Predicting Credit Card Default — A Comparative Classification Analysis

Otajon Yuldashev – 473457

1. Motivation & Problem Definition

Credit risk management is a central problem in modern banking and consumer finance. Financial institutions must decide which clients are likely to default on their credit obligations in order to minimize losses while maintaining competitiveness. Underestimating default risk exposes lenders to significant financial losses, whereas overestimating risk may lead to rejecting creditworthy customers, reducing revenue and customer trust.

The challenge is therefore to balance **risk control and fairness**, ensuring that high-risk clients are identified early without excessively penalizing low-risk individuals. This problem naturally lends itself to a supervised classification framework, where the objective is to predict whether a client will default on their payment in the following month.

2. Technical Challenges

Credit default prediction presents several modeling difficulties:

- **Class Imbalance**: Only around 22% of clients default, making accuracy an unreliable performance measure.
- **Non-linearity**: Default risk does not increase linearly with demographic or financial variables.
- **Behavioral Dependence**: Repayment history variables capture behavioral patterns that compound over time.
- **Asymmetric Costs**: False negatives (missed defaulters) are typically more costly than false positives.

These challenges suggest that linear models may be insufficient and that ensemble methods may offer superior performance.

3. Project Objective

The objective of this project is to **systematically compare three classification models**:

- Logistic Regression (linear baseline),
- Random Forest (bagging-based ensemble),
- Gradient Boosting (boosting-based ensemble),

and identify which approach best captures the non-linear and behavioral nature of credit default risk while maintaining interpretability and robustness.

4. Dataset Description

The analysis uses the **Default of Credit Card Clients Dataset**, containing information on **30,000 credit card clients**.

- **Target variable**:

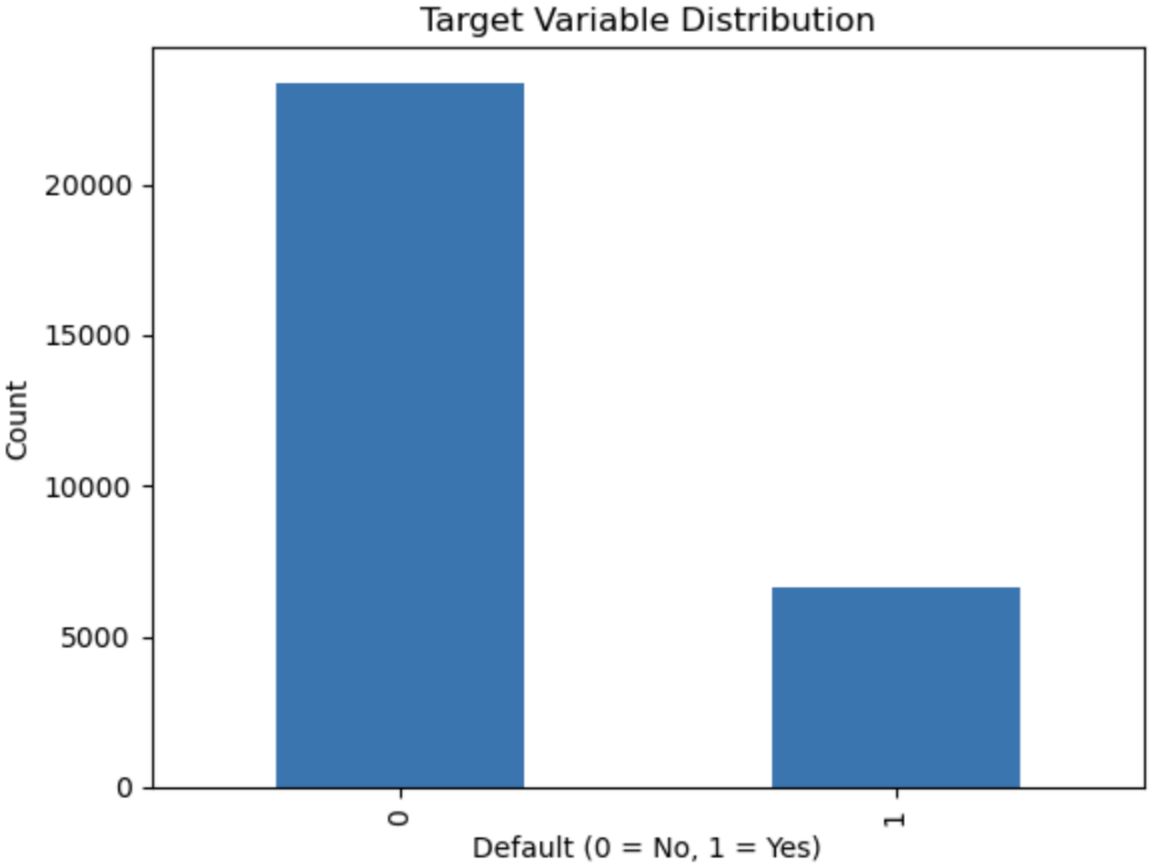
default payment next month (binary: 0 = no default, 1 = default)

