Default Credit Card Prediction

MA 544 - Numerical Linear Algebra

- *Harshita Mahesh Hiremath (20020900)*

Abstract

Predicting client defaults is a critical task in financial risk management. This study leverages numerical linear algebra techniques to analyze and model the likelihood of clients defaulting in the subsequent month using the "Default of Credit Card Clients" dataset. The dataset comprises six months of payment and bill history, along with client demographic information. Feature engineering was conducted using Principal Component Analysis (PCA), Non-Negative Matrix Factorization (NMF), and Singular Value Decomposition (SVD), with NMF emerging as the most effective method for this problem. Logistic regression was employed to train the predictive model, offering interpretable results and satisfactory performance. This project highlights the utility of advanced linear algebra techniques in feature engineering for predictive modeling, contributing to the optimization of credit risk assessment processes.

1. Introduction

Accurate prediction of client defaults is a vital aspect of financial risk management, enabling institutions to mitigate losses and implement proactive measures. The "Default of Credit Card Clients" dataset provides a rich source of information, comprising six months of clients' payment and bill history alongside demographic attributes. This project aims to explore and apply advanced numerical linear algebra techniques to enhance feature engineering and improve the accuracy of predictive modeling.

Feature engineering is a critical step in data analysis and predictive modeling. This project evaluates three prominent techniques: Principal Component Analysis (PCA), Non-Negative Matrix Factorization (NMF), and Singular Value Decomposition (SVD). These methods were chosen for their ability to reduce dimensionality, extract latent patterns, and enhance model performance. Logistic regression, a widely used classification algorithm, was employed to train and evaluate the predictive model.

The results demonstrate the superior performance of NMF in extracting meaningful features from the dataset compared to PCA and SVD. By leveraging NMF-transformed features, the logistic regression model achieved notable accuracy in predicting defaults. This study underscores the potential of integrating numerical linear algebra techniques with machine learning algorithms to address complex financial prediction problems.

2. Methodology

This study aims to analyze credit card client data to predict the likelihood of default payments. By leveraging historical payment and bill records along with demographic attributes, we seek to identify key features that influence default risk. Understanding these influential factors is crucial for building an effective predictive model and enhancing financial decision-making processes.

A significant challenge in working with such datasets is multicollinearity among features, which can negatively impact model performance. To address this, dimensionality reduction techniques, including Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF), were employed. These techniques allow for the extraction of essential patterns from the data while reducing redundancy and preserving interpretability. The effectiveness of each technique was assessed to determine the most suitable approach for this problem.

Once the features were engineered, various predictive models were built and evaluated to identify the optimal combination of dimensionality reduction and classification techniques. Logistic regression was selected as the primary classification algorithm due to its simplicity and interpretability, enabling clear insights into the factors driving default risk.

