

Lead Scoring Presentation

Introduction:-

Problem Statement

The problem statement revolves around X Education, an online education company, striving to enhance its lead conversion rate. Despite receiving a substantial number of leads, the company faces a challenge in efficiently identifying and converting potential leads into paying customers. Presently, only around 30% of leads result in conversion, leading to the need for a more targeted and effective approach to lead management.

The primary issue lies in the company's inability to distinguish between leads that are highly likely to convert and those with lower conversion potential. To address this, X Education aims to build a predictive model that assigns a lead score to each potential customer, indicating the likelihood of conversion. By leveraging data-driven insights and machine learning techniques, the company seeks to prioritize its sales efforts towards leads with the highest probability of conversion, thereby increasing the overall lead conversion rate.

The dataset provided contains various attributes related to leads, such as lead source, website activity, and past interactions. The target variable, 'Converted,' indicates whether a lead has been successfully converted or not.

In summary, the problem statement revolves around optimizing lead conversion processes at X Education through the development of a predictive model that assigns lead scores, ultimately leading to more targeted and efficient sales efforts and a higher overall conversion rate.

Project Objectives

- **Improve Lead Conversion Rate:** Enhance the efficiency and effectiveness of lead conversion processes to increase the overall conversion rate.
- **Identify Potential Leads:** Develop a model to accurately identify potential leads with a high likelihood of conversion, allowing the sales team to prioritize their efforts effectively.

- **Assign Lead Scores:** Implement a lead scoring system that assigns scores to leads based on their conversion probability, enabling the sales team to focus on leads with the highest potential.
- **Optimize Sales Efforts:** Streamline sales efforts by targeting leads more efficiently, reducing unnecessary outreach to leads with lower conversion potential.

Data Overview:-

Data Description

The dataset provided for the lead conversion optimization project at X Education consists of approximately 9000 data points containing various attributes related to leads and their interactions with the company's website and marketing channels. Below is a description of the dataset:

Lead Information:

Personal details of leads such as name, email address, and phone number may be included to identify individual leads.

Other demographic information such as age, gender, and occupation may also be present.

Lead Source:

Indicates the source from which the lead originated, such as organic search, paid advertisements, referrals, or direct traffic.

Website Engagement Metrics:

Total time spent on the website: The cumulative duration of a lead's interaction with the website.

Total visits: The number of times a lead visited the website.

Page views: The number of pages viewed by the lead during their visit.

Lead Activity:

Last activity: The last interaction or activity performed by the lead on the website, such as filling out a form, downloading content, or watching videos.

Last activity date: The timestamp of the last recorded activity by the lead.

Marketing Interactions:

Marketing channels: Information about the marketing channels through which leads were acquired, such as email campaigns, social media platforms, or online advertisements.

Lead Conversion Status:

Target variable: 'Converted' indicates whether a lead was successfully converted into a paying customer (1 for converted, 0 for not converted).

Additional Attributes:

Other attributes may include lead scoring parameters, campaign details, and any specific actions taken by leads during their interactions with the company's marketing materials.

Categorical Variables:

The dataset may contain categorical variables with multiple levels, such as lead source categories or marketing channel names.

Missing Values and 'Select' Levels:

The dataset may have missing values that need to be handled appropriately during preprocessing.

Some categorical variables may contain a level labeled as 'Select,' which represents a null or unspecified value and needs to be addressed.

