LEAD SCORING CASE STUDY

Team Members: Kshitij Bhardwaj, S C Lalith Shekar , Leena Rikhai

TABLE OF CONTENTS

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 gets a lot of leads, its lead conversion rate is very poor at around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
- Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

Objective of the Study:

- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

ANALYSIS APPROACH

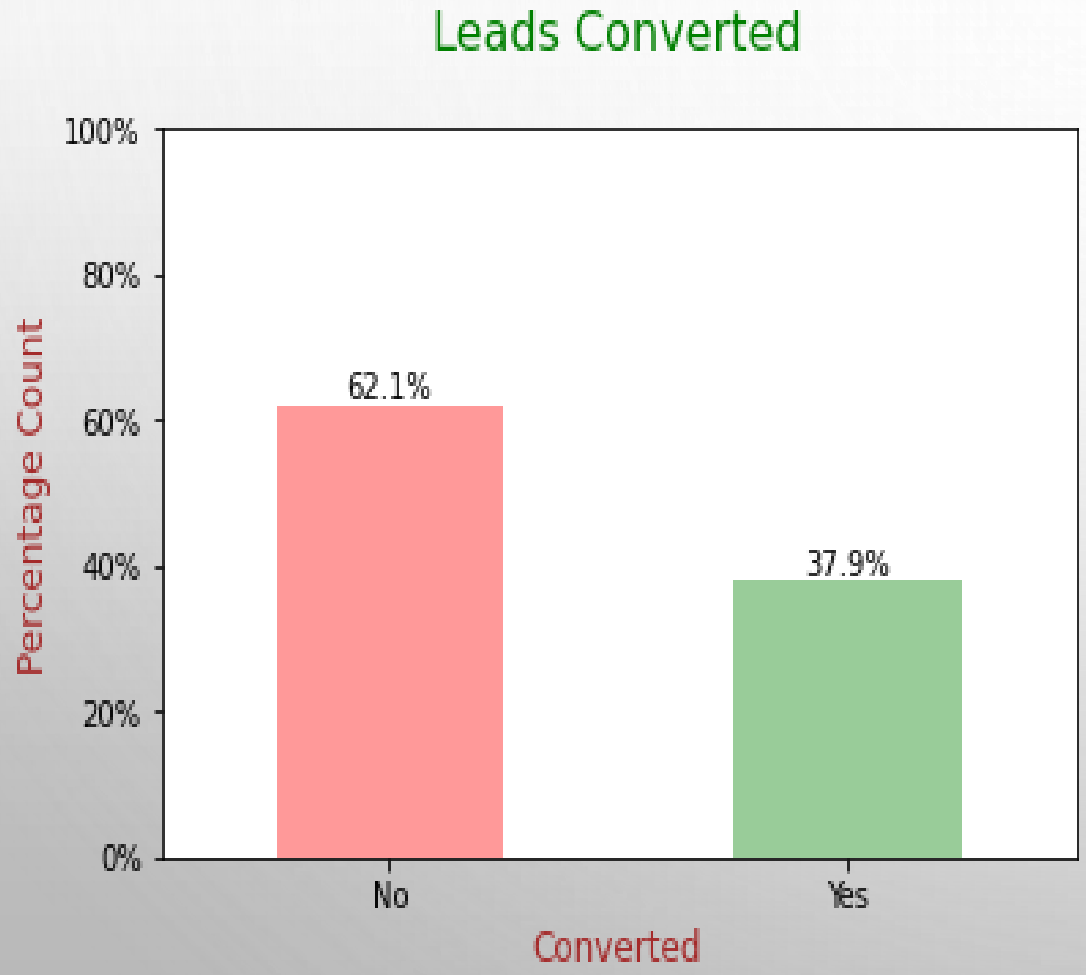
- **DATA CLEANING:**
- LOADING DATA SET, UNDERSTANDING & CLEANING DATA
- **EDA:**
- CHECK IMBALANCE, UNIVARIATE & BIVARIATE ANALYSIS
- **DATA PREPARATION**
- DUMMY VARIABLES, TEST-TRAIN SPLIT, FEATURE SCALING
- **MODEL BUILDING:**
- RFE FOR TOP 15 FEATURE
- **MODEL EVALUATION:**
- CONFUSION MATRIX, CUTOFF SELECTION, ASSIGNING LEAD SCORE
- **PREDICTIONS ON TEST DATA:**
- COMPARE TRAIN VS TEST METRICS, ASSIGN LEAD SCORE AND GET TOP FEATURES

