Exploratory Data Analysis: Default of Credit Card Clients

Arewa Morountudun Ojelade
Department of Data Analytics: University of Maryland Global Campus
DATA 645: Machine Learning
Dr. Firdu Bati
October 28, 2025

Introduction

Credit card balances are unsecured debts that are not backed by an asset. Lenders' recovery prospects depend partly on the type of debt; consequently, it is crucial that lenders accurately assess borrowers' default probabilities for unsecured debt (Chen, 2024). This exploratory data analysis evaluates features that help determine the likelihood of client default using the dataset collected from the 2009 *Default of Credit Card Clients* research. The dataset contains client information from a Taiwanese bank, including credit limit, age, and sex, as well as the six-month payment history from April to September 2005, and the target feature, default status in October.

What is a Default?

A credit card default in the United States occurs after 180 days of missed payments. Each month of nonpayment results in a negative hit to the credit score. As delinquency ages, the total balance owed increases due to aggregated fees and interest. Prolonged nonpayment results in account closure followed by a declaration of loss by the lender. In most cases, the lender sells the debt to a collection agency and recovers most of it, but at other times, the lender pursues legal action on the defaulted account. In either case, considerable profits are lost, and the lender must avoid losses if they are predictable. (Bucci, 2021)

Risk Analysis

The process of analyzing the probability that a borrower will default on a credit card loan is called Credit Default Risk Analysis. There are three primary considerations during analysis: Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD). To make informed decisions on interest rates and credit limit extensions, lenders must leverage statistical and quantitative models to analyze collected data and risk metrics to determine the probability of default at any time. This analysis investigates and prepares the data set for the credit risk model and summarizes key insights.

Exploratory Analysis

The multivariate data were mined from research on customers' credit card default payments in Taiwan, using banking payment data from April to October 2005. The purpose of the dataset is to classify whether customers defaulted in October 2005, leveraging payment history from the previous 6 months. "Among the total 25,000 observations, 5529 observations (22.12%) are the cardholders with default payment (Yeh & Lien, 2009)." The variables in this data set are integer-valued, with 23 features across nine segments, and contain no missing values. The six segments that contain univariate features are *LIMIT_BAL*, *SEX*, *EDUCATION*, *MARRIAGE*, *AGE,* and *default payment next month.* The three multivariate segments include the payment history segment *PAY_0* to *PAY_6*, the bill statement amount segment *BILL_AMT1* to *BILL_AMT6,* and, lastly, the payment amount segment *PAY_AMT1* to *PAY_AMT6*. The features *SEX*, *MARRIAGE*, *EDUCATION,* and *PAY_AMT1* to *PAY_AMT6* are nominal features that were pre-encoded into discrete finite attributes. In the case of *SEX*, the dataset is encoded to label 1 = male and 2 = female. Similarly, *MARRIAGE* is encoded 1 = married, 2 = single, and 3 = others. Lastly, the *EDUCATION* feature is encoded as 1 = graduate school, 2 = university, 3 = high school, and 4 = others (Yeh, 2016).

There are some inaccuracies in the dataset documentation. For example, the documentation states that for the features *PAY_0* to *PAY-6*, which delineate the history of past payments, the encoded measurement scale for the repayment status of *pay duly* is -1, and delinquent payments are represented by a positive value equal to the number of months of delayed payment (Yeh, 2016). However, the dataset tells a different story. There exist unmentioned values of 0 and -2.

This discrepancy requires an assumption about the original intention of encoded values. This report assumes that a value of 0 represents a payment that is duly paid, and the absolute value of the value indicates the number of months of payment delinquency. This assumption is based on the distribution of plotted values for each payment feature, which will be explained in more detail in the preprocessing section of this analysis.

