Credit Cards Payment Default Prediction

Capstone Project by Violeta Furculita

April 2024

Context and Problem Overview

Context:
- Credit cards are lines of credit extended to customers to use for everyday purchases
- Most customers make monthly payments toward the principal and interest charged, where applicable. Some customers may forgo making payments due to financial distress or intent to commit fraud
- Credit cards are a major revenue stream for financial institutions
- Even small increases in default rates can lead to significant losses
- Being able to accurately predict which customers may default allows targeted risk management

Problem:
- Conventional methods of predicting likelihood of default at underwriting or during usage may not be well equipped to predict default
- Identifying at-risk customers early allows contact and/or appropriate loss management
- However, default risk depends on complex interactions of many customer attributes

Understanding the problems

01   Default costs are high for both the bank and the customer. The bank may experience losses, lost revenue, collection costs. The customer may experience credit rating negative impact

02   The conventional approach such as FICO is not well equipped to predict default. Lost of other factors such as demographic, employment, recent behaviors, credit limits, balance levels etc factored together may be able to help.

03   Machine learning is well-suited to solve the problem of default prediction because it has the ability to look at vast troves of data and make predictions that can help at underwriting and during ongoing credit management.

Project objective

Analyze public credit card default data and build machine learning models that can help predict payment default rates on credit cards with reasonably high accuracy of at least 75-80% accuracy, which would be a significant improvement over existing methods.

Target audience

Financial institutions that offer credit cards either directly or through an intermediary to consumers.

Project Methodology

1. Clean, explore and visualize the credit card default dataset to understand trends, patterns and relationships with default behaviors.
2. Engineer relevant features from the raw data for model building.
3. Evaluate various classification algorithms like Logistic Regression, Decision Trees, Neural Networks , K Nearest Neighbors and Gradient Boosting ensembles like XG Boost to identify models that perform best on this problem.
4. Tune model hyperparameters to optimize for key metrics like accuracy and train time.
5. Analyze feature importance to understand driver attributes of default risk.
6. Deploy the best performing model(s) and evaluate on test data to assess predictive power.
7. Present model performance, key findings and recommendations to stakeholders on how predictive analytics can help mitigate future credit card default rates.
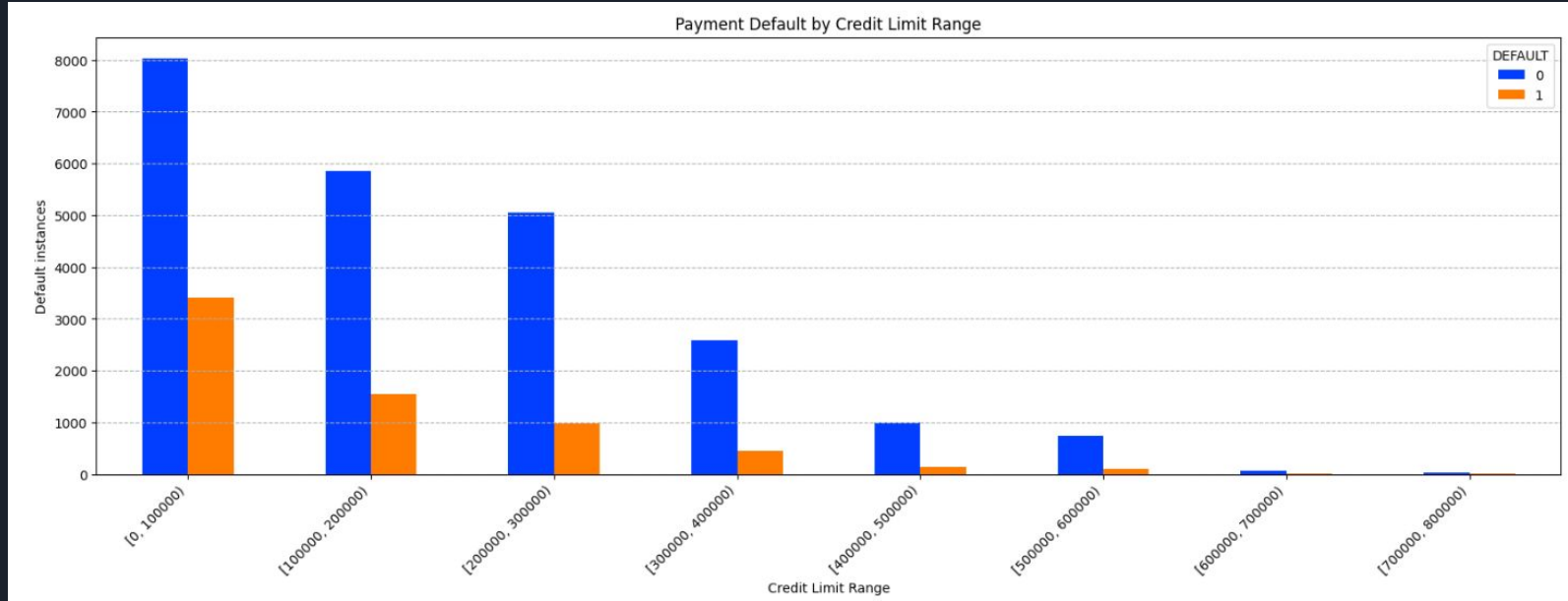
Overview of initial credit card user data

