X - EDUCATION

CAPSTONE PROJECT
LEAD SCORES ANALYSIS

BY: SATHYAVANI K

DATE: 22-05-2025

Problem statement

- Introduction
- **X Education, an online learning platform, offers courses tailored for industry professionals. The company promotes its courses through various websites and search engines like Google.**
- **Visitors to the website may explore courses, watch videos, or fill out inquiry forms. When a visitor provides their email or phone number, they are classified as a lead. Additionally, leads are generated through past referrals.**
- **Once acquired, the sales team engages with these leads through calls and emails to convert them into customers. The company's typical lead conversion rate is around 30%.**
- **Business Goals**
- **X Education aims to identify high-potential leads, also known as “Hot Leads.”**
- **To achieve this, the company requires a predictive model that assigns a lead score, where higher scores indicate a greater likelihood of conversion. This system will help prioritize leads with a higher probability of becoming customers.**
- **The CEO has set a target lead conversion rate of 80%, emphasizing the need for an efficient lead qualification process.**

OVERALL APPROACH

1. Data Cleaning & Preprocessing: Handled missing values, removed irrelevant levels (e.g., ‘Select’), and performed feature engineering

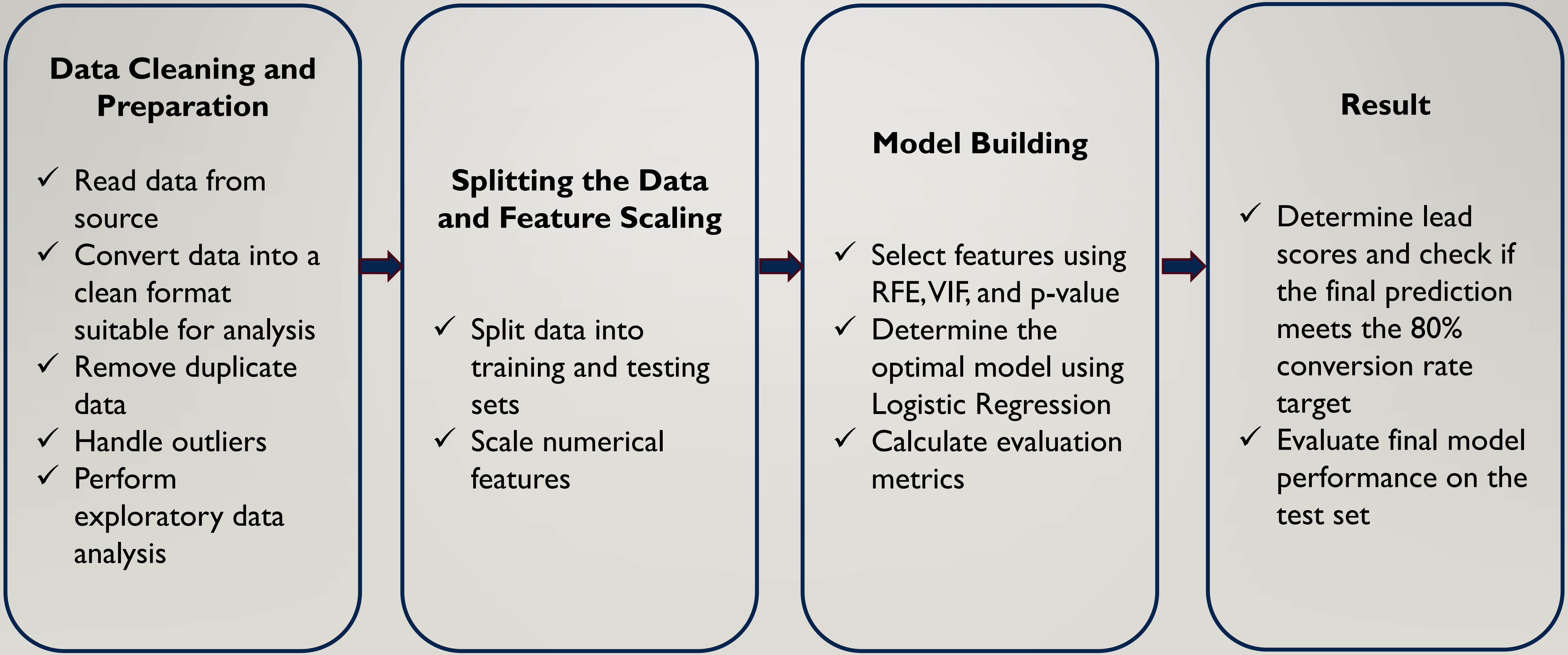
2. Exploratory Data Analysis (EDA): Identified key patterns and correlations using visualizations.

