NC STATE UNIVERSITY

North Carolina State University

Department of Financial Mathematics

FIM 590 004 Risk Management in Commercial Banks

Exposure at Default Modeling

Author:

Zhen He

December 2, 2024

Contents

FIM590-004 Final Project Exposure at Default Modeling

1. Introduction

1.1.Overview

This report explores Exposure at Default (EAD) modeling using a dataset of defaulted loans from Lending Club. It begins with data wrangling, including the selection of relevant loans, removal of high-cardinality and missing-value columns, and transformation of date features into usable formats. These steps ensure a clean dataset for analysis and modeling.

Exploratory analysis focuses on key metrics like the Credit Conversion Factor (CCF) and correlations among numerical features. Ridge Regression is employed for EAD prediction, using a robust preprocessing pipeline for numerical and categorical data. Model performance is evaluated with Mean Absolute Error (MAE), highlighting the best model parameters through validation and test results.

The report identifies significant predictors such as loan issuance years and time since the last payment, emphasizing their impact on EAD. Economic conditions and regulatory changes during periods like the 2007–2008 financial crisis are highlighted as key factors. Visualizations of feature importance provide actionable insights to strengthen risk management practices.

1.2.EAD and CCF

EAD represents the estimated potential loss a bank might incur if a borrower defaults on their obligation. This loss depends on the level of exposure the bank has to the borrower at the time of default, which occurs at an uncertain future date. EAD is calculated by combining the risk already utilized in the operation with a proportion of the undrawn credit risk. (CFI, 2024)

EAD for loans refers to the outstanding loan balance, including accrued interest, at the time of a borrower's default, representing the lender's total credit exposure.

$$EAD = Funded\ Amount \times Credit\ Conversion\ Factor$$

$$Credit\ Conversion\ Factor = \frac{(Funded\ Amount - Received\ Principal)}{Funded\ Amount}$$

This project will focus on the Credit Conversion Factor.

1.3.Key Findings

- The year of loan issuance and the time since the last payment are the most significant factors in predicting Exposure at Default (EAD).
- Loans issued during the financial crisis (2007–2010) show lower exposures due to tightened credit conditions, while loans issued during 2015–2018 have the highest exposures, peaking in 2018.
- Longer durations since the last payment are linked to higher exposures, as borrowers tend to maximize loan utilization closer to default.
- Significant changes in borrowers' FICO Scores are associated with increased exposure, highlighting the role of creditworthiness fluctuations.

2. Data Preparation

2.1.Data Source

The dataset is from LendClub, and downloaded on Kaggle. (George, 2018)

Lending Club is a peer to peer lending company based in the United States, in which investors provide funds for potential borrowers and investors earn a profit depending on the risk they take (the borrower's credit score). Lending Club provides the "bridge" between investors and borrowers.

The data set contains loans from 2007 to 2018, it contains 2260701 observations and 151 features.

2.2.Feature Encoding

After removing irrelevant and highly correlated columns, there are 15 categorical features and 97 numerical features.

Note that when dealing with issuance date and date since last payment, since the year of issuance does not reflect some time continuity or trend (e.g., the older the year, the looser or stricter the economic conditions, lending policies may be), this variable should be used as a categorical variable. In addition, for the date of the last disbursement, it would be more meaningful to calculate the year to date (2018).

In addition, employment lengths that are greater than 10 years are categorized as 10 years.

All the categorical features are being handled through OneHotEncoder to ensure meaningful output.