22% of customers defaulted on their credit card payment
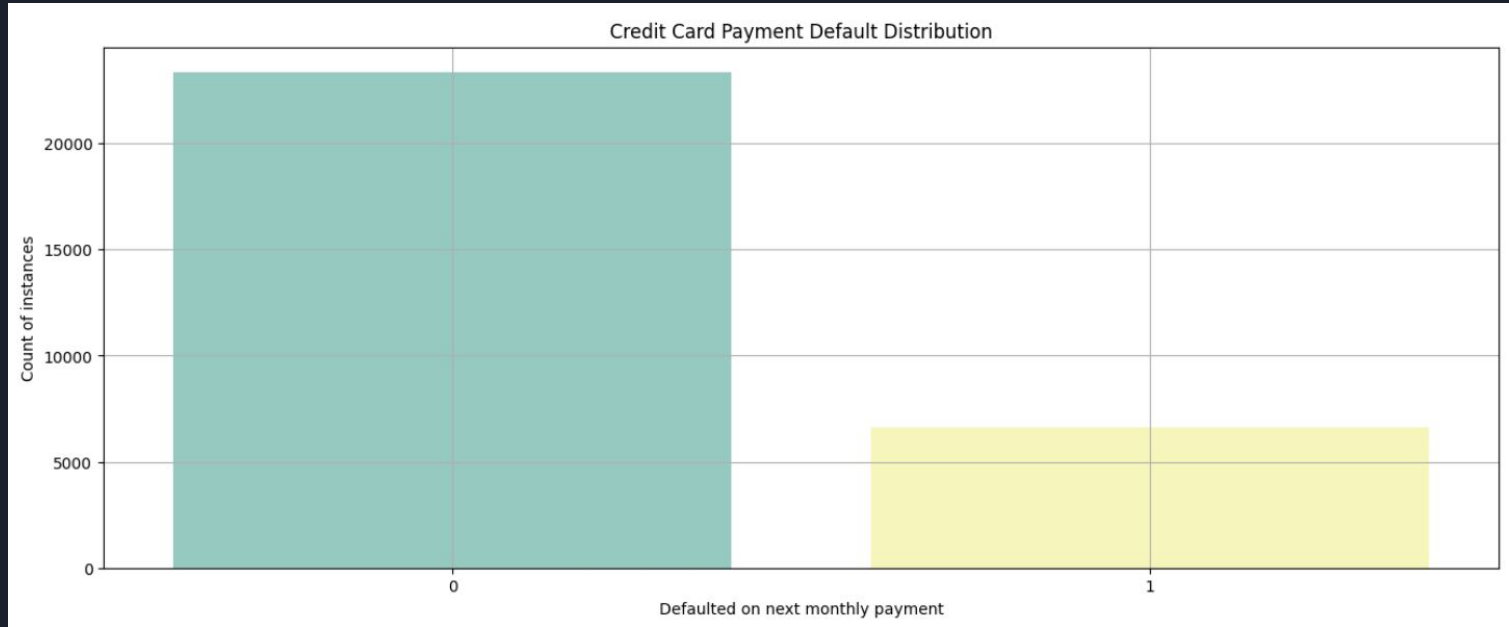
