FICO Score Quantization for Credit Scoring: A Machine Learning Perspective

Franck A. Olilo

Abstract

Credit risk assessment is of paramount importance in the financial industry, guiding lenders in making informed decisions regarding loan approvals and interest rates. This research presents a holistic analysis of a dataset, encompassing the application of logistic regression to construct a probability of default (PD) model, estimating loss given default (LGD), and introducing the innovative use of Weight of Evidence (WoE) and Information Value (IV) to assess and rank the predictive power of diverse risk factors. Additionally, a novel quantization technique is proposed to transform continuous variables into discrete counterparts. Upon concluding this paper, readers are equipped to extend their research efforts by applying the quantization technique to all available risk factors within the dataset, subsequently comparing the predictive performance against their own models.

Introduction

In recent years, the intersection of machine learning (ML) and finance has given rise to new paradigms in credit risk assessment. ML techniques offer the potential to extract intricate patterns from vast datasets, thereby enhancing our ability to predict credit risk with greater accuracy and granularity. This paper embarks on a journey to explore the integration of machine learning into the quantization of FICO scores, aiming to provide a machine learning perspective on credit scoring.

This paper unfolds as follows: Section 2 provides an overview of related work in credit scoring and FICO score quantization. Section 3 presents our methodology, detailing the machine learning algorithms employed and the datasets used for experimentation. Section 4 presents the empirical results of our study and comparative analyses. Section 5 discusses the implications of our findings and their relevance in the context of credit risk assessment. Finally, Section 6 offers concluding remarks and avenues for future research.

By the culmination of this paper, we aim to shed light on the transformative potential of a machine learning perspective in the quantization of FICO scores for credit scoring, ultimately contributing to the evolution of credit risk assessment practices.

1-Dataset

The dataset utilized in this paper comprises 10,000 customers, each characterized by their customer ID, default status, and six distinct risk factors outlined below:

Credit_line: The number of open credit accounts.

Loan_amt_outstanding: The amount of the loan left to pay.

Total Debt_outstanding: Total debt left to be paid by customers.

Income: The customer's income source, which could be their salary, pension, or other assets.

Year_employed: The number of years the customers have been employed.

Fico_score: 300 – 850

Default: 0 or 1

It's essential to note that in real-world scenarios, additional risk factors may come into play, depending on the specific models being

considered, such as mortgage loans, auto loans, commercial loans, etc. As for the source of this dataset, it cannot be disclosed to maintain confidentiality.

	credit_lines_outstanding	loan_amt_outstanding	total_debt_outstanding	income	years_employed	fico_score	default
0	0	5221.545193	3915.471226	78039.38546	5	605	0
1	5	1958.928726	8228.752520	26648.43525	2	572	1
2	0	3363.009259	2027.830850	65866.71246	4	602	0
3	0	4766.648001	2501.730397	74356.88347	5	612	0
4	1	1345.827718	1768.826187	23448.32631	6	631	0
...	...	...	...	...	...	...	...
9995	0	3033.647103	2553.733144	42691.62787	5	697	0
9996	1	4146.239304	5458.163525	79969.50521	8	615	0
9997	2	3088.223727	4813.090925	38192.67591	5	596	0
9998	0	3288.901666	1043.099660	50929.37206	2	647	0
9999	1	1917.652480	3050.248203	30611.62821	6	757	0

Figure 2.1- Partial Dataset

2-Exploratory Data Analysis

Before embarking on the model-building process, the initial step following data loading is data examination. Exploratory data analysis (EDA) proves invaluable for validating model assumptions, identifying influential data points, and pinpointing outliers. During EDA, we also investigate relationships between various variables.

In Figure 2.1, we present an overview of the dataset. Given its substantial size, our primary concern initially should be the detection of missing values. We can affirm that our dataset is devoid of any missing values. Absence of missing data not only streamlines preprocessing efforts but also enhances the robustness of standard deviation calculations as shown below

	credit_lines_outstanding	loan_amt_outstanding	total_debt_outstanding
count	10000.000000	10000.000000	10000.000000
mean	1.461200	4159.677034	8718.916797
std	1.743846	1421.399078	6627.164762
min	0.000000	46.783973	31.652732
25%	0.000000	3154.235371	4199.836020
50%	1.000000	4052.377228	6732.407217
75%	2.000000	5052.898103	11272.263740
max	5.000000	10750.677810	43688.784100

	income	years_employed	fico_score	default
count	10000.000000	10000.000000	10000.000000	10000.000000
mean	70039.901401	4.552800	637.557700	0.185100
std	20072.214143	1.566862	60.657906	0.388398
min	1000.000000	0.000000	408.000000	0.000000
25%	56539.867903	3.000000	597.000000	0.000000
50%	70085.826330	5.000000	638.000000	0.000000
75%	83429.166133	6.000000	679.000000	0.000000
max	148412.180500	10.000000	850.000000	1.000000

Figure 2.2- Mean, Standard deviation, Min,Max

Subsequently, we proceed to assess the assumptions of logistic regression. Most of these assumptions are met, as our response variable is binary, our observations are independent, there are no extreme outliers, and our sample size is sufficiently large. One remaining assumption to consider is multicollinearity, as depicted in Figure 2.3 through a pairwise correlation heatmap.

