Logistic Regression Over UCI German Credit Data

Puhe Hao
Nanjing University of Posts and Telecommunications, Nanjing, China
B20031122@njupt.edu.cn

***Abstract*—Credit risk assessment is a critical aspect of lending decisions for financial institutions, influencing both profitability and risk management strategies. Logistic regression, a widely-used statistical technique, offers a powerful framework for modeling binary outcomes, making it particularly relevant for credit scoring applications. In this study, we leverage logistic regression to analyze the UCI German credit dataset, a comprehensive repository of applicant information commonly used for evaluating credit risk models. Our objective is to develop a logistic regression model capable of predicting the likelihood of credit default based on applicant characteristics. By examining the dataset's socio-economic indicators, credit history, and demographic attributes, we aim to uncover the underlying relationships between these variables and credit risk. Through model development, feature importance analysis, and performance evaluation, we seek to enhance our understanding of credit risk factors and contribute to the refinement of credit scoring methodologies. Ultimately, our study aims to provide insights that can assist lending institutions in making more informed and equitable lending decisions, thereby promoting financial stability and responsible lending practices.**

***Index Terms*—Logistic Regression;UCI German Credit Data**

I. Introduction

In the realm of finance, credit risk assessment plays a pivotal role in decision-making processes for lending institutions. The ability to accurately predict the creditworthiness of applicants not only mitigates potential financial losses but also fosters responsible lending practices. In this context, logistic regression emerges as a powerful tool for modeling the probability of binary outcomes, making it particularly well-suited for credit risk analysis.

This paper focuses on employing logistic regression to analyze the UCI German credit dataset [2], a benchmark dataset widely used for evaluating credit scoring models. The dataset comprises a comprehensive set of features, including socio-economic indicators, credit history, and other demographic attributes, associated with credit applicants. Leveraging this dataset, our study aims to develop a logistic regression model capable of predicting the likelihood of credit default based on applicant characteristics.

The German credit dataset offers a rich source of information, making it conducive to exploring the intricacies of credit risk assessment. By leveraging logistic regression [1], we aim to uncover the underlying relationships between the predictor variables and the likelihood of credit default. Through this analysis, we seek to achieve the following objectives:

- Model Development: Develop a logistic regression model trained on historical credit data to predict the probability of credit default.
- Model Evaluation: Assess the performance of the logistic regression model using appropriate evaluation metrics, such as accuracy, loss.
- Experiment and analysis: we implement our algorithm using pytorch and test our model according to multiple metrics. The results show the efficiency of our model.

II. Background

Logistic regression is a kind of generalised linear model (generalized linear model), which has a lot in common with multiple linear regression analysis. Their models are basically the same, both have $wx+b$, where w and b are the parameters to be solved, and the difference lies in the difference of their dependent variables. Multiple linear regression directly takes $wx+b$ as the dependent variable, i.e., y = $wx+b$, whereas logistic regression corresponds $wx+b$ to a hidden state p through the function L, $p = L(wx+b)$, and then decides the value of the dependent variable according to the magnitude of p versus $1-p$. The logistic regression is a generalised linear model. the value of the dependent variable. If L is a logistic function, it is logistic regression, and if L is a polynomial function, it is polynomial regression.

In more general terms, logistic regression adds a layer of logistic function calls to the linear regression.Logistic regression is mainly for binary prediction, we have talked about the Sigmod function in the activation function, the Sigmod function is the most common logistic function, because the output of the Sigmod function is for the probability of the value between 0 and 1, when the probability is greater than 0.5 predicted 1, less than 0.5 predicted 0.

III. Methodology

A. Data Preprocessing

UCI German Credit is UCI's German Credit dataset with raw and numerical data.The German Credit data is a dataset that predicts the propensity to default on a loan based on an individual's bank loan information and the occurrence of loan delinquency for the applying customer. The dataset contains 1,000 pieces of data in 24 dimensions.

The UCI German credit dataset is loaded and examined for missing values or anomalies. Each feature column undergoes normalization by subtracting the mean and dividing by the standard deviation to ensure uniform scaling. The dataset is

This work was Completed for the graduate course of paper writing in Nanjing University of Posts and Telecommunications .

shuffled to remove any inherent biases, and subsequently split into training and testing sets in a 90:10 ratio.