3. Logistic Regression Model: Built and optimized a model to predict conversion likelihood.

4. Evaluation Metrics: Used Accuracy, Precision, Recall, F1-score, and AUC-ROC for performance assessment.

5. Observation, Recommendation and Conclusion

ANALYTICAL APPROACH



Data Cleaning and Preparation

- ✓ Read data from source
- ✓ Convert data into a clean format suitable for analysis
- ✓ Remove duplicate data
- ✓ Handle outliers
- ✓ Perform exploratory data analysis

Splitting the Data and Feature Scaling

- ✓ Split data into training and testing sets
- ✓ Scale numerical features

Model Building

- ✓ Select features using RFE, VIF, and p-value
- ✓ Determine the optimal model using Logistic Regression
- ✓ Calculate evaluation metrics

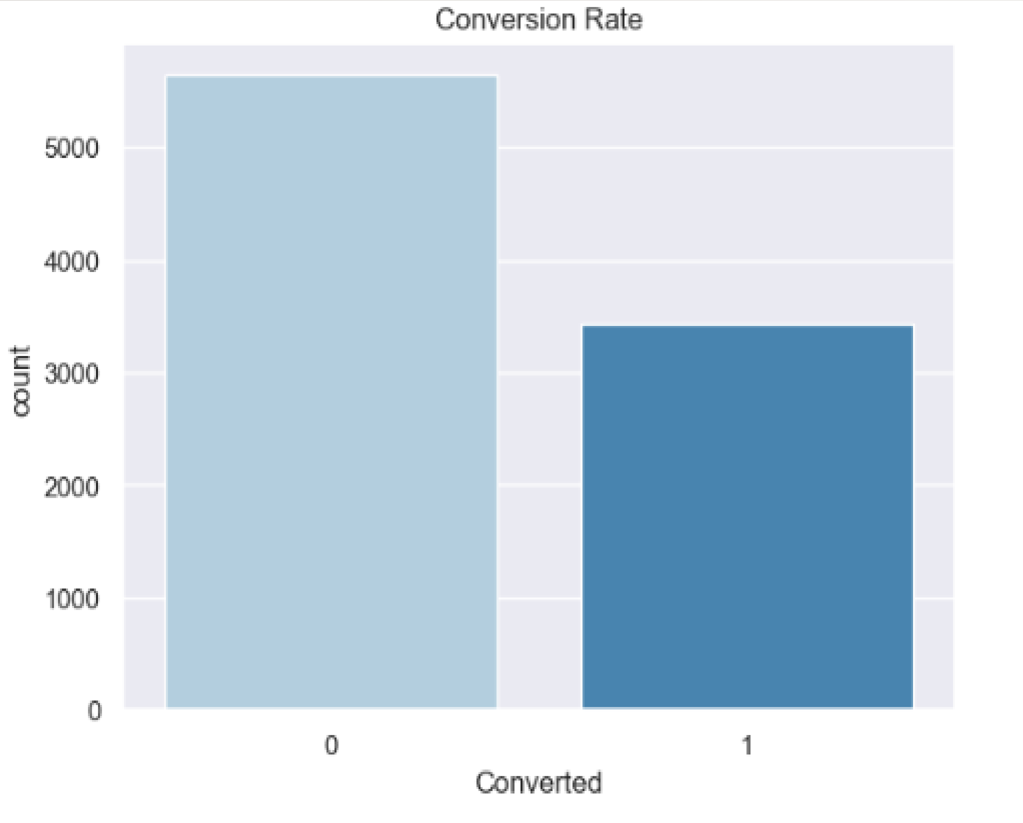
Result

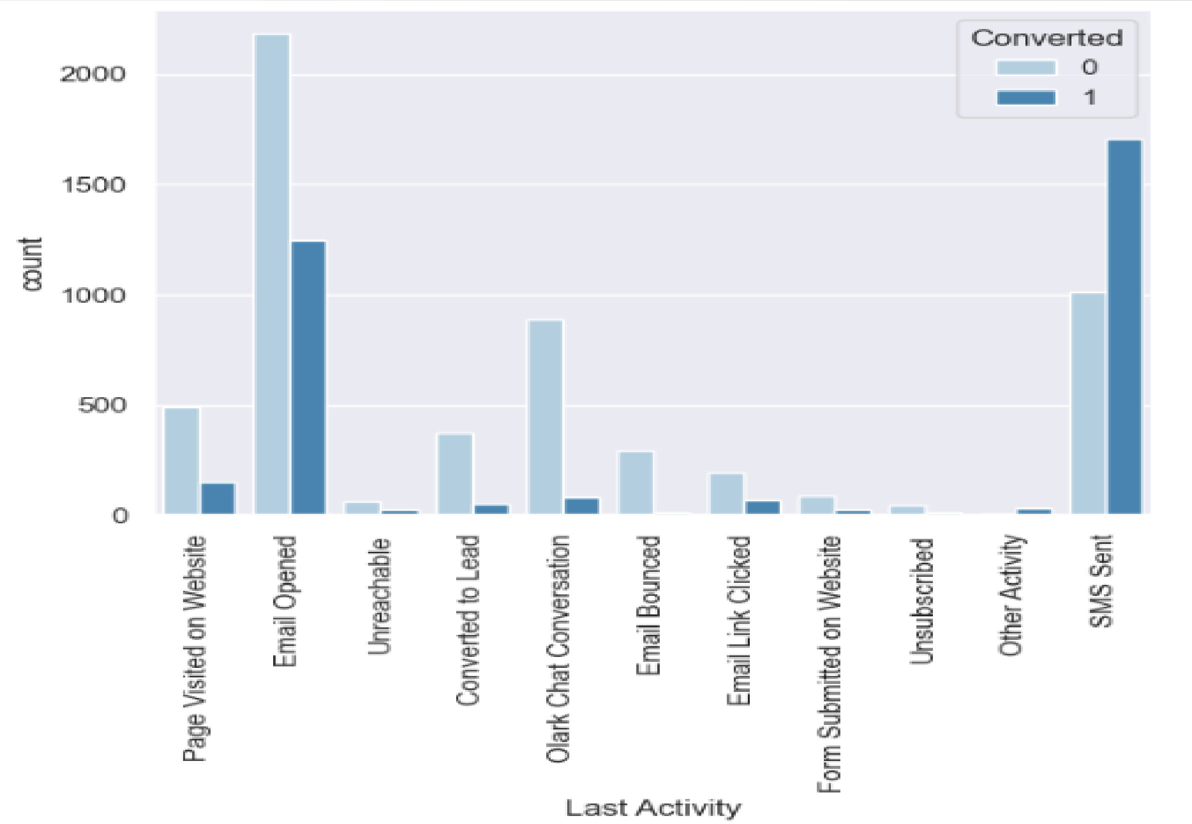
- ✓ Determine lead scores and check if the final prediction meets the 80% conversion rate target
- ✓ Evaluate final model performance on the test set

Data Cleaning & Preprocessing

- **Converting the variable with values YES/NO to 1/0.**
- **Converting the ‘SELECT’ values with NANs.**
- **Dropping the3 columns having>70% of null values**
- **Dropping unnecessary columns.**
- **Dropping the rows as the null values were<2%.**