- **Feature groups**:
    - **Demographic variables**: age, sex, education, marriage
    - **Credit profile**: credit limit (LIMIT_BAL)
    - **Repayment status variables**: PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6
    - **Billing amounts**: BILL_AMT1–BILL_AMT6
    - **Payment amounts**: PAY_AMT1–PAY_AMT6

No missing values are present in the dataset, and all variables are numeric.

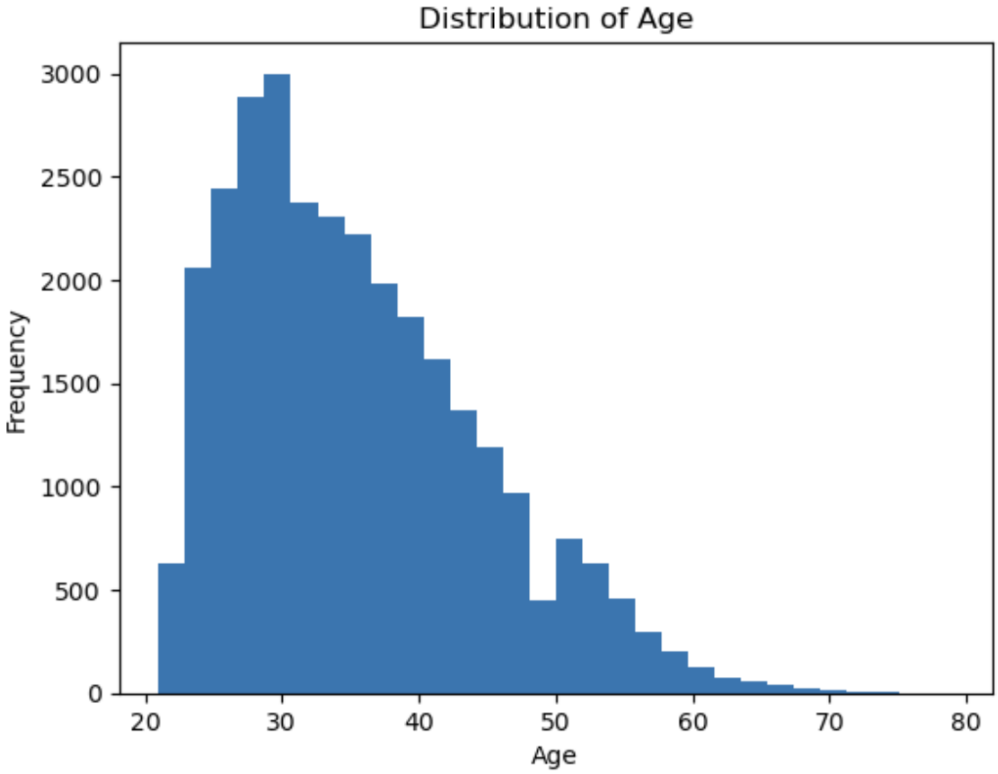
5. Exploratory Data Analysis (EDA)