Once the dataset has been loaded into a pandas DataFrame, the info function provides a statistical summary displayed in Appendix A with the total count of 30,000 records in the data

set. For this report, currency has been converted from New Taiwan Dollar to USD. Details on the conversion are located in the preprocessing section. The *LIMIT_BAL* feature has a mean of $5,526.98 USD, a minimum of $330 USD, and a maximum of $33,000 USD. The first quartile (Q1) is at $1,650 USD, and the third quartile (Q3) is at $7,920 USD. The central tendency, or 50th percentile, is $4,620. The central tendency for the bill amount is the mean value of the 50th percentile of all bill amount features: $\sum_{i=1}^{n} BILL\_AMTi$ = $648.43, where n=6, the total number of previous monthly bill statements in the data set. The central tendency for the payment amount is the mean value of the 50th percentile of all payment amount features: $\sum_{i=1}^{n} PAY\_AMTi$ = $57.24, where n = 6, the total number of previous monthly payments in the dataset. The above is a calculation of the "mean of central tendencies", which is the arithmetic average of the measures of central tendency. On average, customers' payment central tendency is 8.82% of the bill amount.

Figure 1

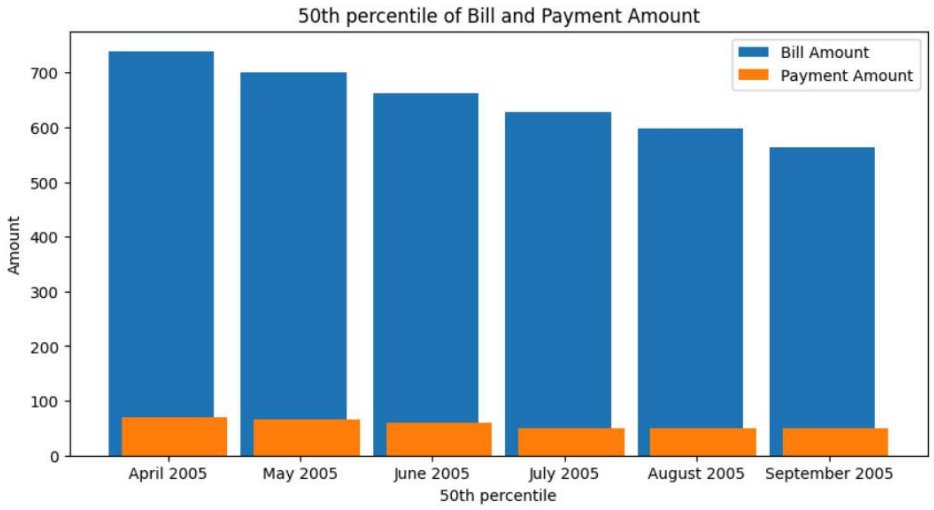
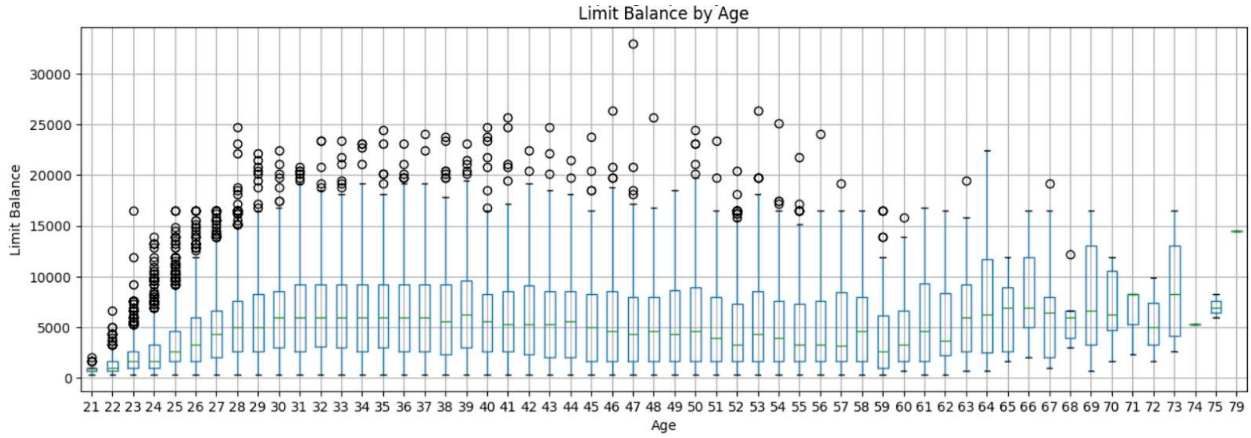


Figure 1 shows the central tendency-dispersion plot for all six months of bill and payment amounts. Appendix A presents the dataset's summary statistics table.

