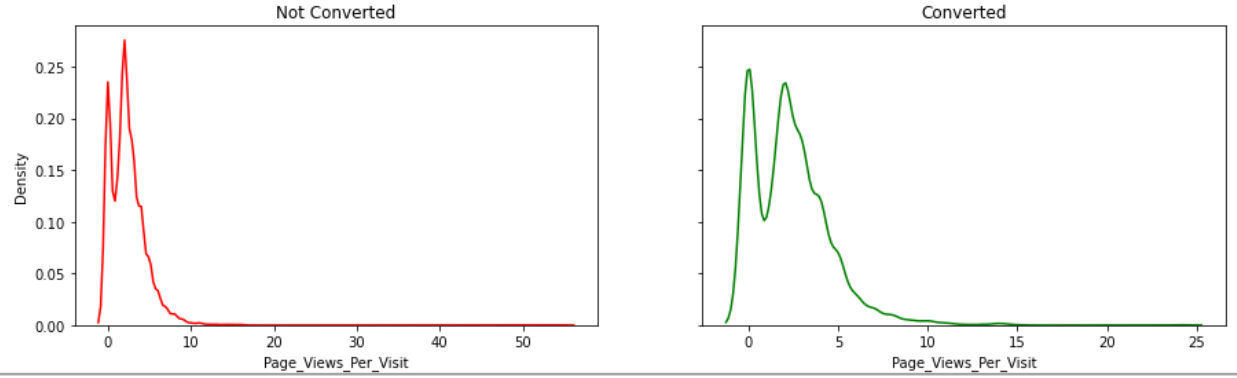
Column : Total Visits



Column : Total Time Spent On website



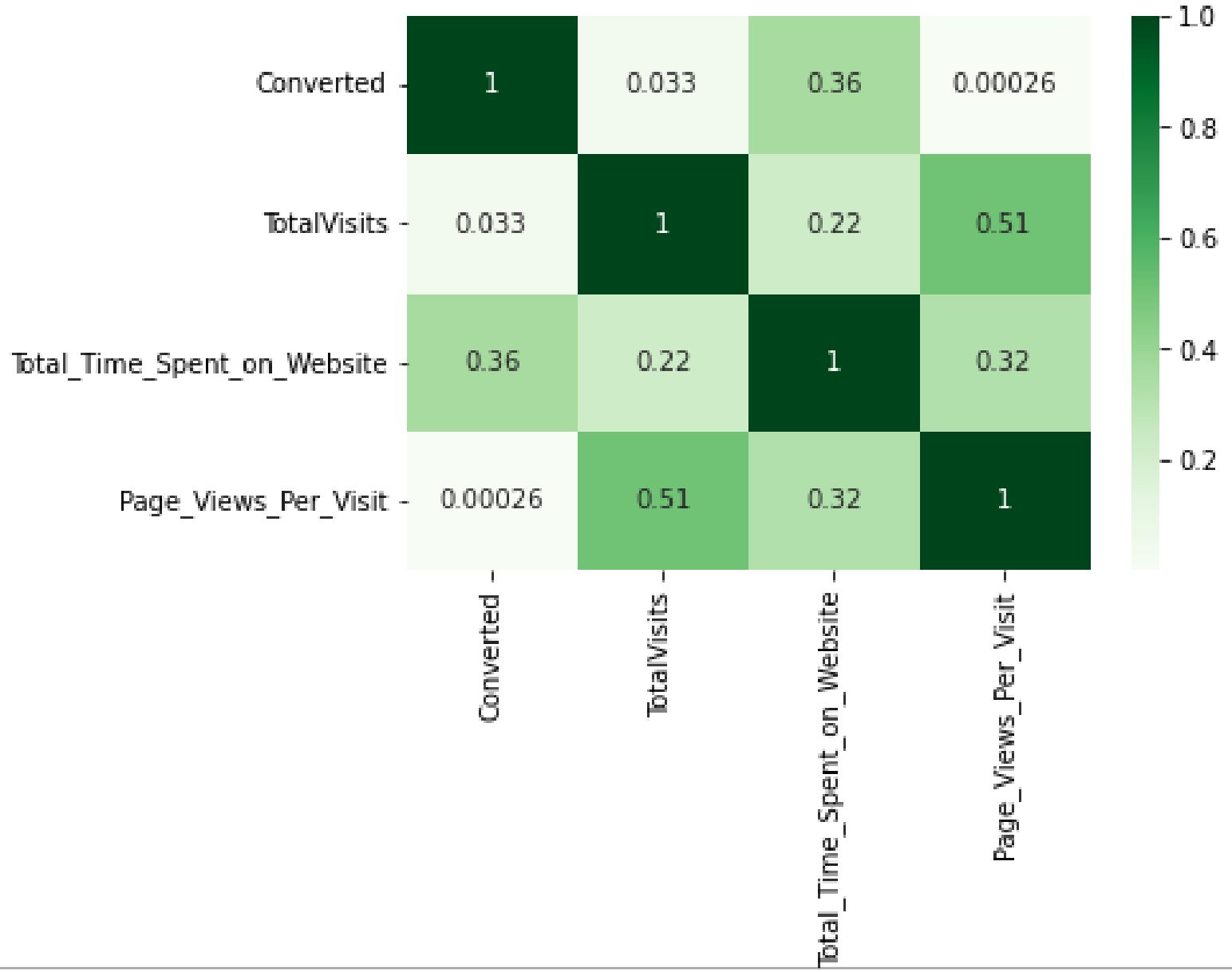
Column : Page Views Per Visit



Exploratory Data Analysis (EDA)

Bivariant Analysis

Analysing numerical variable's using heatmap:



Observation :

- As 'Converted' column is target, we can ignore in correlation
- Columns 'Page_Views_Per_Visit' and 'TotalVisits' seems to be correlated much if we compare but value is not much.
- Hence, no action

Others:

- In Country column, mostly it is India (Higher Percentage).
- Hence we can categorize country column as 'India' and 'Foreign'

Prepare data from Model building

- Transforming values of Country column. Categorizing as ’India’ and ‘Foreign’.
- Column 'A_free_copy_of_Mastering_The_Interview' which has binary value YES/NO. Converted to 1 and 0.
- Function which creates dummy columns to convert categorical columns to numerical. Also, drop the column with maximum length of its name.
- Dropping the columns which are converted using dummy variable method.
- Till here, we have cleaned the data and converted all columns to numerical for further processing. We can now start with Data Modelling part.

Model building

Train Test Split

- 70% as Training dataset and 30% as test dataset with random state as 100.

Feature Scaling

- Here, we will be using MinMaxScaler class to convert numerical columns into same scale as of others categorical columns
- This will convert value between 0 and 1

Model building

RFE (Recursive Feature Elimination)

- Using RFE, we have calculated top 15 features to consider for modelling.
- User defined function named as “build_statsmodel” which takes features dataset and target dataset and returns model
- This function is useful when we have to rebuild the model iteratively.
- First Model summary with columns given by RFE
  - Some of the column still has higher P-Value
  - There are high value of correlations present between 15 features, i.e, there is still come multicollinearity among the features

Manual Feature Elimination

- VIF – Variance Inflation Factor to check the multicollinearity
- User defined function named as “calculate_vif” which takes dataframe as input and list down VIF values of columsn in descending order.
- VIF values seems to be less than 5, however, there are still some features for which P-value is higher which makes it insignificant. Hence dropping such features and rebuilding model. Also, recalculate VIF values.
- By iterating, Model #3 is the model, for which P-values as well as VIF’s are OK.

