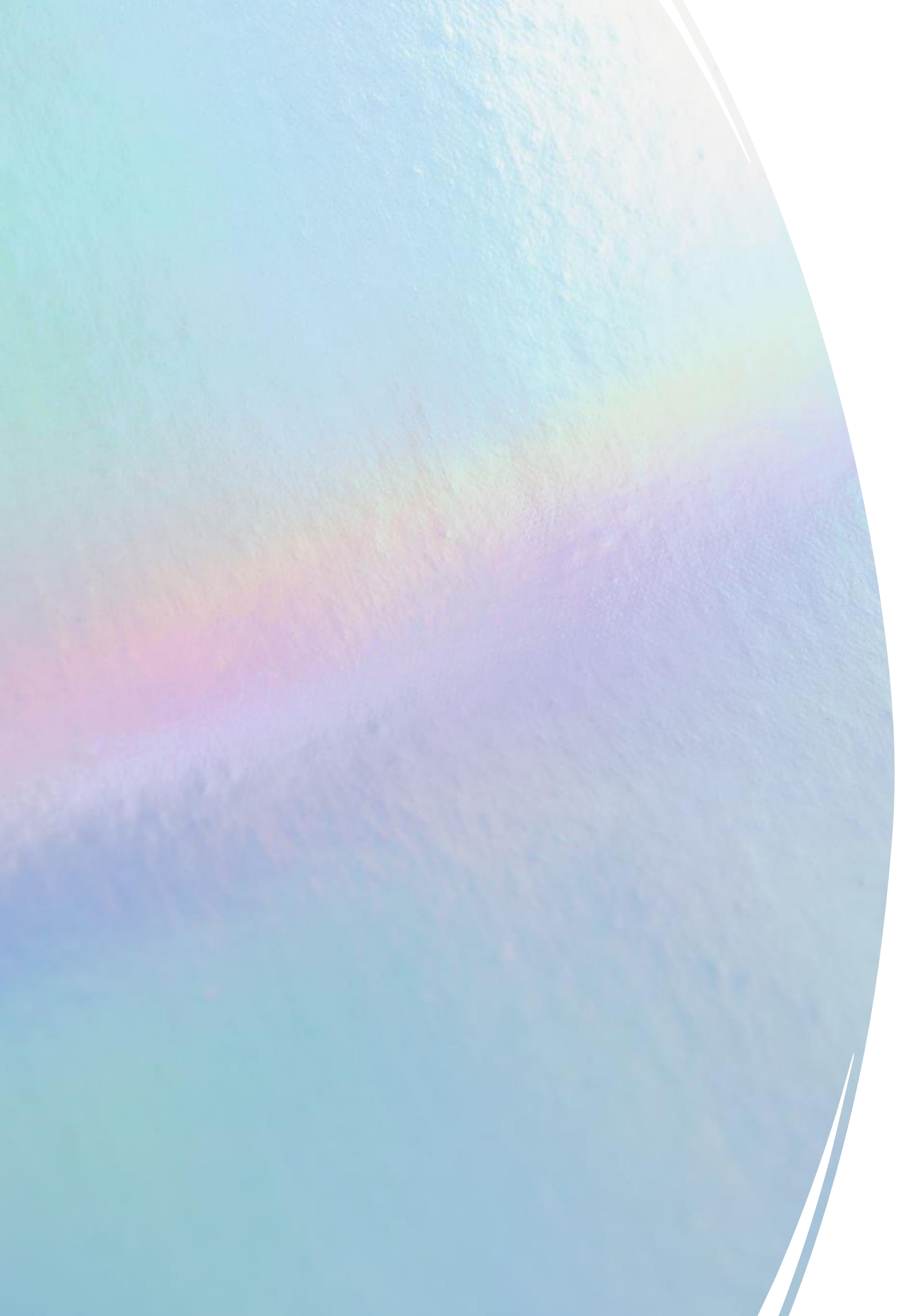


LEAD SCORING CASE STUDY

PROBLEM STATEMENT:

- X Education is an organization which provides online courses to industry professionals. The company marks it's courses on many popular websites like Google. • X Education wants to select most promising leads that can be converted to paying customers. • Although the company generates a lot of Leads, only a few are converted into paying customers, wherein the company wants a higher lead conversion rate. Leads come through numerous modes like email, advertisements on websites, google searches etc. • The company has had 30% conversion rate through the whole process of turning Leads into customers by approaching those Leads which are to be found having interest in taking the course. The implementation process of lead generating attributes are not efficient in helping conversations.

BUSINESS GOAL:

The company wants to build a model for selecting the most promising leads.

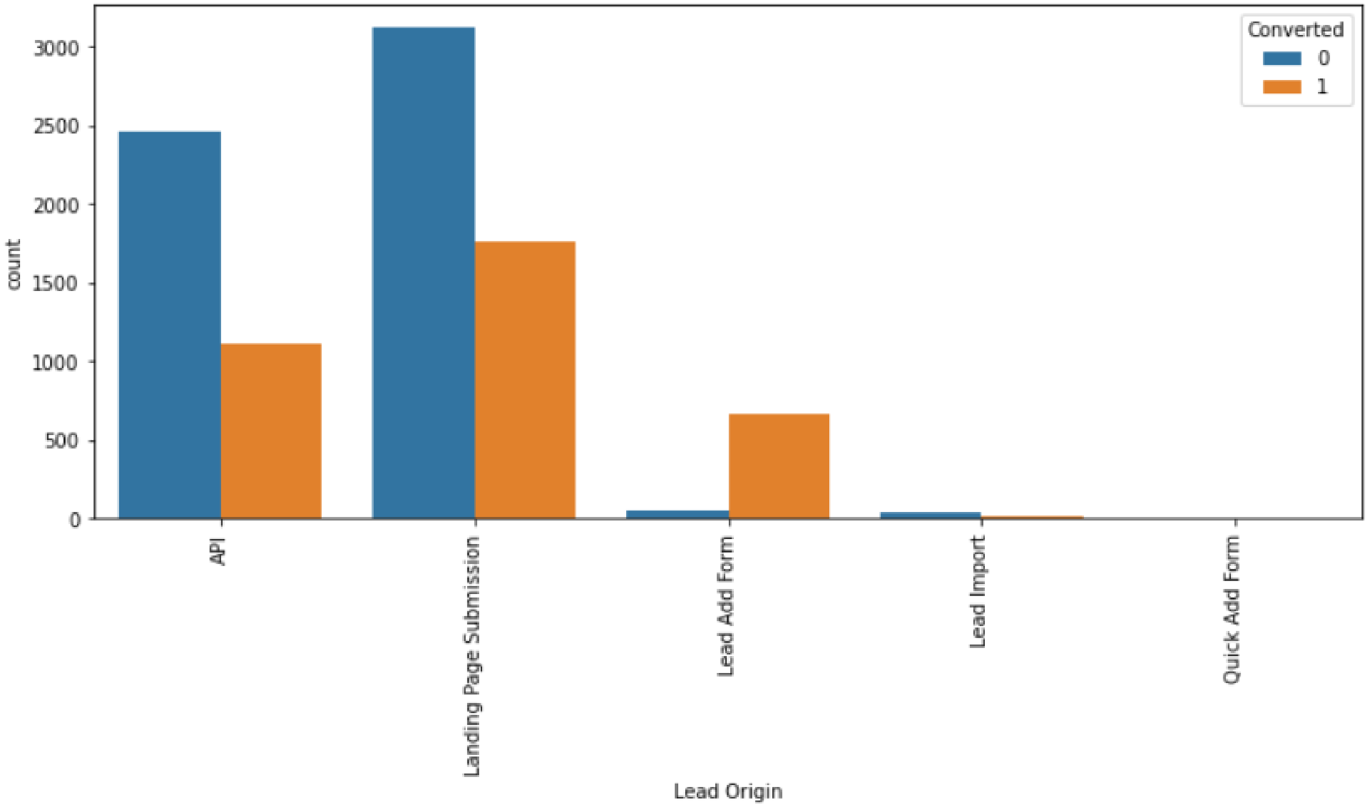
A lead score has to be given to each leads to indicate how promising the lead could be. The higher the lead score, the more promising the lead is to get converted. The lower the score, the lesser the chances of conversion.