The box plot in Figure 2 below visualizes the limit balance feature grouped by age. Notice the logistic trend between age and limit balance. The limit balance slowly increases until age 30, where it levels off, decreasing in the late 40s and then increasing again in the early 60s.

Figure 2



Preprocessing

Data Cleaning

The original dataset contained no missing values, and nominal features were pre-encoded as integer values documented in the research dataset documentation (Yeh, 2016). However, these codes were not standardized. One cleaning activity performed on this dataset was standardization of the encoded features.

SEX: Initially, male was encoded to 1 and female to 2; however, this is not the standard encoding for sex in a dataset. As a result, male was standardized to 0 and female to 1.

```
df.loc[df['SEX'] == 1,'SEX'] = 0 # Standardize Male as 0
df.loc[df['SEX'] == 2, 'SEX'] = 1 # Standardize Female as 1
df.head()
```

MARRIAGE: Marital status was re-encoded and standardized to start at 0 = single, and 2 = other. The value 1 remains unchanged, representing married.

```
df.loc[df['MARRIAGE'] == 2, 'MARRIAGE'] = 0 # Standardize Single as 0
df.loc[df['MARRIAGE'] == 3, 'MARRIAGE'] = 2 # Standardize  as Other as 2
df.head()
```

EDUCATION: Similarly, the education feature was standardized to 0 = high school, 1 = university, 2 = graduate, and 3 = other.

```
df.loc[df['EDUCATION'] == 3, 'EDUCATION'] = 0 # Standardize Highschool as 0
df.loc[df['EDUCATION'] == 2, 'EDUCATION'] = 5 # Placeholder
df.loc[df['EDUCATION'] == 1, 'EDUCATION'] = 2 # Standardize Graduate as 2
df.loc[df['EDUCATION'] == 4, 'EDUCATION'] = 3 # Standardize Other as 3
df.loc[df['EDUCATION'] == 5, 'EDUCATION'] = 1 # Standardize University as 1
df.head()
```

Feature Rename: The ID column was dropped from the dataset because it does not contribute to the target feature. The feature *PAY_0* was renamed *PAY_1*, and the target feature '*default payment next month'* was renamed '*DEFAULT'* to facilitate easier data processing and to align with established feature naming standards.

```
df.drop(columns=['ID'], inplace=True)
df = df.rename(columns={'default payment next month': 'DEFAULT',
                        'PAY_0' : 'PAY_1'})
df.head()
```

Currency Conversion:

The original currency of this dataset is the New Taiwan Dollar (NTD), as the data were collected in Taiwan. However, for a convenient frame of reference, currency amounts such as limit balance, bill, and payment amounts were converted to the United States Dollar (USD) using the conversion rate of 1 NTD = 0.033 USD at the time of publication.

```
col_name = ['LIMIT_BAL','BILL_AMT1','BILL_AMT2','BILL_AMT3','BILL_AMT4','BILL_AMT5','BILL_AMT6','PAY_AMT1','PAY_AMT2','PAY_AMT3','PAY_AMT4','PAY_AMT5','PAY_AMT6']
for col in col_name:
  df[col] = df[col] * .033 #convert into USD
```

Delinquency Distribution:

The Exploratory Analysis section briefly noted the discrepancy between the research dataset documentation and the encoded delinquency values contained in the dataset. The documentation does not refer to the representation of the values 0 and -2. "The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above"

(Yeh, 2016). The explained format, as well as the delinquency distribution, lays the foundation for the following assumptions.