Using correlation as our metric, multicollinearity is identified by a score greater than or equal to 0.8. Figure 2.3 reveals that the following variables exhibit a noticeable correlation:

- Credit line outstanding and default: 0.86

- Loan outstanding and default: 0.84

- Total debt outstanding and loan outstanding: 0.85

- Income and loan outstanding: 0.84

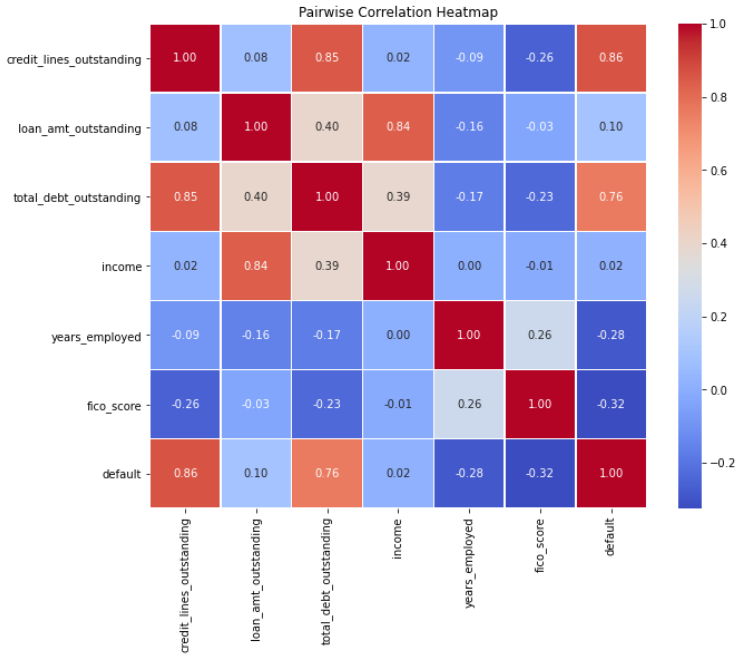

Figure 2.3-Correlation score

Another issue worth examining is the imbalance in observations between defaults and non-defaults, as illustrated in Figure 2.4. An imbalanced dataset can hinder the model's ability to effectively learn from the data, potentially resulting in a lower precision score.

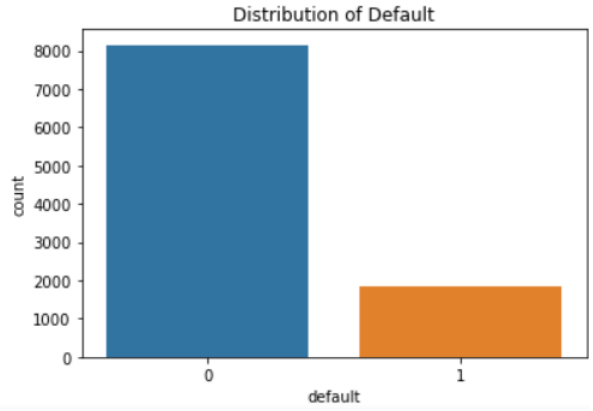

Figure 2.4-Imbalanced Visual

Given the limited number of risk factors available, outright deletion of a risk factor is not a viable option in this study. Therefore, we opt to explore more advanced methods for detecting multicollinearity, such as Variance Inflation Factor (VIF) or Tolerance, and additionally assess the weight of evidence and information value associated with each risk factor to gauge their predictive importance. This approach aligns with the objectives of our research.

3-Feature Selection/ WoE and IV

WoE and IV play a pivotal role in the realm of credit risk, particularly when crafting Probability of Default (PD) models. Let's elucidate these concepts:

- WoE (Weight of Evidence): measures the strength of the relationship between an independent variable and the dependent variable. Moreover, it facilitates the transformation of continuous variables into discrete ones.

- IV (Information Value): IV quantifies the predictive power of a categorical variable, aiding in the ranking of variables based on their significance.