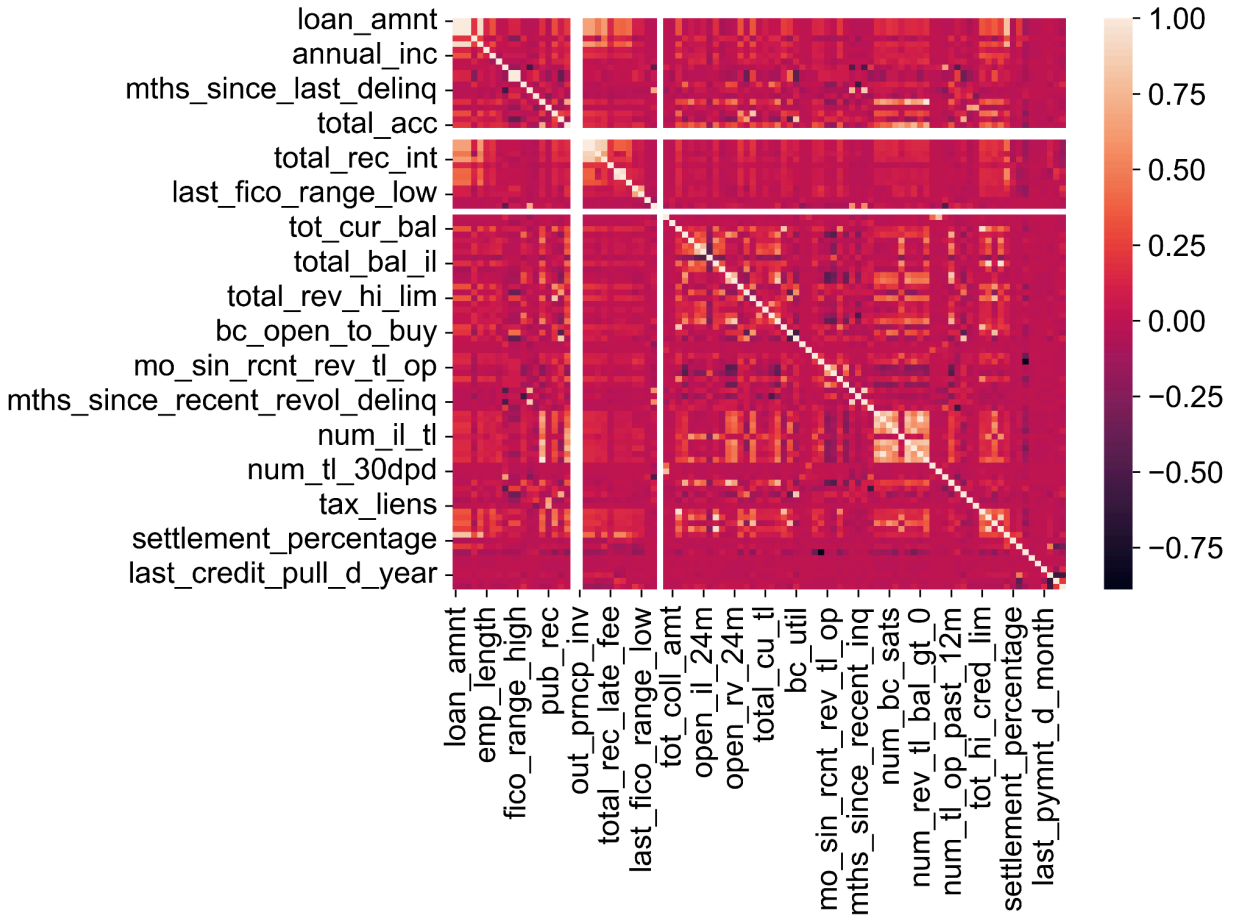
2.3.Handling Missing Values:

For numerical features, missing values are being filled with average. And for categorical features, missing values are being filled with values that appear the most frequent.

2.4.Multicollinearity

When creating a linear model that uses many features to make predictions, some of those features can be highly correlated with each other. This isn't a problem that's going to break the model; it will still make predictions and it might have good performance metrics. However, it is

