Lead Scoring

Manthan Jha

Introduction

Company named X Education gets a lot of leads. However, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Business Problem

- Despite receiving many leads, X Education's lead conversion rate is low

- Identifying potential leads (Hot Leads) can increase conversion rates

- Focusing on potential leads reduces time and resources spent on low-conversion leads

- Higher conversion rates improve sales team efficiency and effectiveness

- The objective is to build a model to assign lead scores and prioritize Hot Leads

Data Description

- "Leads.csv" dataset: 37 columns, 9240 rows
- It is the main dataset for EDA and model building
- Contains features related to leads, target variable (converted)
- Includes lead demographics, lead source, marketing interactions, website activity, communication history, etc.
- "Test.csv" dataset: 36 columns, 2007 rows
- This is a separate dataset for testing the logistic regression model

Data Preprocessing

- Handled missing values: Dropped off columns with more than 40% missing values, impute missing values for remaining columns

- Segment data: Categorized columns into categorical and numerical variables

- Handled outliers: Checked for outliers using box plots, applied capping method to impute outliers.

- Prepared the data for EDA and model development
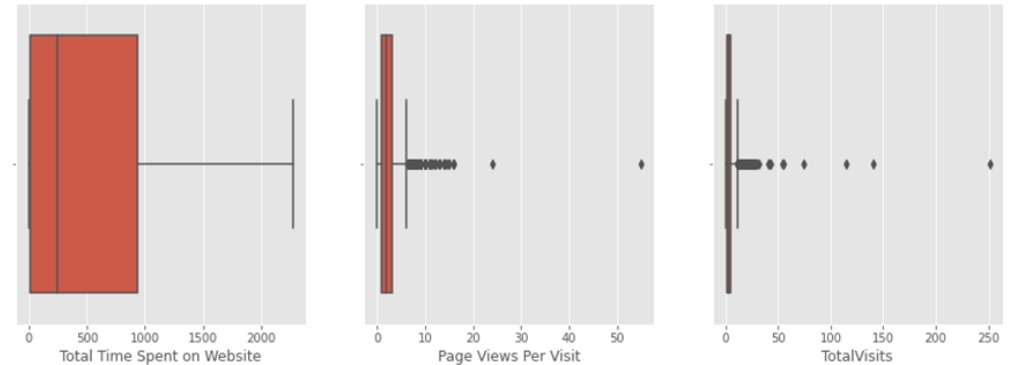
Exploratory Data Analysis (EDA)

- Performed univariate analysis: To understand the distribution and characteristics of individual variables.

- To Identify key features that may influence lead conversion.
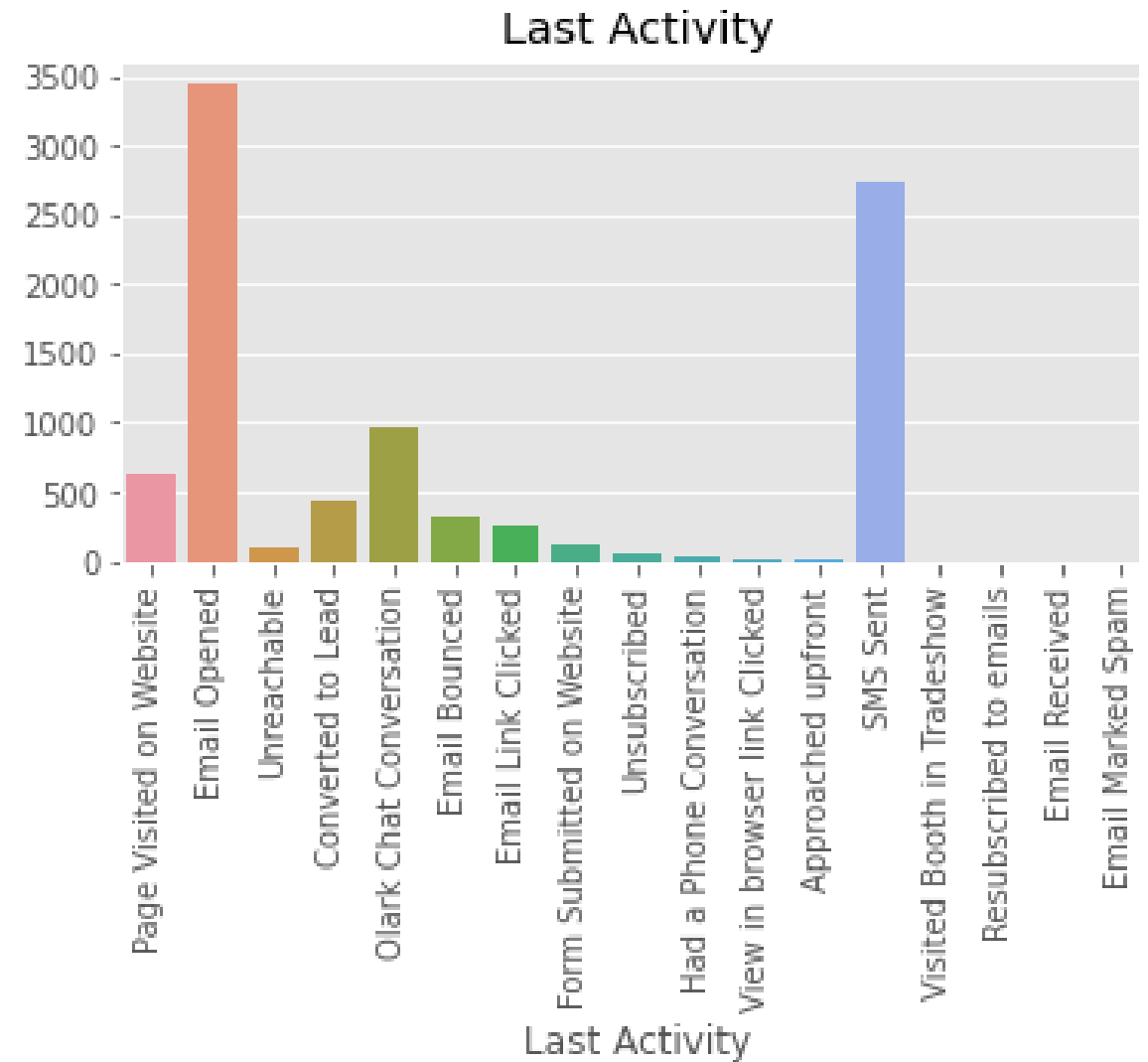
Exploratory Data Analysis (EDA)

- Analyzed continuous variables: Check for skewness and take necessary steps for normalization

- The Continuous columns suggested the data was skewed to the left.
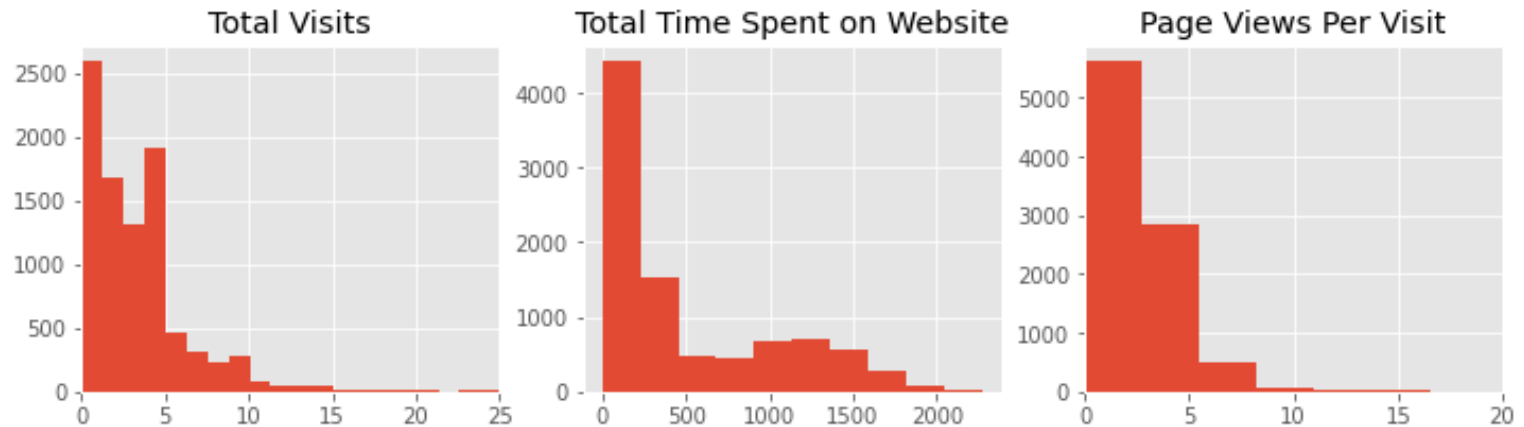
Exploratory Data Analysis (EDA)

