Lead Score Case Study

- **Presented by -**
  - Jaymeen Jethva

Index

- Problem Statement
- Solving Approach
- Data Cleaning
- Data Visualisation
- Data Preprocessing
- Model Building and Evaluation
- Model Performance
- Conclusion
- Recommendations

Problem Statement

- X Education, an online course provider for industry professionals, is facing a challenge in improving its lead conversion rate, which currently stands at around 30%.
- The company collects leads through various channels, including their website and referrals, and aims to identify the most promising leads—referred to as "Hot Leads"—to improve conversion rates.
- The company seeks assistance in building a predictive model that can assign a lead score to each lead.
- The higher the score, the more likely the lead will convert into a paying customer. The goal is to increase the lead conversion rate to approximately 80%, enabling the sales team to focus their efforts on leads that are most likely to convert.

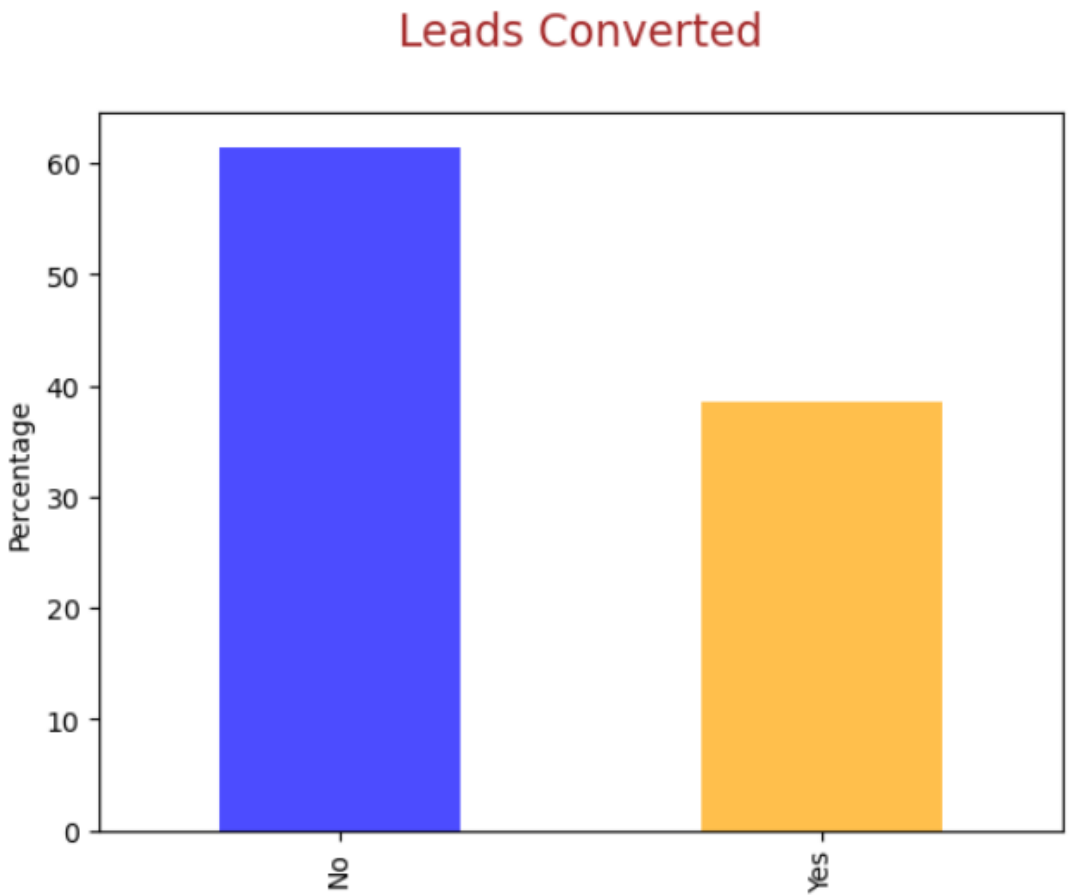
Solving Approach

Data Cleaning

- Columns with 'Select' as value is converted to 'np.nan' value.
- Columns with over 35% missing values are Dropped.
- Missing values in categorical columns were handled based on value counts and considerations.
- Imputation was applied to fill missing values in some categorical variables using mode for categorical variables having less percentage of missing value.
- Columns with no modeling use (Prospect ID, Lead Number) or only one category were dropped.
- Numerical data was imputed using mode and outliers in ‘TotalVisits' and ‘PageViews PerVisit’ were capped.
- Data standardization included fixing invalid values and standardizing casing (e.g., "Google" vs. "google").
- Low-frequency values in categorical variables were grouped into "Others" where applicable.

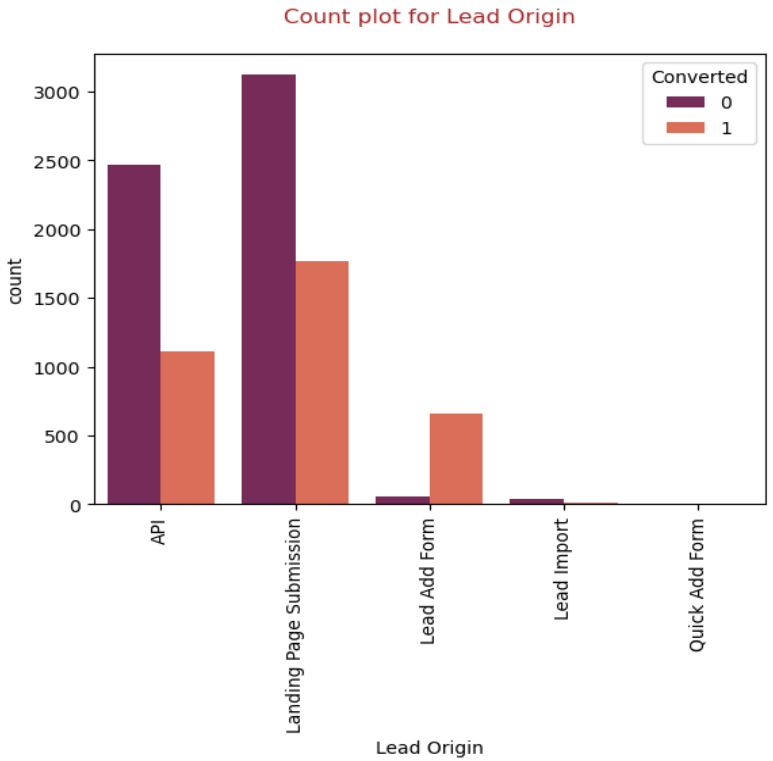
Data Visualisation

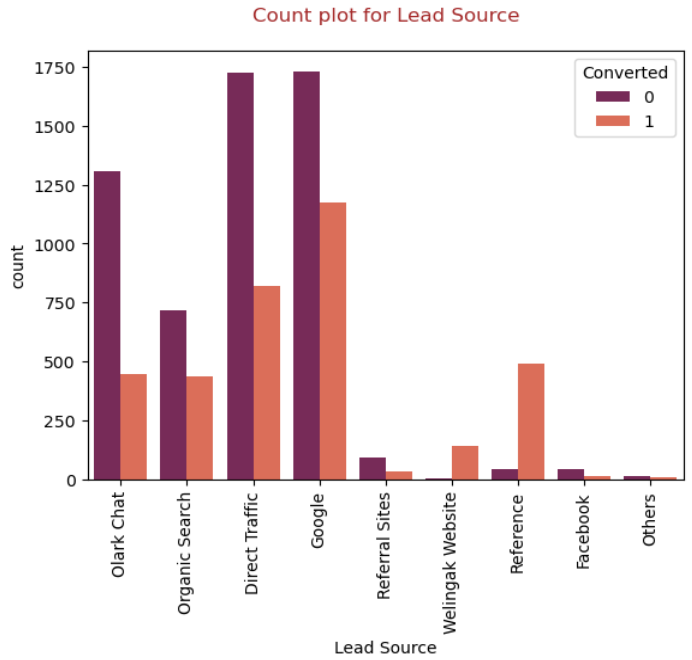
- Data is imbalanced while checking the Target variable
- Conversion rate is of 38.53%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.47% of the people didn't convert to leads. (Majority)



