CS6890: Fraud Analytics
Assignment 3
Cost-Sensitive Regression for Predictive Modeling

Manikanta Vallepu - AI20BTECH11014
Jarupula Saikumar - CS21BTECH11023
Vignan Kota - CS21BTECH11029
Sahil Chandra - CS20BTECH11033
Kalala Abhinav - CO21BTECH11007

1. Problem Statement

The accurate prediction of target variables is a fundamental task in various domains, including finance, healthcare, and marketing. However, traditional regression models often fail to account for the varying costs associated with prediction errors, leading to suboptimal performance in scenarios where misclassifications incur significant financial or societal consequences.

In this research study, we aim to address the challenge of cost-sensitive regression, where **traditional regression models are adapted** to account for varying costs associated with prediction errors. Specifically, we investigate the implementation and performance of two cost-sensitive regression approaches proposed by Bahnsen et al. and Nikou Gunnemann. The primary goal is to evaluate the effectiveness of these approaches in handling datasets with imbalanced costs of prediction errors.

The primary goal is to **evaluate the effectiveness of these approaches in handling datasets with imbalanced costs of prediction errors** and to provide insights into their practical applicability in real-world scenarios.

2. Description of the dataset

The dataset used in this study, named *costsensitiveregression.csv*, comprises several variables that are utilized in the investigation of cost-sensitive regression approaches. The dataset is structured as follows:

2.1. Statistics of the Dataset

Here are some basic statistics of the dataset:

- Number of instances: 147636
- Number of features: 11
- Number of classes (dependent variable): 2
- Mean false negative cost (C_{FN}): 533.4049116733859
- Mean true positive cost (C_{TP}): 6
- Mean false positive cost (C_{FP}): 6
- True negative cost (C_{TN}): 0

2.2. Variables in the Dataset

1. Columns A to K represent the independent variables, denoted as $X_1, X_2, ..., X_{11}$, respectively.
2. Column L corresponds to the dependent variable, denoted as Y.
3. Column M contains the false negative cost, denoted as C_{FN}. The false negative cost varies from row to row based on the risk parameter details, implying that each instance in the dataset is associated with a distinct false negative cost value.
4. Both true positive and false positive costs are constant for all instances in the dataset. The true positive cost (C_{TP}) and false positive cost (C_{FP}) are set to 6 for all observations.
5. The true negative cost (C_{TN}) is also constant for all instances, with a value of 0.

The dataset is essential for evaluating the performance of cost-sensitive regression models, as it provides the necessary features, target variable, and cost information required for training and testing the models.

3. Algorithm Used

Two cost-sensitive regression algorithms were employed in this study to address the challenge of varying costs associated with prediction errors:

3.1. Bahnsen's Approach

Bahnsen et al. proposed a cost-sensitive regression framework that incorporates the costs of prediction errors into the objective function of traditional regression models. The key steps involved in this approach are as follows:

3.1.1 Algorithm

Algorithm 1 Bahnsen's Cost-Sensitive Regression

Training dataset (X, Y), False Negative Costs (C_{FN}), True Positive Cost (C_{TP}), False Positive Cost (C_{FP}) Optimized model parameters θ

Define Cost Function:

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

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

Optimize Cost Function:

Use an optimization algorithm (e.g., gradient descent) to minimize $J(\theta)$

Evaluate Performance:

Assess the performance of the cost-sensitive regression model using relevant evaluation metrics and validation techniques.

3.2. Nikou Gunnemann's Approach

Nikou Gunnemann introduced a cost-sensitive regression approach that adjusts the objective function of regression models to minimize prediction errors according to their associated costs. The key steps involved in this approach are as follows:

3.2.1 Algorithm

Algorithm 2 Nikou Gunnemann's Cost-Sensitive Regression

Training dataset (X, Y), False Negative Costs (C_{FN}), True Positive Cost (C_{TP}), False Positive Cost (C_{FP}) Optimized model parameters θ

Define Cost Function:

$J_c(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i(h_\theta(X_i)C_{TP} + (1 - h_\theta(X_i))C_{FNi}) + (1 - y_i)(h_\theta(X_i)C_{FP} + (1 - h_\theta(X_i))C_{TN}))$

Optimize Cost Function:

Use an optimization algorithm (e.g., gradient descent) to minimize $J_c(\theta)$

Evaluate Performance:

Evaluate the performance of the cost-sensitive regression model on testing data using relevant evaluation metrics.

These algorithms were selected based on their effectiveness in handling cost-sensitive regression tasks and were implemented to investigate their performance on the dataset under consideration.

