Lead Scoring Assignment

Done by:
Aastha Sharma

Problem Statement

- X Education markets online courses to industry professionals, acquiring leads from website visitors, search engines, and referrals.
- Leads are generated when potential customers fill out forms or provide contact details, after which the sales team reaches out via calls and emails.
- The current lead conversion rate is low, with only 30% of leads converting into customers.
- To improve efficiency, the company aims to identify 'Hot Leads'—those with higher conversion potential—so the sales team can focus their efforts on them, improving overall conversion rates.

Steps Followed

1. **Data Collection**:

➢ Gather the dataset containing the dependent variable (binary or categorical outcome) and independent variables (features or predictors).

2. **Data Preprocessing**:

➢ Handle missing values, if any.

➢ Convert categorical variables to dummy variables (one-hot encoding).

➢ Normalize or scale numerical features if required.

➢ Split the data into training and testing sets.

3. **Model Building**:

➢ Import necessary libraries such as sklearn or statsmodels.

➢ Initialize the logistic regression model.

4. **Training the Model**:

➢ Fit the logistic regression model to the training data.

➢ The model learns the relationships between independent variables and the outcome variable.

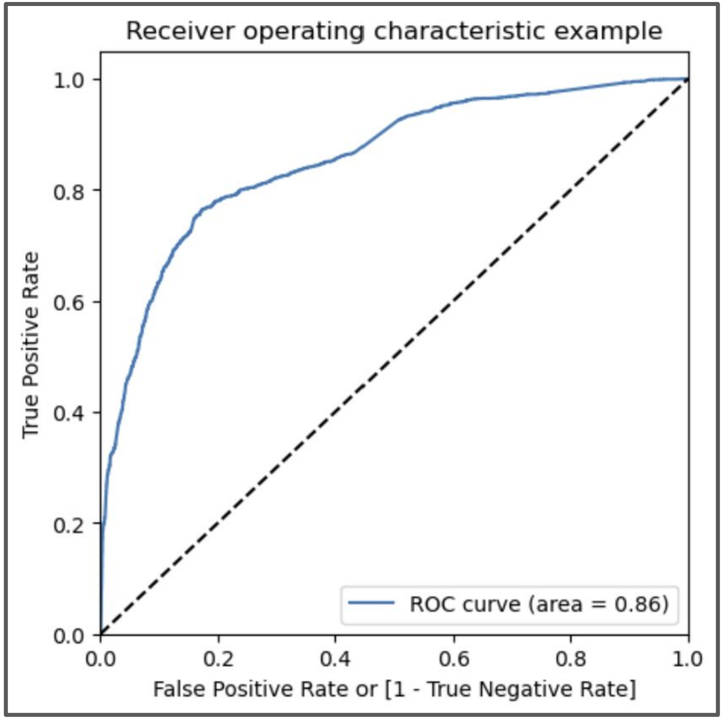
5. **Model Evaluation**:

➢ Predict the probabilities and classes for the testing set.

➢ Evaluate the model using performance metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

OBSERVATIONS

ROC Curve of Training Data



- With an AUC of 0.86, the model in this case performs significantly better than random chance. It has a strong ability to separate positive and negative classes, though there may still be some misclassifications.
- This score suggests that the model is fairly accurate, but there might be room for improvement.

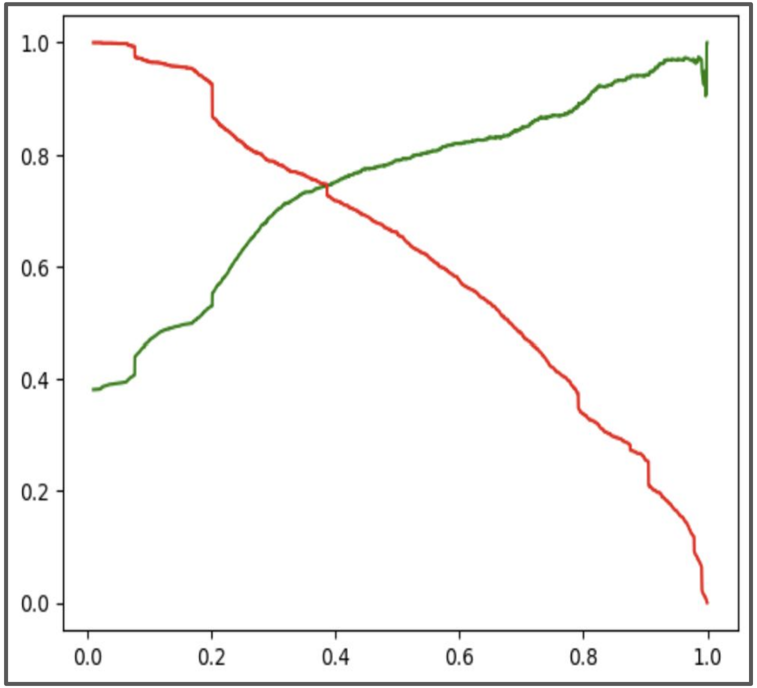
Accuracy, Specificity and Sensitivity of Training Data

Sensitivity (78.71%): Sensitivity measures how well the model identifies actual positives (True Positives). A value of 0.787 indicates that the model correctly identifies around 78.71% of the positive cases. This suggests that the model is relatively good at detecting positive outcomes but still misses about 21.29% of them.

Specificity (78.69%): Specificity refers to the model’s ability to identify true negatives. With a value of 0.786, the model correctly identifies around 78.69% of negative cases. This shows a balanced performance in recognizing both positive and negative cases, though some false positives remain.

Accuracy (78.69%): Accuracy reflects the overall correctness of the model’s predictions, with an accuracy of 78.69%. This means that roughly 78.69% of the total predictions (both positive and negative) are correct. However, it should be noted that accuracy alone doesn't always give a full picture, especially if the classes are imbalanced.

Precision and Recall of Training Data

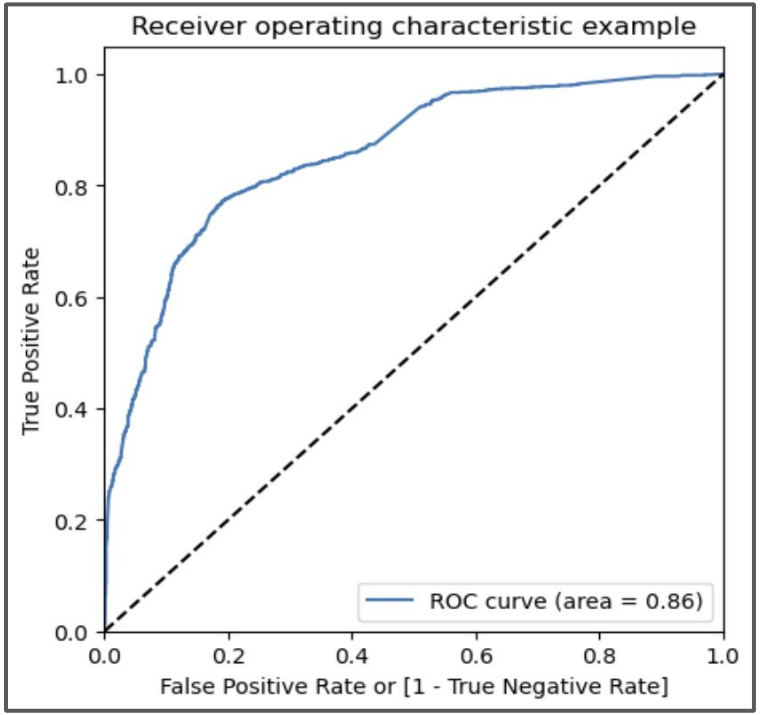


Precision: 69.47% – Of all the positive predictions, 69.47% were correct.

Recall: 78.71% – The model correctly identified 78.71% of the actual positive cases.

This indicates a trade-off between precision and recall, where the model is moderately accurate in predicting positives but still misses some true positive cases.

ROC Curve of Test Data



- The area under the curve (AUC) is 0.86, which suggests that the model has good discrimination ability between positive and negative classes. A value of 1.0 indicates perfect classification, while 0.5 means no better than random guessing.
- With an AUC of 0.86, the model performs well but has room for improvement.

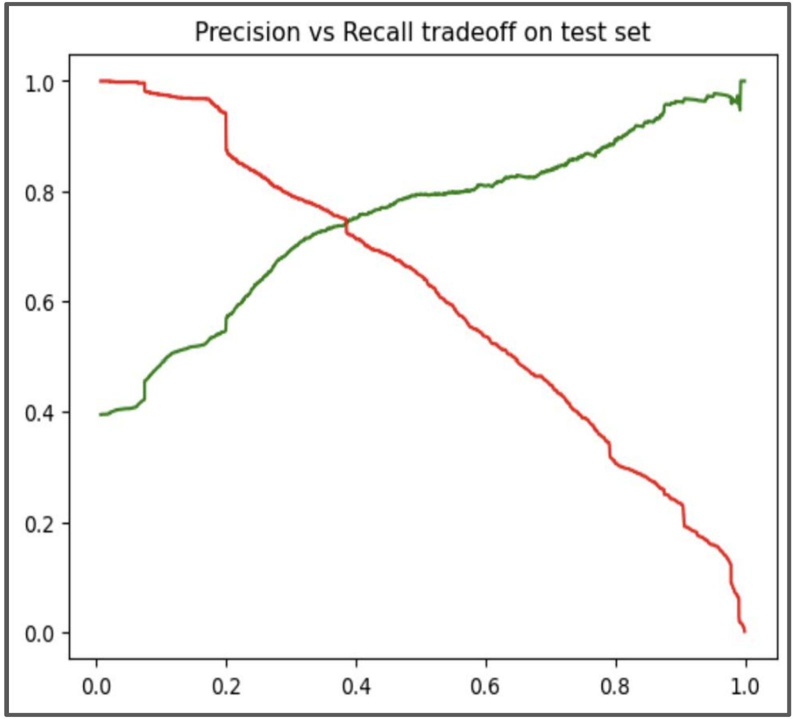
Accuracy, Specificity and Sensitivity of Test Data

Sensitivity (79.27%): This indicates that your model is correctly identifying around 79.27% of the positive cases. It shows the model's ability to capture true positives, meaning it's performing well in identifying the actual positive outcomes.

Specificity (77.04%): The specificity value of 77.04% shows that the model is correctly identifying negative cases about 77.04% of the time. This reflects the model's effectiveness in avoiding false positives and correctly classifying negative cases.

Accuracy (77.92%): An accuracy of 77.92% implies that the model is correctly predicting both positive and negative cases in around 78% of the total cases. While it gives an overall performance measure, accuracy alone might not be the best indicator if there’s class imbalance, so sensitivity and specificity provide additional clarity.

Precision and Recall of Test Data



Precision: 50.44% – Of all the positive predictions, 50.44% were correct.

Recall: 62.31% – The model correctly identified 62.31% of the actual positive cases.

The graph shows a trade-off between precision and recall. As one increases, the other tends to decrease.

Conclusion

- The model demonstrates strong performance with an AUC of 0.86, indicating good discrimination between positive and negative classes. However, there is still potential for enhancement to achieve even higher accuracy and better classification.
- While the accuracy, sensitivity, and specificity are relatively consistent between training and test data, there is a trade-off between precision and recall. The training data shows higher precision compared to test data, while recall is higher in test data. This suggests that improvements in one metric could impact the other.
- To optimize performance, focusing on balancing precision and recall is essential. Adjustments in feature selection, model tuning, or employing techniques such as cross-validation could lead to more reliable predictions and better generalization to unseen data.
- Continued refinement and evaluation of the model can help in addressing these trade-offs and improving overall predictive accuracy, leading to more effective and reliable outcomes.

THANK YOU