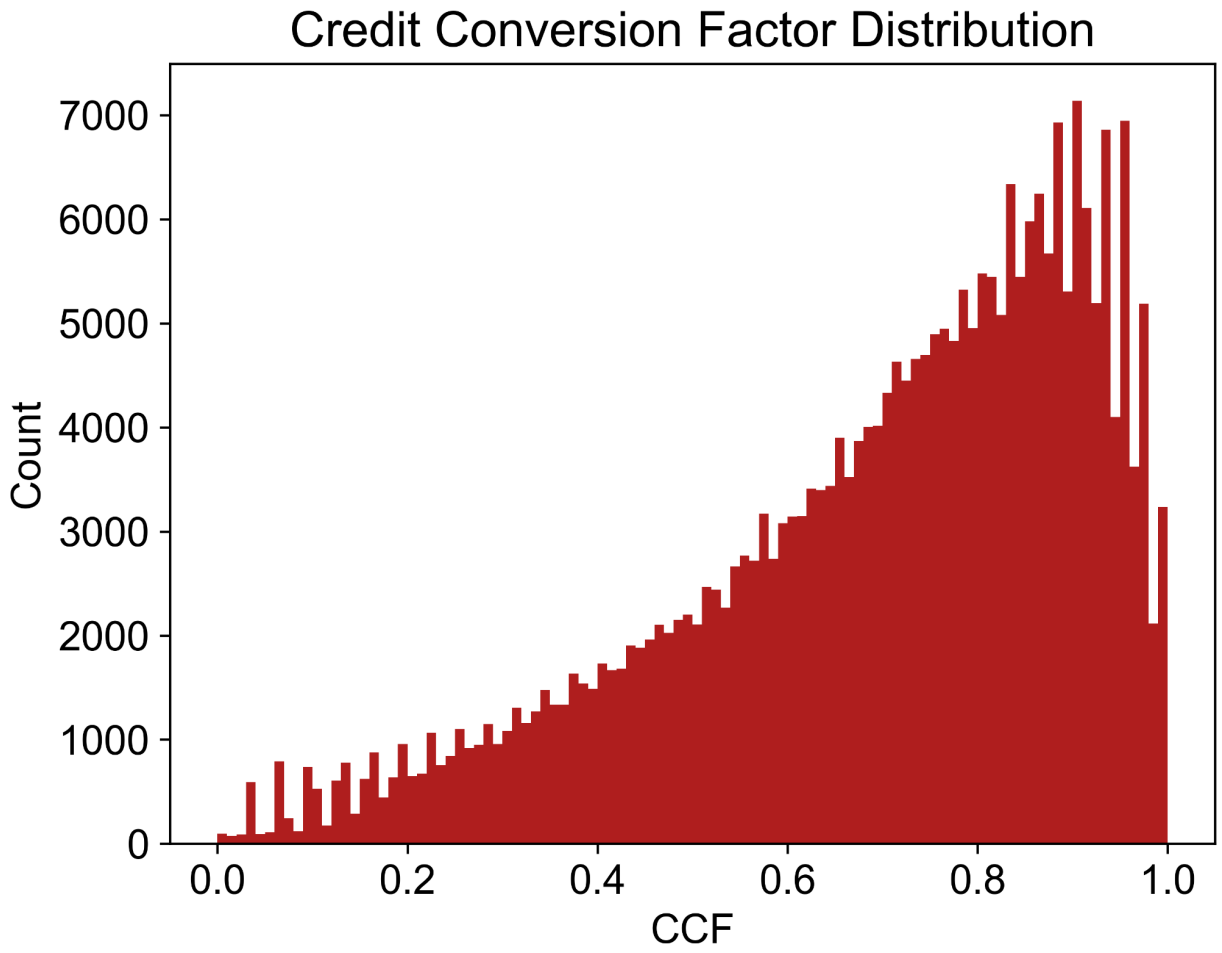
an issue on the interpretation of the coefficients for your model because it becomes hard to tell which features are truly important.



According to this graph, there exists a small number of highly correlated features.

2.5.CCF

The distribution shows a gradual increase in frequency as the CCF value approaches 1, peaking near the upper end of the range. This suggests that a significant portion of observations has higher CCF values, indicating a greater utilization of available credit limits at default.



