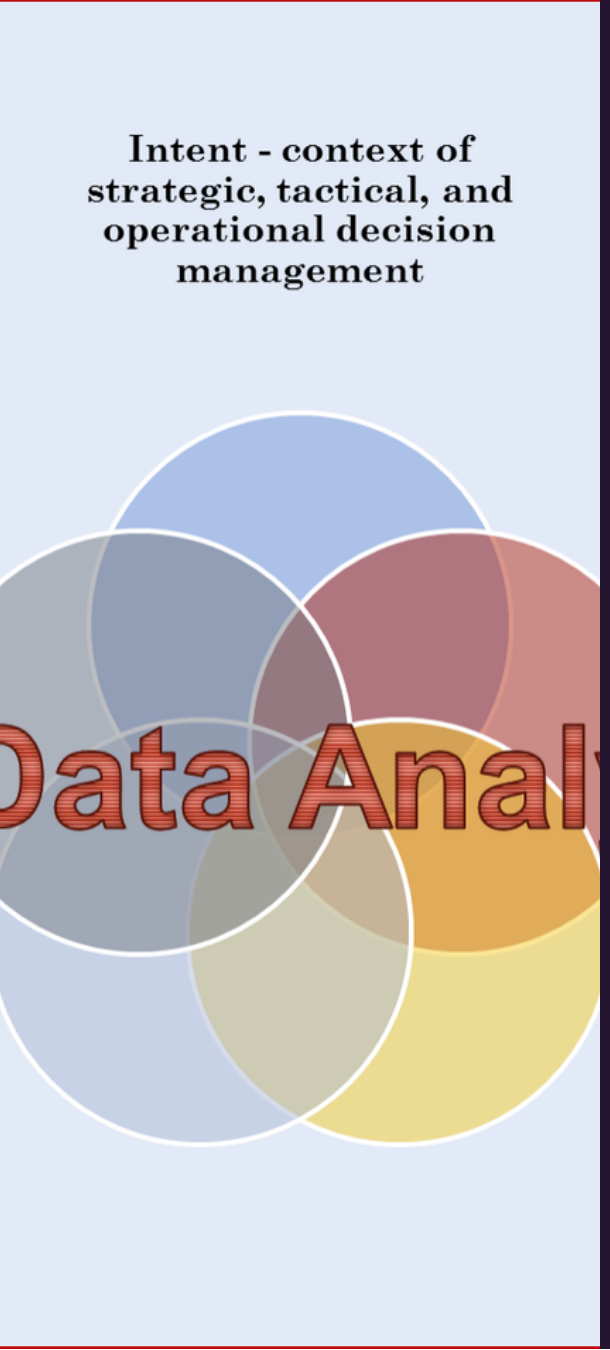
Lead Scoring Analysis: Unlocking Conversion Potential

This presentation aims to unlock the potential for lead conversion through a comprehensive analysis of the lead scoring process.

by Jaimin Sonwane



Problem Statement

Dataset Overview

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Overview of Data

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’

Analysis Approach

1. **Data Exploration**

   Understanding the dataset through visualizations and summary statistics.

2. **Handling 'Select' Levels**

   Transforming unselected categorical variables to ensure clean and accurate analysis.

3. **Logistic Regression Model**

   Building and interpreting a logistic regression model to assign lead scores effectively.

Data Exploration

Distribution of 'Converted' Leads

- The proportion of converted leads compared to non-converted leads
- Variations in conversion rates across different segments or time periods
- Trends in lead conversion over time or in response to specific marketing campaigns or activities

Summary Statistics

- The count, mean, standard deviation, minimum, 25th percentile, 50th percentile (median), 75th percentile, and maximum values are provided for each numerical variable.

'Select' Levels in Categorical Variables

The 'specialization_Select' column contains 7298 occurrences of 'False' and 1942 occurrences of 'True'. This indicates that the majority of leads have not selected a specialization, while a smaller proportion have selected a specialization.

Data Cleaning

1 Approach Explanation

Handling 'Select' Levels: The 'Select' levels in categorical variables are often treated as missing values and can be imputed using appropriate methods such as mode imputation, label encoding, or one-hot encoding..

