Credit Card Default Risk Analysis

By: Denzel Chike

Date: 05/08/2025

University of North Texas

Background / Problem statement

- My project focuses on credit card users who are at risk of default. A default means someone who has failed to make the minimum required monthly payment for the credit card. The consequences that come with missing a payment are late fees, interest rises, might get sent to collections, legal action, and more. This is a major concern for banking organizations and customers who bank with them. I chose to analyze the UCI Credit Card Default dataset from Kaggle to identify patterns and predictors for default behavior by using machine learning predictive model techniques. I used feature engineering methods to get new variables, like credit utilization and payment history, that is categorized. This helps by improving the prediction strength and providing a better understanding of the model. I used both Logistic Regression and Random Forest models that were trained and used metrics like accuracy, precision, recall, and a confusion matrix. The results showed that customer financial behavior, including using their credit cards a lot and paying bills late, can determine whether or not they'll miss their upcoming payment or default on their credit. This analysis project has good insights for banks to see who might not pay their loans, which helps them make smarter decisions on who should get approved for credit. Keywords: credit risk, logistic regression, random forest, default, prediction, payment history, and credit utilization.

Dataset Overview

- The data I am using for this project is from Kaggle. This data is called the UCI Credit Card Default dataset. The UCI Credit Card Default data contains 30,000 samples and 23 attributes. The metadata consists of demographic information like age, gender, marital status, and education. The other attributes are financial limits, which are credit limit, bill amount, previous payments, and repayment history of six months. The purpose is to have a binary label that shows if the client missed their payment, also known as default, that was due the next month. I did preprocessing that handled missing values and also removed invalid rows. I added a new feature called credit utilization, which finds the ratio of the bill amount to the credit limit. I used bins to categorize the function into low, medium, high, and very high usage. PAY_0 was also categorized into "Current or Early" and "Delayed" to see who paid early and who hit default.

Methodology

- There were many studies that explored and used machine learning models like random forest, logistic regression, and neural networks for credit scoring and predicting default. Logistic Regression is a statistical method that is used to predict the probability of a binary outcome based on one or more variables. In our case, we will be using this to predict who will hit default. Logistic Regression has always been a popular method to investigate things like credit and fraud. One existing methodology that I found is an article called "Prediction of Default Probability of Credit-Cards Bills" by Yuhan Ma. They created a credit scoring model to predict credit card default by using a dataset of 30,000 clients. They used the XGBoost machine learning algorithm, which stands for Extreme Gradient Boosting, that creates and continuously generates simple decision trees to find the most impactful factors affecting default risk. They cleaned the data and removed outliers and trained the model, and achieved an AUC of 0.779. What makes my project stand out is by use of feature engineering methods like credit utilization and transforming repayment status into categorical variables. I also analyzed demographics and trends for specific genders, which I believe are overlooked in quantitative models.

Five Contributions

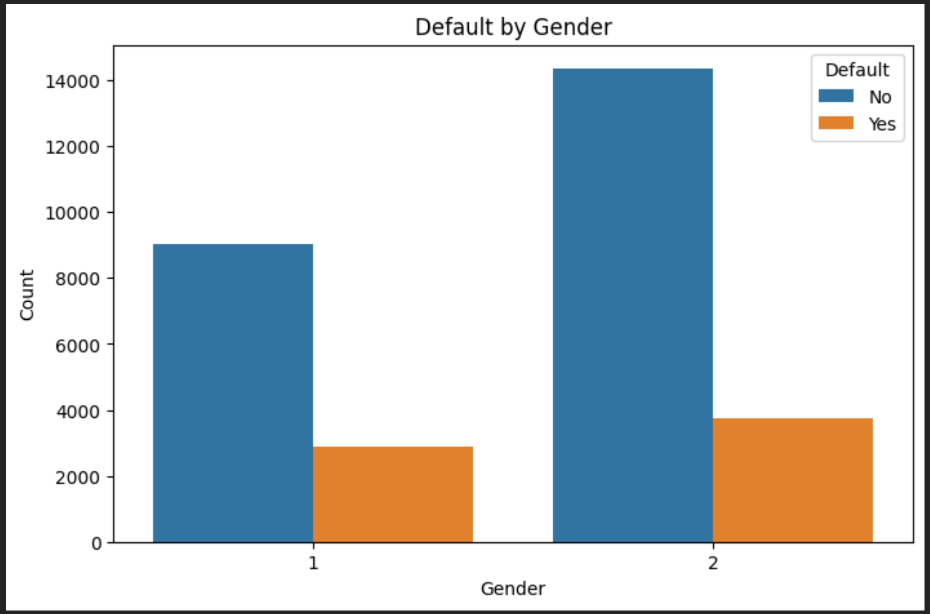
- **Feature Engineering**
- This first contribution was finding a way to tell a story which was by transforming raw columns into features rather than using the dataset as it came. I created a ratio called the credit utilization ratio that showed how much credit was given to each customer and categorized them by low, medium, high, and very high. By combining this feature with repayment behavior helps show which users were at default risk. Last but not least, payment history was grouped into labels to show which users were paying their bills and falling behind.
- **Data Visualization of Trends**
- I used Seaborn and Matplotlib for the visualization of different variables to detect default patterns. With the help of credit utilization groups, I was able to find that users using more than their credit limit were at high risk of default. By including other variables like gender, I was able to find that male users had a higher chance of hitting the default than female users. Providing this visualization will help explain the data to stakeholders.
- **Model Comparison: Logistic Regression vs Random Forest**
- Instead of sticking with one predictive model, I used two different predictive models, which are Logistic Regression and Random Forest. It helped me understand the balance between interpretability and predictive power. My accuracy rate for Logistic Regression was 81%, but it failed to identify users who could be at risk of default, cause the recall for the positive class was low. Whereas the Random Forest model did better at finding high-risk defaulters.
- **Focused Evaluation Metrics: What Matters**
- Getting predictive power is not good enough in the finance world. What's more important is the recall, which tells you how good a model did to identify defaulters. Confusion matrices and classification reports helped with accuracy which led us to find that Random Forest reduced false negatives better than Logistic Regression. Failing to catch a user at risk of hitting the default can lead to risky credit. That's why it is very important that I focus on the right metrics in order for this project to match real-world priorities.
- **Feature Important for Real World Explanation**
- Random Forest helped extract feature importance scores. By having the Random Forest model create a bunch of decision trees, it was able to tell me which variable the model relied on the most. The variables that the model relied on the most were credit utilization and repayment history ranked. This helps share insights with users before making any credit decision.