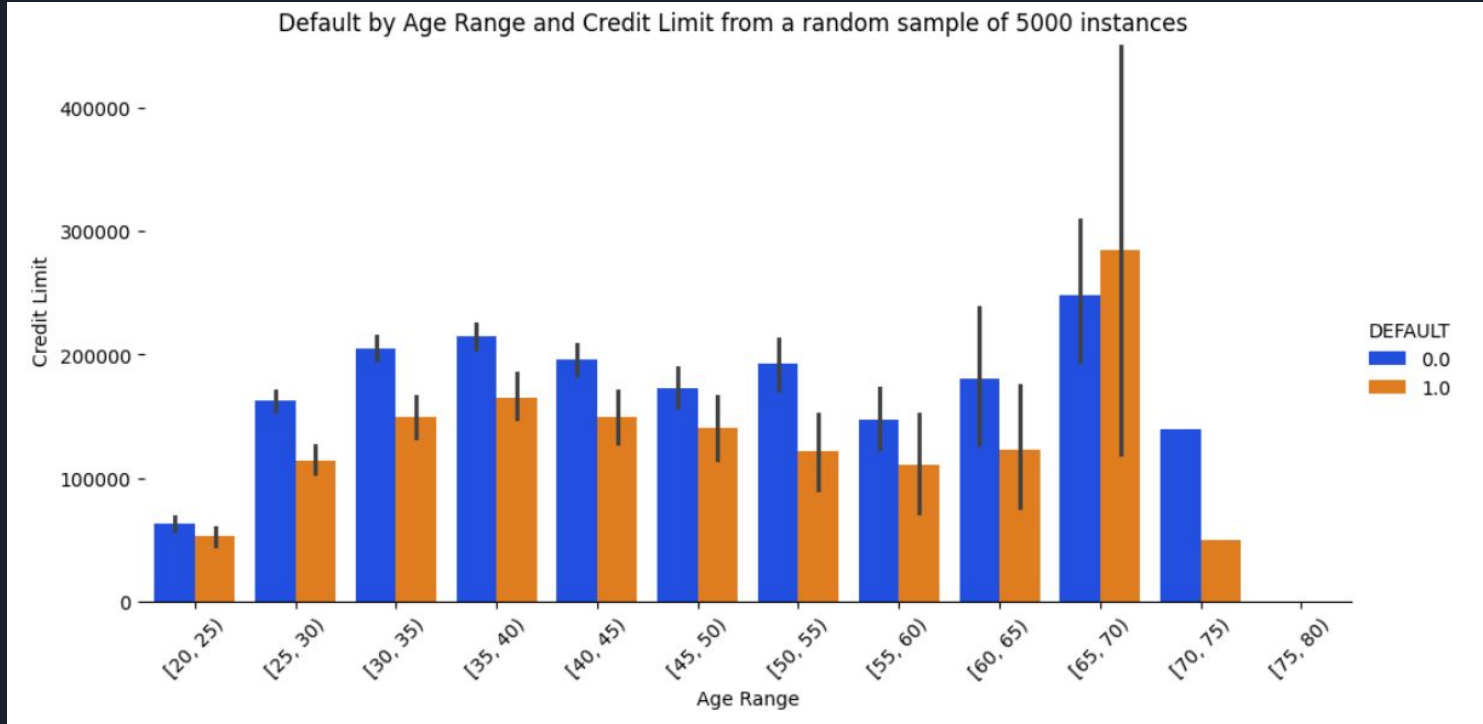# Overview of initial credit card user data