DATA CLEANING

- **"Select"** level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that don't add any insight or value to the study objective
- Imputation was used for some categorical variables.
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
- Numerical data was imputed with mode after checking distribution.

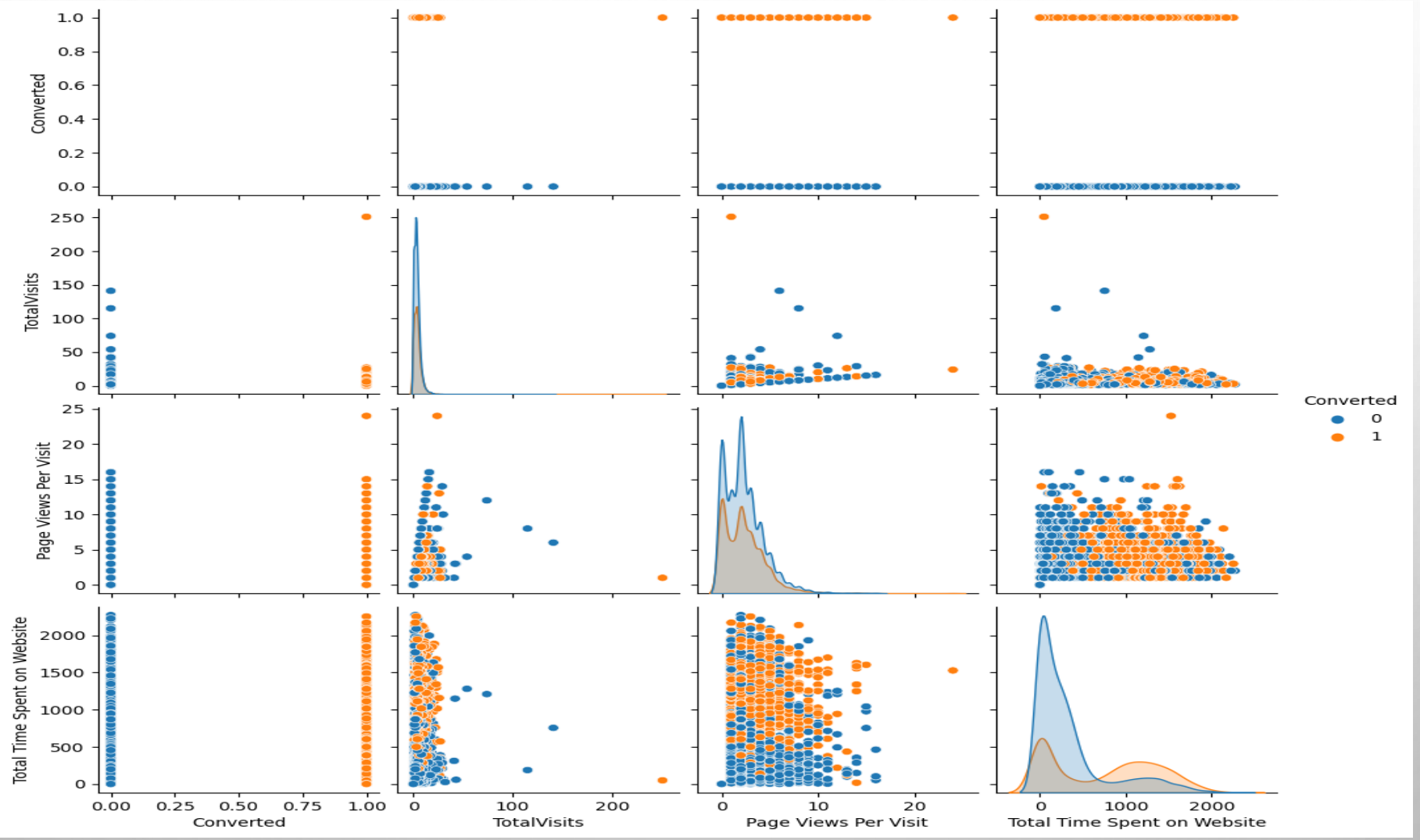
EDA

- Data is imbalanced while analyzing target variable.

