X Education - Lead Scoring Case Study

Detection of Hot Leads to concentrate more of marketing efforts on them, improving c onversion rates for X Education

Team Members: Ramya, Vekata Rama Krishna , Preethi

Table of Contents

- Background of X Education Company
- Problem Statement & Objective of the Study
- Suggested Ideas for Lead Conversion
- Analysis Approach
- Data Cleaning
- EDA
- Data Preparation
- Model Building (RFE & Manual fine tuning)
- Model Evaluation
- Recommendations

Background of X Education Company

- An education company named X Education sells online courses to industry professionals.
- On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%.

Problem Statement & Objective of the Study

Problem Statement

- X Education, despite generating a substantial number of leads, is grappling with a low lead conversion rate of approximately **30%**. This inefficiency in the conversion process is a significant concern for the organization.
- The primary objective of X Education is to enhance the efficiency of the lead conversion process. The strategy to achieve this involves identifying and focusing on the most promising leads, referred to as **Hot Leads**.
- The sales team at X Education is keen on knowing these potential leads. The identification of these leads will allow the team to concentrate their communication efforts on this select group, rather than reaching out to everyone. This targeted approach is expected to improve the lead conversion rate significantly.

Objective of the Study

- The study aims to assist X Education in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company requires a model that assigns a lead score to each lead. The customers with a higher lead score should have a higher conversion chance, and the customers with a lower lead score should have a lower conversion chance.
- The CEO has given a ballpark of the target lead conversion rate to be around **80%**. This model is expected to help achieve this target by enabling a more focused and efficient lead conversion process.

Suggested Ideas for Lead Conversion

Leads Grouping

- Leads are grouped based on their propensity or likelihood to convert.
- This results in a focused group of hot leads.

Better Communication

- We could have a smaller pool of leads to communicate with, which would allow us to have a greater impact.

Boost Conversion

- We would have a greater conversion rate and be able to hit the 80% objective since we concentrated on hot leads that were more likely to convert.

Since we have a target of 80% conversion rate, we would want to obtain a high **sensitivity** in obtaining hot leads.

Analysis Approach

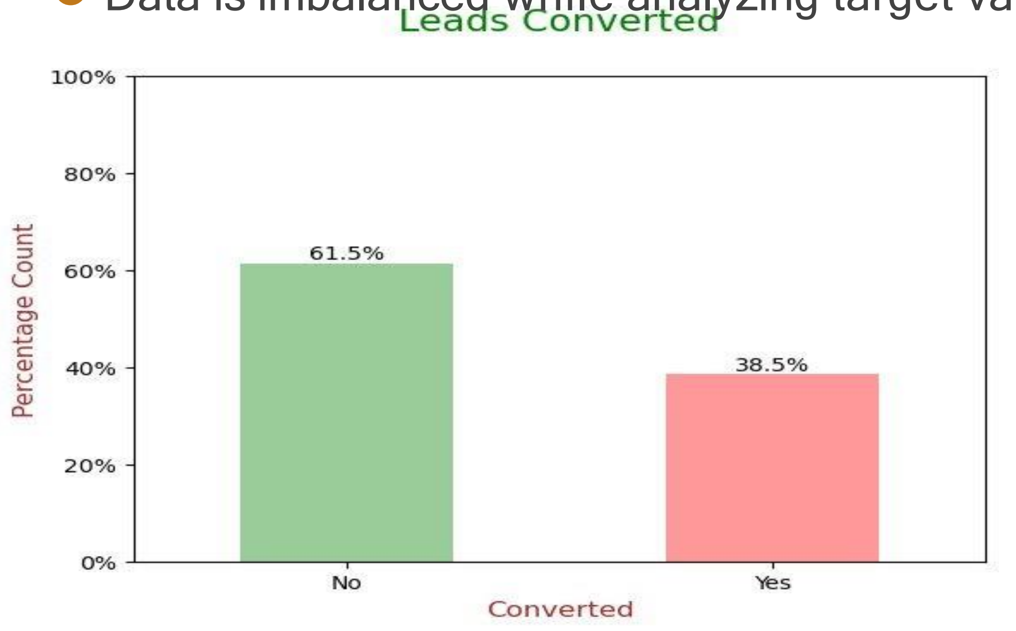
The approach to analyze this problem can be broken down into the following steps:

- **Data Understanding**: The first step is to understand the data provided by X Education. This includes understanding the variables, their significance, and the distribution of values.
- **Data Cleaning**: This step involves cleaning the data by handling missing values, outliers, and incorrect data entries.
- **Exploratory Data Analysis (EDA)**: In this step, we will perform a univariate and bivariate analysis. We will try to identify patterns, trends, and correlations in the data that might influence the lead conversion rate.
- **Feature Engineering**: Based on the EDA, we will create new features that might help improve the performance of the model. This could involve creating binary flags, bucketing variables, creating interaction terms, etc.
- **Model Building**: We will use a suitable machine learning algorithm to build a predictive model. The choice of model will depend on the nature of our data and our target variable. We will also ensure that the model does not overfit or underfit the data.
- **Model Evaluation**: We will evaluate the model using appropriate metrics that align with the business objective. In this case, since we are interested in the probability of conversion, metrics like Area Under the ROC curve (AUC-ROC) could be used.
- **Lead Scoring**: Based on the model's predictions, we will assign a lead score to each lead. The leads with higher scores will be the ones that the sales team should focus on.
- **Model Validation**: We will validate the model using a different dataset to ensure that it generalizes well to unseen data.
- **Implementation and Monitoring**: Post-validation, the model will be implemented and its performance will be monitored over time. Necessary adjustments will be made based on the feedback from the sales team and the model's performance.

This approach will help X Education in identifying the most promising leads and achieving their target lead conversion rate of around 80%.

EDA

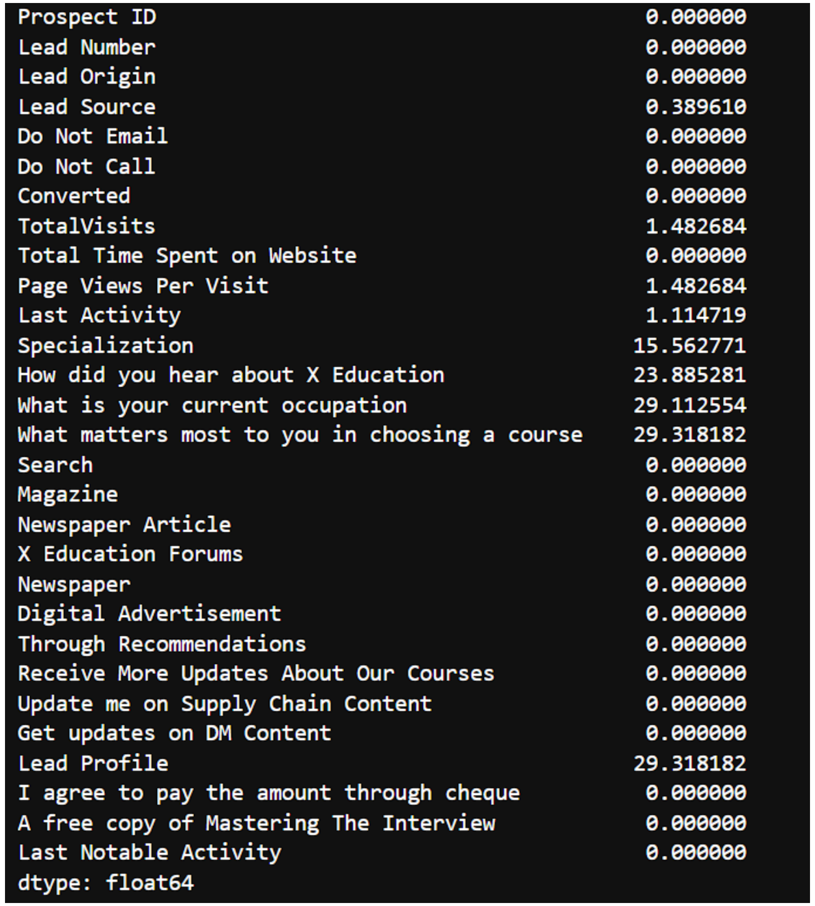
- Data is imbalanced while analyzing target variable.



- Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)

- While 61.5% of the people didn't convert to leads. (Majority)

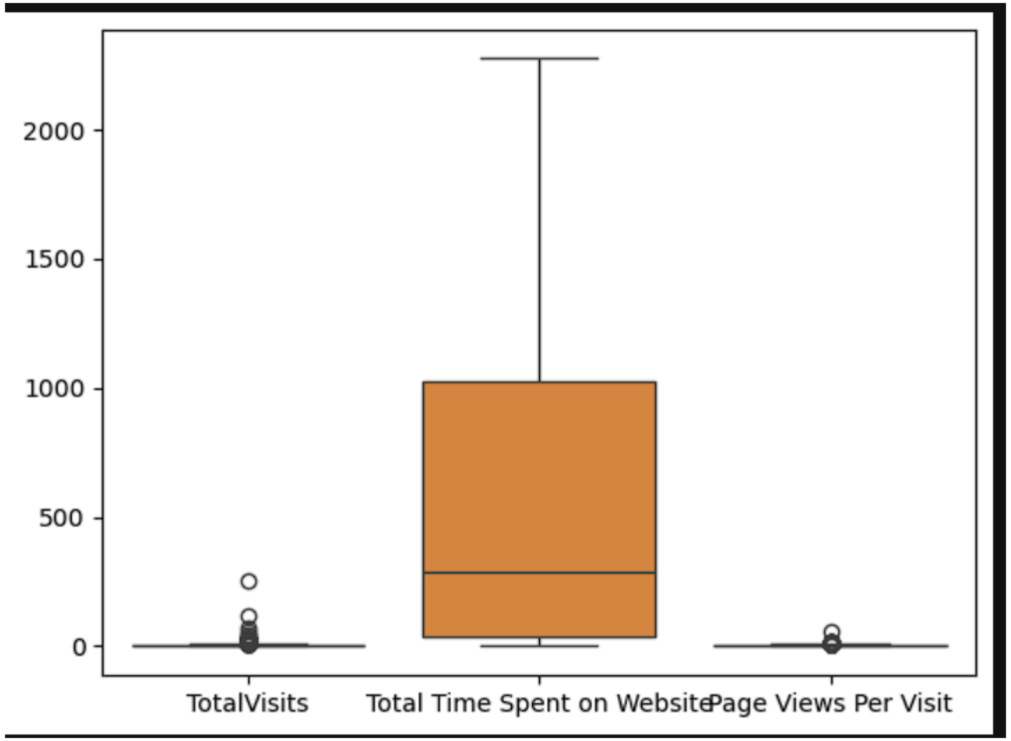
- Data have null vlaues which need to taken care



```
Prospect ID                                      0.000000
Lead Number                                      0.000000
Lead Origin                                      0.000000
Lead Source                                      0.389610
Do Not Email                                     0.000000
Do Not Call                                      0.000000
Converted                                        0.000000
TotalVisits                                      1.482684
Total Time Spent on Website                      0.000000
Page Views Per Visit                             1.482684
Last Activity                                    1.114719
Specialization                                  15.562771
How did you hear about X Education              23.885281
What is your current occupation                 29.112554
What matters most to you in choosing a course   29.318182
Search                                           0.000000
Magazine                                         0.000000
Newspaper Article                                0.000000
X Education Forums                               0.000000
Newspaper                                        0.000000
Digital Advertisement                            0.000000
Through Recommendations                          0.000000
Receive More Updates About Our Courses           0.000000
Update me on Supply Chain Content                0.000000
Get updates on DM Content                        0.000000
Lead Profile                                    29.318182
I agree to pay the amount through cheque         0.000000
A free copy of Mastering The Interview           0.000000
Last Notable Activity                            0.000000
dtype: float64
```