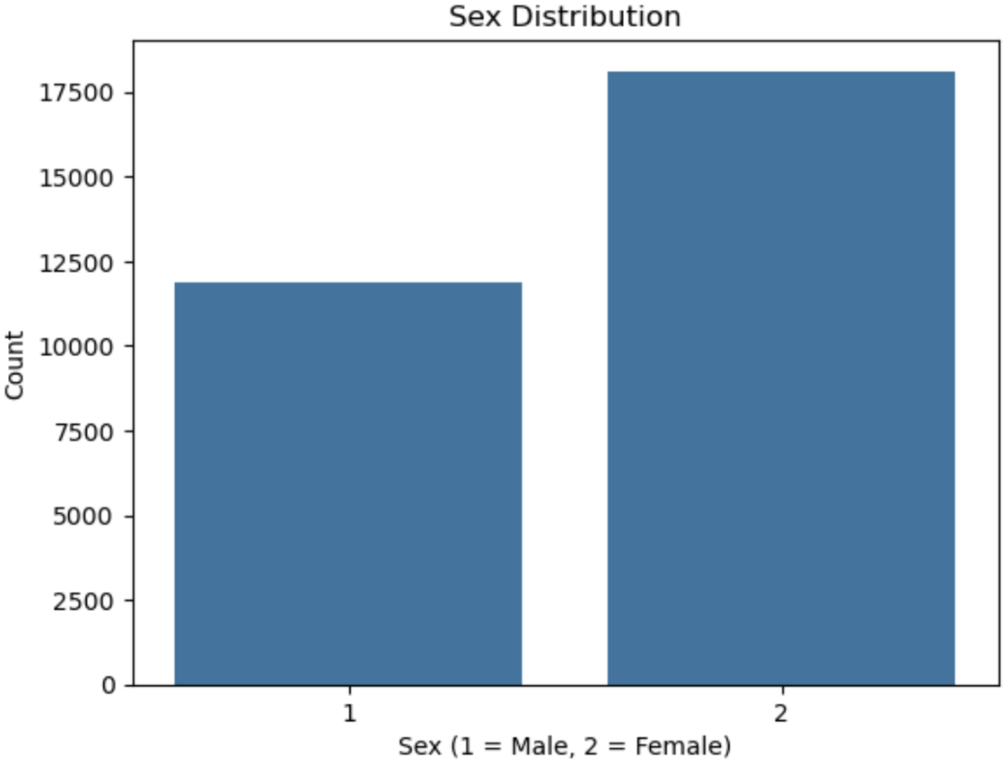
Target Distribution

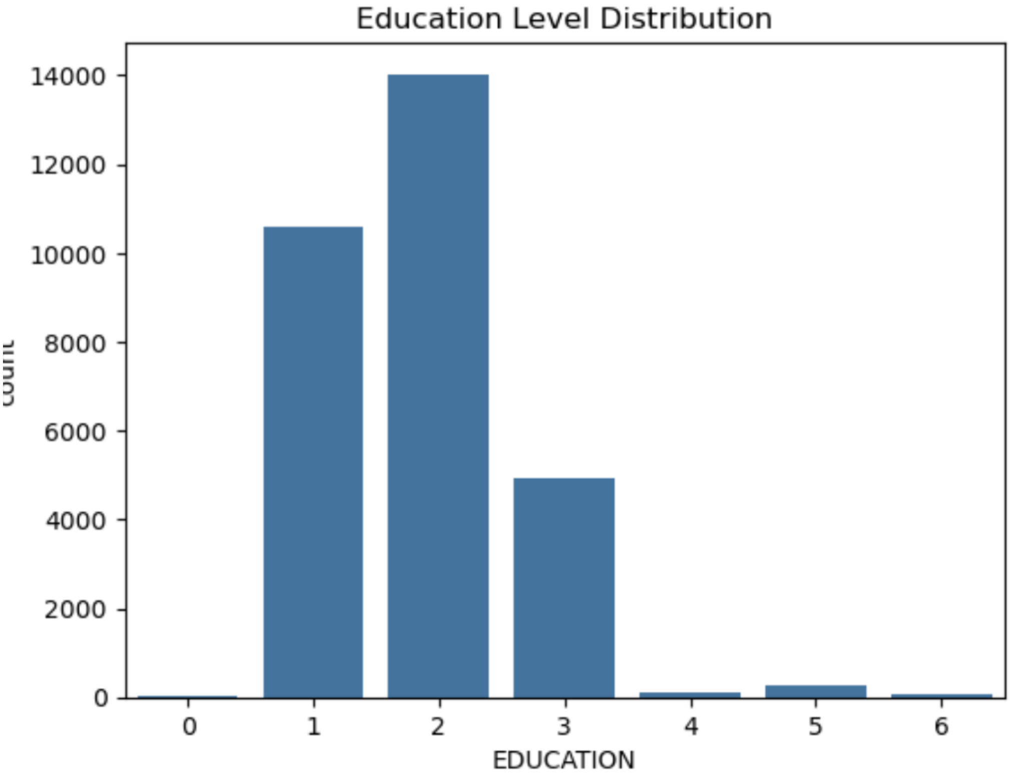


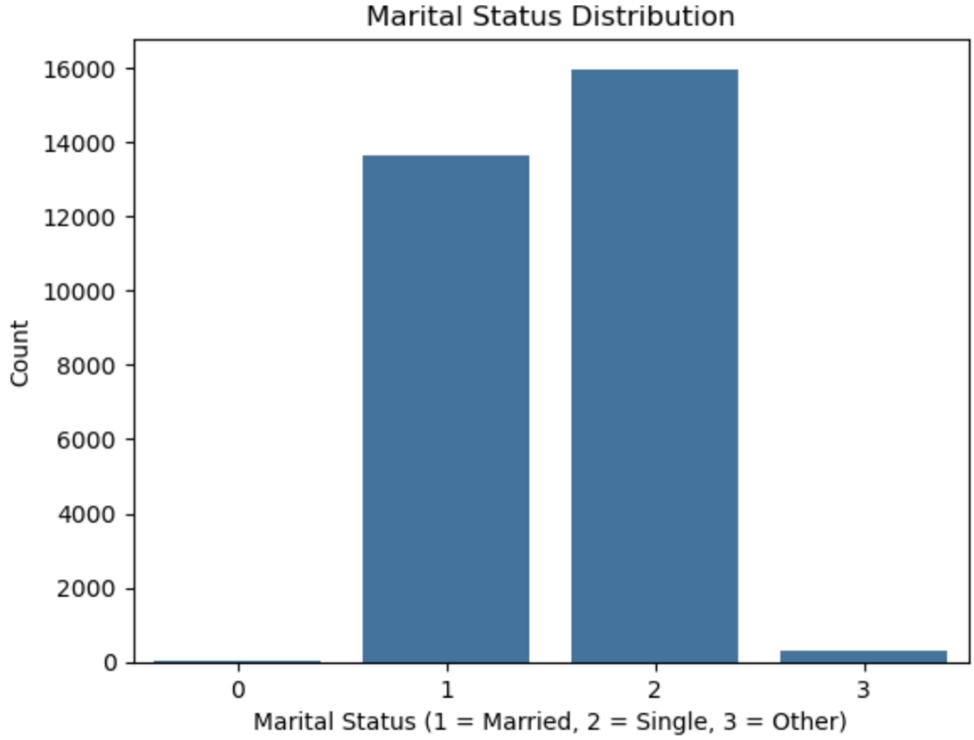
Approximately **77.9% of clients do not default**, while **22.1% default**, confirming a moderate class imbalance. This reinforces the need to evaluate models using recall and F1-score rather than accuracy alone.

