LEAD SCORING CASE STUDY

TEAM MEMBERS

Subha R

Subha Meenakshy C

Sujan Kumar Kummara

CONTENTS

PROBLEM STATEMENT

- X Education, an online course provider, currently experiences a lead conversion rate of only **30%.**
- Many professionals visit the website, fill out forms, and are categorized as leads, but only a small portion convert into customers.
- The sales team currently engages with all acquired leads equally, leading to inefficient use of resources.
- To improve conversion rates, X Education needs to identify and prioritize **Hot Leads**—those most likely to convert.
- Focusing efforts on high-potential leads will improve conversion rates, reduce wasted outreach, and increase overall sales efficiency.

BUSINESS OBJECTIVE

- **Objective:** X Education seeks a model that assigns a lead score between 0 and 100 to each lead, helping them identify Hot Leads and enhance conversion rates.
- **Goal:** The CEO aims to achieve a lead conversion rate of 80%.
- **Future Considerations:** The model should be capable of addressing future challenges, such as:
  - **Peak Time Management:** Defining actions needed during high-demand periods to optimize operations.
  - **Maximizing Resource Utilization:** Ensuring full manpower is efficiently utilized to target the right leads.
  - **Post-Target Strategy:** Outlining approaches to maintain and grow conversions after the target is achieved.

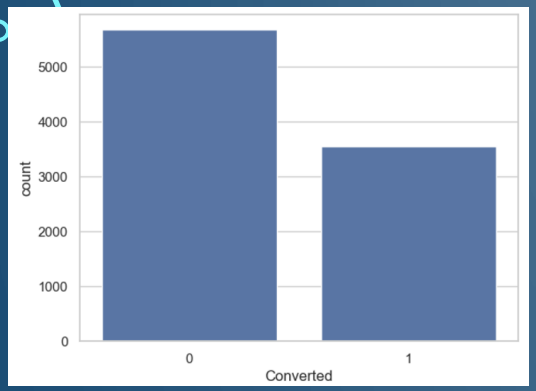
PROBLEM APPROACH

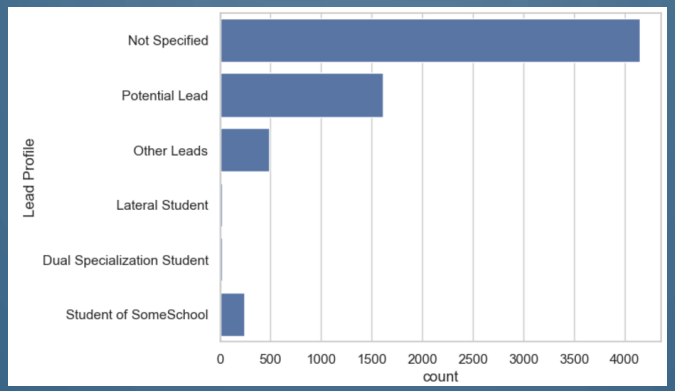
1. Importing inspecting the data
2. Data preparation
3. EDA
4. Dummy variable creation
5. Test-Train split
6. Feature scaling
7. Correlations
8. Model Building (RFE, R^2, VIF, and p-values)
9. Model Evaluation
10. Making predictions on test set

