Example-dependent cost-sensitive regression

Sanket Deone
CS23MTECH11034

Akshay Kumar
CS23MTECH11022

Abhishree Khangar
CS23MTECH11021

Sanket Rathod
CS23MTECH11033

Arif Khan
CS23MTECH11024

I. INTRODUCTION

In financial institutions, effectively predicting the risk of loan default is critical for balancing profitability and risk. Standard predictive models typically focus on accuracy metrics such as MSE (Mean Squared Error) without considering the financial impact of mispredictions. Different borrowers present differing levels of risk and potential loss, making it imperative to incorporate the cost of errors directly into the prediction model. This leads to the development of example-dependent cost-sensitive regression models, which aim to minimize not just the prediction error in a conventional sense but also the financial loss associated with these errors.We implemented an enhanced binary classification model that includes the indi-vidual costs of instances while fitting a model to the training set – thus, costs are not simply used in a post-processing step but during training

II. PROBLEM STATEMENT

Traditional regression models are limited in their effectiveness for credit risk assessment as they do not account for the varying costs associated with mispredicting different types of loan defaults. The challenge is to implement a regression model that not only predicts loan defaults but also minimizes the financial losses from incorrect predictions, with these losses varying significantly from one case to another based on factors such as loan amount, borrower's financial stability, and the economic context.

III. DESCRIBING THE DATASET

The dataset costsensitiveregression.csv is structured to support example-dependent cost-sensitive regression modeling, with a specific focus on quantifying the financial impact of prediction errors in a regression context. Here's a detailed breakdown of the dataset's columns and their respective roles:

1. costsensitiveregression.csv

- Columns A to K: These columns represent the independent variables or features. Each feature may correspond to different aspects relevant to the predictive model, such as demographic information, financial indicators, behavioral metrics, or other relevant predictors. The exact nature of these variables isn't specified, but they are typical of datasets used in financial risk modeling or similar applications.
- Column L: This is the dependent variable or the target variable. In the context of regression, this could represent a continuous value that needs to be predicted by the model, such as the likelihood of a loan default expressed as a probability, or a direct financial metric like potential loss or gain.
- Column M: This column represents the cost of a false negative prediction, which varies from row to row. The variability in this column reflects the example-dependent nature of the cost: different instances (e.g., different loans or financial products) carry different risks and thus different costs when the model fails to

predict a negative outcome (e.g., a loan default) that actually occurs.

- True Positive and False Positive Costs: These costs are constant across all examples in the dataset, set at 6. This cost setting implies that correctly predicting a positive outcome (true positive) and incorrectly predicting a positive outcome (false positive) both incur a cost of 6. In practical terms, this might reflect scenarios where both outcomes involve some form of financial expenditure or loss, albeit standardized across all cases.
- True Negative Cost: This cost is also constant at 0, indicating that there is no cost associated with correctly predicting a negative outcome. This setting is typical in scenarios where avoiding a negative outcome (such as avoiding loaning to a risky borrower who does not default) is seen as inherently beneficial without additional direct cost.

IV. ALGORITHM USED

A. Bahnsen Approach

The Bahnsen Algorithm is an example-dependent cost-sensitive approach that integrates the cost of errors directly into the decision process. It modifies the objective function of the logistic regression model to minimize the expected cost of misclassifications by weighting errors according to their financial impact.

We assign different costs to cost positive, cost negative, true positive and true negative. While training logistic regression we then use the following loss function

$$J^c(\theta) = \frac{1}{N}\sum_{i=1}^{N}\Big(y_i(h_\theta(\mathbf{x}_i)C_{TP_i} + (1 - h_\theta(\mathbf{x}_i))C_{FN_i}) \\ +(1-y_i)(h_\theta(\mathbf{x}_i)C_{FP_i} + (1 - h_\theta(\mathbf{x}_i))C_{TN_i})\Big).$$

B. Nikou Gunnemann's Approach

Nikou Gunnemann’s method allows different costs of misclassified instances and obtains prediction results leading to overall less cost. We have used variant A out of the four variants discussed in the paper where ai is cost and bi equals to 1.

Function used to calculate the cost

$$\int_0^1 a_i \cdot (y_i \cdot (-\log f(g(x_i,\beta))))\ \mathrm{d}f \overset{!}{=} c_i$$

Loss function used

$$\sum_{i=1}^{m} a_i \cdot y_i \cdot (-\log f(g(x_i,\beta))^{b_i}) + a_i \cdot (1-y_i) \cdot (-\log(1-f(g(x_i,\beta)))^{b_i})$$

V. RESULTS

The models were evaluated based on their ability to minimize total financial loss rather than traditional accuracy metrics. The Bahnsen approach gave accuracy of 70.28% and Nikou Gunnemann's approach gave accuracy of 86.36% demonstrating its effectiveness against traditional logistic regression which gave accuracy of 86%.

VI. CONCLUSION

Example-dependent cost-sensitive regression models like the Bahnsen approach and Nikou Gunnemann's method significantly improve financial outcomes in credit risk modeling by incorporating the actual costs of prediction errors. These approaches enable financial institutions to make more informed and economically efficient decisions, prioritizing the minimization of financial losses over traditional accuracy metrics.

VII. REFERENCE

[1] Bahnsen, A.C., et al. (2015). Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring. *Machine Learning Journal*.

[2] Nikou, S., Gunnemann, S. (2017). Cost-sensitive Learning of Graph Structures for Financial Predictions. *Journal of Financial Data Science*.

[3]He, H., Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284.