Lead Score Assignment

- **Anuj Sharma**
- **Gurveen Gill**
- **Shaktiprasanna**

DS C47

Upgrad & IIITB | Data Science Program

28th February 2023

Agenda

01. Problem Statement & Goal

02. Methodology

03. Exploratory Data Analysis

04. Model Training and Assessment

05. Summary

A. Problem Statement

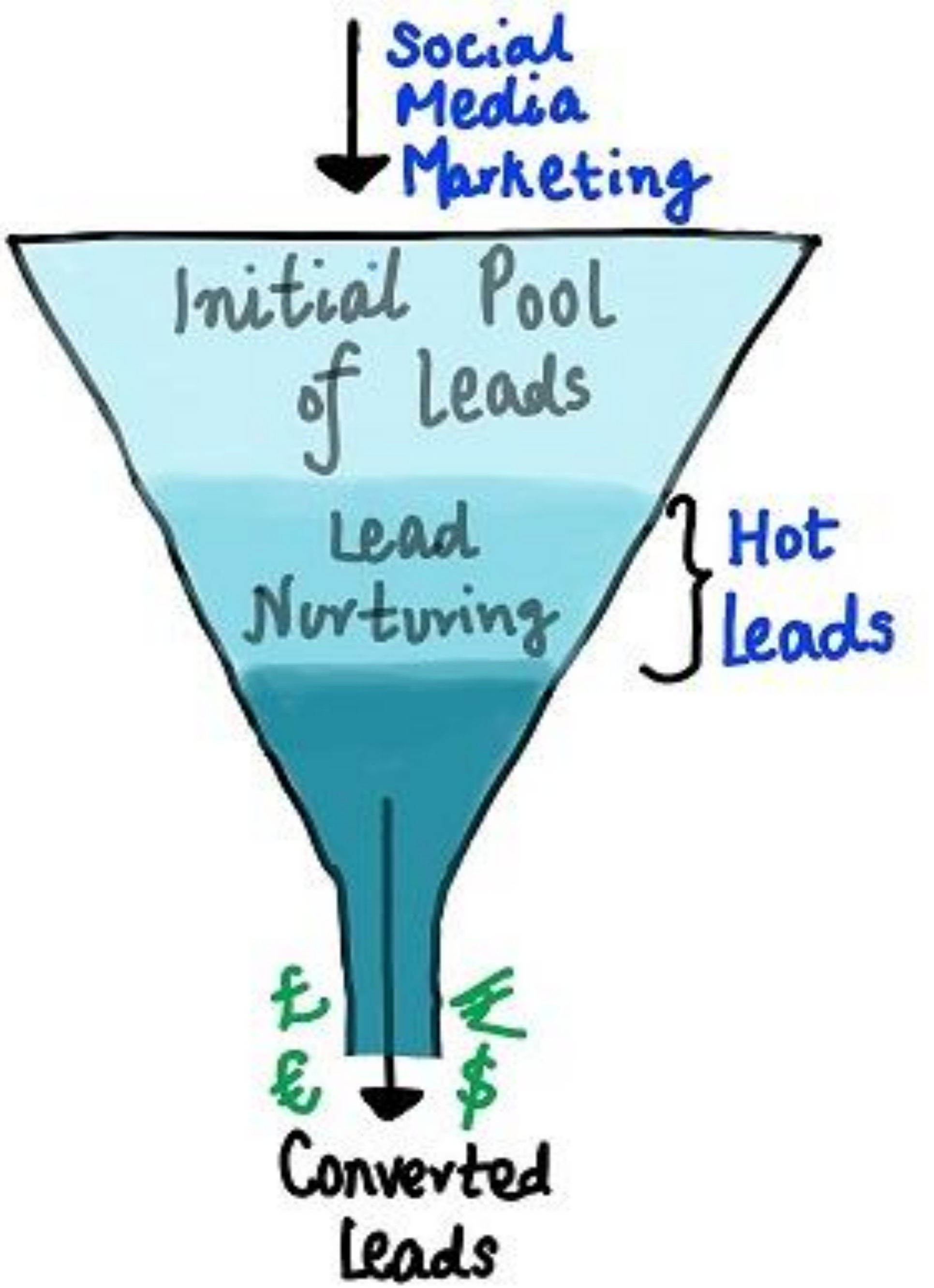
X Education sells online courses to industry professionals and gets a lot of leads on its website every day.

The company wishes to identify the most potential leads, known as 'Hot Leads', to increase its poor lead conversion rate of around 30%.

A lead conversion process can be represented using a funnel, where many leads are generated initially, but only a few become paying customers.

X Education needs a model to assign a lead score to each lead to identify the customers with a higher conversion chance.

The target lead conversion rate is around 80%.



A. GOAL

Develop a logistic regression model to assign lead scores between 0 and 100

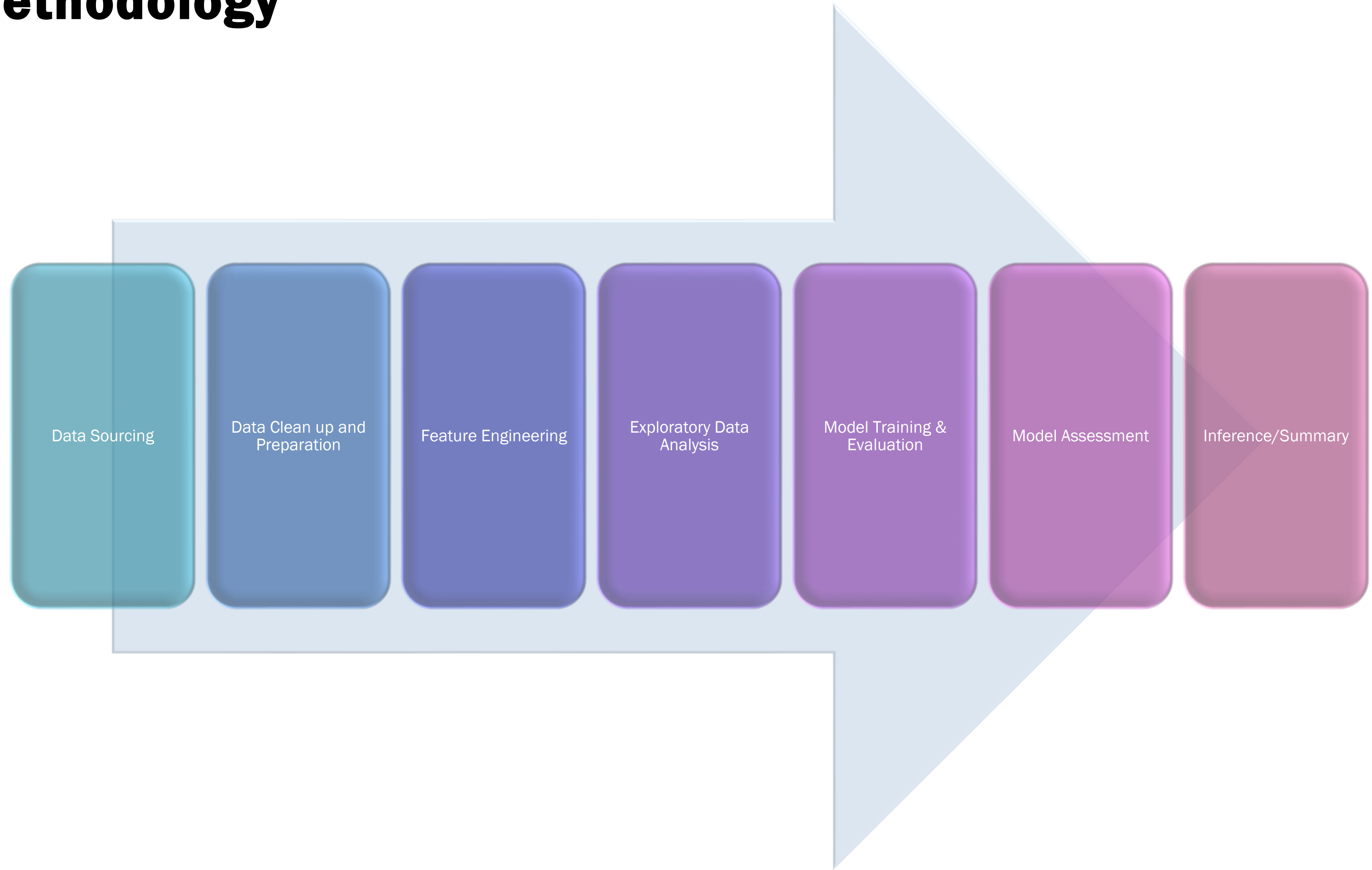
Use lead scores to target potential leads, with higher scores indicating hotter leads more likely to convert and lower scores indicating colder leads less likely to convert

Account for additional problems presented by the company that the model should be able to adjust to in the future

Fill out the separate document based on the logistic regression model from step one

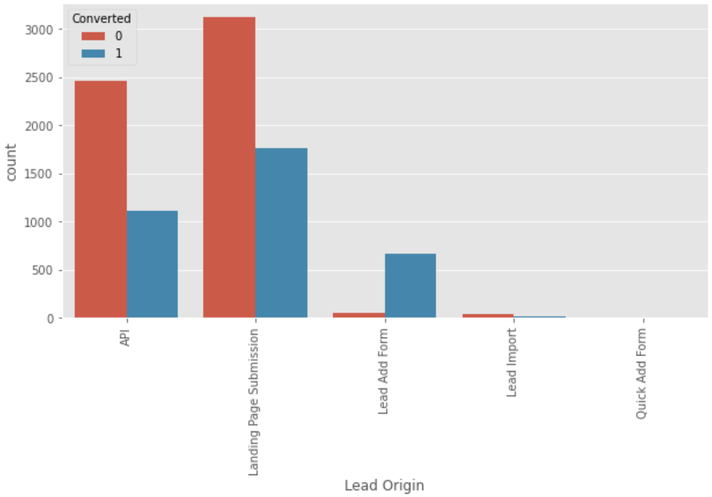
Include all findings and recommendations in the final PowerPoint presentation

B. Methodology

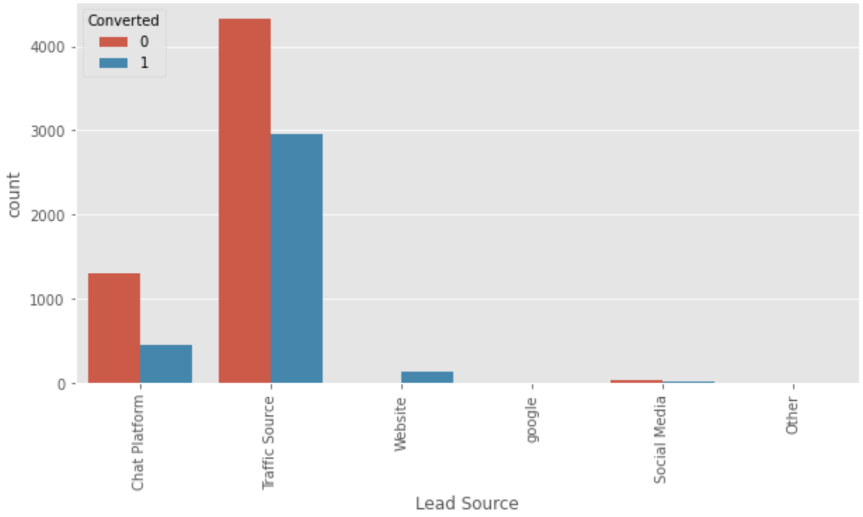


Data Sourcing
Data Clean up and Preparation
Feature Engineering
Exploratory Data Analysis
Model Training & Evaluation
Model Assessment
Inference/Summary

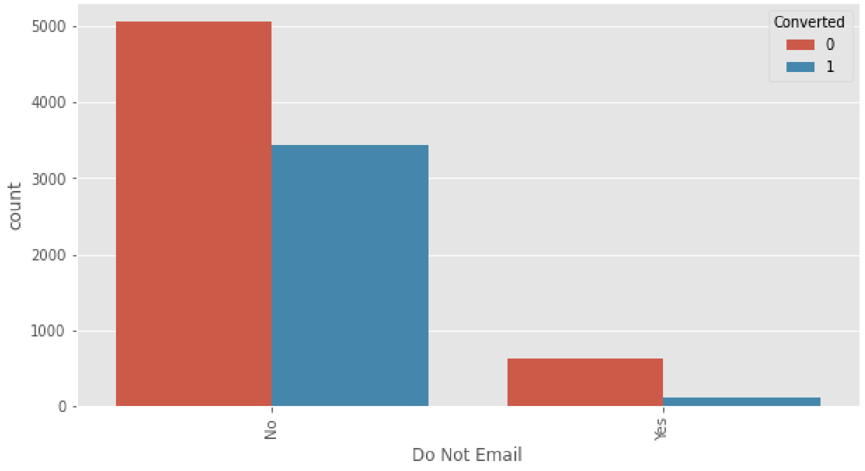
C. Exploratory Data Analysis



- Based on Lead Origin column, API and Landing Page Submission bring a higher number of leads as well as conversions. Lead Add Form has a very high conversion rate, but the count of leads is not very high. Lead Import and Quick Add Form get very few leads.

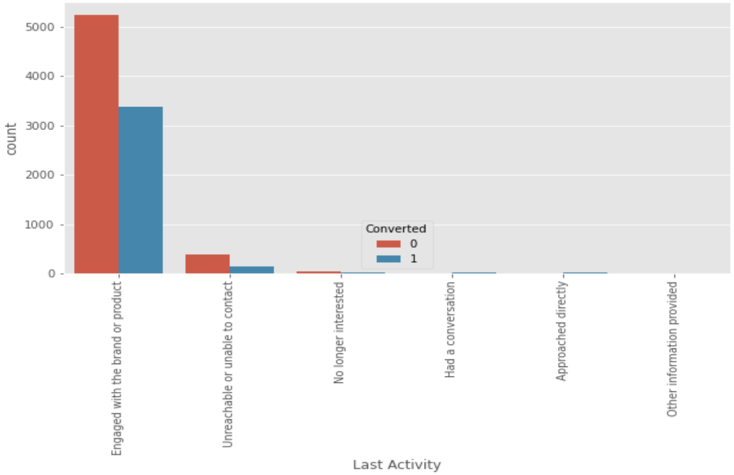


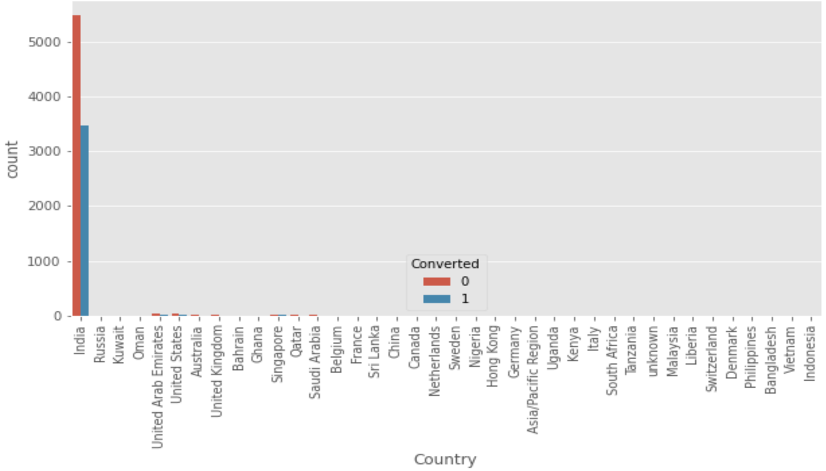
- Based on Lead Source column, it is observed that most leads are sourced from chat platforms and traffic sources such as Google, Direct Traffic, Organic Search, Reference, Referral Sites, Bing, Click2call, Pay per Click Ads, welearnblog_Home, WeLearn, blog, and NC_EDM.

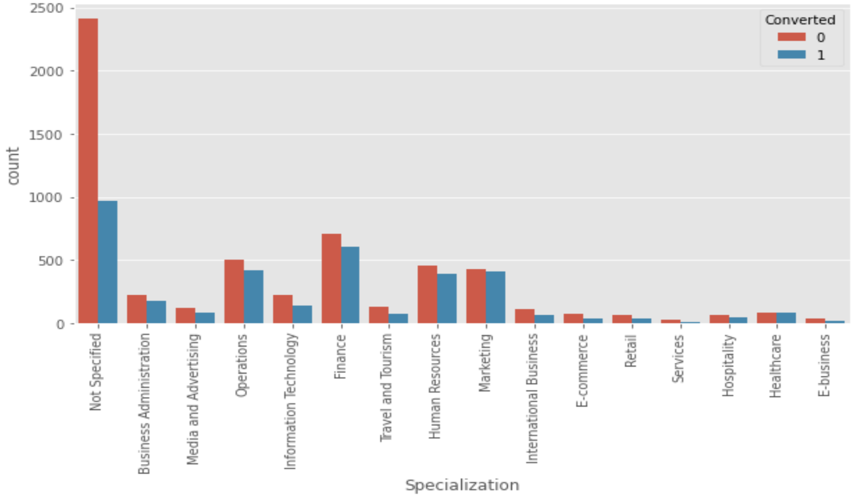


- Interaction Preference column suggests that about 90% of users prefer not to be contacted via phone or email regarding the product.

C. Exploratory Data Analysis





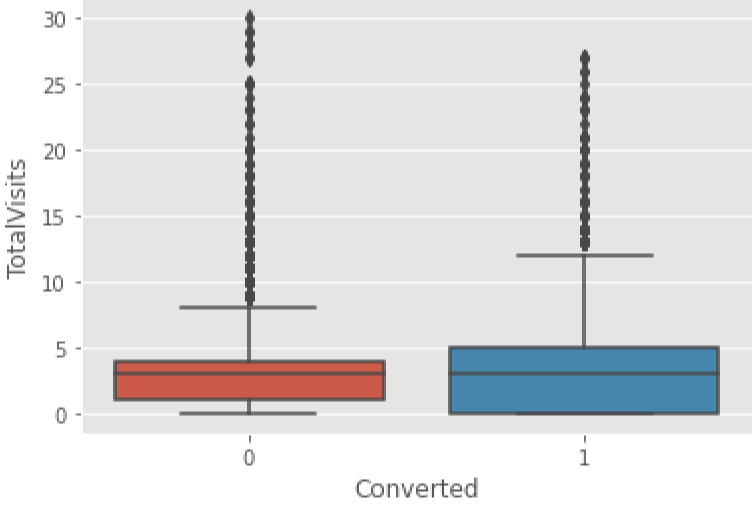


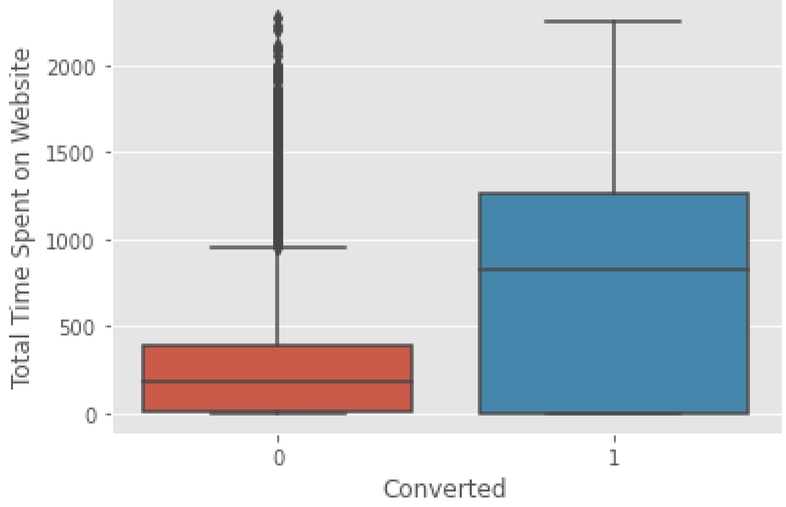
- Last Activity column analysis indicates that more than 80% of users who have been engaged through activities such as "Email Opened", "SMS Sent", "Resubscribed to emails", "Visited Booth in Tradeshow", "Email Received", "View in browser link Clicked", "Form Submitted on Website", "Olark Chat Conversation", "Page Visited on Website", "Converted to Lead", and "Email Link Clicked" are most likely to get converted.

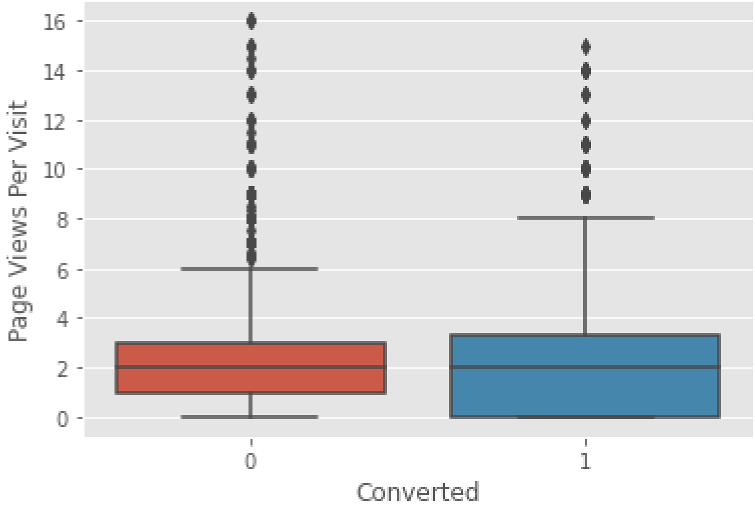
- Country column analysis shows that 90% of the users who got converted and were not converted belong from India, suggesting people from India have a high chance of getting converted.

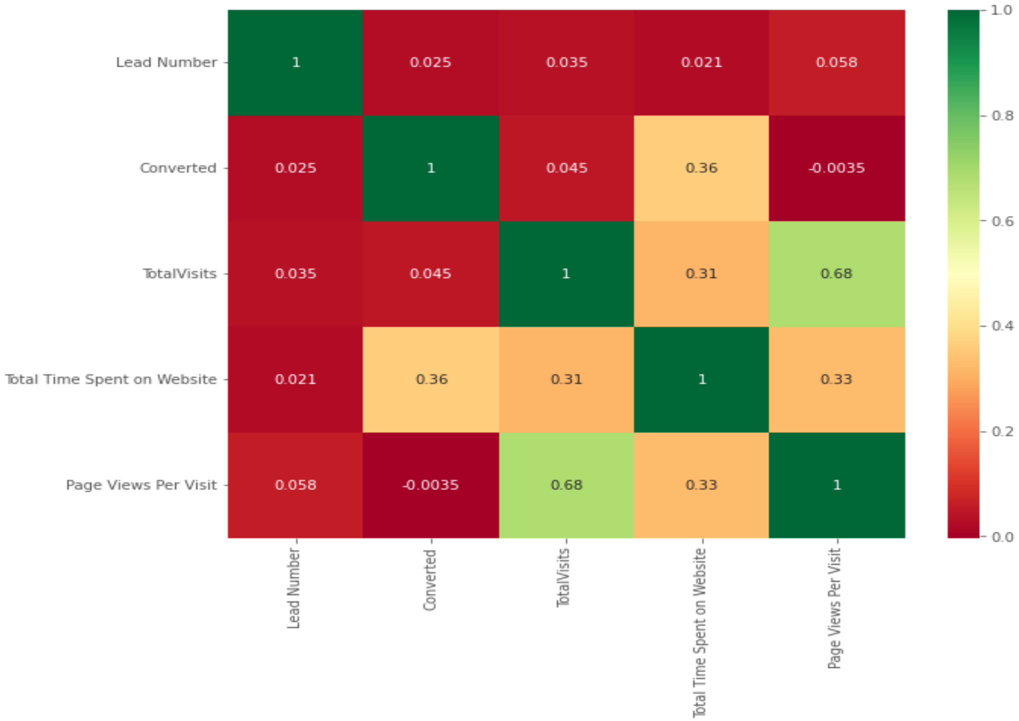
- Specialization column suggests that people enquiring for courses on business management like operations, finance, human resource, market are more likely to get converted.

C. Exploratory Data Analysis







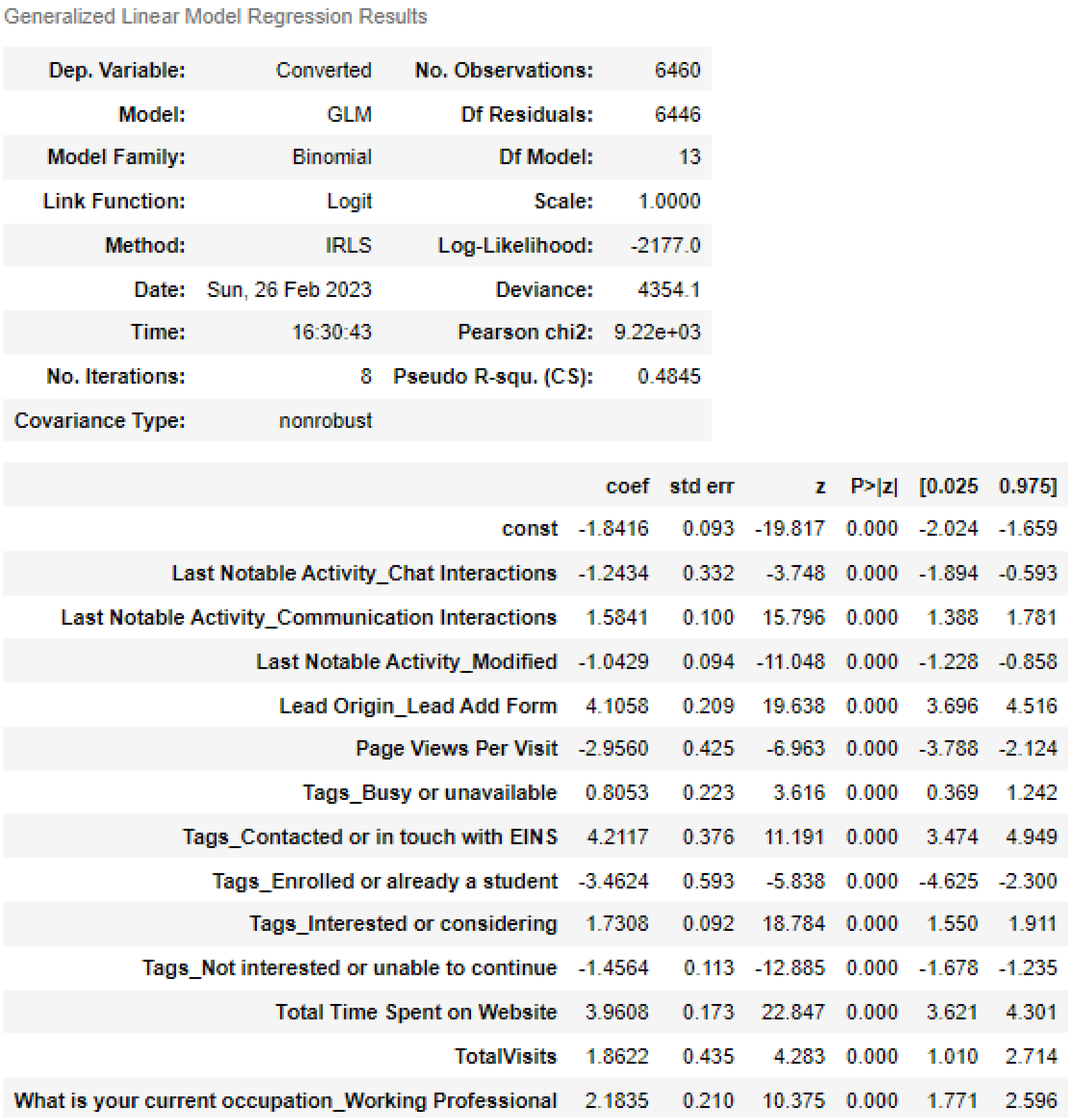


	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Lead Number	1	0.025	0.035	0.021	0.058
Converted	0.025	1	0.045	0.36	-0.0035
TotalVisits	0.035	0.045	1	0.31	0.68
Total Time Spent on Website	0.021	0.36	0.31	1	0.33
Page Views Per Visit	0.058	-0.0035	0.68	0.33	1

Summary:-

- The Total Visits and Page View Per Visit columns have a significant number of outliers. However, after removing the outliers, there was a positive change observed in the correlation matrix.
- There is a strong relationship between the Total Visits and Page View Per Visit columns.
- Individuals who spend more time on the website are more likely to be converted.

