Cost-Sensitive Regression via Genetic Algorithms

Team Members: CS21BTECH11060, MA22BTECH11015, AI22BTECH11020

March 26, 2025

1 Introduction

This report summarizes a cost-sensitive regression implementation using Genetic Algorithms for fraud detection. The dataset consists of 11 independent features (columns A to K), a binary dependent label (column L), and a row-dependent false negative cost (column M). The true positive and false positive costs are constant (6), while the true negative cost is 0.

2 Methodology

2.1 Data Preprocessing

The data is read from `costsensitiveregression.csv`. Columns are reordered so that the label becomes the first column and the false negative cost ("FNC") is separated from the feature set. The dataset is split into training (80%) and testing (20%) sets.

2.2 Cost Functions

Two cost functions are implemented:

- **Bahnsen Approach:**

  $$L(y, \hat{y}) = y\,(\hat{y} \cdot TP + (1 - \hat{y}) \cdot FN) + (1 - y)\,(\hat{y} \cdot FP + (1 - \hat{y}) \cdot TN)$$

  where $TP = FP = 6$, $TN = 0$, and FN is row-dependent.

- **Nikou Approach:**

  $$b = (1 - y)\,\Gamma^{-1}(FP) + y\,\Gamma^{-1}(FN), \quad a = \frac{1}{\Gamma(b + 1)}$$

  with the loss defined as

  $$L(y, \hat{y}) = a\,y\,(-\log(\hat{y}))^b + a\,(1 - y)\,(-\log(1 - \hat{y}))^b\,.$$

2.3 Genetic Algorithm

A Genetic Algorithm is used to optimize the weight vector for each approach. The GA utilizes:

- **Population Size:** 50
- **Iterations:** 300
- **Mutation Probability:** 0.01

Fitness is defined as $1/(1+\text{loss})$ and candidate solutions are evolved through crossover and mutation.

3 Results

The algorithm selects the best weight vector based on maximum fitness on the training set. Testing is then performed using both cost functions, and the convergence of the loss over iterations is plotted for analysis.

3.1 Bahnsen Approach

The Bahnsen approach results are shown in the plot:

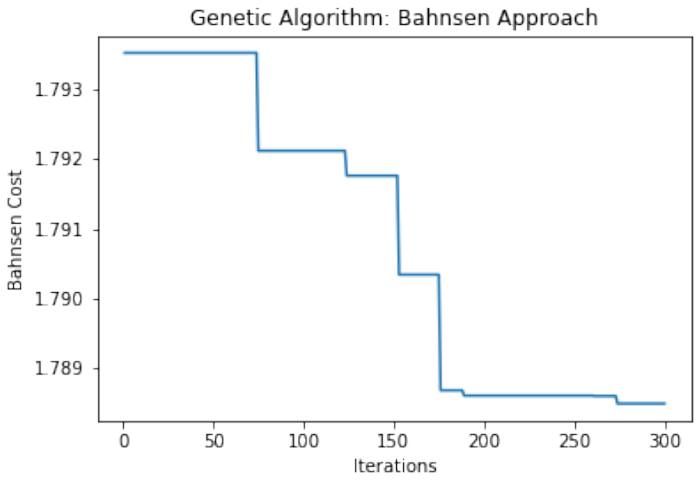

Figure 1: Genetic Algorithm Optimization for Bahnsen Approach

4 Conclusion

Both the Bahnsen and Nikou cost-sensitive approaches were implemented and optimized via a Genetic Algorithm. The results indicate that the GA effectively minimizes the cost-sensitive loss. Future work may refine these methods further for improved performance on larger datasets.