Summary Report: A Data-Driven Framework for Behavioral Patterns in Credit Risk Prediction

Introduction:

Credit card defaults represent a major financial risk for banks and credit institutions. With the rapid growth of consumer credit, identifying customers likely to default has become crucial for sustainable financial operations.
This project presents a **machine learning-based predictive framework** to assess the risk of credit card default using behavioral, demographic, and transactional data. The model analyzes **repayment history**, **bill and payment amounts**, and **financial attributes** to generate an early warning system for high-risk customers. By combining **data-driven insights** with **behavioral analysis**, this project helps institutions make informed lending decisions, reduce losses, and manage credit portfolios more effectively.

Objectives:

The key objectives of this project are:

1. **To predict the likelihood of a customer defaulting on their credit card payment in the next month.**
2. **To identify key behavioral and financial factors influencing credit risk.**
3. **To assist financial institutions in making data-driven lending and risk management decisions.**

📁 Dataset Overview

The dataset contains **30,000 customer records** with **25 attributes**, representing the demographic, financial, and payment behavior of credit card clients in Taiwan. Each record corresponds to one customer, and the **target variable** is:
Will_Default_Next_Month → 1 if the customer defaulted, 0 otherwise.

Category	Description

Category	Description
Demographic Attributes	Gender_Code, Education_Level, Marital_Status, Age_Years
Credit Attributes	Credit_Limit (assigned credit limit in NT dollars)
Repayment Status	Repay_Sep to Repay_Apr (payment delay history over 6 months, where -1 = paid on time, 1 = 1-month delay, etc.)
Billing Amounts	BillAmt_Sep to BillAmt_Apr (total monthly bill amount)
Payment Amounts	PaidAmt_Sep to PaidAmt_Apr (actual monthly repayment amount)
Target Variable	Will_Default_Next_Month (1 = Default, 0 = No Default)

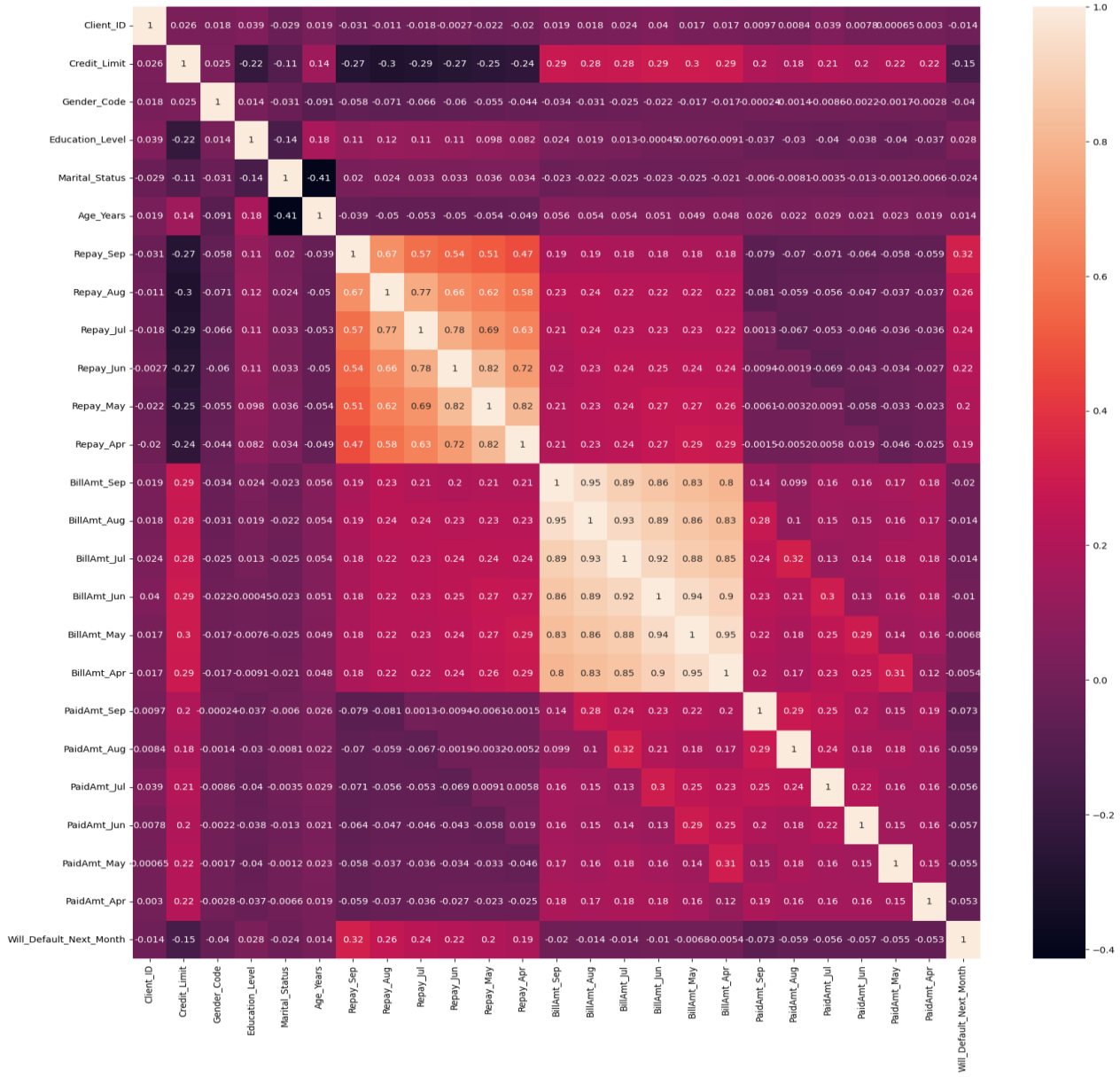
```
 1   Credit_Limit              30000 non-null  int64
 2   Gender_Code               30000 non-null  int64
 3   Education_Level           30000 non-null  int64
 4   Marital_Status            30000 non-null  int64
 5   Age_Years                 30000 non-null  int64
 6   Repay_Sep                 30000 non-null  int64
 7   Repay_Aug                 30000 non-null  int64
 8   Repay_Jul                 30000 non-null  int64
 9   Repay_Jun                 30000 non-null  int64
 10  Repay_May                 30000 non-null  int64
 11  Repay_Apr                 30000 non-null  int64
 12  BillAmt_Sep               30000 non-null  int64
 13  BillAmt_Aug               30000 non-null  int64
 14  BillAmt_Jul               30000 non-null  int64
 15  BillAmt_Jun               30000 non-null  int64
 16  BillAmt_May               30000 non-null  int64
 17  BillAmt_Apr               30000 non-null  int64
 18  PaidAmt_Sep               30000 non-null  int64
 19  PaidAmt_Aug               30000 non-null  int64
 20  PaidAmt_Jul               30000 non-null  int64
 21  PaidAmt_Jun               30000 non-null  int64
 22  PaidAmt_May               30000 non-null  int64
 23  PaidAmt_Apr               30000 non-null  int64
 24  Will_Default_Next_Month   30000 non-null  int64
```

