Fraud Analytics Assignment 2 : Comparison of Cost sensitive classification and Logistic Regression for Classification

Team Members:
CS22BTECH11043 : Nethi Keerthana
CS22BTECH11012 : Bolla Nehasree
ES22BTECH11025 : N. Krishna Chaitanya

March 26, 2025

1 Problem Statement

The goal of this study is to compare the performance of Genetic Algorithm-based classifier which is a Cost senstive classification model with a standard Logistic Regression model. We evaluate both methods based on accuracy and cost-sensitive loss across various threshold values.

2 Dataset Description

Columns A to K are independent variables, Column L is the dependent variable,Column M is the false negative cost, varying from row to row based on the business details,True Positive and False Positive cost is constant for all, which is 3 and True Negative cost is constant for all, which is 0.

3 Algorithms

3.1 Genetic Algorithm for cost sensitive classification Classification

The Genetic Algorithm (GA) optimizes the weight vector of a logistic model by evolving a population through selection, crossover, and mutation. The objective function minimizes a weighted cost-sensitive loss.

Cost Sensitive Cost Function:

$$J^c(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i(h_\theta(x_i)C_{TP_i} + (1 - h_\theta(x_i))C_{FN_i})$$

$$+(1 - y_i)(h_\theta(x_i)C_{FP_i} + (1 - h_\theta(x_i))C_{TN_i}))\,.$$

$$J_i(\theta) = \begin{cases} C_{TP_i} & \text{if } y_i = 1 \text{ and } h_\theta(x_i) \approx 1 \\ C_{TN_i} & \text{if } y_i = 0 \text{ and } h_\theta(x_i) \approx 0 \\ C_{FP_i} & \text{if } y_i = 0 \text{ and } h_\theta(x_i) \approx 1 \\ C_{FN_i} & \text{if } y_i = 1 \text{ and } h_\theta(x_i) \approx 0. \end{cases}$$

3.2 Standard Logistic Regression

Standard logistic regression is a binary classification model that predicts probabilities using the sigmoid function. It learns a linear decision boundary by minimizing binary cross-entropy loss through optimization techniques like gradient descent.

Logistic regression Cost Function:

$$J_i(\theta) = -y_i \log(h_\theta(x_i)) - (1 - y_i)\log(1 - h_\theta(x_i)).$$

- If $y_i = 0$ and $h_\theta(x_i) \neq 0$, then

$$J_i(\theta) = -(0)\log(0) - (1 - 0)\log(1 - (0)) \approx 0.$$

- If $y_i = 0$ and $h_\theta(x_i) \approx 1$, then

$$J_i(\theta) = -(0)\log(1) - (1 - 0)\log(1 - (1)) \to \infty.$$

- If $y_i = 1$ and $h_\theta(x_i) \approx 0$, then

$$J_i(\theta) = -(1)\log(0) - (1 - 1)\log(1 - (0)) \to \infty.$$

- If $y_i = 1$ and $h_\theta(x_i) \approx 1$, then

$$J_i(\theta) = -(1)\log(1) - (1 - 1)\log(1 - (1)) \approx 0.$$

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} J_i(\theta).$$

4 Algorithm :

1. Load dataset and split into training and testing sets.
2. Scale features using standardization.
3. Train the Genetic Algorithm and Logistic Regression models.
4. Evaluate accuracy and cost-sensitive loss for different thresholds.
5. Compare average training and testing costs across models.
6. Visualize performance using plots.

5 Results and Analysis

5.1 Cost VS Offset values

Figure 2 shows the effect of varying the decision threshold on cost-sensitive loss for both methods.

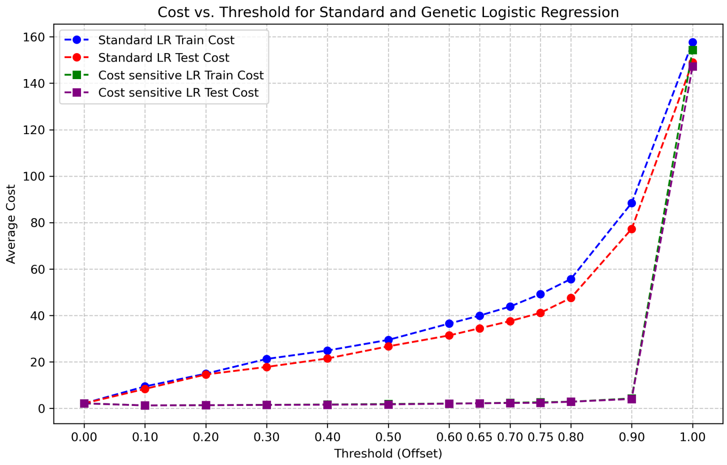

Figure 1: Comparison of Average Costs for Genetic Algorithm and Logistic Regression

6 Comparison and Observation :

The plot illustrates the impact of varying thresholds on the average cost for both Standard Logistic Regression (LR) and Cost-sensitive Logistic Regression (CSLR). In the case of Standard LR, represented by the blue and red lines, both training and testing costs increase as the threshold rises. In contrast, Cost-sensitive LR, depicted by the green and purple lines, maintains consistently lower and nearly constant costs across different threshold values. This indicates that the cost-sensitive approach effectively minimizes misclassification penalties. Overall, the comparison highlights that Cost-sensitive LR achieves significantly lower costs than Standard LR, demonstrating its efficiency in handling misclassification costs and improving overall cost-effectiveness.Average costs for cost sensitive method for different offsets can also be seen to be similar till 0.9. Also training and testing cost for CSLR is almost the same as seen in the graph which suggests that model is free from over fitting or underfitting.

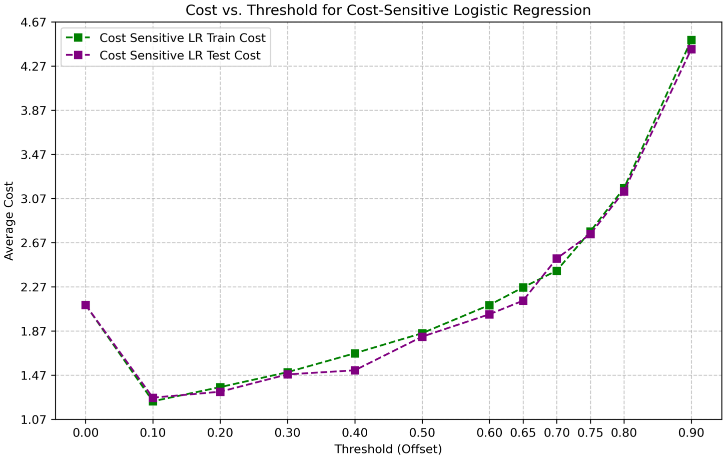

Figure 2: Average Costs for Cost sensitive logistic regression

7 Observation :

Cost is initially high at low thresholds but drops sharply around 0.10-0.20, indicating an optimal range for offset.

After 0.20, cost steadily increases, due to more mis classification costs probably dire to false negatives.

Train and test costs follow similar trends, showing good generalization.

Best threshold: 0.10-0.20 for minimizing cost.