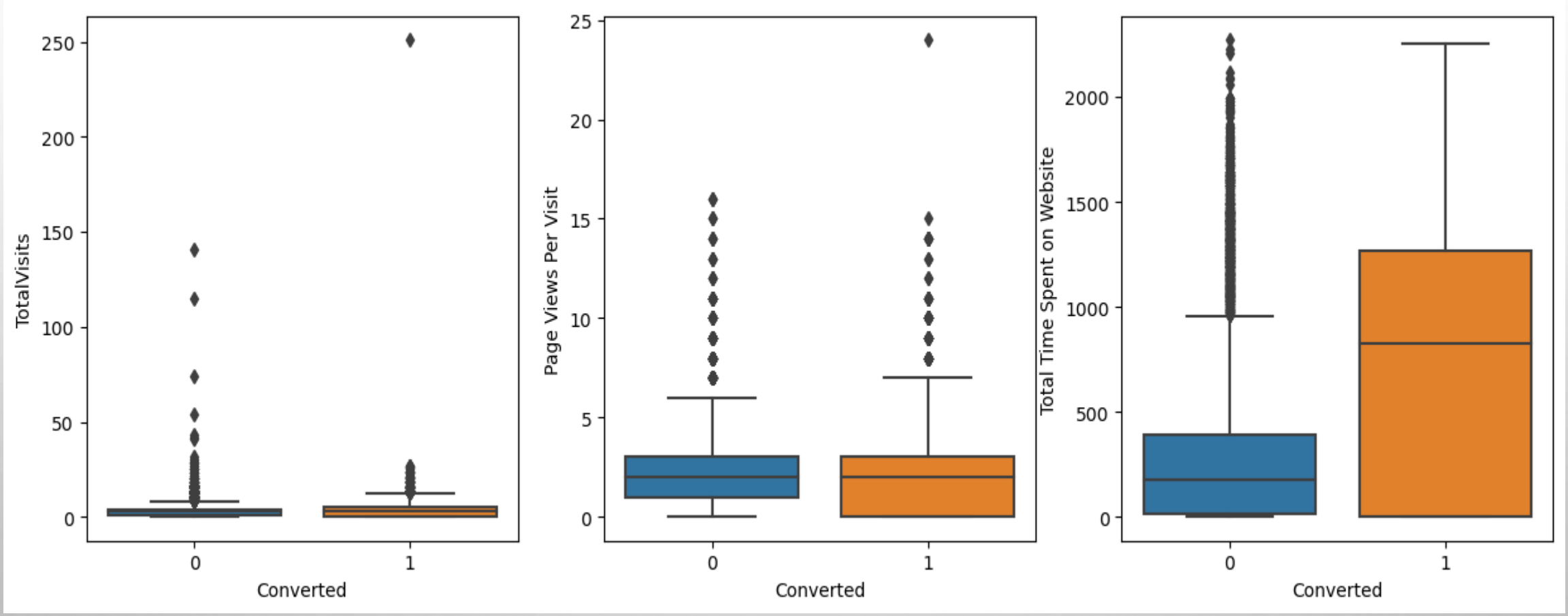


- Conversion rate is of 37.9%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 62.1% of the people didn't convert to leads. (Majority)

