LOGISTIC REGRESSION ASSIGNMENT

(Lead Scoring case study)

Submitted by

Rekha V S

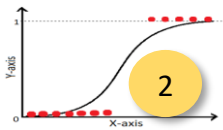
PROBLEM STATEMENT & BUSINESS OBJECTIVE

Problem statement:

- An education company 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%.

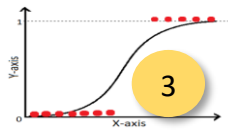
Business Objective:

- X Education need to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company has to build a model 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.
- A ballpark of the target lead conversion rate to be around 80%.



METHODOLOGY & WAY FORWARD

- 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 the model should be able to adjust to if the company's requirement changes in the future so they need to be handled as well.



STEPS FOLLOWED

Data Understanding & Exploration
- Reading the data from source
- Understanding the data and inspecting the data frame

Data Cleaning
- Missing value imputation
- Outlier identification & treatment
- Treatment for 'Select' response

Exploratory Data Analysis
- Categorical variable analysis
- Numerical variable analysis

Data Preparation
- Dummy variable creation
- Train & Test data split
- Variable scaling

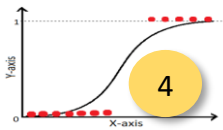
Model building
- Feature selection using RFE
- Manual feature elimination based on p-values and VIFs

Model Evaluation
- Evaluation using various evaluation metrics

Drawing ROC Curve
- Checking the area under ROC curve

Finding optimal cut off point
- Finding the optimal probability threshold

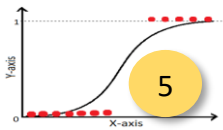
Prediction on test data
- Making prediction on test data
- Evaluate the final prediction on the test set
- Determine the lead score and check for final prediction's conversion rate



DATA UNDERSTANDING & EXPLORATION

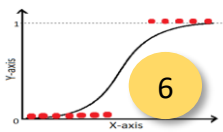
- Read the data from the 'Leads' CSV file and understood the meaning of variables from 'Leads Data Dictionary'
- The dataset contains total of 9240 rows and 37 columns
- There are no duplicates in the data
- There are missing values in the dataset, with few columns having >25% of null values which needs to be treated
- Datatypes of the column : 4 float64, 3 int64, 30 object (7 Numeric and 30 categorical columns)
- There are 4 columns with 'Select' which needs to be handled

Note: Only key points/insights from the analysis is added in the presentation in the best possible '**concise**' way for all the steps.
Other analysis are detailed in the notebook



DATA CLEANING

- Dropped the columns not useful for analysis like Record index/unique IDs ['Lead Number', 'Prospect ID'], Internally assigned scores ['Asymmetrique Activity Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Index', 'Asymmetrique Profile Score']
- Replaced 'Select' with NaN as select indicates no options were selected ['Specialization', 'How did you hear about X Education', 'Lead Profile', 'City']
- Treated the missing value
  a) Dropped the columns with missing values > 50% [E.g., 'Lead profile', 'Lead Quality etc]
  b) Column by column analysis and imputation with apt means like median [E.g., 'Total visits'] or with Others/Not specified etc [E.g., Specializaton', 'Tags', etc] wherever the data is not available, but cannot be imputed with other entries
- Outlier treatment ['TotalVisits', 'Total Time Spent on Website','Page Views Per Visit']
  a) Removed top 5% of the column outlier values considering difference between 95% quantile and max value
  b) No capping done at bottom 5% as the value is same as min value for all columns

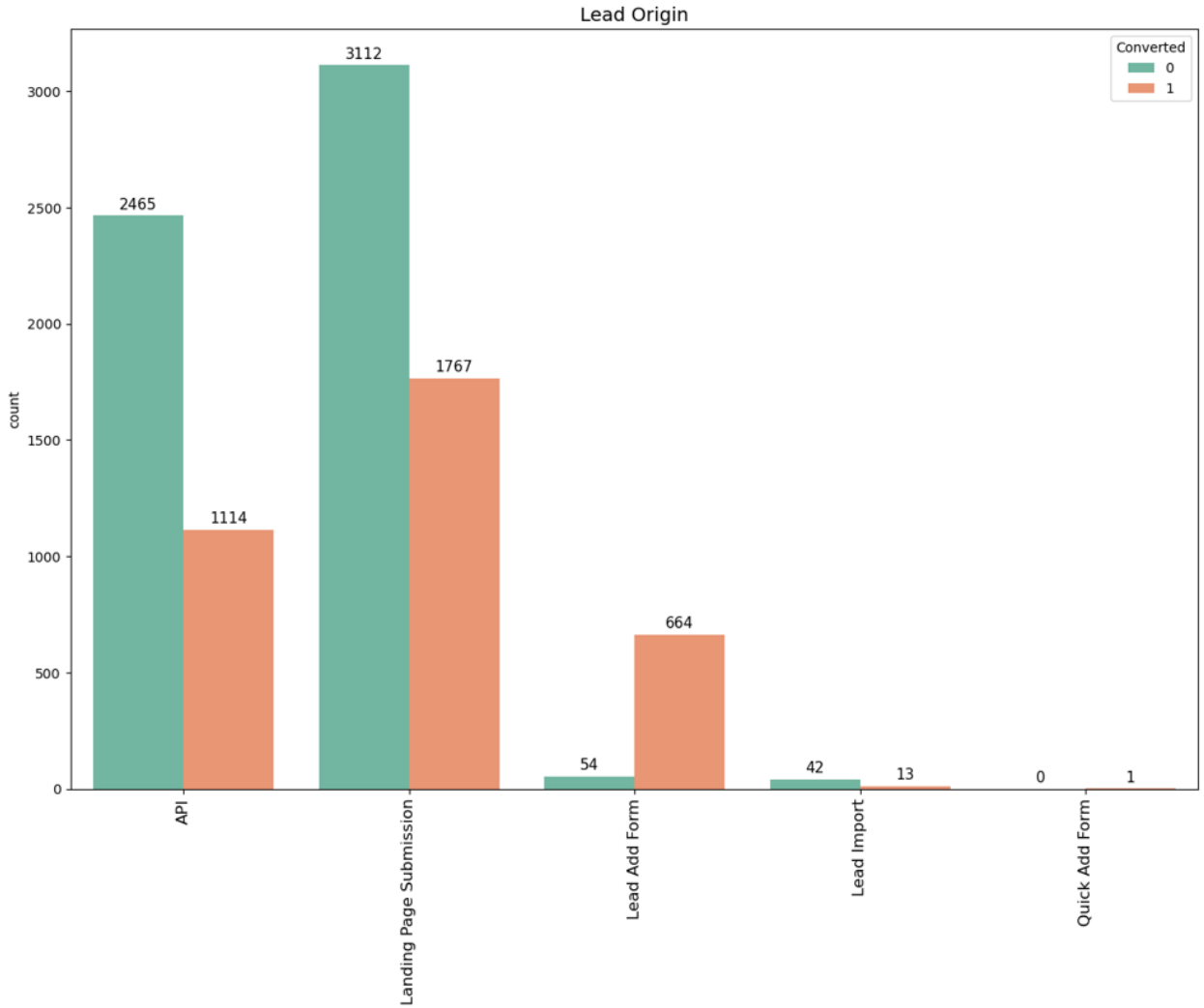


EXPLORATORY DATA ANALYSIS – CATEGORICAL VARIABLES

- 'Converted' is the target variable which indicates whether a lead has been successfully converted or not.

