UNIVERSITY OF BIRMINGHAM

School of Computer Science
College of Engineering and Physical Sciences

MSc. Project

Machine Learning & Deep Learning Approaches to Predict Credit Card Default

Submitted in conformity with the requirements
for the degree of MSc. Artificial Intelligence & Computer Science
School of Computer Science
University of Birmingham

Sarathkumar Padinjare Marath Sankaranarayanan
Student ID: 2359859
Supervisor: Dr.Kashif Rajpoot

September 2022

Abstract

The material contained within this report has not previously been submitted for a degree at the University of Birmingham or any other university. The research reported within this report has been conducted by the author unless indicated otherwise.

Keywords Credit Card Default Prediction, Ensemble Learning

Sarathkumar Padinjare Marath Sankaranarayanan

Declaration

The material contained within this report has not previously been submitted for a degree at the University of Birmingham or any other university. The research reported within this report has been conducted by the author unless indicated otherwise.

Signed Sarathkumar Padinjare Marath Sankaranarayanan

Sarathkumar Padinjare Marath Sankaranarayanan

"You have to learn the rules of the game.
And then you have to play better than anyone else"

Albert Einstein

MSc. Project
Machine Learning & Deep Learning Approaches to Predict Credit Card Default

Sarathkumar Padinjare Marath Sankaranarayanan

Contents

Table of Abbreviations

SVM	Support Vector Machine
ANN	Artificial Neural Network
GBDT	Gradient Boosting Decision Tree
GRU	Gated Recurrent Unit
LGBM	Light Gradient Boosting Machine
XGBoost	Xtreme Gradient Boosting Machine
GRU	Gated Recurrent Unit
CV	Cross Validation

List of Figures

1 Introduction

This section will introduce the user to definitions of terms relevant for understanding the problem, discuss the motivation behind the problem, the aim & approach taken to solve the problem, and the structure of this report.

1.1 Definitions

1.1.1 Credit Card Statement Date

The credit card statement date is the date on which the statement/bill is generated every month. Any transaction conducted on the card post billing date will reflect in the next month's credit card statement.

1.1.2 Delinquent Account

A credit card account is considered delinquent if the customer has failed to make the minimum monthly payment for 30 days from the original due date.

1.1.3 Delinquency Rate

The percentage of credit card accounts within a financial institution's portfolio whose payments are delinquent.

$$DelinquencyRate = \left(\frac{NumberOfDelinquentCreditCardAccounts}{TotalNumberOfCreditCardAccount} \right) * 100 \tag{1}$$

1.1.4 Credit Card Default

The customer is considered as defaulting customer in the event of nonpayment of the due amount in 120 days after the latest statement date.

1.2 Motivation

Delinquency rates & credit card default rates are directly proportional. According to the figure 1, the delinquency rates were at an all-time high just before the recession started in 2008; moreover, this was the same time when more & more customers began to default on credit card payments.

Predicting credit defaults is essential for managing risk in the consumer lending industry. Credit default prediction enables lenders to make the best possible lending decisions, improving customer satisfaction and fostering strong company economics.