CCF Stats:

Number of Loans	269320
Mean	0.697208
Standard Deviation	0.218809
Min	0.000000
25% Quantile	0.566590
50% Quantile	0.747175
75% Quantile	0.871866
Max	1.000000

2.5.Data Normalization

Data standardization is an essential step in our modeling process to ensure that each feature contributes equitably to the model's predictions. The dataset comprises features measured on different scales. To enhance the model's efficiency and accurately identify the most significant features based on their regression coefficients, it is necessary to normalize all features to a consistent scale. In this study, it achieved standardization by applying the z-score normalization method:

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}$$

Where Z_i is the z-score vector for the i-th feature, X_i is the i-th feature vector, μ_i is the mean of the i-th feature, and σ_i is the standard deviation of the i-th feature. This method is applied through StandardScaler.

3. Model Explanation

There are 269320 observations and 112 features, to train the model, 60% of the data is in the training set, 20% in the validation set and 20% in the test set.

3.1.Model

Ridge Regression:

Ridge regression is a form of regularization method used in machine learning. It adds a regularization term to the loss function by adding a multiplication of a parameter λ and the sum of the squares of the model coefficients, which penalizes large coefficient values. This constraint prevents the model from fitting the training data too perfectly, improving its generalization to new, unseen data. (Wikipedia, 2024) The formula is given by:

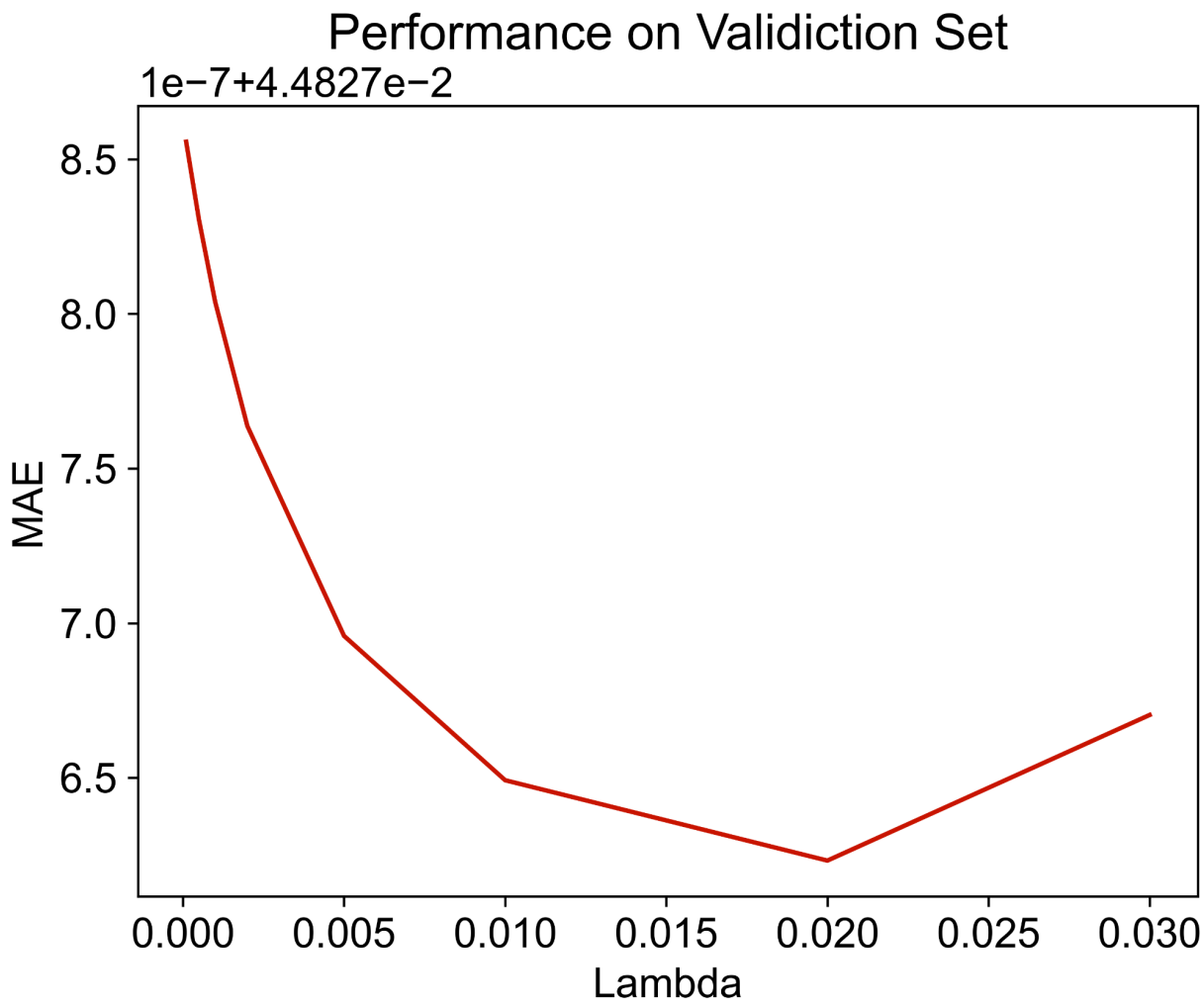
$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - (y^{(i)}))^2 + \lambda\sum_{j=1}^{n}\theta_j^2$$