4. Results

4.1. Bahnsen's Approach

Bahnsen's cost-sensitive regression approach achieved the following results:

- Accuracy: 0.8633500406393931

4.1.1 Classification Report:

	precision	recall	f1-score	support
0	0.88	0.93	0.90	20685
1	0.80	0.72	0.76	8843
accuracy			0.86	29528
macro avg	0.84	0.82	0.83	29528
weighted avg	0.86	0.86	0.86	29528

4.1.2 Confusion Matrix:

The confusion matrix visually represents the performance of the classifier by showing the counts of true positive, true negative, false positive, and false negative predictions.

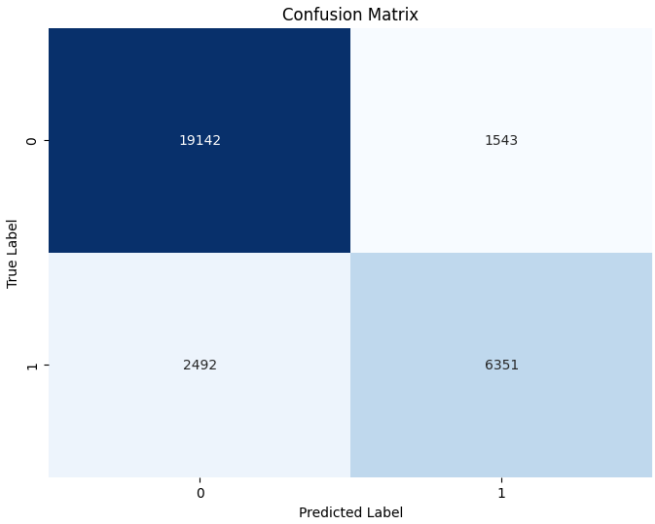

Figure 1. Confusion Matrix for Bahnsen's Approach

4.1.3 ROC Curve:

The ROC curve illustrates the trade-off between true positive rate and false positive rate across different threshold values. A higher area under the curve (AUC) indicates better performance of the classifier.

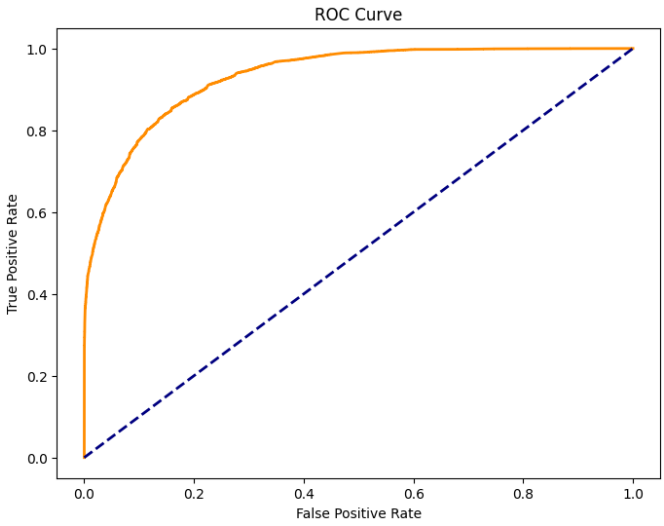

Figure 2. ROC Curve for Bahnsen's Approach

4.1.4 Precision-Recall Curve:

The precision-recall curve shows the trade-off between precision and recall for different threshold values. A higher area under the curve (AUC) indicates better performance of the classifier.

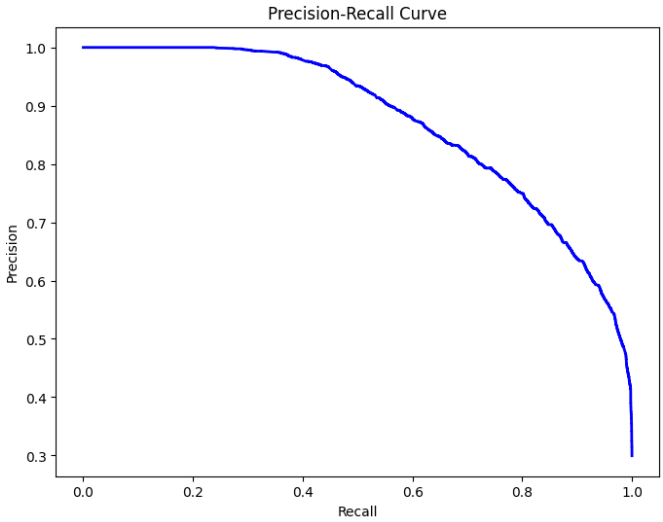

Figure 3. Precision-Recall Curve for Bahnsen's Approach

4.2. Nikou Gunnemann's Approach

Nikou Gunnemann's cost-sensitive regression approach yielded the following results:

- Mean Squared Error (MSE): 13.617555378438482
- Mean Absolute Error (MAE): 0.814404996885878
- Root Mean Squared Error (RMSE): 3.6901972004810912

4.2.1 Scatter Plot of Predicted vs. Actual Values:

The scatter plot compares the predicted values generated by the model with the actual values from the test dataset. It helps visualize the correlation between predicted and actual values.