EDA – BIVARIATE ANALYSIS FOR NUMERICAL VARIABLES

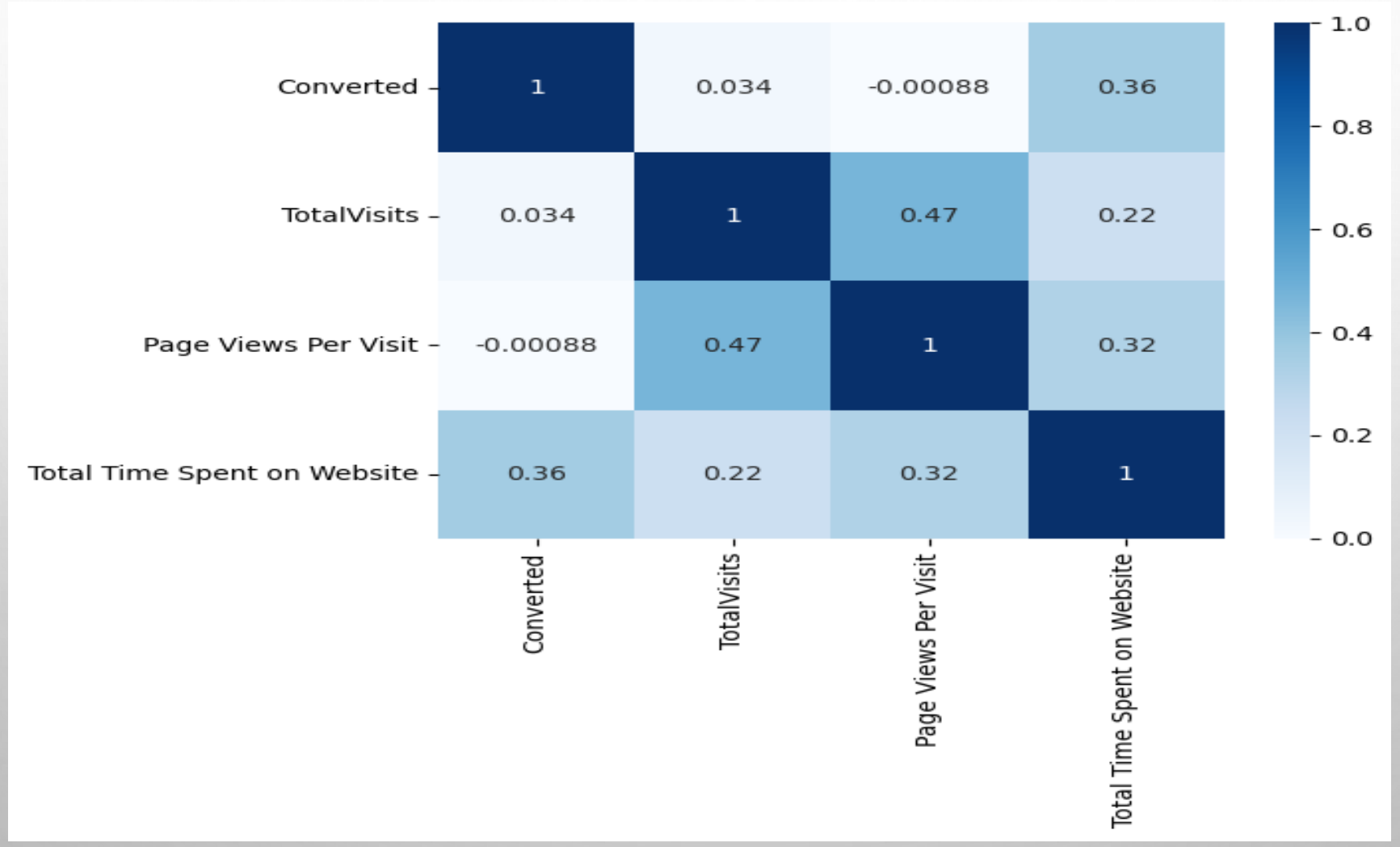


EDA – BIVARIATE ANALYSIS FOR NUMERICAL VARIABLES



- Boxplot with Converted as hue

CORRELATION BETWEEN NUMERICAL VARIABLES



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.
- After analyzing the dataset, we started the work on a Logistic Regression Model for binary classification, as it showed promise in handling the target variable, "Converted." We divided the data into training and testing sets and trained the model on the former. To assess its performance, we evaluated sensitivity, specificity, and other relevant metrics.
- Hence it is important to perform **Recursive Feature Elimination** (RFE) and to select only the important columns.