  [0: Not converted into lead & 1: Converted into lead]

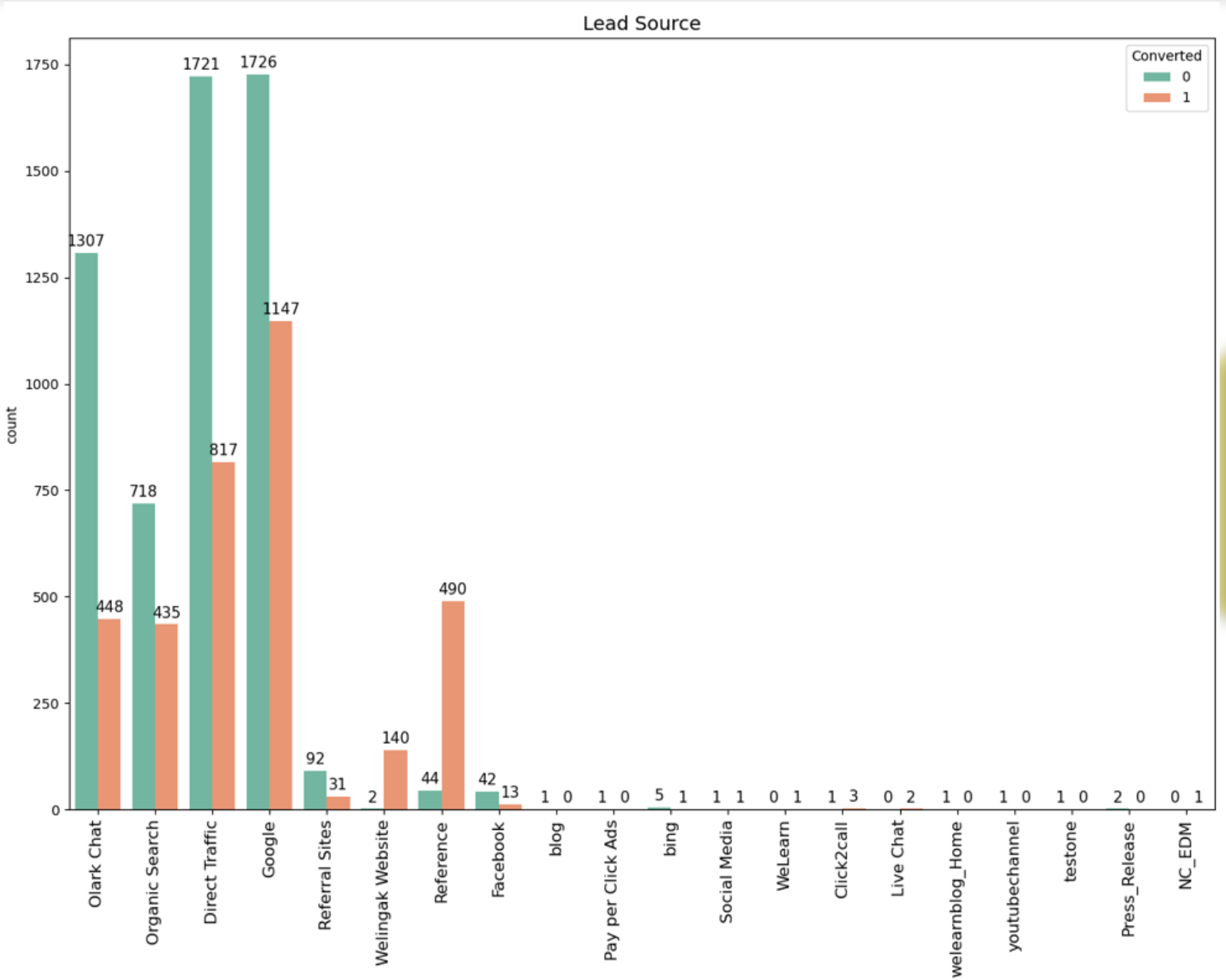
- The lead conversion rate is around 39%



Origin of the leads are maximum from 'API' and 'Landing Page Submission', but their conversion ratio is only 31% and 36% respectively.

