STEVENS
INSTITUTE OF TECHNOLOGY
1870

Credit Card Default Prediction

MA 544 - Numerical Linear Algebra

- *Harshita Mahesh Hiremath*

Problem Overview

The problem focuses on predicting whether **credit card clients** will **default on their payments** in the next month. The goal is to analyze patterns in the data and build a predictive model that identifies clients at **high risk of default**, enabling better risk management and financial decision-making.

Objectives

- Analyze credit card client data to predict **default payments**
- Identify key features influencing **default risk**
- Use **dimensionality reduction** (PCA, NMF) to handle multicollinearity.
- Build and evaluate predictive model amongst various dimensionality reduction techniques
- Segment clients into clusters for **risk profiling**

Dataset Overview

- **Total Records**: 30,000 credit card clients.
- **Target Variable**: `default.payment.next.month` (1 = Default, 0 = No Default).
- **Key Features**:
- **Demographic Information**:
  - `LIMIT_BAL` (Credit limit), `AGE`, `SEX`, `EDUCATION`, `MARRIAGE`.
- **Payment Behavior**:
  - `PAY_0` to `PAY_6` (Payment status for the last 6 months).
    - Values: -1 (Paid on time), 1+ (Delay in months).
- **Billing History**:
  - `BILL_AMT1` to `BILL_AMT6` (Bill amounts for the last 6 months).
- **Payment Amounts**:
  - `PAY_AMT1` to `PAY_AMT6` (Amount paid in the last 6 months).
  - **Data Type**: Mostly numeric with no null values.
  - **Goal**: Predict the likelihood of default in the next month.

Principal Component Analysis

- PCA with 15 components retained reasonable variance (error = 318.85) but lost some information.
- Achieved **51%**, balancing performance but still insufficient.
- **Class 0 (Non-Default)**: High **precision (0.87)** but low **recall (0.44)**, missing many non-defaults.
- **Class 1 (Default)**: High **recall (0.75)** but low **precision (0.28)**, with many false positives.
- Struggles to balance precision and recall; needs feature refinement or class balancing techniques.

Singular Value Decomposition

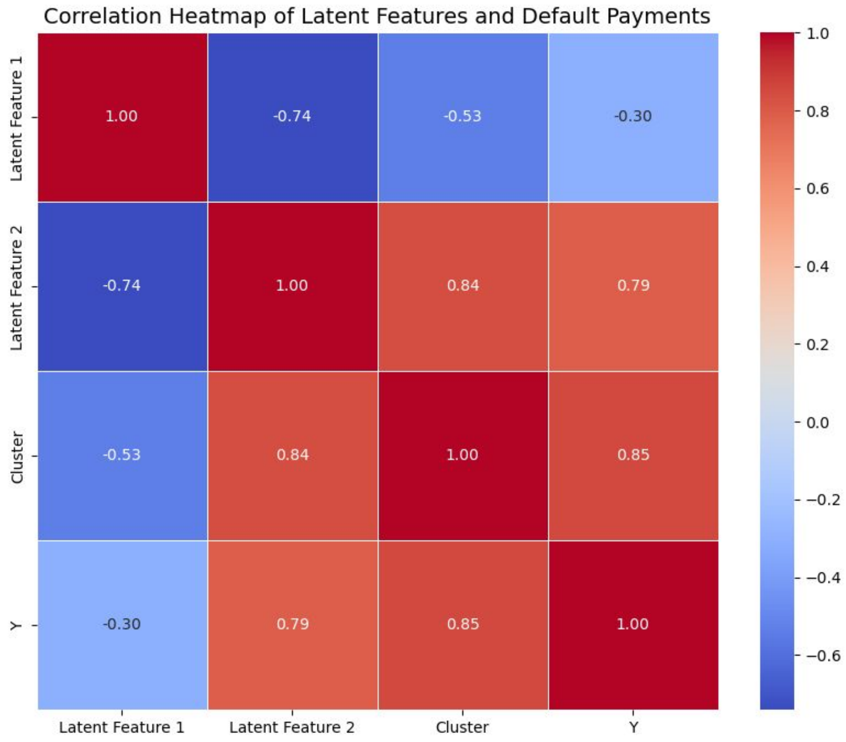
- Model achieved **78% accuracy**, reducing **23 features to 10** via SVD, but only predicts the majority class (`Y=0`).
- **Class 0 (Non-Default)**: Perfect recall (**1.00**) at the cost of ignoring defaults.
- **Class 1 (Default)**: Fails entirely, with **precision, recall, and F1-score = 0.00**.
- **SVD Limitation**: Latent features fail to capture patterns critical for distinguishing defaults.
- Suggests potential **loss of minority-specific information** during dimensionality reduction.