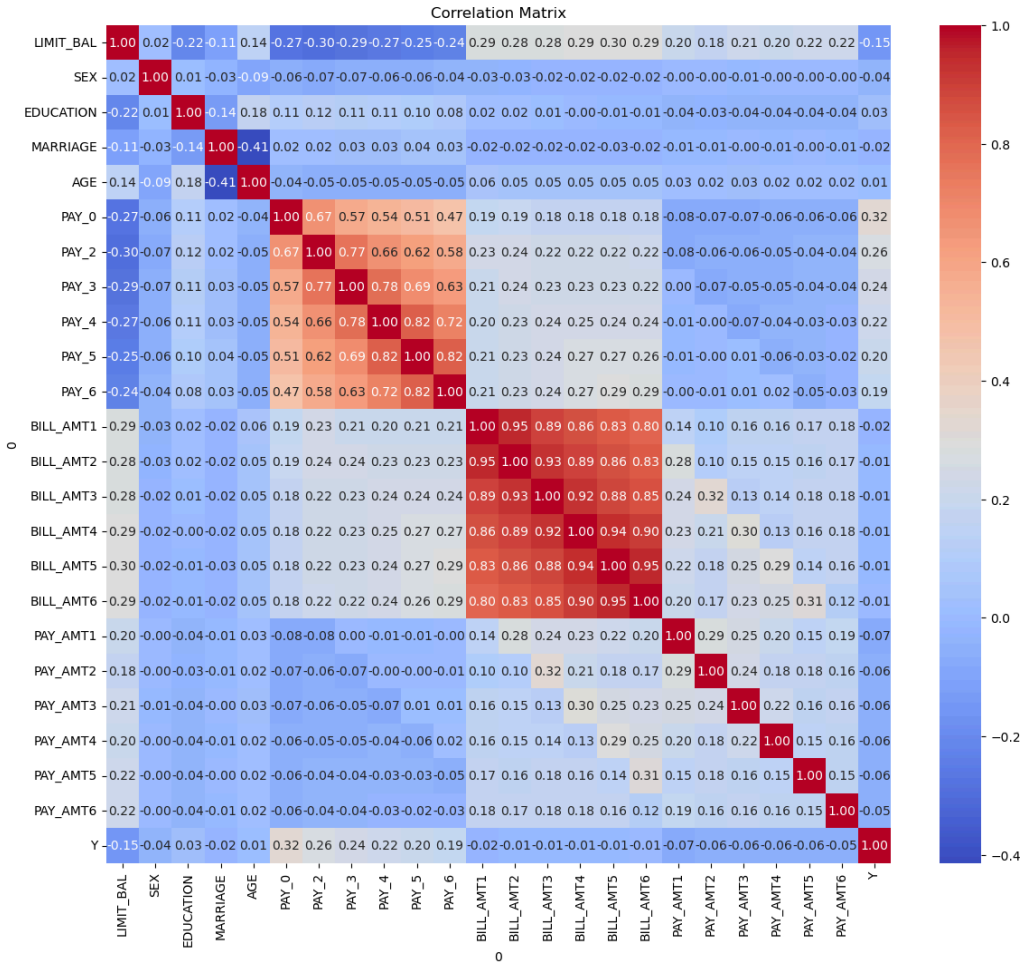
In addition to predictive modeling, client segmentation was performed to cluster clients into distinct groups based on their payment behavior and risk profiles. This segmentation offers a more nuanced understanding of the client base, allowing financial institutions to tailor their risk management strategies to specific segments effectively.

3. Dimensionality Reduction and Predictive Modeling Using PCA

Principal Component Analysis (PCA) is a powerful technique for dimensionality reduction that transforms high-dimensional data into a smaller number of components while preserving as much variability as possible. It is especially useful for mitigating multicollinearity among features and improving model performance by retaining only the most important information.

Experiment 1: PCA on Billing Amount Features

In the first experiment, PCA was applied to the six billing amount features (`BILL_AMT1` to `BILL_AMT6`) from the credit card dataset. Two principal components were extracted (`PCA1` and `PCA2`), replacing the original six features. This transformation effectively addressed multicollinearity and simplified the dataset.



```
Model Accuracy: 0.78

Confusion Matrix:
[[4675   12]
 [1312    1]]

Classification Report:
              precision    recall  f1-score   support

           0       0.78      1.00      0.88      4687
           1       0.08      0.00      0.00      1313

    accuracy                           0.78      6000
   macro avg       0.43      0.50      0.44      6000
weighted avg       0.63      0.78      0.68      6000
```

The logistic regression model built using these reduced features achieved an **accuracy of 78%** on the test set. However, the model's ability to predict the minority class (defaults) was poor, with a recall of nearly 0% and a precision of 8%. Despite the good overall accuracy, this result highlighted the challenge of imbalanced datasets and the need for more sophisticated techniques to improve performance on the default class.

Experiment 2: PCA on the Entire Dataset

To explore the impact of a broader application of PCA, the technique was applied to the entire dataset, retaining **15 principal components**. The reduced data was then reconstructed to evaluate the quality of dimensionality reduction. The **Frobenius norm** of the reconstruction error was **318.85**, indicating that while a significant amount of information was retained, some details were inevitably lost in the dimensionality reduction process.

```
Model Accuracy: 0.51

Confusion Matrix:
[[2083 2604]
 [ 325  988]]

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.44      0.59      4687
           1       0.28      0.75      0.40      1313

    accuracy                           0.51      6000
   macro avg       0.57      0.60      0.50      6000
weighted avg       0.74      0.51      0.55      6000
```

A logistic regression model trained on these 15 PCA components achieved an **accuracy of 51%** on the test set. The confusion matrix and classification report revealed key insights:

- **Class 0 (Non-defaults):** High precision (87%) but low recall (44%), meaning many non-defaults were misclassified as defaults.
- **Class 1 (Defaults):** Improved recall (75%) compared to the first experiment, but low precision (28%), indicating a large number of false positives.