Model building

Final Model

Summary

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6337
Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2628.7
Date:	Tue, 09 Aug 2022	Deviance:	5257.4
Time:	17:26:56	Pearson chi2:	6.25e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0149	0.181	0.082	0.934	-0.340	0.370
TotalVisits	5.9231	2.037	2.908	0.004	1.931	9.915
Total_Time_Spent_on_Website	4.6598	0.166	28.016	0.000	4.334	4.986
Lead_Origin_Lead_Add_Form	2.2139	0.226	9.781	0.000	1.770	2.658
Lead_Source_Welingak_Website	1.9647	0.755	2.601	0.009	0.484	3.445
Last_Activity_Email_Bounced	-2.0924	0.372	-5.626	0.000	-2.821	-1.363
Last_Activity_Olark_Chat_Conversation	-1.2718	0.163	-7.803	0.000	-1.591	-0.952
Last_Activity_SMS_Sent	1.2366	0.074	16.677	0.000	1.091	1.382
Country_NA	1.5704	0.112	13.990	0.000	1.350	1.790
What_is_your_current_occupation_NA	-3.6120	0.187	-19.285	0.000	-3.979	-3.245
What_is_your_current_occupation_Student	-2.2023	0.273	-8.068	0.000	-2.737	-1.667
What_is_your_current_occupation_Unemployed	-2.3365	0.174	-13.410	0.000	-2.678	-1.995
Last_Notable_Activity_Had_a_Phone_Conversation	3.6812	1.120	3.287	0.001	1.486	5.876
Last_Notable_Activity_Unreachable	2.0852	0.486	4.290	0.000	1.133	3.038

Variance Inflation Factor

	Feature	VIF
10	What_is_your_current_occupation_Unemployed	2.56
7	Country_NA	2.38
1	Total_Time_Spent_on_Website	2.04
2	Lead_Origin_Lead_Add_Form	1.81
8	What_is_your_current_occupation_NA	1.79
0	TotalVisits	1.60
6	Last_Activity_SMS_Sent	1.56
5	Last_Activity_Olark_Chat_Conversation	1.43
3	Lead_Source_Welingak_Website	1.31
4	Last_Activity_Email_Bounced	1.07
9	What_is_your_current_occupation_Student	1.06
12	Last_Notable_Activity_Unreachable	1.01
11	Last_Notable_Activity_Had_a_Phone_Conversation	1.00

Model building

Predictions

- Use predict() method of final model on train dataset
- Create New Data frame with columns as “Converted” (from training dataset) and “Conversion_probability” (From model prediction)
- Use default Threshold as 0.5.
- Here we have chosen threshold as 0.5 at random.
- Sample Dataset:

	Converted	Conversion_prob	predicted
0	0	0.061875	0
1	0	0.163485	0
2	0	0.619053	1
3	1	0.640064	1
4	1	0.912853	1

- The predicted values definitely will have some errors, like:
  - - 'Converted' customers are predicted as 'Not Converted’
  - - 'Not Converted' customers are predicted as 'Converted'

Model Evaluation

- The simplest model evaluation metric for classification models is accuracy - it is the percentage of correctly predicted labels
- **Accuracy** = (Correctly predicted labels)/(Total number of labels)
- Accuracy = 81%

Confusion Matrix

Actuals/Predicted	Not Converted	Converted
Not Converted	3444	461
Converted	741	1705

- Sensitivity - Number of actual Yeses correctly predicted/Total number of actual Yeses

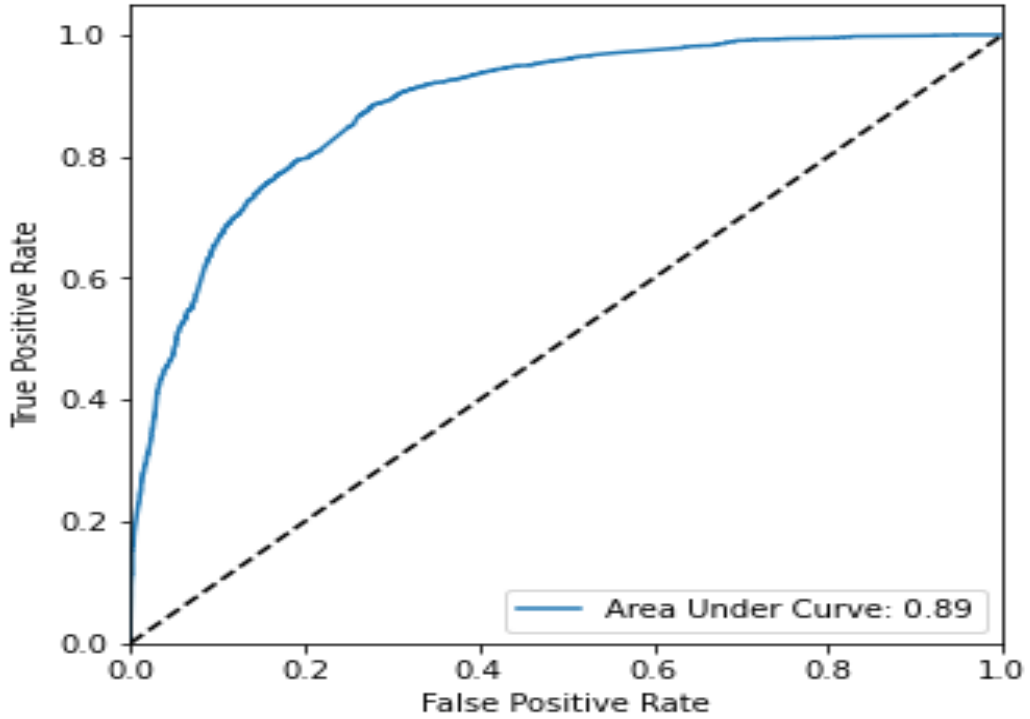
  Sensitivity = TP/(TP+FN)
- Specificity - Number of actual Noes correctly predicted/Total numer of actual Noes