1. Delinquency values are positive.
2. Repayment status of *pay duly* is represented by 0

To rectify this discrepancy, all delinquency values were set to the absolute value of the original value, and for this report, zero is recognized as *pay duly.* Figure 3-2 displays the distribution of payment delinquency based on the assumptions made in this preprocessing step.

Figure 3-1

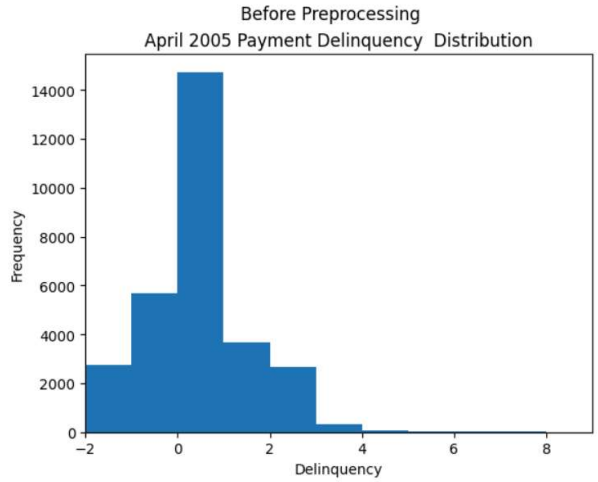
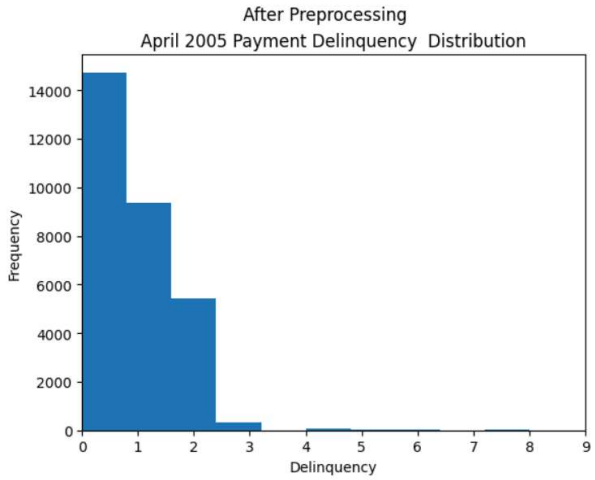


Figure 3-2



Outliers:

Outliers were handled by leveraging the Z-score outlier formula to eliminate the bill and payment averages $\geq \ |3\sigma|$ (standard deviations).

Figure 4-1

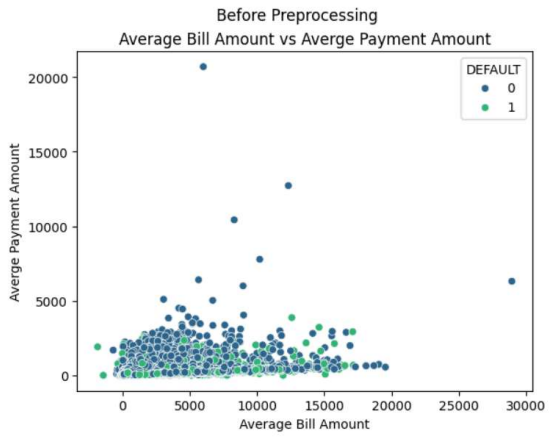
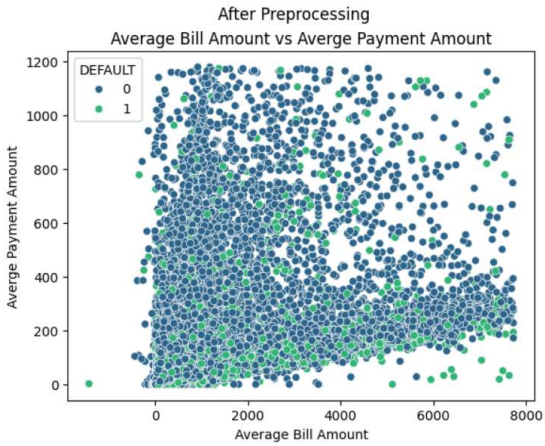