2 Transforming 'Select' to Null Values

Identify 'Select' Levels

Replace 'Select' with Null

Verify Transformation

Handle Missing Values

Logistic Regression Model

Features Used

Converted

Lead score

Time on website spent

Last notable activity SMS Sent

Coefficients and Significance

Discuss the significance of coefficients and their impact on lead score assignment.

Lead Score Assignment Logic

Engagement Level

Behavioural Data

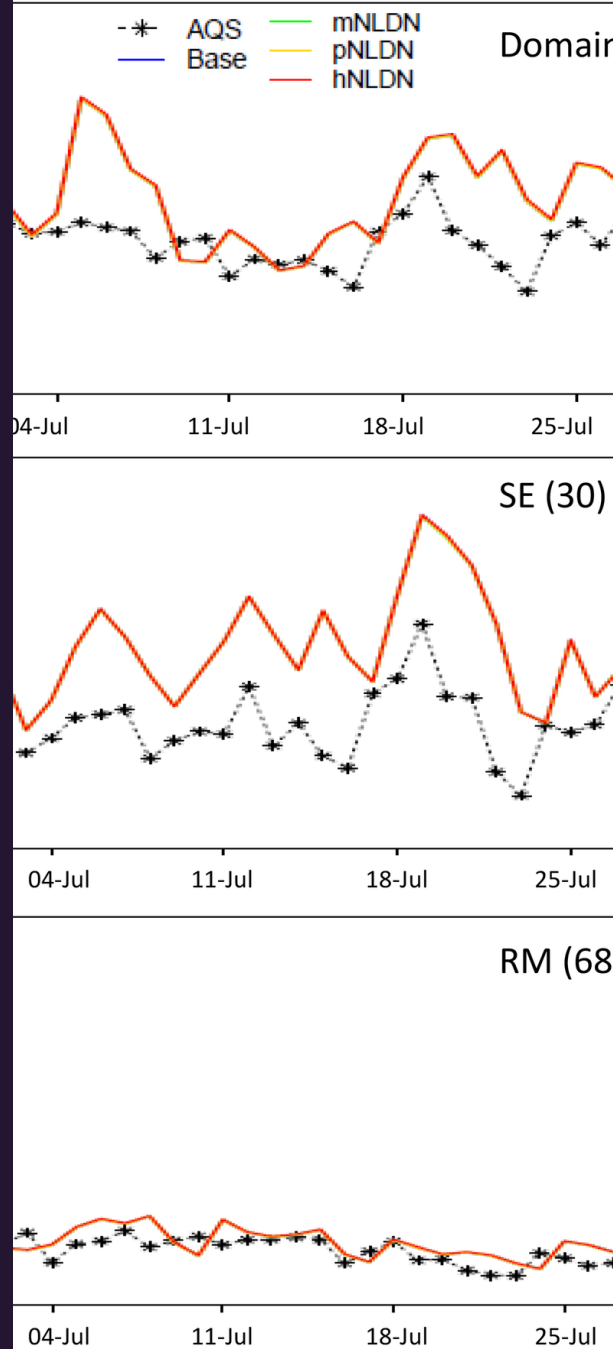
Lead Source and Campaigns

Scoring Model

Model Evaluation Metrics

1 Classification Report

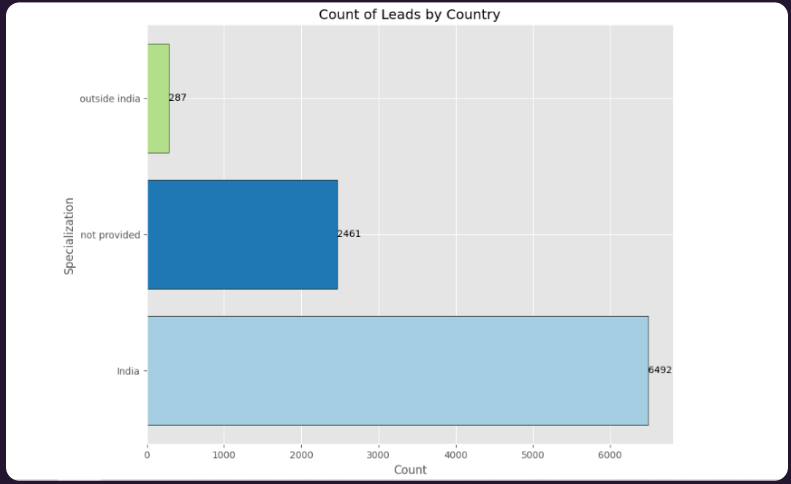
- Precision for class 0 (not converted): 0.83
- Recall for class 0: 0.88
- F1-score for class 0: 0.85
- Precision for class 1 (converted): 0.80
- Recall for class 1: 0.72
- F1-score for class 1: 0.76
- Overall accuracy: 0.82



