Lead Scoring Case Study

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.

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%.

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.

Approach followed for solving case study

- Data Understanding
- Data Pre-Processing
  - Data Cleaning
  - Handling Outliers
  - Explanatory Data Analysis
  - Data Preparation for Model building
- Model Building
  - Test-Train Split
  - Feature Scaling
  - Feature selection using RFE
  - Model prediction on fitted model
- Model Evaluation
  - Finding the Optimal Threshold
  - Precision and Recall
  - Prediction on Test set
- Model Summary

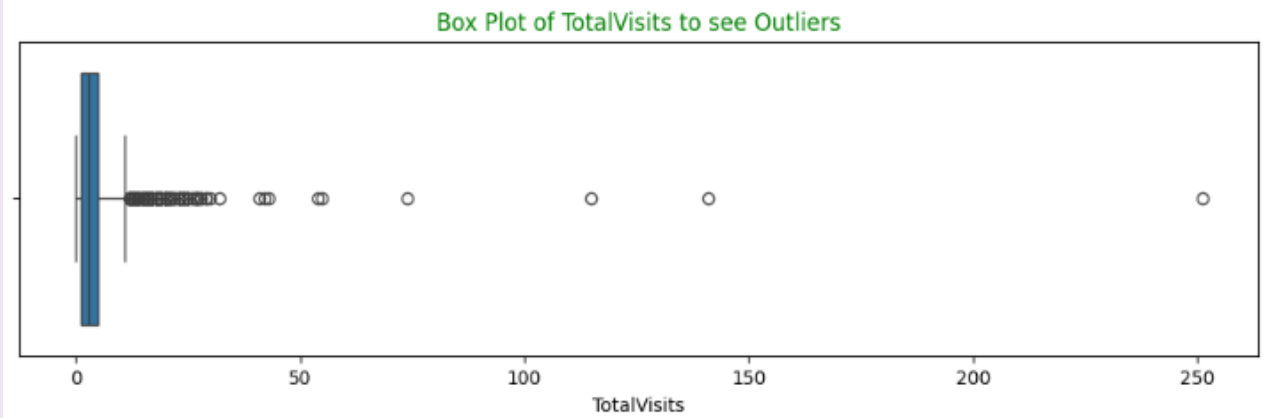
Data Understanding and Data cleaning

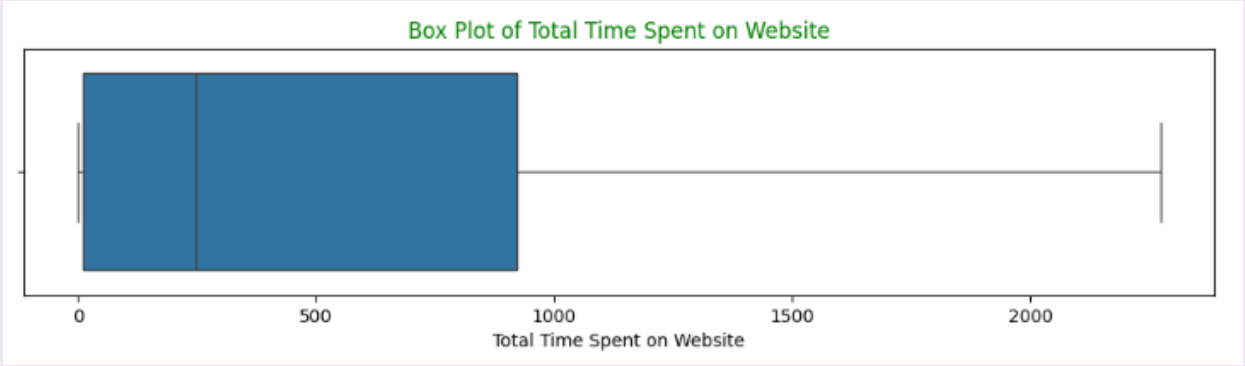
- There are 9240 entries in our dataset and 37 variables to analyze.
- ‘Prospect ID’ and ‘Lead Number’ are unique identifiers assigned to each lead.
- ‘Converted’ attribute is our target variable based on which we can study if a lead has converted or not.
- There are few categorical values that have 'select’ category which we treat as missing value or missing data and we replace this with NaN.
- There are many columns, that have high missing values even above 40% which would be more than 3500 entries of the 9240 entries we have in our dataset, imputing them would create bias and may affect the models predictive power, so we drop them. We also drop ‘Country’, ‘City’ columns, since they are location parameters.
- Columns ‘Specialization’, ‘Last Activity’, ‘Last Notable Activity’, ‘What is your current occupation’ have various categories, so we combine a few, for easy interpretation.
- Many columns have single label which creates bias, thus we drop them.
- Columns having missing values below 5%, we remove the rows.

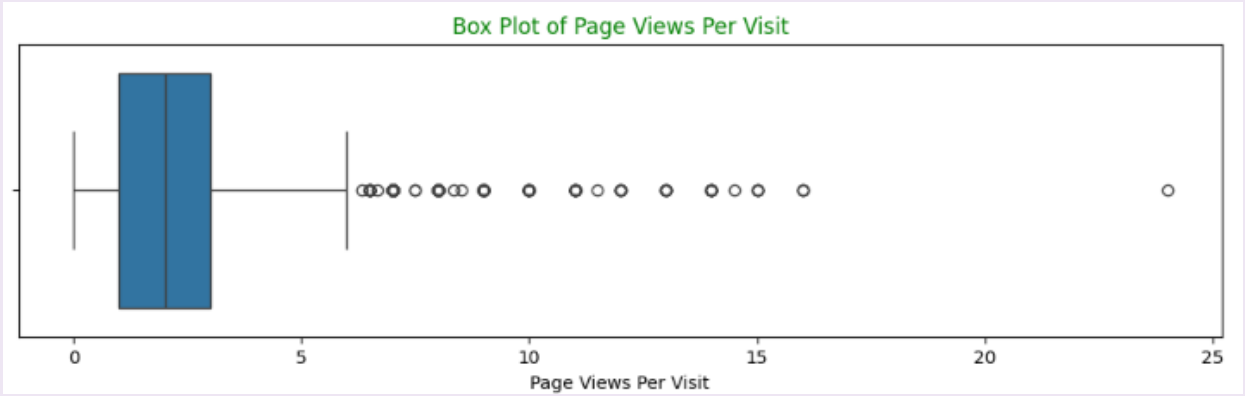
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype  
---  ------                                         --------------  -----  
 0   Prospect ID                                    9240 non-null   object 
 1   Lead Number                                    9240 non-null   int64  
 2   Lead Origin                                    9240 non-null   object 
 3   Lead Source                                    9204 non-null   object 
 4   Do Not Email                                   9240 non-null   object 
 5   Do Not Call                                    9240 non-null   object 
 6   Converted                                      9240 non-null   int64  
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64  
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object 
 11  Country                                        6779 non-null   object 
 12  Specialization                                 7802 non-null   object 
 13  How did you hear about X Education             7033 non-null   object 
 14  What is your current occupation                6550 non-null   object 
 15  What matters most to you in choosing a course  6531 non-null   object 
 16  Search                                         9240 non-null   object 
 17  Magazine                                       9240 non-null   object 
 18  Newspaper Article                              9240 non-null   object 
 19  X Education Forums                             9240 non-null   object 
 20  Newspaper                                      9240 non-null   object 
 21  Digital Advertisement                          9240 non-null   object 
 22  Through Recommendations                        9240 non-null   object 
 23  Receive More Updates About Our Courses         9240 non-null   object 
 24  Tags                                           5887 non-null   object 
 25  Lead Quality                                   4473 non-null   object 
 26  Update me on Supply Chain Content              9240 non-null   object 
 27  Get updates on DM Content                      9240 non-null   object 
 28  Lead Profile                                   6531 non-null   object 
 29  City                                           7820 non-null   object 
 30  Asymmetrique Activity Index                    5022 non-null   object 
 31  Asymmetrique Profile Index                     5022 non-null   object 
 32  Asymmetrique Activity Score                    5022 non-null   float64
 33  Asymmetrique Profile Score                     5022 non-null   float64
 34  I agree to pay the amount through cheque       9240 non-null   object 
 35  A free copy of Mastering The Interview         9240 non-null   object 
 36  Last Notable Activity                          9240 non-null   object 
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