Nonnegative Matrix Factorization

- The model achieved **96% accuracy** with strong performance across classes.
- Used just **two latent features** generated by **Non-Negative Matrix Factorization (NMF)**.
- Captured **critical latent patterns** in client behavior effectively.
- Differentiated defaults ($Y=1$) with **high precision (0.98)** and **recall (0.84)**.
- Demonstrates the **power of dimensionality reduction** for predictive modeling.

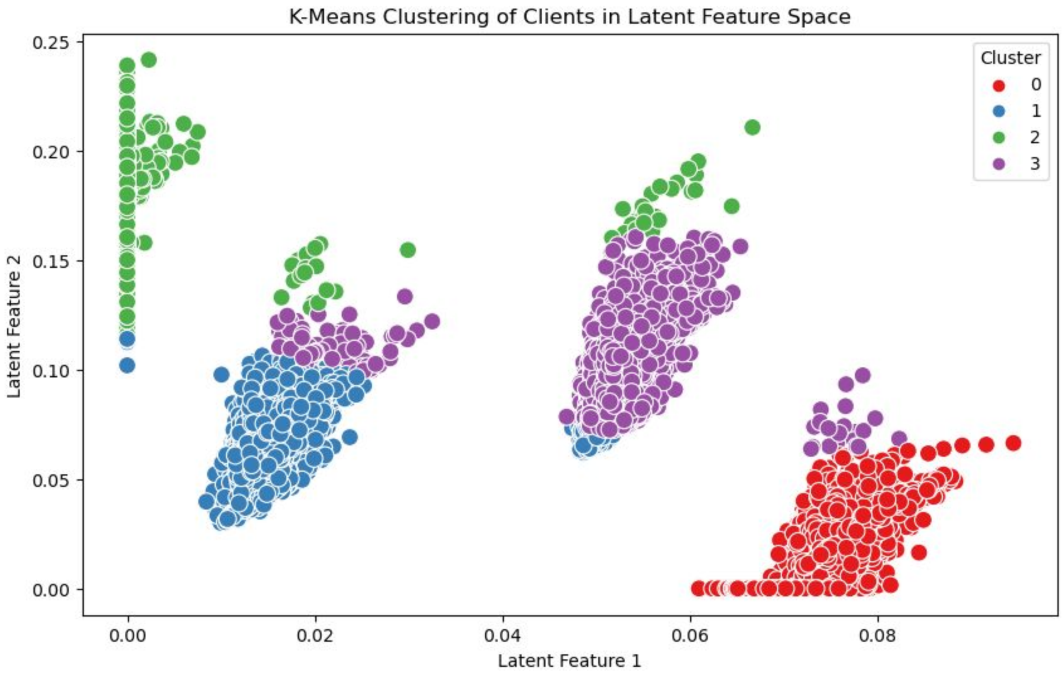
Correlation Matrix Analysis of Latent Features

- Latent Feature 1 has a weak or negative correlation, it suggests this feature is not strongly linked to default behavior.
- Latent Feature 2 has a strong positive correlation (e.g., 0.79), it means clients with higher values for this feature are more likely to default.