Whereas 'Lead Add Form' gives a good conversion rate of around 93%

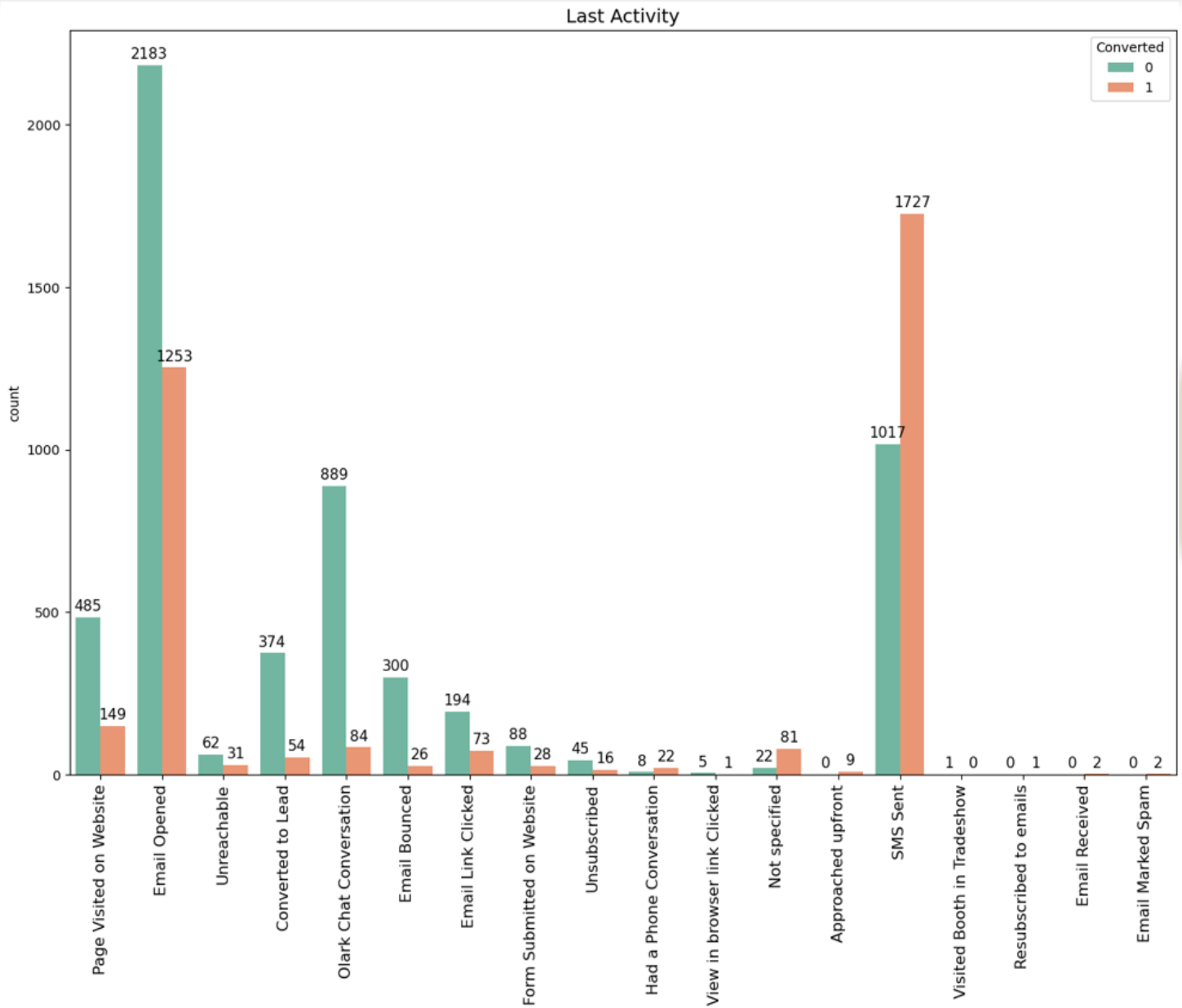
EXPLORATORY DATA ANALYSIS – CATEGORICAL VARIABLES



'Direct Traffic' and 'Google' are the source of highest number of leads.

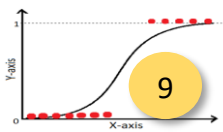
But 'Welingak Webiste' and 'Reference' shows the highest conversion rate of 99% and 92% respectively

EXPLORATORY DATA ANALYSIS – CATEGORICAL VARIABLES

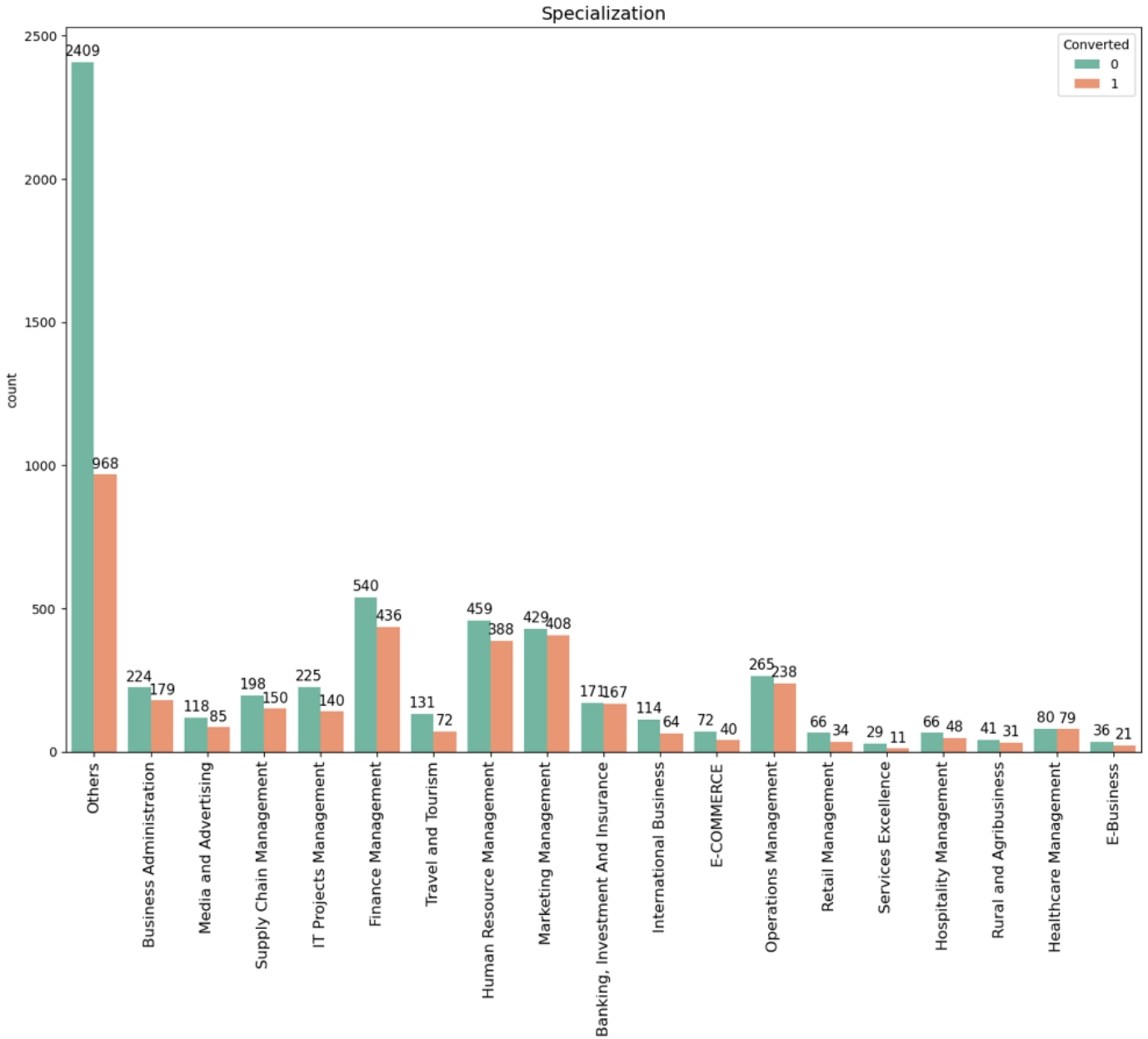


Most leads have their 'Email opened' as last activity.