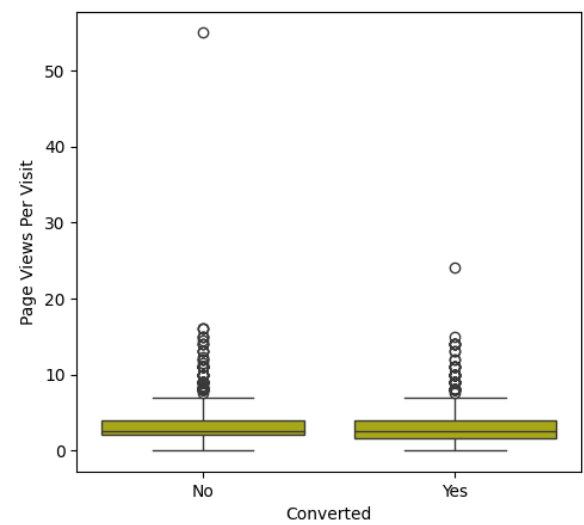
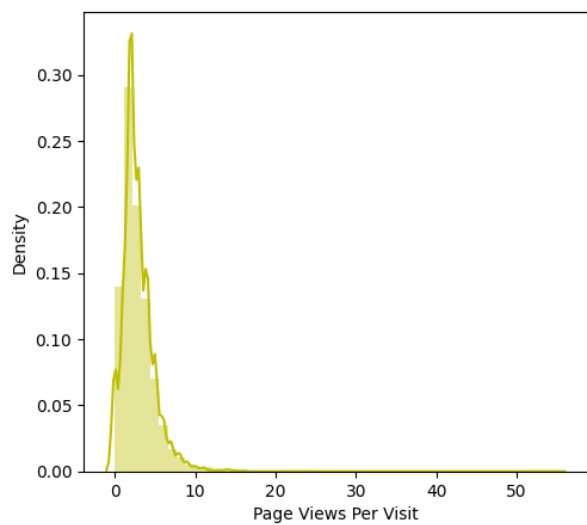
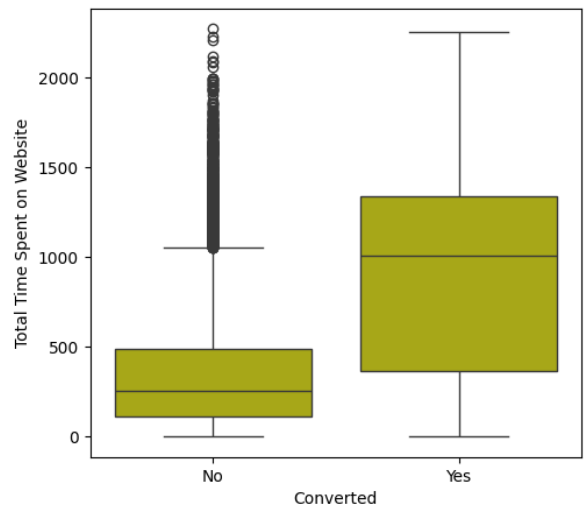
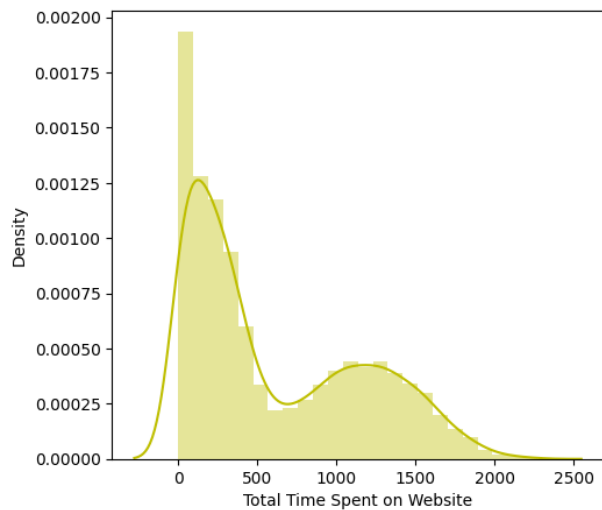
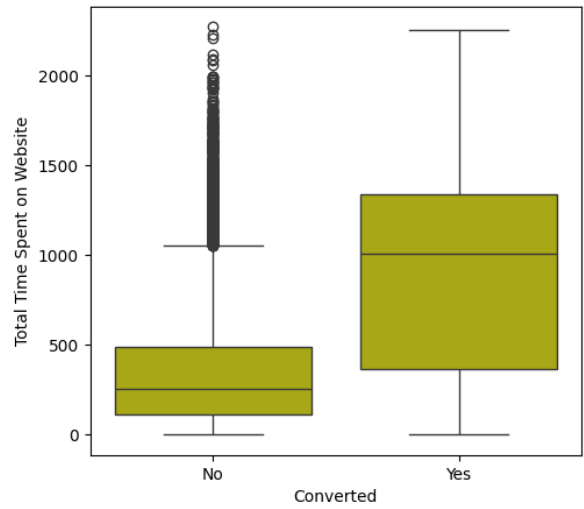
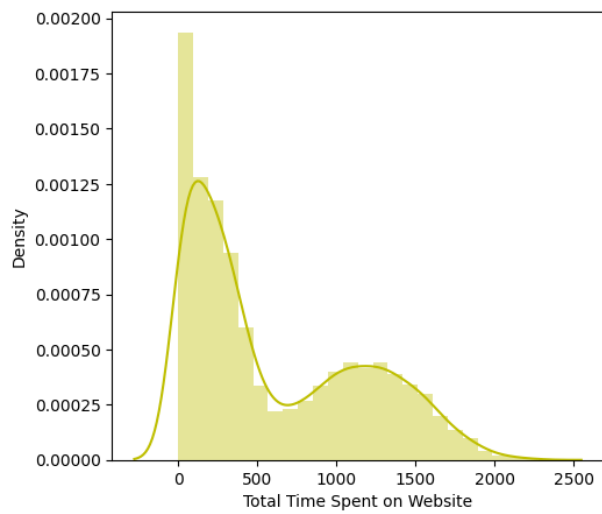
Overall, the dataset provides valuable insights into the behavior and characteristics of leads interacting with X Education's online platform and marketing efforts. Analyzing this data can help identify patterns, trends, and predictors of lead conversion, ultimately informing the development of a predictive model to optimize lead conversion processes.