In Figure 3.1, we present the IV values for each of our risk factors. Notably, we can discern that "credit line outstanding" and "total debt outstanding" possess IV scores exceeding 1, indicating a potentially substantial predictive power. Conversely, "loan amount outstanding" exhibits an approximate score of 0.02, reflecting a relatively weak association with the dependent variable. The remaining three risk factors fall within a favorable range.

```
Information Value (IV) for each feature
credit_lines_outstanding    12.026096
loan_amt_outstanding         0.020835
total_debt_outstanding       1.064116
income                       0.003133
years_employed               0.304898
fico_score                   0.223547
dtype: float64
```

Figure3.1- Risk Factors IV scores

It's important to note that at this stage, we refrain from making hasty decisions regarding data transformation or removal. Such actions would be contingent on the accuracy of the logistic regression model. Only if we obtain notably poor AUC or accuracy scores would we contemplate such modifications.

4-Logistic Regression: PD Model

With all our initial assumptions verified, we are poised to commence constructing our model, the purpose of which is to forecast the likelihood of a customer defaulting. This prediction will subsequently serve as the foundation for estimating the potential loss incurred in the event of a default. It is worth noting that, based on our current understanding, we have opted to incorporate all the available risk factors into our model.

During the Exploratory Data Analysis (EDA) phase, we detected an imbalance within our dataset. Consequently, our approach involves an initial split of the dataset into training and testing subsets. Subsequently, we employ Synthetic Minority Over-sampling Technique (SMOTE) to address this imbalance, ensuring the training data is appropriately balanced. We should also know that WoE transformation of variables could also be used to manage imbalanced without adding more data set

Figure 4.1 visually represents the Receiver Operating Characteristic (ROC) curve of our logistic regression model. The curve illustrates the model's performance, which appears to be nearly flawless, a conclusion corroborated by

metrics such as the Area Under the Curve (AUC), precision, and Gini score.

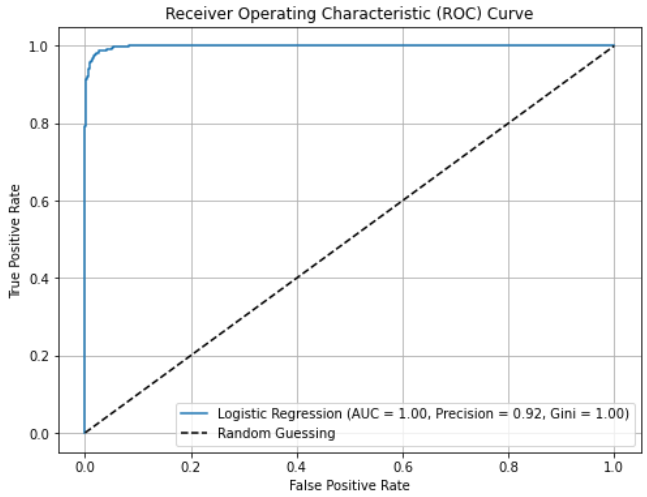

Figure 4.1-ROC curve

Utilizing the provided code, specifically the "loan properties function," one can calculate the anticipated loss by inputting a customer's loan properties. It is essential to note that in these calculations, we have considered a recovery rate of 10%, which plays a pivotal role in determining the estimated loss.

5-FICO Quantization

Once again, during the Exploratory Data Analysis (EDA) phase, we identified correlations among certain risk factors. It's noteworthy that no feature engineering was applied to any of these factors. However, one potential transformation we contemplated was converting continuous variables into discrete ones. For the scope of this paper, we've elected to execute this transformation exclusively for the FICO SCORE.

We harbor the suspicion that the FICO score holds significant predictive power concerning a customer's likelihood of default. To harness this potential, we're embarking on the creation of a rating map that discretizes the FICO score into distinct buckets. This process is commonly known as quantization. Quantization serves as a mechanism to optimize various characteristics of the resulting buckets, often through metrics like Mean Squared Error (MSE) or log likelihood.

To elucidate, let's consider the FICO score range, which spans from 300 to 850. Our aim is to segregate this range into three distinct buckets:

- **High Risk**: Scores ranging from 300 to 599.

- **Moderate Risk**: Scores ranging from 600 to 699.

- **Low Risk**: Scores ranging from 700 to 850.

While this initial classification appears reasonable, we decided to employ quantization to identify the optimal FICO score boundaries by comparing the MSE or log likelihood. For the purposes of this paper, we have chosen to work with five buckets. Initially, the MSE of the initial classification was 6.34. However, after utilizing our "find_optimal_boundaries" function in the code, we managed to identify optimal boundaries that resulted in an MSE of 2.84, enhancing the precision of our FICO score-based risk assessment. Future researcher can research the impact of this classification on the initial AUC score.

Conclusion/Future Work

Our research demonstrates FICO score quantization's potential in revolutionizing credit risk assessment. By leveraging advanced algorithms and data-driven insights, credit scoring can become more precise and adaptable. Machine learning's transformative power in finance is evident. Future research avenues include applying WoE techniques to categorize all risk factors, comparing predictive performance, and advancing credit risk assessment methodologies. These efforts will contribute to ongoing enhancements in lending decisions, benefiting financial institutions, regulators, and borrowers alike.