But highest number of conversion, i.e., 63% are happening from 'SMS sent'

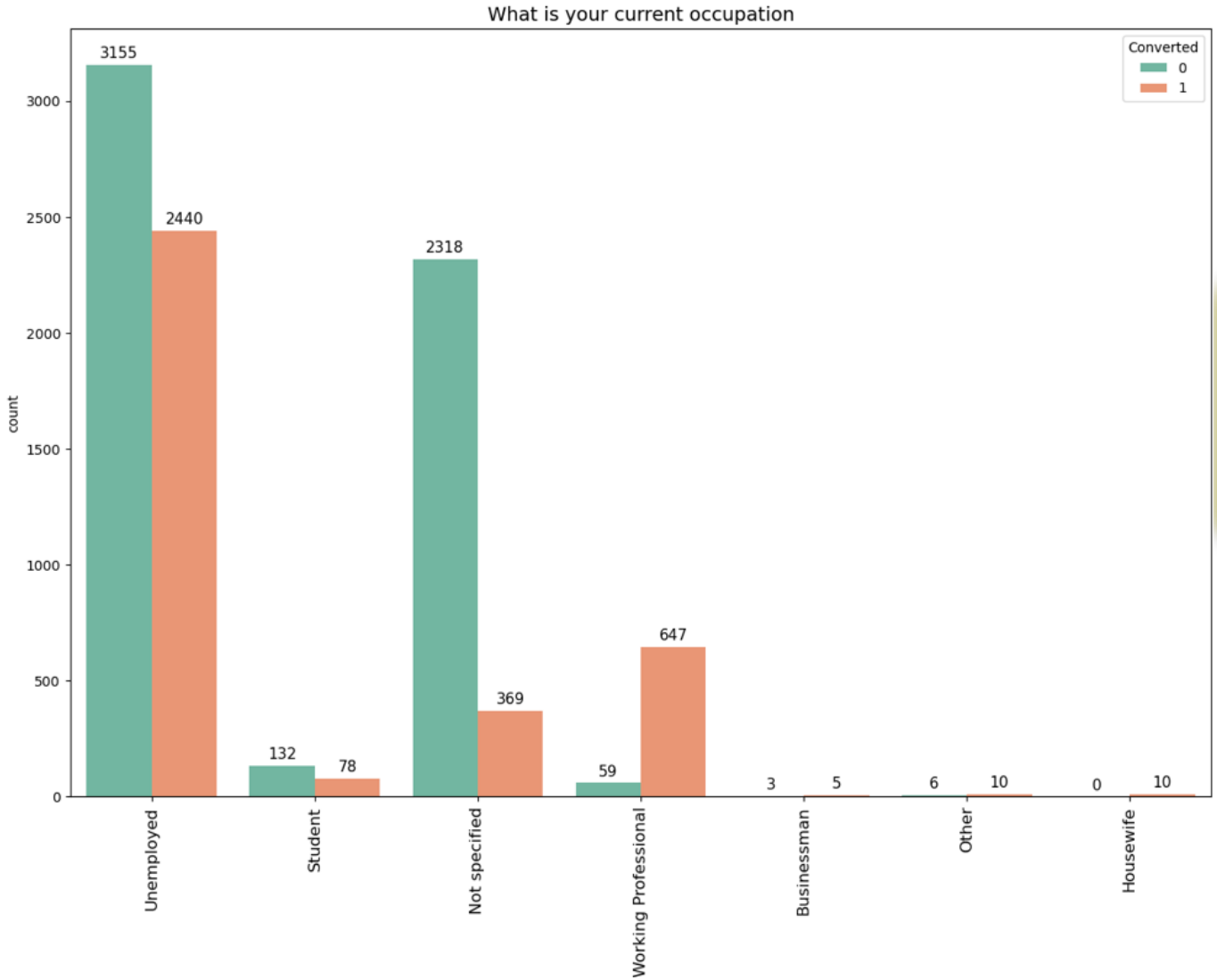


EXPLORATORY DATA ANALYSIS – CATEGORICAL VARIABLES



'Management' specialization altogether have more number of leads. Conversion rates are mostly similar across different specializations.

EXPLORATORY DATA ANALYSIS – CATEGORICAL VARIABLES



'Unemployed' persons submit higher number of leads, but conversion rate is only 44% .

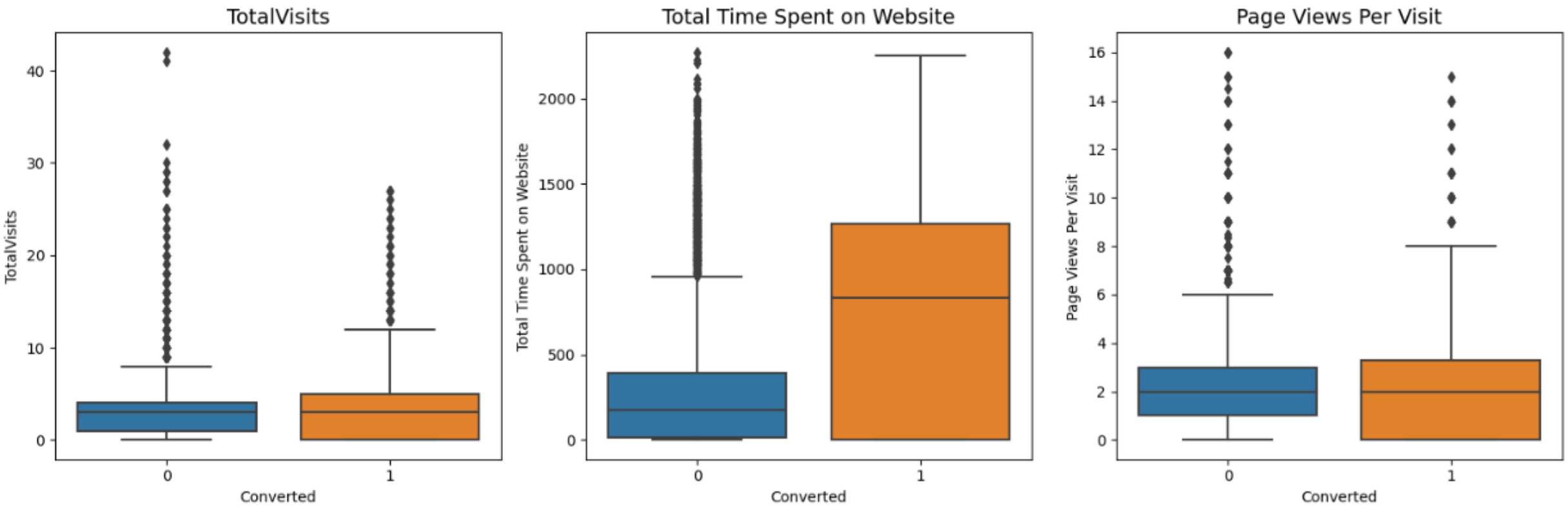
However, 'Working Professional' are easier to convert with highest conversion rate of 92%

Note:

No much inference can be drawn from other categorical variables as most of the entries are 'No'. Hence not included here.