Statistics

Some key statistics of the behavior of the various feature variables with the target variable “Converted”.



Distribution and the box-plots of some of the feature variables that underwent outlier treatment by IQR Method



All the rows with 'Select' values were either dropped or imputed with a different value according to its composition within the column. This category is to be dropped as it is essentially a dummy/null category which shows up while selecting the categories from a drop-down menu.

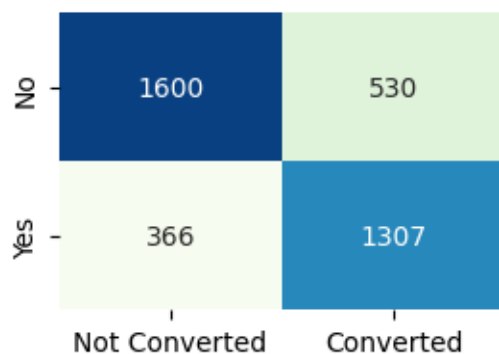
I have dropped features that do not provide meaningful information like Prospect ID, Lead Number, Do Not Email...etc.

Also, I have replaced categories having less count with 'Others'.

I have also imputed features which have less missing values with the mode of that column.

With the remaining columns I used Recursive Feature Elimination to eliminate redundant columns.

Confusion Matrix



	Not Converted	Converted
No	1600	530
Yes	366	1307

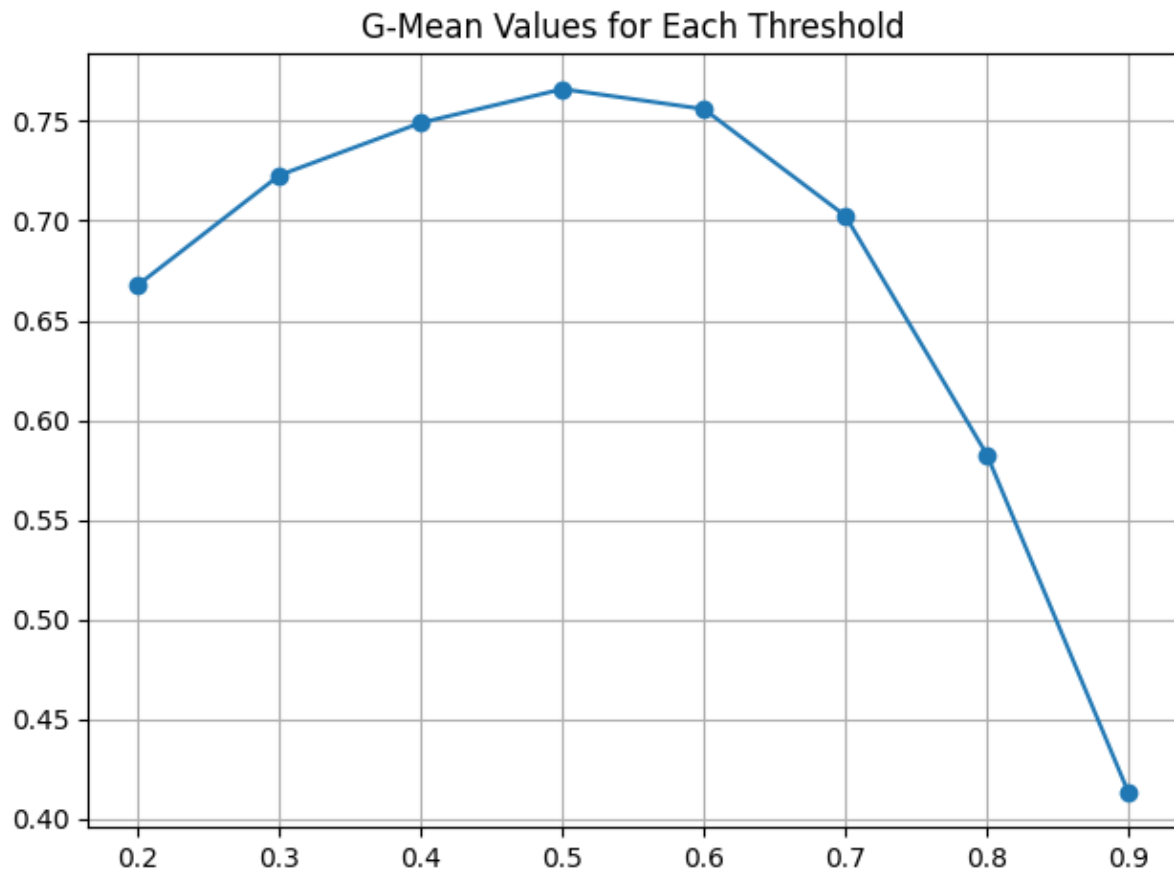
True Positives (TP): There are 1,307 instances where the model correctly predicted 'Yes' when the actual outcome was 'Yes.' These represent successful predictions of positive cases.

False Positives (FP): There are 530 instances where the model incorrectly predicted 'Yes' when the actual outcome was 'No.' These are cases where the model falsely identified a negative case as positive.

False Negatives (FN): There are 366 instances where the model incorrectly predicted 'No' when the actual outcome was 'Yes.' These are cases where the model failed to identify a positive case.

True Negatives (TN): There are 1,600 instances where the model correctly predicted 'No' when the actual outcome was 'No.' These represent successful predictions of negative cases.