The built model should have a conversion rate of around 80% or more.

STRATEGY

Importing Data and Libraries
Checking the Dataframe
Preparation the Data
Exploratory Data Analysis
Outlier Detection and Treatment
Creating Dummy Variables
Train - Test Split
Feature Scaling
Model Building and Feature Selection using RFE
Creating Confusion Matrix
Plotting the ROC Curve and finding optimal cutoff point
Precision and Recall and F1 Score
Assigning Lead Score with respect to Lead_Num_ID

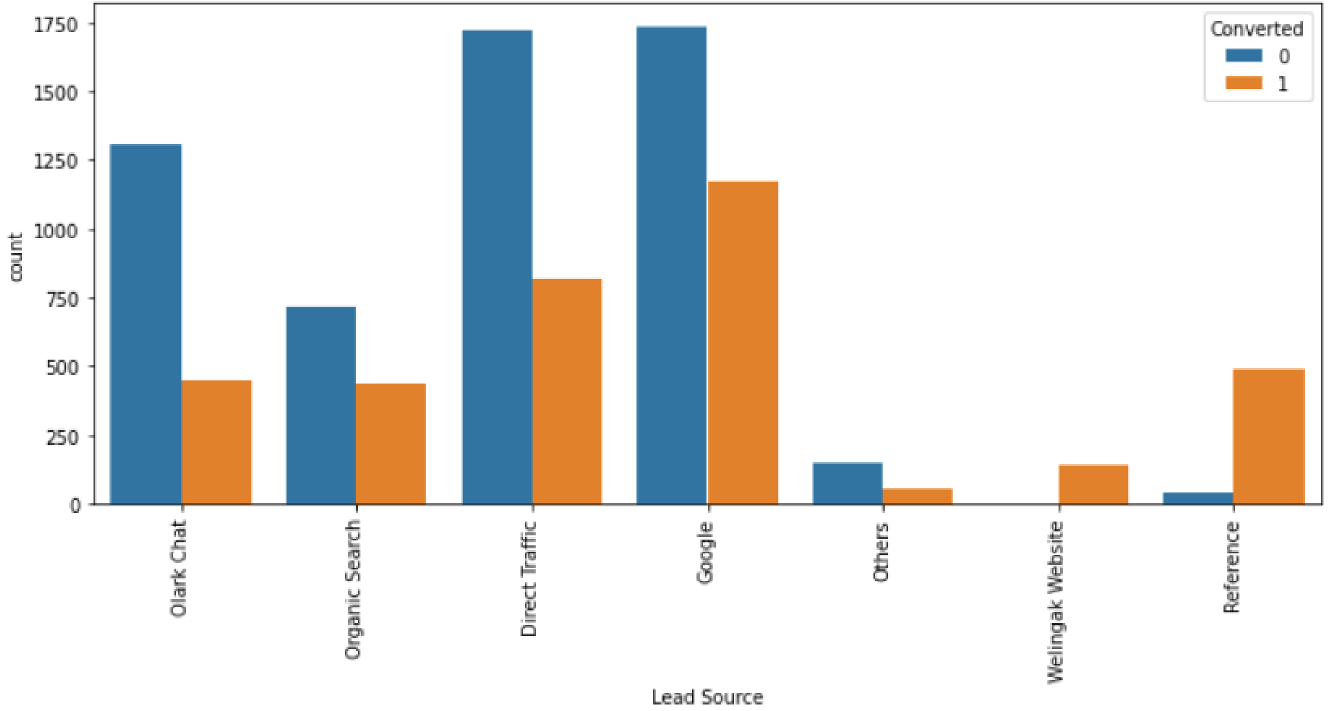
EDA



- **Lead Origin v/s Converted:**

- Leads who originate from Lead Add Form have high percentage of conversion compared to API and Landing Page Submissions.