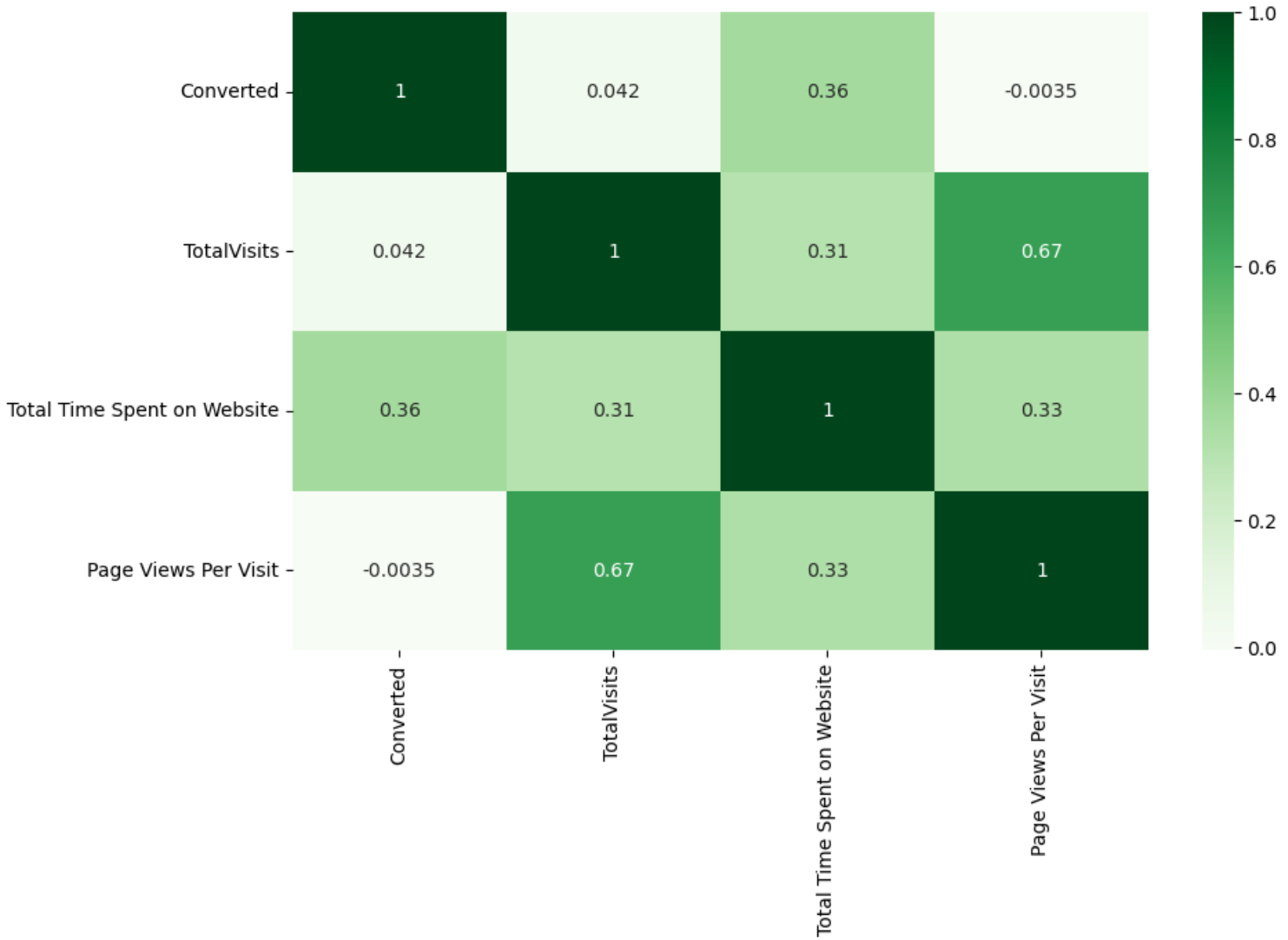
After analysis & visualization, few categorical variables are dropped as they are not significant for analysis and will not add information to the model

EXPLORATORY DATA ANALYSIS – NUMERICAL VARIABLES



- The Conversion rates are higher for all 3 - 'Total Visits', 'Total Time Spent on Website' and 'Page Views Per Visit'
- 'TotalVisits' & 'Page Views Per Visit' have almost same median values for both outputs of leads. So, its inconclusive
- However, we can infer that people spending more time on the website are more likely to be converted.

EXPLORATORY DATA ANALYSIS – CORRELATION MATRIX



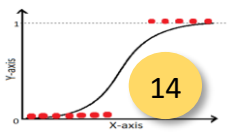
- We can observe a high correlation of 0.67 between 'TotalVisits' and 'Page Views per Visit'
- Among all numeric variables, Target variable 'Converted' has high correlation of 0.36 with 'Total Time Spent on Website'. This is relatable to the observation drawn from previous boxplot.

DATA PREPARATION

- Before dummy variable creation, some of the categorical variables have Yes/No as responses. Converted them into 1s and 0s

  ['Do Not Email', 'Do Not Call', 'Search', 'A free copy of Mastering The Interview']

- Created dummy variables for categorical columns and concatenated with original dataset. Dropped original columns after dummy variable creation
- The data is split into train and test set [70: 30 respectively]
- Used standard scaler for scaling the numerical columns
- Fit and transform done for training data
- Conversion rate is around 39% . This is neither exactly 'balanced' nor heavily imbalanced. Hence have not done any special treatment for this dataset.