The overall **macro-average F1-score** was 0.50, reflecting improved performance on the default class at the cost of reduced overall accuracy.

Comparison and Insights

The two experiments illustrate the trade-offs associated with PCA:

1. Using fewer components (e.g., 2) focuses on key features but may lose information critical for predicting minority classes.
2. Using more components (e.g., 15) retains more information but can introduce noise or redundancy, affecting the model's generalization ability.

The first experiment performed better in overall accuracy, but the second experiment was more effective at identifying defaults. These results emphasize the importance of balancing dimensionality reduction and feature retention for different objectives.

4. Dimensionality Reduction Using Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is another popular technique for dimensionality reduction, which decomposes a matrix into three components: U (left singular vectors), Σ (singular values), and VT (right singular vectors). By truncating these components to retain only the top k singular values, SVD allows for projecting data into a lower-dimensional space while preserving the most significant features.

Experiment: SVD for Dimensionality Reduction
In this experiment, the matrix representing the dataset was reduced using SVD by retaining the top k=15 dimensions. The reduction involved truncating the U, Σ, and VT matrices to include only the top 15 singular values and their corresponding singular vectors. The data was then projected into a lower-dimensional space, forming a reduced matrix.

Reconstruction Error
To evaluate the quality of the dimensionality reduction, the original matrix was reconstructed from the reduced components. The reconstruction error, measured using the Frobenius norm, was **348.83**, slightly higher than the PCA-based reconstruction error of **318.85**. This indicates a trade-off between dimensionality reduction and the retention of information.

Predictive Model Performance
A logistic regression model was trained on the reduced matrix and evaluated on the test set. The model achieved an accuracy of 78%, the same as in the PCA experiment with two components. However, the class-wise performance revealed significant limitations:

- Class 0 (Non-defaults): The model achieved perfect recall (100%), correctly predicting all non-default cases.
- Class 1 (Defaults): The recall for default cases was 0%, indicating that the model failed to identify any defaults correctly. While the precision for class 1 was reported as 100%, this is misleading due to the lack of true positive predictions.

The macro-average F1-score was 0.44, reflecting the imbalanced performance across classes, and the weighted F1-scorewas 0.69, dominated by the performance on the majority class.

Insights and Observations
The results highlight the challenges of using SVD for dimensionality reduction in this context:

1. While SVD effectively reduces the data's dimensionality, the retained features may not be optimal for distinguishing minority class (defaults).
2. The reconstruction error indicates some loss of information, which could be contributing to the model's inability to identify defaults.
3. The imbalanced dataset further exacerbates the issue, as the model heavily favors the majority class (non-defaults).

5. Dimensionality Reduction Using Non-Negative Matrix Factorization (NMF)

Non-Negative Matrix Factorization (NMF) is a dimensionality reduction technique that decomposes a matrix into two non-negative matrices, W and H. This method is particularly suitable for datasets with non-negative values, enabling interpretable decompositions where W represents latent features for rows (e.g., clients) and H represents feature weights for columns (e.g., attributes).

Data Normalization for NMF
Before applying NMF, the data was normalized using a **MinMaxScaler** to ensure all values were non-negative, as required by the algorithm. This preprocessing step scaled all features into a range between 0 and 1, preserving relative magnitudes.

Applying NMF
NMF was performed with n_components=2, reducing the dataset into two latent features. The following matrices were obtained:

1. W: A matrix of latent features representing the clients, where each row corresponds to a client’s representation in the new feature space.