- Conducted bivariate analysis: Explore relationships between variables and their impact on lead conversion.

- These columns give insights into the matter showing certain significant factors that might affect the model.
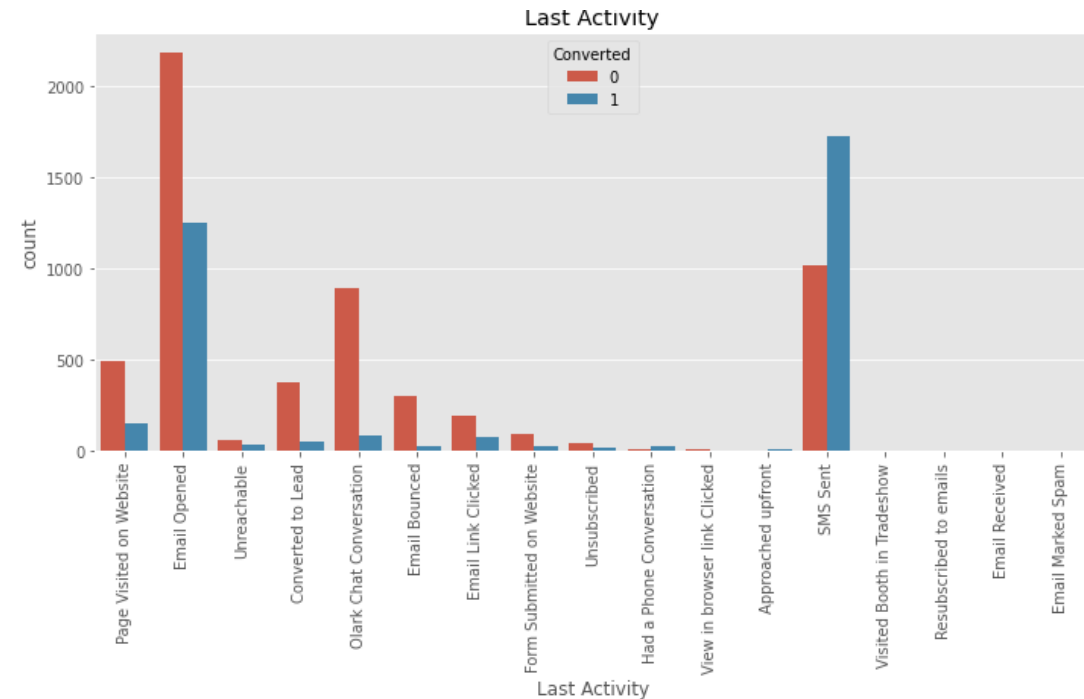
Model Development

- Created dummy variables: Converted categorical variables into binary indicators using one-hot encoding

- Splited the dataset into training and testing data: Allocate data for model training and evaluation

- Scaled numerical variables: Standardize numerical features to have a mean of 0 and standard deviation of 1

- Built a logistic regression model using StatsModels

- Selected optimal features using Recursive Feature Elimination (RFE) and variance inflation factor (VIF)

Model Evaluation

Model was evaluated on basis of certain metrics:

- Accuracy: 90 % approx

- sensitivity: 0.86

- specificity: 0.95

- false_postive_rate: 0.041

- positive_predictive_value: 0.92

- negative_predictive_value: 0.92

Receiver operating characteristic
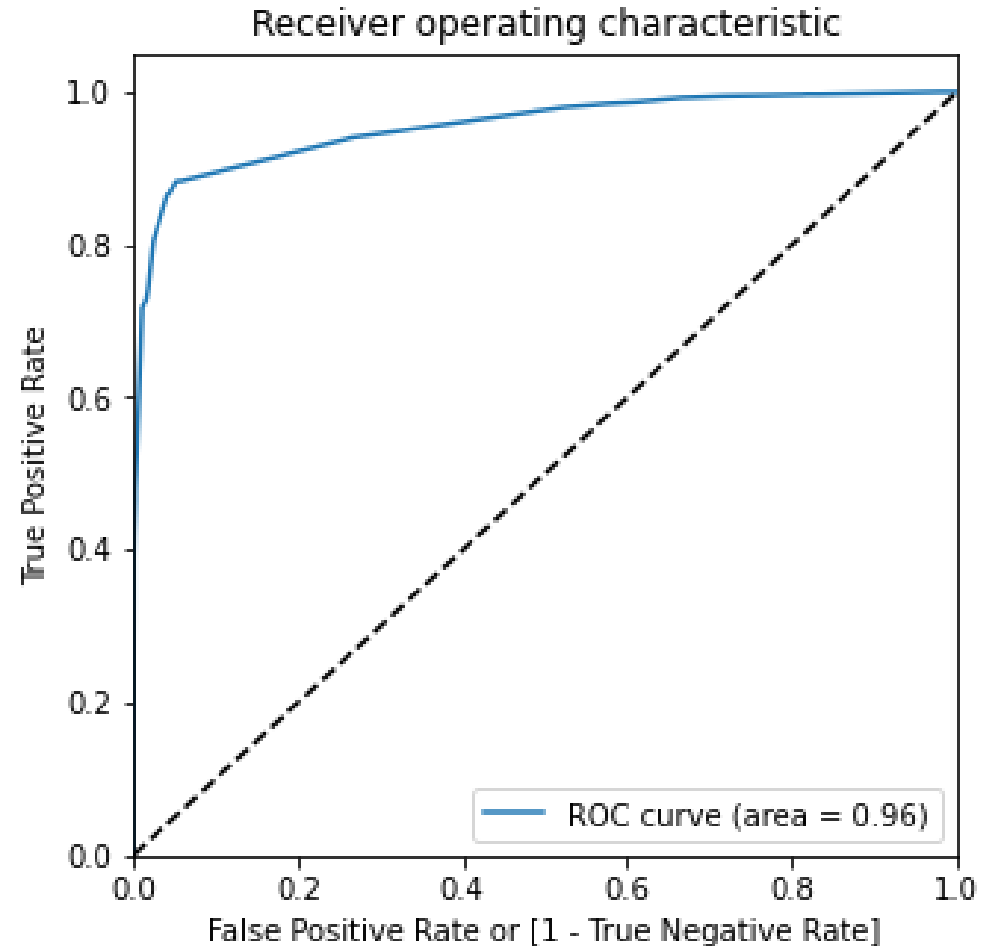
Model Testing

- Applied the trained logistic regression model on the "Test.csv" dataset
- Cleaned the test data by handling missing values and outliers
- Created dummy variables for categorical features
- Predicted the probability of lead conversion using the logistic regression model
- Evaluated the model's performance on the test dataset

Recommendations

Based on the data analysis and understanding, here are five key recommendations for X Education to improve lead conversion rates:

- Focus on Lead Nurturing: Prioritize personalized communication channels such as emails, SMS messages, and targeted newsletters to engage with potential leads and provide relevant information.

- Utilize Automated Communication: Implement automated SMS messages and email workflows to target leads with high conversion potential, based on the predictive model. Personalize these messages to encourage leads to take action.

- Collaborate with Data Science Team: Work closely with the data science team to fine-tune the predictive model and gather feedback for continuous improvement. Incorporate insights from the team to enhance the accuracy and effectiveness of the model.

- Offer Incentives and Discounts: Develop strategies to provide discounts or incentives to create urgency and motivate potential leads to make a purchase decision.

- Diversify Communication Channels: Expand communication efforts beyond phone calls by leveraging email, social media, and chatbots. Engage with leads through their preferred channels to increase response rates and improve conversion rates.

Thank you