Case Study : Leads Scoring

Optimizing Lead Conversion for X Education

-By Anushree, Prashant Kumar, Steven H.

Introduction

- X Education, an online course provider, attracts numerous industry professionals to its website each day. professionals explore courses after discovering them through various online channels like Google.
- Upon visiting the website, they may browse courses, fill out forms, or watch videos. Those who provide contact details like email addresses or phone numbers are categorized as leads.
- The company also receives leads through referrals. However, only a fraction of these leads are converted into paying customers, with a typical conversion rate of 30%. X Education aims to improve lead conversion efficiency by identifying potential leads, or 'Hot Leads,' to increase the conversion rate to approximately 80%.

Objectives:

- **Identify Hot Leads:** Develop a lead scoring model to identify potential leads with a higher likelihood of conversion based on their demographics, online behavior, and interactions with the X Education platform.
- **Optimize Resource Allocation:** Prioritize high-quality leads for targeted marketing campaigns and personalized engagement strategies to maximize conversion rates while minimizing resource wastage.
- **Enhance Business Performance:** By improving lead conversion efficiency, X Education aims to enhance its business performance, increase revenue, and establish itself as a leader in the online education industry.
- **Key Questions:**

Dataset Description:

- The dataset used in our analysis comprises information collected from various online channels by X Education.
- It includes a comprehensive array of features that provide insights into leads' demographics, online behavior, and interactions with the X Education platform.
- **Features:**
- **Lead Origin:** Indicates the original source through which the lead was acquired, such as 'Direct Traffic,' 'Organic Search,' or 'Referral Sites.'
- **Last Activity:** Records the last known activity of the lead, whether it's visiting the website, filling out a form, or engaging with marketing content.
- **Current Occupation:** Specifies the current professional status or occupation of the lead, such as 'Working Professional,' 'Student,' or 'Unemployed.'
- **Tags:** Describes any tags or labels associated with the lead, providing additional context or categorization.
- **Lead Add Form:** Indicates whether the lead was generated through a specific lead capture form on the X Education website.
- **Email Preference:** Reflects the lead's preference regarding email communication, such as 'Opted-in' or 'Opted-out.'

Preprocessing Steps:

- **Handling Missing Values:** Initial preprocessing involved identifying and addressing missing values within the dataset. 'Select' values were replaced with NaN to facilitate imputation.
- **Column Reduction:** Columns with significant missing values (>40%) or those containing only one unique value were dropped to streamline the dataset.
- **Encoding Categorical Variables:** Categorical variables were encoded using one-hot encoding to transform them into a format suitable for model training.

Preprocessing Steps:

- The objective of analyzing this dataset is to develop a robust lead scoring model that accurately predicts the likelihood of lead conversion based on the provided features.
- By understanding the characteristics and behaviors of potential leads, X Education aims to prioritize high-quality leads and optimize its conversion strategies for maximum efficiency and profitability.

Exploratory Data Analysis (EDA)

Data Exploration:

- **Dataset Size:** The dataset comprises 9240 records of leads collected from various online channels.
- **Feature Overview:** We analyzed 37 features to gain insights into lead demographics, behavior, and interactions with the X Education platform.

Exploratory Data Analysis (EDA)

Missing Values Handling:

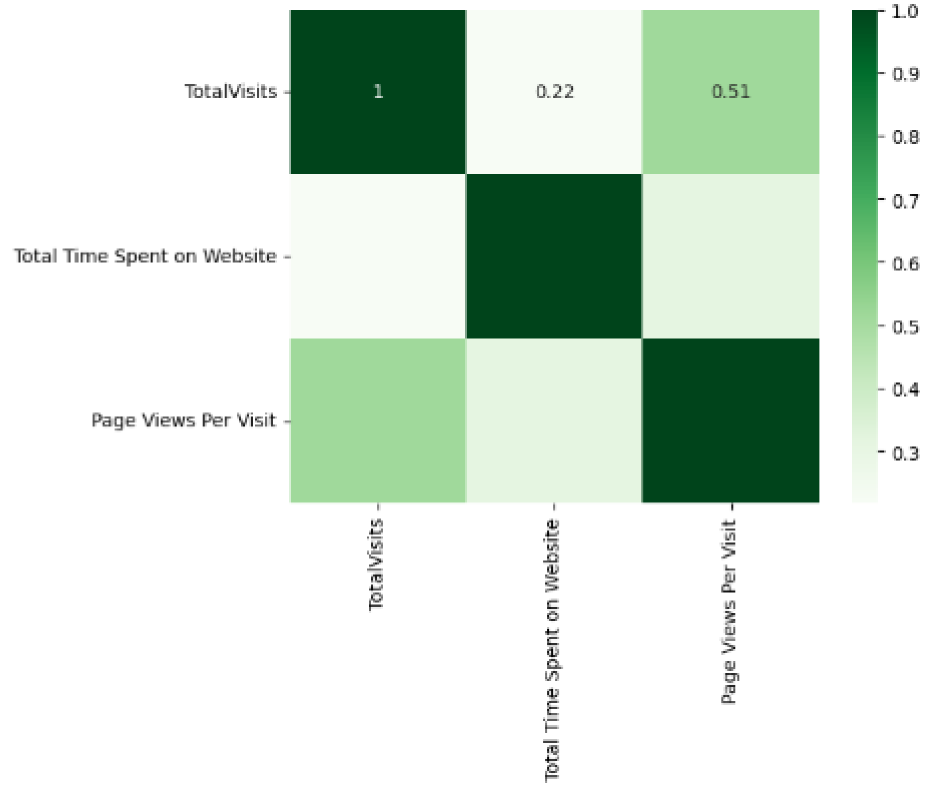
- **Identification:** Initially, we identified missing values within the dataset, with 'Select' values replaced with NaN for clarity.
- **Treatment:** Features [‘How did you hear about X Education’, ‘Lead Profile’, ‘Lead Quality’, ‘Asymmetrique Profile Score’, ‘Asymmetrique Activity Score’, ‘Asymmetrique Activity Index’ and ‘Asymmetrique Profile Index’] with significant missing values (>40%) were dropped, while categorical features ['City', 'Specialization', 'Tags', 'What matters most to you in choosing a course', 'What is your current occupation', 'Country', 'Page Views Per Visit', 'TotalVisits', 'Last Activity', 'Lead Source']were imputed using the mode, and numerical features were imputed using the median.