Segmented Univariate Analysis

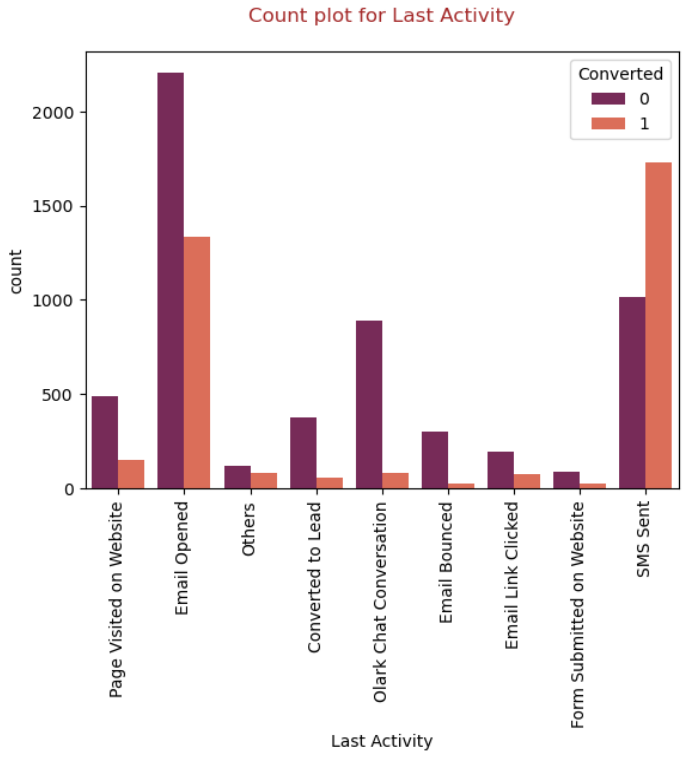
- **Insights:**
- Chances of getting converted are higher for customers who has origin from 'Lead Add Form' and origin for Majority of people are 'Landing Page Submission' and 'API'
- Majority of lead source are 'Google' and 'Direct Traffic'. Chances of getting converted are higher for 'Reference'

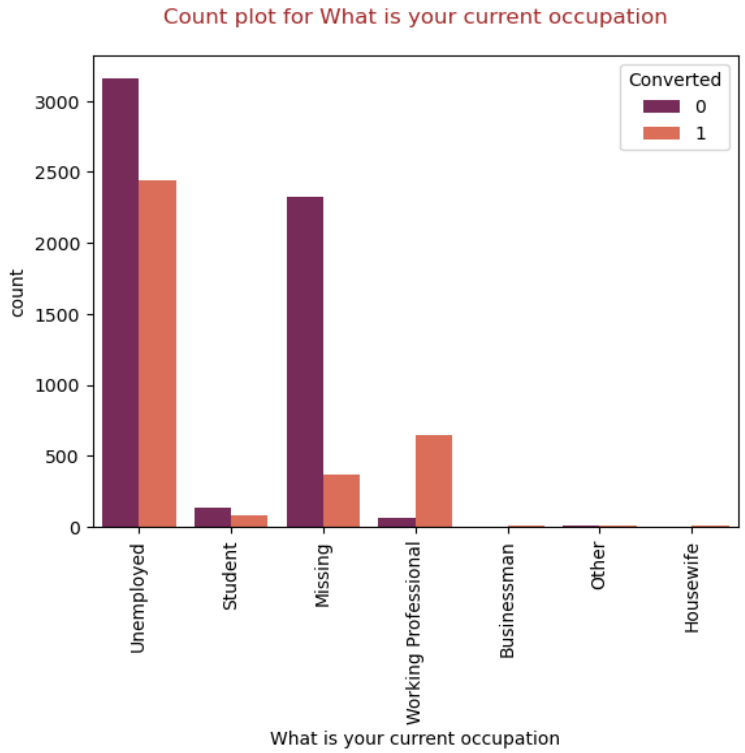




Segmented Univariate Analysis

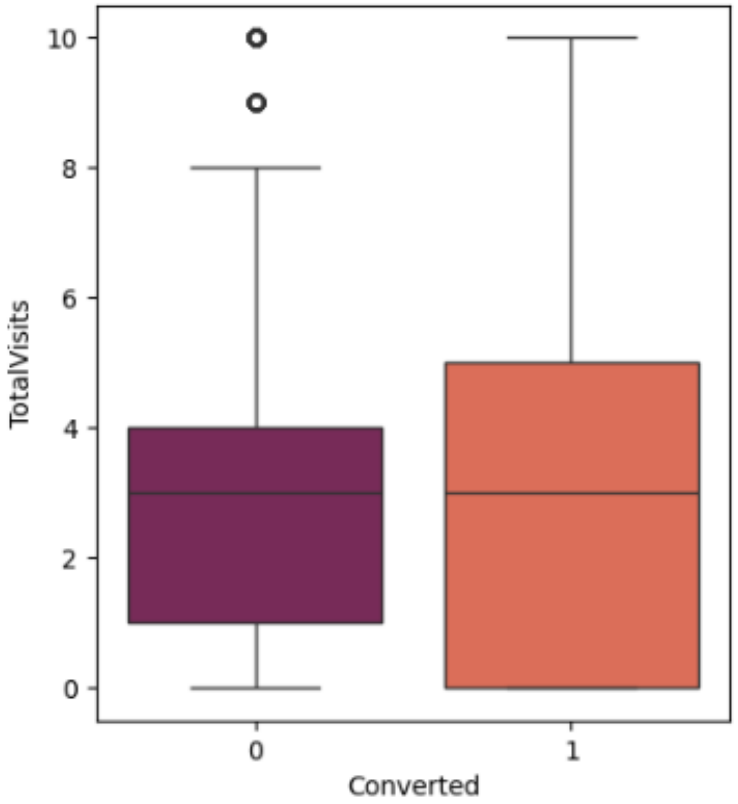
- **Insights:**
- Majority of people are with last activity as 'Email opened' but chances of getting converted are higher for people with last activity as 'SMS sent'
- Majority of people are Unemployed but chances of getting converted for working professionals are higher