Handling Outliers

- We have three numerical columns, ‘TotalVisits’, ‘'Total Time Spent on Website’ and ‘Page Views Per Visit’ to check for outliers presence.
- From the boxplot of ‘TotalVisits’ we can visualize that there are outlier present, but after studying associated features associated with these outliers we observed leads who have visited the website more than 20 times have been converted but the rest seem like errors in the dataset so we drop lead_score['TotalVisits'] < 30.
- There are no outliers for 'Total Time Spent on Website’, here we don't need outlier treatment.
- For ‘Page Views Per Visit’, above the 99th percentile there are many leads that have been converted, so removing those datapoints, might lead to unnecessary data loss, so we remove only the max datapoint to avoid inaccuracies.

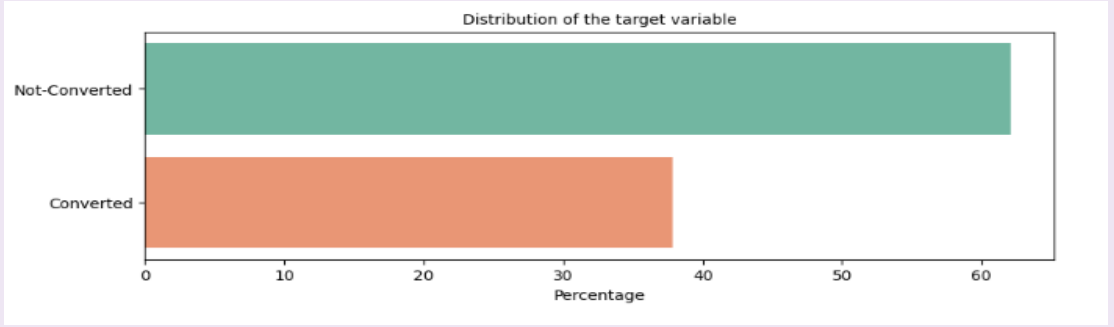


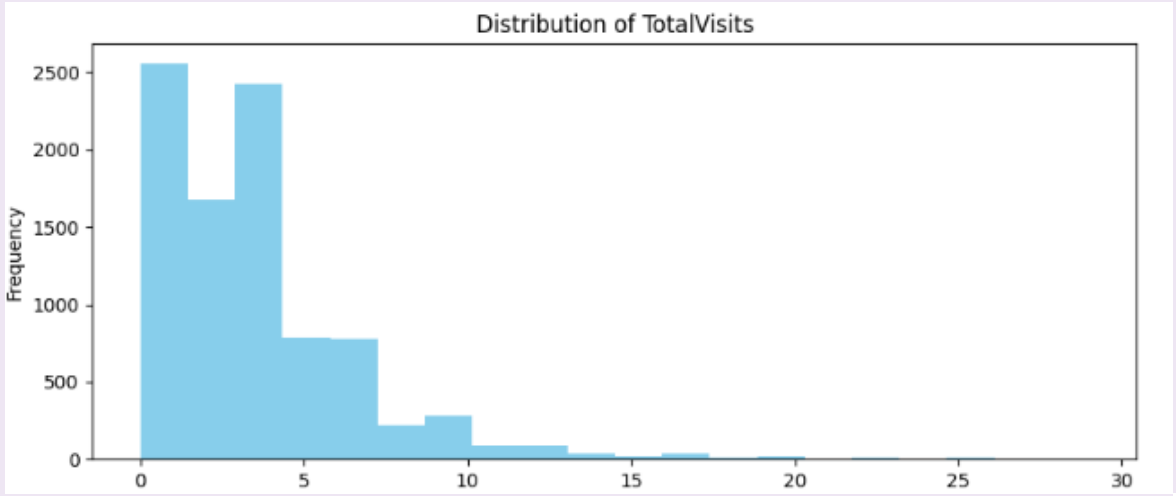




Explanatory Data Analysis

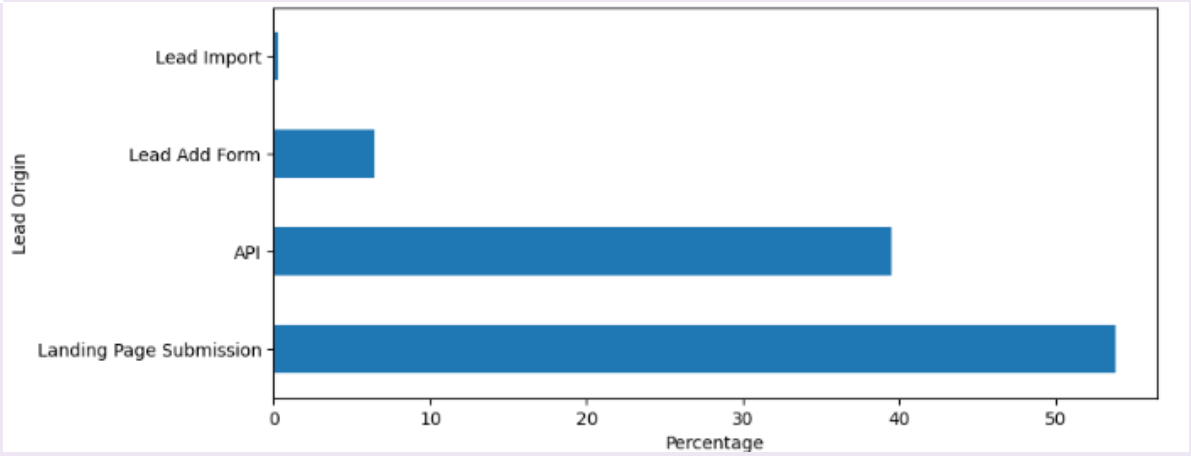
Our target variable is the 'Converted' column, which has binary values, from the percentage distribution, we observe approx. 38% conversion rate.

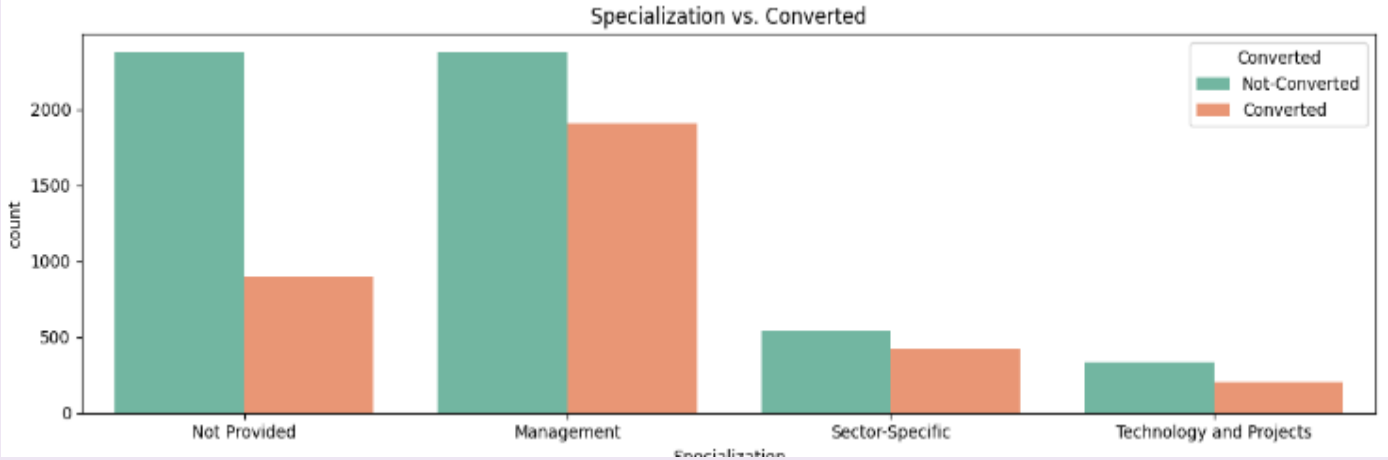




The histogram of ‘TotalVisists’ shows a right-skewed distribution, where max frequency distribution is accumulated for number of total visits between 0 to 5.

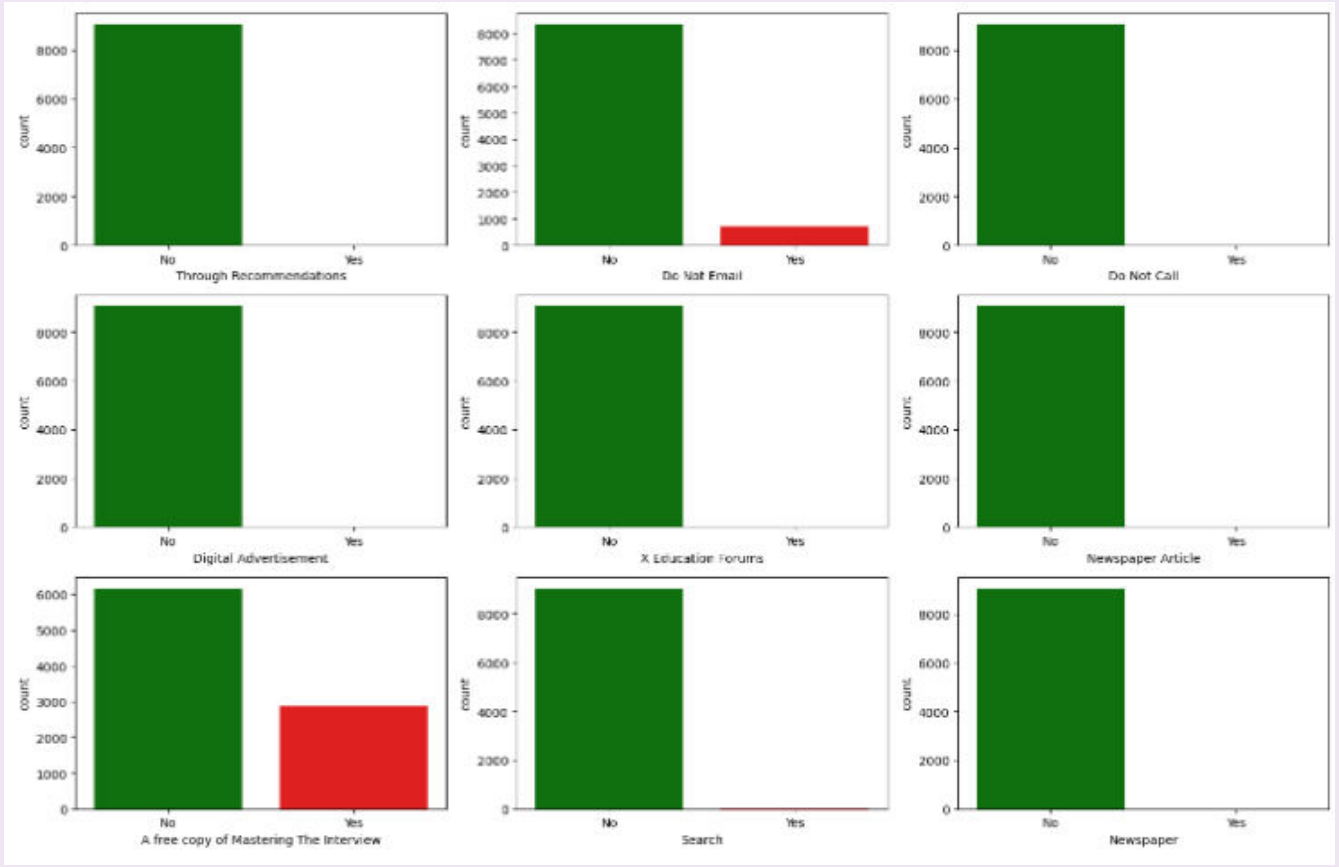
The percentage distribution of 'Lead Origin' variable has 'Landing Page Submission' category with more than 50% of leads being identified. 'API' category also has identified significant number of leads, 'Lead Import' has the lowest percentage to help identify leads.