Feature Selection and Reduction

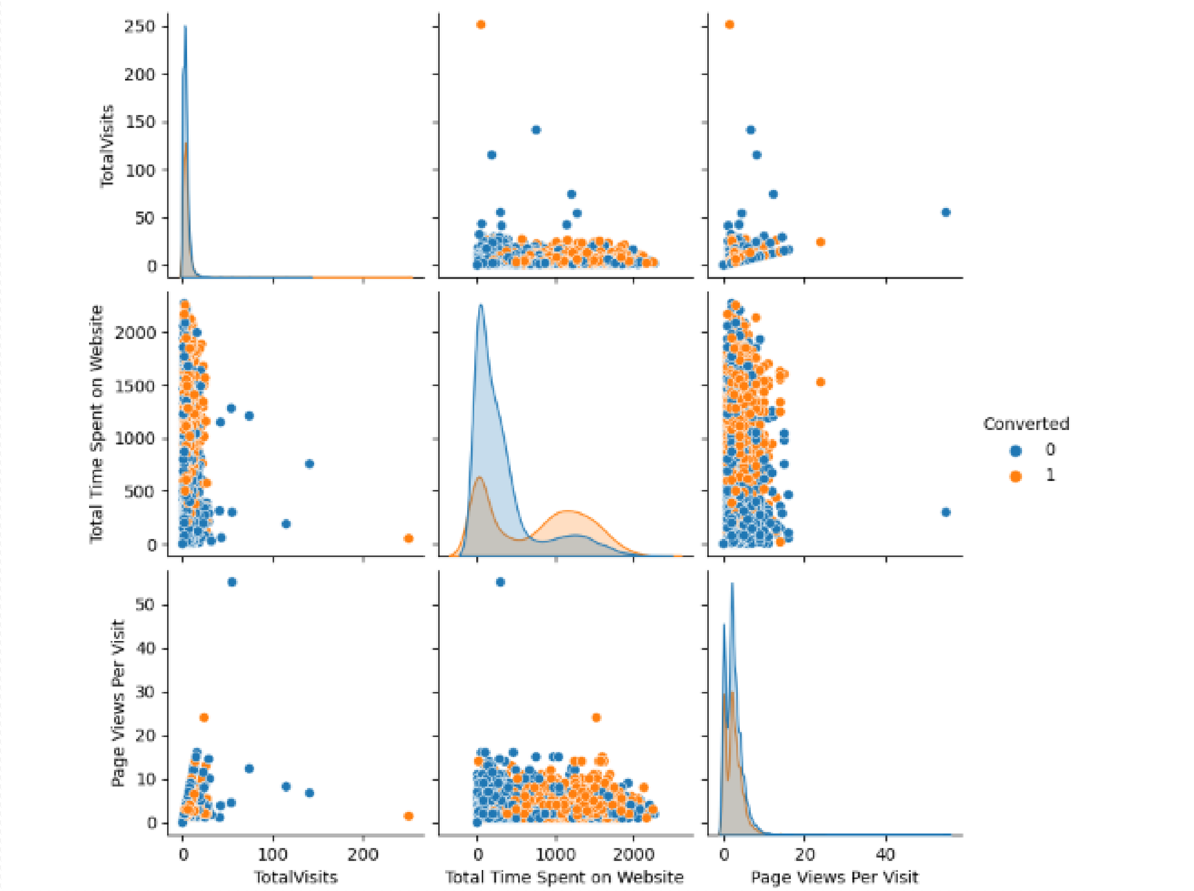
- **Single-Value Columns:** Columns containing only one unique value or redundant information, ["Prospect ID", "Lead Number"] were dropped to streamline the dataset.
- **Class Imbalance Check:**
- We assessed the distribution of the target variable ('Converted') to identify any class imbalance issues and ensure adequate representation of both converted and non-converted leads.

Correlation Analysis:

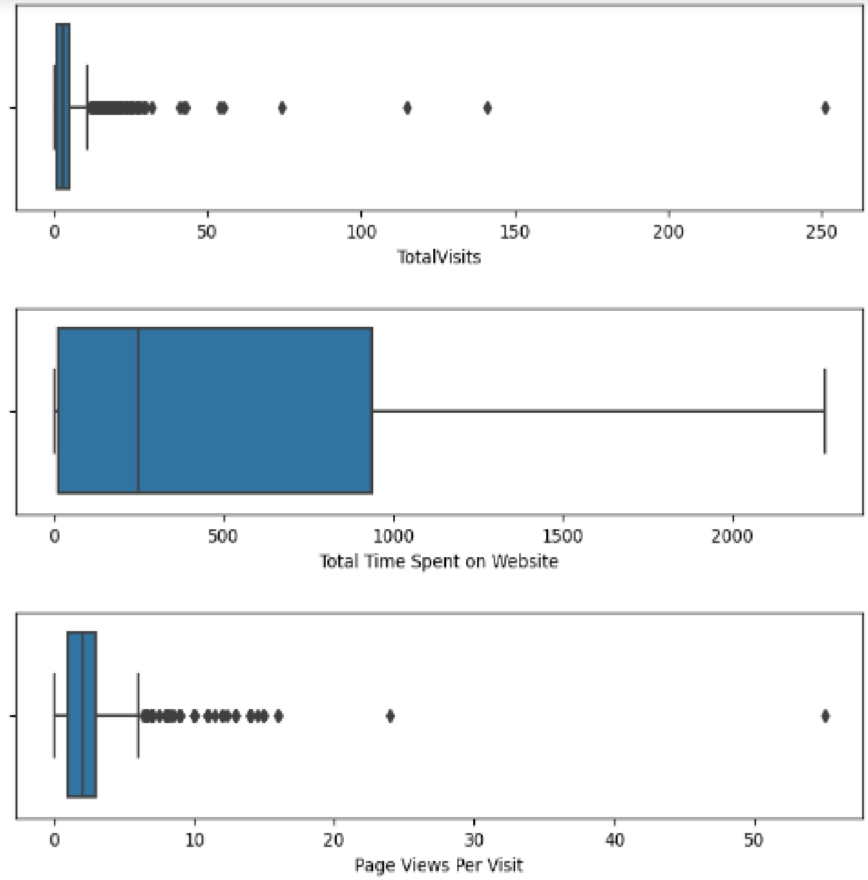
We examined the correlation between numerical features to identify potential multicollinearity issues.