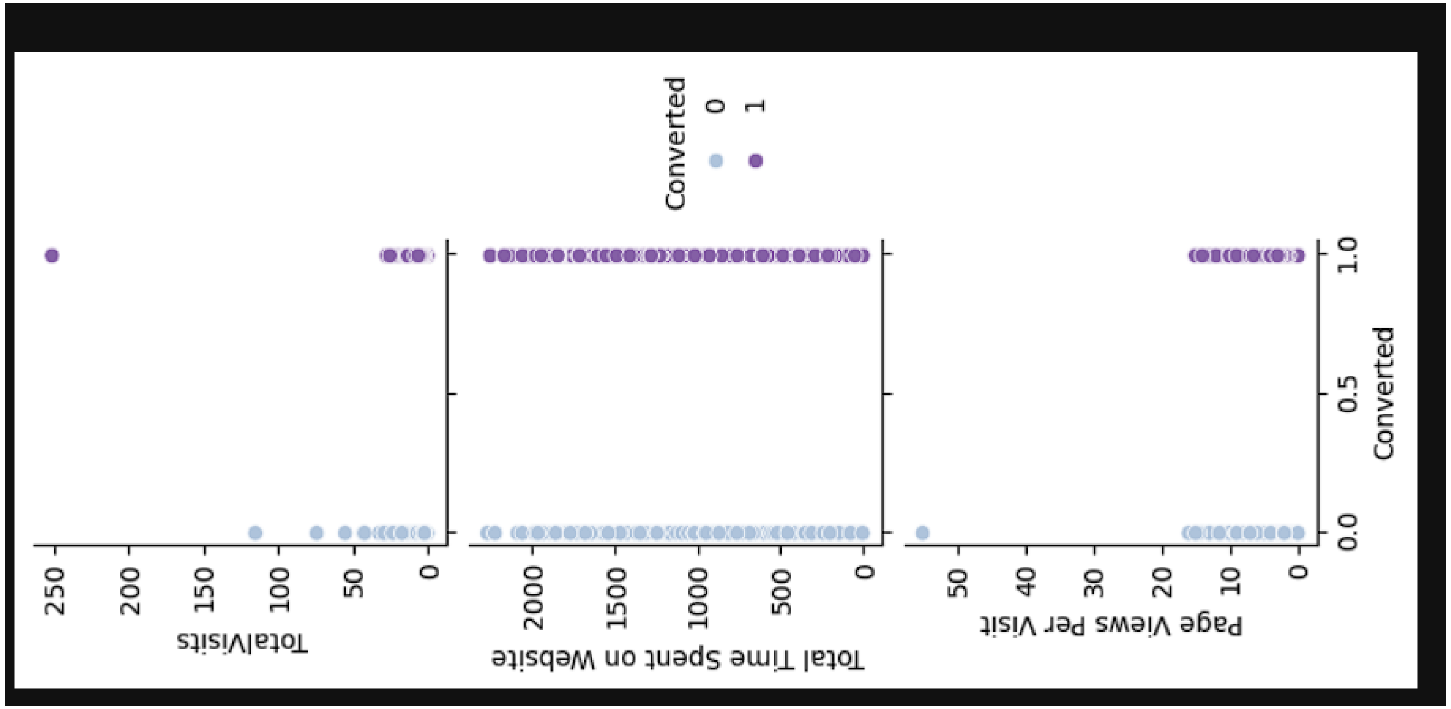
EDA- Data Cleaning

- Outliers



- **There is no Much Outliers**

Comparing the Continous variables vs Target Variable



Data Preparation before Model building

- Binary level categorical columns were already mapped to 1 / 0 in previous steps
- Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation
- Splitting Train & Test Sets
    - 70:30 % ratio was chosen for the split
- Feature scaling
    - Standardization method was used to scale the features
- Checking the correlations
    - Predictor variables which were highly correlated with each other were dropped (Lead Origin_Lead Import and Lead Origin_Lead Add Form).

Model Building

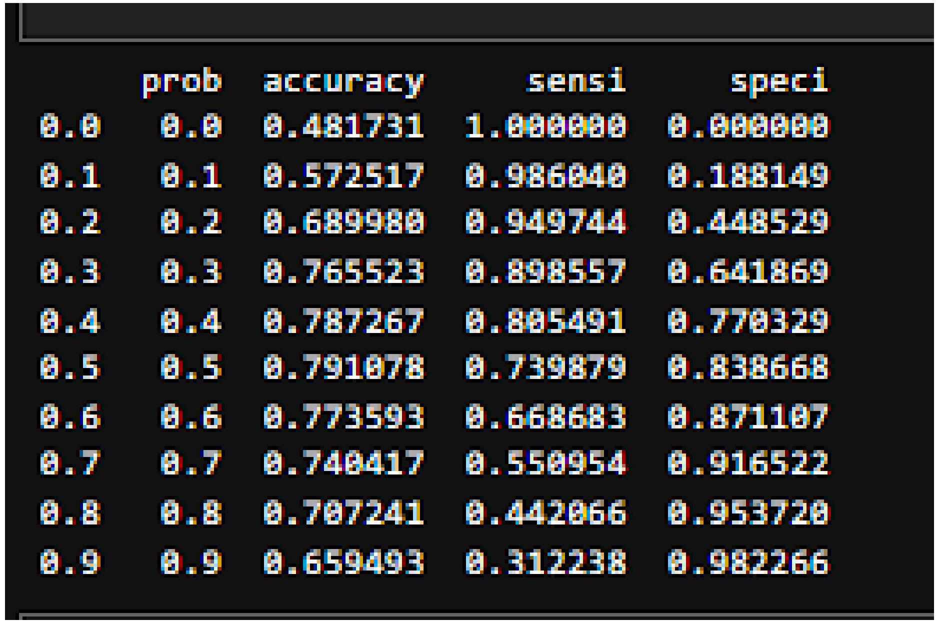
- **Feature Selection**
- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform **Recursive Feature Elimination** (RFE) and to select only the important columns.
- Then we can manually fine tune the model.
- RFE outcome
  - Pre RFE – 48 columns & Post RFE – 10 columns