Geometric Mean

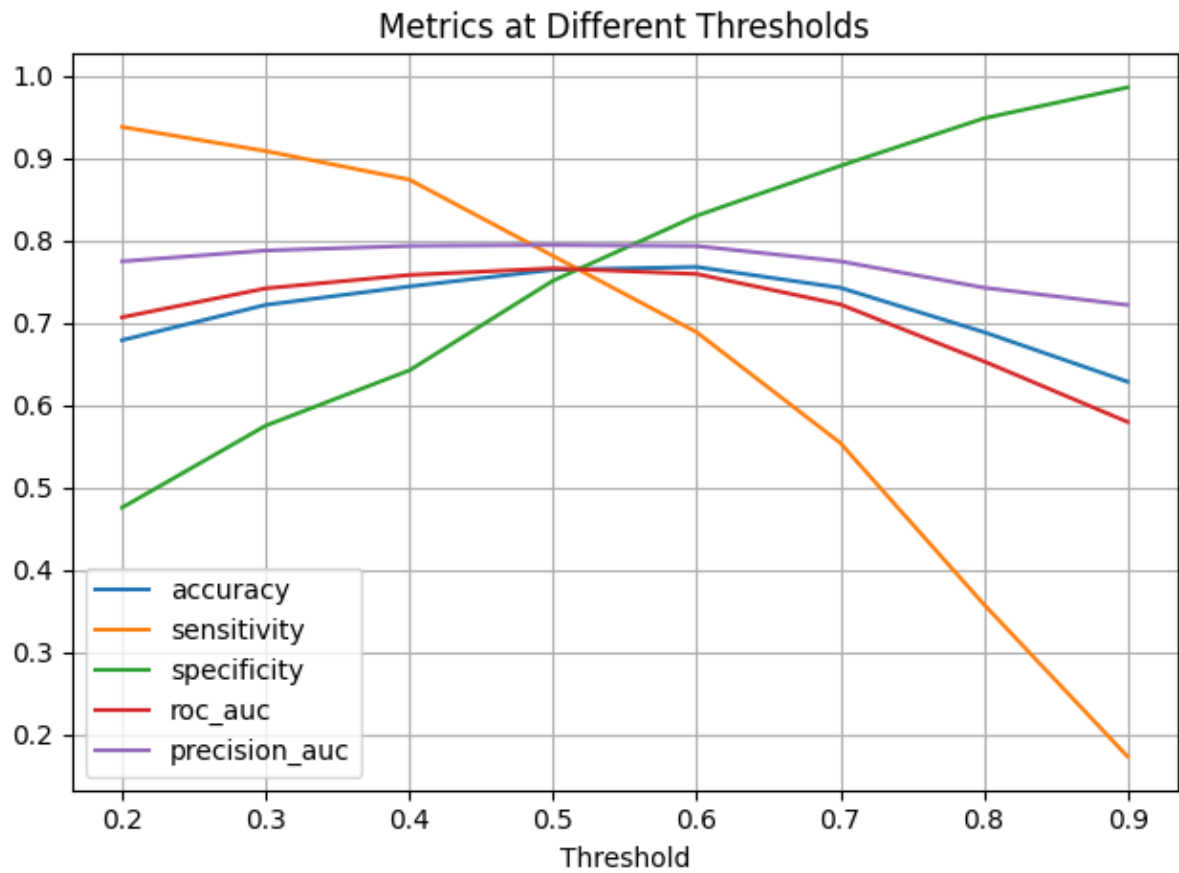


The geometric mean is a metric used to evaluate the performance of a binary classification model across different thresholds for class probability cutoffs. It combines sensitivity (true positive rate) and specificity (true negative rate) into a single measure, making it useful for imbalanced datasets where one class dominates the other.

If the G-Mean values increase slightly and then steeply decrease as the threshold changes, it suggests that there is an optimal threshold where the model achieves a balance between sensitivity and specificity. Initially, as the threshold increases, sensitivity typically decreases while specificity increases, leading to a slight increase in G-Mean. However, beyond a certain threshold, increasing the cutoff further leads to a rapid decline in sensitivity, resulting in a sharp decrease in G-Mean.