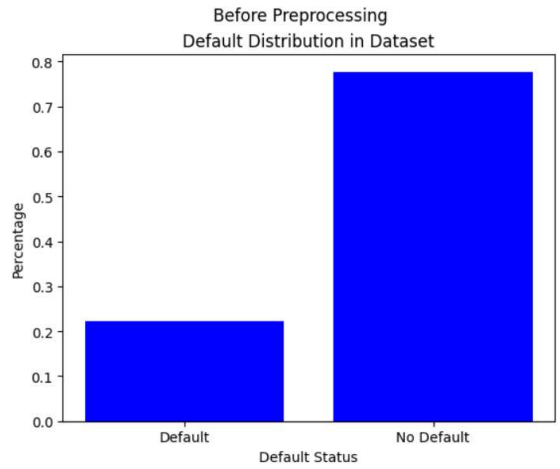
Figure 4-2



Data Transformation

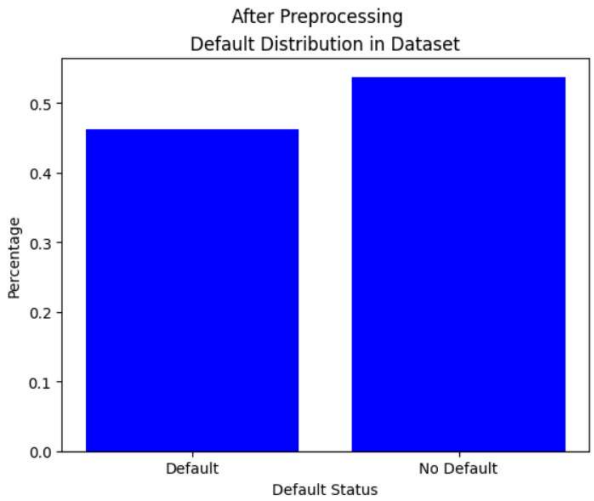
Tiling: The dataset's default target feature accounts for only 22% of default instances. To ensure there are enough default instances for an accurate model, the dataset was split into training and test sets to prevent data leakage, retaining the initial default status distribution. Next, the training set was separated by default status, and the training dataframe with status 1 (default) was tiled or expanded by a multiple of 3, increasing the proportion of positive default status to 46%.

Figure 5-1




```
0.22 of the dataset has defaulted
0.78 of the dataset has not defaulted
```

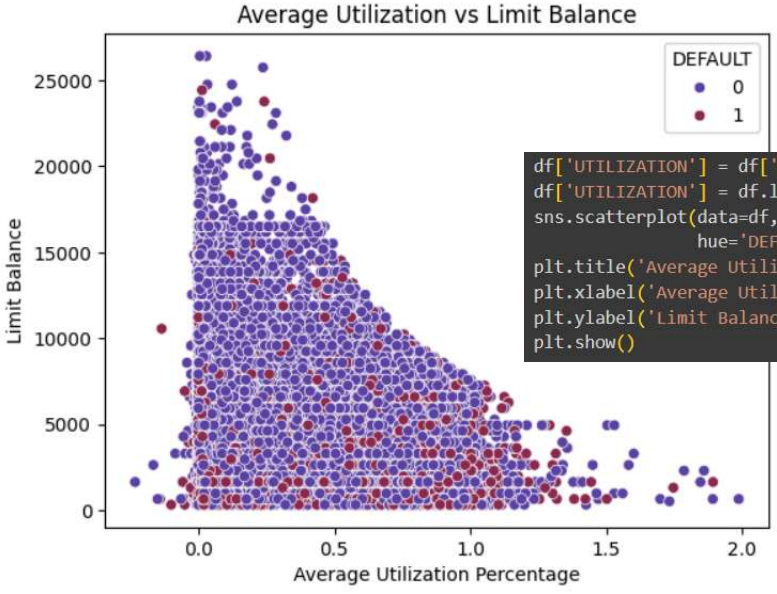
Figure 5-2




```
0.46 of the dataset has defaulted
0.54 of the dataset has not defaulted
```

Feature Construction: The quotient of each customer's average bill amount and the limit balance was used to construct the utilization percentage feature. As the average utilization percentage increases, the maximum limit balance decreases. There is a higher concentration of defaults at lower balance limits.