Model Building

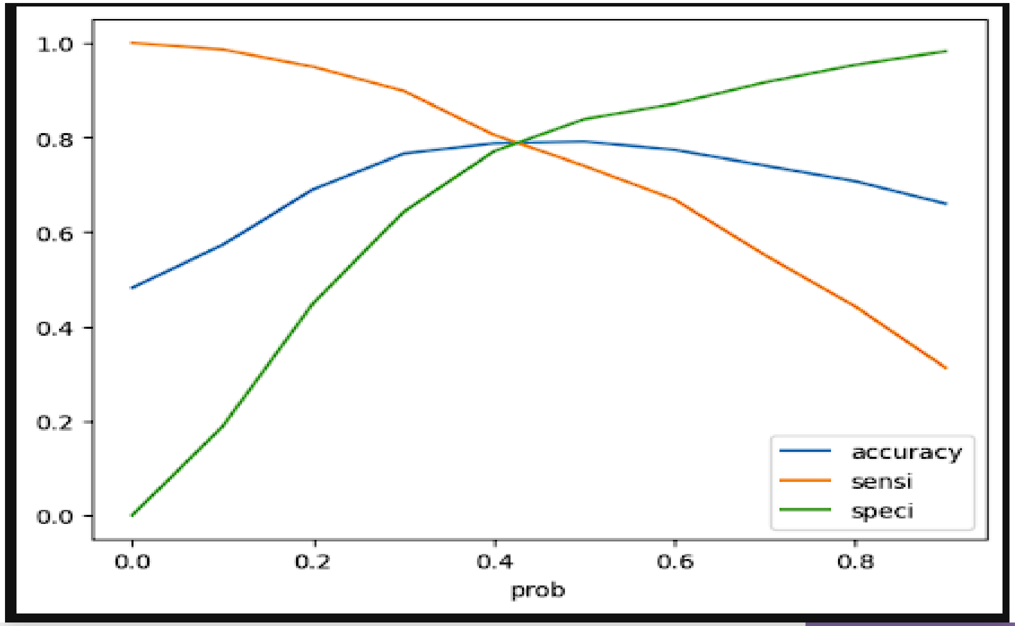
- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Model 4 looks stable after four iteration with:
  - significant p-values within the threshold (p-values < 0.05) and
  - No sign of multicollinearity with VIFs less than 5
- Hence, **logm4** will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