Users with credit limits under 200k default the most

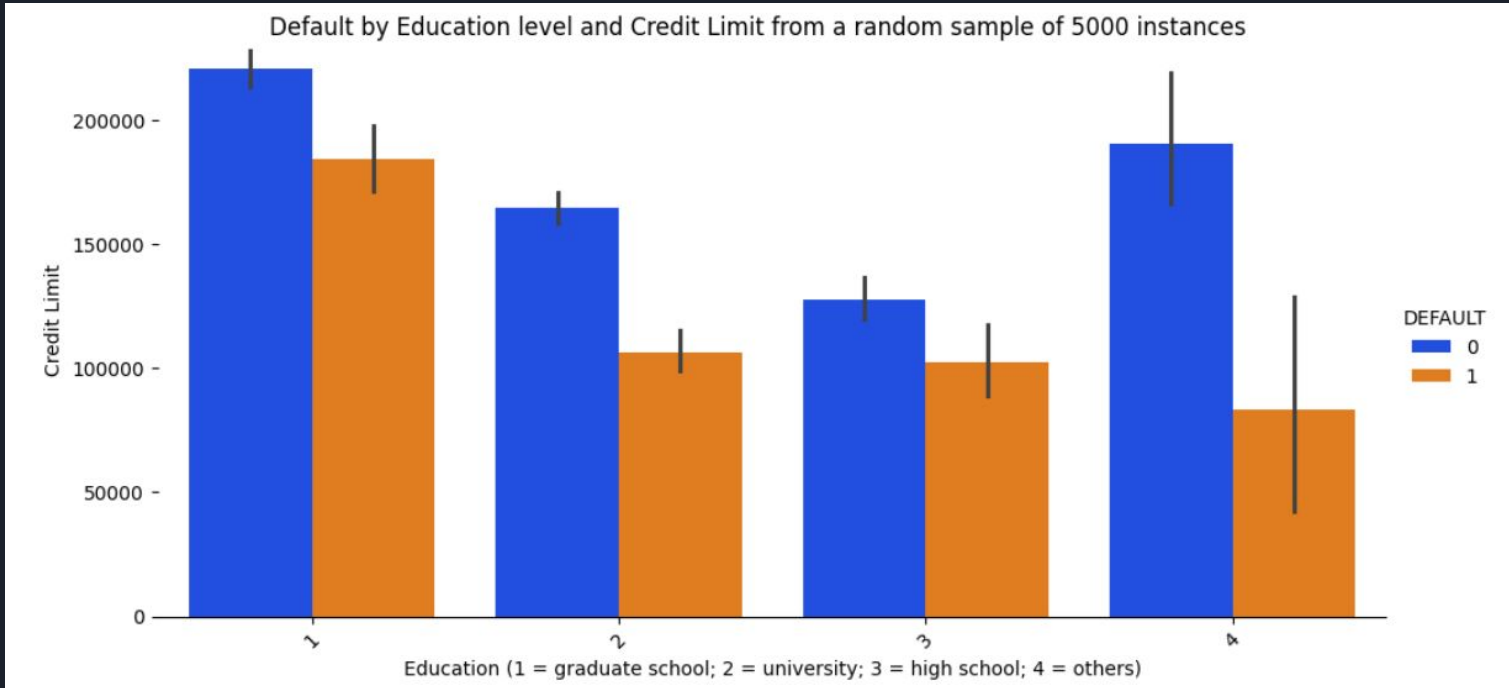
Payment Default by Credit Limit Range

Overview of initial credit card user data

Users with ages between 65-70 and credit limit above 200k have a high default rate

Default by Age Range and Credit Limit from a random sample of 5000 instances

Overview of initial credit card user data

Users with graduate degrees and credit limits above 100k tend to default more than users from other education groups

Default by Education level and Credit Limit from a random sample of 5000 instances

Preliminary Findings From User Data

- Payment default rate is high at over 20% of the user base

- Users with credit limits under 100k have the highest default rate followed by those with credit limits under 200k

- For user with ages higher than 65 and credit limits higher than 200k the default rate is higher as compared to sage group age and lower credit limits

- Most defaults happen in the population aged 25-55

- Users with credit limits higher than 100k and graduate degrees tend to default more than users with the same credit limit and different degree types

Preliminary Findings From Building Models

Using supervised learning models we can achieve robust accuracy and training times for our models that predict payment default. Specifically Logistic Regression and Neural Networks can produce prediction accuracy of 78% with a fast training time between 0.01 and 0.08 seconds.

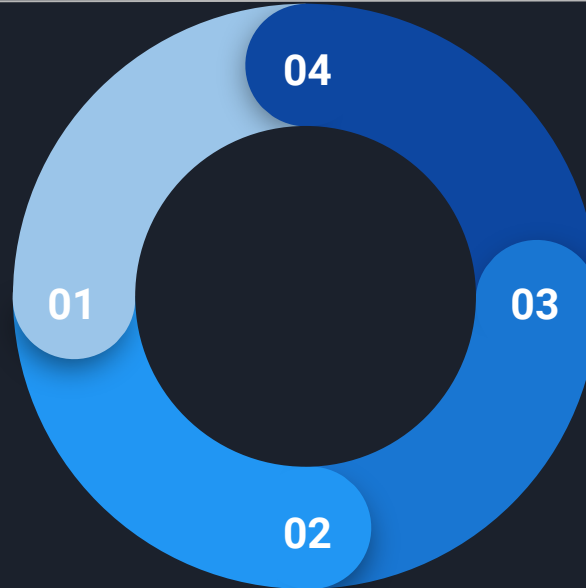|	Train Time	Train Accuracy	Test Accuracy
Logistic Regression	0.0848323	0.776881	0.776714
Decision Tree	0.388209	1	0.724238
Support Vector Machines	9.36027	0.817417	0.812571
Neural Networks	0.0135406	0.835702	0.787762
K Nearest Neighbor	20.6464	0.835667	0.812286

Recommendations and next steps

Fine tune models

Further model fine tuning is required to achieve higher accuracy rate on predictions

Use Ensembles

We will further explore model ensembles to help us achieve higher predictive rates

04

01

03

02

Deploy

Select a model(s) that deliver the best accuracy and train time to deploy and be used by the credit card company

Get feedback

As more user data gets collected, ingest it in the models deployed to check whether predictive accuracy increases