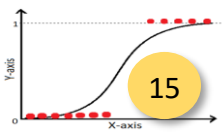


MODEL BUILDING

- Generalized Linear Model (GLM) from Stats models library is used to build logistics regression model
- Considering a large number of variables in the dataset and insignificance of many towards building the model, initial feature selection was made through RFE (Recursive Feature Elimination) technique. Top 15 features were selected through RFE
- From the top 15 features, final variable selection was made manually by running iterations to eliminate features considering p-values and VIF (Variance Inflation Factor)
- Final model built has all p-values almost zero and VIF < 2

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0405	0.056	-18.642	0.000	-1.150	-0.931
Total Time Spent on Website	1.1235	0.040	28.162	0.000	1.045	1.202
Lead Origin_Lead Add Form	3.4429	0.198	17.352	0.000	3.054	3.832
Lead Source_Olark Chat	1.3845	0.104	13.325	0.000	1.181	1.588
Lead Source_Welingak Website	2.0365	0.743	2.741	0.006	0.580	3.493
Last Activity_Converted to Lead	-1.3257	0.222	-5.985	0.000	-1.760	-0.892
Last Activity_Email Bounced	-1.8372	0.291	-6.320	0.000	-2.407	-1.267
Last Activity_Not specified	-1.6239	0.461	-3.524	0.000	-2.527	-0.721
Last Activity_Olark Chat Conversation	-1.4031	0.164	-8.559	0.000	-1.724	-1.082
Last Activity_SMS Sent	1.1000	0.074	14.776	0.000	0.954	1.246
What is your current occupation_Not specified	-1.3049	0.087	-14.974	0.000	-1.476	-1.134
What is your current occupation_Working Professional	2.3460	0.174	13.450	0.000	2.004	2.688
Last Notable Activity_Unreachable	1.7890	0.490	3.649	0.000	0.828	2.750

	Features	VIF
2	Lead Source_Olark Chat	1.64
1	Lead Origin_Lead Add Form	1.60
7	Last Activity_Olark Chat Conversation	1.40
9	What is your current occupation_Not specified	1.32
3	Lead Source_Welingak Website	1.27
0	Total Time Spent on Website	1.25
8	Last Activity_SMS Sent	1.24
10	What is your current occupation_Working Profes...	1.16
6	Last Activity_Not specified	1.15
4	Last Activity_Converted to Lead	1.02
5	Last Activity_Email Bounced	1.02
11	Last Notable Activity_Unreachable	1.00



MODEL EVALUATION

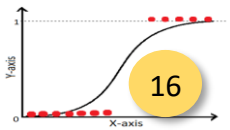
- The model built using selected features underwent prediction in training set first
- Selected an arbitrary cut-off probability point of 0.5 to find the predicted labels
- Below metrics shows the result of evaluation

```
# Evaluating other metrics

TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

```
sensitivity = TP/float(TP+FN)
specificity = TN/float(TN+FP)
precision = TP/float(TP+FP)
accuracy = float(TP+TN)/float(TP+FP+TN+FN)

print("Accuracy: ", round(accuracy,3))
print("Sensitivity: ", round(sensitivity,3))
print("Specificity: ", round(specificity,3))
print("Precision: ", round(precision,3))
```

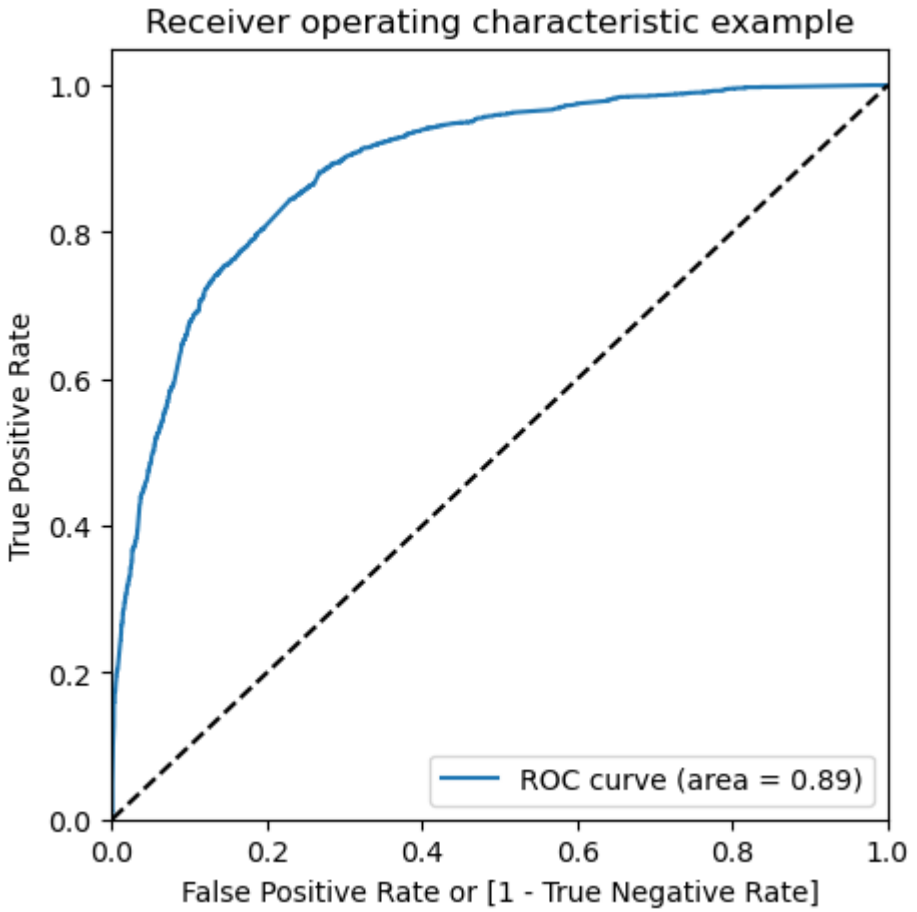
```
Accuracy:  0.816
Sensitivity:  0.708
Specificity:  0.885
Precision:  0.798
```

- All the metrics are above 80% except for Sensitivity which is 71%. This may be due to arbitrarily chosen cut-off of 0.5. Using ROC curve, will optimize the cut-off point for a better Sensitivity



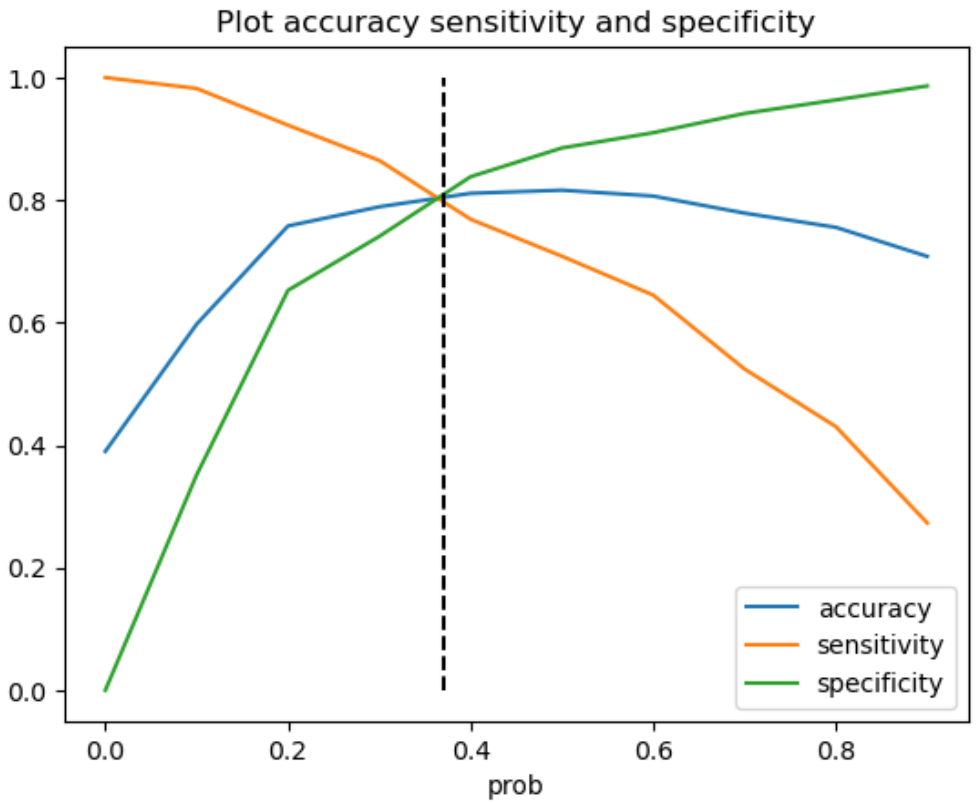
DRAWING ROC CURVE

- The ROC (Receiver Operating Characteristic) curve is used to evaluate the performance of the model built. It shows the trade-off between true positive rate (TPR) and false positive rate (FPR) across different classification thresholds.
- The area under the curve of the ROC is 0.89 which is quite good and hence the model built seems to be good.

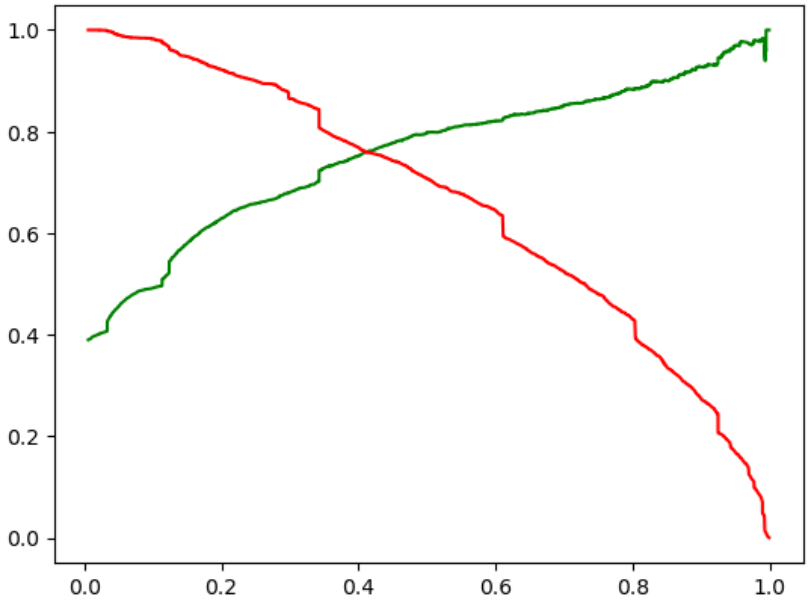


FINDING OPTIMAL CUT OFF POINT

- From the curve, 0.37 is the optimal cutoff probability arrived at.
- The evaluation metrics with 0.37 cut off shows improvement in sensitivity which is 79% now



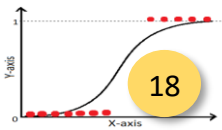
Precision and Recall Trade-off



Evaluation metrics