Correlation between Numerical Features



Checking outliers in Numerical features



Feature Selection

- Utilized techniques such as Recursive Feature Elimination (RFE) to rank and select relevant features based on their impact on the target variable.
- The final set of selected features included 'Lead Origin_Lead Add Form', 'Do Not Email_Yes', 'Last Activity_Converted to Lead', among others.

Building Models

- Evaluate the trained model's performance on the test dataset using appropriate evaluation metrics such as R-squared, Mean Squared Error (MSE), and Adjusted R-squared.
- Compare the model's performance against baseline models and assess its ability to generalize to new data.

Building Models:

- **Model 1:** Identified key features such as lead origin, email activity, occupation, and lead tags, achieving an accuracy of 79.17% on the training set.
- **Model 2:** Improved accuracy by dropping the "Tags_Interested in Next Batch" feature, maintaining a sensitivity of 93.34% and specificity of 70.43%.
- **Model 3:** Enhanced performance further by removing the "Tags_Lateral Student" feature, maintaining high sensitivity and specificity levels.
- **Model 4:** Increased precision by excluding the "Tags_Wrong Number Given" feature, optimizing the balance between sensitivity and specificity.

Building Models:

- **Model 5:** Improved precision and accuracy by eliminating the "Tags_Invalid Number" feature, achieving balanced performance metrics.
- **Model 6:** Refined the model by excluding the "Last Notable Activity_Had a Phone Conversation" feature, maintaining high accuracy and precision.
- **Model 7:** Enhanced specificity by dropping the "Tags_Switched Off" feature, optimizing the model for better classification of non-converted leads.
- **Model 8:** Further improved specificity by removing the "Last Notable Activity_Email Bounced" feature, refining the model's ability to identify non-converted leads accurately.
- **Model 9:** Achieved optimal balance between sensitivity and specificity by excluding the "Tags_Ringing" feature, maximizing precision and accuracy for lead classification.

Model no. 9 (Final Model)

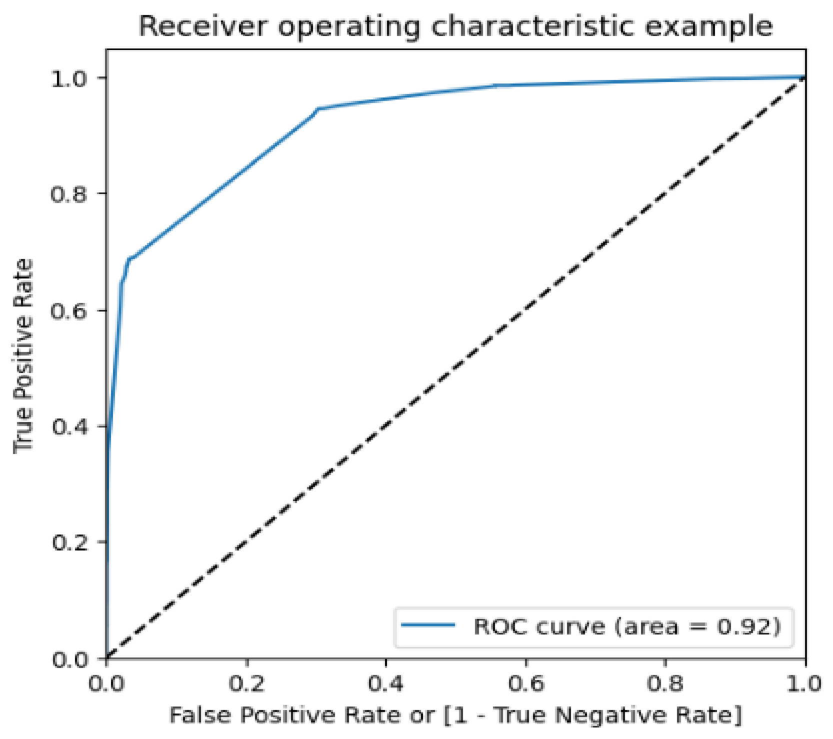
- **Features Included:** Lead Origin_Lead Add Form, Do Not Email_Yes, Last Activity_Converted to Lead, Last Activity_Olark Chat Conversation, What is your current occupation_Unemployed, What is your current occupation_Working Professional, Tags_Busy, Tags_Closed by Horizzon, Tags_Lost to EINS, Tags_Will revert after reading the email, Tags_in touch with EINS, Last Notable Activity_SMS Sent.
- **Accuracy:** 79.17%
- **Sensitivity:** 93.34%
- **Specificity:** 70.43%
- **Optimal Cutoff Threshold:** 0.38