Figure 6




```
df['UTILIZATION'] = df['AVG_BILL_AMT']/df['LIMIT_BAL']
df['UTILIZATION'] = df.loc[df['UTILIZATION'] <= 2, 'UTILIZATION']
sns.scatterplot(data=df, x='UTILIZATION', y='LIMIT_BAL',
                hue='DEFAULT', palette='twilight')
plt.title('Average Utilization vs Limit Balance')
plt.xlabel('Average Utilization Percentage')
plt.ylabel('Limit Balance')
plt.show()
```

Data Reduction

PCA: Figure 7-1 displays the explained variance for each component in the training dataset. In Figure 7-2, the 3D graph showcases the top 3 principal components, which explains 49.49% of the variance, meaning these three components account for half of the dataset.

Figure 7-1

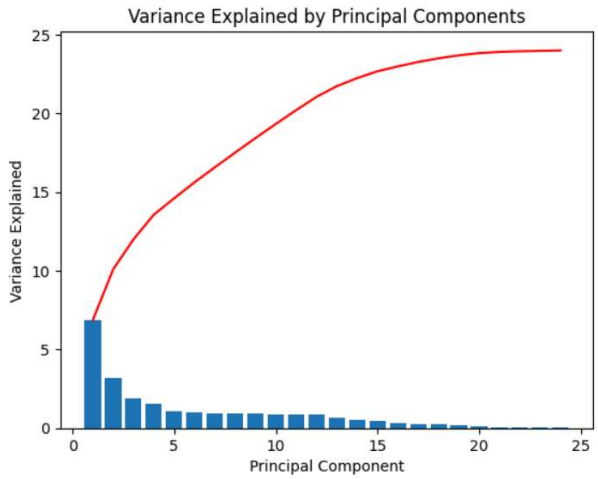
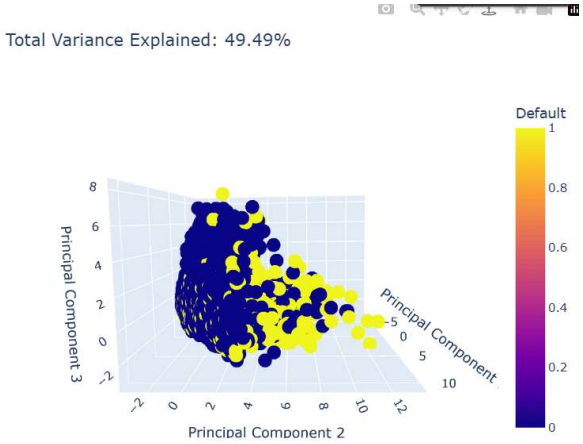
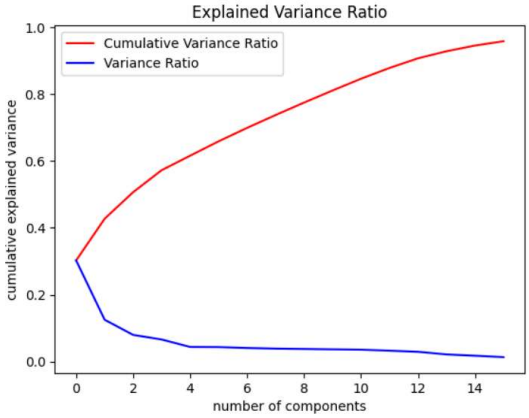


Figure 7-2



Principal Component Analysis (PCA) with a 95% variance ratio was used to select the most important components for dimensionality reduction after applying the StandardScaler function to the training and test sets separately. Figure 8 compares the variance ratios of each selected component with the cumulative ratio. The dataset has been effectively reduced from 24 to 16 features, which summarize 95% of the dataset.

