LEAD SCORING CASE STUDY

Presented By

Karishma Sheikh

Nitesh Chaudhari

Aishwarya Mali

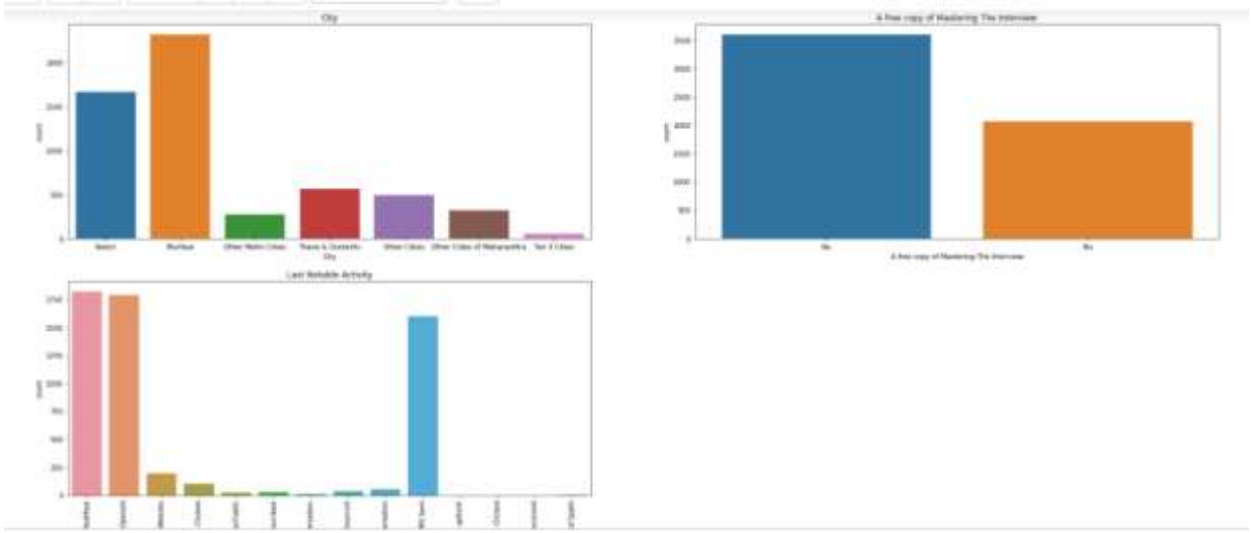
STEP 1:
DATA CLEANING

- Checked percentage of missing values in all columns and dropped columns which are having high percentage(above 40%) of missing values and dropped them.
- We dropped highly skewed columns.
- Dealt with select levels separately.
- At the end after cleaned the data we got ~ 61 % of data (5666/9240)*100

STEP 2:

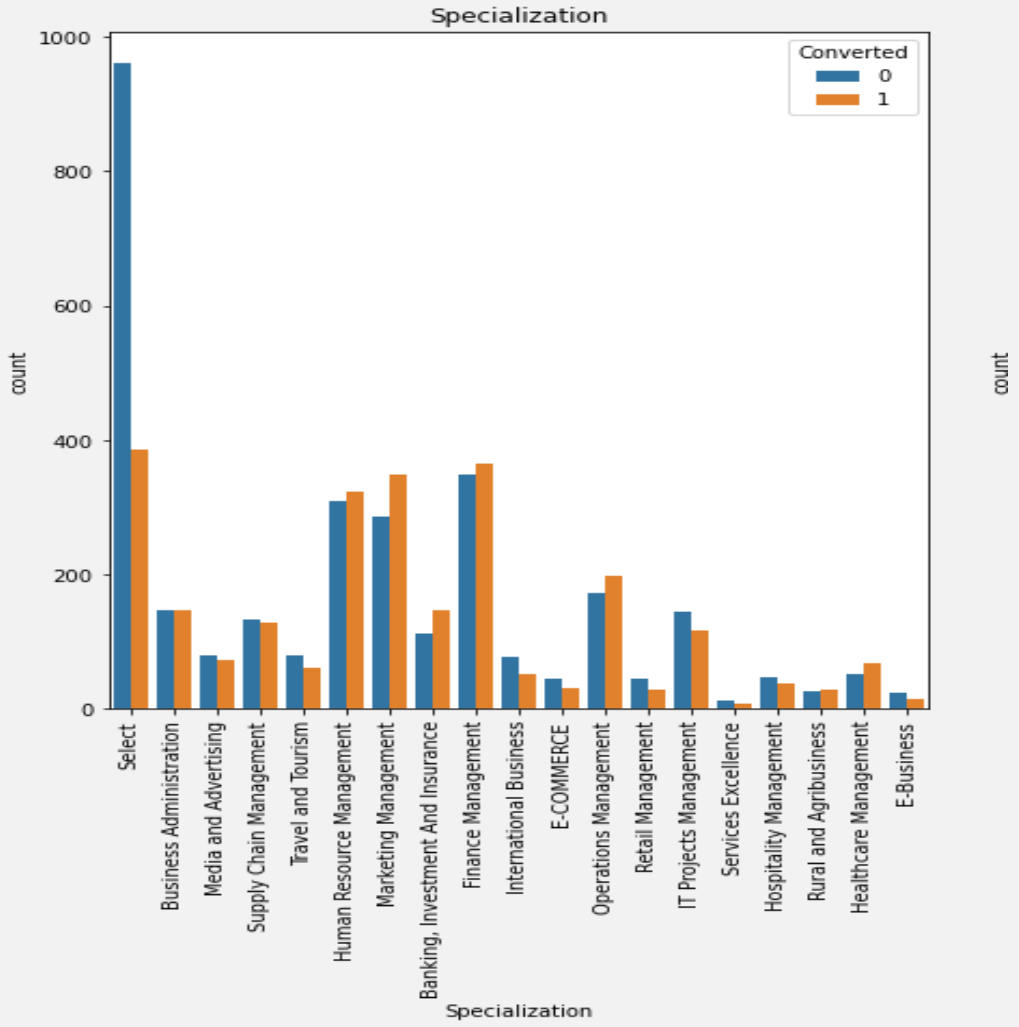
DATA PREPARATION-EDA

WE ANALYZED ALL THE CATEGORICAL VARIABLES THROUGH COUNT SUB PLOTS

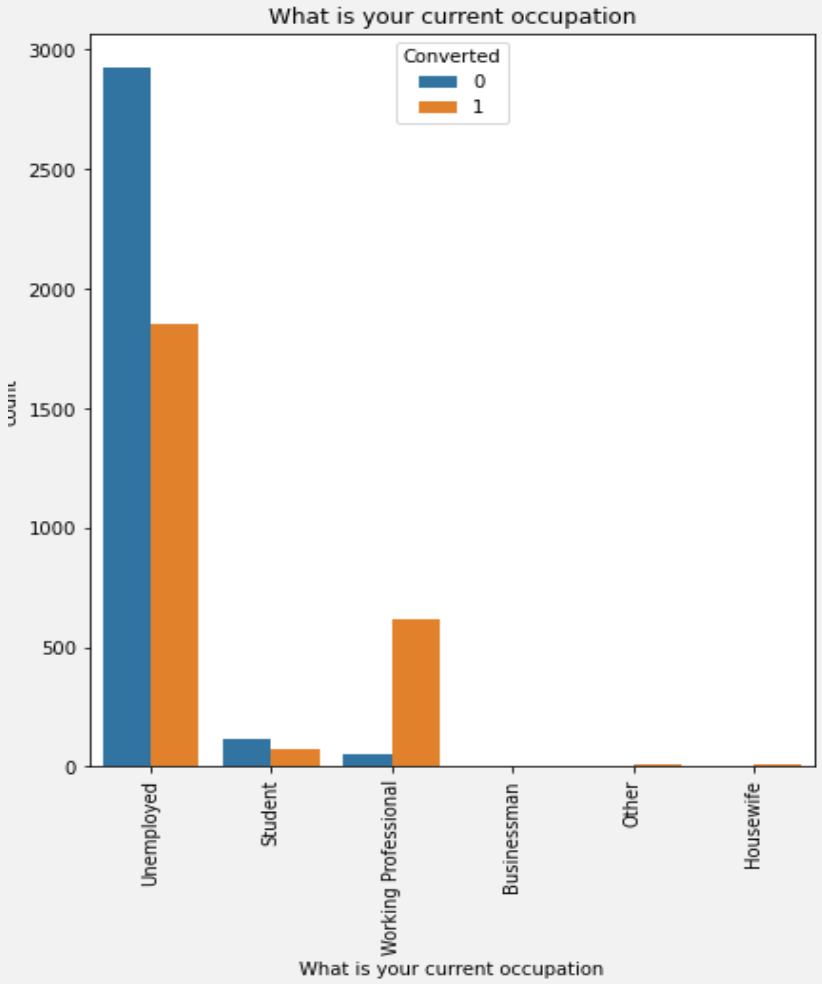


Relating all the categorical variables to Converted