3.2.Tuning Hyper-Parameters

To tune hyper-parameters, this project employs Mean Absolute Value to calculate the performance on the validation set.

MAE reflects the size of the discrepancies between real values and gives insight into how effectively the model forecasts the target variable.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y^{(i)} - y_{predict}^{(i)}\right|$$



This figure shows that when lambda is 0.02, the model performs best.

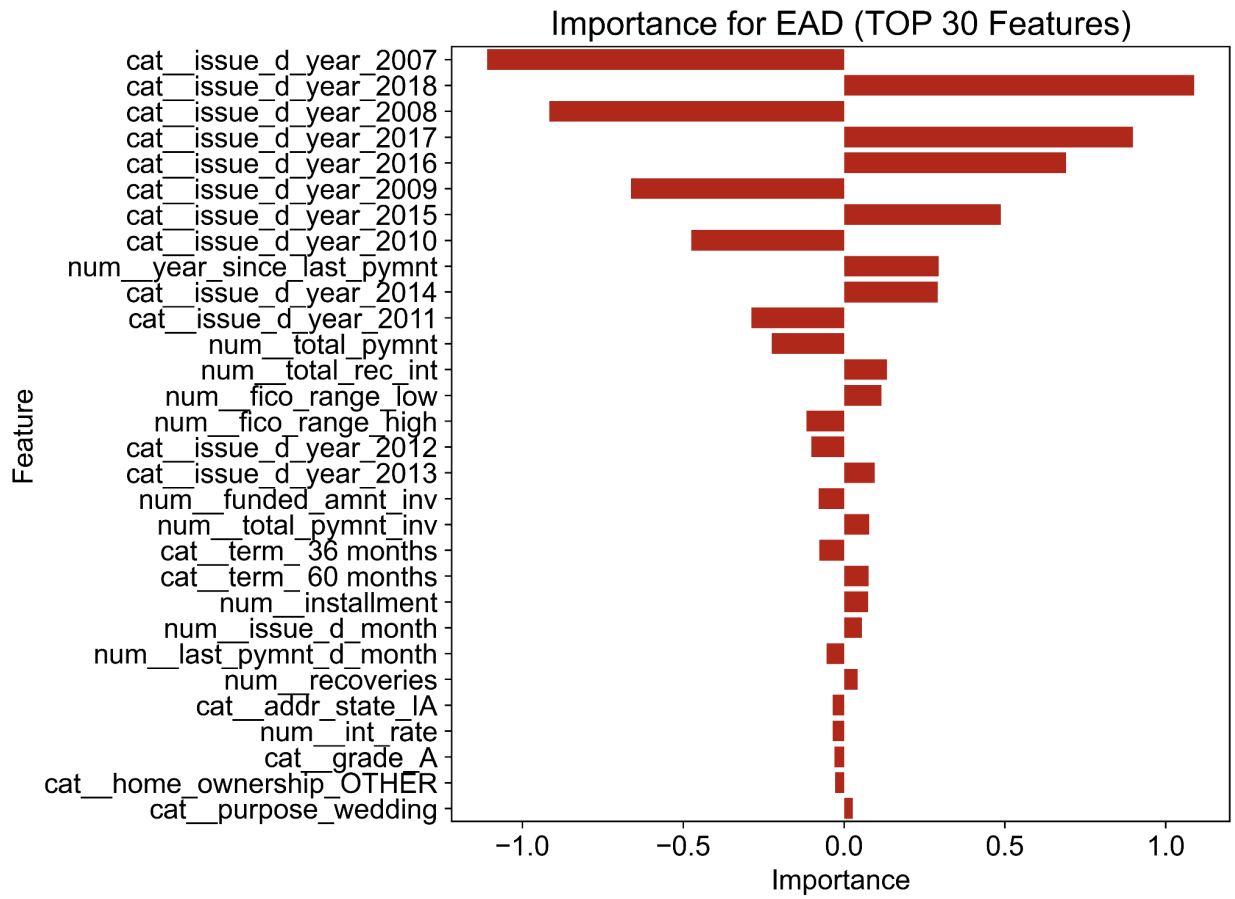
3.5.Performance Summary

Mean absolute error on the test set is 0.04463607783075181, which represents a reanable performance.