Exploratory data analysis

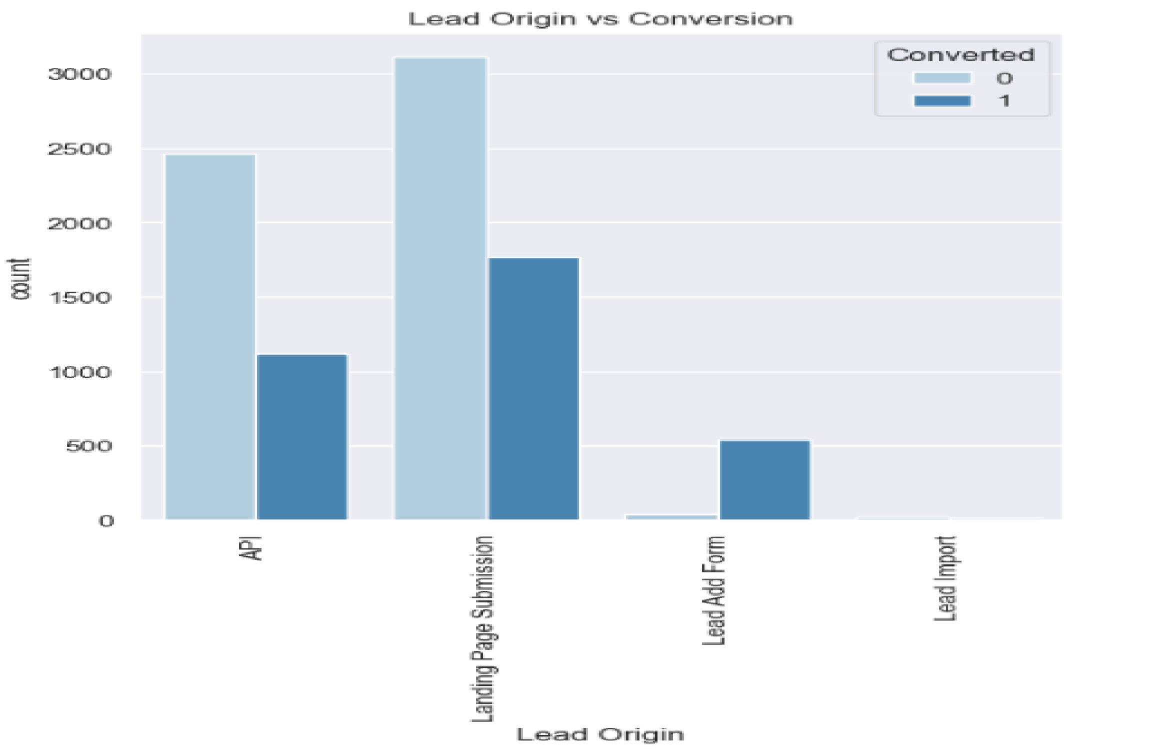




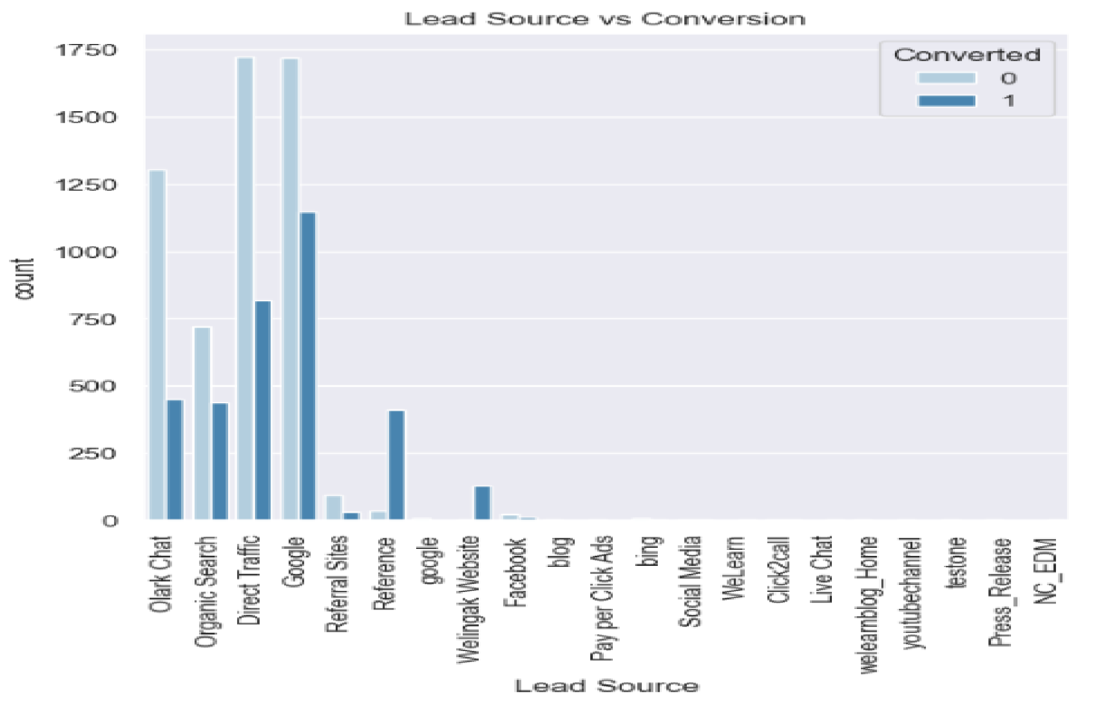
- **We have around 30% conversion rate**