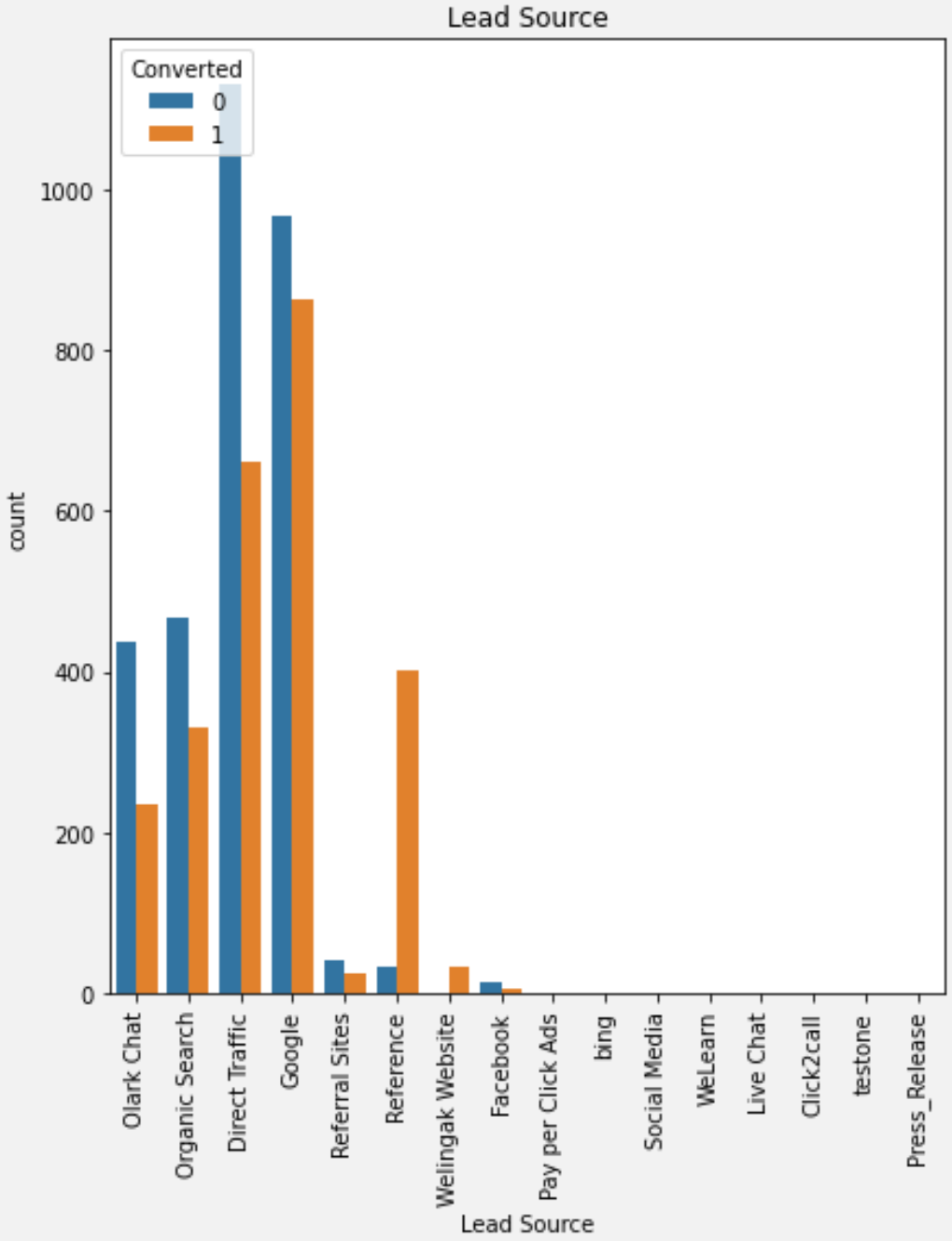
The leads having *'Specialization'* selected as 'Finance Management', 'Marketing Management' and 'Human Resource Management' are higher in number compared to all the other categories.