Confusion Matrix & Evaluation Metrics with 0.353 as cutoff

Model Evaluation
Train Data Set

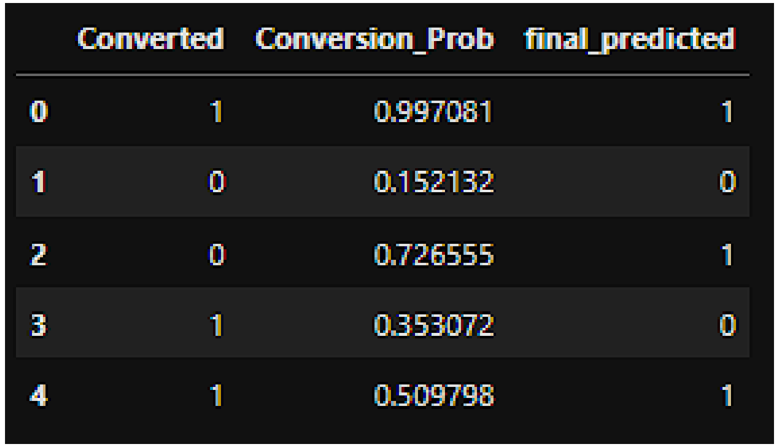


```
     prob  accuracy     sensi     speci
0.0   0.0  0.481731  1.000000  0.000000
0.1   0.1  0.572517  0.986040  0.188149
0.2   0.2  0.689980  0.949744  0.448529
0.3   0.3  0.765523  0.898557  0.641869
0.4   0.4  0.787267  0.805491  0.770329
0.5   0.5  0.791078  0.739879  0.838668
0.6   0.6  0.773593  0.668683  0.871107
0.7   0.7  0.740417  0.550954  0.916522
0.8   0.8  0.707241  0.442066  0.953720
0.9   0.9  0.659493  0.312238  0.982266
```

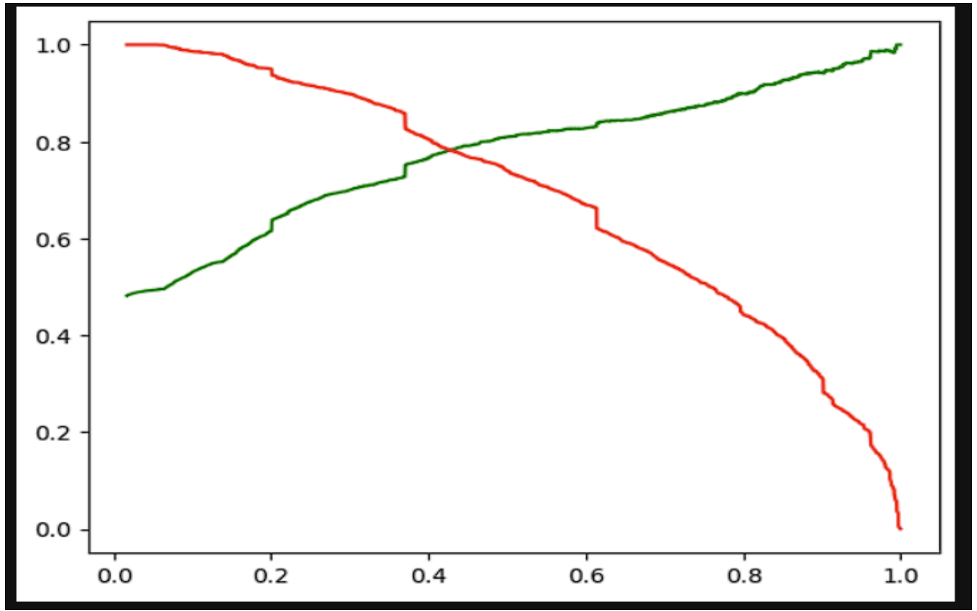


Confusion Matrix & Evaluation Metrics with 0.4 as cutoff

Model Evaluation
Train Data Set



	Converted	Conversion_Prob	final_predicted
0	1	0.997081	1
1	0	0.152132	0
2	0	0.726555	1
3	1	0.353072	0
4	1	0.509798	1