Some Visualizations

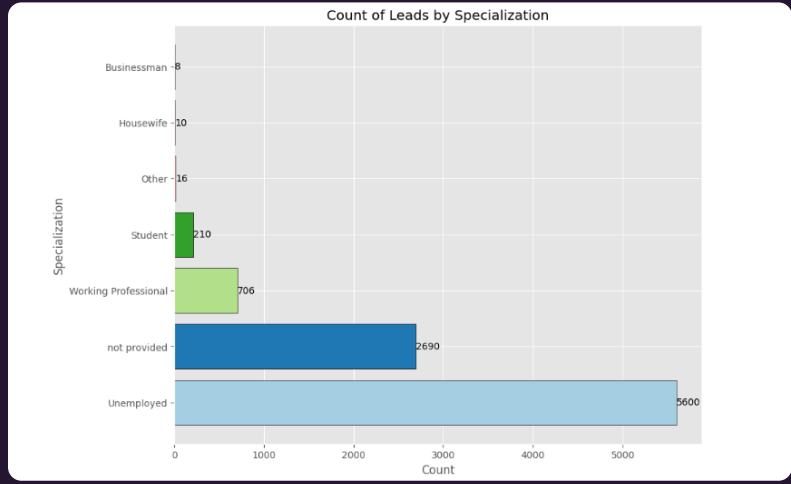
1

Count of Leads By Country

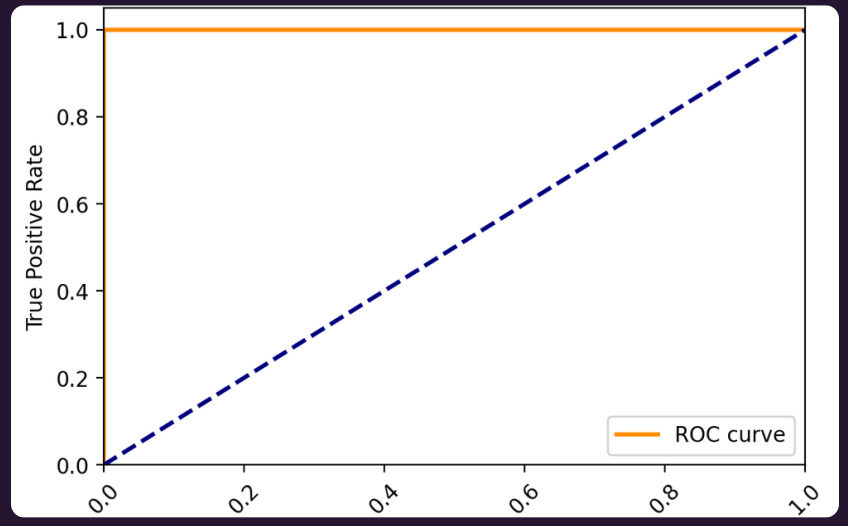


2

Count of Leads By Specialization

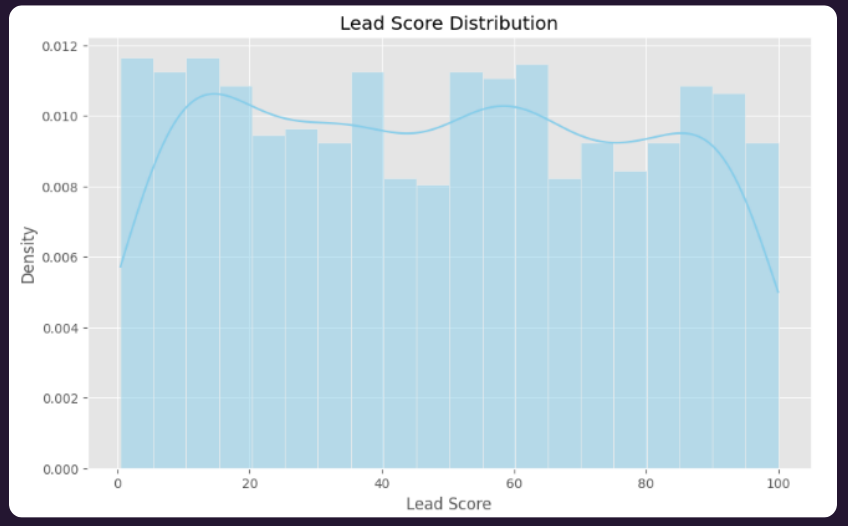


Visual Summarization



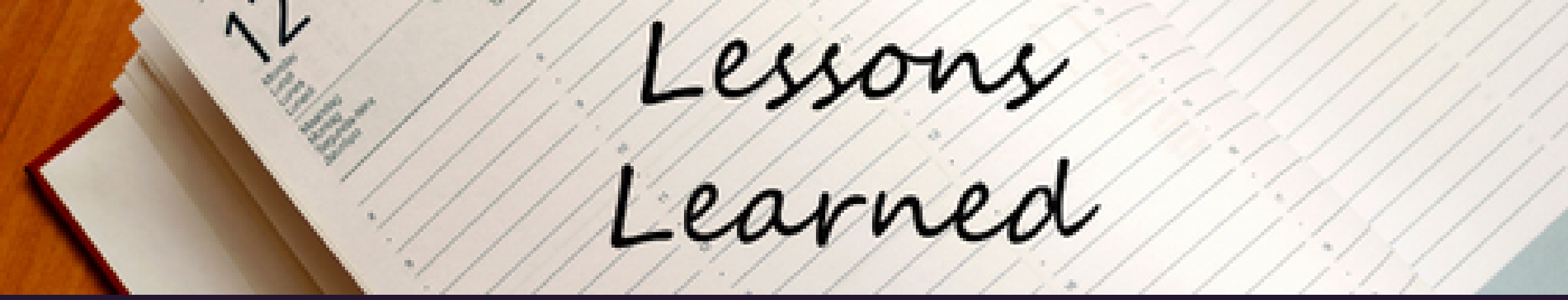
ROC Curve

A visual representation of the trade-off between sensitivity and specificity for the model.



Lead Score Distribution

Highlight the distribution pattern of lead scores obtained from the analysis.



Lessons Learned

Challenges Faced

1. **Data Quality:** Ensuring the quality and consistency of the data was a significant challenge. Strategies such as data cleaning, outlier detection, and missing value imputation were employed to address this issue.
2. **Model Interpretation:** Interpreting the results of the logistic regression model and understanding the significance of the coefficients required careful consideration. Strategies such as visualizing feature importance and conducting thorough model evaluation were employed to gain a deeper understanding.
3. **Visualization Complexity:** Creating clear and informative visualizations to communicate the results effectively was a challenge. Strategies such as using appropriate visualization techniques and providing detailed explanations were employed to address this challenge.

Insights Gained

1. **Feature Importance:** The analysis revealed the significance of different features in predicting lead conversions. Certain features demonstrated a higher coefficient value, indicating their strong impact on the prediction process.
2. **ROC Curve Analysis:** The ROC curve provided a visual representation of the trade-off between sensitivity and specificity for the model. This allowed for a clear understanding of the model's performance in distinguishing between positive and negative classes.
3. **Logistic Regression Model:** The logistic regression model was effective in lead scoring and provided valuable insights into the factors influencing lead conversion.
4. **Business Impact:** The analysis highlighted the impact of lead scores on conversion rates, providing actionable insights for optimizing lead conversion strategies.

Thank You ...!!!