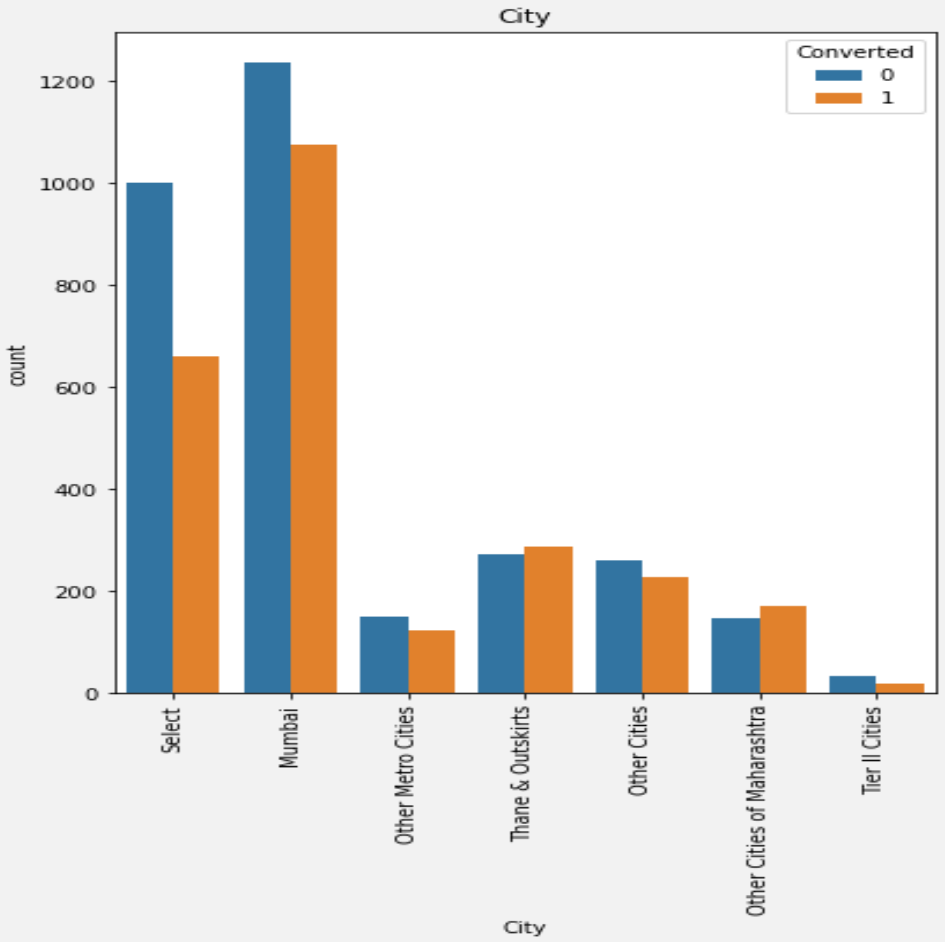


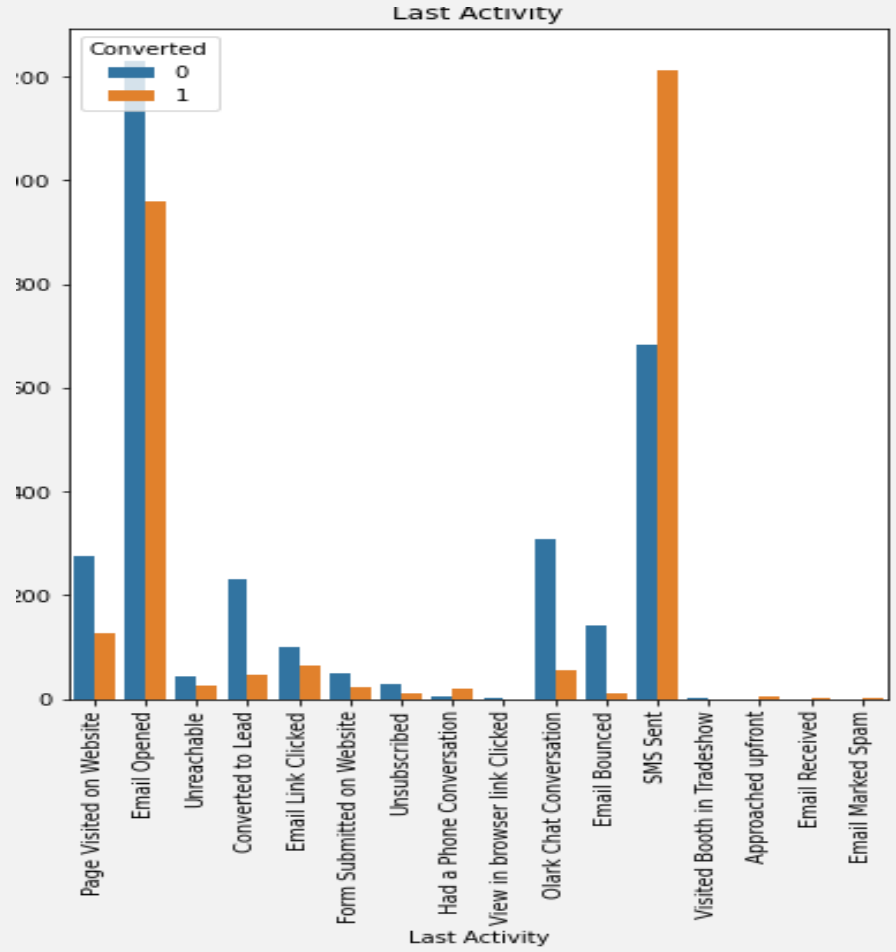
Among all the categories of ‘What is your current occupation’, working professionals and unemployed seems to be higher in number.



The category where the ‘Lead Source’ is reference is found to be having higher number of conversions followed by google and direct_traffic categories.