- **The count of leads whose last activity was "Email Opened" is the highest.**
- **The conversion rate is highest when the last activity was "SMS Sent".**

EXPLORATORY DATA ANALYSIS



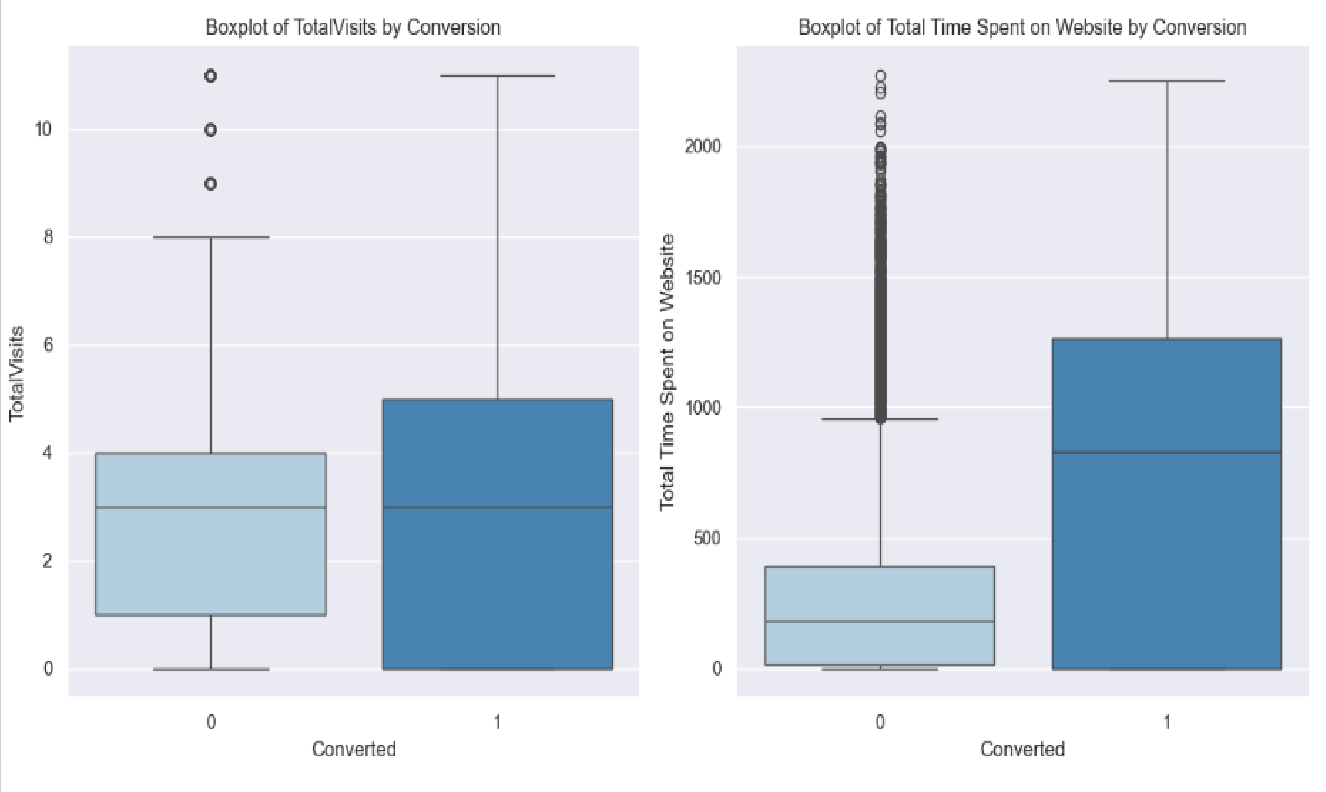
Lead Origin vs Conversion

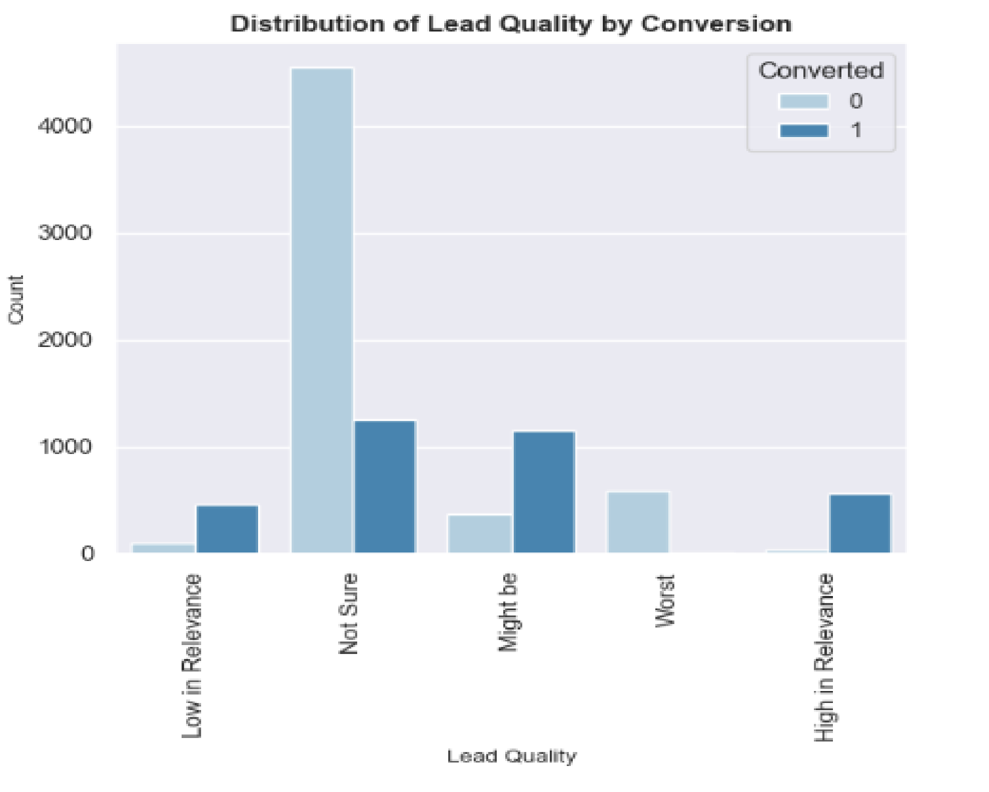


Lead Source vs Conversion

- The count of leads from Google and Direct Traffic is maximum.
- The conversion rate of the leads from Reference and Welingak Website is maximum.
- API and Landing Page Submission have a lower conversion rate (~30%) but a considerable number of leads.
- The count of leads from the Lead Add Form is low, but the conversion rate is very high.