Demographic Variables



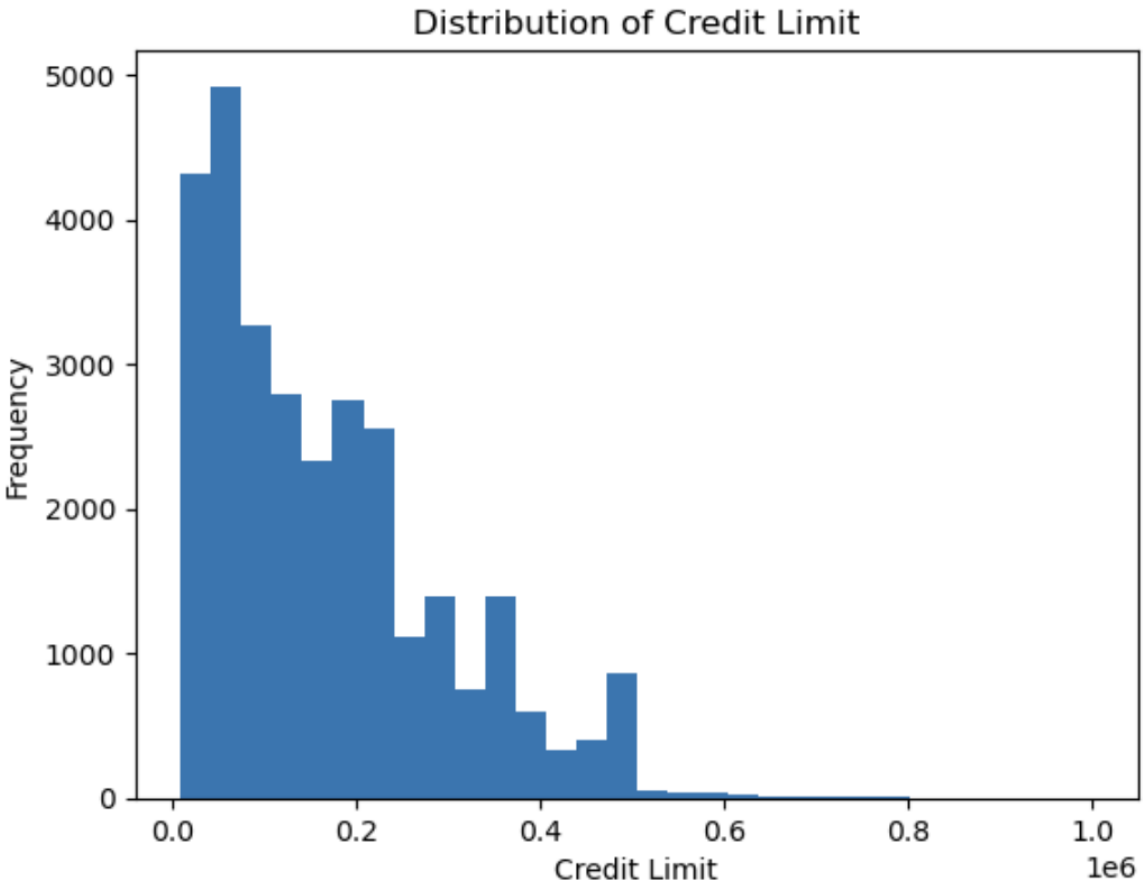






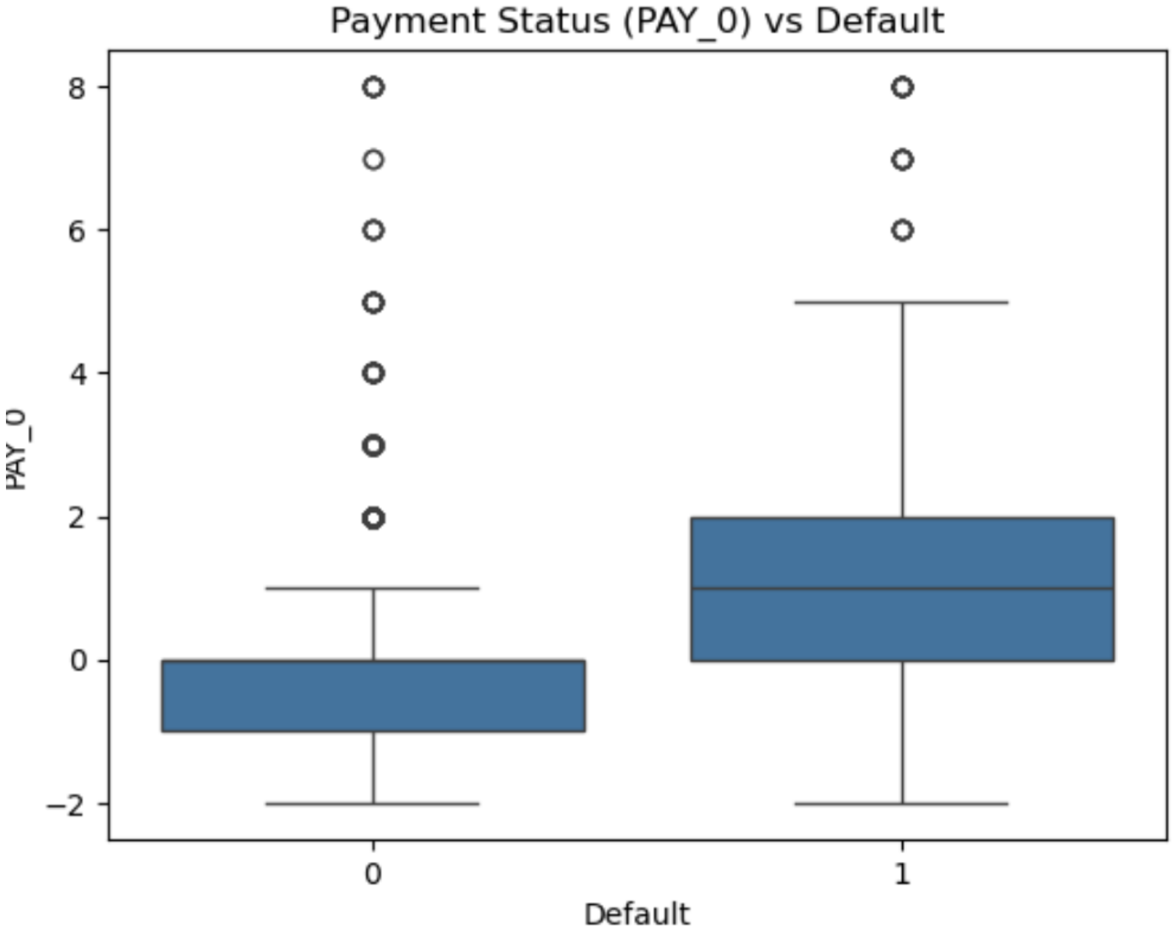
Age is right-skewed, with most clients between 25 and 45 years old. Sex, education, and marital status show uneven distributions but only weak associations with default. These variables provide contextual information but are not strong predictors on their own.