K-Means Clustering

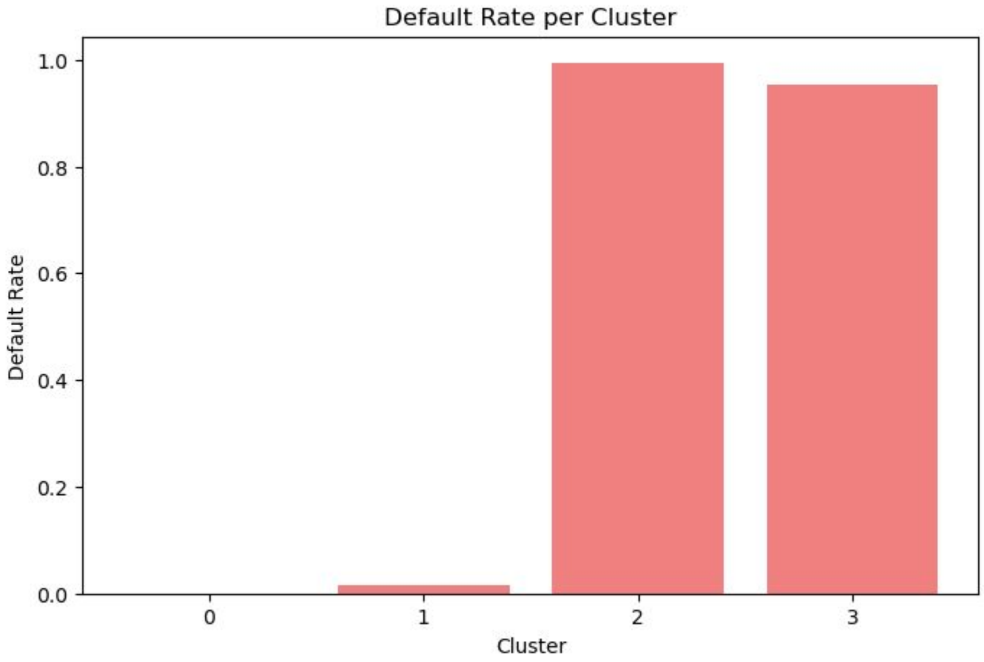
- **4 Clusters Identified**: Based on `BILL_AMT1` and `PAY_AMT1`.
- **Cluster 0 (Dark Purple)**: Low bills and payments → Likely low spenders or at-risk clients.
- **Clusters 1 & 2 (Green/Blue)**: Moderate bills and payments → Consistent behavior.
- **Cluster 3 (Yellow)**: High bills and significant payments → High spenders.
- **Outliers**: Few points in Cluster 0 show extreme payments or bills.
- **Insight**: **High bill amounts but low payments** (Clusters 0/1) may signal **default risk**.



Default Rate Per Cluster

- **Default Rates**: Clusters **2** and **3** show the **highest default rates** (~90%), indicating high-risk groups.
- **Cluster 0**: Very low default rate, suggesting minimal risk.
- **Bill vs Payment Behavior**:
    - Clients in **Cluster 3** (yellow) and **Cluster 2** show **high bills but low payments**, correlating with higher defaults.
    - Clusters with consistent payments exhibit lower default risk.

Insight: Focus on **Clusters 2 and 3** for intervention, as they represent high-risk clients with poor repayment behavior.



Conclusion

Class Imbalance: The dataset exhibited significant class imbalance, leading to poor model performance on the minority class ($Y=1$) in earlier approaches.
NMF Outperformed PCA and SVD:

- **NMF** effectively captured latent patterns, achieving **96% accuracy** with strong recall (**0.84**) for defaults.
- In contrast, **PCA** and **SVD** failed to retain critical information, leading to imbalanced and lower performance.

Key Insight: NMF's latent features provided better representation of client behaviors, making it the most effective technique for handling this problem.