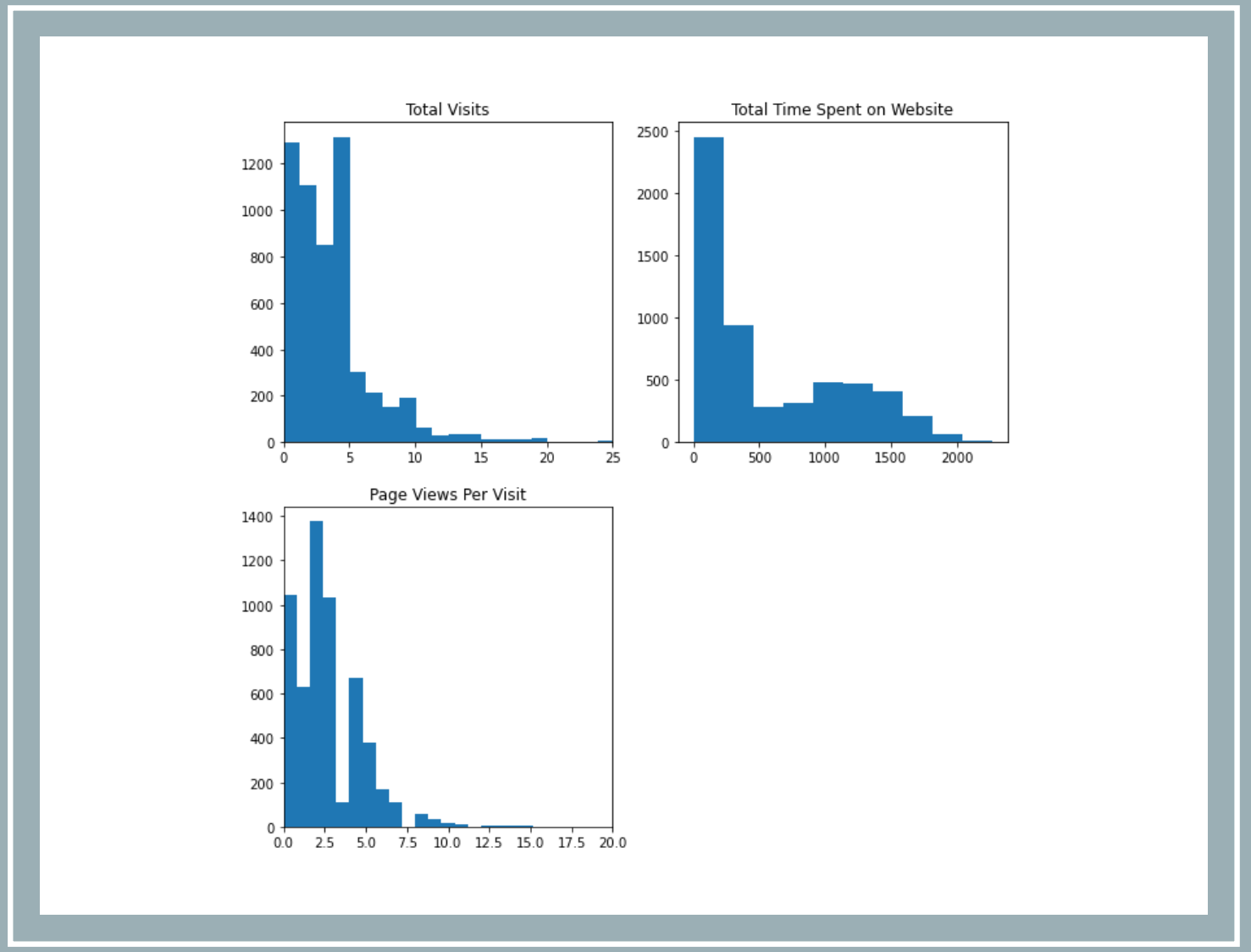




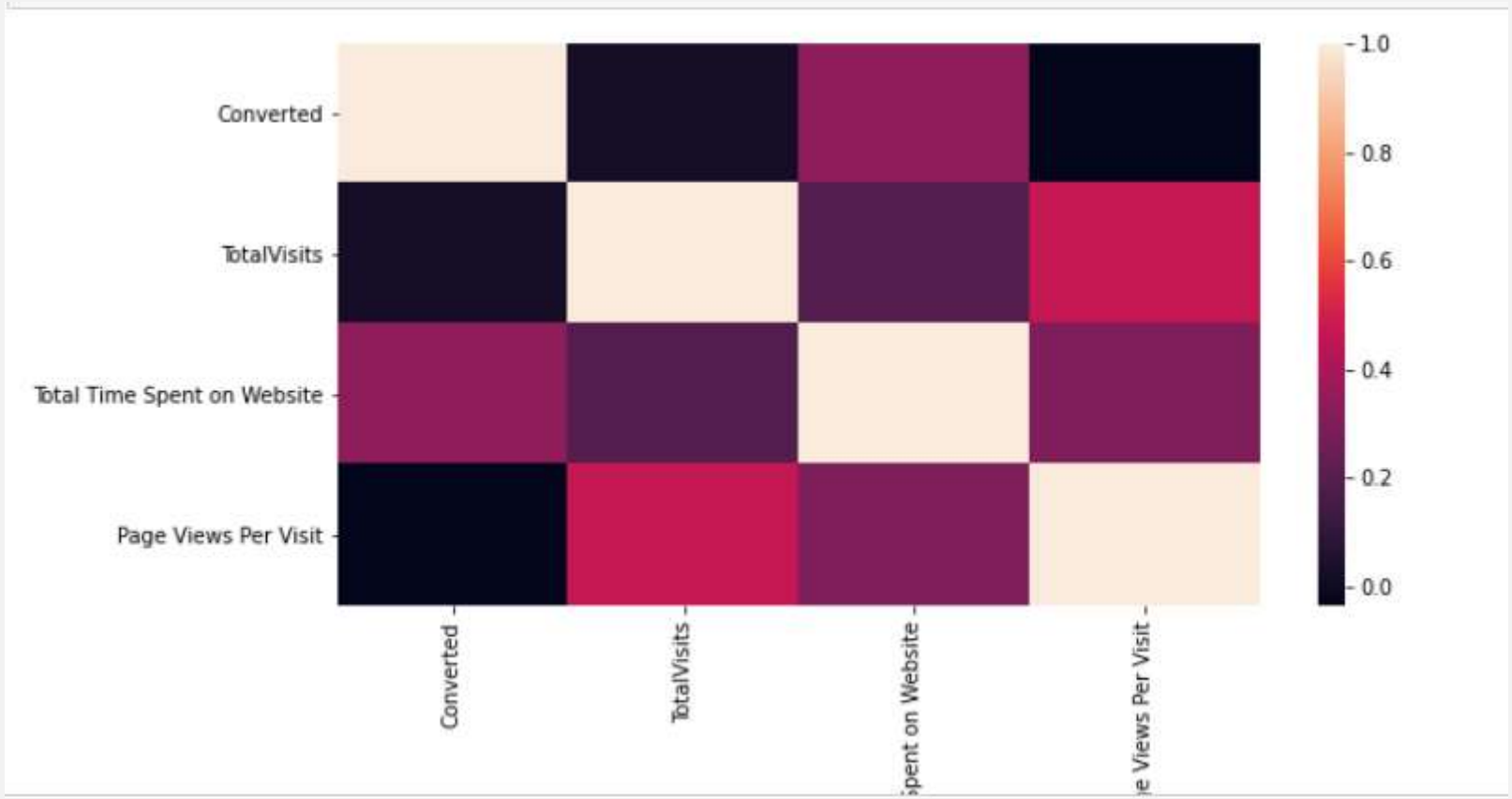


1. Most of the leads are from “MUMBAI” City.
2. SMS and Email are mostly preferred by leads as per Last activity column.
3. Found that Working professionals are getting converted at higher rates.

WE ANALYZED NUMERICAL VARIABLES



Correlation between variables was analyzed using heat map



Created Dummy Variables

1. Merged all the values which do have very less percentage of occurrences into one value.
2. Dropped the columns which was generated by sales team.
3. We created Dummy Variables for categorical columns.