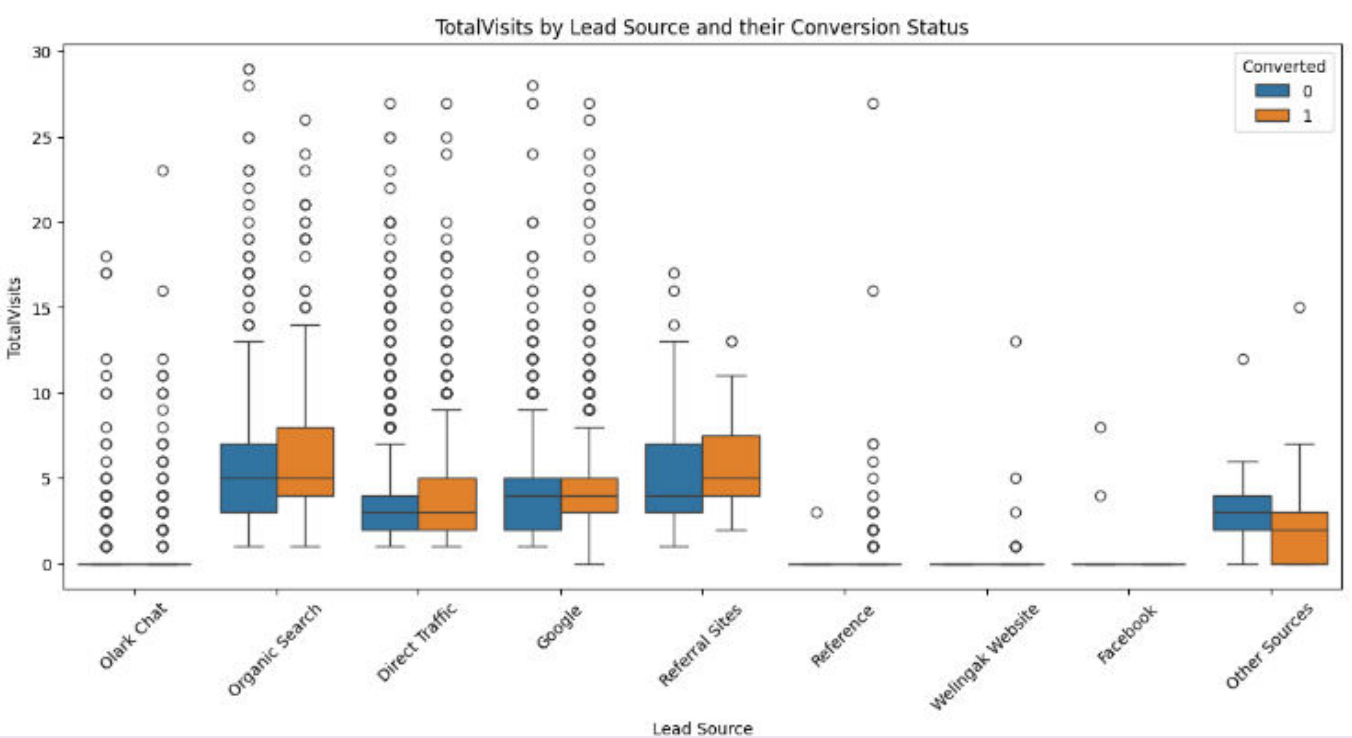




It appears that the majority of customers did not acknowledge seeing the advertisement across all the channels listed in the columns. This is indicated by the dominance of the green bars No over the red bars Yes in each plot.
To simplify our model we drop these columns except variable 'A free copy of Mastering The Interview' and 'Do Not Email'

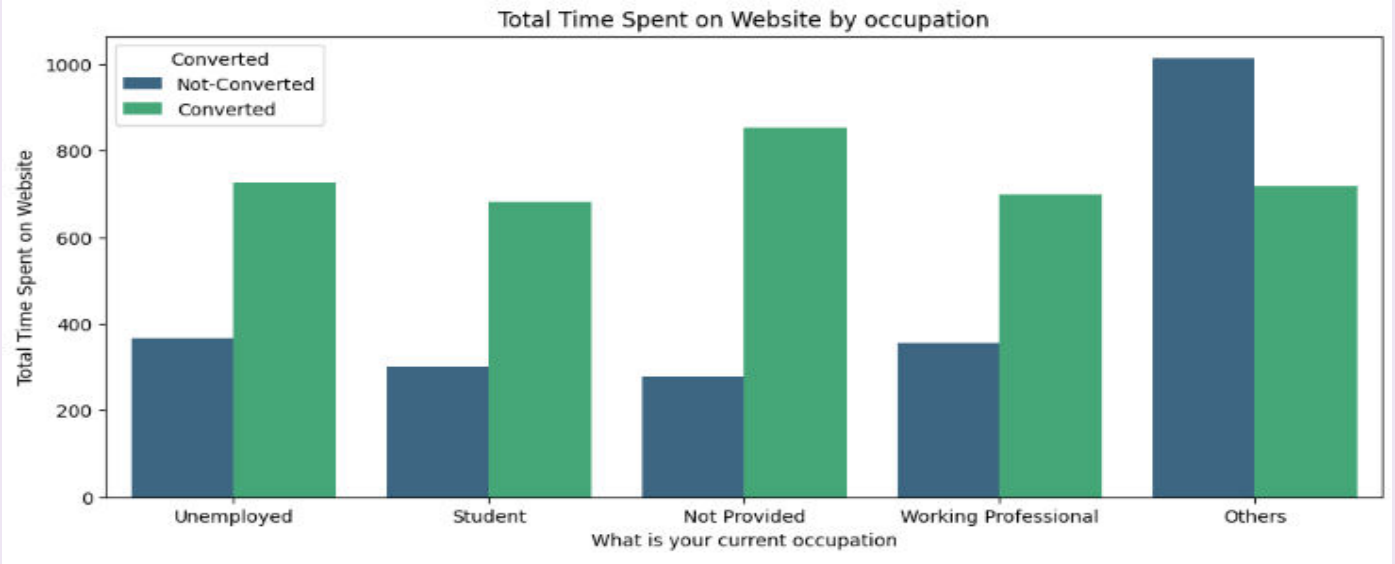
'Management' category shows a highest converted leads, compared to rest of the specialization categories, it might also be due to combined similar specialization across different fields for that category.
Rest specializations in 'Sector-Specific', 'Technology and Projects' sort of have a relatively balanced distribution between converted and non-converted leads.
Many of the leads who have not specified any Specialization, that we can see under the 'Not Provided' category, the conversion rate for this category is higher for the non-converted ones, but still has a reasonable conversion rate.