4. Report Finding

4.1.Overview

Issue Date and Year Since Last Payment play tremendous roles in predicting EAD.



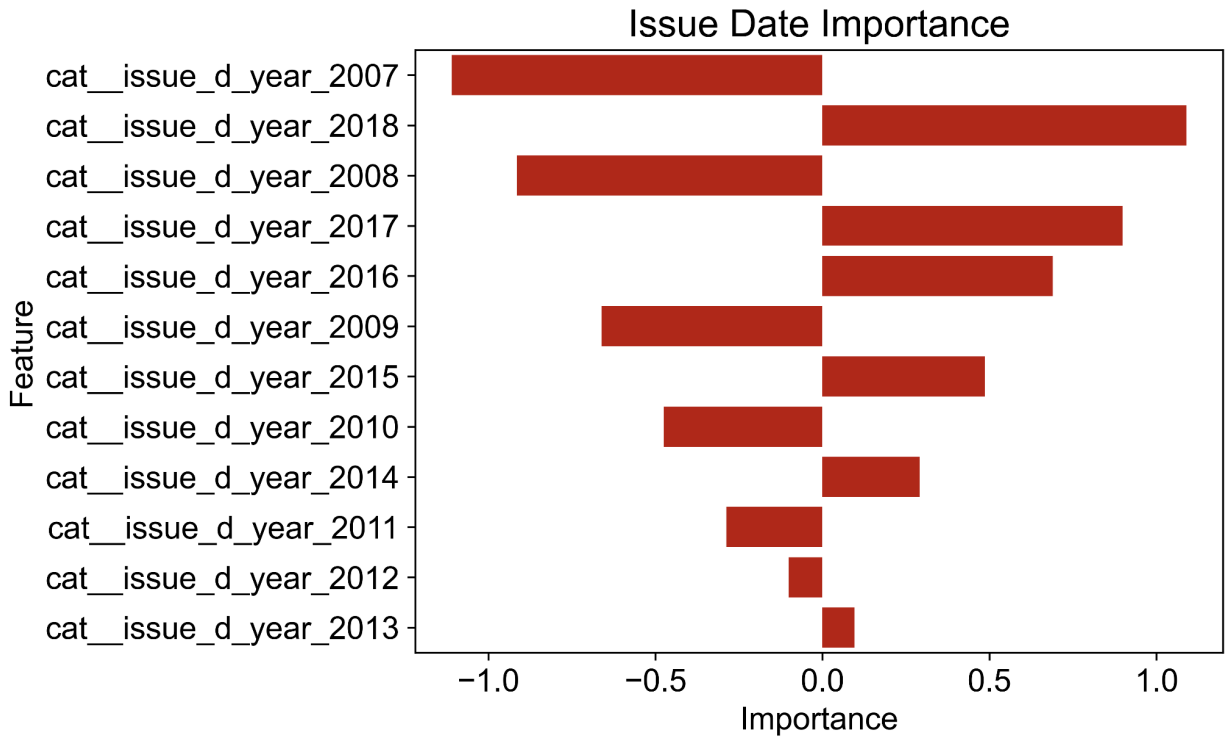
The bar chart illustrates the top 30 most significant features contributing to the Exposure at Default (EAD) model, ranked by their importance. Features related to the loan issuance year, such as "cat__issue_d_year_2007" and "cat__issue_d_year_2018," dominate the ranking, reflecting their critical role in predicting EAD. The feature "num__year_since_last_pymnt" also exhibits high importance, emphasizing the influence of borrower payment history.

