Table of Contents

Lead Scoring Case Study - X Education

X Education aims to improve its lead conversion rate (~30%) by identifying high-potential leads. This project builds a Logistic Regression Model to assign a Lead Score (0-100) to prioritize leads, increasing efficiency for the sales team.

Steps Followed

1. Data Cleaning

- Handled null values and replaced unnecessary "Select" options with NULL.
- Dropped columns with more than 40% null values. For columns with less than 40% missing values, imputed categorical data using the mode and numerical data using the mean.
- Standardized text data by converting all-uppercase and all-lowercase values to proper case format.

2. Data Transformation

- Changed multicategory labels into binary variables in the form of '0' and '1'.
- Created dummy variables.
- Checked for outliers and applied IQR-based outlier treatment by capping values above the upper bound and flooring values below the lower bound for numerical columns, with validation checks for any remaining outliers.

3. Data Preparation

- Split the dataset into training (70%) and testing (30%) sets.

- Scaled the dataset using `StandardScaler()`.
- Plotted a heatmap to analyze correlations and removed highly correlated variables.

4. Model Building

- Built the model using Recursive Feature Elimination (RFE) with 15 variables.
- Removed insignificant variables based on P-Value scores, one by one.
- Checked the Variance Inflation Factor (VIF) for each variable, ensuring all variables had a VIF score < 5.0 before proceeding.

5. Model Evaluation

- Determined the optimal probability cutoff by analyzing accuracy, sensitivity, and specificity.
- Identified the best cutoff point and used it for final predictions.
- Evaluated the model's precision, recall, accuracy, sensitivity, and specificity.
- Predicted values on the test set, achieving an accuracy above 80%.
- Conducted model evaluation on the test set, confirming an acceptable accuracy and recall/sensitivity above 80%.
- Assigned lead scores to the test dataset, indicating that high lead scores represent hot leads, while low lead scores are not hot leads.

6. Conclusion

Key insights from the study:

- The test set achieved accuracy and recall/sensitivity in an acceptable range.
- The model demonstrated stability and adaptability, allowing it to adjust to future changes in business requirements.
- **Recommendation for a high conversion rate:**
    1. **Lead Score Utilization:** Sales teams should prioritize hot leads, saving effort on cold leads. Leads with a score above 30 are more likely to convert.
    2. **Optimizing Sales Team Efforts:**
        - If Lead Score > 70 → High-priority leads requiring immediate actions (emails, calls, messages).
        - If Lead Score is between 30-70 → Medium-priority leads should receive promotional emails and discounted offers.
        - If Lead Score < 30 → Minimal efforts and remarketing.
    3. **Sales Strategy Adjustments:**
        - Leads tagged as "Closed by Horizon" and "Will revert after reading email" have a strong positive correlation with conversion and should be prioritized for follow-ups.
        - Leads tagged as "Busy" should be followed up at different times/days to increase response rates.
    4. **Variable Thresholding Based on Lead Sources:**

- The lead score threshold can be adjusted based on the source of leads to improve accuracy.

5. **Rebuilding and Recalibrating the Model:**
    - Customer behavior evolves over time, requiring periodic model recalibration.
    - Additional features may need to be introduced, and some features may become obsolete.
    - The model should be updated periodically to maintain accuracy and recall performance.