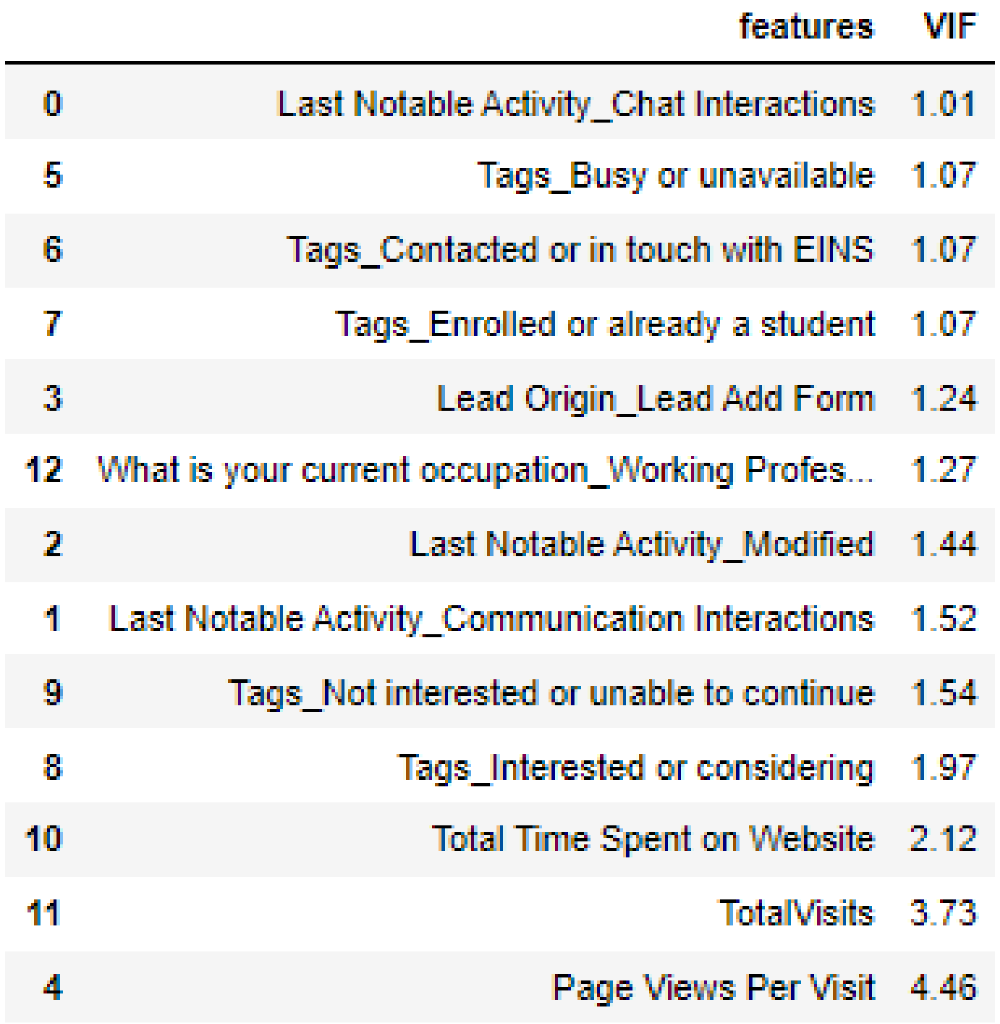
D. Model Training and Assessment – GLM Results



Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6460
Model:	GLM	Df Residuals:	6446
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2177.0
Date:	Sun, 26 Feb 2023	Deviance:	4354.1
Time:	16:30:43	Pearson chi2:	9.22e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.4845
Covariance Type:	nonrobust		

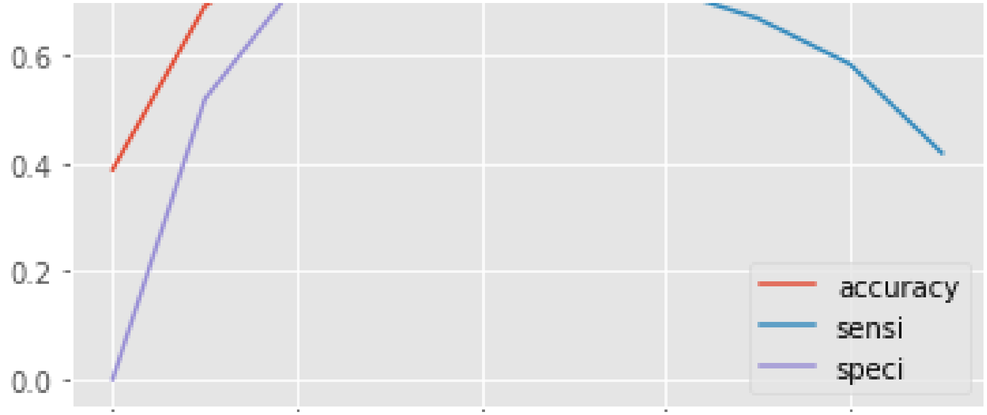
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.8416	0.093	-19.817	0.000	-2.024	-1.659
Last Notable Activity_Chat Interactions	-1.2434	0.332	-3.748	0.000	-1.894	-0.593
Last Notable Activity_Communication Interactions	1.5841	0.100	15.796	0.000	1.388	1.781
Last Notable Activity_Modified	-1.0429	0.094	-11.048	0.000	-1.228	-0.858
Lead Origin_Lead Add Form	4.1058	0.209	19.638	0.000	3.696	4.516
Page Views Per Visit	-2.9560	0.425	-6.963	0.000	-3.788	-2.124
Tags_Busy or unavailable	0.8053	0.223	3.616	0.000	0.369	1.242
Tags_Contacted or in touch with EINS	4.2117	0.376	11.191	0.000	3.474	4.949
Tags_Enrolled or already a student	-3.4624	0.593	-5.838	0.000	-4.625	-2.300
Tags_Interested or considering	1.7308	0.092	18.784	0.000	1.550	1.911
Tags_Not interested or unable to continue	-1.4564	0.113	-12.885	0.000	-1.678	-1.235
Total Time Spent on Website	3.9608	0.173	22.847	0.000	3.621	4.301
TotalVisits	1.8622	0.435	4.283	0.000	1.010	2.714
What is your current occupation_Working Professional	2.1835	0.210	10.375	0.000	1.771	2.596