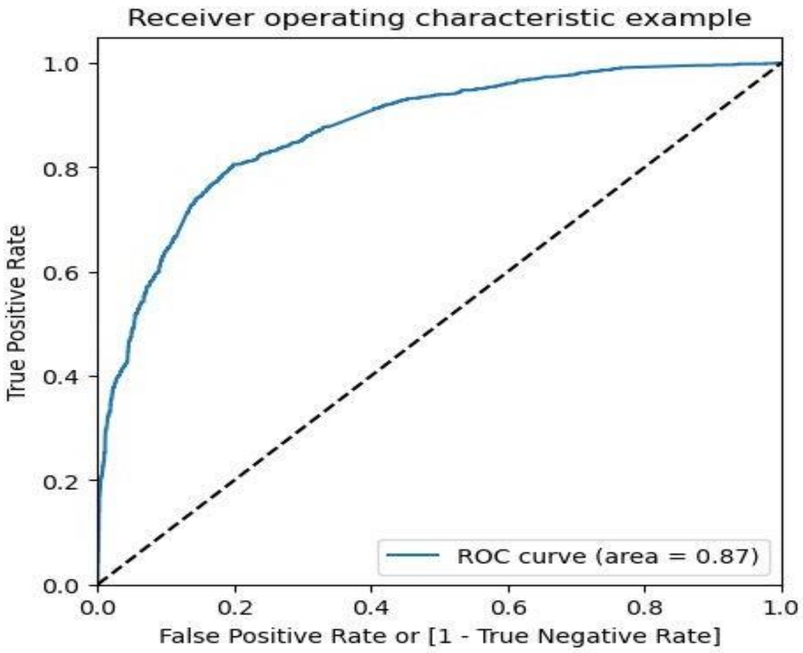


Model Evaluation

Area under ROC curve (AUC): The value of 0.88 out of 1 indicates a good predictive model. A higher AUC suggests better performance in distinguishing between positive and negative instances.

Curve position: The curve is positioned close to the top left corner of the plot. This placement signifies a model with:

1. High true positive rate (sensitivity) at all threshold values.
2. Low false positive rate (1-specificity) at all threshold values.



1.Area under ROC curve (AUC): The value of 0.87 out of 1 indicates a good predictive model. A higher AUC suggests better performance in distinguishing between positive and negative instances.

2.Curve position: The curve is positioned close to the top left corner of the plot. This placement signifies a model with:

1. High true positive rate (sensitivity) at all threshold values.
2. Low false positive rate (1-specificity) at all threshold values.

Recommendation based on Final Model

As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.

- **Last Activity_SMS Sent (Coefficient: 1.52)**: This feature has the highest positive coefficient. It suggests that leads who have received SMS messages are more likely to convert.
- **Lead Origin_Lead Add Form (Coefficient: 1.51)**: Leads originating from the “Lead Add Form” have a strong positive impact on conversion. Focusing on this source can improve lead conversion rates.
- **Last Notable Activity_Modified (Coefficient: 1.50)**: Leads with a “Modified” last notable activity are positively associated with conversion. Pay attention to follow-up actions after initial contact.
- **Lead Source_Olark Chat (Coefficient: 1.34)**: Leads coming from the “Olark Chat” source have a significant positive effect on conversion. Engaging with chat leads effectively can boost conversions.
- **Lead Source_Welingak Website (Coefficient: 1.30)**: The “Welingak Website” as a lead source is also impactful. Prioritize efforts related to this source.
- **Total Time Spent on Website (Coefficient: 1.24)**: The time spent by leads on the website positively influences conversion. Enhance website engagement and user experience.
- **Do Not Email_Yes (Coefficient: 1.09)**: Leads who prefer not to receive emails still contribute positively to conversion. Respect their communication preferences.
- **What is your current occupation_Student (Coefficient: 1.04)**: Students show a slight positive impact on conversion. Tailor marketing strategies accordingly.
- **Last Activity_Had a Phone Conversation (Coefficient: 1.01)**: Although a small effect, leads with phone conversations have a positive association with conversion.
- **Last Notable Activity_Unreachable (Coefficient: 1.01)**: Similarly, “Unreachable” last notable activity has a minor positive impact.

Recommendation based on Final Model

To Increase Lead Conversion Rates:

Focus on Features with Positive Coefficients:

- Prioritize marketing efforts around features that have positive coefficients in the model.
- These features contribute significantly to lead conversion. Tailor your strategies accordingly.

Attract High-Quality Leads from Top-Performing Sources:

- Identify the lead sources that consistently yield high conversion rates.
- Allocate resources to attract more leads from these sources.

Optimize Communication Channels:

- Analyze which communication channels (email, SMS, phone calls) have the most impact on lead engagement.
- Optimize your approach based on this analysis.

Incentivize Referrals:

- Encourage existing leads to refer others.
- Offer incentives or discounts for successful referrals that convert into leads.

Target Working Professionals Aggressively:

- Working professionals tend to have higher conversion rates.
- Tailor marketing efforts to address their needs and financial situations.

Areas of Improvement:

Analyse Negative Coefficients in Specialization Offerings:

- Investigate features related to specialization offerings that have negative coefficients.
- Understand why these features negatively impact conversion and explore ways to improve them.

Review Landing Page Submission Process:

- Evaluate the user experience during the landing page submission process.
- Identify any friction points or areas where potential leads drop off.
- Optimize the submission process for better conversion rates.

Thank You!