Categorical variables, such as loan terms ("cat__term_36 months"), and numerical factors, including payment amounts ("num__total_pymnt" and "num__total_rec_int"), further contribute to the model's predictions. The results underscore the interplay between historical financial trends and loan-specific attributes in determining exposure levels. Lesser influential features,

such as loan purpose or homeownership status, have smaller but non-negligible impacts. This highlights the multidimensional nature of EAD prediction.

Since Issue Date is the most important, the next section analyzes issue date.

4.2.Issue Date Analysis



It is worth noting that loans issued during and after the financial crisis (2007, 2008, 2009, 2010) have smaller exposures, which is in line with financial intuition, as credit funding was tightened and borrowers were scrutinized more stringently after this period. Such a positive contribution (negative coefficient) to the risk exposure decreases year by year.

In 2015, 2016, 2017, and 2018, presumably because of more aggressive quantitative easing by central banks and deregulation, loans issued during this period had the largest exposure and peaked in 2018, which is the final cutoff time for the dataset.

4.3.Other Feature Analysis

The time since the last payment also has an impact on exposure, with the larger the value, the greater the exposure. This leads to the conclusion that when close to an event of default, borrowers will utilize the loan as much as possible, which increases the risk exposure.

In addition, the change in FICO Score also has an effect, when the change in FICO Score of the borrower is large, it increases the risk exposure.

5. Summary

This paper investigates the modeling approach for Exposure at Default (EAD), utilizing the defaulted loan dataset provided by Lending Club for the analysis. The first step of the study is a thorough preparation of the data, including screening for relevant features, handling missing values, and standardizing the data via z-score to ensure fairness and consistency of model predictions across features. Credit Conversion Factor (CCF) is the core metric of the analysis, defined as the ratio of used credit to total available credit, which reflects the borrower's behavioral pattern at default.

To predict EAD, this paper employs a ridge regression model and determines the best parameter configuration through hyperparameter tuning to optimize the model's performance on the validation set. The results show that the year of loan disbursement and the last repayment interval are the key variables in predicting EAD. Loans disbursed during the financial crisis (2007-2010) have low risk exposure due to tighter credit policies and stricter borrower scrutiny, while the risk exposure of loans disbursed during 2015-2018 rises significantly, peaking in 2018, consistent with more accommodative economic policies and de-regulation trends during that period.

In addition, the analysis found that the longer the time to the last repayment, the greater the exposure, suggesting that borrowers tend to maximize the use of their credit lines when they are close to default. In addition, significant changes in borrowers' FICO credit scores are an important factor in risk exposure, with larger score fluctuations tending to be accompanied by higher risk. These results reveal the combined impact of borrower behavior and external economic conditions on risk exposure.

The findings fully reflect the multidimensional nature of EAD prediction, which is not only influenced by macroeconomic trends but also driven by individual borrower characteristics. The findings of this paper provide valuable practical recommendations for financial institutions to strengthen their risk management frameworks, optimize their credit policies, and more accurately predict potential default losses, thereby improving overall financial soundness and market resilience.

6. Reference

CFI. (2024). *Exposure at Default (EAD)*. https://corporatefinanceinstitute.com/resources/commercial-lending/exposure-at-default-ead/

George, N. (2018). *All Lending Club loan data: 2007 through current Lending Club accepted and rejected loan data.* Kaggle. https://www.kaggle.com/datasets/wordsforthewise/lending-club

Wikipedia. (2024). *Ridge regression.* https://en.wikipedia.org/wiki/Ridge_regression