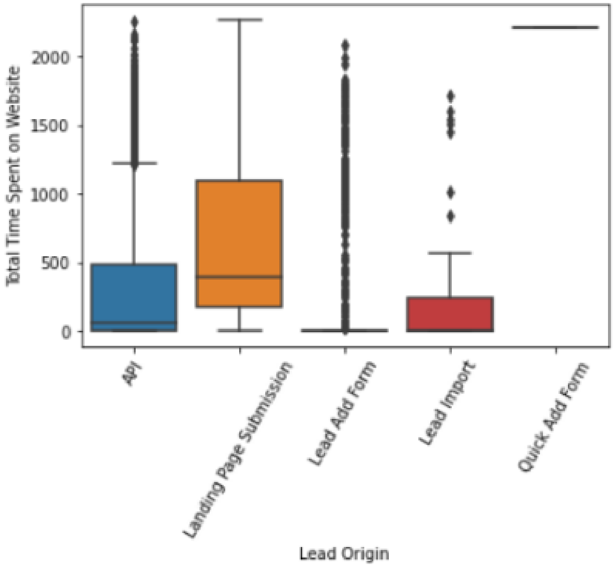
EDA



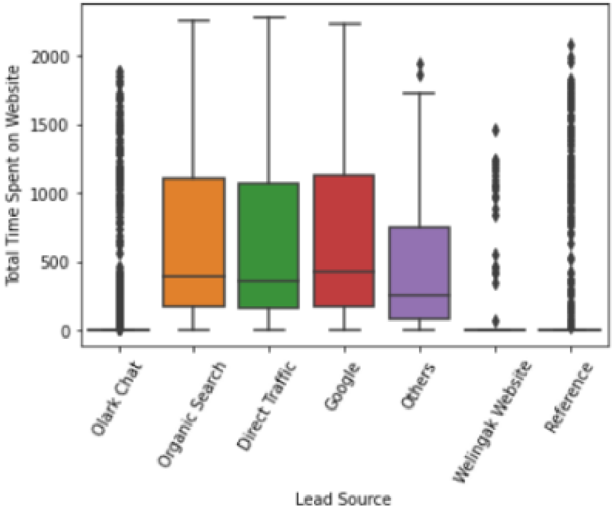
- **Lead Source v/s Converted:**

- Conversion rate of Reference source is maximum alongside the ones which originates directly from Google.
- Other sources such as Direct traffic and Olark chat etc. have comparatively low conversion rates.