Credit Limit



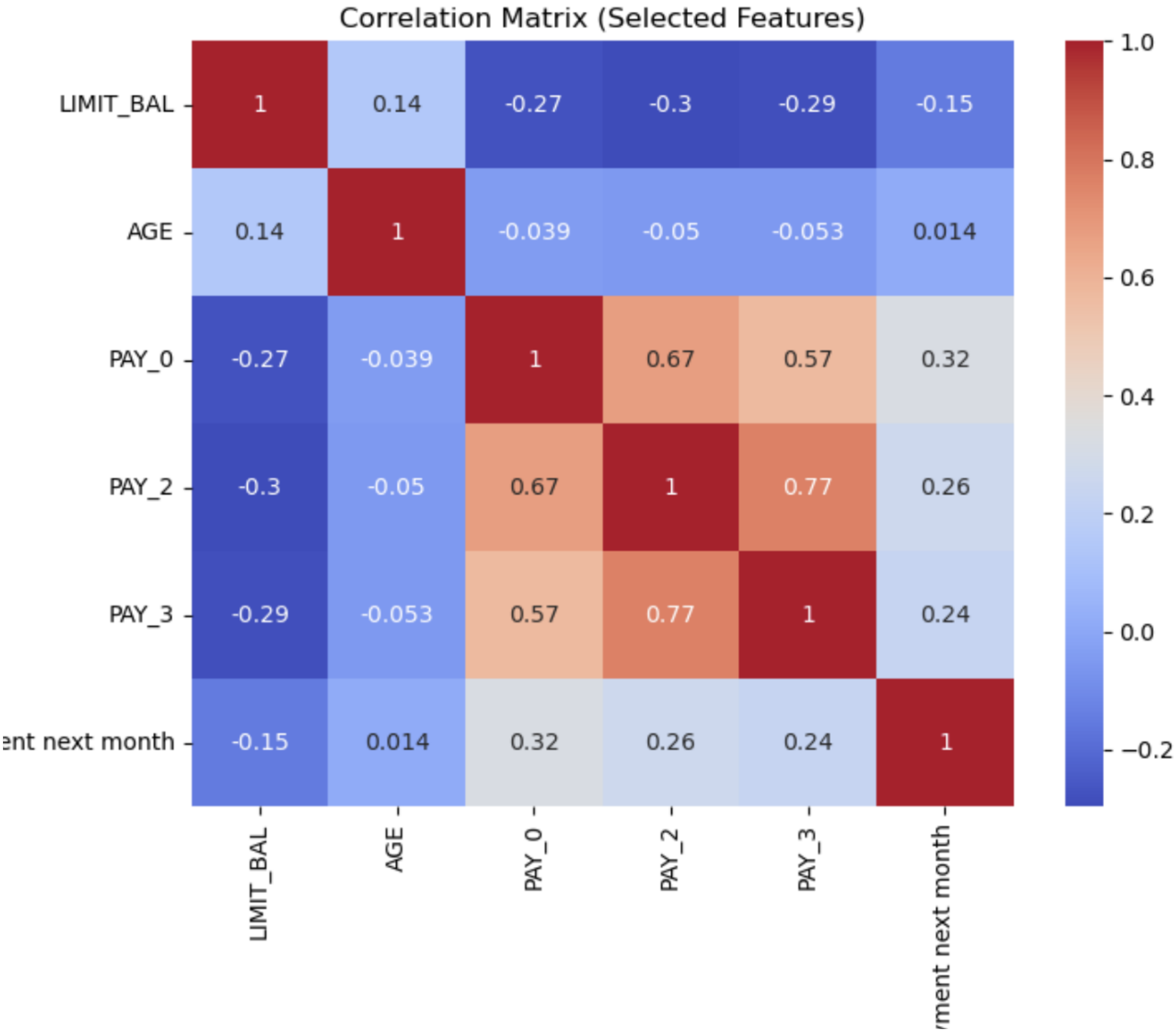
Credit limits are heavily right-skewed. Clients with lower limits tend to default more frequently, suggesting that credit capacity plays a protective role.

Repayment Behavior



Repayment status variables (especially PAY_0) show a strong relationship with default. Clients who default typically exhibit recent payment delays, while non-defaulters generally have on-time or early payments. This confirms repayment history as the most informative predictor of default risk.

Correlation Analysis



Correlation analysis reveals that repayment variables have the strongest association with default, while demographic variables show weak linear relationships. This suggests that non-linear models are better suited for this task.

6. Data Preprocessing

The preprocessing steps were designed to ensure fairness and reproducibility:

- The ID variable was excluded from modeling.
- Data was split into training (80%) and test (20%) sets using stratified sampling.
- Feature scaling was applied only for Logistic Regression via a Pipeline.