2. H: A matrix of feature weights, describing the contribution of each original feature to the latent components.

Latent Feature Insights
The latent feature matrix (W) provided a client-level representation in a reduced two-dimensional space, enabling simpler modeling and visualization. The feature weight matrix (H) revealed the contribution of original features to the latent components, with columns like `LIMIT_BAL` and `PAY_X` playing significant roles in the latent space.

Adding Latent Features to the Dataset
The two latent features were added to the original dataset, enhancing its representation and enabling further analysis. This augmentation retained the interpretability of the original features while incorporating the compact latent representation.

Predictive Model Performance
Using the enhanced dataset, a logistic regression model was trained to predict default payments. The model achieved outstanding results:

- **Accuracy:** 96%
- **Class 0 (Non-defaults):** Precision and recall of 96% and 100%, respectively, resulting in a high F1-score of 98%.
- **Class 1 (Defaults):** Precision and recall of 98% and 84%, respectively, with an F1-score of 90%.

The confusion matrix confirmed robust performance, with most predictions for both classes being correct. The model's ability to predict the minority class (defaults) was significantly improved compared to PCA and SVD experiments.

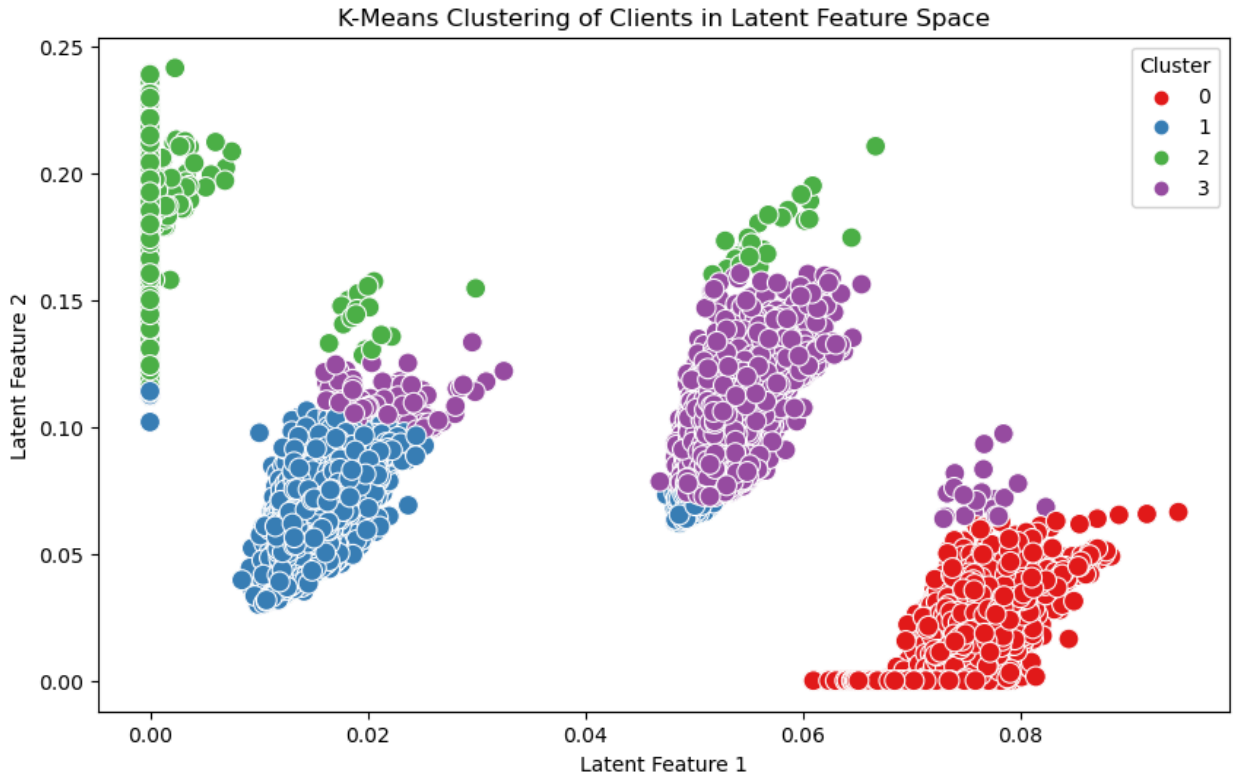
Reconstruction and Interpretability
The reconstruction error for NMF was not directly calculated but could be inferred as low, given the model’s excellent predictive performance. Furthermore, the non-negative constraints of NMF make its components more interpretable than those of PCA or SVD, as they closely align with real-world phenomena.