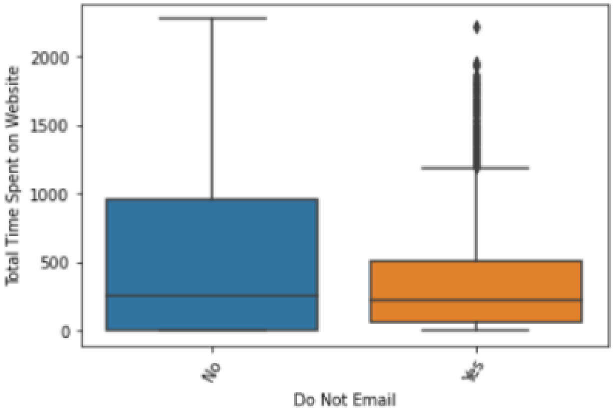
BIVARIATE ANALYSIS



Lead Origin Vs Total Time Spent on Website

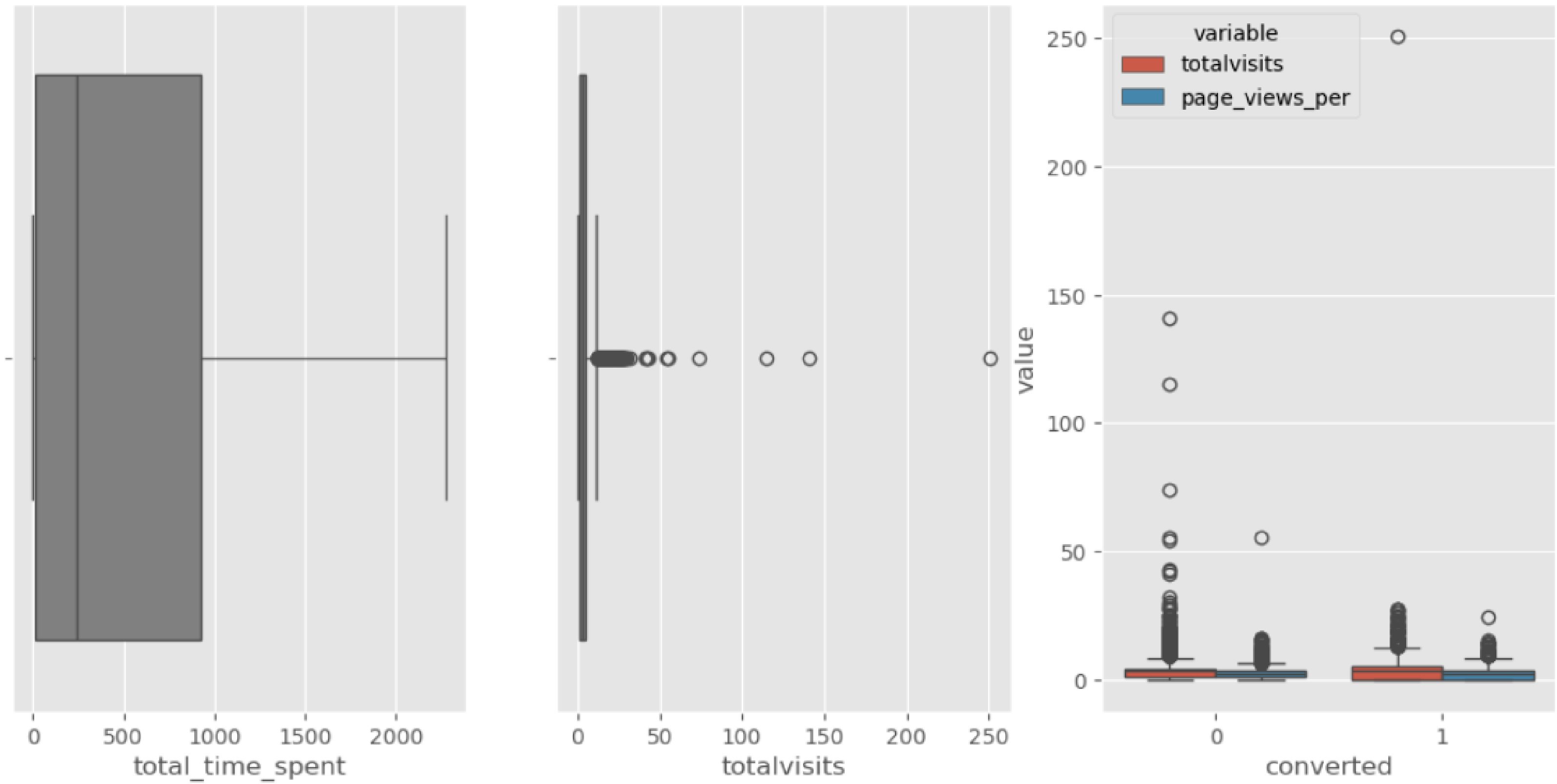


Lead Source Vs Total Time Spent on Website



Do Not Email Vs Total Time Spent on Website

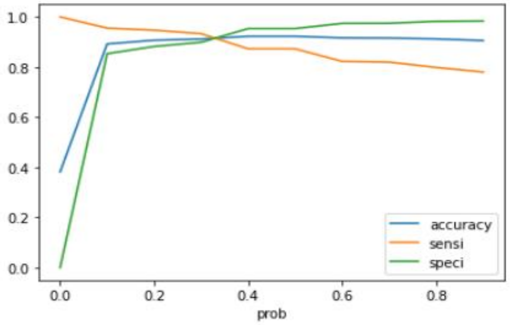
BIVARIATE

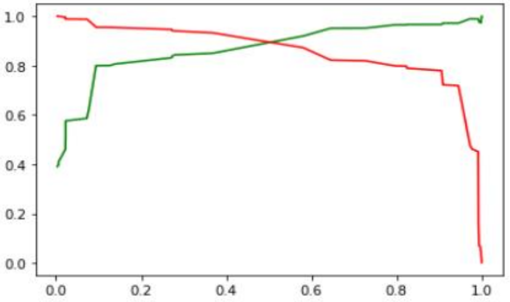

0
500
1000
1500
2000
total_time_spent
0
50
100
150
200
250
totalvisits
variable
🟥 totalvisits
🟦 page_views_per
250
200
150
100
50
0
value
0
1
converted

MODEL BUILDING

- Splitting into train and test set.