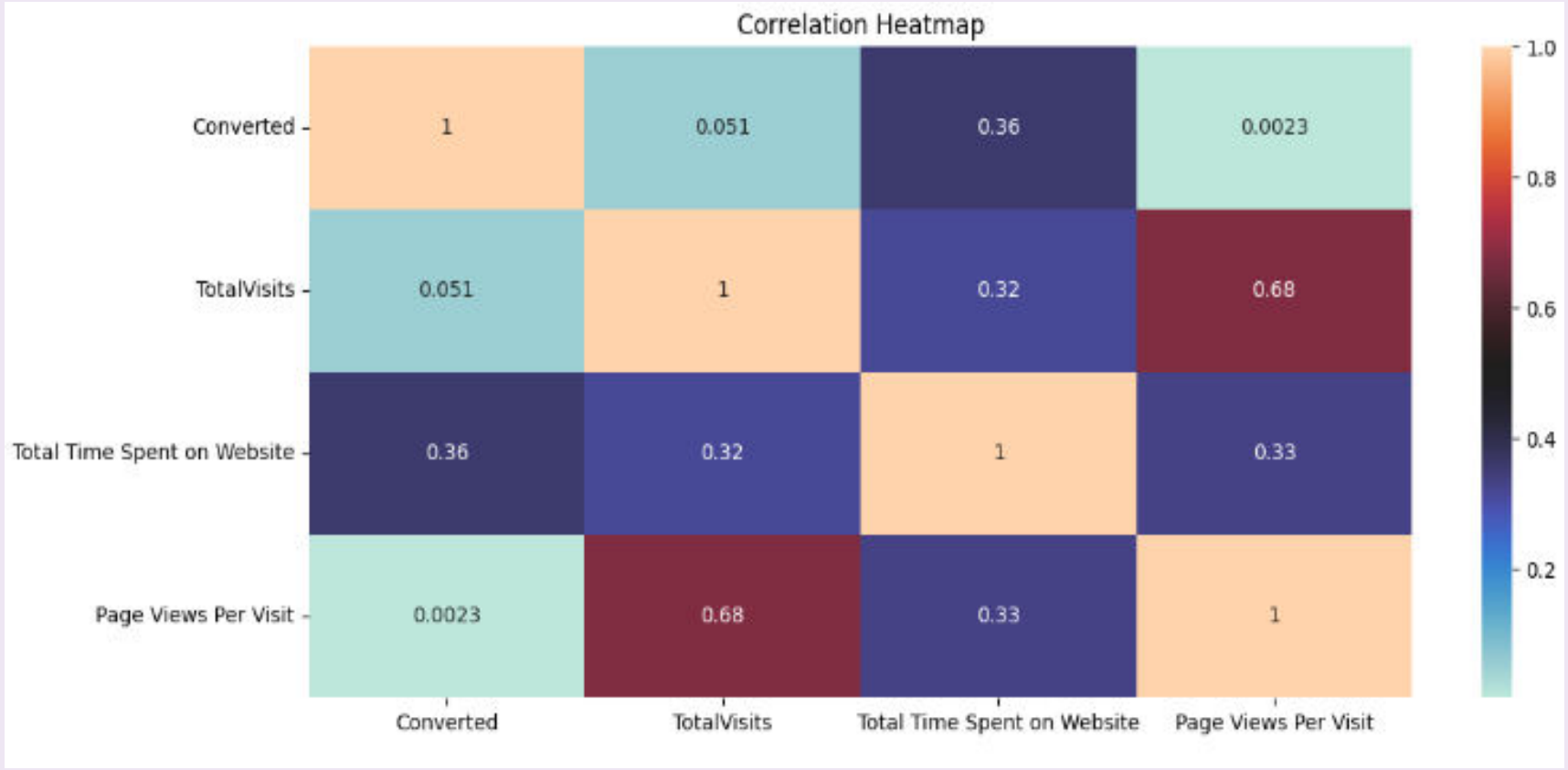




For 'Olark Chat', both converted and non-converted leads have a similar distribution of total visits.
Category 'Organic Search' converted leads have a almost similar median number of total visits compared to non-converted leads.
For 'Direct Traffic' source and 'Google' source the medians seem to be almost same for number of visits. although 'Direct Traffic' source seem to have a higher max visits for converted leads.
For 'Referral Sites', Converted leads have a slightly higher median number of visits compared to non-converted leads.
For 'Reference' there are more visits form the Converted leads compared to the non-converted ones; and for 'Welingak Website' and 'Facebook' both converted and non-converted leads have very few visits, lastly 'Other Sources' have quite some visits with the median of non-converted leads being higher than converted.

Except 'Others' category, almost all types of occupation field , those who have been converted to lead, have spent almost equal amount of time on the website, which is almost more than 700 minutes; and the ones who did not get converted have spent less time comparared to the converted ones.





- There is a weak positive correlation between TotalVisits and conversion rate.
- Total Time Spent on Website has a relatively moderate positive correlation (0.31) with Converted, might suggests that customers who spend more time on the website are more likely to convert.
- Page Views Per Visit has a very weak positive correlation (0.002) with the target variable, so it doesn't seem to have much influence on the conversion leads.

Data Preparation for Model Building

- We evaluate the data types of all the columns to ensure all columns are correctly formatted before proceeding.
- Then create dummy variable for all the categorical variables present in our dataset, the ones showing ‘object’ datatypes.
- Apply one-hot encoding specifying prefixes and dropping one category to avoid multicollinearity, which will then be our reference category.
- Also as shown, the complete dataset is of ‘object’ data type, so we convert from 'object' to 'integer' format, so this format that can be provided to machine learning algorithms for better prediction performance.

```
Lead Origin                                object
Lead Source                                object
Do Not Email                               object
Converted                                   int64
TotalVisits                               float64
Total Time Spent on Website                 int64
Page Views Per Visit                      float64
Last Activity                              object
Specialization                             object
What is your current occupation            object
A free copy of Mastering The Interview     object
Last Notable Activity                      object
dtype: object
```