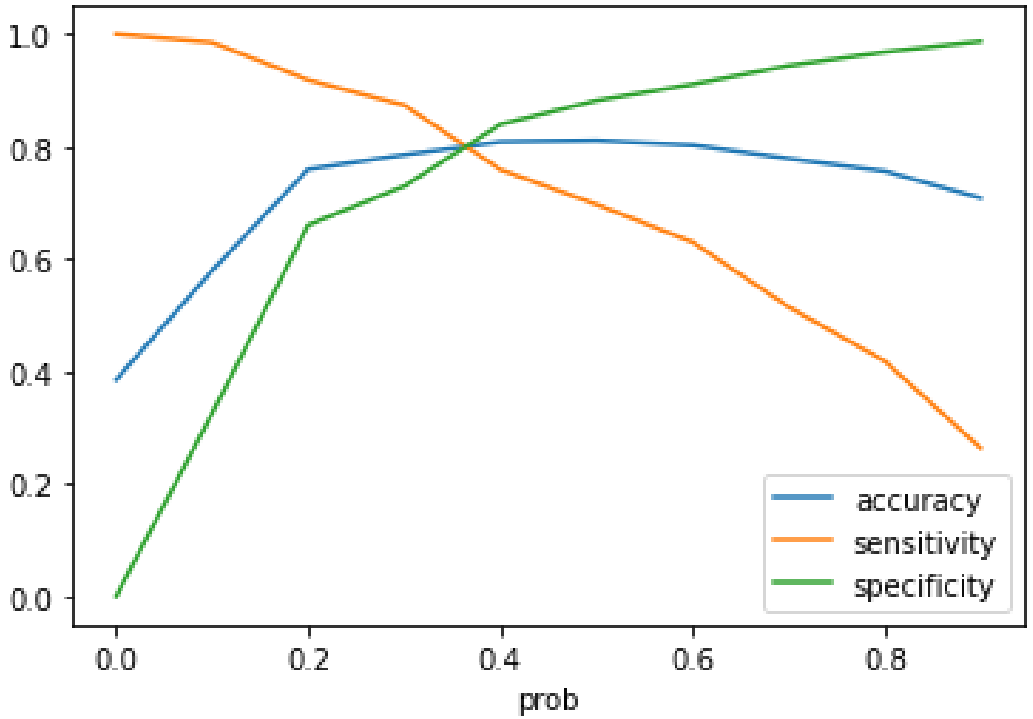
  Specificity = TN/(TN+FP)
- False Positive Rate , FPR = FP/(TN+FP)
- Positive Predictive Value, PPV = TP/(TP+FP)
- Negative Predictive Value, NPV = TN/(TN+FN)
- Even though accuracy is 81%, the sensitivity is 69%
- **i.e. out of total predicted values, 69% are correctly predicted my model**

Model Evaluation

ROC Curve

- ROC (Receiver Operating Characteristic) Curve shows trade off between True Positive Rate (Sensitivity) and False Positive Rate (1 - Specificity)