B. Logistic Regression Model

A logistic regression model is implemented using PyTorch, a popular deep learning framework. The model architecture consists of a single linear layer followed by a sigmoid activation function, designed to predict the probability of credit default based on applicant features. The model is defined to have 24 input features corresponding to the attributes provided in the German credit dataset.

Distinguish between training set and test set, since there is no validation set here, we directly use the accuracy of the test set as a criterion for judging the goodness of the test set. The Distinction rule is 900 for training, 100 for testing. The format of german.data-numeric is that the first 24 columns are 24 dimensions, and the last one is the label to be typed (0, 1), so we distinguish the data and label together

C. Training Procedure

The logistic regression model is trained using stochastic gradient descent with the Adam optimizer. The training loop iterates over a predefined number of epochs, with the model parameters updated in each iteration to minimize the cross-entropy loss between predicted and actual credit default labels. During training, the model's performance is evaluated on the training data using predefined evaluation metrics, including loss and accuracy.

D. Model Evaluation

Following model training, the performance of the logistic regression model is assessed on an independent test dataset. The trained model's predictions are compared against the ground truth labels to compute evaluation metrics such as accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC). These metrics provide insights into the model's ability to accurately classify credit applicants as either low or high risk of default.

IV. Experiments Analysis

We implement The model using pytorch. The experiment test training loss and accuracy.The training loss (blue line) represents how well the model is fitting the training data. As the model learns from the training data, we expect the training loss to decrease over epochs. This decrease indicates that the model is becoming more accurate in predicting the training labels. The test loss (red dots connected by a line) represents how well the model generalizes to unseen data. A decreasing test loss indicates that the model is generalizing well to new examples and is not overfitting to the training data. In the initial epochs, both training and test losses might be high as the model starts learning. As training progresses, the losses typically decrease and stabilize. If the training loss continues to decrease while the test loss starts increasing, it suggests overfitting, where the model memorizes the training data too well and fails to generalize.

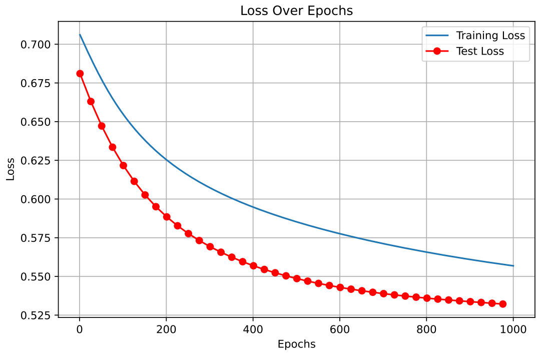

Fig. 1: Loss performance

As shown in Fig 2, accuracy (green dots connected by a line) represents the proportion of correctly classified examples out of the total examples in the test dataset. An increasing accuracy over epochs indicates that the model is improving in its ability to correctly classify examples. Similar to loss, accuracy may start low and increase as the model learns from the training data. However, if the training accuracy increases significantly while the test accuracy stagnates or decreases, it suggests overfitting. High accuracy along with low loss indicates that the model is performing well both on the training and test datasets, striking a good balance between fitting the training data and generalizing to unseen data.

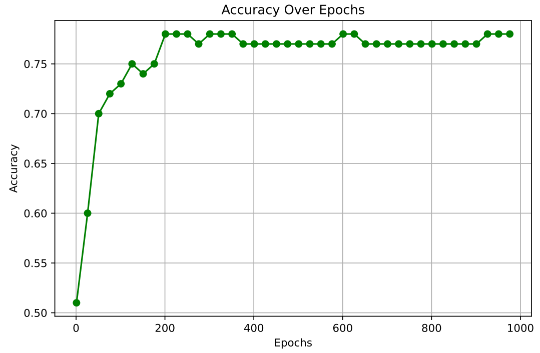

Fig. 2: Accuracy performance

V. Conclusion

In this study, we employed logistic regression to analyze the UCI German credit dataset and predict the likelihood of credit default based on applicant characteristics. Through model development, evaluation, and interpretation of results, we gained valuable insights into credit risk assessment and predictive modeling techniques in the financial industry.

References

[1] https://web.stanford.edu/ jurafsky/slp3/5.pdf
[2] https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data