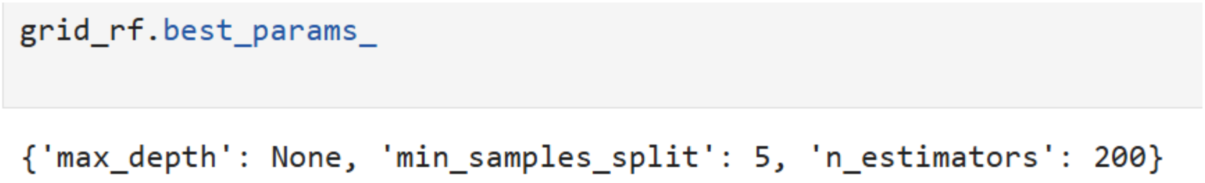
- Tree-based models were trained on raw features, as they are robust to feature scaling.

7. Model Training & Hyperparameter Tuning

Logistic Regression

Used as a linear baseline with standardized features. While interpretable, it is limited in capturing non-linear patterns.

Random Forest

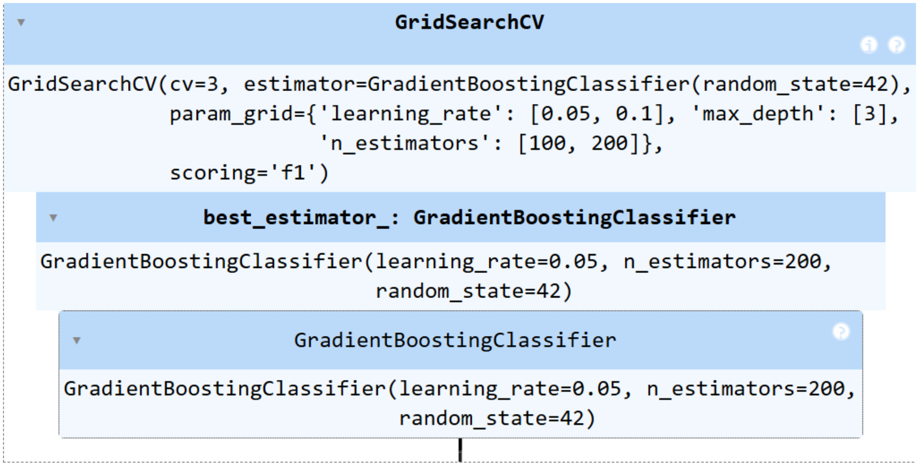


```
grid_rf.best_params_
```

```
{'max_depth': None, 'min_samples_split': 5, 'n_estimators': 200}
```

A bagging-based ensemble model capable of capturing complex interactions. Hyperparameters such as number of trees, depth, and minimum samples per split were tuned using GridSearchCV.

Gradient Boosting



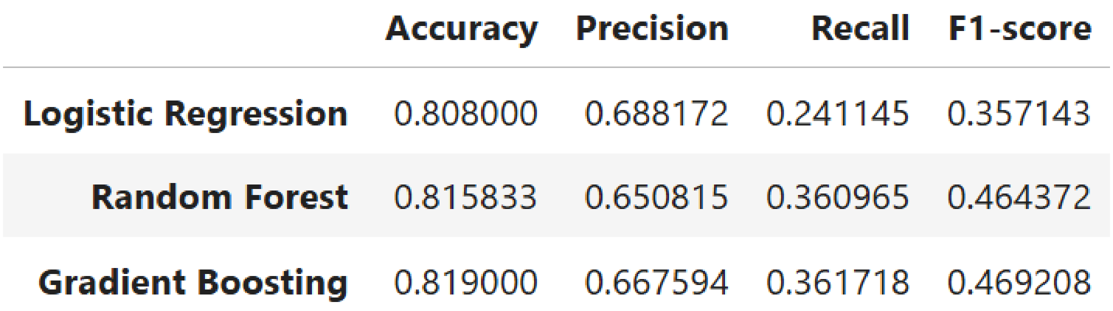
```
GridSearchCV
GridSearchCV(cv=3, estimator=GradientBoostingClassifier(random_state=42),
             param_grid={'learning_rate': [0.05, 0.1], 'max_depth': [3],
                         'n_estimators': [100, 200]},
             scoring='f1')
best_estimator_: GradientBoostingClassifier
GradientBoostingClassifier(learning_rate=0.05, n_estimators=200,
                           random_state=42)
GradientBoostingClassifier
GradientBoostingClassifier(learning_rate=0.05, n_estimators=200,
                           random_state=42)
```

A boosting-based ensemble model that sequentially focuses on difficult observations. Hyperparameters including learning rate, number of estimators, and tree depth were optimized using GridSearchCV.