[69]:

	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google	Lead Source_Olark Chat	Sou
0	0	0.0	0	0.0	0	0	0	0	0	0	1	
1	0	5.0	674	2.5	0	0	0	0	0	0	0	
2	1	2.0	1532	2.0	1	0	0	1	0	0	0	
3	0	1.0	305	1.0	1	0	0	1	0	0	0	
4	1	2.0	1428	1.0	1	0	0	0	0	1	0	

]:
```
Leads.shape
```

[70]:
```
(5666, 38)
```

Test and Train Split

- The dataset has been split into training and testing sets.(70:30)

```
# Scale the three numeric features present in the dataset

scaler = MinMaxScaler()

X_train[['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']] = scaler.fit_transform(X_train[['TotalVisits

X_train.head()
```

- Performed Scaling and calculated conversion rate 45%

```
]:  # Conversion Rate
    Converted = (sum(Leads['Converted'])/len(Leads['Converted'].index))*100
    Converted

[75]: 45.287680903635724
```


CORRELATION AMONG VARIABLES
I. HEAT MAPS ARE USED TO FIND THE CORRELATION.
Converted
TotalVisits
Direct Traffic
Lead Import
Lead Add Form
Page Submission
Reference
Referral Sites
Social Media
Google
Facebook
Olark Chat
Organic Search
Website
Not Email_Yes
Housewife
Other
Student
Unemployed
Professional
Interview_Yes
Insurance
Administration
Management
Management
Management
Management
Management
Management
Management
City_Mumbai
Other Cities
Maharashtra
Metro Cities

- Since the number of variables are pretty high, a table can be viewed for a clear understanding. Hence plotted a table with correlations.

	Converted	TotalVisits	Time Spent on Website	Views Per Visit	Origin_Landing Page Submission	Origin_Lead Add Form	Origin_Lead Import	Source_Direct Traffic	Lead Source_Facebook	Source_
Converted	1.000000	0.025141	0.339829	-0.034967	-0.077101	0.289411	-0.014648	-0.114030	-0.016894	0
TotalVisits	0.025141	1.000000	0.190621	0.469510	0.260163	-0.209166	-0.041440	0.061676	-0.040332	0
Total Time Spent on Website	0.339829	0.190621	1.000000	0.298349	0.270559	-0.237685	-0.058459	0.106475	-0.057904	0
Page Views Per Visit	-0.034967	0.469510	0.298349	1.000000	0.454317	-0.354026	-0.065082	0.080194	-0.061640	0
Lead Origin_Landing Page Submission	-0.077101	0.260163	0.270559	0.454317	1.000000	-0.378994	-0.076329	0.496776	-0.072313	0
Lead Origin_Lead Add Form	0.289411	-0.209166	-0.237685	-0.354026	-0.378994	1.000000	-0.018471	-0.205972	-0.018908	-0
Lead Origin_Lead Import	-0.014648	-0.041440	-0.058459	-0.065082	-0.076329	-0.018471	1.000000	-0.041483	0.976922	-0
Lead Source_Direct Traffic	-0.114030	0.061676	0.106475	0.080194	0.496776	-0.205972	-0.041483	1.000000	-0.042463	-0
Lead Source_Facebook	-0.016894	-0.040332	-0.057904	-0.061640	-0.072313	-0.018908	0.976922	-0.042463	1.000000	-0
Lead Source_Google	0.025200	0.080467	0.196653	0.176280	0.054282	-0.207813	-0.042127	-0.469759	-0.043123	1
Lead Source_Olark Chat	-0.074936	-0.244352	-0.297373	-0.422931	-0.459064	-0.109124	-0.022374	-0.249488	-0.022902	-0
Lead Source_Organic	-0.030853	0.194595	0.065042	0.312443	0.028845	-0.122794	-0.024731	-0.275771	-0.025315	-0

Covariance Type:	nonrobust					
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6879	0.614	-2.748	0.006	-2.892	-0.484
TotalVisits	2.4443	0.852	2.868	0.004	0.774	4.114
Total Time Spent on Website	4.4288	0.192	23.104	0.000	4.053	4.805
Lead Origin_Landing Page Submission	-0.8903	0.149	-5.962	0.000	-1.183	-0.598
Lead Origin_Lead Add Form	24.9805	3.16e+04	0.001	0.999	-6.19e+04	6.2e+04
Lead Source_Olark Chat	1.0875	0.161	6.746	0.000	0.772	1.404
Lead Source_Reference	-21.1128	3.16e+04	-0.001	0.999	-6.2e+04	6.2e+04
Lead Source_Welingak Website	0.2713	3.46e+04	7.85e-06	1.000	-6.78e+04	6.78e+04
Do Not Email_Yes	-1.4230	0.208	-6.835	0.000	-1.831	-1.015
What is your current occupation_Housewife	23.1980	3.16e+04	0.001	0.999	-6.19e+04	6.2e+04
What is your current occupation_Student	-0.6884	0.638	-1.080	0.280	-1.938	0.561
What is your current occupation_Unemployed	-0.6289	0.602	-1.045	0.296	-1.809	0.551
What is your current occupation_Working Professional	2.1664	0.631	3.435	0.001	0.930	3.403

```
# Fit a logistic Regression model on X_train after adding a constant and output the summary

X_train_sm = sm.add_constant(X_train)
logm2 = sm.GLM(y_train, X_train_sm, family = sm.families.Binomial())
res = logm2.fit()
res.summary()
```

MODEL BUILDING

- Built a model using logistic regression
- We selected a small set of features from lot of variables in dataset using RFE , for our calculation we used 20 variables.
- The variables were selected by RFE , We used these variables to create a logistic regression model using statsmodels.

