International Journal of Informatics and Data Science Research
ISSN 2997-3961 (Online)
Vol. 2, No. 6, Jun 2025,
Available online at: https://scientificbulletin.com/index.php/IJIDSR

Advancing Credit Scoring: Practical Methodologies for Modern Risk Modeling with Machine Learning

Anvar Zokhidov
Former Senior AI Manager, Tenge Bank (Halyk Group) | Member, AICA.uz

Article information:
Manuscript received: 4 Apr 2024; **Accepted:** 10 May 2024; **Published:** 25 Jun 2025

Abstract: In today's world, credit scoring is an essential part of financial institutions, banks and companies that grant loans to customers as it allows to assess how likely a customer will return their loan on time. A credit scoring model is a tool that calculates the Probability of Default on a loan.

Introduction

Not long ago, people would just manually check out a few basic factors of their clients' background. Then traditional mathematical models appeared that started calculating the Probability of Default that greatly helped in scoring. Time has passed and now more and more machine learning models are replacing traditional scoring methods. This is because these models have revolutionized the way of scoring by being able to take as an input a client's data such as age, region, credit history and behavioral data stacked together and give a precise number as an output showing how likely the client will go default. One of the mostly used models nowadays is Logistic Regression.

The role of Logistic Regression

Logistic Regression is considered almost a default algorithm among banks due to its simplicity, effectiveness and interpretability. It is also easy to train and deploy on many platforms by being not robust. However, arguably the biggest advantage of Logistic Regression over other, more complex, models is its temporal stability – that is, its ability to maintain performance when used to data from future time periods, outside the training window. Therefore, it relatively does not need to be re-trained frequently.

Logistic Regression is a supervised machine learning algorithm used for classification that predicts a discrete or categorical outcome. The key idea is to model the probability that the target variable equals 1 given the input features. The relationship is expressed as:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

- $P(Y = 1 \mid X)$ is the probability of the event occurring (e.g., default).
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for the input features $X_1, X_2, \ldots, X_n$.

The exponentiated value in the logistic formula represents a linear regression model with a linear combination of the input features. However, instead of using this linear output directly

to predict an outcome, logistic regression applies the exponential function and then the sigmoid (logistic) function to transform this value into a probability between 0 and 1. The closer the value to 1 the higher the probability of an event occurring and vice versa.

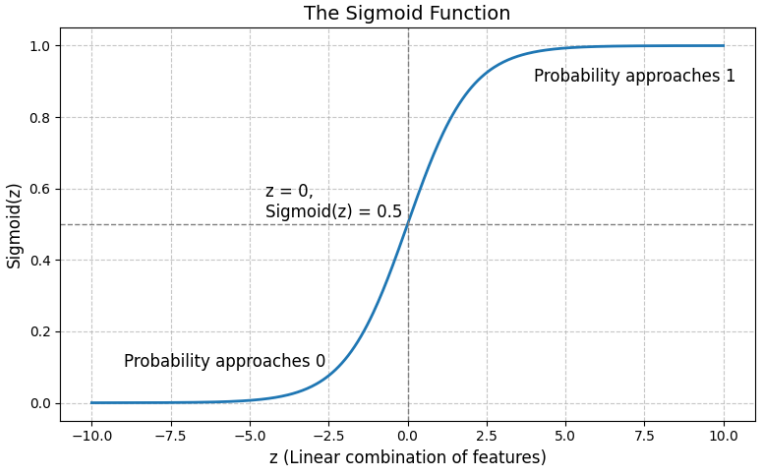

Figure 1: The Sigmoid Function.

As a result, logistic regression can be viewed as using a linear regression model as an intermediate step, and then wrapping it in a non-linear function (the sigmoid) to produce probabilities that are suitable for classification problems—such as credit scoring.

Data Collection and Preparation

An essential foundation for building an effective credit scoring model is the quality and preparation of data. Banks use different types of data based on their own business model and protocols. Common client data to be used are usually age, marital status, official income, region and credit history.