Model no. 9 (Final Model)

Interpretation of Coefficients:

- **Lead Origin_Lead Add Form:** This coefficient indicates the impact of leads generated through the 'Lead Add Form' origin on the probability of conversion. A positive coefficient suggests that leads from this origin are more likely to convert.
- **Do Not Email_Yes:** A positive coefficient indicates that leads who opted not to receive emails are less likely to convert compared to those who opted to receive emails.
- **Last Activity_Converted to Lead:** This coefficient signifies the impact of leads who were last engaged in activities converting them to leads. A positive coefficient suggests that leads engaged in this activity are more likely to convert.
- **Last Activity_Olark Chat Conversation:** Positive coefficient indicates that leads engaged in Olark Chat Conversations are more likely to convert, as they are actively engaging with the platform.
- **What is your current occupation_Unemployed:** This coefficient indicates how being unemployed affects the likelihood of lead conversion. A positive coefficient suggests that unemployed leads are more likely to convert.
- **What is your current occupation_Working Professional:** A positive coefficient suggests that leads who are working professionals are more likely to convert compared to other occupations.
- **Tags_Busy:** This coefficient indicates that leads tagged as 'Busy' are more likely to convert, as they may have shown interest despite being occupied.
- **Tags_Closed by Horizzon:** Positive coefficient suggests that leads closed by 'Horizzon' are more likely to convert, indicating a high probability of successful closure.
- **Tags_Lost to EINS:** Leads tagged as 'Lost to EINS' have a positive coefficient, indicating that they are more likely to convert compared to other tags.
- **Tags_Will revert after reading the email:** Positive coefficient suggests that leads expected to revert after reading emails are more likely to convert, indicating a proactive response.
- **Tags_in touch with EINS:** Positive coefficient indicates that leads in touch with 'EINS' are more likely to convert, as they are actively engaged with the platform.
- **Last Notable Activity_SMS Sent:** This coefficient suggests that leads who received SMS notifications as their last notable activity are more likely to convert.

Plotting ROC Curve



ROC Curve value should be close to 1 and we are getting the value 0.92 which is close enough and its a good value indicating a good predictive model

Evaluating Model

Final Observations: Lets compare the values for Train and Test

- ***Train Data:***

- Accuracy : 79.17%
  Sensitivity : 93.34%
  Specificity : 70.43%

- ***Test Data:***

- Accuracy : 79.43%
  Sensitivity : 93.69%
  Specificity : 70.12%

Final Result

- The Model Seems to predict the conversion rate very well and we should be able to give the CEO confidence in making good calls based on this model

According to our final model, the following are predictor variables :

- Lead Origin_Lead Add Form
  Do Not Email_Yes
  Last Activity_Converted to Lead
  Last Activity_Olark Chat Conversation
  What is your current occupation_Unemployed
  What is your current occupation_Working Professional
  Tags_Busy
  Tags_Closed by Horizzon
  Tags_Lost to EINS
  Tags_Will revert after reading the email
  Tags_in touch with EINS
  Last Notable Activity_SMS Sent