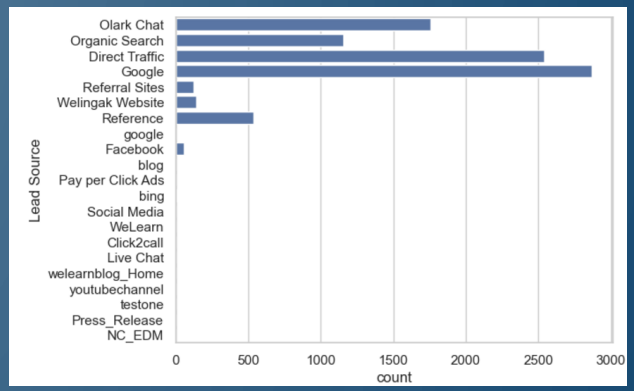
DATA PRE-PROCESSING

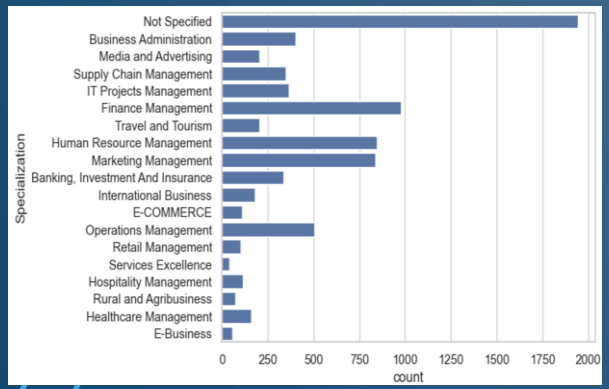
- **Dataset Overview:** The dataset consists of 37 rows and 9,240 columns.
- **Dropped Features:** Removed single-value features such as “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”, “Chain Content”, “Get updates on DM Content”, and “I agree to pay the amount through cheque”.
- **Irrelevant Identifiers:** Excluded “Prospect ID” and “Lead Number” as they are not necessary for the analysis.
- **Low-Variance Features:** Dropped features with insufficient variance after reviewing value counts for object-type variables, including “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, and “Digital Advertisement”.
- **Missing Data Handling:** Removed columns with over 35% missing values, such as 'How did you hear about X Education' and 'Lead Profile'.