Model Building

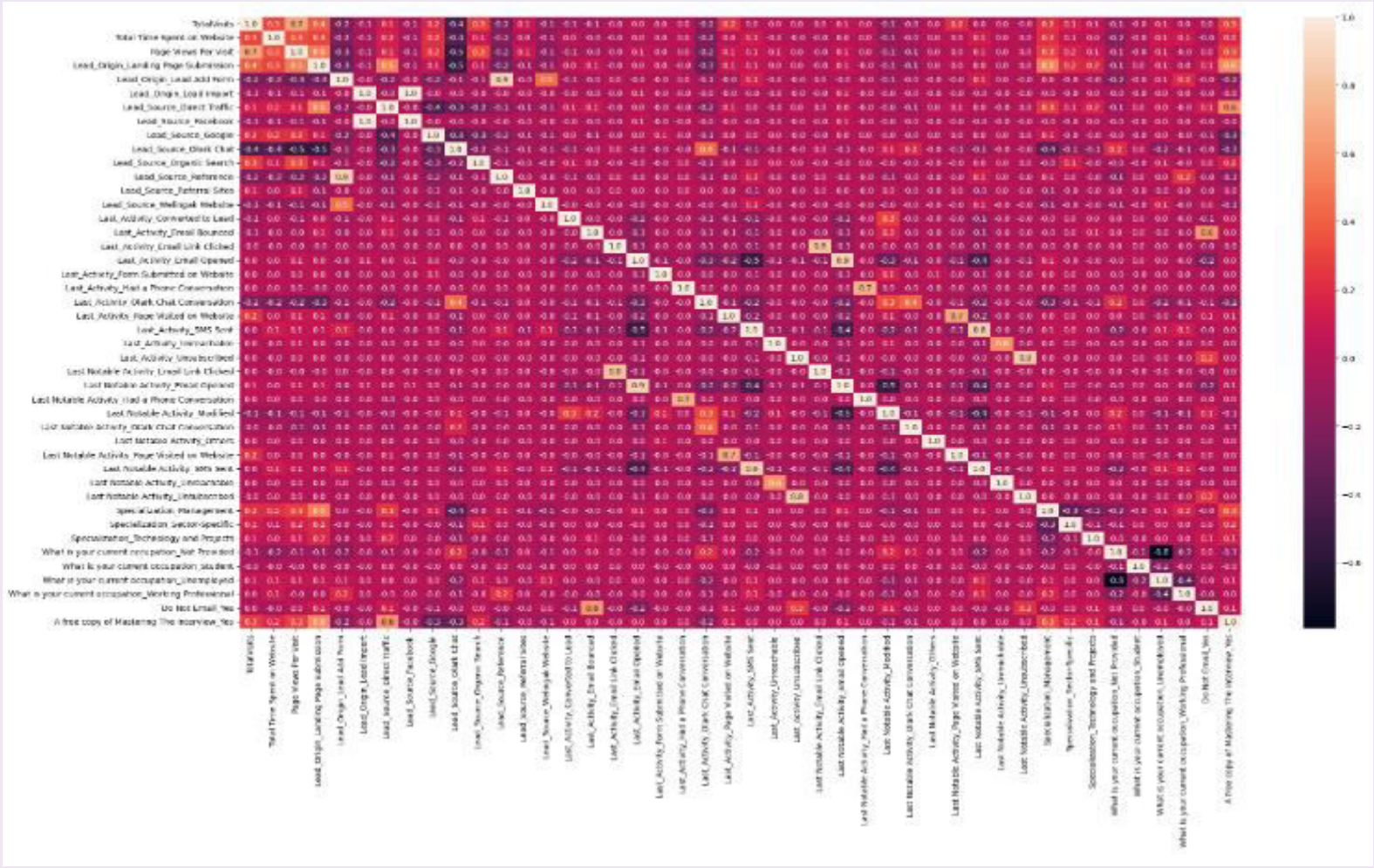
- To define Features and Target, create X by selecting all features except the target variable and assign the target variable to y.
- Test-Train Split: Split the dataset into training and testing sets using a 70:30 ratio and set a random state of 100 to ensure reproducibility.

Model Building

Feature Scaling: Standardize X_train using StandardScaler from the scikit-learn library to normalize numerical variables

Correlation Analysis

- Analyze the correlation between all features using .corr() to understand relationships among variables.
- Given the number of features, prioritize feature selection using Recursive Feature Elimination (RFE) for efficiency.



Model Building

Feature Selection with RFE: Implement Recursive Feature Elimination (RFE) to select the most relevant features for this instantiate LogisticRegression() as the estimator within RFE and we want 15 variables to be selected based on their relevance.

```
Index(['Total Time Spent on Website', 'Lead_Origin_Lead Add Form',
       'Lead_Source_Direct Traffic', 'Lead_Source_Google',
       'Lead_Source_Organic Search', 'Lead_Source_Referral Sites',
       'Lead_Source_Welingak Website',
       'Last_Activity_Had a Phone Conversation',
       'Last_Activity_Olark Chat Conversation', 'Last_Activity_SMS Sent',
       'Last Notable Activity_Had a Phone Conversation',
       'Last Notable Activity_Unreachable',
       'What is your current occupation_Not Provided',
       'What is your current occupation_Working Professional',
       'Do Not Email_Yes'],
      dtype='object')
```

These are the variables that RFE selected

These variables excluded due to low relevance or high correlation.

```
Index(['TotalVisits', 'Page Views Per Visit',
       'Lead_Origin_Landing Page Submission', 'Lead_Origin_Lead Import',
       'Lead_Source_Facebook', 'Lead_Source_Olark Chat',
       'Lead_Source_Reference', 'Last_Activity_Converted to Lead',
       'Last_Activity_Email Bounced', 'Last_Activity_Email Link Clicked',
       'Last_Activity_Email Opened', 'Last_Activity_Form Submitted on Website',
       'Last_Activity_Page Visited on Website', 'Last_Activity_Unreachable',
       'Last_Activity_Unsubscribed',
       'Last Notable Activity_Email Link Clicked',
       'Last Notable Activity_Email Opened', 'Last Notable Activity_Modified',
       'Last Notable Activity_Olark Chat Conversation',
       'Last Notable Activity_Others',
       'Last Notable Activity_Page Visited on Website',
       'Last Notable Activity_SMS Sent', 'Last Notable Activity_Unsubscribed',
       'Specialization_Management', 'Specialization_Sector-Specific',
       'Specialization_Technology and Projects',
       'What is your current occupation_Student',
       'What is your current occupation_Unemployed',
       'A free copy of Mastering The Interview_Yes'],
      dtype='object')
```

Model Building

Building model on selected features:

- Using the Generalized Linear Model (GLM) from the statsmodels library, specifically with the Binomial family to handle binary outcomes, we added a constant term for the intercept and trained the model on the training data (X_train).
- The model summary in the image shows insights into feature coefficients and we see that two columns 'Last Notable Activity_Had a Phone Conversation', 'Last_Activity_Had a Phone Conversation ' have high p-value (>0.05).
- We also check for VIF values, high VIF values(>5) are to be dropped from the model.

Drop features one by one with high p-values and also check their vif values, after dropping variable, build model again and analyze.

	Features	VIF
7	Last_Activity_Had a Phone Conversation	2.18
10	Last Notable Activity_Had a Phone Conversation	2.17
1	Lead_Origin_Lead Add Form	1.53
9	Last_Activity_SMS Sent	1.47
12	What is your current occupation_Not Provided	1.45
3	Lead_Source_Google	1.39
2	Lead_Source_Direct Traffic	1.33
6	Lead_Source_Welingak Website	1.32
0	Total Time Spent on Website	1.24
13	What is your current occupation_Working Profes...	1.19
8	Last_Activity_Olark Chat Conversation	1.18
4	Lead_Source_Organic Search	1.15
14	Do Not Email_Yes	1.12
5	Lead_Source_Referral Sites	1.01
11	Last Notable Activity_Unreachable	1.01