F1-score was used as the primary tuning metric to balance precision and recall under class imbalance.

8. Model Evaluation

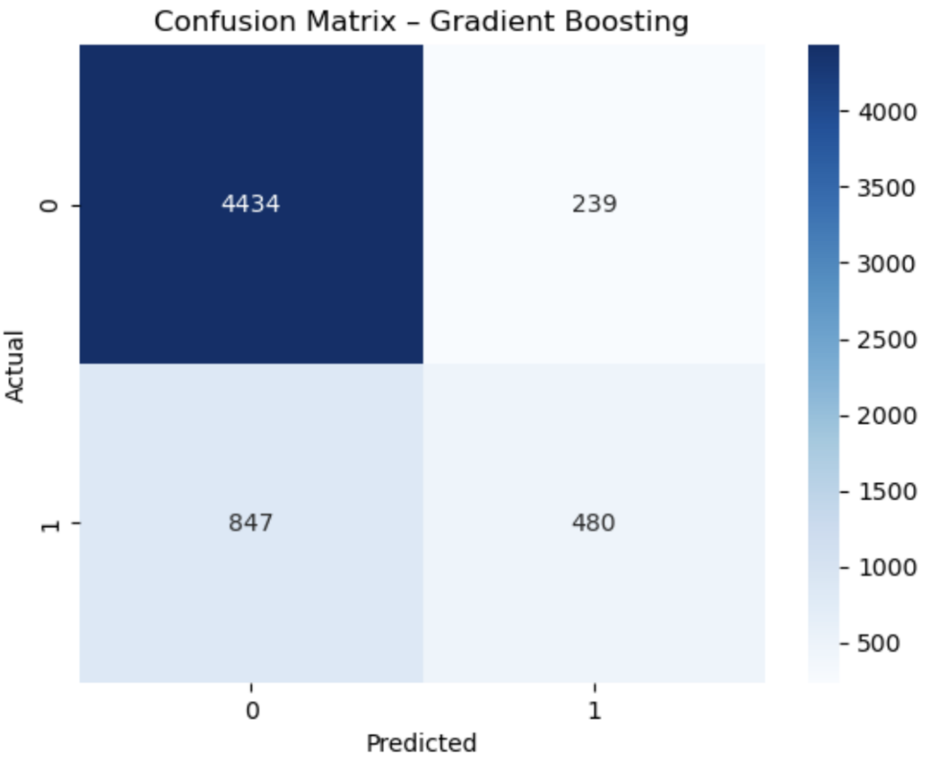
Evaluation was performed on a held-out test set using accuracy, precision, recall, and F1-score.



	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.808000	0.688172	0.241145	0.357143
Random Forest	0.815833	0.650815	0.360965	0.464372
Gradient Boosting	0.819000	0.667594	0.361718	0.469208

- **Logistic Regression** achieved reasonable accuracy but very low recall for defaulters, making it unsuitable for risk management.
- **Random Forest** significantly improved recall and F1-score, demonstrating the value of non-linear modeling.
- **Gradient Boosting** achieved the highest F1-score and provided the best overall balance between identifying defaulters and avoiding false alarms.

9. Error Analysis



The confusion matrix for the Gradient Boosting model shows strong identification of non-defaulters and improved detection of defaulters compared to the baseline. While false negatives remain, their number is reduced relative to Logistic Regression, indicating better risk control.

From a business perspective, this improvement reduces exposure to undetected high-risk clients while maintaining reasonable approval rates.

10. Key Findings

1. Repayment history variables are the dominant drivers of default risk.
2. Demographic variables contribute limited predictive power.
3. Linear models fail to capture complex financial behavior.
4. Ensemble methods significantly outperform the linear baseline.
5. Gradient Boosting provides the best trade-off between precision and recall.

11. Ethical Considerations

Credit default models may reflect historical biases embedded in financial data. False positives can unjustly restrict access to credit, while false negatives increase financial risk. Consequently, such models should be used as **decision-support tools** rather than fully automated decision systems, with appropriate monitoring and transparency.

12. Conclusion

This analysis demonstrates that credit default prediction is inherently non-linear and driven primarily by historical repayment behavior. Ensemble models, particularly Gradient Boosting, outperform linear approaches by effectively capturing interaction effects and behavioral patterns. The selected model offers a robust and interpretable solution suitable for real-world credit risk assessment, provided it is used responsibly.