Lead Scoring Case Study

Nishant Anand: nishantax2024@email.iimcal.ac.in
Nishant Gaurav: nishantgaurav@gmail.com
Parul Chopra: parul6588@yahoo.co.in
Group: DS71

Problem Statement

- X Education sells online courses to industry professionals. Leads come from various sources (Google, referrals, etc.) and engage through browsing, form submissions, and calls. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- Current lead conversion rate is low (~30%).
- Goal: Identify ‘Hot Leads’ to increase conversion efficiency.
- Target lead conversion rate: **80%**

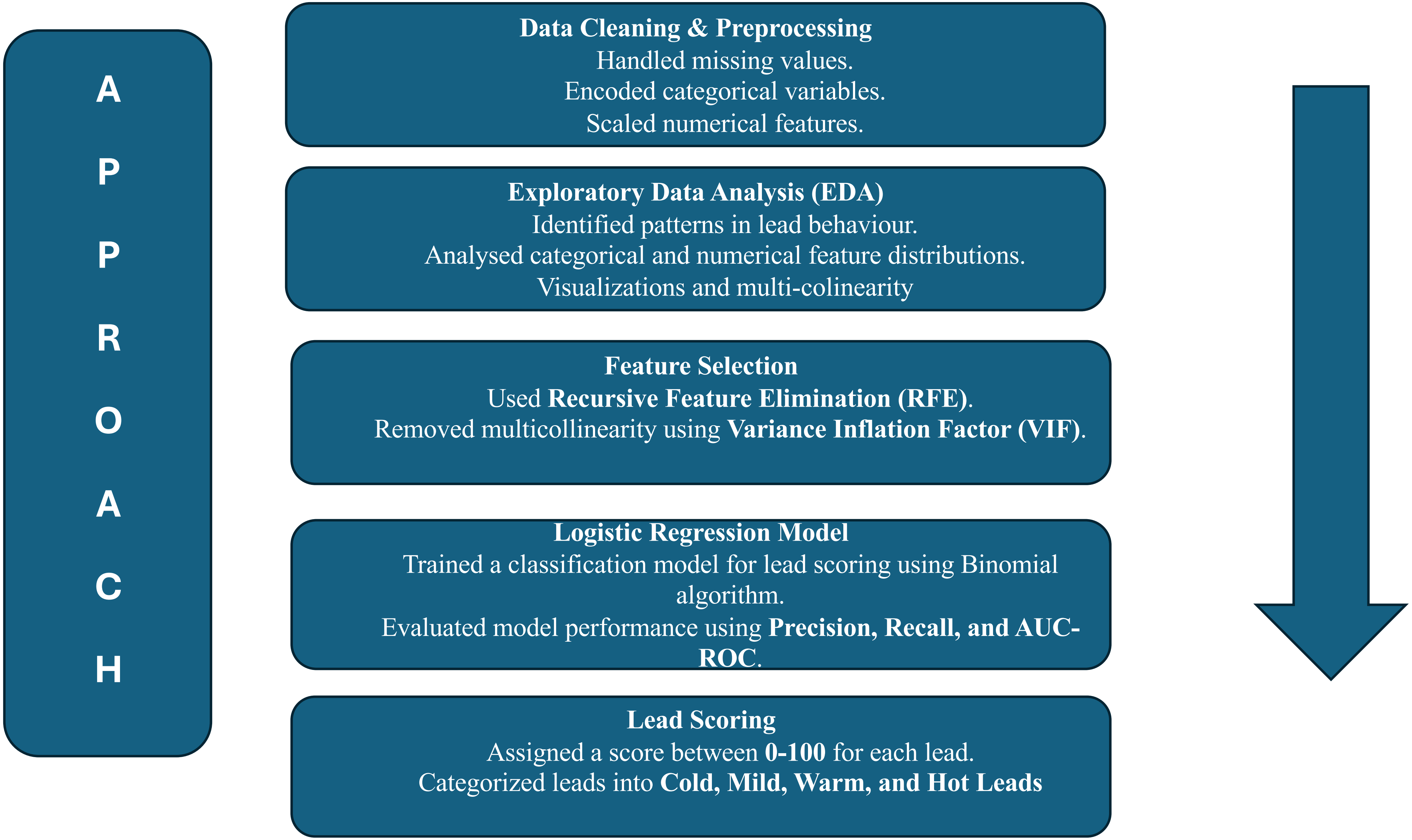
Business Understanding

- Leads come from multiple sources: **Google, Referrals, Social Media, etc.**
- Leads engage in various activities: **Browsing, Filling Forms, Watching Videos, etc.**
- Sales team follows up with leads via **calls and emails**.
- Challenge: **Prioritizing the right leads for higher conversions.**
- **Goal:** Identify **Hot Leads** to improve efficiency and reach an **80% conversion rate**.
- A logistic regression model is built to assign lead scores (0-100) for prioritization.

Social Media Marketing
Initial Pool of Leads
Lead Nurturing
Hot leads
£ € € $
Converted leads

Data Overview

- Dataset contains **~9000 leads**.
- Target variable: **Converted** (1 = Converted, 0 = Not Converted)
- Other Key variables:
  - **Lead Source** (Google, Referral, etc.)
  - **Lead Origin**
  - **Total Time Spent on Website**
  - **Total visits**
  - **Last Activity** (Form Submission, Call, etc.)
  - **What is your current occupation**
- Some Categorical variables include "Select" values that act as missing data, since user has not selected any option from dropdown menu of these variables. These are handled as NULL values.



APPROACH

Data Cleaning & Preprocessing

Handled missing values.

Encoded categorical variables.

Scaled numerical features.

Exploratory Data Analysis (EDA)

Identified patterns in lead behaviour.

Analysed categorical and numerical feature distributions.

Visualizations and multi-colinearity

Feature Selection

Used **Recursive Feature Elimination (RFE)**.

Removed multicollinearity using **Variance Inflation Factor (VIF)**.

Logistic Regression Model

Trained a classification model for lead scoring using Binomial algorithm.

Evaluated model performance using **Precision, Recall, and AUC-ROC**.

Lead Scoring

Assigned a score between **0-100** for each lead.

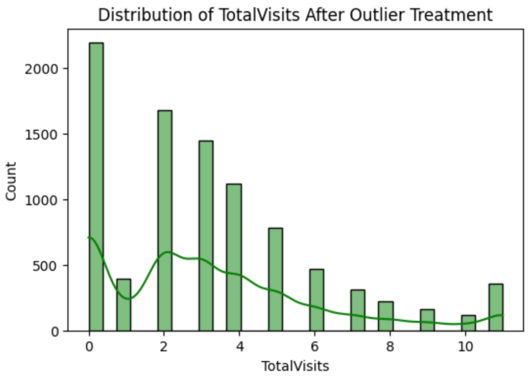
Categorized leads into **Cold, Mild, Warm, and Hot Leads**

Data Cleaning & Preprocessing

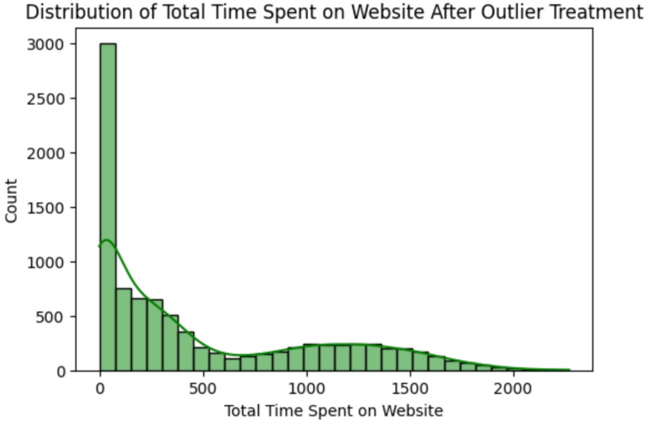
- Drop "Prospect ID" and "Lead Number" as these are the ID cols
- Check for missing values
- Replace 'Select' to NaN in cols "Specialization", "City", "How did you hear about X Education" and "Lead Profile"
- Drop cols with majority of the values as NULL (> 30%)
- Identify the categorical and numerical cols
- Replace missing values with median in case of Numerical columns and Mode in case of Categorical Columns

Exploratory Data Analysis

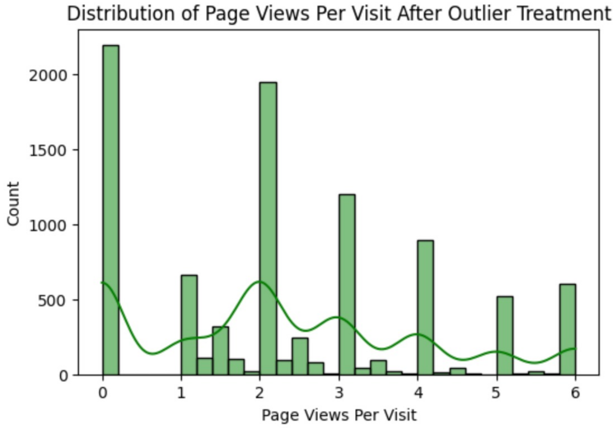
Numeric Variables after outlier treatment



Distribution of TotalVisits After Outlier Treatment



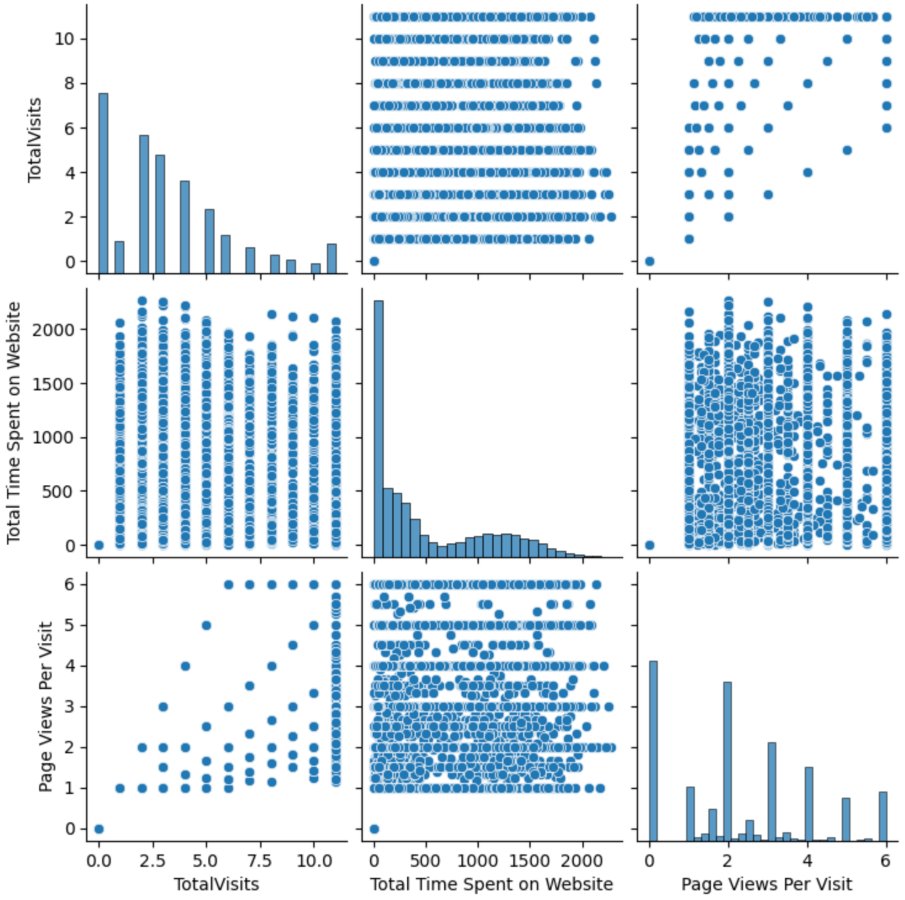
Distribution of Total Time Spent on Website After Outlier Treatment



Distribution of Page Views Per Visit After Outlier Treatment

Univariate Analysis

Bivariate Analysis

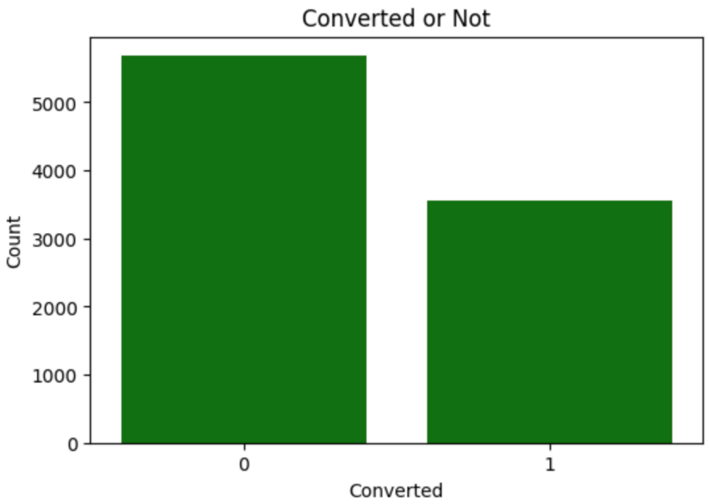


Total Visits vs. Total Time Spent: No strong correlation—some leads with many visits still spend little time.

Page Views vs. Total Visits: Some clustering, indicating that leads who revisit also engage with multiple pages.

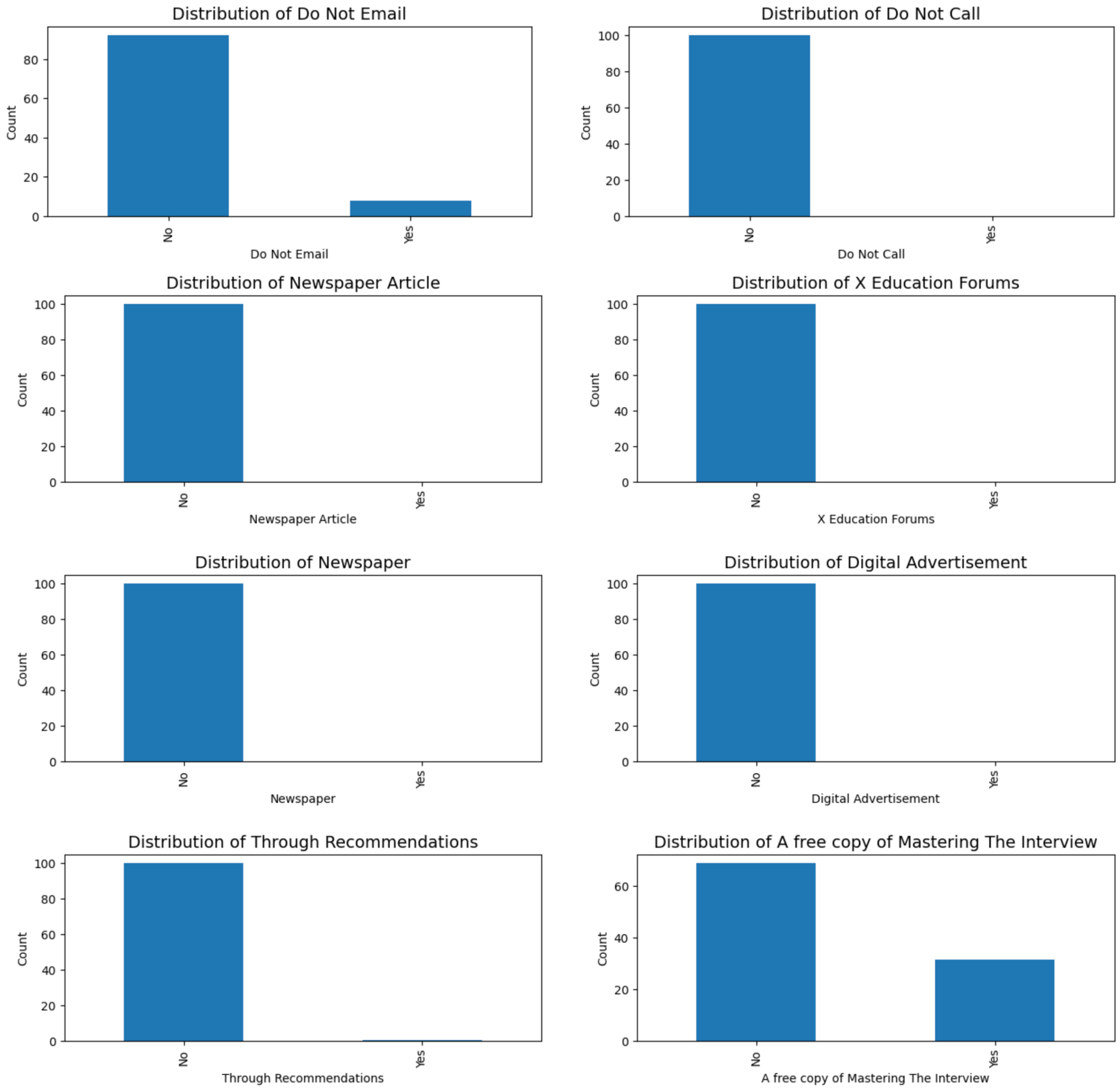
Categorical variables with Binary values

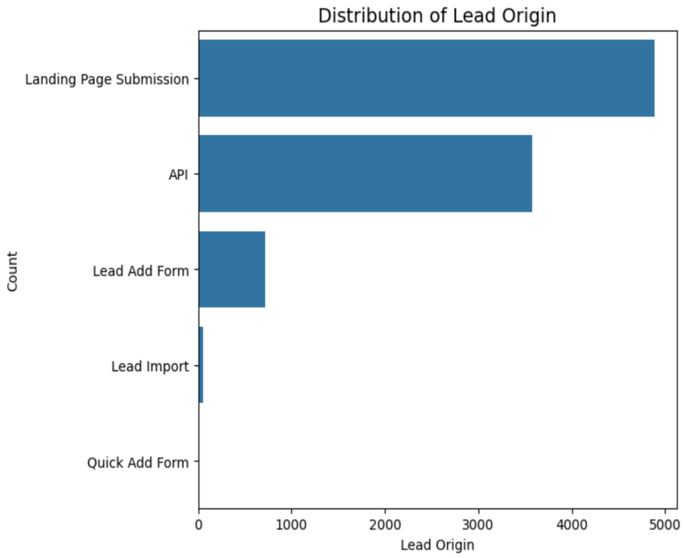
Target Variable

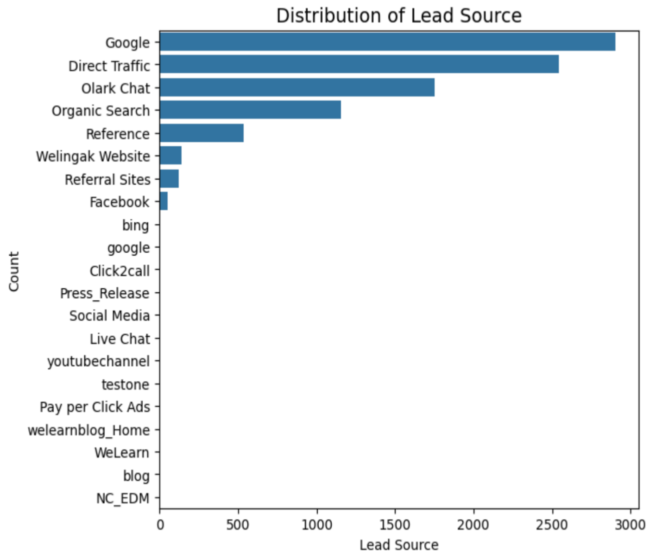


A significant number of leads **did not convert**, indicating room for optimization in the **sales funnel**.

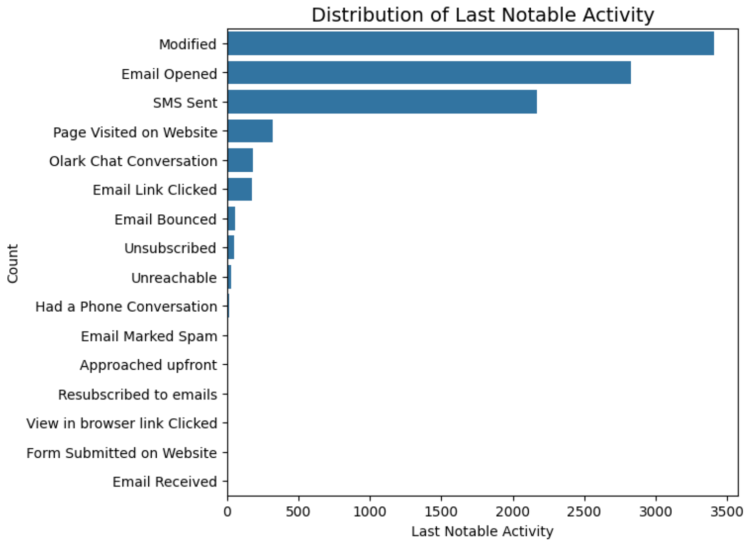
Most leads have "No" for these categories, meaning very few leads **engage via newspaper ads, recommendations, or digital ads**.

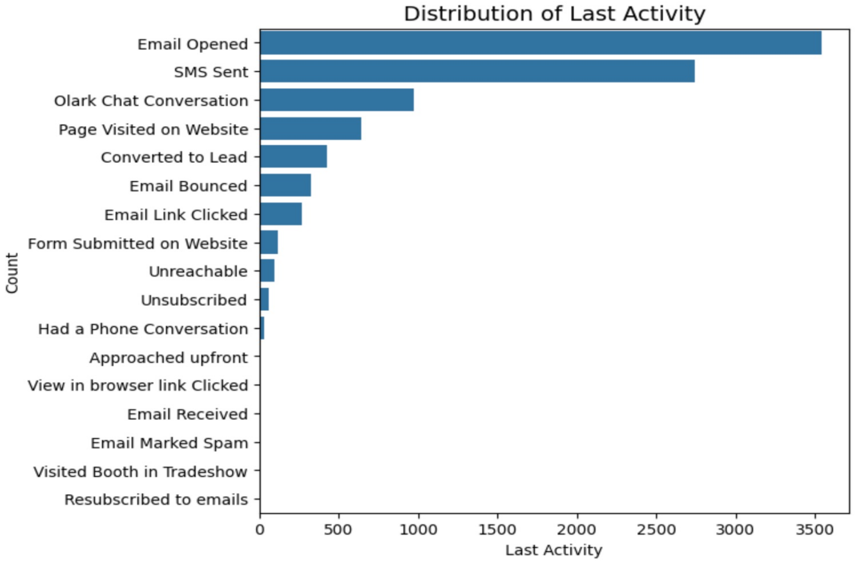


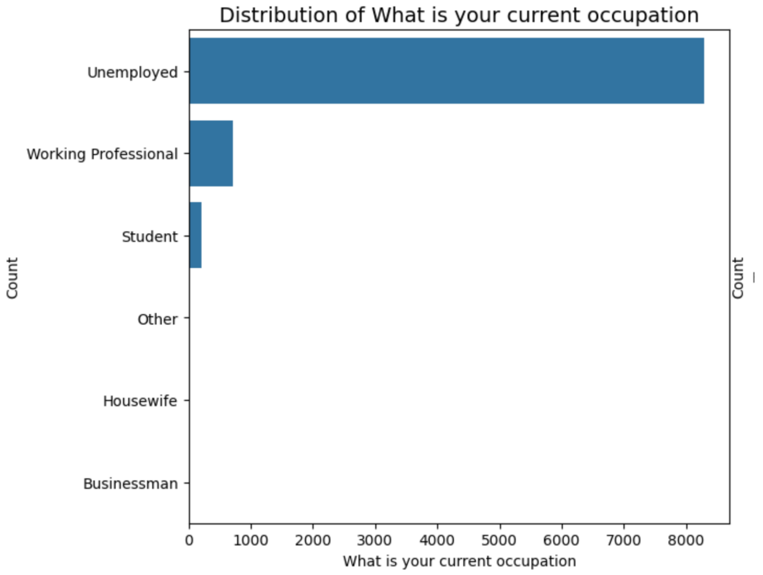

Distribution of Lead Origin

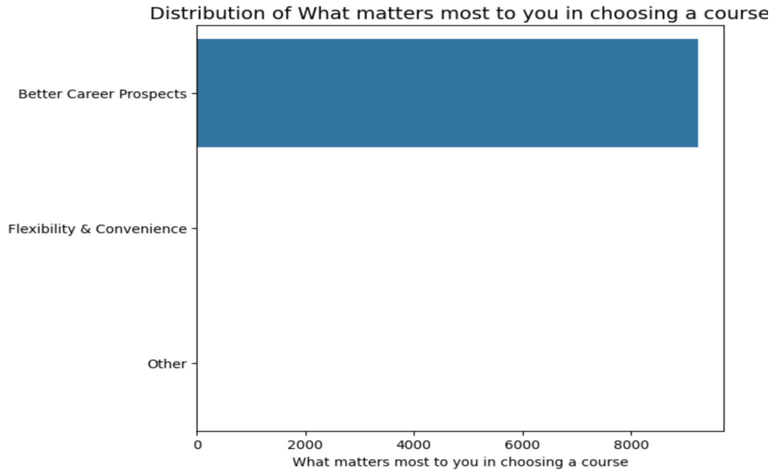

Distribution of Lead Source

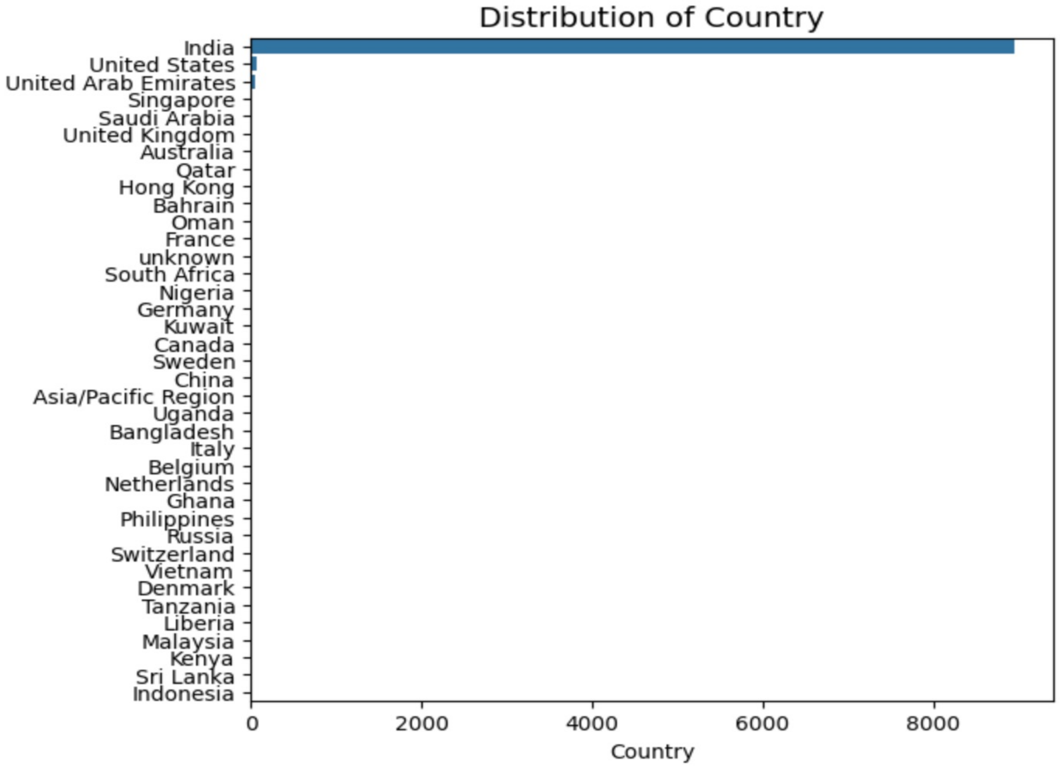
Most leads come from **Landing Page Submissions, Google and Lead Add Form. Email opening & SMS responses** are the most common last notable activities.


Distribution of Last Notable Activity

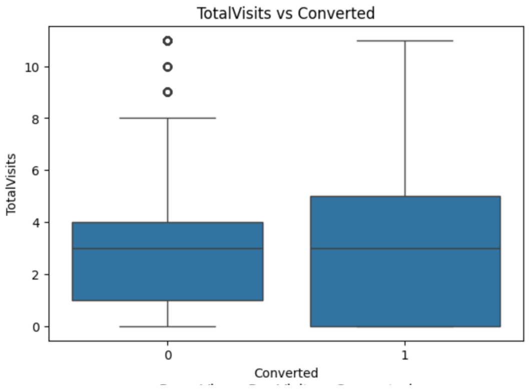

Distribution of Last Activity

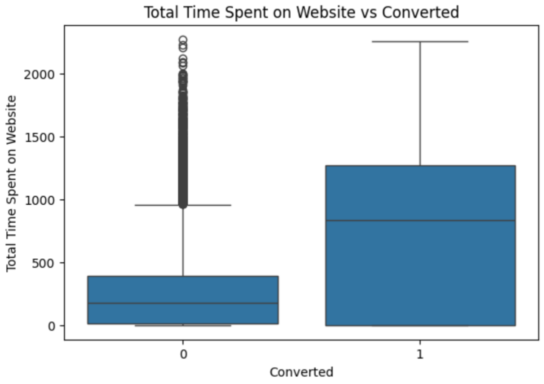


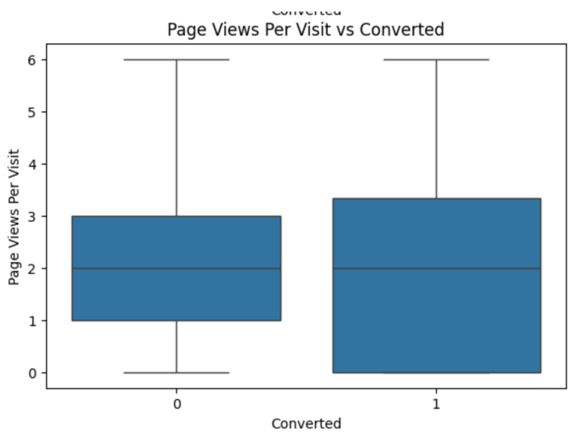




It can be seen that **unemployed as well as some working professionals from India dominate**, primarily choosing courses for **better career prospects.**

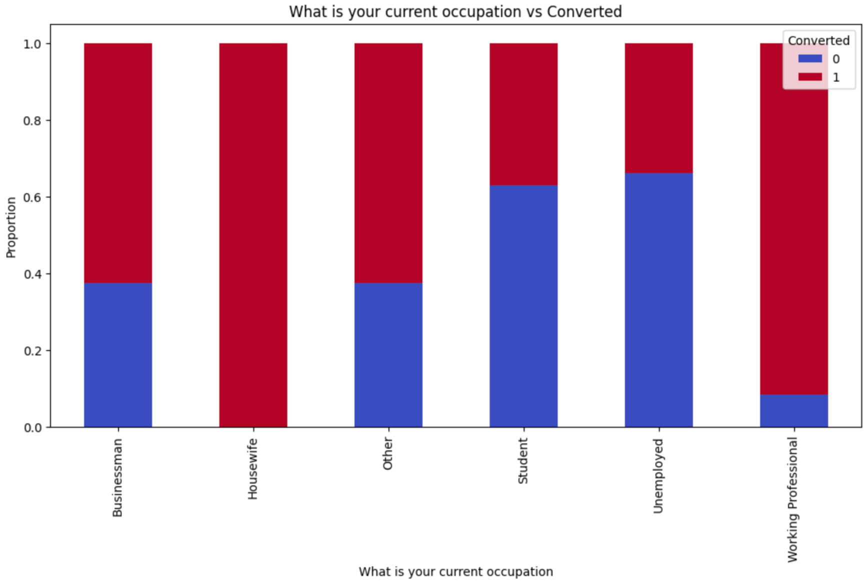


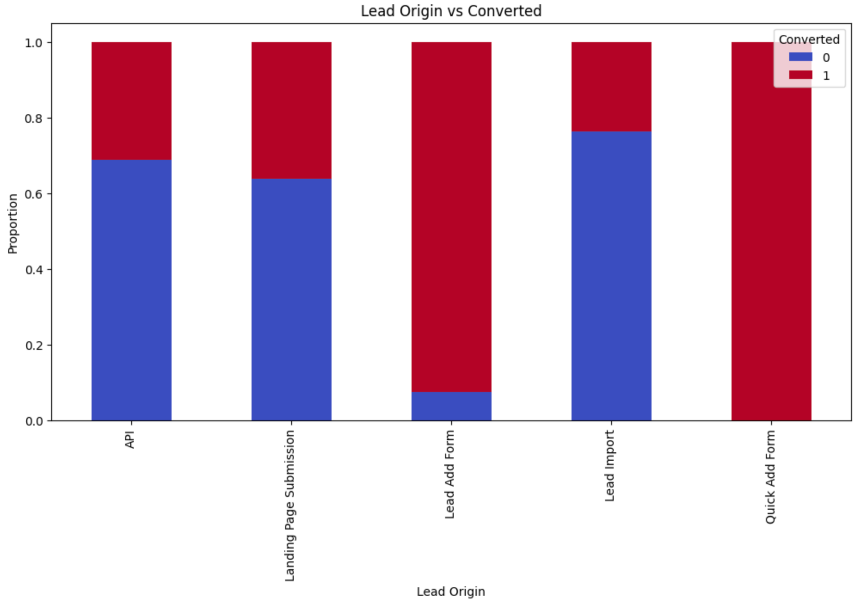




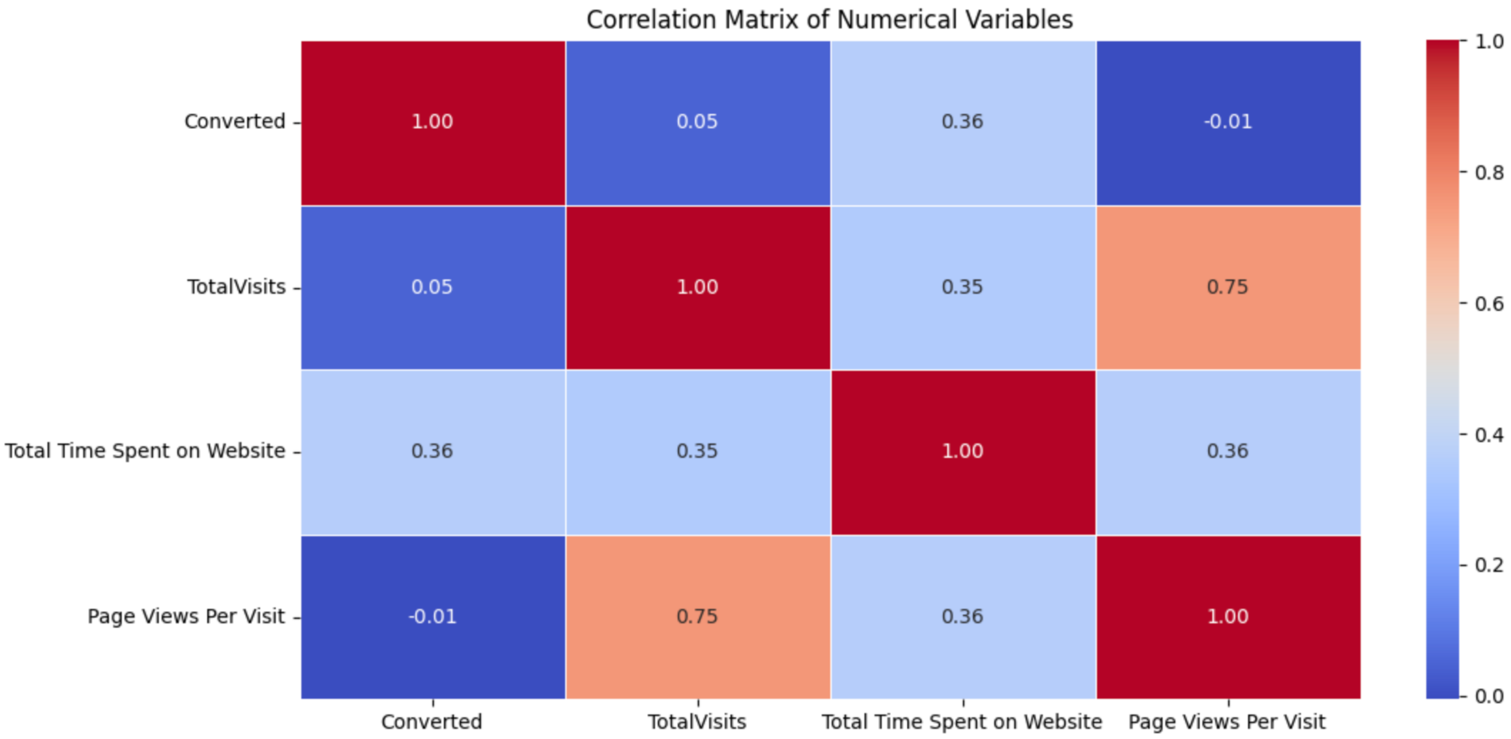
Bivariate Analysis

Leads who **spend more time on the website** and have **higher page views per visit** show a greater likelihood of conversion. **Students and working professionals** convert more frequently, while **Lead Add Form origins** have higher conversion rates than other lead sources.





Multivariate analysis



Total Time Spent on Website has the strongest positive correlation (**0.36**) with conversion. It indicates higher engagement and increases the likelihood of conversion.

Page Views Per Visit and Total Visits are highly correlated with each other(0.75) but are not directly having any impact on conversion.

Further Steps for Data Preparation

- For categorical variables with multiple levels, dummy features (one-hot encoded) were created

- Test-Train Split done
  - Putting response variable to y ‘Converted’
  - Putting feature variable to X, dropping target variable ‘Converted’ from X
  - Splitting the data into TRAIN (70%) and TEST (30%)

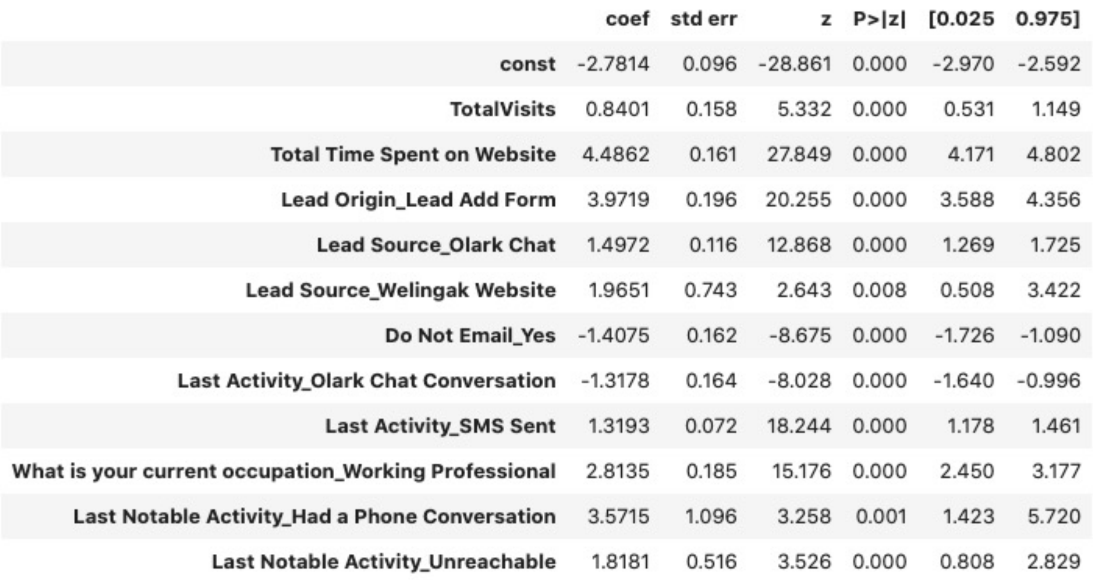
- Feature Scaling: Minmax scaling for numerical variables

Model Building

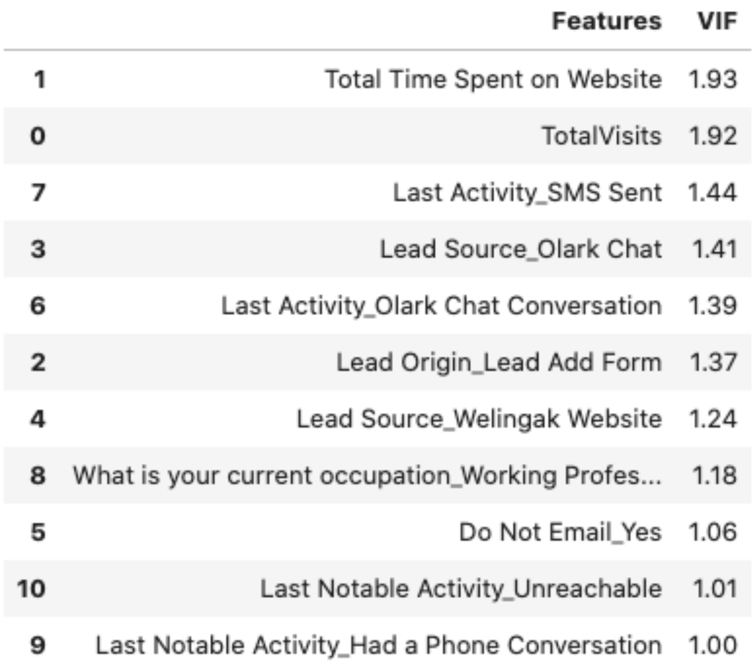
- Model Building: Models built by feature Selection Using RFE
    - 15 features selected;
    - Generalized Linear Model

- Recursive Feature Elimination and Variance Inflation Factor (variables with VIF >5 dropped) helped choosing most powerful and statistically significant (p-value<0.05) features.

Final p-values < 0.05 and VIFs are < 5. Model trained



	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7814	0.096	-28.861	0.000	-2.970	-2.592
TotalVisits	0.8401	0.158	5.332	0.000	0.531	1.149
Total Time Spent on Website	4.4862	0.161	27.849	0.000	4.171	4.802
Lead Origin_Lead Add Form	3.9719	0.196	20.255	0.000	3.588	4.356
Lead Source_Olark Chat	1.4972	0.116	12.868	0.000	1.269	1.725
Lead Source_Welingak Website	1.9651	0.743	2.643	0.008	0.508	3.422
Do Not Email_Yes	-1.4075	0.162	-8.675	0.000	-1.726	-1.090
Last Activity_Olark Chat Conversation	-1.3178	0.164	-8.028	0.000	-1.640	-0.996
Last Activity_SMS Sent	1.3193	0.072	18.244	0.000	1.178	1.461
What is your current occupation_Working Professional	2.8135	0.185	15.176	0.000	2.450	3.177
Last Notable Activity_Had a Phone Conversation	3.5715	1.096	3.258	0.001	1.423	5.720
Last Notable Activity_Unreachable	1.8181	0.516	3.526	0.000	0.808	2.829



	Features	VIF
1	Total Time Spent on Website	1.93
0	TotalVisits	1.92
7	Last Activity_SMS Sent	1.44
3	Lead Source_Olark Chat	1.41
6	Last Activity_Olark Chat Conversation	1.39
2	Lead Origin_Lead Add Form	1.37
4	Lead Source_Welingak Website	1.24
8	What is your current occupation_Working Profes...	1.18
5	Do Not Email_Yes	1.06
10	Last Notable Activity_Unreachable	1.01
9	Last Notable Activity_Had a Phone Conversation	1.00

New column created: 'predicted' with 1 if Converted_Prob > 0.5 else 0

Top Features

The top three variables

	Features	Coefficient
1	Total Time Spent on Website	4.4862
2	Lead Origin_Lead Add Form	3.9719
3	Last Notable Activity_Had a Phone Conversation	3.5715

The top three categorical/dummy variables

	Features	Coefficient
1	Lead Origin_Lead Add Form	3.9719
2	Last Notable Activity_Had a Phone Conversation	3.5715
3	What is your current occupation_Working Professional	2.8135

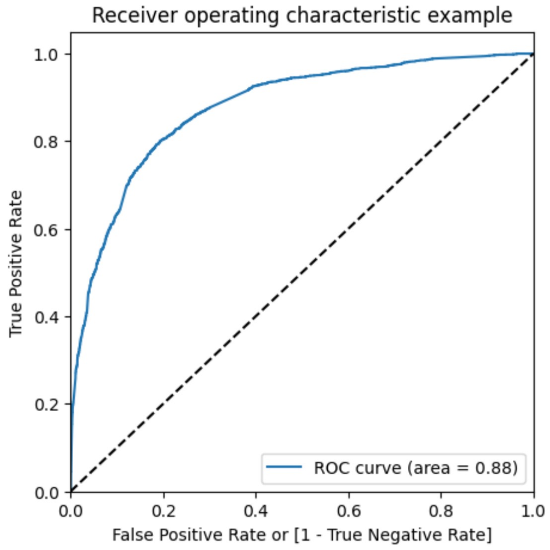
P-value < 0.001 for all the above variables

Confusion matrix

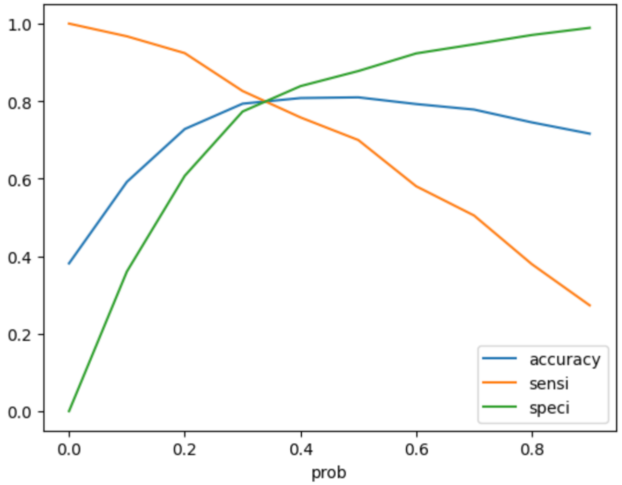
Predicted	Converted	
	Yes	No
Yes	3512	490
No	741	1725

Attribute	Percentage
Sensitivity	69.95
Specificity	87.75
Positive predictive value	77.87
Negative predictive value	82.57

The model has **high specificity (87.75%)**, meaning it accurately identifies non-converting leads, but **moderate sensitivity (69.95%)**, indicating some converting leads may be misclassified.
The **ROC-AUC score of 0.88** implies overall good model performance.



Finding Optimal Cutoff Point



Attribute	Percentage
Sensitivity / Recall	81.02
Specificity	79.13
False postive rate	20.86
Positive predictive value / Precision	70.525
Negative predictive value	87.12

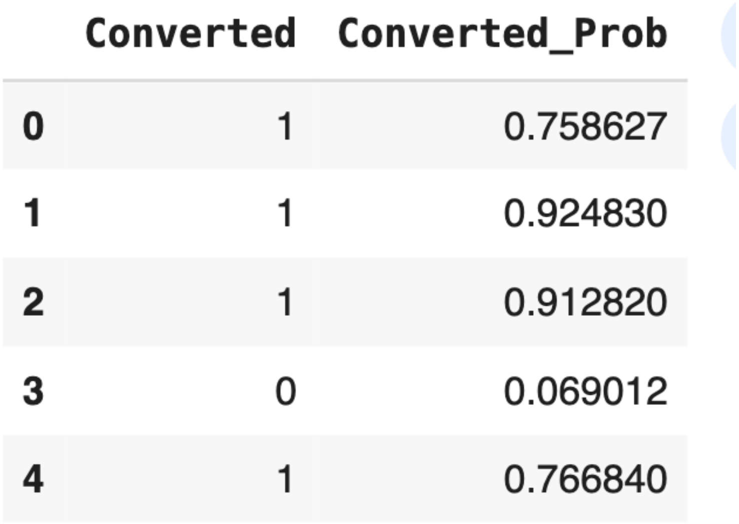
From the curve above, 0.33 is the optimum point to take it as a cutoff probability.

At this cutoff probability **0.33**, it helps in balancing **sensitivity (81.02%) and specificity (79.13%)** to improve lead conversion predictions.

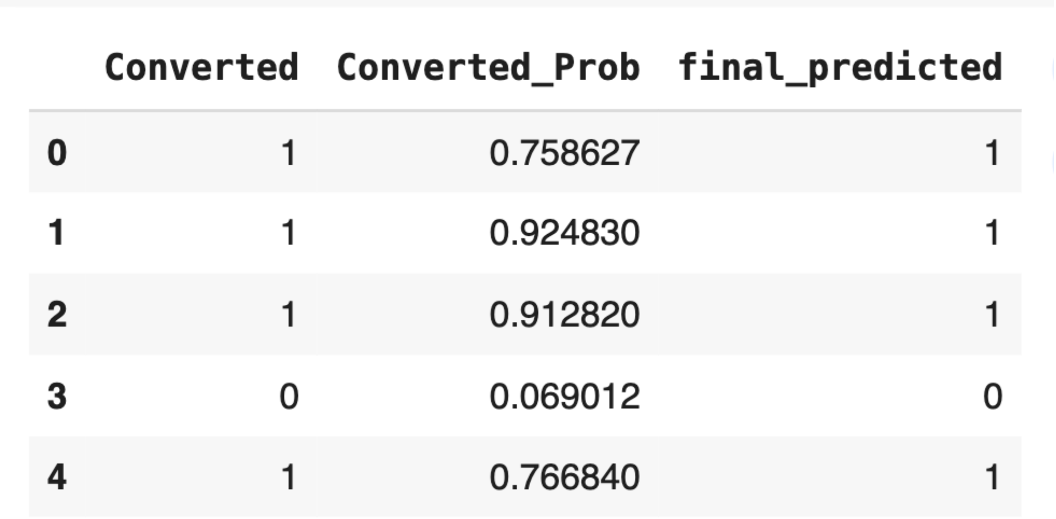
Making predictions on the test set

Appending y_test_df and y_pred_1

y_pred_final['final_predicted'] = y_pred_final.Converted_Prob.map(lambda x: 1 if x > 0.33 else 0): Overall accuracy: 80.33



	Converted	Converted_Prob
0	1	0.758627
1	1	0.924830
2	1	0.912820
3	0	0.069012
4	1	0.766840



	Converted	Converted_Prob	final_predicted
0	1	0.758627	1
1	1	0.924830	1
2	1	0.912820	1
3	0	0.069012	0
4	1	0.766840	1

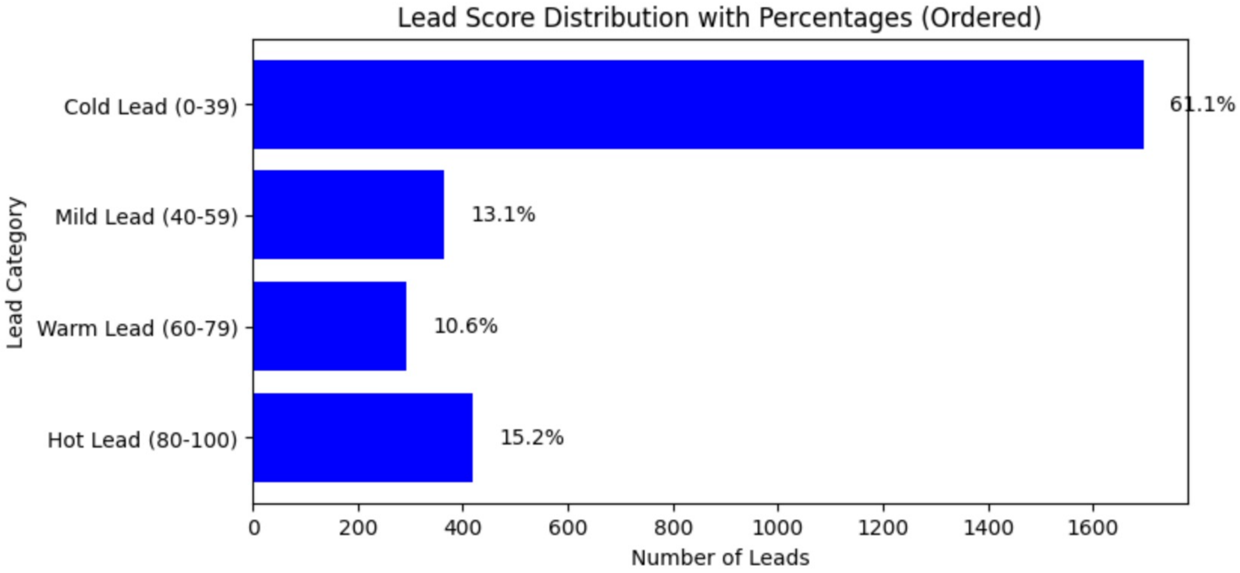
Using a **0.33 probability threshold** leads are classified. Higher **Converted_Prob values (>0.33) correspond to higher conversion likelihood**, effectively distinguishing between potential and non-converting leads.

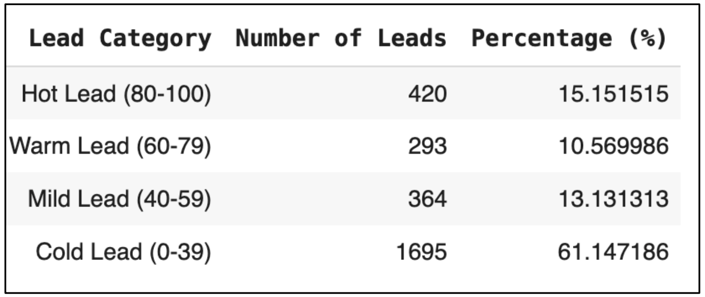
Test data set

Attribute	Percentage
Sensitivity / Recall	80.45
Specificity	80.26
False postive rate	19.73
Positive predictive value / Precision	72.68
Negative predictive value	86.28

On the test data set, the model achieves **80.45% recall**, effectively identifying most actual conversions, and **72.68% precision**, implying that about **73% of predicted conversions are correct**.
With **86.28% negative predictive value**, it is reliably able to filter out non-converting leads.

Lead Score Distribution & Categorization





Lead Category	Number of Leads	Percentage (%)
Hot Lead (80-100)	420	15.151515
Warm Lead (60-79)	293	10.569986
Mild Lead (40-59)	364	13.131313
Cold Lead (0-39)	1695	61.147186

Maximum number of leads (61.1%) fall into the **Cold Lead category (0-39 score)**, while only **15.2% are Hot Leads (80-100 score)**.
If **Hot and Warm leads** are prioritized, it can improve conversion efficiency and reduce wasted efforts on low-probability leads.

Business Recommendations

Hot & Warm Leads given top priority for Sales Efforts

Focus should be made on direct sales calls and personalized follow-ups on **Hot (80-100) and Warm (60-79) leads**, as they have the highest potential to convert.

Contact to Cold Lead (0-39) may be automated through **email campaigns and retargeting ads** instead of sales calls to save time and resources.

Enhance Website Engagement to Boost Conversion of Leads

Since **Total Time Spent on Website strongly correlates with conversion**, improve website **landing page user experience, add chatbots with immediate resolution of queries, and personalize content** to increase engagement

Optimize Ad Spend Based on High-Performing Lead Sources

Google and Direct Traffic bring the most leads. More budget to be allocated for marketing here.
Reduce investment in low-impact sources like newspaper ads and social media, which have minimal conversions.

Refine Lead Scoring Model & Thresholds Dynamically

The **0.33 cutoff improves recall and lead identification**, but periodic **re-evaluation is needed** based on changing lead behaviours and needs of the company.
Frequently track performance metrics (conversion rate, precision, recall) and adjust thresholds accordingly.

Improve Lead Nurturing via Email & SMS Automation

Most leads **open emails or receive SMS**, but engagement drops afterward. Implement **drip email campaigns, personalized recommendations, and time-sensitive offers, discounts, loyalty benefits to those who have already taken up courses with X Education** and keep leads engaged.
Create targeted courses and promotions for **working professionals**, as they show a higher likelihood of conversion.

Conclusion

- **Lead scoring model** enhances the efficiency of sales by recognizing and prioritizing top-converting leads, maximizing marketing and sales efforts. Business can **optimize efforts** based on model insights.
- Future work: The use of AI-driven recommendations for **lead nurturing** and more optimized lead engagement techniques will continue to drive conversion rates and growth for the business.

Thank You