MODEL BUILDING

- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Model 5 looks stable after five iteration with:
  - significant p-values within the threshold (p-values < 0.05) and
  - No sign of multicollinearity with VIFs less than 5
- Hence, **logm5** will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

MAKING PREDICTIONS

Predicted Probabilities on Train Set

	Actual_Churn_Flag	Predicted_Probability
7113	1	0.994111
4749	0	0.039261
7986	0	0.039261
1281	1	0.996947
7346	1	0.996765

Creating new column 'predicted' with
1 if Churn_Prob > 0.5 else 0

	Actual_Churn_Flag	Predicted_Probability	predicted
7113	1	0.994111	1
4749	0	0.039261	0
7986	0	0.039261	0
1281	1	0.996947	1
7346	1	0.996765	1

Appending y_test_df and y_prediction

	Converted	LeadID	0
0	1	4703	0.987342
1	0	5544	0.377657
2	0	5520	0.072540
3	1	1342	0.001074
4	0	4101	0.017074

MODEL EVALUATION

Train Dataset

```
Accuracy: 0.9258384506376949
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.96      0.94      3936
           1       0.93      0.87      0.90      2415

    accuracy                           0.93      6351
   macro avg       0.93      0.91      0.92      6351
weighted avg       0.93      0.93      0.93      6351

Confusion Matrix:
 [[3790  146]
 [ 325 2090]]
```

Sensitivity (Recall): 0.865424430641822

Specificity: 0.9629065040650406

False Positive Rate (FPR): 0.03709349593495935

Positive Predictive Value (Precision): 0.9347048300536672

Negative Predictive Value (NPV): 0.9210206561360875

Test Dataset

```
Accuracy: 0.9213813372520205
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.92      0.94      1702
           1       0.87      0.93      0.90      1020

    accuracy                           0.92      2722
   macro avg       0.91      0.92      0.92      2722
weighted avg       0.92      0.92      0.92      2722

Confusion Matrix:
 [[1564  138]
 [  76  944]]
```