Default by Gender - Visualization

- **Graph Conclusion:**

- This is a bar chart where 1 is males and 2 is Females. Although the bar chart shows that Females have more defaulters you have ratio that theres more Females that are NOT at risk of default than Males. But gender by itself was not a strong predictor.

Gender
Female 0.207763
Male 0.241672

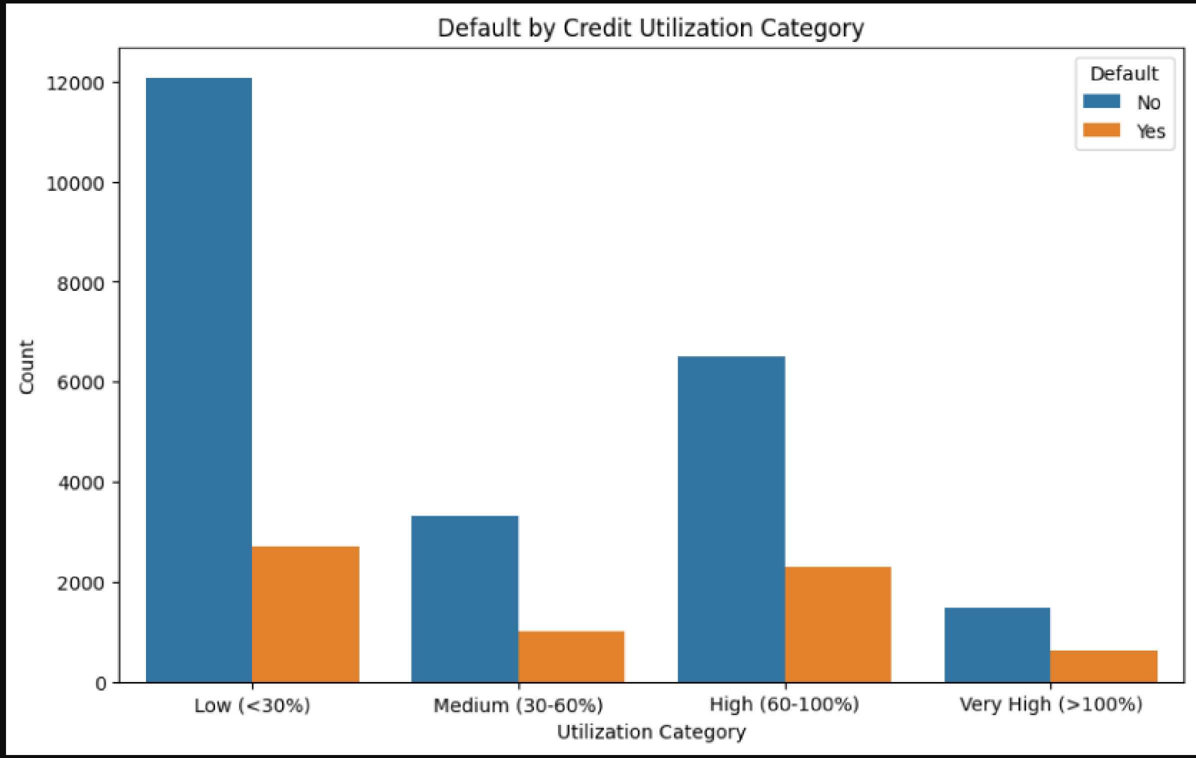


Default by Utilization Category - Visualization

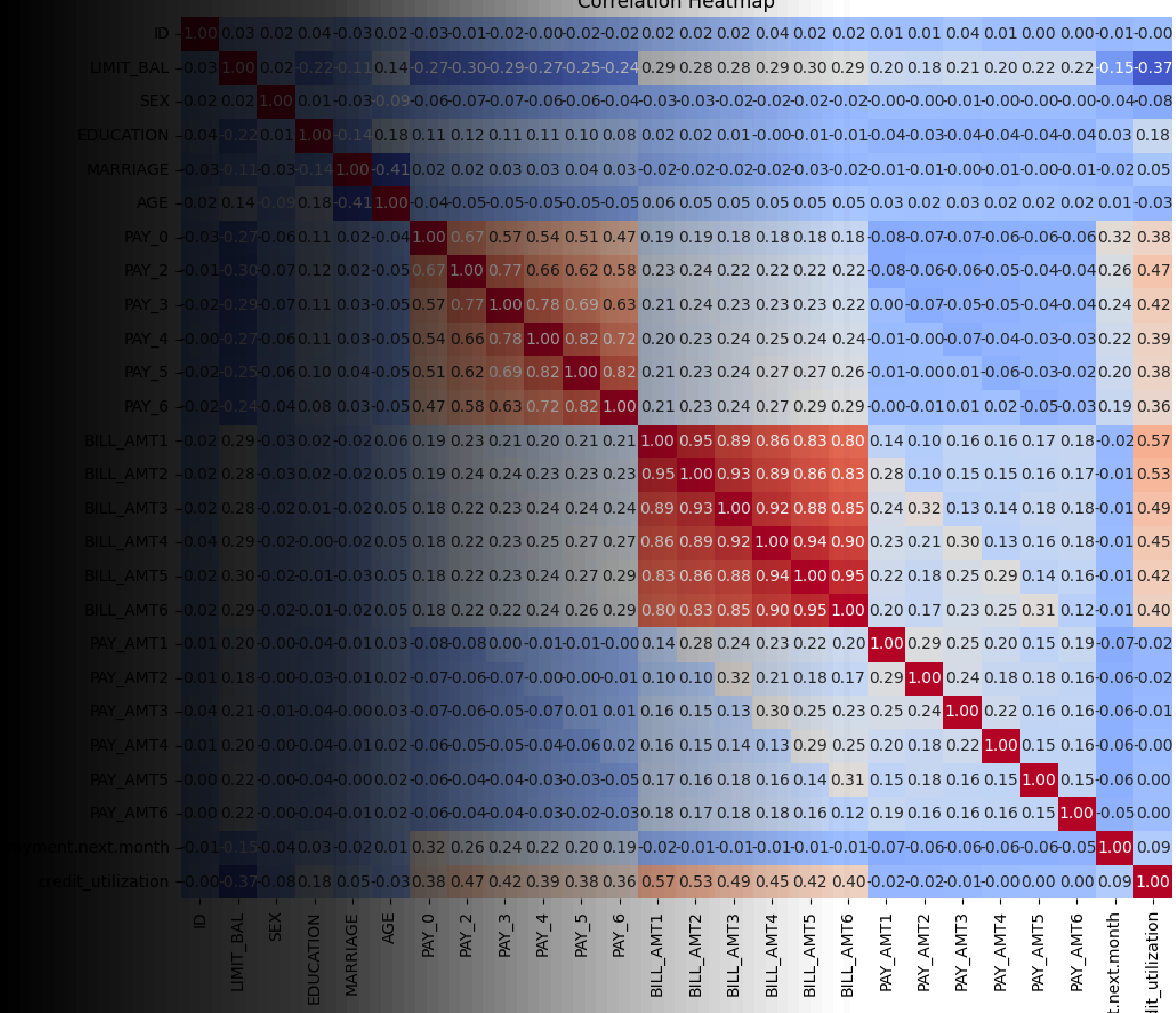
- Graph Conclusion:
  - Default increases when credit utilization category increase. It shows that the very high category has the highest number of defaults.

Utilization Category:

Low (<30%)	0.183060
Medium (30-60%)	0.232002
High (60-100%)	0.260983
Very High (>100%)	0.300709



Correlation Heatmap - Visualization



- **Graph Conclusion:**
  - The heatmap shows the correlation between default staus and payment delays (PAY_0 & PAY_2) is positive while bill and payments correlation is negative.

Limitations & Next Steps

- One limitation was that the dataset is a bit outdated, meaning the way users use credit and borrow money has changed from then to today. A second limitation that I realized is that there's no constant update on a user's income, credit limit, bills, and more. Not seeing how these certain variables go up and down made it difficult to see if they could have a chance of hitting default. The final third limitation was that the dataset didn't have a balanced mix of users who paid and didn't pay their bills. It involved more non-defaulters than defaulters, which makes it harder for predictive models to find the people who might default.

Self Evaluation

- I was responsible for everything that is included in this project. I remained organized although I did this project by myself.