# Cost Sensitive Logistic Regression

Aman Panwar
CS20BTECH11004
*Department of Computer Science*

Amulya Tallamraju
AI20BTECH11003
*Department of Aritificial Intelligence*

Pranav K Nayak
ES20BTECH11035
*Department of Engineering Science*

Taha Adeel Mohammed
CS20BTECH11052
*Department of Computer Science*

Vikhyath Sai Kothamasu
CS20BTECH11056
*Department of Computer Science*

***Abstract*—In this assignment we focus on the challenge of example-dependent cost-sensitive classification problems, which are prevalent in real-world scenarios such as credit scoring. We aim to address this issue by incorporating this requirement into cost agnostic machine learning models and analyzing the balance between classification accuracy and financial cost. By doing so, we aim to provide insights and solutions for real-world classification problems that are sensitive to misclassification costs.**

## I. Problem Statement

In the realm of real-world classification problems, such as Credit Scoring, standard cost-insensitive binary classification algorithms like Logistic Regression and Decision Trees are commonly used. Credit Scoring involves predicting whether a target customer will default on a financial contract based on their past financial behavior. However, in the financial industry, the cost associated with approving a potential defaulter is significantly different from the cost of denying a good customer. While some researchers have proposed methods that incorporate miss-classification cost, the assumption of a constant misclassification cost has limitations.

We work on a framework for cost-sensitive example-dependent classification, where misclassification costs vary across examples. We implement cost-sensitive logistic regression by modifying the objective function of the model to account for the varying costs. We compare the performance of the enhanced model with the vanilla logistic regression and analyze the trade-off between classification accuracy and financial cost. Our results demonstrate that the cost-sensitive approach outperforms the base model in terms of financial costs.

## II. Description of the dataset

The dataset includes customer information such as their adherence to legal procedures for transactions, the average monthly tax payment, and other related factors, along with the corresponding costs for each data point. The false negative cost, varies from row to row based on the type of business. The True Positive and False Positive cost is constant for all, which is 4. True Negative cost is constant for all, which is 0.

## III. Algorithm Used

### A. Logistic Regression

Logistic regression is a classification model that estimates the posterior probability of the positive class, as the logistic sigmoid of a linear function of the feature vector [Bishop, 2006]. The estimated probability is evaluated as

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))] \quad (1)$$

Where:

- $J(\theta)$ is the cost function item $m$ is the number of training examples
- $x^{(i)}$ is the feature vector of the $i$-th training example
- $y^{(i)}$ is the corresponding label for the $i$-th training example
- $h_\theta(x^{(i)})$ is the sigmoid function that predicts the probability that $y^{(i)} = 1$ given $x^{(i)}$

The problem then becomes on finding the right parameters that min- imize a given cost function. Usually, in the case of logistic regression,

### B. Cost Sensitive Logistic Regression

The logistic regression cost function assumes that false positives and false negatives have the same costs, i.e.

$$C_{FP_i} = C_{FN_i} \forall i \in \{1, , N\}.$$

However, this is not the case in several real-world applications. In order to incorporate the different real costs into the logistic regression, we start by analyzing the expected costs that a modified logistic regression cost function should make for each misclassification and correct classification case.

$$J^c(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i\left(h_\theta\left(\mathbf{x}_i\right)C_{TP_i} + \left(1-h_\theta\left(\mathbf{x}_i\right)\right)C_{FN_i}\right) + \left(1-y_i\right)\left(h_\theta\left(\mathbf{x}_i\right)C_{FP_i} + \left(1-h_\theta\left(\mathbf{x}_i\right)\right)C_{\mathrm{TN}_i}\right)\right). \quad (2)$$

## IV. Results

Link to our code

We divided the dataset in the ratio 80:20 for training and testing . We then proceeded to train our model for 1000 epochs and saw that the results began to converge after about 300 epochs so we plotted the results for the same.