Sensitivity (Recall): 0.9254901960784314

Specificity: 0.918918918918919

False Positive Rate (FPR): 0.08108108108108109

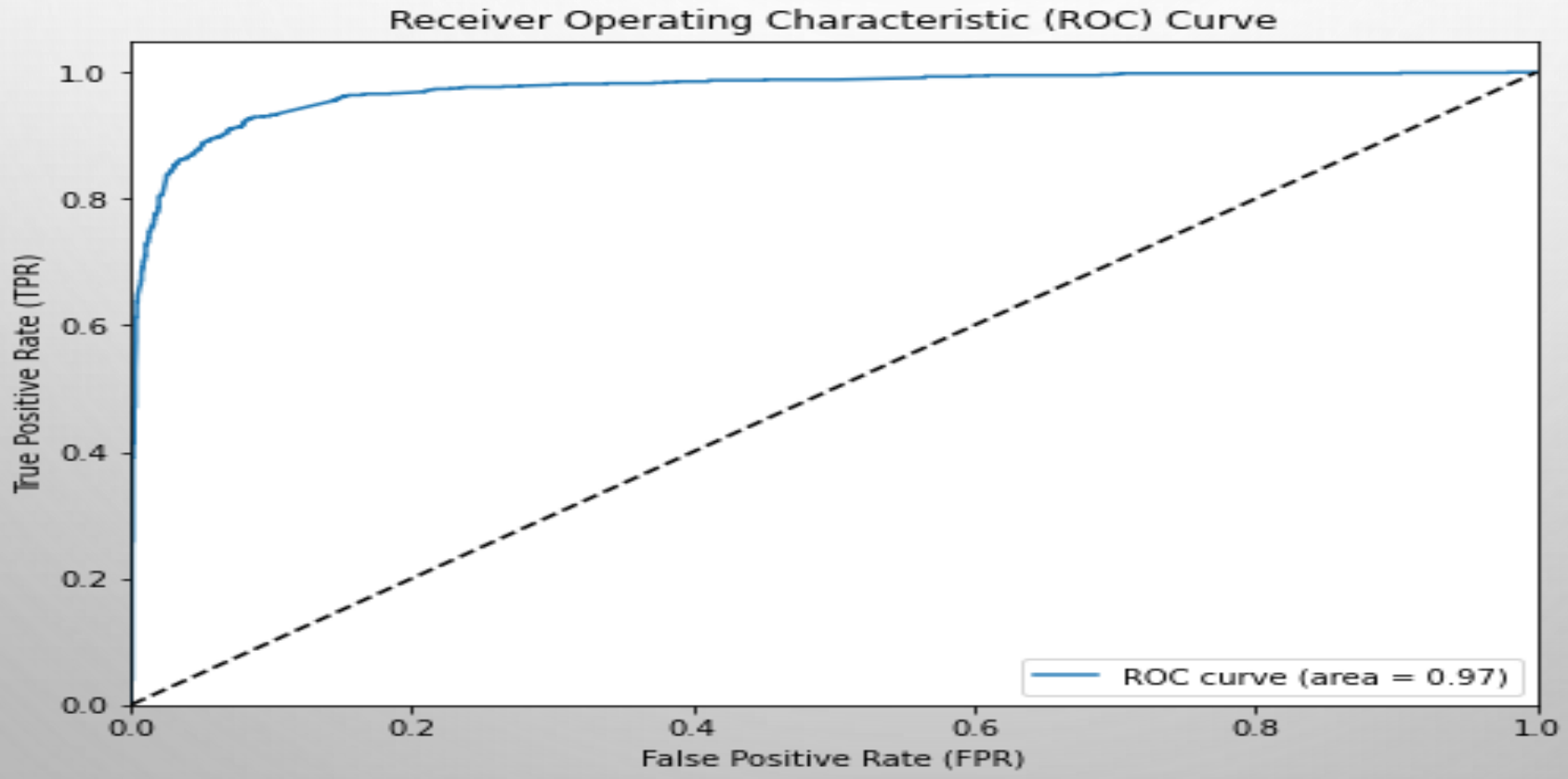
Positive Predictive Value (Precision): 0.8724584103512015

Negative Predictive Value (NPV): 0.9536585365853658

MODEL EVALUATION

ROC Curve

Area under ROC curve is 0.97 out of 1 which indicates a good predictive model.



RECOMMENDATION BASED ON FINAL MODEL

Strategies for Aggressive Lead Conversion during Intern Hiring: During the period of intern hiring, when aggressive lead conversion was the objective, we proposed the following strategies:

1. Prioritize Potential Leads: Identify leads with a high probability of conversion and concentrate efforts on them.
2. Automate Communication: Use automated marketing tools to reach out to a broader audience without compromising personalization.
3. Segment and Personalize Messages: Tailor communication based on leads' preferences, interests, and buying behavior.
4. Train the Sales Team: Provide the sales team with specialized training to handle increased lead interactions efficiently.

RECOMMENDATION BASED ON FINAL MODEL

Strategies to Minimize Useless Phone Calls during Target Achievement: During times when the company had achieved its targets ahead of schedule and aimed to minimize unnecessary phone calls, we recommended the following strategies:

1. Focus on Qualified Leads: Prioritize leads that have shown strong interest and engagement during previous interactions.
2. Use Marketing Automation: Implement marketing automation to deliver relevant and timely content to potential customers.
3. Set Qualification Criteria: Establish clear qualification criteria to filter out leads with a low probability of conversion.
4. Provide Valuable Content: Share valuable resources with leads to keep them engaged and interested in the company's offerings.
5. Stay Engaged through Updates: Keep leads informed about updates and new developments without the need for direct phone calls.

CONCLUSION

Our team's hardwork and continuous learning in building the lead conversion model provided valuable insights into the power of data-driven decision-making and personalized communication. By implementing the recommended strategies during different business phases, X Education can improve its sales and marketing approaches and achieve higher lead conversion rates.

THANK YOU!