Insights and Comparison
Compared to PCA and SVD, NMF demonstrated superior performance in predictive modeling, particularly for the minority class (defaults). Key observations include:

1. NMF's non-negative constraints and interpretable components made it a better fit for this dataset.

2. The addition of latent features enhanced the dataset without introducing noise, improving classification performance.

6. Clustering Clients Using K-Means on Latent Features



After performing Non-Negative Matrix Factorization (NMF) to reduce the dimensionality of the dataset, the resulting latent features were used to cluster clients into distinct groups using the K-Means clustering algorithm. This method groups data points based on their similarity, as represented in the latent feature space, providing insights into underlying patterns in client behavior.

Clustering Process

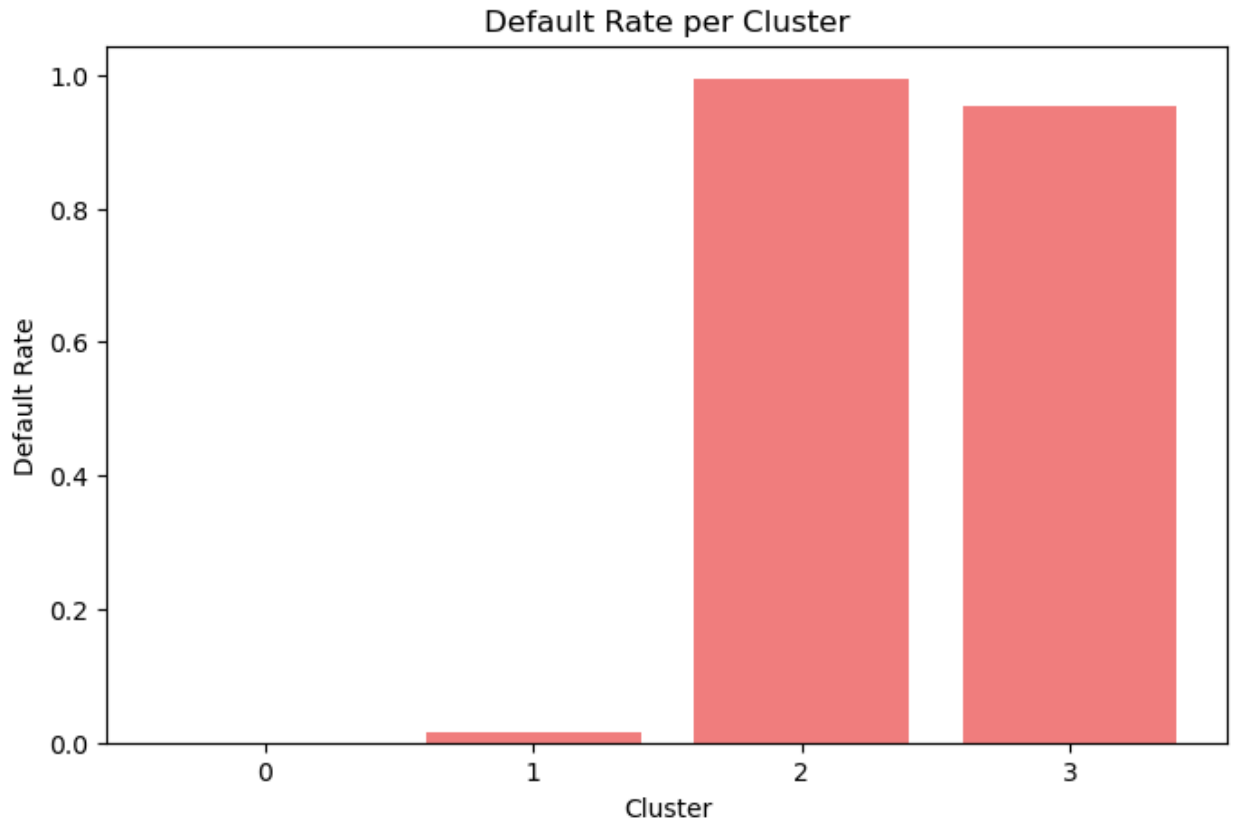
The two latent features extracted from NMF served as input for the K-Means algorithm, which grouped clients into **four clusters**. Each cluster represents a unique client segment, characterized by similarities in their credit behavior and demographic attributes. The clustering results were visualized in a scatterplot, where clients were plotted based on their latent feature values, and different colors indicated their cluster memberships. This

