

Summary Report: Logistic Regression Model for Lead Scoring

Business Context

X Education, which is an education company is facing a challenges lead conversion. They have asked us to help them with a solution to help them identify the most promising leads, which are highly likely to be converted. The objective is to assign a lead score to each lead, indicating their likelihood of conversion. The leads with higher scores should be prioritized as they are more likely to convert into paying customers. The target lead conversion rate is set around 80%.

Data Description

The dataset provided by X Education consists of approximately 9,000 leads with various attributes, including:

- Lead Source, Total Time Spent on Website, Total Visits, Last Activity and many more.
- Target variable: 'Converted' (1 for converted, 0 for not converted)

A notable challenge in the data is the presence of a 'Select' level in many categorical variables, which acts as a proxy for null values and needs to be handled appropriately.

Following is the approach used to build a logistic regression model to calculate the probability/lead conversion score,

Data Preprocessing

1. Handling Missing Values:

- Columns with high missing values are removed from the analysis
- Categorical columns with 'Select' as a value were imputed with a new category 'missing'
- Records with missing values in numerical columns are excluded, while missing values in categorical columns are retained with a 'missing' category if the variable is found significant

2. Encoding Categorical Variables & Test-Train Split:

- Utilized `pd.get_dummies` to convert categorical variables into dummy/indicator variables, ensuring no multicollinearity issues by dropping one level from each category.
- Split the dataset into training and testing subsets (e.g., 70-30 or 80-20 split).

3. Scaling:

- Standardized numerical features to ensure they are on the same scale. (using Min-Max scaling)

Model Building

4. Feature Selection:

- Used Recursive Feature Elimination (RFE) to select relevant features.
- Assessed multicollinearity using Variance Inflation Factor (VIF).
- Selected important features: 'Lead Origin', 'Lead Source', 'Do Not Email', 'Converted', 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit', 'Last Activity', 'Specialization', 'What is your current occupation', 'A free copy of Mastering The Interview', and 'Last Notable Activity'.





5. Model Training:

- Trained the logistic regression model on the preprocessed dataset.
- Evaluated the significance of each feature using the p-value and removed features with high p-values (>0.05).
- Assessed multicollinearity using Variance Inflation Factor (VIF) and removed highly collinear variables.
- Utilized techniques like cross-validation to ensure model robustness.

6. Model Evaluation:

- Evaluated model performance using metrics such as accuracy, precision, recall, and ROC-AUC.
- The precision-recall tradeoff chart was particularly useful in determining the optimal cut-off probability to achieve the target lead conversion rate.

7. Model Performance Metrics:

- Confusion matrix:
 -  True negatives (TN): **1949**
 -  False positives (FP): **392**
 -  False negatives (FN): **539**
 -  True positives (TP): **1614**
- Model accuracy is **80%**
- Sensitivity (Recall) was approximately **75%**, specificity around **83%**.

8. Finding the Optimal Cutoff Point:

- Created an ROC curve (AUC = 0.87).
- Optimal cutoff point: Around **42%** for improved performance.
- At this cutoff, sensitivity increased to **79.7%**, specificity at **79%**, precision at **77.7%**.

9. Making Predictions on the Test Set:

- Scaled test set features and selected relevant columns.
- Predicted conversion probabilities.
- Evaluated model performance:
 - 📊 Accuracy: **77%**
 - 📊 Sensitivity: **79%**
 - 📊 Specificity: **76%**
 - 📊 Precision: **76%**

10. Precision and Recall View:

- Built a model using precision-recall trade-off.
- Optimal threshold (training set): 0.42.
- Precision: ~**80%**, Recall: ~**79%**.
- Confusion matrix:
 - 📊 TN: **1850**
 - 📊 FP: **491**
 - 📊 FN: **435**
 - 📊 TP: **1718**

11. Prediction on the Test Set:

- Predicted conversion probabilities on the test set.
- Accuracy on the test set: ~**78%**.
- Confusion matrix:
 - 📊 True negatives (TN): 770
 - 📊 False positives (FP): 218
 - 📊 False negatives (FN): 201
 - 📊 True positives (TP): 737

Key Findings

1. Significant Variables:

- Total Time Spent on Website: Coefficient = 4.562, z-value = 24.41
- Lead Origin_Lead Add Form: Coefficient = 3.8964, z-value = 15.08
- Total Visits: Coefficient = 3.0001, z-value = 5.009

These variables have the highest absolute coefficients and z-values, indicating their strong influence on lead conversion.

2. Top Categorical/Dummy Variables:

- Lead Origin_Lead Add Form: Coefficient = 3.8964
- What is your current occupation_Working Professional: Coefficient = 2.4917
- Source_Welingak Website: Coefficient = 2.2903

These categorical variables should be focused on to increase the probability of lead conversion.

3. Optimal Cut-off Probability:

- The model has an optimal performance at 42% threshold
- The precision-recall tradeoff indicated that a cut-off probability around 80% yields a high precision (>90%), making it ideal for prioritizing leads when its important ensure higher lead conversion due to limited bandwidth.

Recommendations

- Target leads with high probability scores (predicted as 1 by the model) as they are the most promising.
- The precision-recall tradeoff chart suggests using higher cut-off, for e.g. above 80% probability to ensures much higher conversion.
- Segment and prioritize leads based on key variables like Total Time Spent on Website, Lead Originating from Add Form, leads having higher Total Visits, Working Professionals, and Sourced from Welingak Website for aggressive Lead Conversion During Intern Period.
- Score each segment based on their respective values to tailor communication and engagement strategies.