Machine Learning and Statistical Methods in Credit Risk Management

Jorge Reyes

University of Tennessee, Knoxville

COSC 522 Machine Learning

December 3, 2024

Contents

1 Introduction

Abstract

The goal of this project is to showcase machine learning and statistical methods in credit risk applications. For this project, we define credit risk as the possibility that a lender will incur a financial loss due to the borrower defaulting or not meeting its obligations. This paper will outline how to implement credit risk ML and statistical models, why those models were chosen, how to measure performance, and finally, how to apply our results. While we do explain the motivation and background of the topics covered in this paper, this paper will not be covering the mathematical theory behind each topic. The first part of any data science project consist of collecting data and then conducting pre-processing operations. This project is no different. In fact, more resources were spent understanding and cleaning the dataset than any other step in the project. This dataset, although the exact source is unknown, comes from a kaggle repository and is rated as 'gold' by kaggle admins due to its usability. The dataset is split into two tables, one contains fields outlining the life of the loan and the other houses all customer characteristics. We will be referring to the aforementioned tables as 'credit' and 'app' respectively throughout this paper. One last note on the dataset, there is not a clear definition of if/when a default occurred. Knowing this, we chose to identify 'bad' borrowers as those having an overdue balance greater than 60 days. This gives leniency to the borrower and allows us to include more records for our target class than having, for example, a cutoff of overdue balance greater than 90 days. Next, we will be creating two classification models, a logistic regression and a random forest classifier, that will predict whether customers are 'good' or 'bad' borrowers. Before creating our time-to-default (in this case, time-to-bad?) model, we will be running a special type of EDA called vintage analysis. Our goal for the vintage analysis is to show the cumulative distribution of 'bad' borrowers in our loan portfolio (ie. dataset) by their origination date. For our time-to-default models, we will be aiming at predicting the probability that a borrower, given their characteristics, will remain 'good' as the loan approaches 5 years on the books. As with any ML or statistical model, we will be implementing various performance measures and scores. In our final section, we will be applying our findings and commenting on the results as it relates to managing the risk of a simple loan portfolio. To conclude the research, we will dedicate a section on limitations, final conclusions, and future work.

2 Setting the Stage

2.1 Introduction to the Data

As stated in the introduction, the dataset is split into two parts. Describing the ’credit’ table is an appropriate starting point because it required a great deal of examination before any cleaning was done. The ’credit’ table contains a primary key labeled ID that will act as a customer ID and can be used to join to the ’app’ table. Table 1 shows the last two fields of the ’credit’ table, leaving out the month and ID fields, that allows us to define a ’bad’ borrower. For example, if we were to single out a borrower and for every month (month stated as -3, -2, etc., where 0 is the maturity month in the table) we track the status, we can see that if status code is either 2, 3, 4, or 5, we consider that borrower as ’bad’. What we have just explained is that we are making the assumption that a borrower is not allowed repayments after more than 60 days. The reason for this is to allow for more straightforward modeling and because we are essentially treating a ’bad’ borrower event as a default.

Loan Status Code	**Loan Status Description**
0	1-29 days past due
1	30-59 days past due
2	60-89 days overdue
3	90-119 days overdue
4	120-149 days overdue
5	Bad debts, write-offs - more than 150 days
C	paid off that month
X	No loan for the month

Table 1: Sample of credit table

We will discuss the second part of the dataset, table named ’app’, further in the feature engineering section. Feel free to check out the data source using the link in the bibliography linked here [1].

2.2 Pre-Processing Highlights

If you are interested in following all of the pre-processing actions, annotations were made to the code throughout the process, but we will not cover everything in this paper as that is a daunting and time-consuming task. Much of the pre-processing actions were done to manipulate the ’credit’ table to get ’bad’ borrower indicator, origination month, loan term, and month-on-books fields. Because we created a month-on-books field (mainly for the Survival Analysis model), we are in a sense treating the entire dataset as a loan portfolio which was personally helpful given my background in finance. We can define the month-on-books field as the month the borrower turned ’bad’. As you may have guessed, this field was not used as a feature in the classification models.

2.3 Feature engineering and EDA

3 Dealing with Class Imbalance

Describe the methods used in your research. Include details about data collection, experimental design, tools, or techniques employed.

4 Classification Models

Present the findings of your research. Use tables, figures, and charts to illustrate your results if necessary.

5 Time-To-Default: Survival Analysis

Interpret your results and discuss their implications. Compare your findings with those of previous studies.

6 Applying Results

Summarize the key points of your research. State the main conclusions and suggest potential areas for future research.

7 Conclusion

Summarize the key points of your research. State the main conclusions and suggest potential areas for future research.

References

[1] rikdifos. *Loan Approval Prediction.* `https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction?select=application_record.csv`. Accessed: 2025-01-15. 2023.