- Scale and fit transform variables in train set.
- Build the first model.
- Use RFE to eliminate less relevant variables.
- Build the next model.
- Eliminate variables based on high p-values.
- Check VIF value for all the existing columns.
- Predict using train set.
- Evaluate accuracy and other metric.
- Predict using test set.
- Precision and recall analysis on test predictions.

MODEL EVALUATION AND PREDICTION



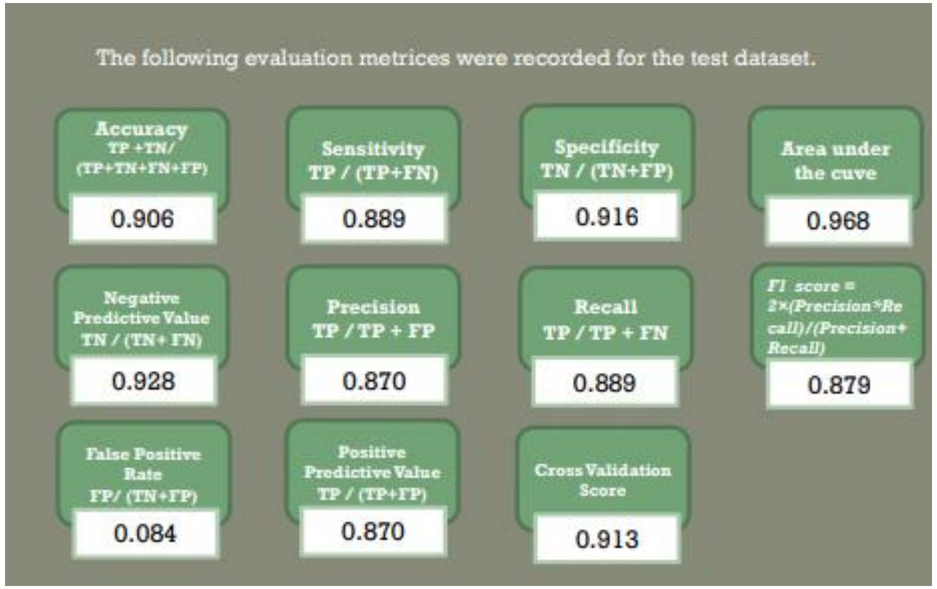


Accuracy, Sensitivity and Specificity:

- 88% Accuracy
- 86% Sensitivity
- 88% Specificity

Precision, Recall and F1 score:

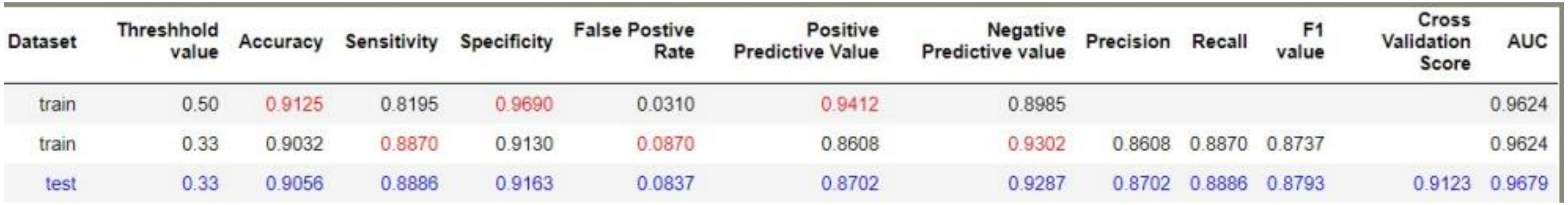
- 82.7% Precision
- 86% Recall
- 84.52% F1 score



The following evaluation metrices were recorded for the test dataset.

Accuracy **TP +TN/ (TP+TN+FN+FP)** 0.906	**Sensitivity** **TP / (TP+FN)** 0.889	**Specificity** **TN / (TN+FP)** 0.916	**Area under the cuve** 0.968
Negative Predictive Value **TN / (TN+ FN)** 0.928	**Precision** **TP / TP + FP** 0.870	**Recall** **TP / TP + FN** 0.889	***F1 score = 2×(Precision*Recall)/(Precision+ Recall)*** 0.879
False Positive Rate **FP/ (TN+FP)** 0.084	**Positive Predictive Value** **TP / (TP+FP)** 0.870	**Cross Validation Score** 0.913	

FINAL PARAMETERS



Dataset	Threshhold value	Accuracy	Sensitivity	Specificity	False Postive Rate	Positive Predictive Value	Negative Predictive value	Precision	Recall	F1 value	Cross Validation Score	AUC
train	0.50	0.9125	0.8195	0.9690	0.0310	0.9412	0.8985					0.9624
train	0.33	0.9032	0.8870	0.9130	0.0870	0.8608	0.9302	0.8608	0.8870	0.8737		0.9624
test	0.33	0.9056	0.8886	0.9163	0.0837	0.8702	0.9287	0.8702	0.8886	0.8793	0.9123	0.9679

Statistical Significance: All variables have p-values less than 0.05, indicating that the relationships observed are statistically significant.

Multicollinearity: Low Variance Inflation Factor (VIF) values suggest minimal multicollinearity among the features, which is important for ensuring the reliability of the model's coefficients.

Correlation: The heatmap supports your findings by visually indicating that features are not highly correlated with one another.

Model Performance: An overall accuracy of 0.9056 at a probability threshold of 0.33 on the test dataset indicates a strong model performance, suggesting that it can effectively classify or predict outcomes based on the features.

SUMMARY

- – Leads that spend more than average time on the website are promising and should be prioritized for targeted outreach to improve conversion rates.
- – SMS messages have shown to significantly impact lead conversions.
- – Landing page submissions are an effective way to capture more leads.
- – Leads specializing in marketing management and human resources management have higher conversion rates, making them valuable prospects.
- – Referrals and offers for referring new leads can boost conversion rates.
- – Alert messages or notifications have been observed to improve lead conversion rates.
- – The model's threshold was selected based on Accuracy, Sensitivity, Specificity, and Precision-Recall curves
- – The model achieved an approximate 88% accuracy on both train and test data, suggesting that the Logistic Regression model is a good fit.
- – The model demonstrates 83% recall and around 87% precision meaning it is about 83% effective in predicting positive outcomes (Hot Leads).
- – Overall, this model has proven to be accurate and reliable for identifying potential leads.