VIF Value

- VIFs seem to be in a decent range except for three variables.
- Let's first drop the variable Lead Origin_Lead Add Form since it has a high p-value as well as a high VIF.

ut[84]:

	Features	VIF
3	Lead Origin_Lead Add Form	68.38
5	Lead Source_Reference	62.58
2	Lead Origin_Landing Page Submission	7.65
10	What is your current occupation_Unemployed	7.13
6	Lead Source_Welingak Website	6.21
15	City_Mumbai	4.97
0	TotalVisits	2.46
1	Total Time Spent on Website	2.40
11	What is your current occupation_Working Profes...	2.09
4	Lead Source_Olark Chat	1.90
19	City_Thane & Outskirts	1.90
16	City_Other Cities	1.79
17	City_Other Cities of Maharashtra	1.54
18	City_Other Metro Cities	1.46

```
X_train.drop('Lead Origin_Lead Add Form', axis = 1, inplace = True)
```

```
# Refit the model with the new set of features

logm1 = sm.GLM(y_train,(sm.add_constant(X_train)), family = sm.families.Binomial())
logm1.fit().summary()
```

[86]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	3966
Model:	GLM	Df Residuals:	3946
Model Family:	Binomial	Df Model:	19
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1865.1
Date:	Wed, 12 Jan 2022	Deviance:	3730.2
Time:	17:10:54	Pearson chi2:	4.41e+03
No. Iterations:	22		
Covariance Type:	nonrobust		

REFITTING MODEL

- Dropped columns having higher VIF and p values and we refitted model with new features and checked for VIF values, as some columns have high VIF value we were repeating the process until optimal values are obtained.

FINAL VIF VALUE

- After dropping columns with high VIF values and refitting the model with new features of values. We finally got VIF values

	Features	VIF
0	TotalVisits	2.14
1	Total Time Spent on Website	2.10
7	City_Mumbai	1.92
11	City_Thane & Outskirts	1.27
5	What is your current occupation_Working Profes...	1.23
8	City_Other Cities	1.18
3	Lead Source_Reference	1.15
9	City_Other Cities of Maharashtra	1.13
10	City_Other Metro Cities	1.10
4	Do Not Email_Yes	1.07
6	Specialization_Banking, Investment And Insurance	1.06
2	Lead Source_Olark Chat	1.02

MODEL EVALUATION

- Predicted the probabilities on the train set
- Created a new dataframe containing the actual conversion flag and the probabilities predicted by the model.

```
y_train_pred_final['Predicted'] = y_train_pred_final.Conversion_Prob.map(lambda x: 1 if x > 0.5 else 0)

# Let's see the head
y_train_pred_final.head()
```

)]:

	Converted	Conversion_Prob	Predicted
0	0	0.613285	1
1	0	0.199808	0
2	0	0.226315	0
3	1	0.888413	1
4	0	0.100274	0

CONFUSION MATRIX, SENSITIVITY AND ACCURACY

- WE CREATED CONFUSION MATRIX AND CHECKED OVERALL ACCURACY(79%),SENSITIVITY(72%) AND SPECIFICITY (84%)

```
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

```
16]: # Calculate the sensitivity

     TP/(TP+FN)
```

```
t[116]: 0.7291553133514986
```

```
17]: # Calculate the specificity

     TN/(TN+FP)
```

```
t[117]: 0.8446738620366026
```

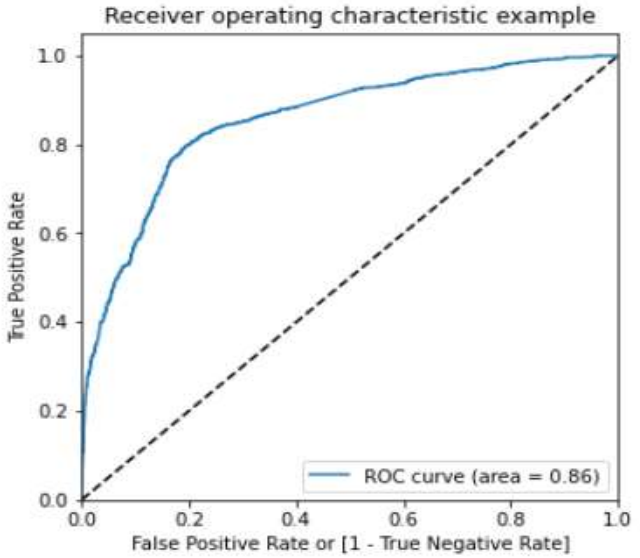
```
# Create confusion matrix

confusion = metrics.confusion_matrix(y_train_pred_fi
print(confusion)
```

```
[[1800  331]
 [ 497 1338]]
```

```
# Predicted     not_churn    churn
# Actual
# not_churn        2543       463
# churn            692        1652
```

ROC CURVE

WE FOUND OPTIMAL CUTOFF USING ROC CURVE AND CREATED COLUMNS WITH DIFFERENT PROBABILITY CUT OFF (0.42).THE AREA UNDER CURVE IS FOUND TO BE 0.86 WHICH IMPLIES THAT THE MODEL IS GOOD AS THE 'AUC' IS HIGHER.