✅ **No missing values were found**, and all columns are numerical, which simplifies preprocessing and modeling.

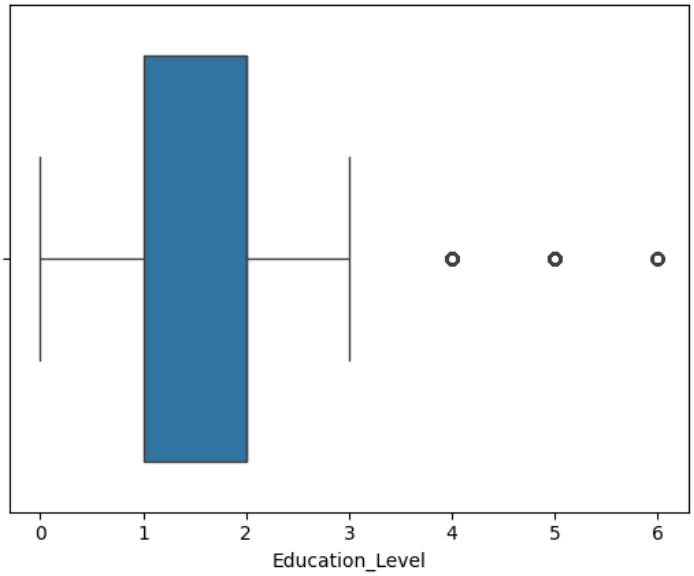
📊 Exploratory Data Analysis (EDA)

🔹 Correlation Analysis

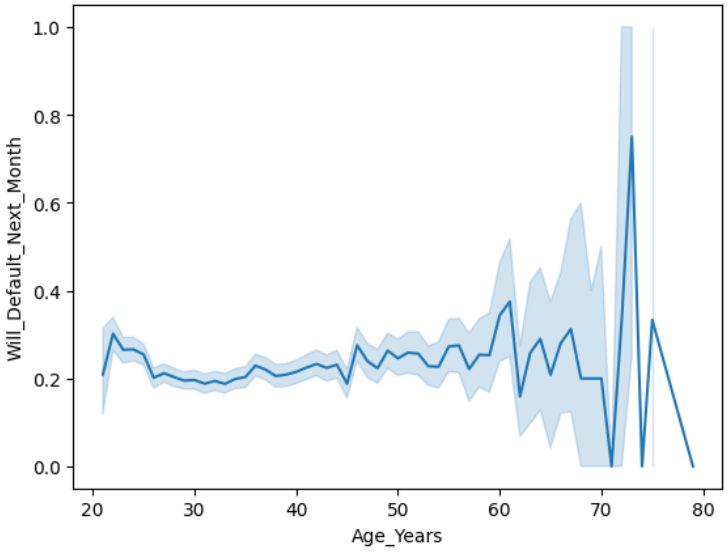
- Strong positive correlations between consecutive **Repay** and **Bill Amount** features indicate consistent financial behavior.
- **Repay_Sep** and **Repay_Aug** show moderate positive correlation with Will_Default_Next_Month (~0.32), confirming that **payment delays directly increase default risk**.
- **Credit_Limit** is negatively correlated (-0.15) with the target, implying **higher credit limits reduce the likelihood of default**.



