Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Solution Summary:

Step1: **Data Exploration**.
Read and analyze the data.

Step2: **Data Cleaning**:
We dropped the variables that had high percentage of NULL values in them. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.

Step3: **Exploratory Data Analysis**

- Descriptive Statistics: Analyze the distribution of variables and understand the relationships between them.

- Visualizations: Use plots (e.g., histograms, scatter plots) to visualize data patterns and correlations.

Step4: **Data Preparation**

- Feature Engineering: Create new features from existing ones, encode categorical variables, and normalize/standardize numerical features.

- Train-Test Split: Split the data into training and testing sets, typically with an 80-20 or 70-30 ratio.

Step5: **Feature selection using RFE**:
Using the Recursive Feature Elimination we went ahead and selected the 20 top important features. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.

Step6: **Model Building**:

Feature Selection: Choose the most relevant features for the model to avoid overfitting and reduce dimensionality.

Model Selection: Choose logistic regression as the model due to its suitability for binary classification problems.

Step 7: **Model Training**:

Fit the logistic regression model to the training data using appropriate techniques, such as Maximum Likelihood Estimation (MLE).

Step 8: **Model Evaluation**:

Performance Metrics: Use metrics like Accuracy, Precision, Recall, F1-Score, and ROC-AUC to evaluate the model's performance on the test data.

Confusion Matrix: Analyze the confusion matrix to see how well the model classifies the true positive and negative cases.

Step 9: **Results on Observations**:

Coefficients Analysis: Interpret the logistic regression coefficients to understand the impact of each feature on the outcome.

Odds Ratio: Calculate and interpret the odds ratios for the features to understand their effect on the probability of the event.

Step 10: **Conclusion on Hypotheses**:

Hypothesis 1: Leads that spend more time on the website are more likely to convert.

Validated: The analysis supports this hypothesis, as leads with higher engagement (spending more time on the website) show a higher likelihood of conversion. This is reflected in the positive coefficient for the "Total Time Spent on Website" variable, indicating a positive relationship between website engagement and conversion likelihood.

Hypothesis 2: Certain lead sources (e.g., referrals, specific ads) have higher conversion rates.

Validated: The model's output shows that certain lead sources, such as "Welingak Website" and "Reference," have strong positive coefficients, suggesting higher conversion rates. This confirms that not all lead sources are equal in conversion potential, and targeting those with higher conversion rates can improve overall efficiency.

Hypothesis 3: Demographic factors like occupation and country might influence conversion likelihood.

Partially Validated: The analysis confirms that occupation, specifically leads identified as "Working Professionals," is a significant factor influencing conversion likelihood, as indicated by its positive coefficient. However, the hypothesis regarding country influence wasn't explicitly tested or observed in the provided data, so further analysis would be needed to explore that aspect.

Overall Conclusion:

The data analysis strongly supports the first two hypotheses, providing clear evidence that website engagement and certain lead sources are crucial factors in predicting lead conversion.

The third hypothesis is partially validated, with occupation being an influential factor, though further exploration is needed for other demographic variables like country.

These insights can be directly applied to refine X Education's lead conversion strategy, making it more targeted and effective.

Based on the analysis and model performance, several key insights can be drawn-

- **Model Effectiveness:** The model demonstrates strong predictive accuracy, aligning closely with the business objective of improving lead conversion rates to 80%. The consistency in performance metrics between the training and test datasets indicates that the model is reliable and generalizes well, making it a solid tool for guiding sales strategies.

- **Targeted Sales Efforts:** The model's output has identified 386 leads with a higher likelihood of conversion. By focusing on these leads, X Education can optimize sales efforts and improve efficiency, addressing the current inefficiency problem where the sales team spends time on leads with lower conversion potential.

- **Validated Hypotheses:** The analysis supports the hypotheses that, Leads spending more time on the website are more likely to convert. Certain lead sources, like referrals and specific online channels, have higher conversion rates. **Demographic factors**, such as occupation, play a significant role in conversion likelihood.

Strategic Recommendations: The recommendations based on the model's output provide actionable steps for X Education to enhance their lead conversion strategy. By prioritizing leads from high-conversion sources, targeting working professionals, and focusing on engaged leads, the company can better allocate resources and increase its conversion rate.

- **Potential for Business Impact:** By implementing these targeted strategies, X Education can significantly increase its lead conversion rate, moving closer to the desired 80%. This improvement will not only increase revenue but also optimize the sales team's efforts, ensuring that time and resources are spent on the most promising leads.

Overall, the model provides a clear pathway to achieving the business goal of a higher lead conversion rate, supporting strategic decision-making and enabling the company to make more informed calls on which leads to pursue.

Submitted by DSC 65:
Rafida Khannum
Keshavi Gupta
Venkat Aditya