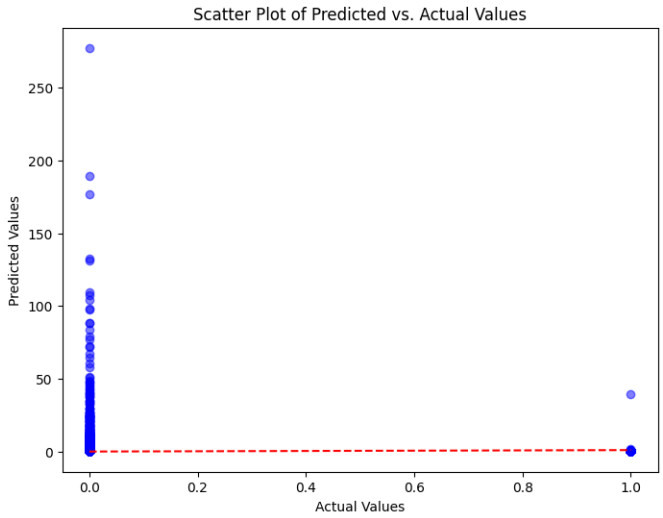

Figure 4. Scatter Plot of Predicted vs. Actual Values for Nikou Gunnemann's Approach

4.2.2 Residual Plot:

The residual plot shows the difference between predicted and actual values (residuals) plotted against the predicted values. It helps assess the homoscedasticity and linearity assumptions of the regression model.

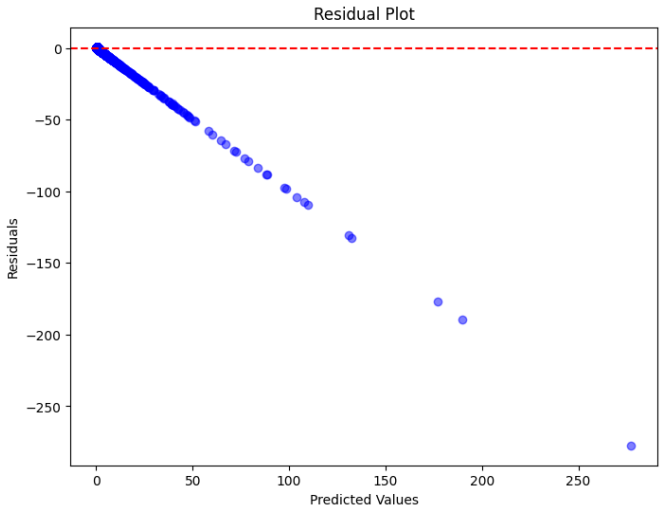

Figure 5. Residual Plot for Nikou Gunnemann's Approach

4.2.3 Distribution of Residuals:

The distribution of residuals visualizes the distribution of errors made by the regression model. It helps assess the normality assumption of the residuals.

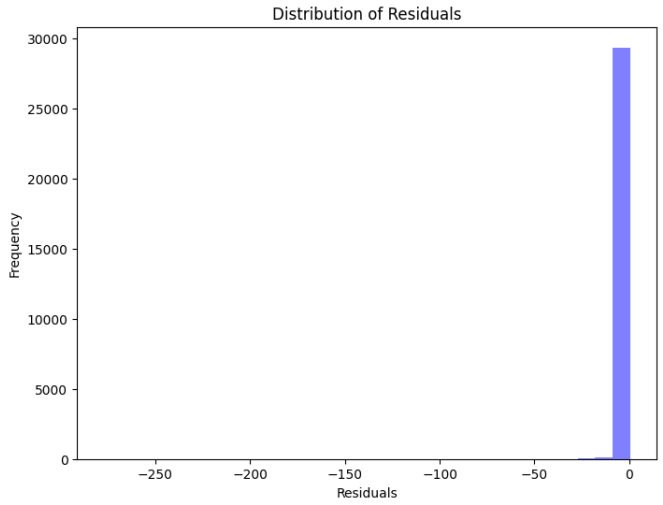

Figure 6. Distribution of Residuals for Nikou Gunnemann's Approach