```
sensitivity = TP/float(TP+FN)
specificity = TN/float(TN+FP)
precision = TP/float(TP+FP)
accuracy = float(TP+TN)/float(TP+FP+TN+FN)

print("Accuracy: ", round(accuracy,3))
print("Sensitivity: ", round(sensitivity,3))
print("Specificity: ", round(specificity,3))
print("Precision: ", round(precision,3))
```

```
Accuracy:  0.808
Sensitivity:  0.787
Specificity:  0.821
Precision:  0.738
```



PREDICTION ON TEST DATA

- Made prediction on test dataset with 0.37 cut off. Evaluation metrics showed results almost same as training dataset
- Final prediction on conversion rate is close to 80%
- Hence model built seem to do a pretty good job

Evaluation metrics

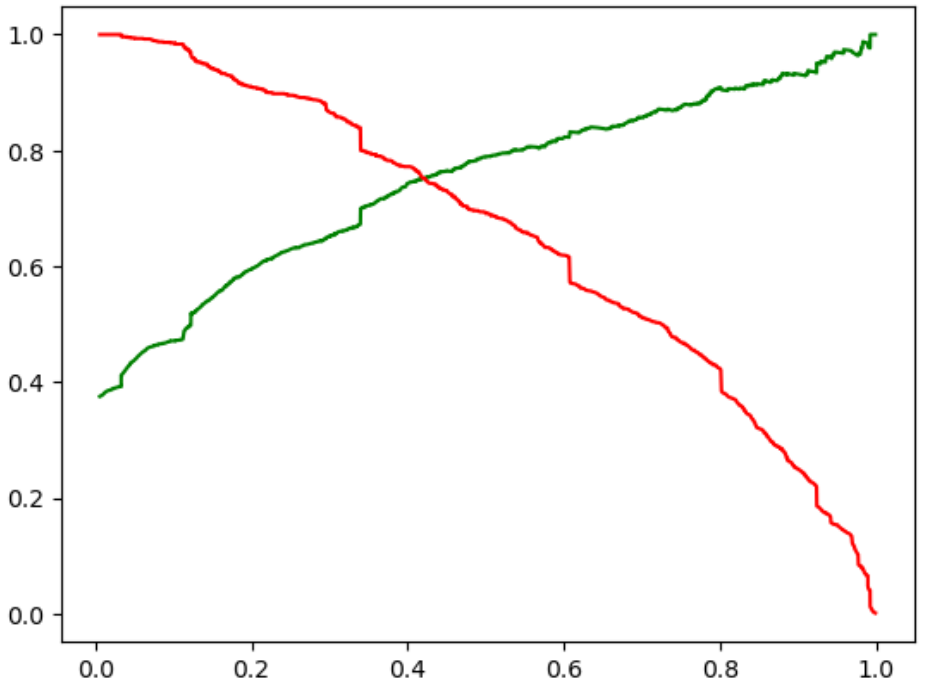
```
sensitivity = TP/float(TP+FN)
specificity = TN/float(TN+FP)
precision = TP/float(TP+FP)
accuracy = float(TP+TN)/float(TP+FP+TN+FN)

print("Accuracy: ", round(accuracy,3))
print("Sensitivity: ", round(sensitivity,3))
print("Specificity: ", round(specificity,3))
print("Precision: ", round(precision,3))
```