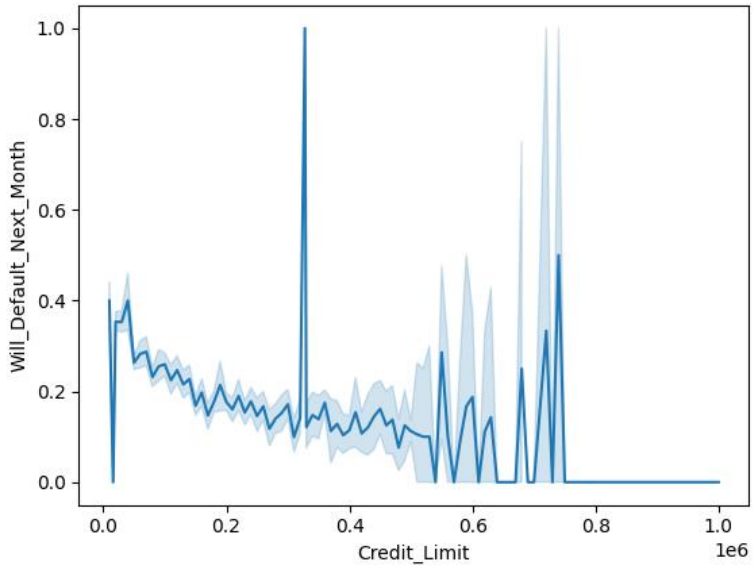
- **Education Level Distribution**
  - Majority of customers belong to **Graduate or University** levels (1–2).
  - Few outliers exist in levels 4–6 (others/unknown), but they don't significantly affect the model.



- **Age vs Default Rate**
  - Default rate is relatively stable (20–30%) between **ages 25–55**.
  - Slight spike observed in customers above **60 years**, suggesting **higher financial instability** in older clients.
  - Younger customers (<25) show volatile repayment patterns due to limited credit experience.



- **Credit Limit vs Default Rate**
  - Default probability **decreases as credit limit increases**.
  - Clients with **lower credit limits (<200k NT$)** are more prone to defaults.
  - Indicates that **financially sound customers with high credit trust** exhibit stronger repayment consistency.



Methodology

The project follows a complete end-to-end ML workflow:

1. Data Preprocessing

- Checked for missing values (none found).
- Applied **feature scaling** using StandardScaler for normalization.
- Split the dataset into **80% training** and **20% testing** subsets for model validation.

2. Feature Engineering

- Analyzed correlations to drop redundant variables if necessary.
- Treated repayment features as behavioral predictors and retained demographic ones for interpretability.

3. Model Development

Implemented and trained **seven supervised machine learning algorithms**:

- Logistic Regression
- Support Vector Machine (SVM)

- K-Nearest Neighbors (KNN)
- Decision Tree
- Random Forest
- Gradient Boosting
- XGBoost

Each model was validated using **cross-validation** to ensure reliability and prevent overfitting.

4. Model Evaluation

Evaluated models using:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**

Model	accuracy	precision	recall	f1_score
Logistic_Regression	0.716167	0.725969	0.716167	0.720759
SVM	0.816167	0.797776	0.816167	0.792174
K_Nearest_Neighbors	0.792667	0.770422	0.792667	0.775416
Decision_Tree	0.716167	0.725969	0.716167	0.720759
Random_Forest	0.815167	0.796565	0.815167	0.795250
Gradient_Boosting	0.818333	0.800658	0.818333	0.796889
XGBoost	0.810000	0.790060	0.810000	0.790542

🏆 **Best Performing Model**

Gradient Boosting Classifier achieved the highest accuracy (**81.83%**) and balanced precision-recall performance.

This indicates its strong capability to generalize and detect both defaulters and non-defaulters accurately.

💡 Insights & Interpretations

1. **Repayment history** is the strongest predictor — late or skipped payments correlate directly with higher default risk.
2. **Higher credit limits** are associated with more disciplined repayment behavior.
3. **Education and marital status** moderately affect default risk — single or less-educated customers tend to default slightly more often.
4. **Age factor** reveals that mid-age customers are the most stable, while very young or older clients are riskier.
5. **Gradient Boosting and SVM** emerged as robust models for financial prediction, outperforming traditional logistic models.

🚀 Conclusion

This project successfully developed a **credit default prediction model** using machine learning techniques.
It enables financial institutions to:

- Identify high-risk clients in advance.
- Take preventive measures (e.g., reduce credit limits or adjust loan terms).
- Optimize lending policies to reduce non-performing assets (NPAs).

The framework demonstrates that **data-driven behavioral modeling** can greatly enhance **risk management** and **decision-making** in modern banking systems.

🧰 Tech Stack

Languages: Python
Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, XGBoost
Algorithms: Logistic Regression, SVM, Random Forest, Gradient Boosting, XGBoost
Evaluation Metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC