Lead Scoring Logistic Regression Model Case Study

ANKIT KUMAR
CHETNA JOSHI
HARSHIT GOYAL

D70 JULY 2024

“ Wisdom is knowing what to do next. Skill is knowing how to do it. Virtue is doing it. ”

by Thomas Jefferson

Business Objective

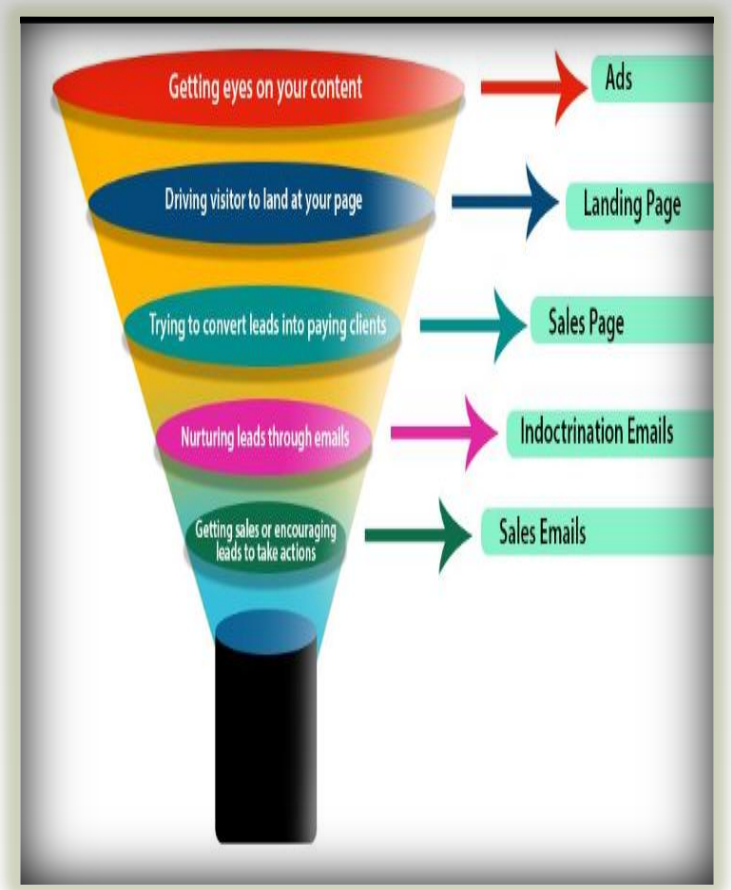
X Education(Lets name it as) has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company requires you to build a model wherein you need 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.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Identification of such promising leads using Logistic regression Model is the aim of this case study.

Domain Understanding



Any education company can sell 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 can then market 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.

A typical lead conversion process can be represented using the funnel beside
As you can see, there are a lot of leads generated in the initial stage but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

Problem Statement

An education company named X Education similarly sells online courses to industry professionals.

The typical lead conversion rate at X education is around 30%.

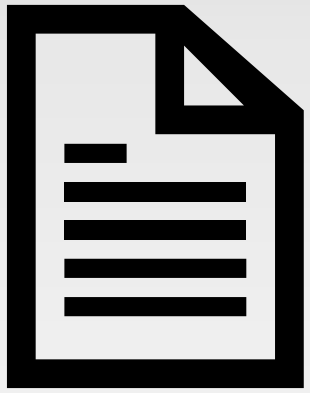
Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Data sourcing and Data understanding

We are provided with a lead's dataset from the past with around 9000 data points.

This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. They may or may not be useful in ultimately deciding whether a lead will be converted or not.

The target variable, in this case, is the column 'Converted' which can tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.



Data sourcing and Data understanding contd.

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	...	Get updates on DM Content	Lead Profile	City	Asymmetrique Activity Index	Asymmetrique Profile Index
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	...	No	Select	Select	02.Medium	02.Medium
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	...	No	Select	Select	02.Medium	02.Medium
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	...	No	Potential Lead	Mumbai	02.Medium	01.High
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	...	No	Select	Mumbai	02.Medium	01.High
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	...	No	Select	Mumbai	02.Medium	01.High

5 rows × 37 columns

- The original dataset provided contains data for 9240 leads along with 37 columns
- Different features include Lead Origin, Lead Score, Total Visits, Do not Email, Do not Call, Total Time spent on webpage, City etc

Some columns description.

Variables	Description
Prospect ID	A unique ID with which the customer is identified.
Lead Number	A lead number assigned to each lead procured.
Lead Origin	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
Lead Source	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
Converted	The target variable. Indicates whether a lead has been successfully converted or not.
TotalVisits	The total number of visits made by the customer on the website.
Do Not Email	An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.
Do Not Call	An indicator variable selected by the customer wherein they select whether of not they want to be called about the course or not.
Total Time Spent on Website	The total time spent by the customer on the website.
City	The city of the customer.
What is your current occupation	Indicates whether the customer is a student, unemployed or employed.

EDA PROCESS AND SOLUTION APPROACH

Data Sourcing , Cleansing and Preparation

Fixing datatypes & rows.	Missing Values	Handling Outliers	Exploratory Data Analysis	Feature standardization

Feature Scaling , Splitting Train and Test data sets

Feature scaling of Numeric data	Splitting the data into train and test sets

Model Building

Feature selection using RFE	Determine the Optimal model using Logistic Regression	Evaluate the Model using Metrics Accuracy, sensitivity, precision and recall

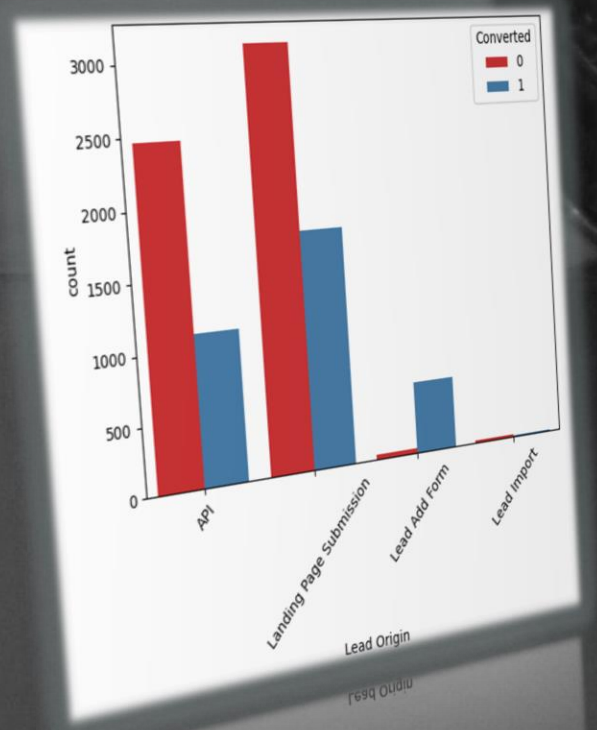
Result

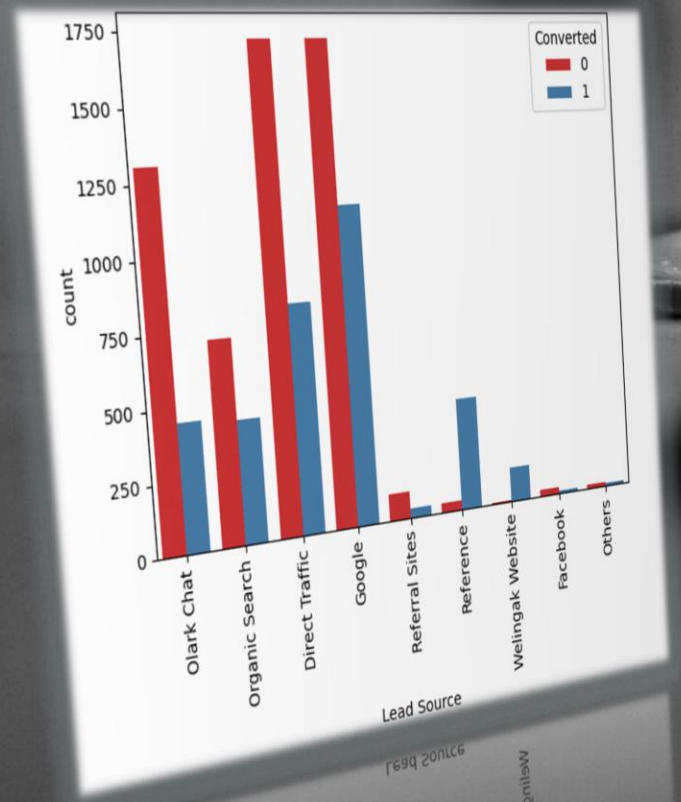
Determine the lead score and check if final predictions results in 80% lead conversion.	Evaluate the final prediction the test set using cut off threshold from sensitivity and specificity metrics.

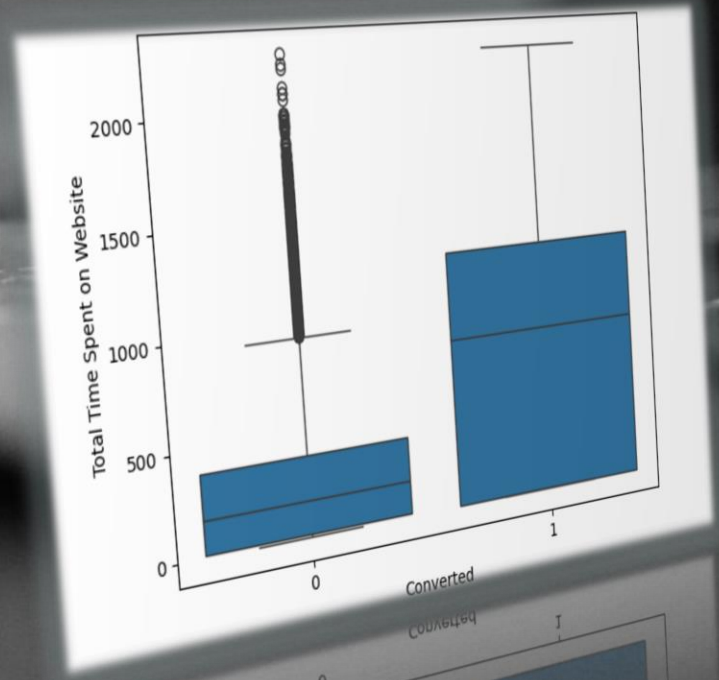
Data Cleansing

- The original dataset provided contains data for 9240 leads along with 37 columns
- Different features include Lead Origin, Lead Score, Total Visits, Do not Email, Do not Call, Total Time spent on webpage, City etc.
- Some of the data needed cleaning steps to remove columns that are not very useful for the model building due to class imbalance, lot of null values etc. (more than 25% of the rows)

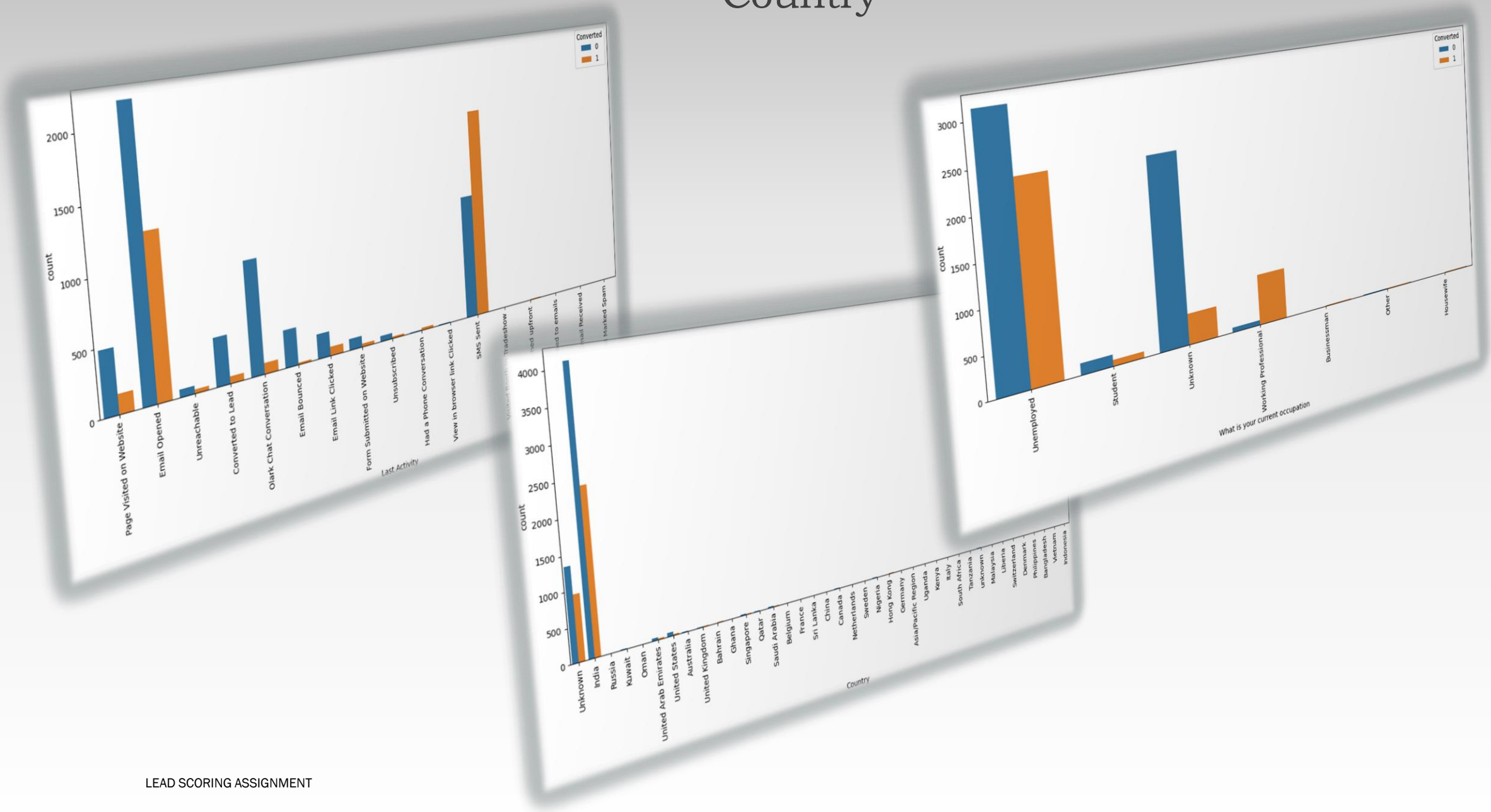
Distribution of the datapoints for Lead Origin Lead source and Total time spent on Website



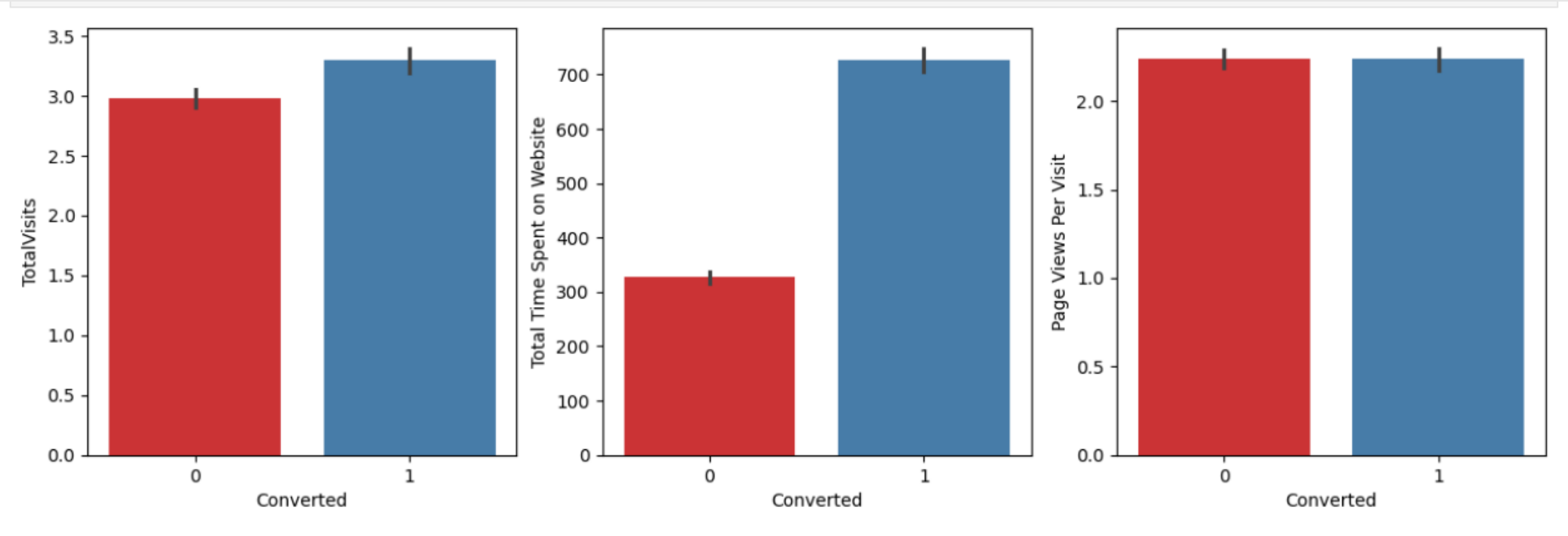




Distribution of Last Activity , What is your current occupation and Country

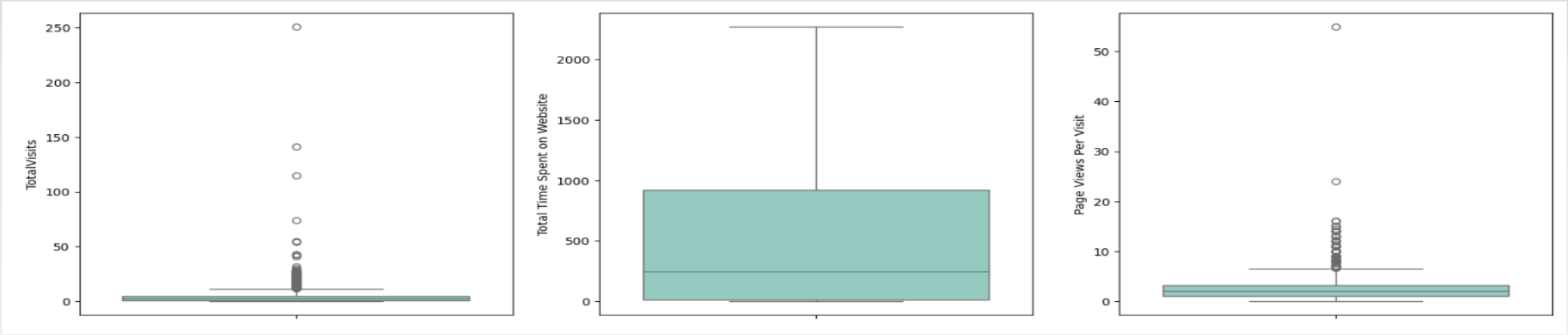


The conversion rated were high for Total Visits, Total Time Spent on Website and Page Views Per Visit

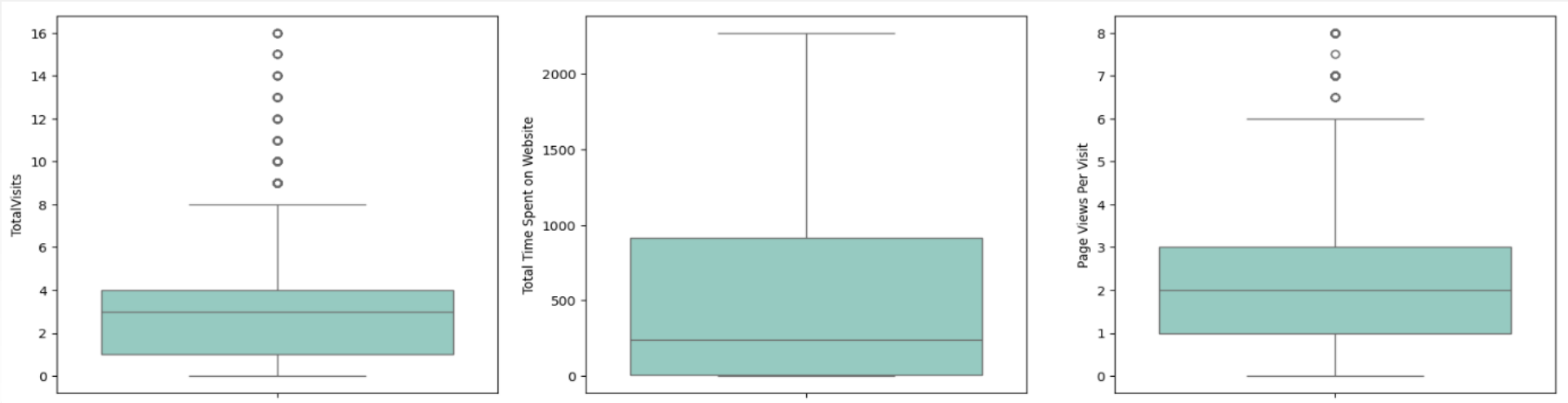


Handling Outliers.