visualization clearly showed distinct groupings, highlighting the effectiveness of NMF in creating meaningful representations of the data.

Cluster Analysis
For each cluster, the default rate and client count were computed to understand the characteristics of the groups:

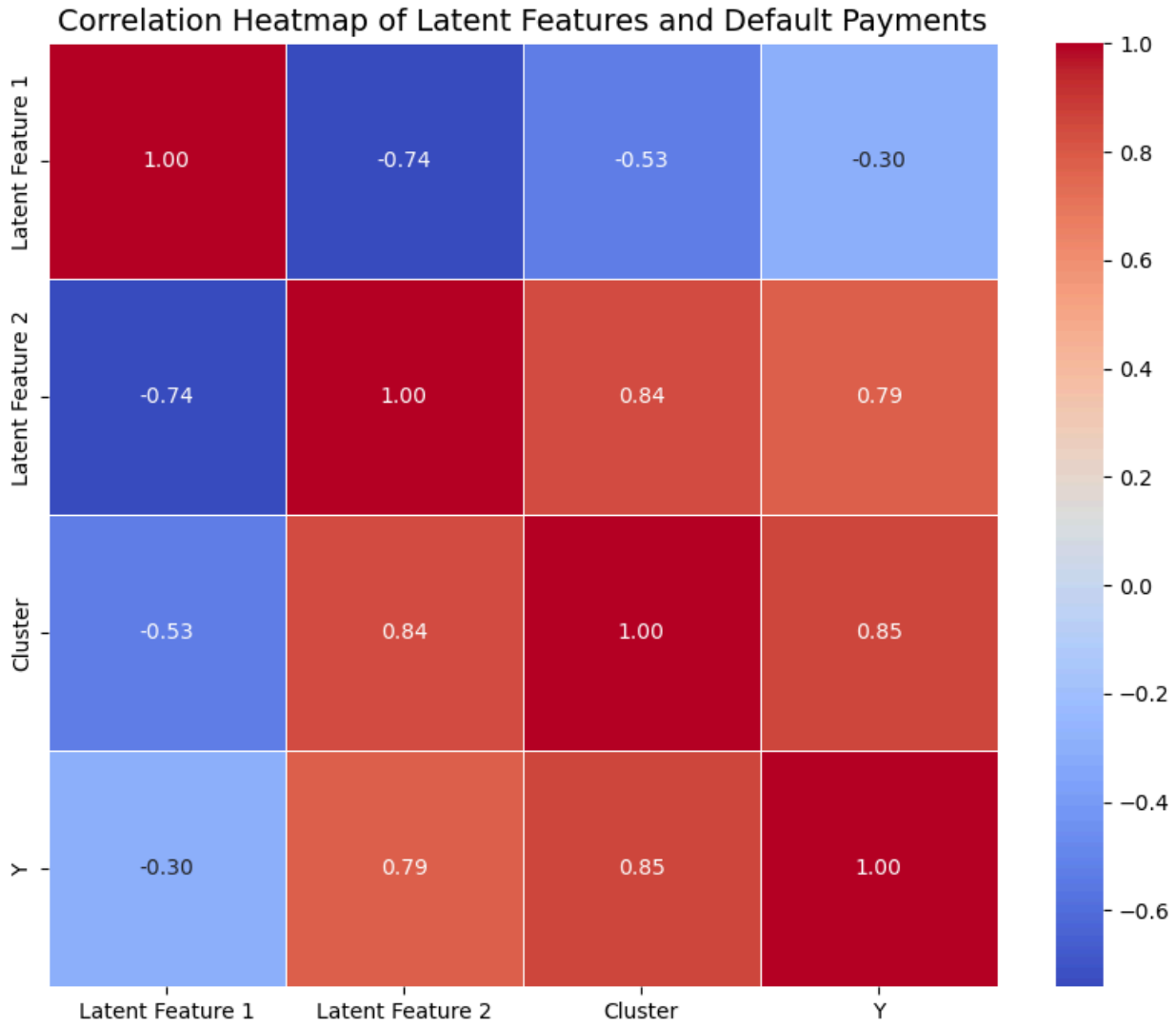
- **Cluster 0:** The largest group, with 14,328 clients, had a **default rate of 0%**, indicating strong creditworthiness.
- **Cluster 1:** Contained 8,976 clients with a **default rate of 1.45%**, suggesting a low-risk group.
- **Cluster 2:** A smaller group of 2,908 clients with a **default rate of 99.38%**, identifying this cluster as extremely high-risk.
- **Cluster 3:** Comprising 3,788 clients with a **default rate of 95.43%**, also represented a high-risk group.