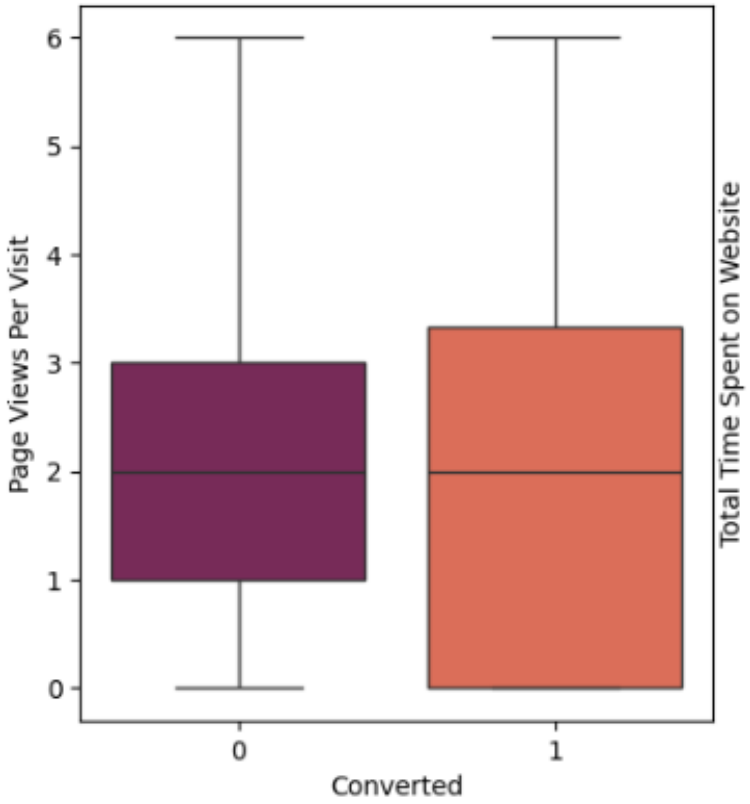


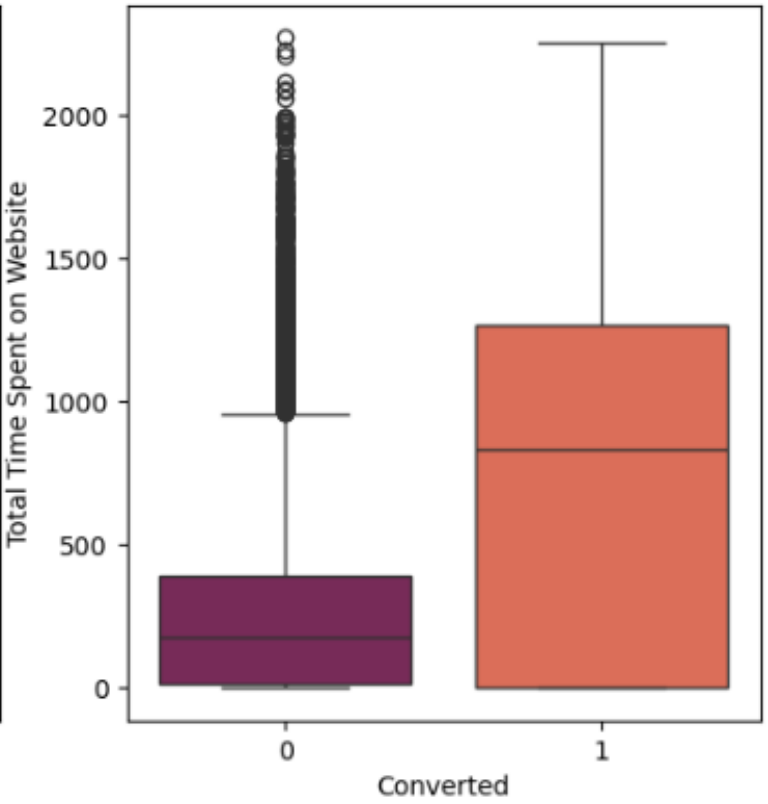


Segmented Univariate Analysis for Numerical feature