Figure 8



The trade-off between data loss and data compression must be considered when testing model accuracy. PCA can help reduce dimensionality; however, it may not justify the reduction in accuracy.

Conclusion

In conclusion, risk analysis plays a critical role in lenders' decision-making when extending credit to a borrower. Since credit card balances are unsecured, uncollateralized debt, it is essential that lenders accurately assess borrowers' default risk. There are three primary considerations during risk analysis: Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD). This Exploratory Data Analysis (EDA) focuses on PD analysis. The multivariate dataset used for this analysis was collected from the *Default of Credit Card Clients* research conducted in 2009, which included client information from a Taiwanese bank spanning a six-month payment history from April to September 2005, with the default status in October 2005.

The provided raw dataset had been pre-cleaned to some extent, with pre-encoded nominal features like *SEX*, *EDUCATION*, *MARRIAGE*, etc., and no null values. However, the pre-encoded values were not standardized and consisted of labeling inaccuracies due to discrepancies in the documentation. Consequently, additional cleaning was required to standardize the encoded nominal values. The distribution of payment status features laid the foundation for the data-driven assumption that status 0 represents *pay duly.* This assumption, along with the original intent expressed in the research documentation, was used to rectify the discrepancy between the research documentation and the pre-encoded nominal values for PAY_1 to PAY_6.

Feature construction was implemented for the *UTILIZATION* feature by dividing the average bill amount by the limit balance. To ensure sufficient default instances in the dataset for an optimized model, records with a default status of 1 (default) were tiled to increase the proportion from 22% to 46%. Tiling was Principal Component Analysis (PCA) conducted after scaling the dataset with StandardScaler, selecting 16 features from 24, summarizing 95% of the dataset. Although PCA is useful for feature selection, it can degrade model accuracy, underscoring its limitations in this regard. Some potential areas for future improvement include splitting the dataset into training, cross-validation, and test sets. The dataset is now standardized, preprocessed, transformed, and ready for modeling.

References

Bucci, S. (2021, February 11). *Credit card default: How it happens, what to do about it*. Bankrate; Bankrate.com. https://www.bankrate.com/credit-cards/advice/credit-card-default/#score

C, P. (2022, August 3). *EDA - Exploratory Data Analysis: Using Python Functions*. Digital Ocean. https://www.digitalocean.com/community/tutorials/exploratory-data-analysis-python#exploratory-data-analysis-eda

CFI Team. (n.d.). *Credit Risk Analysis Models*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/commercial-lending/credit-risk-analysis-models/

Chen, J. (2024, June 29). *Exploring the Types of Default and the Consequences*. Investopedia. https://www.investopedia.com/terms/d/default2.asp

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning*. Springer Nature. https://www.statlearning.com/

Mulla, R. (2021, December 31). *Exploratory Data Analysis with Pandas Python 2023*. Www.youtube.com. https://www.youtube.com/watch?v=xi0vhXFPegw

ostwalprasad. (2025). *2019-01-20-PCA Using Python*. GitHub. https://github.com/ostwalprasad/ostwalprasad.github.io/blob/master/jupyterbooks/2019-01-20-PCA%20using%20python.ipynb

Ryan & Matt Data Science. (2023, September 7). *PCA Analysis in Python Explained (Scikit - Learn)*. YouTube. https://www.youtube.com/watch?v=6uwa9EkUqpg

Simplilearn. (2018). Data Science With Python | Python for Data Science | Python Data Science Tutorial | Simplilearn. In *YouTube*. https://www.youtube.com/watch?v=mkv5mxYu0Wk

Stanford. (2014). *Data Mining*. http://infolab.stanford.edu/~ullman/mmds/ch1.pdf

Yeh, I-Cheng. (2016, January 25). *UCI Machine Learning Repository*. Archive.ics.uci.edu. https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

Yeh, I-Cheng., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020