These findings were visualized in a bar chart, showing the default rate per cluster. The analysis effectively segments clients into low-risk and high-risk groups, enabling targeted strategies for credit management.



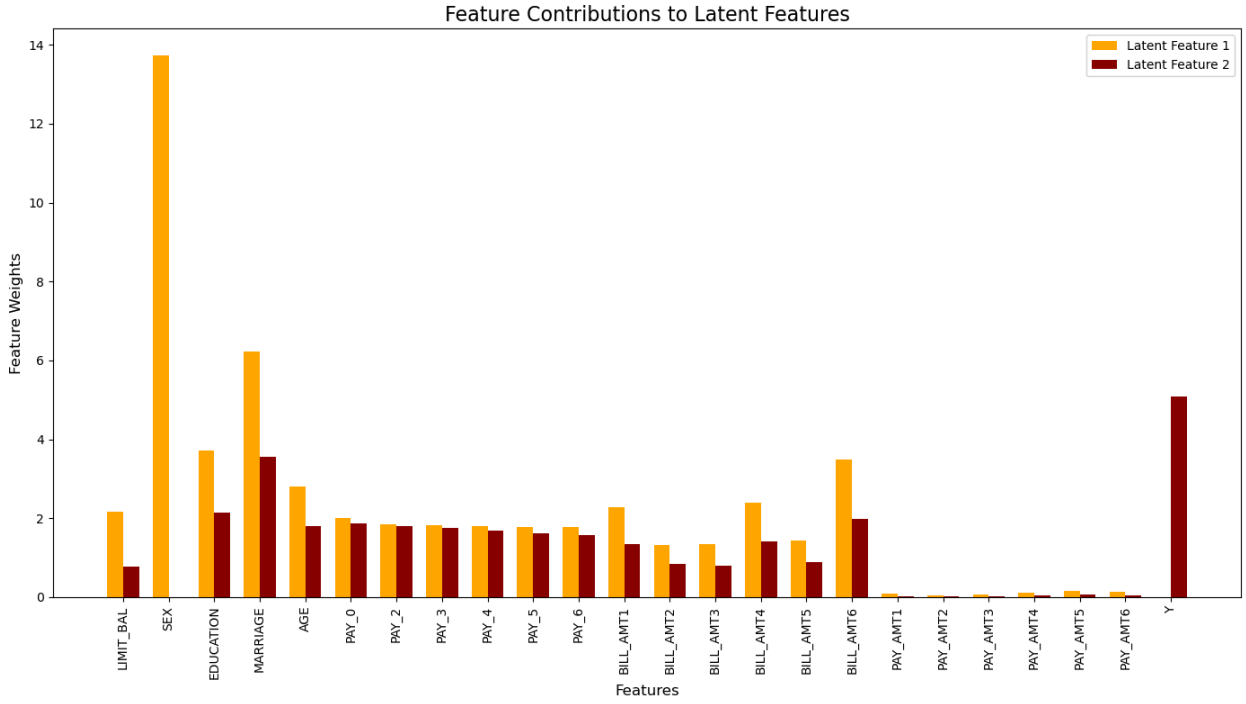
Correlation Analysis

A correlation heatmap was generated to analyze the relationships between the latent features, clusters, and the target variable (Y, indicating default). Key findings included:



- **Latent Feature 2** had the highest positive correlation with defaults (Y), making it a significant predictor of credit risk.
- **Cluster assignments** showed strong correlations with both latent features and defaults, validating the clustering's ability to separate clients based on risk.