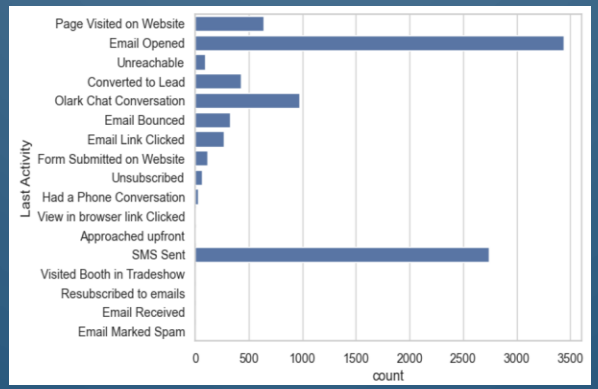
EDA – UNIVARIATE ANALYSIS

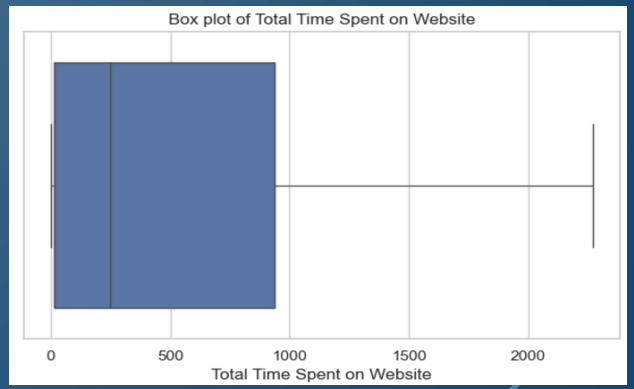




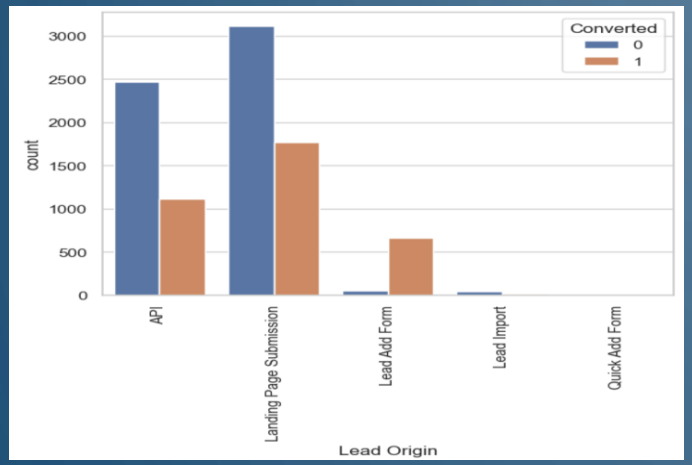


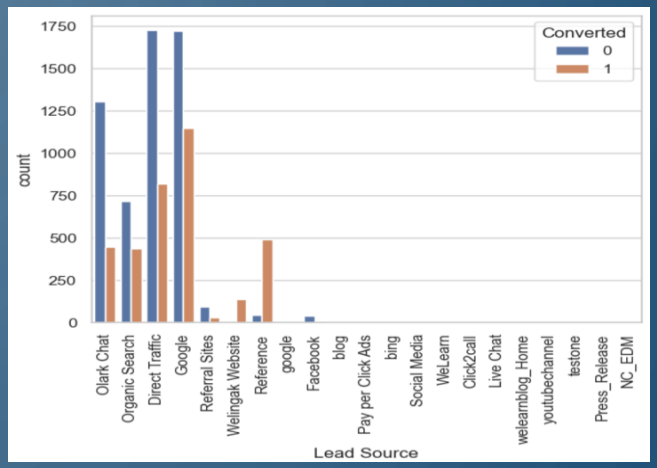


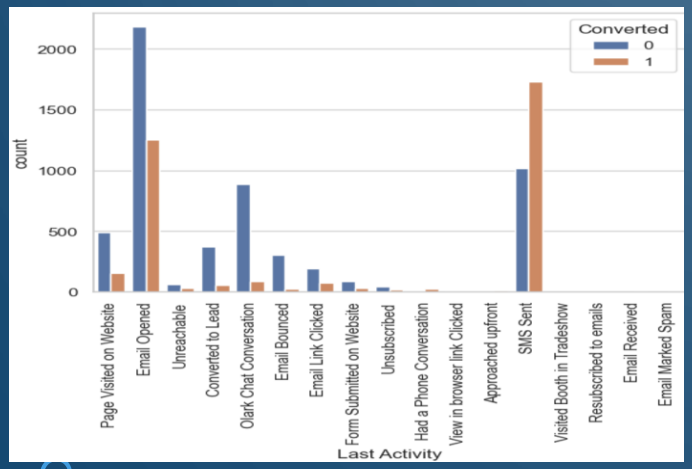


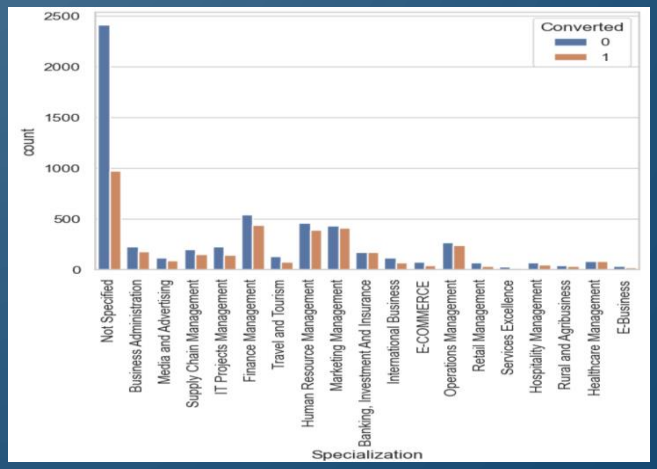


EDA – BIVARIATE ANALYSIS

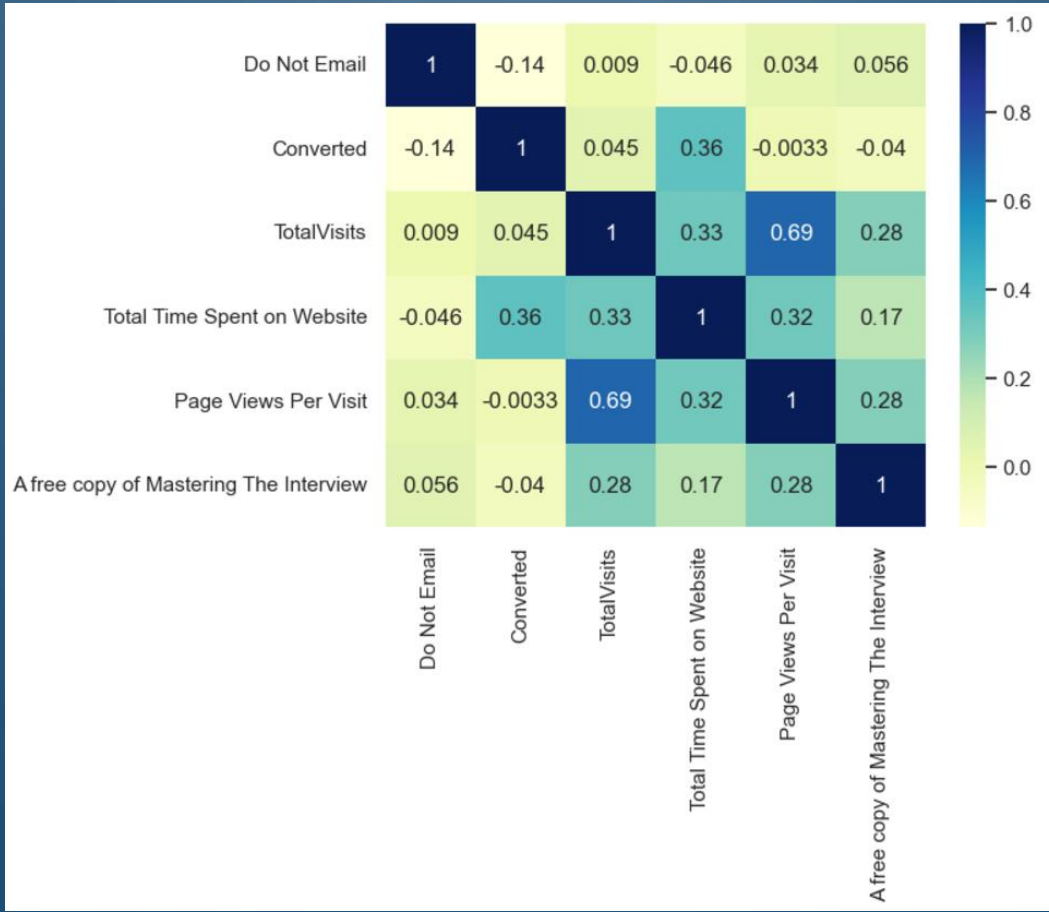








EDA - CORRELATION



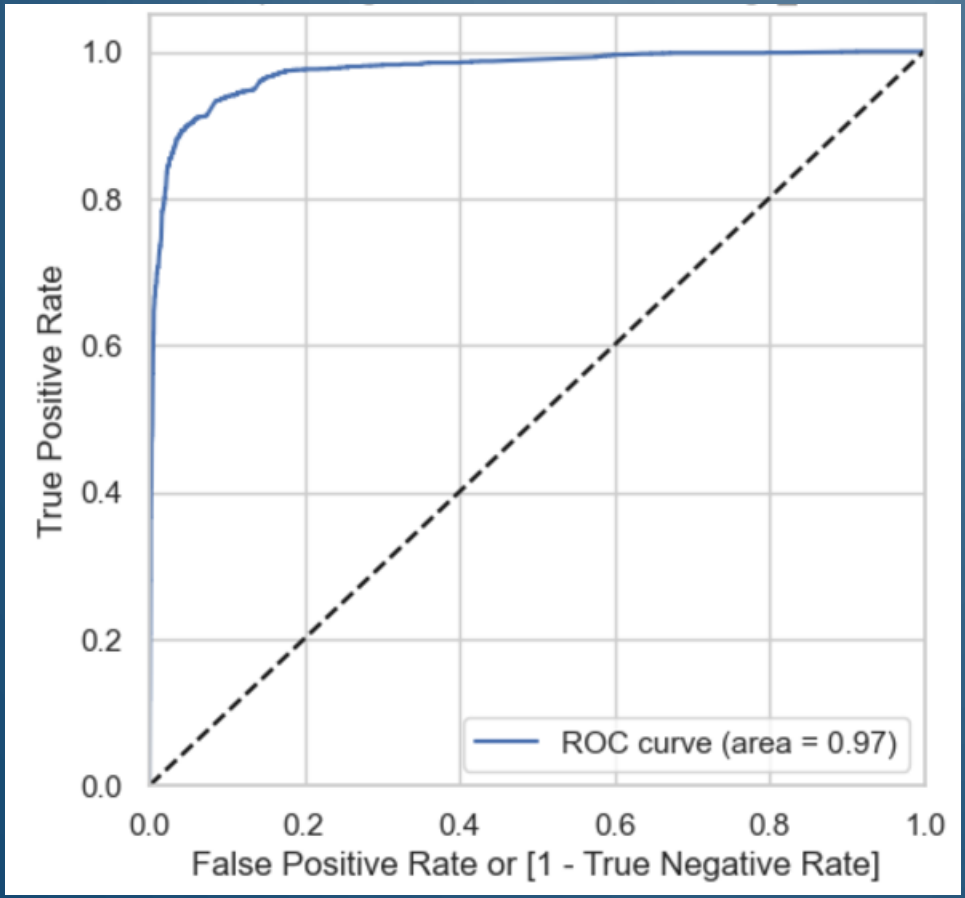