From the following view, it can be seen that outlier exists in the columns TotalVisits and Page Views Per Visit columns which can be handled



After clearing outliers



Final Data Preview

- Columns that had very high class imbalance, with only a very few data points on one class and that are not very critical to the analysis were removed. Example of this was for example : „I agree to pay the amount through cheque“, all the data points were „No“.
- Based on the distribution of data points and the business understanding we could already remove some of the features that are not vital for the model.
- After the data cleaning, 95% of the data was still available for the model building.
- As part of the Data Preparation, we converted some binary variables (Yes/No) to 0/1 for the columns ('Do Not Email’, 'Do Not Call’) and encoded other object columns to binary fields.

Data Splitting and Feature Scaling

The data was split into two datasets : training and testing datasets in a ratio of 70:30

For standardization of the datapoints, Standard Scalar scaling was applied on the numerical variables to transform the features of this dataset so that they have a mean of zero and a standard deviation of one.

This avoids the bias in the model coefficients if some points are very different in scale.

Using RFE feature selection method of Logistic Regression we choose the most optimal 15 variables for our Model building

Model Building using RFE Model 1

- The initial model was built choosing 15 variables using the RFE technique First model obtained.
- Some of the variables were insignificant due to high p-value.
- VIF was also calculated to find the multi-colinearity existing between features, if any.
- Some features were eliminated in this two ways

Generalized Linear Model Regression Results			
Dep. Variable:	Converted	No. Observations:	6204
Model:	GLM	Df Residuals:	6188
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1
Method:	IRLS	Log-Likelihood:	-2716.2
Date:	Sat, 25 Jan 2025	Deviance:	5432.5
Time:	23:22:20	Pearson chi2:	6.28E+03
No. Iterations:	19	Pseudo R-squ. (CS):	0.3667
Covariance Type:	nonrobust		

	Features	VIF
10	What is your current occupation	30.49
12	Lead Profile	19.69
14	Last Notable Activity	16.11
9	How did you hear about X Education	15.86
7	Last Activity	12.89
8	Specialization	8.51
13	City	5.47
0	Lead Origin	3.72
1	Lead Source	3.4
6	Page Views Per Visit	2.33
4	TotalVisits	2.21
11	What matters most to you in choosing a course	2.06
5	Total Time Spent on Website	1.27
2	Do Not Email	1.16
3	Do Not Call	1

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.2052	0.377	-11.167	0	-4.943	-3.467
Lead Origin	0.4269	0.073	5.838	0	0.284	0.57
Lead Source	0.2934	0.022	13.493	0	0.251	0.336
Do Not Email	-1.5382	0.168	-9.138	0	-1.868	-1.208
Do Not Call	19.5431	1.21E+04	0.002	0.999	-2.36E+04	2.36E+04
TotalVisits	0.103	0.048	2.133	0.033	0.008	0.198
Total Time Spent on Website	1.0015	0.039	25.928	0	0.926	1.077
Page Views Per Visit	-0.4484	0.052	-8.648	0	-0.55	-0.347
Last Activity	0.1468	0.015	9.496	0	0.117	0.177
Specialization	-0.0131	0.008	-1.722	0.085	-0.028	0.002
How did you hear about X Education	0.0123	0.019	0.655	0.513	-0.024	0.049
What is your current occupation	0.8619	0.072	11.955	0	0.721	1.003
What matters most to you in choosing a course	-0.6912	0.037	-18.435	0	-0.765	-0.618
Lead Profile	-0.4244	0.036	-11.709	0	-0.495	-0.353
City	0.0148	0.019	0.763	0.446	-0.023	0.053
Last Notable Activity	0.0379	0.017	2.219	0.027	0.004	0.071

Final Model

- After Multiple iterations of Modelling and choosing the Optimal parameters based on Lowest p-value and Lowest VIF value. We arrived at the Model 7.
- It has in total 9 variables.

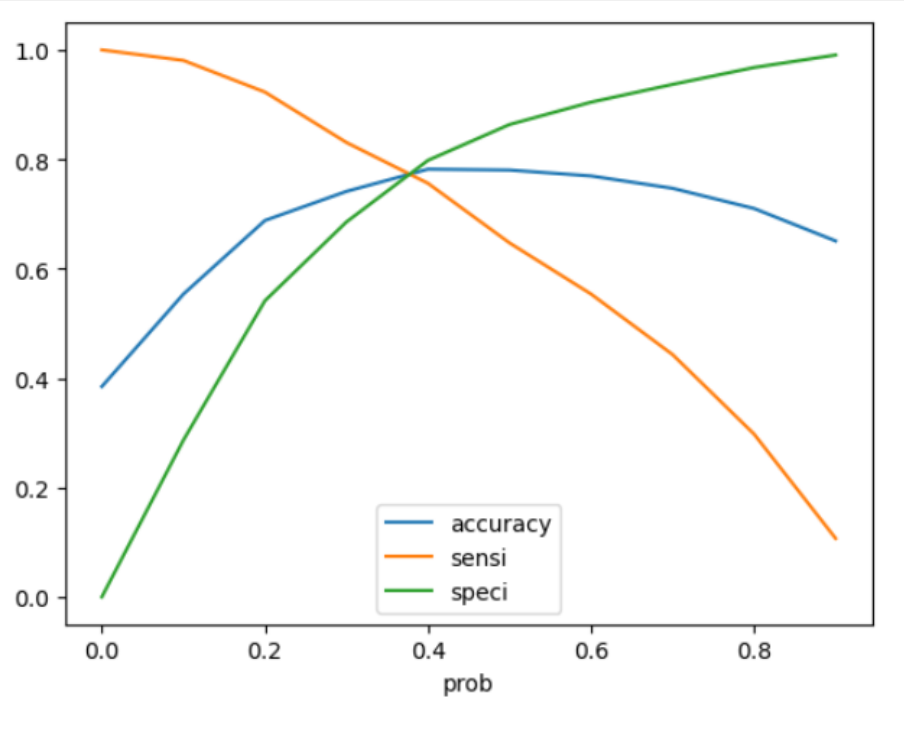
Generalized Linear Model Regression Results			
Dep. Variable:	Converted	**No. Observations:**	6204
Model:	GLM	**Df Residuals:**	6194
Model Family:	Binomial	**Df Model:**	9
Link Function:	Logit	**Scale:**	1
Method:	IRLS	**Log-Likelihood:**	-2896.5
Date:	Sun, 26 Jan 2025	**Deviance:**	5793
Time:	00:11:46	**Pearson chi2:**	6.16E+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.3289
Covariance Type:	nonrobust		

	Features	VIF
7	Specialization	4.25
6	Last activity	4.16
1	Lead source	2.99
5	Page views per visit	2.25
3	Totalvisits	2.17
0	Lead origin	2.14
8	What matters most to you in choosing a course	1.50
4	Total time spent on website	1.24
2	Do not email	1.10

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9773	0.128	-15.389	0	-2.229	-1.725
Lead Origin	0.4157	0.064	6.521	0	0.291	0.541
Lead Source	0.3169	0.02	15.888	0	0.278	0.356
Do Not Email	-1.5359	0.163	-9.43	0	-1.855	-1.217
TotalVisits	0.0808	0.047	1.73	0.084	-0.011	0.172
Total Time Spent on Website	1.0158	0.038	26.939	0	0.942	1.09
Page Views Per Visit	-0.4906	0.05	-9.869	0	-0.588	-0.393
Last Activity	0.1614	0.009	17.382	0	0.143	0.18
Specialization	-0.0319	0.007	-4.744	0	-0.045	-0.019
What matters most to you in choosing a course	-0.5362	0.029	-18.768	0	-0.592	-0.48

Model Evaluation - Sensitivity and Specificity on Train Data Set

- In order to select optimal cutoff for the predicted odds, we plot the graphs between the performance characteristics – accuracy, sensitivity and specificity vs probability range from 0 to 1.
- The point of intersection of the three curves is the optimal point.



Confusion Matrix	3297	521
	842	1544

Overall Accuracy	78 %
Sensitivity	64.7 %
Specificity	86.35 %

Model Evaluation - Sensitivity and Specificity on Test Data Set

- In order to select optimal cutoff for the predicted odds, we plot the graphs between the performance characteristics – accuracy, sensitivity and specificity vs probability range from 0 to 1.

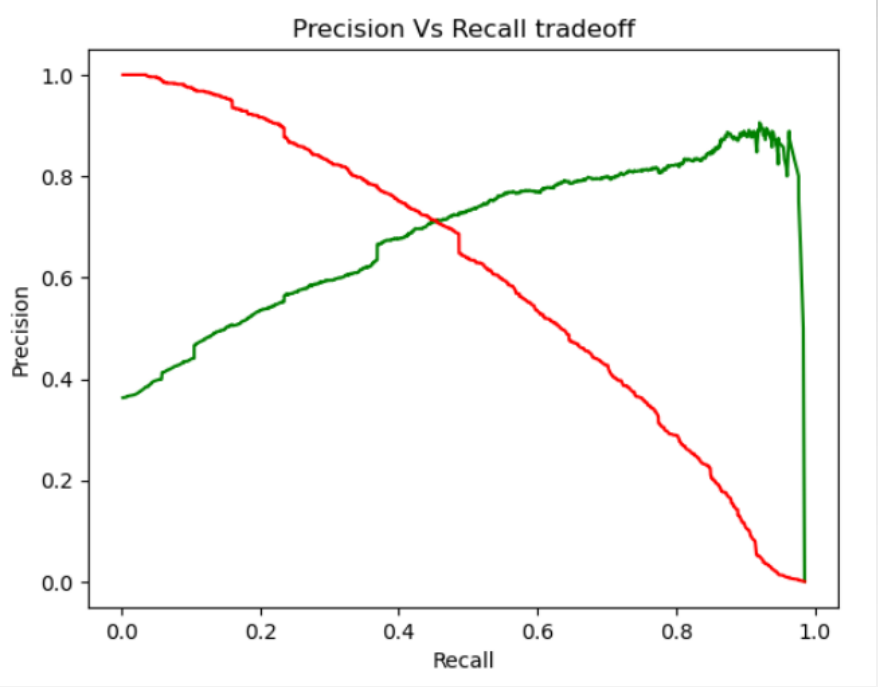
Overall Accuracy	74 %
Sensitivity	79 %
Specificity	71 %

Confusion Matrix

1209	**493**
195	770

Precision and Recall view

- The precision and recall of the model was calculated for training and test datasets after making the predictions in the testing dataset using the optimal cutoff of 0.35 that we calculated in the previous step.
- Now, recall our business objective – because it gives the idea about the number of false negatives. High recall means that there is low number of false negatives.
- Recall could be considered here as more valuable than precision because lower precisions implies lower hot lead predictions, but we don't want to left out any hot leads which are willing to get converted hence our focus on this will be more on Recall than Precision.
- From the Precision and Recall curve, choosing the Cut-off as 0.42. The probabilities above 0.42 are predicted therefore as converted in the testing dataset.
- Precision = 60.9%
- Recall = 79.9 %



Conclusion

- Lead scores are calculated at 100 * predicted odds
- The model evaluation steps revealed that the accuracy, precision and recall parameters are acceptable values. The Recall score is a greater than precision score as well. This fits the business needs for the future.
- The model evaluation steps concludes that the model is in stable state.
- Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :
  1. Total Time Spent on website
  2. Lead Origin
  3. Lead Source

Recommendations to the Company Strategy

- A good strategy to employ would be to focus on the variables that increase the lead scores. These are Total Time Spent on website , Lead Source and Lead Origin
- In order to increase the time that a user spends on webpage, the company to employ more web developers and UI/UX designers to improve the experience for the user and thereby luring the users to spend more time on the webpage, exploring the contents.
- It is also important not to waste a lot of time on some factors that do not contribute much or negatively affect the lead scores.
- The model predictions of lead score less than 50 could be disregarded and the company could focus more on the 50-100
  lead scores.
- When there is only limited time to focus on phone calls, it would be a good idea to consider only lead scores above 85,
  which means that the probability that the lead is hot is very high and is worth a shot to try.
- The variables to focus during such times are mentioned above. For example, more marketing could be employed in Welingak Website and more contents could be added in the webpage and improve the UI/UX to get the users spend more time on the webpage

Thank you