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6343
Model:	GLM	Df Residuals:	6327
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2570.6
Date:	Mon, 17 Jun 2024	Deviance:	5141.2
Time:	20:27:59	Pearson chi2:	6.51e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.4046
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2413	0.096	2.511	0.012	0.053	0.430
Total Time Spent on Website	1.1240	0.041	27.627	0.000	1.044	1.204
Lead_Origin_Lead Add Form	2.3415	0.237	9.884	0.000	1.877	2.806
Lead_Source_Direct Traffic	-1.6174	0.118	-13.678	0.000	-1.849	-1.386
Lead_Source_Google	-1.2206	0.113	-10.822	0.000	-1.442	-1.000
Lead_Source_Organic Search	-1.5247	0.137	-11.146	0.000	-1.793	-1.257
Lead_Source_Referral Sites	-1.7313	0.374	-4.630	0.000	-2.464	-0.998
Lead_Source_Welingak Website	1.7627	0.755	2.335	0.020	0.283	3.242
Last_Activity_Had a Phone Conversation	0.5551	1.014	0.547	0.584	-1.432	2.542
Last_Activity_Olark Chat Conversation	-1.3552	0.168	-8.054	0.000	-1.685	-1.025
Last_Activity_SMS Sent	1.3344	0.075	17.694	0.000	1.187	1.482
Last Notable Activity_Had a Phone Conversation	2.4990	1.578	1.584	0.113	-0.594	5.592
Last Notable Activity_Unreachable	1.8396	0.501	3.672	0.000	0.858	2.822
What is your current occupation_Not Provided	-1.3428	0.089	-15.020	0.000	-1.518	-1.168
What is your current occupation_Working Professional	2.4144	0.179	13.509	0.000	2.064	2.765
Do Not Email_Yes	-1.4269	0.171	-8.345	0.000	-1.762	-1.092

Model Building

Final Model

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6343
Model:	GLM	**Df Residuals:**	6328
Model Family:	Binomial	**Df Model:**	14
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2571.9
Date:	Mon, 17 Jun 2024	**Deviance:**	5143.8
Time:	20:28:00	**Pearson chi2:**	6.50e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4043
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2398	0.096	2.496	0.013	0.051	0.428
Total Time Spent on Website	1.1226	0.041	27.612	0.000	1.043	1.202
Lead_Origin_Lead Add Form	2.3385	0.237	9.871	0.000	1.874	2.803
Lead_Source_Direct Traffic	-1.6144	0.118	-13.661	0.000	-1.846	-1.383
Lead_Source_Google	-1.2165	0.113	-10.795	0.000	-1.437	-0.996
Lead_Source_Organic Search	-1.5287	0.137	-11.174	0.000	-1.797	-1.261
Lead_Source_Referral Sites	-1.7487	0.377	-4.641	0.000	-2.487	-1.010
Lead_Source_Welingak Website	1.7661	0.755	2.339	0.019	0.286	3.246
Last_Activity_Had a Phone Conversation	1.8129	0.782	2.319	0.020	0.281	3.345
Last_Activity_Olark Chat Conversation	-1.3548	0.168	-8.054	0.000	-1.685	-1.025
Last_Activity_SMS Sent	1.3339	0.075	17.694	0.000	1.186	1.482
Last Notable Activity_Unreachable	1.8396	0.501	3.672	0.000	0.858	2.821
What is your current occupation_Not Provided	-1.3407	0.089	-15.005	0.000	-1.516	-1.166
What is your current occupation_Working Professional	2.4127	0.179	13.497	0.000	2.062	2.763
Do Not Email_Yes	-1.4265	0.171	-8.345	0.000	-1.762	-1.092

	Features	VIF
1	Lead_Origin_Lead Add Form	1.53
9	Last_Activity_SMS Sent	1.47
11	What is your current occupation_Not Provided	1.45
3	Lead_Source_Google	1.39
2	Lead_Source_Direct Traffic	1.33
6	Lead_Source_Welingak Website	1.32
0	Total Time Spent on Website	1.24
12	What is your current occupation_Working Profes...	1.19
8	Last_Activity_Olark Chat Conversation	1.18
4	Lead_Source_Organic Search	1.14
13	Do Not Email_Yes	1.12
5	Lead_Source_Referral Sites	1.01
7	Last_Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

All variables have VIF value below 5 and p-value below 0.05, thus no issues of multicollinearity observed, thus we can proceed with prediction using this model.

Model Building

Model prediction on fitted model:

- After fitting the logistic regression model, we use it to predict the probability of leads converting based on the training data (X_train). Using the predict method on the fitted model (res), and we structured them into a DataFrame (y_train_pred_final).
- To classify, we applied a random threshold value of 0.5, so predictions exceeding this threshold were classified as conversions (Conversion_pred = 1), while those below were classified as non-conversions (Conversion_pred = 0).

	Converted	Converted_prob	Index_ID	Conversion_pred
0	0	0.296220	3114	0
1	0	0.148214	6838	0
2	0	0.111723	8263	0
3	1	0.715434	1307	1
4	1	0.384494	2492	0

```
[[3489  438]
 [ 719 1697]]
```

- To evaluate the model's performance create a confusion matrix, to see a detailed breakdown of correct and incorrect predictions.
- The accuracy score, calculated as the proportion of correctly predicted labels against the total predictions, was found to be 81.76%.
- Accuracy score will not be sufficient as we can see the confusion matrix, the ones predicted wrong are left out from the accuracy metrics, thus we need other evaluation metrics.

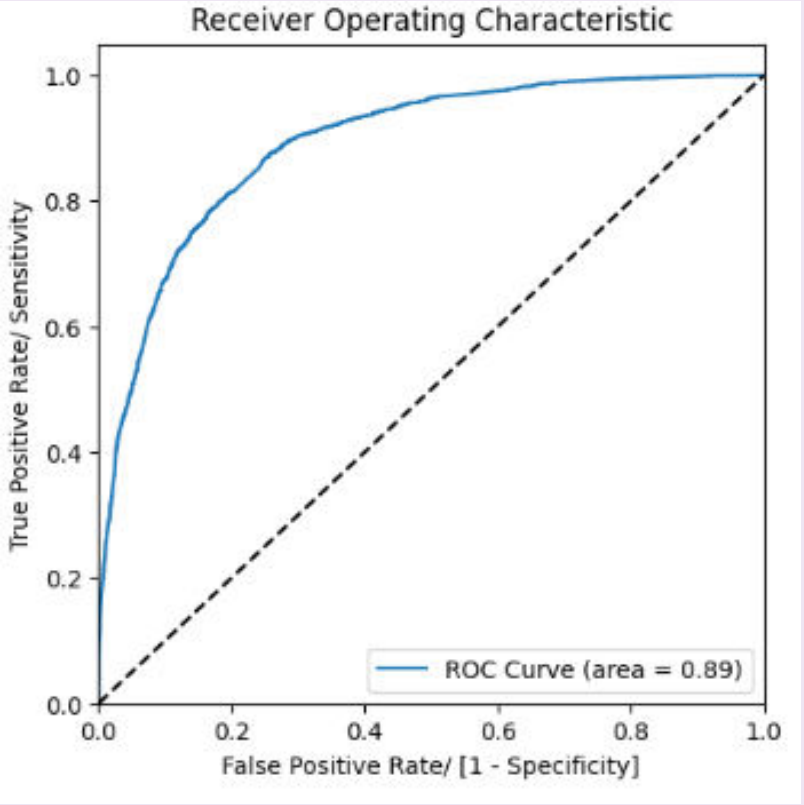
Model Evaluation

We calculated other metrics on the 0.5 cutoff and observed their values as:

Sensitivity: 70.24%
Specificity: 88.84%
Positive Predictive Value (PPV): 79.48%
Negative Predictive Value (NPV): 82.91%
False Positive Rate (FPR): 11.15%

Receiver Operating Characteristic (ROC) Curve:

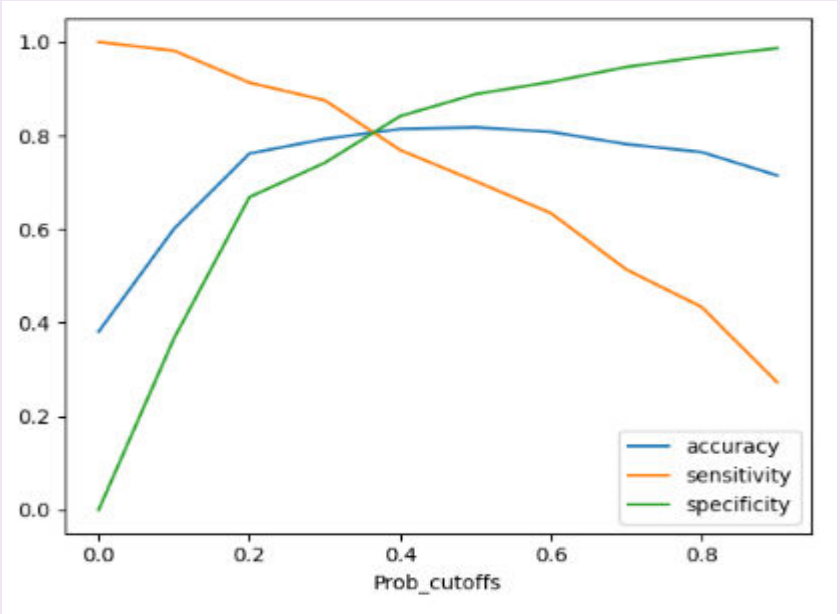
- To visualize the trade-off between sensitivity and specificity across various thresholds, we plotted an ROC curve.
- The curve showcases our model's performance compared to a random classifier (the diagonal line).
- As shown Area Under the Curve is 0.89.
- Our model's curve is positioned towards the upper-left corner, indicating good predictive capability.

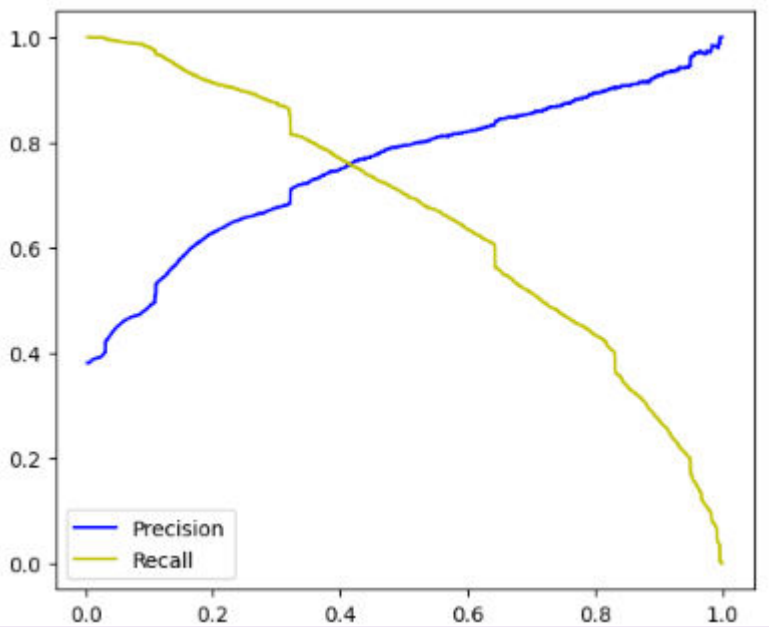


Model Evaluation

Optimal Threshold Determination:

- To enhance model performance, we identified the optimal threshold point by evaluating different probability cutoffs ranging from 0.0 to 0.9.
- After analyzing the metrics at various thresholds, we selected 0.37 as the optimal point, balancing sensitivity, specificity, and accuracy.
- Calculated the metrics again on this cutoff as:
  Accuracy: 81.17, Specificity: 82.58% and Sensitivity: 78.89%





Precision and Recall Evaluation

- Get precision and recall scores from Sklearn library to provide additional evaluation metrics.
  Precision: 73.55
  Recall: 78.97
- Plot the curve, analyzed the trade-off between precision and recall across different threshold values. Chose 0.42 as the cutoff to balance precision and recall effectively.

Model Evaluation

- **Prediction on Test Set:**
- Applied the same scaler used for training data to transform numerical features on the test set.
- Applied the trained logistic regression model (res) to predict probabilities of conversion on the test dataset.
- Set a decision threshold of 0.42 for predicting conversions based on optimal precision-recall trade-off.
- Created a binary column (predicted_conversion) indicating conversions based on the threshold.
- Calculated lead scores by converting predicted probabilities into percentage shown in column (Lead_score).

	Index_ID	Converted	Converted_prob	predicted_conversion	Lead_score
0	2803	1	0.187880	0	18.79
1	5125	0	0.030497	0	3.05
2	3981	1	0.092731	0	9.27
3	4569	1	0.949017	1	94.90
4	5289	0	0.023433	0	2.34

The metrics score we obtained on the test dataset are:

Accuracy: 80.28%
Sensitivity: 72.46%
Specificity: 84.95%
Precision: 74.21

Lead scores above 42 are our potential leads and the model is 80.28% accurate in its predictions overall.

Recommendations

- Focus should be on leads with high predicted probabilities (lead scores above 42), these are the ones with highest likelihood to convert; thus priority should be given to them, constant interactions and maximize phone calls should be made to these leads and tracking their responses will help in better follow-ups.
- The leads that are identified through ‘Add Form’ origin, the ones who have responded positively with interactions through ‘SMS sent’ and ‘Had a Phone Conversation’, also the ones with ‘Working Professional’ as specified in their occupation, all these features seem to contribute highly towards leads conversion.
- Updating the data and try setting different threshold accordingly balancing the precision and recall metrics to ensure effectiveness in lead classification, may also help in better predicting which leads can we put in the ‘Hot Leads’ class or ‘Cold Leads’ class for increasing follow-up actions.
- Trying to pursued leads to select proper option and provide details about them as much as possible will make our prediction model better and thus we can better identify the ones that are most likely to convert.