The data preparation phase is equally important and involves several key steps. Data cleaning is performed to remove outliers and duplicate records. Preprocessing involves normalizing numerical values into a common range as well as encoding categorical features. Addressing missing values is critical, as incomplete data can break the model; common techniques include removal or taking the mean, median and mode of the data points. Feature engineering is a key step as it allows to extract valuable information from behavioral data and use them as new features to the model.

Throughout the data collection and preparation process, it is mandatory to account for data privacy. Clients' personal and financial information must be handled securely and in compliance with relevant data protection regulations. Encryption of this information while the dataset is being prepared to fit the model is what differs financial models from others.

Building the Credit Scoring Model

Once data collection and preparation are complete, the next stage is building the credit scoring model.

The initial step is to split the dataset into training and testing subsets. The training set is used to train the logistic regression model and the testing set is kept for evaluating how well the model performs on unseen data, which is called generalization. Splitting varies depending on the task, dataset and the business goal but a typical split is 70% training and 30% testing if the dataset is large enough, or 80% training and 20% testing if it is small.

After splitting the dataset, the logistic regression model is trained using the training set. The model learns by adjusting its parameters (coefficients) and minimizing a loss function which

is the log-loss or binary cross-entropy for logistic regression.

The process involves:

- ✓ Feeding the input features (e.g., age, region, income, credit history) to the model.
- ✓ Using the sigmoid function to produce a probability of default.
- ✓ Iteratively adjusting the coefficients to reduce prediction errors.

The final step involves evaluating the model's performance. In credit scoring evaluating metrics can be Accuracy, Precision and Recall and F1 Score but the widely used one is ROC-AUC Score (Receiver Operating Characteristic – Area Under the Curve).

Interpreting and Explaining the Model

Interpretability is one of the key reasons Logistic Regression remains widely used in credit scoring, especially when explaining stakeholders why an exact decision has been made.

In logistic regression, its coefficients directly influence a prediction. The larger the absolute value of a coefficient, the more influence that feature has on the predicted probability of default.

- ➢ A positive coefficient increases the likelihood of default (i.e., pushes the probability closer to 1).
- ➢ A negative coefficient decreases the likelihood of default (i.e., pushes the probability closer to 0).

Ranking the learned coefficients of the model helps in identifying the most and least valuable features of the model. Regularization helps with that, especially the Lasso also known as L1 regularization.

Finally, based on the business need a threshold between 0 and 1 is picked up for the model which means the model’s prediction above the threshold will reject a customer’s application for a loan and the prediction below the threshold will approve it.

Conclusion

Despite Logistic Regression being a simple algorithm, it is still widely used for credit scoring in many financial institutions and banks. This is due to its stability and interpretability when trying to explain the stakeholders the logic behind a decision made. Deploying Logistic Regression is fairly easy in almost any decision-making system which makes it even more of a favorite option among other algorithms.

Reference

1. IBM. (n.d.). *Logistic regression*. IBM. https://www.ibm.com/think/topics/logistic-regression
2. Wikipedia contributors. (2024, June 16). *Logistic regression*. In *Wikipedia*. https://en.wikipedia.org/wiki/Logistic_regression
3. Mathew, A. (2017). *Credit scoring using logistic regression* [Master’s project, San Jose State University]. SJSU ScholarWorks. https://doi.org/10.31979/etd.3czc-rhe3
4. Jurafsky, D., & Martin, J. H. (2025). Logistic regression (Chapter 5). In *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed., draft, January 12, 2025). Stanford University. https://web.stanford.edu/~jurafsky/slp3/5.pdf
5. Dembinsky, D., Lucieri, A., Frolov, S., Najjar, H., Watanabe, K., & Dengel, A. (2025, June 18). *Unifying VXAI: A systematic review and framework for the evaluation of*

explainable AI (arXiv:2506.15408) [Preprint]. arXiv. https://arxiv.org/abs/2506.15408

6. Breck, E., Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2019). Data validation for machine learning. In *Proceedings of the 2nd SysML Conference (MLSys 2019)*. https://mlsys.org/Conferences/2019/doc/2019/167.pdf
7. Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489
8. MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (pp. 281–297). University of California Press. https://projecteuclid.org/euclid.bsmsp/1200512992