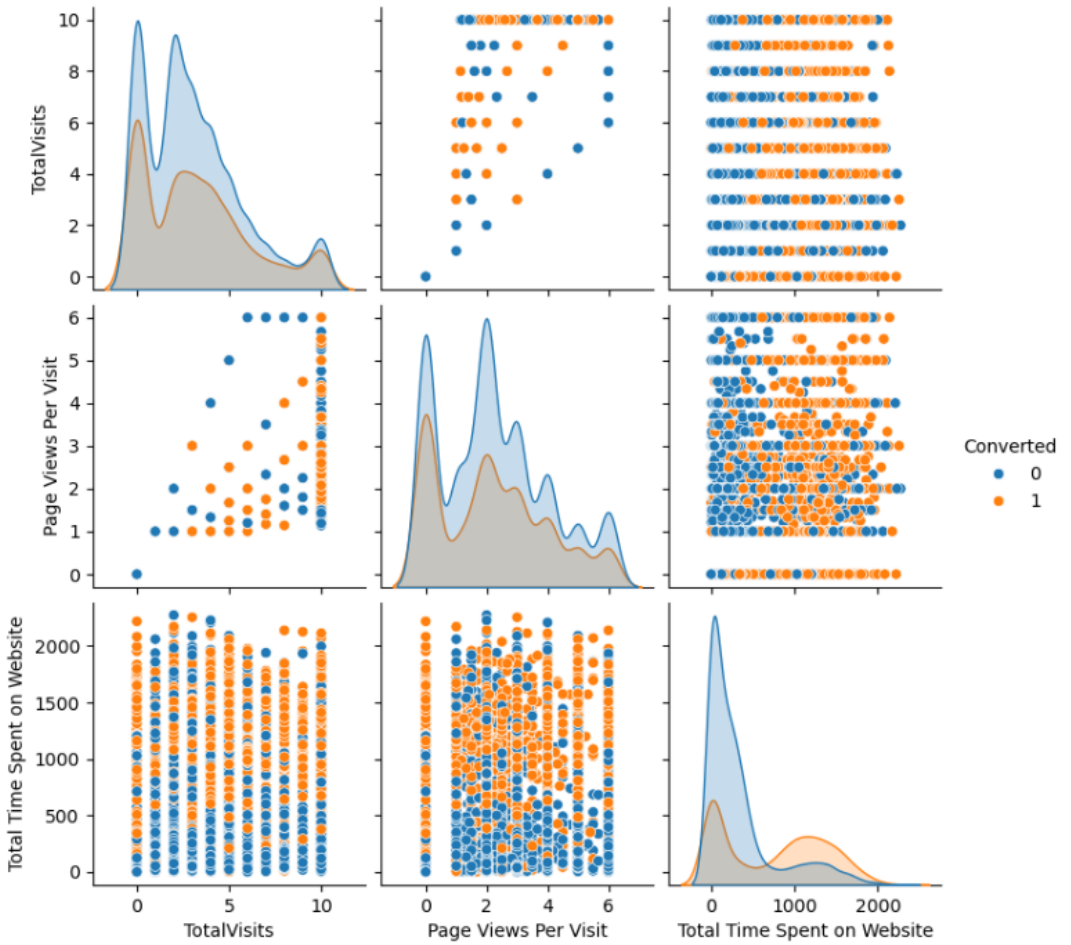
- **Insights:**
- People who spend more time on website they have higher chances of getting converted