```
Accuracy:  0.808
Sensitivity:  0.787
Specificity:  0.821
Precision:  0.738
```

Lead score

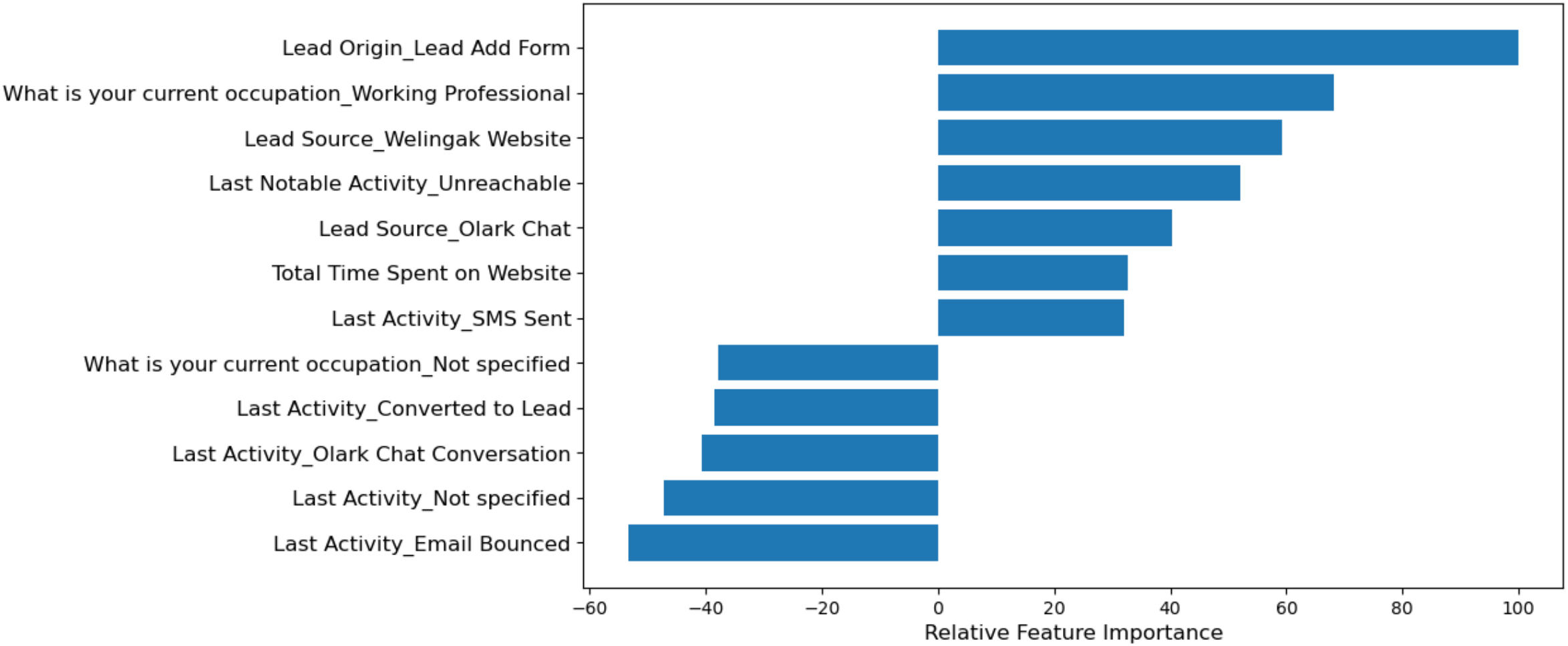
	Converted	Lead ID	Conversion_prob	Final_predicted	lead_score
0	1	4869	0.953295	1	95
1	1	8687	0.883449	1	88
2	1	6309	0.989326	1	99
3	1	8158	0.339436	0	34
4	1	2509	0.744199	1	74
5	1	7812	0.364178	0	36
6	0	435	0.186696	0	19
7	0	7843	0.351955	0	35
8	0	5442	0.484400	1	48
9	1	2307	0.914192	1	91
10	1	6697	0.604194	1	60
11	1	4264	0.557367	1	56
12	0	7825	0.687731	1	69
13	1	4058	0.989326	1	99
14	1	3069	0.992145	1	99
15	0	589	0.032158	0	3
16	0	991	0.300360	0	30
17	0	6817	0.066300	0	7
18	0	5481	0.067326	0	7
19	0	5978	0.494024	1	49

Precision and Recall Trade-off



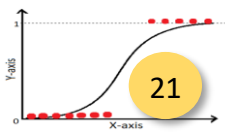
RELATIVE FEATURE IMPORTANCE

- The variables playing a key role in conversion is shown in the graph below
- These 12 variables are sorted in the order and shows the positive / negative impact on the conversion probability



CONCLUSION

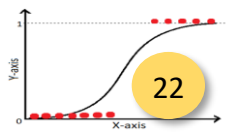
- The final model built has all p-values as almost zero and VIFs are also very low. Hence there is hardly any multicollinearity
- The metrics show a pretty good achievement of Accuracy : 81%, Sensitivity: 79%, Specificity: 82% & Precision: 74%. The metrics for training dataset is also almost same
- The lead score calculated on both training & test dataset shows a conversion rate on final predicted model around 80%
- Hence the overall model built seems good
- 12 features selected as most important ones in predicting the conversion is listed in the graph. Features having positive and negative impact on the conversion probability are also included in the graph



RECOMMENDATIONS

- Since 'API' and 'Landing Page Submission' generate the most leads but have less conversion rates, target to increase the conversion rate from these origins. Also focus on increasing leads generated using 'Lead Add Form' as the conversation rate is pretty high at 93%
- To improve overall lead conversion rate, focus on improving lead conversion of olark chat and generate more leads from welingak website as the conversion rate is too high at 99%
- Company should focus on working professionals as they are easier to convert with good conversion rate of 92%
- Focus on total time spent on website as it has a positive correlation with conversion. Also, attention should be given to the prospects whose last activity is 'SMS sent' as 63% of conversions are happening based on it

The X Education can concentrate on the above to increase their lead conversion chances further and have high number of potential buyers for their courses.



THANK YOU