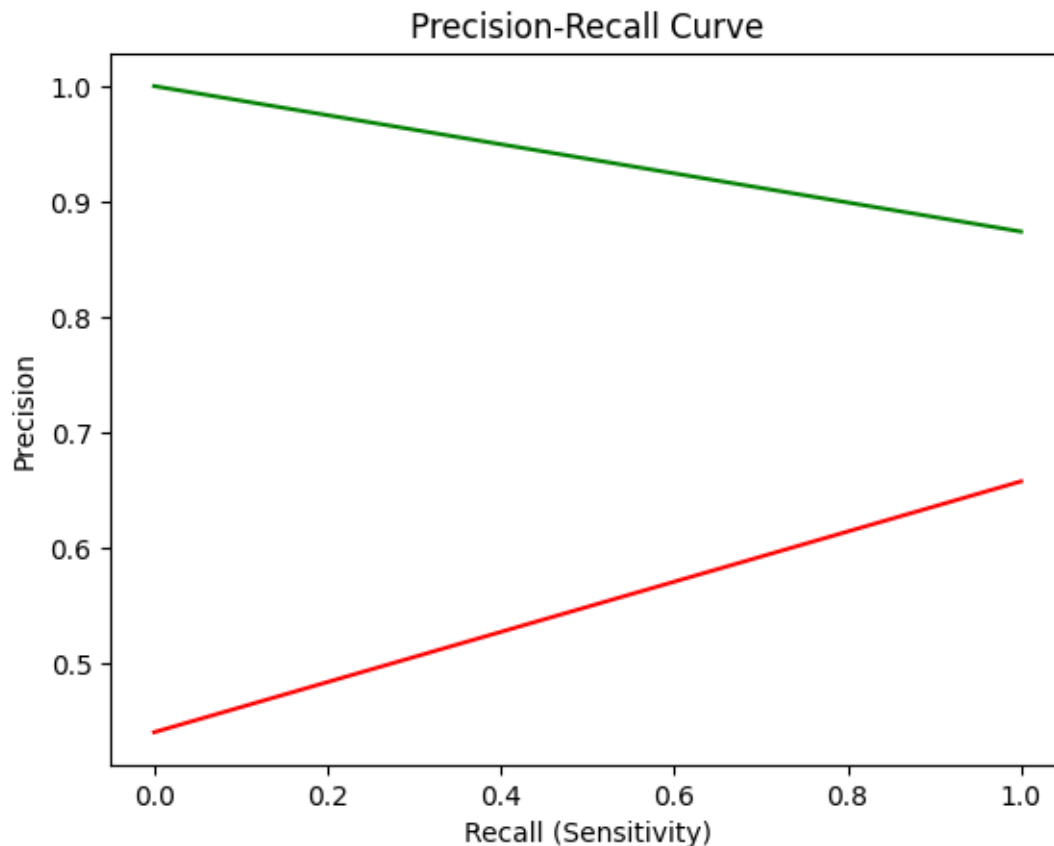
This phenomenon indicates a trade-off between sensitivity and specificity. A threshold too low may result in high sensitivity but low specificity, leading to a high rate of false positives. Conversely, a threshold too high may yield high specificity but low sensitivity, resulting in a high rate of false negatives. The optimal threshold strikes a balance between these competing objectives, maximizing the overall performance of the classification model.

Metrics at different thresholds



	accuracy	sensitivity	specificity	roc_auc	gmeans	precision_auc
0.2	0.678938	0.937836	0.475587	0.706712	0.667849	0.774662
0.3	0.721799	0.908548	0.575117	0.741832	0.722856	0.787792
0.4	0.744149	0.873879	0.642254	0.758066	0.749168	0.793368
0.5	0.764397	0.781231	0.751174	0.766203	0.766055	0.794479
0.6	0.767815	0.688583	0.830047	0.759315	0.756014	0.793239
0.7	0.742572	0.554094	0.890610	0.722352	0.702483	0.774697
0.8	0.688667	0.358039	0.948357	0.653198	0.582708	0.742650
0.9	0.628451	0.173341	0.985915	0.579628	0.413400	0.721626

Precision-Recall Curve



The precision-recall curve is slowly converging. This indicates that the model's performance is improving gradually as you change the threshold for classifying positive instances.

Business Insights

Top three variables contributing most towards lead conversion probability:

Feature	Coefficient
Lead Origin_Lead Add Form	3.688055
TotalTimeSpentonWebsite_bins_minute	2.430137
Lead Source_Olark Chat	1.297931

The business should focus on these variables.

Potential leads which originate through Add Forms should be prioritized.

People who spend a lot of time on the website should be prioritized.

Leads whose source are Olark Chat should be prioritized.