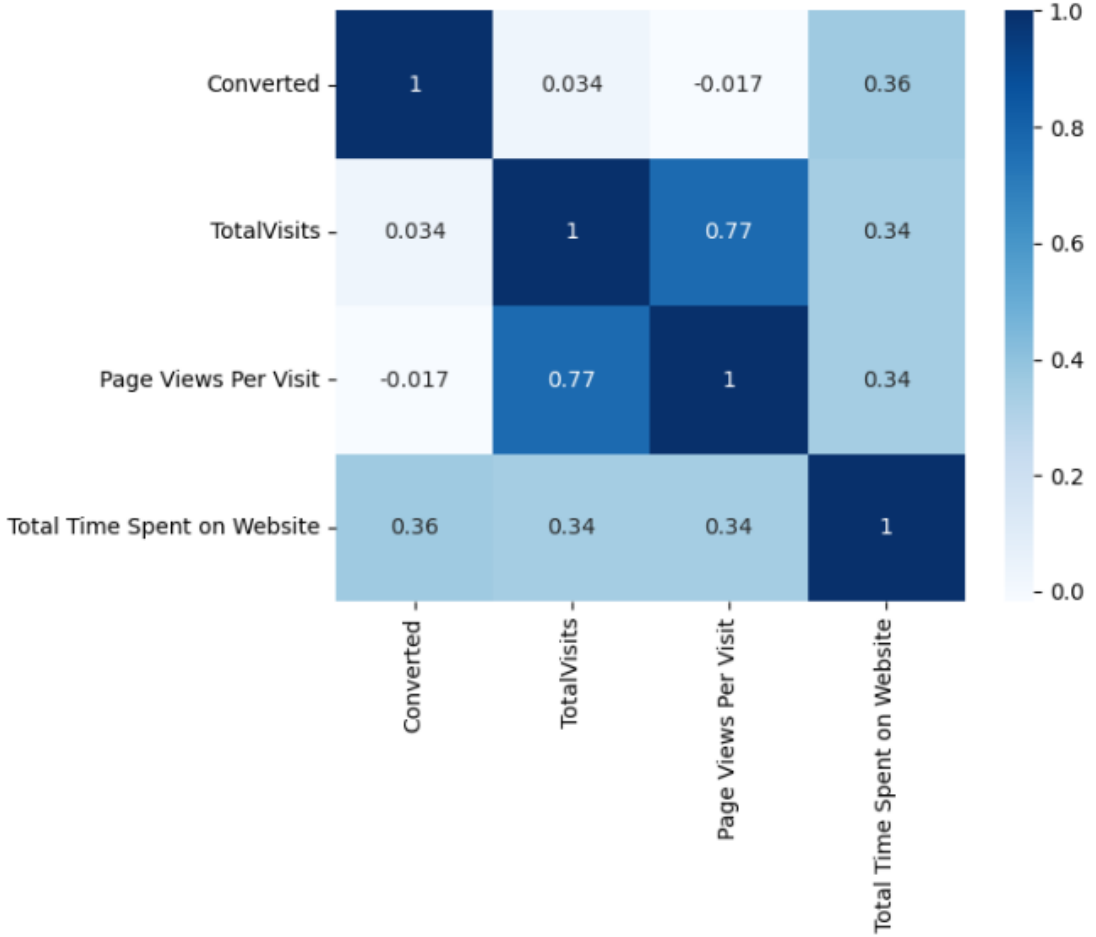




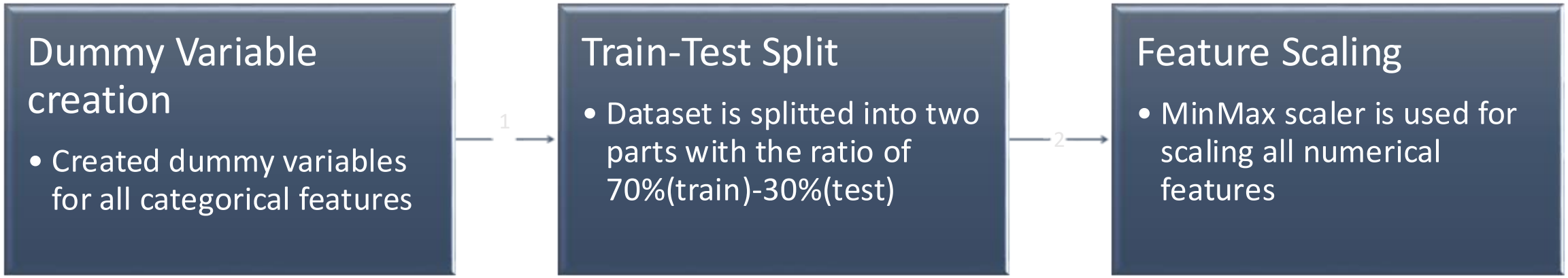


Bivariate Analysis & Correlation Matrix





Data Preprocessing



Model Building and Evaluation

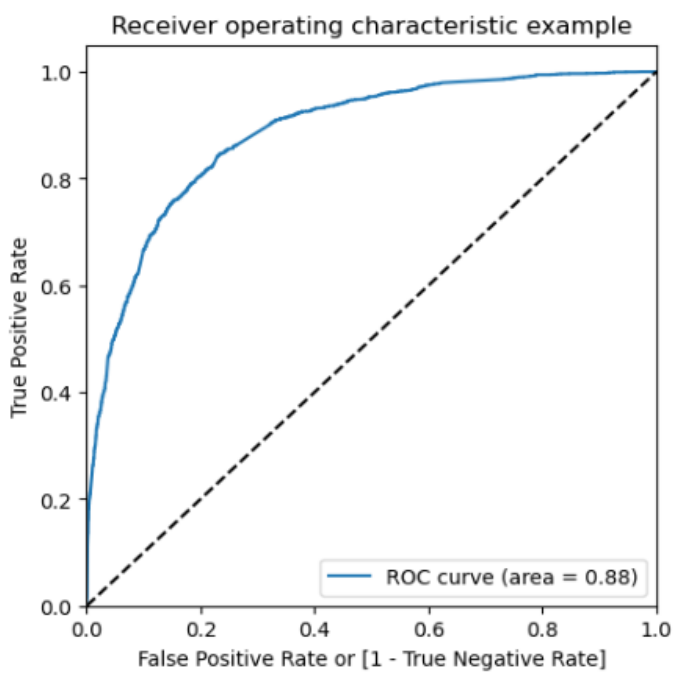
- Developed a Logistic Regression model using all features as the initial input. Subsequently, applied Recursive Feature Elimination (RFE) and a manual elimination approach to identify and remove less significant features, enhancing model performance and interpretability