EXPLORATORY DATA ANALYSIS



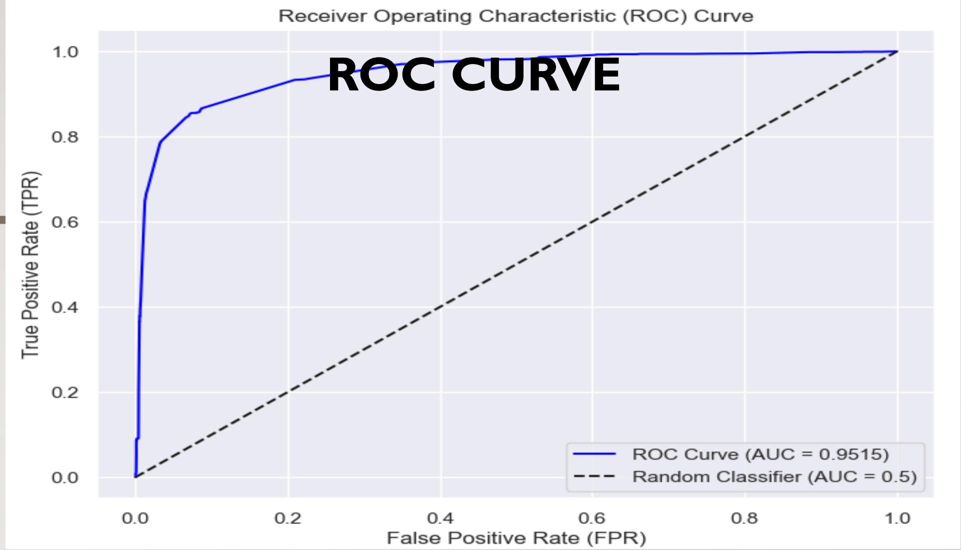


- **The median of both conversion and non-conversion is the same, making it inconclusive for decision-making.**
- **Users who spend more time on the website are more likely to convert.**

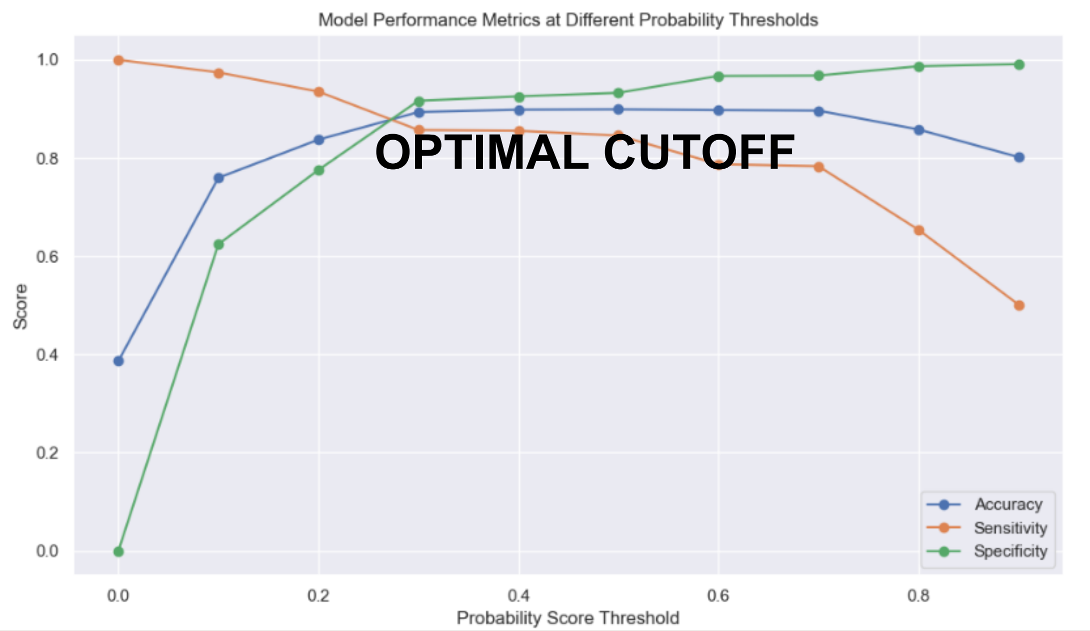
- **Most leads fall under "Not Sure," but their conversion rate is low.**
- **Leads labeled "High in Relevance" have the highest conversion rate.**
- **Focusing on high-relevance leads can improve overall conversions.**

MODEL BUILDING

- SPLITTING THE DATA INTO TEST AND TRAINING SETS
- WE HAVE CHOSEN THE TRAIN_TEST SPLIT RATIO AS 70:30
- USING RFE TO CHOOSE TOP 15 VARIABLES
- BUILD MODEL BY REMOVING THE VARIABLE
  WHOSE p-VALUE > 0.05 AND VIF > 10
- PREDICTIONS ON TEST DATASET
- OVERALL ACCURACY IS 91.33%



ROC CURVE



OPTIMAL CUTOFF

Conclusion

The logistic regression model is used to predict the probability of conversion of a customer.

While we have calculated both sensitivity-specificity as well as Precision-Recall metrics, we have considered the optimal cut-off on the basis of sensitivity-specificity for final prediction.

