CS6890: Fraud Analytics - Assignment 2
Example Dependent Cost Sensitive Classification

Nitya Bhamidipaty - CS21BTECH11041
Akshay Santoshi - CS21BTECH11012

March 2025

1 Data Description

- Columns A to K are independent variables.
- Column L is the dependent variable i.e Status
- 147636 data points are given.
- The number of points with status 0 is 103554 which is much larger than status 1 (44082). This shows that the data set is biased.
- Null values are not present in the data set.
- The average FNC is 533 with a standard deviation of 8774.
- The train-test split taken is 80-20.

1.1 Possible Fraud Detection Insights

- Highly skewed distributions: Many features have long tails, meaning that a small percentage of users exhibit extreme behavior. These users might warrant further investigation.
- Zeros dominate: Many fraud-related variables (PFD, PFG, SFD, SFG) have 0 as their 75th percentile, meaning that fraud cases are rare.
- Withdrawal patterns may signal fraud: Some accounts have very high WP and WS values, which could be linked to suspicious transactions.

2 Approach 1: Vanilla Logistic Regression

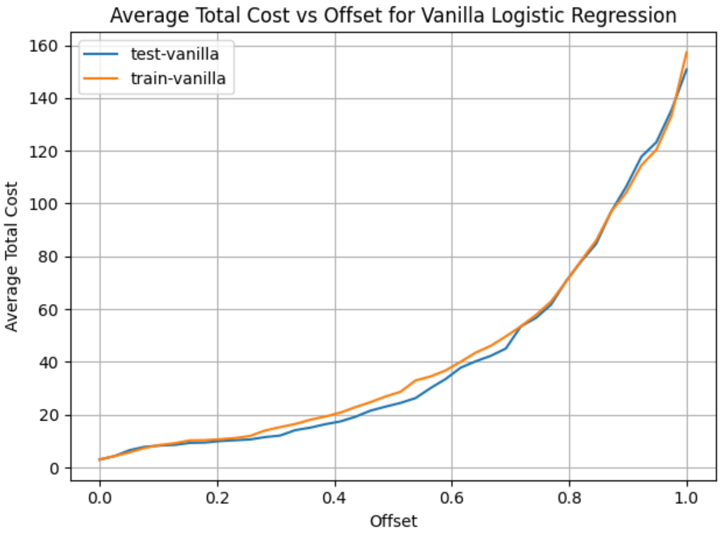

Figure 1: Vanilla Logistic Regression Cost vs Offset

- The average training and testing cost is close, indicating that there is no overfitting.
- Higher offset values result in higher costs, suggesting that as a certain threshold (offset) increases, the logistic regression model incurs more cost.
- The graph shows that 0.5 is not an optimal threshold in terms of cost, although it may be for accuracy (since logistic regression optimizes accuracy).

3 Approach 2: Bahnsen's Logistic Regression

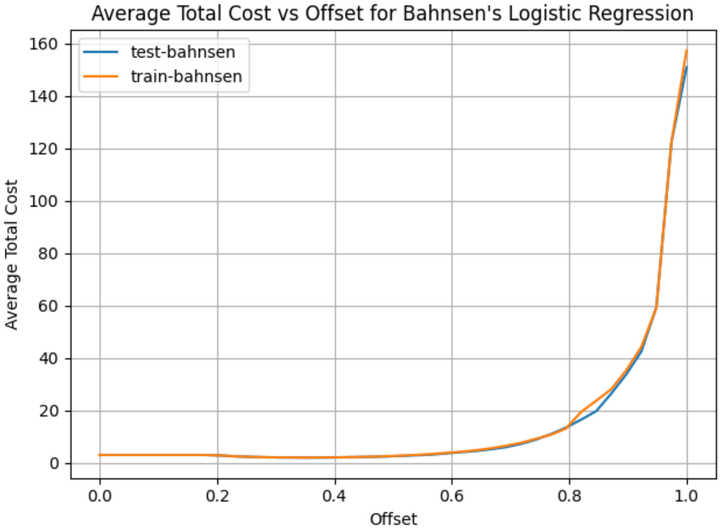

Figure 2: Bahnsen's Logistic Regression

- The average training and testing cost is close, indicating that there is no overfitting.
- The graph shows that 0.5 is not an optimal threshold in terms of cost.
- At threshold 1.0 every point is classified as 0 (negative). The FNC of all data points whose status is 1 gets added in the total cost, because of this the Average Total cost becomes high.

4 Comparison of Models

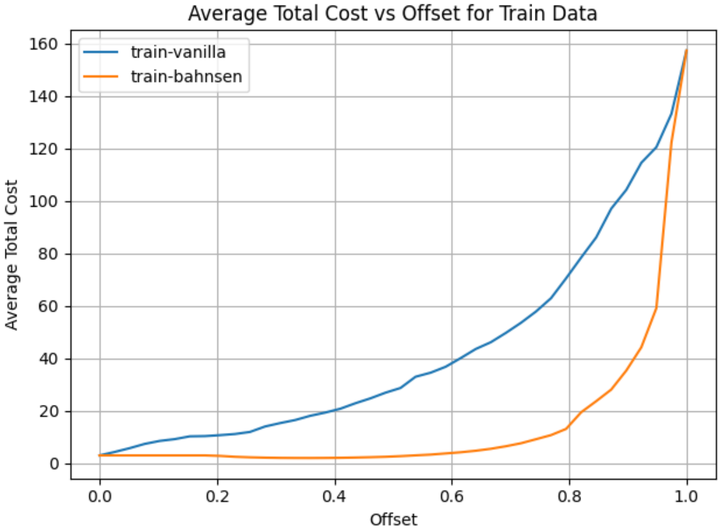

Figure 3: Comparing average cost for train data

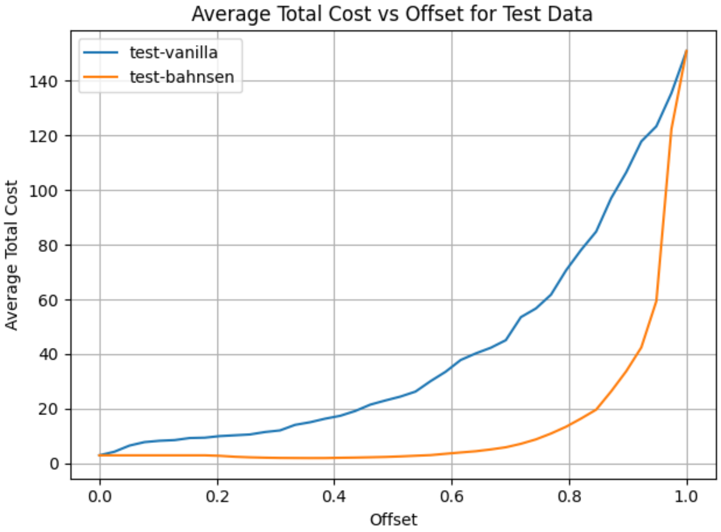

Figure 4: Comparing average cost for test data

- In these plots, the cost-sensitive model clearly performs better for both train and test data.

- The cost-sensitive model is more flat in comparison to the logistic regression model, showing that it is not that sensitive to the offset.

5 Results

Overall, the Cost-Sensitive model performs better as indicated by the graphs.