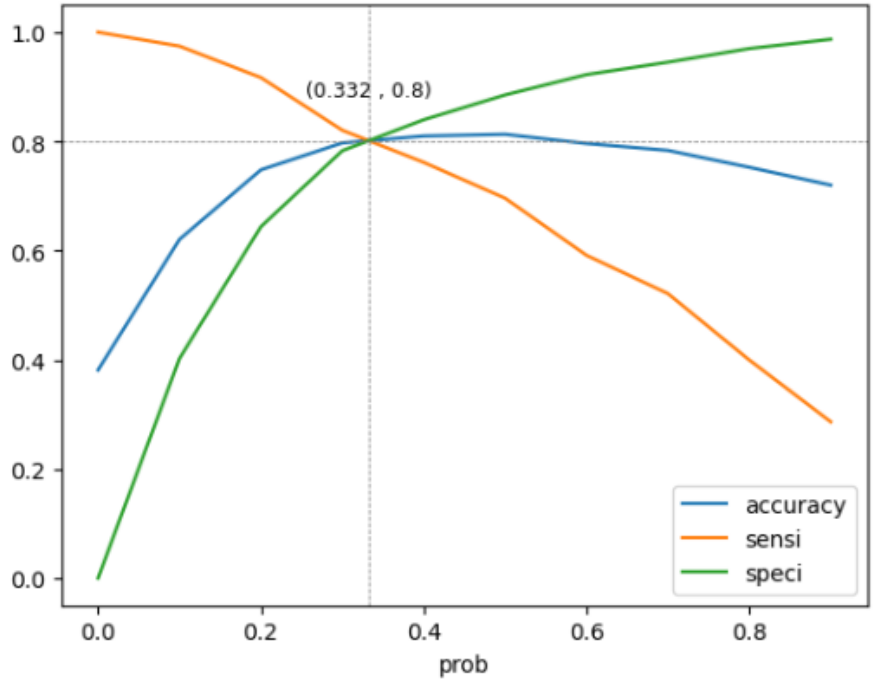
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6454
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2686.7
Date:	Mon, 20 Jan 2025	Deviance:	5373.4
Time:	22:27:51	Pearson chi2:	6.94e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3926
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.7684	0.122	-30.789	0.000	-4.008	-3.529
TotalVisits	0.8734	0.153	5.694	0.000	0.573	1.174
Total Time Spent on Website	4.5204	0.165	27.472	0.000	4.198	4.843
What is your current occupation_Student	1.2276	0.235	5.214	0.000	0.766	1.689
What is your current occupation_Unemployed	1.1475	0.085	13.534	0.000	0.981	1.314
What is your current occupation_Working Professional	3.6975	0.197	18.805	0.000	3.312	4.083
Lead Origin_Lead Add Form	3.8340	0.199	19.237	0.000	3.443	4.225
Lead Source_Olark Chat	1.4767	0.117	12.570	0.000	1.246	1.707
Lead Source_Welingak Website	1.9099	0.743	2.569	0.010	0.453	3.367
Do Not Email_Yes	-1.3264	0.165	-8.026	0.000	-1.650	-1.002
Last Activity_SMS Sent	1.3898	0.073	19.012	0.000	1.246	1.533
Last Notable Activity_Had a Phone Conversation	3.6484	1.129	3.231	0.001	1.435	5.862
Last Notable Activity_Olark Chat Conversation	-0.7921	0.339	-2.336	0.019	-1.457	-0.128
Last Notable Activity_Unreachable	2.1014	0.532	3.951	0.000	1.059	3.144