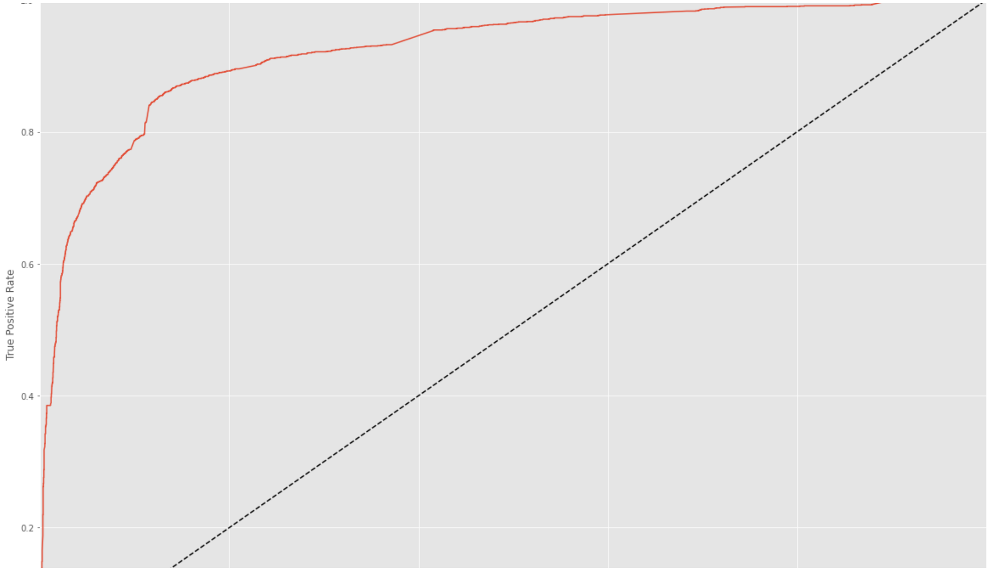
	features	VIF
0	Last Notable Activity_Chat Interactions	1.01
5	Tags_Busy or unavailable	1.07
6	Tags_Contacted or in touch with EINS	1.07
7	Tags_Enrolled or already a student	1.07
3	Lead Origin_Lead Add Form	1.24
12	What is your current occupation_Working Profes...	1.27
2	Last Notable Activity_Modified	1.44
1	Last Notable Activity_Communication Interactions	1.52
9	Tags_Not interested or unable to continue	1.54
8	Tags_Interested or considering	1.97
10	Total Time Spent on Website	2.12
11	TotalVisits	3.73
4	Page Views Per Visit	4.46

C. Model Training and Assessment – Cutoff Point & ROC Curve

From the curve above, 0.4 is the optimum point to take it as a cutoff probability.

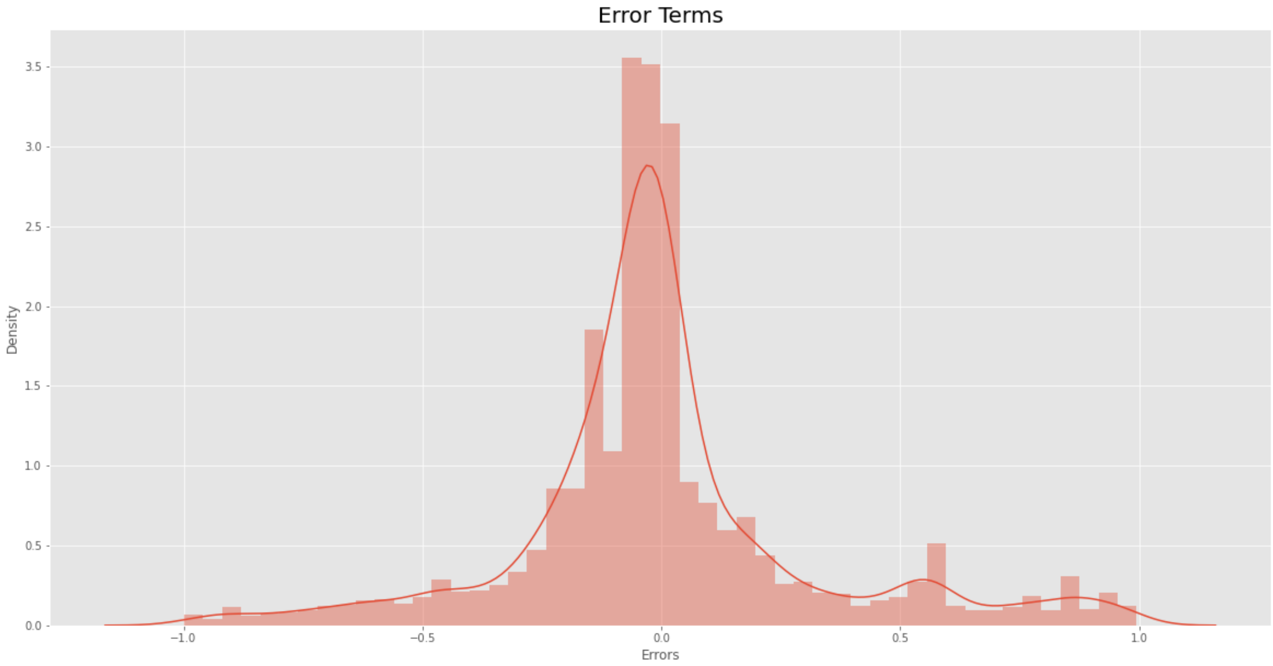


The ROC Curve should be a value close to 1. We are getting a good value of 0.97 indicating a good predictive model.

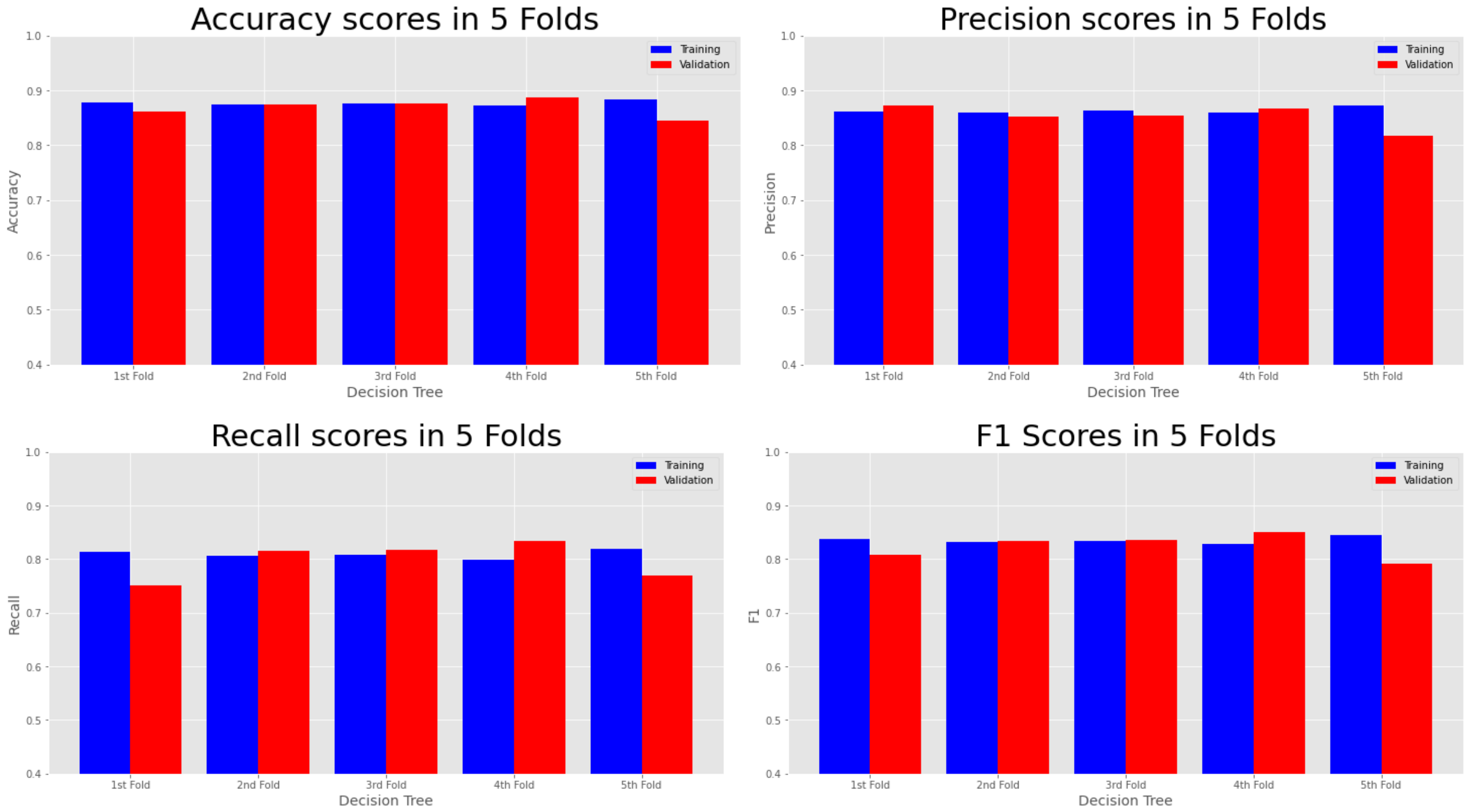


C. Model Training and Assessment – Error Residuals

Error Residual Following Normal Distribution - It indicates that this assumption is met and therefore the model inferences such as confidence intervals and predictions are valid.



C. Model Training and Assessment – 5-fold cross-validation



Accuracy scores in 5 Folds

Precision scores in 5 Folds

Recall scores in 5 Folds

F1 Scores in 5 Folds

Summary

The analysis was conducted using a Logistic model to predict the conversion rate for a company. After running the model on the test data, the following observations were made:

1. The accuracy for the train data was 83.29% while the sensitivity and specificity were 83.70% and 83.66%, respectively.
2. The accuracy for the test data was 84.78%, with sensitivity and specificity values of 83.98% and 85.26%, respectively.
3. To verify the accuracy of the model, 5-fold cross-validation was used to train the model across different combinations of data.
4. The model performed well in predicting the conversion rate, and the CEO can have confidence in making good decisions based on the model's predictions.

Thank you