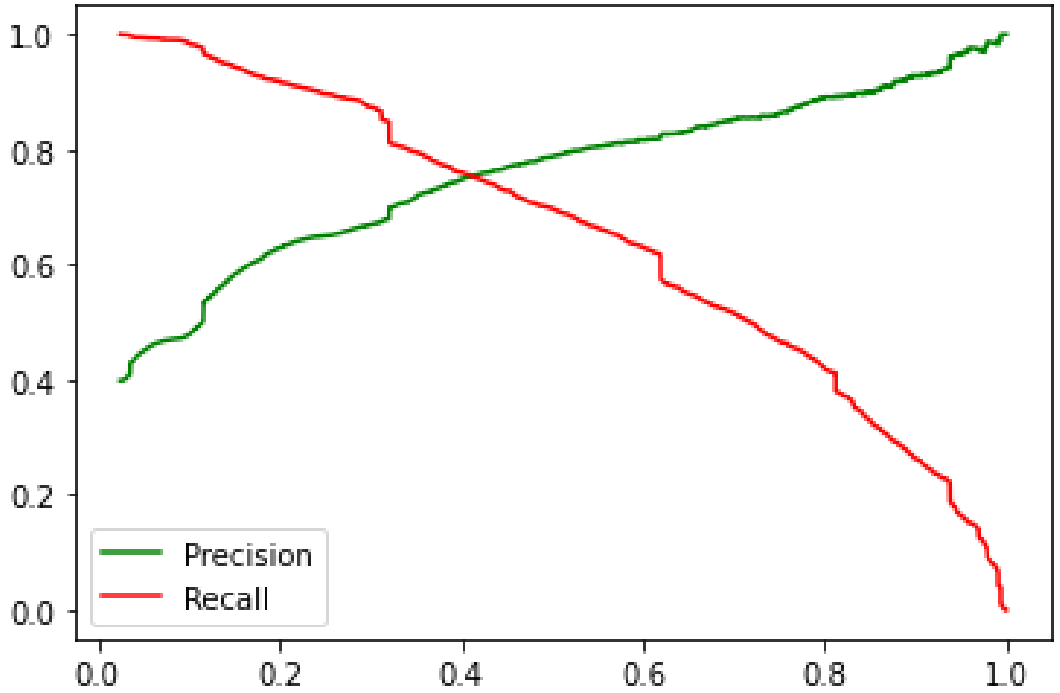


- The Area under ROC curve is 0.89, which is good
- Threshold for max ROC, by choosing 10 different threshold between 0 and 1
- From second graph, **0.37 seems to be optimal threshold.**

Model Evaluation

Precision Recall Curve

- Precision - Out of total Yeses predicted, how many are actually Yes
  - Precision = TP/(TP + FP)
- Recall - Out of total Yeses actual, how may yeses are predicted
  - Recall = TP/(TP + FN)



- Precision of final model : 72%
- Recall of final model : 78%

Prediction and Evaluation on test/unseen dataset

- Using final model, i.e. model number 3 to predict the targets of test dataset.
- Doing feature scaling of test dataset by using scalar class object of train dataset
- Using only those columns of test dataset which are used from model building from train dataset.
- To predict, converted or not from conversion probability, used threshold as 0.37 as given by ROC curve.
- Using final model, metrics of test/unseen dataset

```
Precision             : 72.2% -> Total number of actual 1's from all predicted 1's
Recall/Sensitivity    : 78.0% -> Total numbers of predicted 1's from actual no. of 1's
Specificity           : 83.0% -> Total numbers of predicted 0's from actual no. of 0's
Accuracy              : 81.0% -> Total predicted 1's and 0's from actual 1's and 0's
```

Conclusion

It was found out that variables that mattered the most in the potential buyers are:

- Total number of visits.
- The total time spent on website.
- The last notable activity
  - Had a phone conversation
  - Unreachable
- When the lead source was :
  - Wellingak website
- When the last activity was :
  - SMS
  - Olark chat conversation
  - Email_Bounced
- When the lead origin is lead add form.
- When the current occupation was :
  - Student
  - Unemployed
  - Other

Keeping the above mentioned points in mind the X Education can increase all the potential buyers to change their mind and buy there courses. Company shouldn't make call to the leads whose last activity was "Olark Chat Conversation", "Email bounced" as well as whose current occupation is "student", "NA", and "unemployed" as they are almost unlikely to convert.