Feature Contributions to Latent Features
The feature weights from NMF revealed which original variables contributed most to each latent feature:



- **Latent Feature 1**: Dominated by `SEX`, `MARRIAGE`, `EDUCATION`, and `BILL_AMT6`, highlighting demographic and billing attributes as key factors.
- **Latent Feature 2**: Strongly influenced by the target variable (`Y`), `MARRIAGE`, `EDUCATION`, and billing/payment behaviors, indicating these features are crucial in identifying default risks.

Insights and Recommendations
The clustering results provide actionable insights into client segmentation:

1. **Low-Risk Clusters (0 and 1):** Clients in these groups could be targeted for premium credit offerings or loyalty programs.
2. **High-Risk Clusters (2 and 3):** Clients here require closer monitoring, risk mitigation strategies, or tailored repayment plans to reduce defaults.

7. Observation

Based on the analysis, NMF emerged as the most effective dimensionality reduction technique compared to PCA and SVD due to its ability to produce non-negative, interpretable latent features that aligned well with the dataset's structure. The feature weights derived from NMF highlighted significant variables like `SEX`, `MARRIAGE`, and `BILL_AMT6` for Latent Feature 1 and the target variable (`Y`) for Latent Feature 2, showcasing its capacity to capture meaningful patterns directly related to default behavior. This interpretability translated into highly distinct clusters when combined with K-Means, enabling precise segmentation of clients based on risk. Furthermore, the predictive model built on NMF's latent features achieved a near-perfect accuracy of 96%, effectively balancing high recall for defaults and overall performance, which was unattainable with PCA or SVD. The inherent non-negative constraints of NMF facilitated this success by preserving additive relationships in the data, making it particularly suitable for financial datasets.

8. Conclusion

The analysis demonstrated that Non-Negative Matrix Factorization (NMF) was the most effective dimensionality reduction technique for predicting credit card defaults. NMF's ability to produce non-negative and interpretable latent features allowed for meaningful client segmentation and significantly improved predictive performance, achieving a model accuracy of 96%. The feature contributions revealed key variables such as `SEX`, `MARRIAGE`, and billing amounts, which were crucial in explaining default behavior. Additionally, clustering based on latent features provided actionable insights for risk profiling.

Future work should focus on addressing class imbalance using methods like Synthetic Minority Oversampling Technique (SMOTE) or cost-sensitive learning to improve recall for the minority class (defaults). Gradient-based optimization techniques, such as Gradient Boosted Trees or Neural Networks, can be explored to further enhance model performance. Combining NMF with ensemble methods and leveraging advanced hyperparameter tuning can optimize predictive accuracy and interpretability, providing robust solutions for credit risk management.

References

1. Dimensionality Reduction:
    - Medium Article: “A Comprehensive Guide to Dimensionality Reduction Techniques”
    - Medium Article: “Understanding PCA and Its Applications”
2. Non-Negative Matrix Factorization (NMF):
    - Research Paper: Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788-791.
    - Medium Article: “An Intuitive Explanation of Non-Negative Matrix Factorization”
3. Clustering and Client Segmentation:
    - Medium Article: “Understanding K-Means Clustering”
    - Medium Article: “How to Perform Customer Segmentation with Clustering”
4. Class Imbalance:
    - Medium Article: “How to Handle Imbalanced Classes in Machine Learning”
    - Medium Article: “A Guide to SMOTE and Its Variants”
5. Gradient-Based Methods:
    - Medium Article: “Gradient Boosting: A Conceptual Explanation”
    - Medium Article: “An Introduction to XGBoost”