SUMMARY OF EDA

- Conversion Rate: 38.5% leads converted.
- Total Time Spent on Website shows a strong positive correlation with conversions.
- Lead Source impacts conversion likelihood (e.g., Google and Direct Traffic).
- Last Notable Activity is a significant predictor of conversions.
- Class imbalance observed: Majority leads are not converted (~61.5%).

MODEL BUILDING

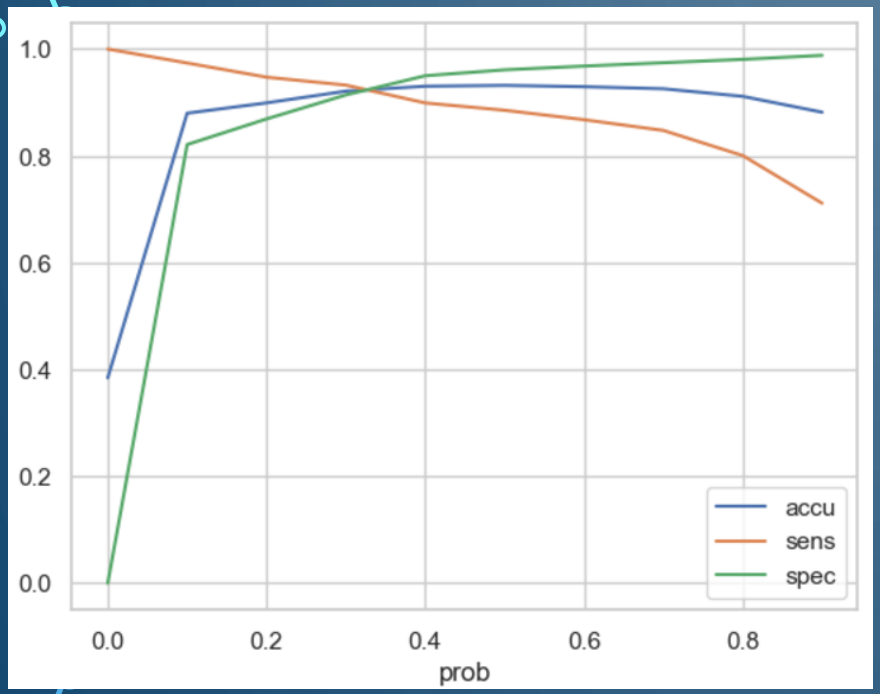
- All the numerical data are normalized
- For categorical variables, dummy variables are created
- Shape of data used for model building: **(9240,111)**
- Splitting the Data into Training and Testing Sets with 70:30 ratio
- Running RFE for selecting 15 significant variables
- Optimizing the model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5 in each step.
- Final model obtained in 4^{th} iteration with training accuracy of 92.9%.

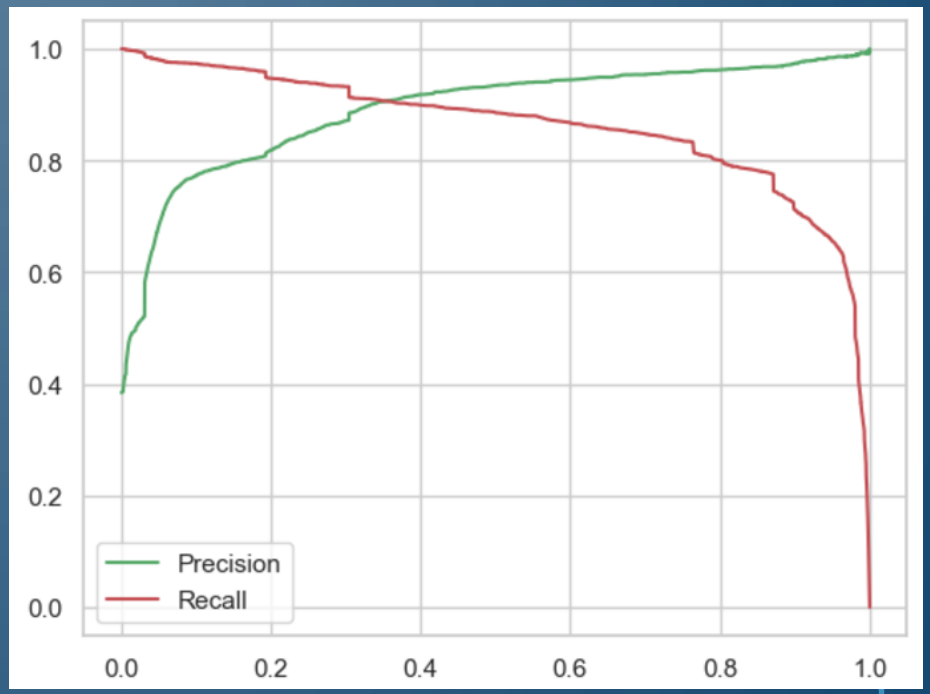
MODEL EVALUATION – ROC CURVE



A ROC-AUC score of **0.97** is very strong, indicating that the model is performing excellently in distinguishing between the positive and negative classes.

OPTIMAL CUT-OFF





The optimal cut-off of 0.37 is obtained based on the balance between accuracy, sensitivity, specificity, and the precision-recall trade-off.

MODEL METRICS

Training Data

- Accuracy: 92.86%
- Sensitivity: 90.31%
- Specificity: 94.49%
- Precision: 91.04%
- Recall: 90.31%

Test Data

- Accuracy: 92.17%
- Sensitivity: 89.57%
- Specificity: 93.82%
- Precision: 90.16%
- Recall: 89.57%

FINAL FEATURES LIST

- Total Time Spent on Website
- Lead Source_Welingak Website
- Last Activity_SMS Sent
- Tags_Busy
- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Not Specified
- Tags_Ringing
- Tags_Will revert after reading the email
- Occupation_Student
- Occupation_Unemployed
- Occupation_Working Professional

CONCLUSIONS

- **Key Predictors:** Important features influencing conversion include Total Time Spent on Website, Lead Source (Welingak Website), Last Activity (SMS Sent), and specific Tags (e.g., “Busy”, “Closed by Horizzon”).
- **Occupation Impact:** Leads with occupations like Student, Unemployed, and Working Professional show higher conversion rates.
- **Model Performance:**
  - The model demonstrates high accuracy on both training and test data, indicating it is reliable in predicting lead conversions.
  - It achieves strong sensitivity, ensuring that most potential leads are correctly identified, while maintaining high specificity, minimizing false positives.
  - Precision and Recall values reflect a well-balanced model, accurately identifying true positive leads without overwhelming the sales team with irrelevant calls.

RECOMENDATIONS

- **Focus on High-Engagement Leads:** Prioritize leads with higher website engagement and specific tags like “Will revert after reading email” for better conversion chances.
- **Leverage Lead Source & Tags:** Target leads from high-conversion sources (e.g., Welingak Website) and key tags (e.g., “Closed by Horizzon”).
- **Optimize Sales Team’s Efforts:** Direct resources towards high-probability leads and use automated channels like email for lower-scoring leads.
- **Monitor and Adjust the Model:** Continuously update the model based on evolving lead behaviors and patterns for better predictions.