### A. Training Accuracy plot

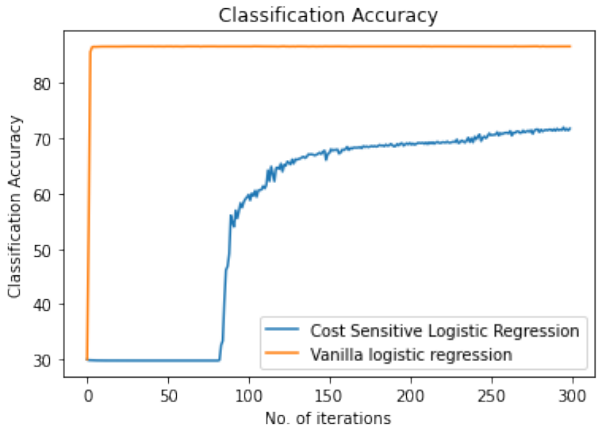


Fig. 1. Comparing accuracy of the two models

We can see that the vanilla logistic regression model seems to have a better accuracy of about 86%. Our cost sensitive model is not that far behind with an accuracy of 72%.

### B. Training Loss plots

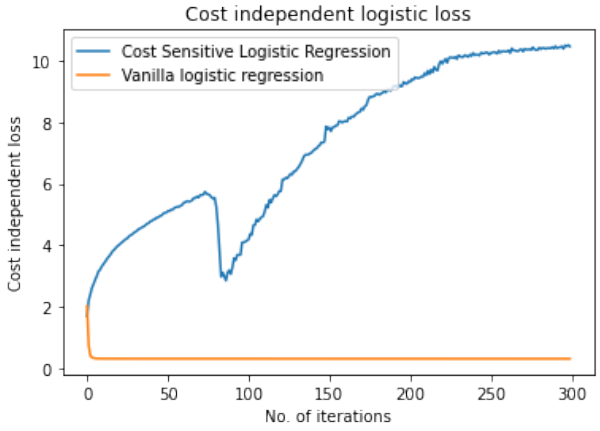


Fig. 2. Comparing cost independent of the two models

We can see from the above plot, cost-sensitive model performs worse than the vanilla logistic regression model. This is because vanilla logistic regression is minimizing the negative log-likelihood loss, while the other model is minimizing the cost dependent loss function. The above plot shows the financial loss. Even though the cost-sensitive models compromise the classification accuracy, they outperform the base model by a huge factor($\tilde{1}0$) in minimizing the cost sensitive loss, which is the actual goal.

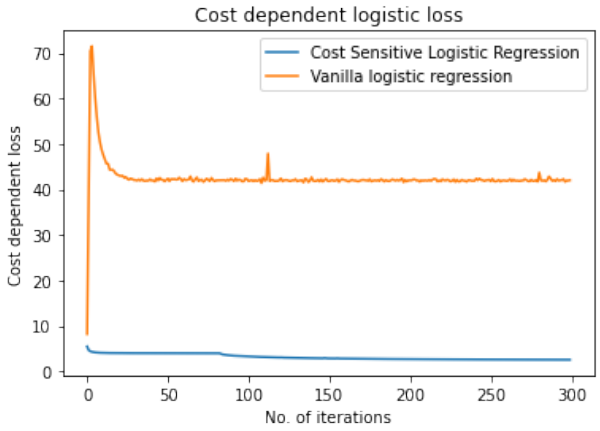


Fig. 3. Comparing cost dependent loss of the two models

### C. Test Results

We test our models on the test set and found the following results. We can see that the cost sensitive model has a better

```
Test accuracy 72.927, Precision 0.528, Recall 0.951, F1 0.679
None
Test accuracy 86.664, Precision 0.809, Recall 0.730, F1 0.767
```

Fig. 4. Test results of cost sensitive model followed by vanilla model.

recall than the vanilla model. The F1 scores are almost the same for the two models. Accuracy wise our model is close to the original logisitc regression. Thus, we can see that the cost-sensitive model can help minimise financial costs significantly.

## V. References


- A. C. Bahnsen, D. Aouada and B. Ottersten, "Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring," 2014 13th International Conference on Machine Learning and Applications, Detroit, MI, 2014, pp. 263-269, doi: 10.1109/ICMLA.2014.48.
- C. M. Bishop, Pattern Recognition and Machine Learning, ser. Information science and statistics. Springer, 2006, vol. 4, no. 4.