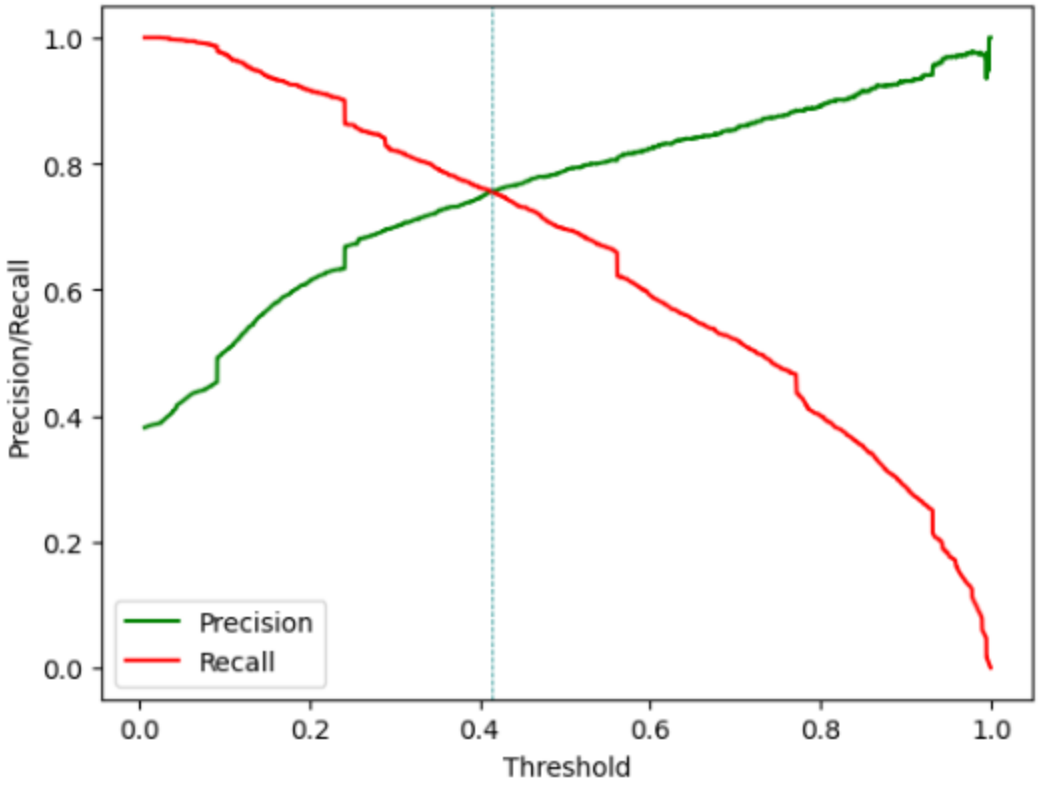
Model Building and Evaluation





- 0.332 is the approximate point where all the curves meet, so 0.332 seems to be our Optimal cutoff point for probability threshold

Model Building and Evaluation



- Using a precision-recall cutoff of 0.413, the True Positive Rate (Sensitivity/Recall) has dropped to about 75%, while the business objective requires this metric to be closer to 80%.
- We have successfully attained the desired metrics of 80% for both Sensitivity and Specificity by implementing a cutoff threshold of 0.332. Consequently, we shall continue with the sensitivity-specificity analysis to determine the optimal cutoff for making final predictions.

Model Performance

- **Train Dataset**

True Negative : 3211

True Positive : 1981

False Negative : 485

False Positive : 791

Model Accuracy : 0.8027

Model Sensitivity : 0.8033

Model Specificity : 0.8023

Model Precision : 0.7146

Model Recall : 0.8033

Model True Positive Rate (TPR) : 0.8033

Model False Positive Rate (FPR) : 0.1977

- **Test Dataset**

True Negative : 1369

True Positive : 883

False Negative : 212

False Positive : 308

Model Accuracy : 0.8124

Model Sensitivity : 0.8064

Model Specificity : 0.8163

Model Precision : 0.7414

Model Recall : 0.8064

Model True Positive Rate (TPR) : 0.8064

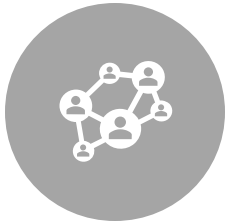
Model False Positive Rate (FPR) : 0.1837

Conclusion

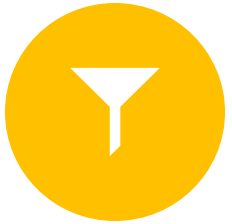
- The model achieved a sensitivity of 80.33% in the train set and 80.64% in the test set, using a cut-off value of `0.332`.
- Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting
- The CEO of X Education had set a target sensitivity of around 80%.
- The model also achieved an accuracy of 80.29% on training dataset, which is in line with the study's objectives.

RECOMMENDATIONS

Focus on features with positive coefficients for targeted marketing strategies.



Develop strategies to attract high-quality leads from top-performing lead sources.



Focus on creating more interactive and engaging content to increase the time spent by leads on the website, as it strongly correlates with conversions.

Engage working professionals with tailored messaging.

Provide proactive and personalized interactions via Olark Chat and Welingak Website to increase its effectiveness in driving conversions.

Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

Thank You