```
y_train_pred_final.head()
```

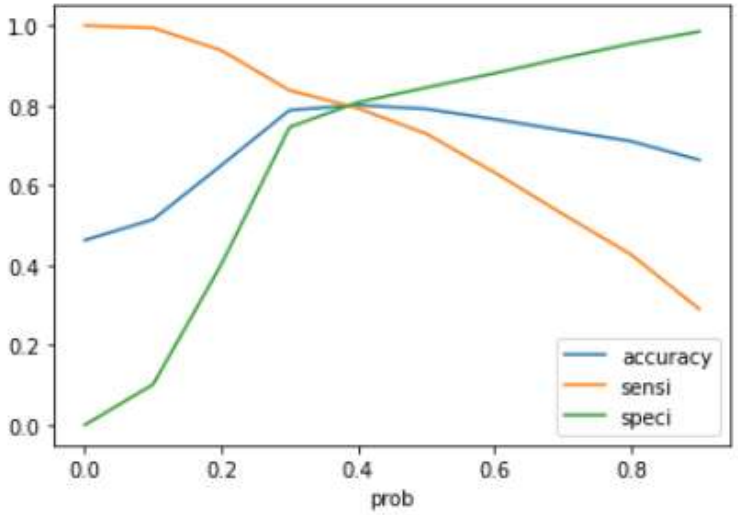
21]:

	Converted	Conversion_Prob	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0	0.613285	1	1	1	1	1	1	1	1	0	0	0
1	0	0.199808	0	1	1	0	0	0	0	0	0	0	0
2	0	0.226315	0	1	1	1	0	0	0	0	0	0	0
3	1	0.888413	1	1	1	1	1	1	1	1	1	1	0
4	0	0.100274	0	1	1	0	0	0	0	0	0	0	0

ACCURACY, SENSITIVITY, SPECIFICTY AND TRADE OFF

CREATED A DATA FRAME TO SEE THE VALUES OF ACCURACY, SENSITIVITY, AND SPECIFICITY AT DIFFERENT VALUES OF PROBABILITY CUTOFFS AND CHECKED CONFUSION MATRIX, ACCURACY, SENSITIVITY AND SPECIFICITY.

At around 0.42, you get the optimal values of the three metrics. So let's choose 0.42 as our optimal cutoff point.




```
TN = confusion2[0,0] # true negatives
FP = confusion2[0,1] # false positives
FN = confusion2[1,0] # false negatives
```

```
# Calculate Sensitivity

TP/(TP+FN)
```

```
3]: 0.7803814713896458
```

```
# Calculate Specificity

TN/(TN+FP)
```

```
9]: 0.8165180666353824
```

PREDICTIONS ON TEST SET

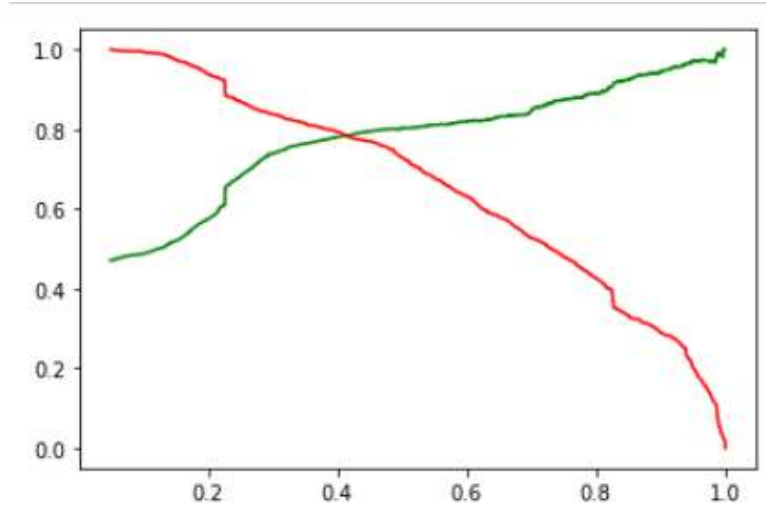
- Added a constant and Dropped the required columns in test set and Made predictions on the test set and stored it in the variable and checked all evaluation metrics

- We got the metrics as Accuracy (79%),Sensitivity(78%),specificity(81%)

```
[149]: 0.7917647058823529

]:  confusion2 = metrics.confusion_matrix(y_pred_final['Converted'], y_pred_final.final_predicted )
    confusion2

[150]: array([[765, 204],
              [150, 581]], dtype=int64)

]:  # Calculate sensitivity
    TP / float(TP+FN)

[151]: 0.7803814713896458

]:  # Calculate specificity
    TN / float(TN+FP)

[152]: 0.8165180666353824
```

PRECISION AND RECALL TRADEOFF

THE TRAINING MODEL WAS BUILD USING THE PRECISION-RECALL VIEW AND PERFORMED ALL THE EVALUATION METRICS.




```
# Calculate Precision

TP/(TP+FP)
```

```
55]: 0.7855183763027976
```

```
# Calculate Recall

TP/(TP+FN)
```

```
56]: 0.7803814713896458
```

- PREDICTIONS ON TEST SET

  - We made predictions on test set and created data frame and checked all the evaluation metrics we got precision (74%) and Recall (79%)

```
                [150, 581]], dtype=int64)

[180]:  TP = confusion2[1,1] # true positive
        TN = confusion2[0,0] # true negatives
        FP = confusion2[0,1] # false positives
        FN = confusion2[1,0] # false negatives

[181]:  # Calculate Precision

        TP/(TP+FP)

Out[181]: 0.7401273885350318

[182]:  # Calculate Recall

        TP/(TP+FN)

Out[182]: 0.7948016415868673
```

INFERENCES

Top Three Variables

- Total Time Spent on Website

- This has got positive correlation which implies that it has positive contribution towards the conversion. Higher the time spent on the website, higher is the probability of the lead getting converted as a prospect.

- Lead Source_Reference

- If the source of the lead is via Reference, then there is a higher probability that the lead would convert, as the referrals not only provide for cashbacks along with provided assurances from current users. As a referrer mostly refers known people who mostly are bound by trusted reviews from these current users.

- **What is your current occupation_Working Professional**

- If the lead is a working professional, then the conversion probability is found to be higher. Given the reason that most of the working professionals tend to opt for the upskilling courses to attain better career prospects provided they do stand in a better position to bear their fees as well compared to students who already are enrolled with some other programs and mostly dependents.

Hence these leads need to be pursued as they depict higher probability of conversion.