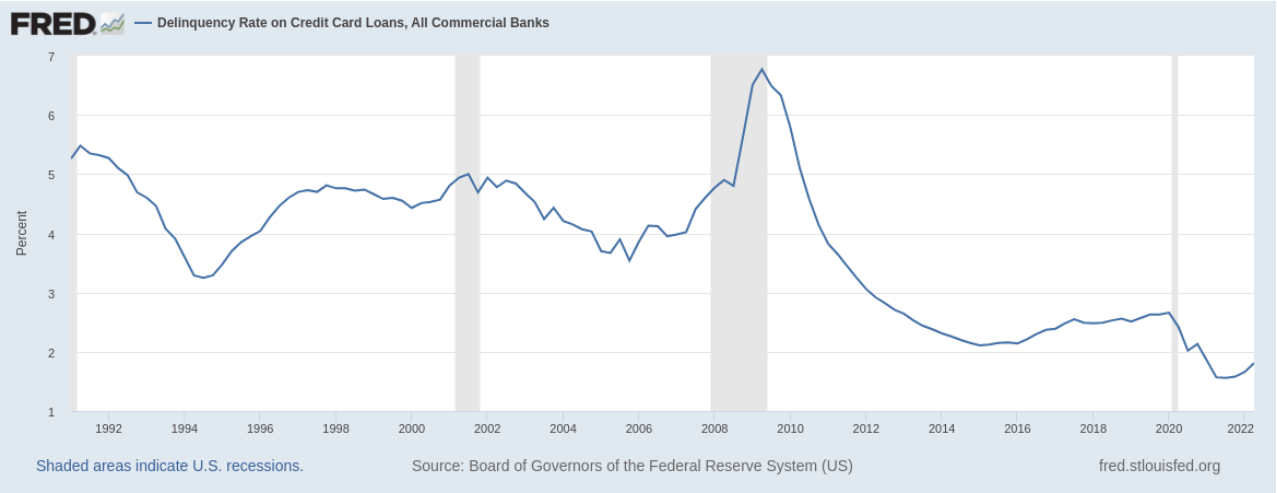

Figure 1: Delinquency rate on credit card loans for the period 1992-2022(Board of Governors of the Federal Reserve System (US) 2022).

Existing models can be used to manage risk. However, developing models that perform better than those in use is feasible.

1.3 Aim & Approach

The objective of this project was to explore different machine learning algorithms & deep learning architectures on the American Express default prediction dataset(American Express 2022) to predict if a customer will default on the payment in the future. The project work started by developing a model using classic machine learning algorithm Support Vector Machine (SVM) followed by creating multiple models using Random Forest Classifier & Gradient Boosting Decision Tree (GBDT) algorithms. Artificial Neural Network (ANN), Gated Recurrent Unit (GRU) & Custom ensemble model created by model combining ANN, GRU and GBDT were created as part of exploring deep learning architectures. Finally created a lean model, using less features & optimzed parameters using GBDT which provided comparable performances to the previously explored models.

1.4 Structure of Report

The remainder of the report is structured as follows: in Section 2 background information on different machine learning & deep learning algorithms along with metrics explanation is provided. Then in Section 3 a literature review related to the credit card default prediction research is given. Section 4 & 5 provides detailed explanations on the dataset, tools & software used in the project, and methodology followed for creating the models. Model evaluation results and the comparison is given in Section 6. Finally Section 7 discusses the conclusion of the project.

2 Background Knowledge

This section provides the reader with the required background information on the machine learning algorithms, deep learning architectures & metrics.

3 Literature Review

asfsf

3.1 Summary

dfsadfsd

4 Materials

4.1 Primary Dataset

The primary dataset contains 190 aggregated profile features of 458913 American Express customers at each statement date for 13 months. Features are anonymized and normalized, and fall into the following general categories:

- D_* = Delinquency variables
- S_* = Spend variables
- P_* = Payment variables
- B_* = Balance variables
- R_* = Risk variables

This dataset(American Express 2022) was released as part of the "American Express - Default Prediction" hosted in Kaggle by the American Express team.

4.2 Secondary Dataset

The secondary dataset was derived from primary dataset by applying the below mathematical aggregate operations to the numerical features.

- Minimum
- Maximum
- Mean
- Last Value
- Standard Deviation

Aggregate for the categorical features were taken by the applying below operations.

- Last Value
- Count
- Unique Value Count

The secondary dataset contains 920 features and 458913 records.

4.3 Tools & Software

The primary programming language used for the implementation of this project is Python version 3.7. Data analysis and manipulation is done using Pandas(1.3.5), seaborn(0.11.2) & Dask(2.12.0) packages. Scikit Learn(1.0.2) package is used for create, train & evaluate machine learning models. ANN & GRU models were created using Tensorflow (2.8.2).Google colab was used to train the model in cloud and Github was used as the version control & project management software.

5 Methodology

asfsf

5.1 Summary

dfsadfsd

6 Results & Discussions

asfsf

6.1 Summary

dfsadfsd

7 Conclusion & Summary

asfsf

7.1 Summary

dfsadfsd

References

American Express (2022), 'American Express Default Precition Dataset', Available: `https://www.kaggle.com/competitions/amex-default-prediction/data`.

Board of Governors of the Federal Reserve System (US) (2022), 'Delinquency Rate on Credit Card Loans, All Commercial Banks [DRCCLACBS]', Available: `https://fred.stlouisfed.org/series/DRCCLACBS`. Data retrieved from FRED, Federal Reserve Bank of St. Louis;.

8 Appendix One: Code

8.1 Directory Structure

8.2 Running the Provided Code