Lead Score calculated shows the conversion rate of the final predicted model is around 92% in test data as compared to 95% in train data.

In business terms, this model has the capability to adjust with the company’s requirements in the coming future.

TOP variables that contribute to lead getting converted in the model are:

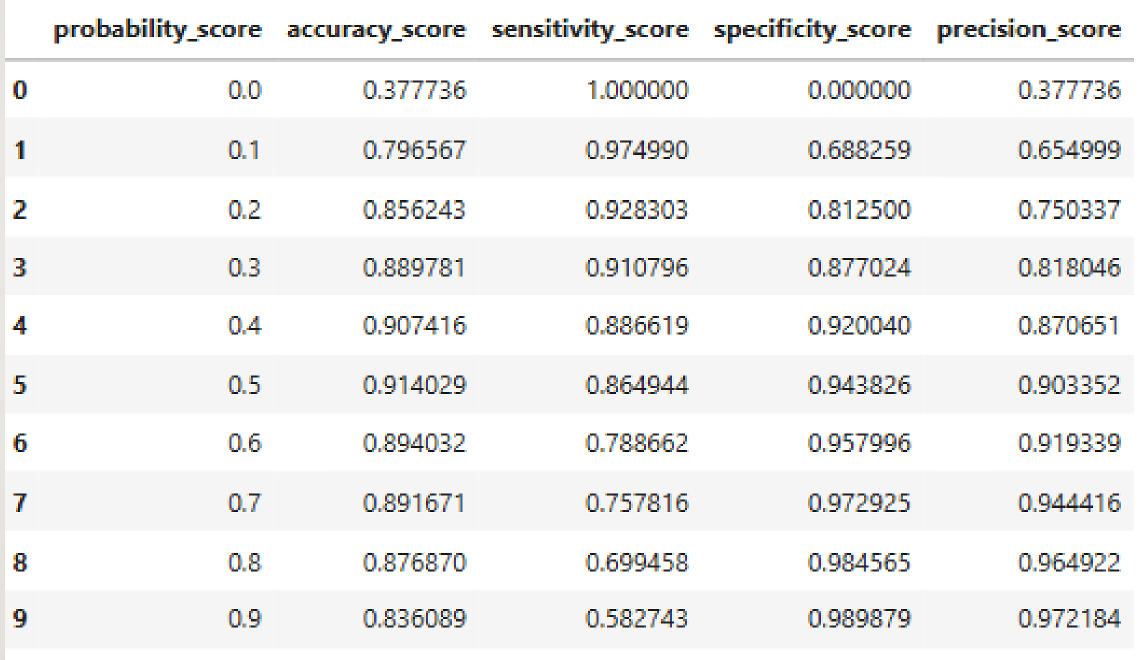
- Tags Lost to EINS
- Tags Closed by Horizon
- Lead Quality Worst

Hence, overall this model seems to be good.

Model evaluation

- ➢ CALCULATED ACCURACY, SENSITIVITY, AND SPECIFICITY FOR VARIOUS PROBABILITY CUTOFFS FROM 0.1 TO 0.9

- ➢ AS PER THE GRAPH AND LOOKING AT THE OTHER SCORES, IT CAN BE SEEN THAT THE OPTIMAL POINT IS 0.45



	probability_score	accuracy_score	sensitivity_score	specificity_score	precision_score
0	0.0	0.377736	1.000000	0.000000	0.377736
1	0.1	0.796567	0.974990	0.688259	0.654999
2	0.2	0.856243	0.928303	0.812500	0.750337
3	0.3	0.889781	0.910796	0.877024	0.818046
4	0.4	0.907416	0.886619	0.920040	0.870651
5	0.5	0.914029	0.864944	0.943826	0.903352
6	0.6	0.894032	0.788662	0.957996	0.919339
7	0.7	0.891671	0.757816	0.972925	0.944416
8	0.8	0.876870	0.699458	0.984565	0.964922
9	0.9	0.836089	0.582743	0.989879	0.972184

TRAIN DATA CONFUSION MATRIX

PREDICTED ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	3636	222
CONVERTED	324	2075

PERFORMANCE METRIC

ACCURACY	91.40%
PRECISION	90.34%
SENSITIVITY	86.49%
SPECIFICITY	94.38%