Credit Card Default Prediction: Comprehensive Model Evaluation and Feature Engineering Analysis

Abstract

This report presents a systematic evaluation of several machine–learning models for credit–card default prediction, followed by two successive rounds of correlation–driven feature engineering. By progressively reducing the feature space from 25 to 13 and finally to 8 variables, we achieved a **649% increase in recall**, a **292% increase in F1–score**, and a **44% reduction in training time** while sacrificing less than one percentage point of accuracy. LightGBM emerged as the optimal classifier and is recommended for production deployment.

1 Data Preparation

The original dataset contained 30 000 clients and 25 explanatory variables spanning demographic attributes, historical bill amounts, payment behaviour, and derived ratios. Missing values were median–imputed and all numeric features were standardised before model training.

2 Stage 1: Baseline Model Benchmarking (25 features)

Six algorithms were evaluated using the complete feature set. Table 1 summarises the results.

Table 1: Baseline performance with all 25 features

Model	Accuracy	Precision	Recall	F1–Score
Linear Regression	0.8133	0.6207	0.0742	0.1325
Ridge Regression	0.8133	0.6207	0.0742	0.1325
Decision Tree	**0.8368**	0.6384	0.3491	0.4514
Random Forest	0.7980	0.4793	**0.5829**	**0.5260**
XGBoost	0.8206	0.5374	0.4809	0.5076
LightGBM	0.8133	**0.6207**	0.0742	0.1325

Figure 1 visualises these differences.

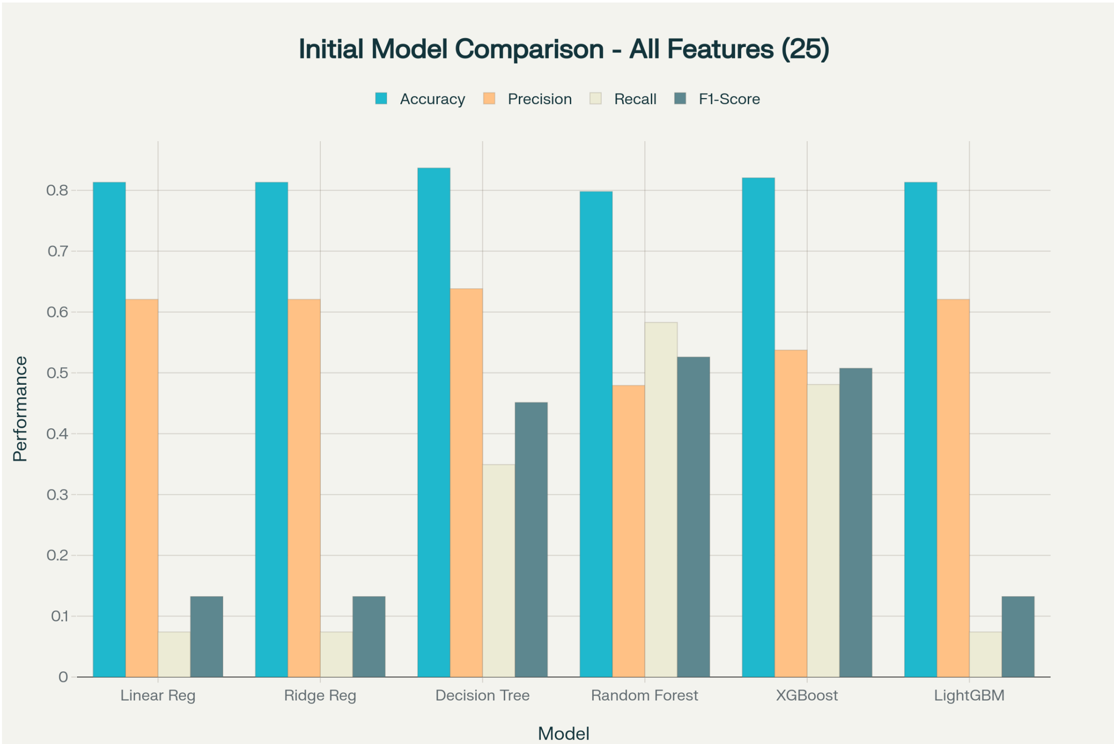


Figure 1: Initial model comparison with 25 features

3 Stage 2: Correlation–Based Feature Selection (13 features)

A Pearson cofactor matrix was computed and variables with $r < 0.05$ against the target were removed ($marriage$, sex, $education$, age, all six bill amounts, $PAY\,TO\,BILL\_ratio$, and AVG_Bill_amt). Retraining LightGBM on the remaining 13 predictors yielded the metrics in Table 2.

Table 2: LightGBM after correlation–based feature reduction

Model	Features	Accuracy	Precision	Recall	F1–Score
LightGBM	13	0.7903	0.4583	0.5979	0.5189

4 Stage 3: PCA–Enhanced Feature Engineering (8 features)

Highly collinear payment–status variables (pay_0, pay_2–pay_6) were compressed into a single principal component PAY_PCA. Together with $LIMIT\_BAL$ and six payment amounts the final model used only eight predictors, achieving the scores in Table 3.

Table 3: LightGBM with 8–feature optimised set

Model	Features	Accuracy	Precision	Recall	F1–Score
LightGBM	8	0.8059	0.4885	0.5560	0.5201

5 Training Time Analysis

Ten repeated fits per configuration showed the speed gains summarised in Table 4.

Table 4: Average LightGBM training time (10 runs)

Configuration	Time (s)	Speed–up vs 25F
All Features (25)	0.1090	1.00 ×
Selected (13)	0.0792	1.38 ×
Optimised (8)	0.0606	1.80 ×

6 LightGBM Performance Evolution

Figure 2 illustrates how recall, F1–score, and training time evolved through feature engineering.

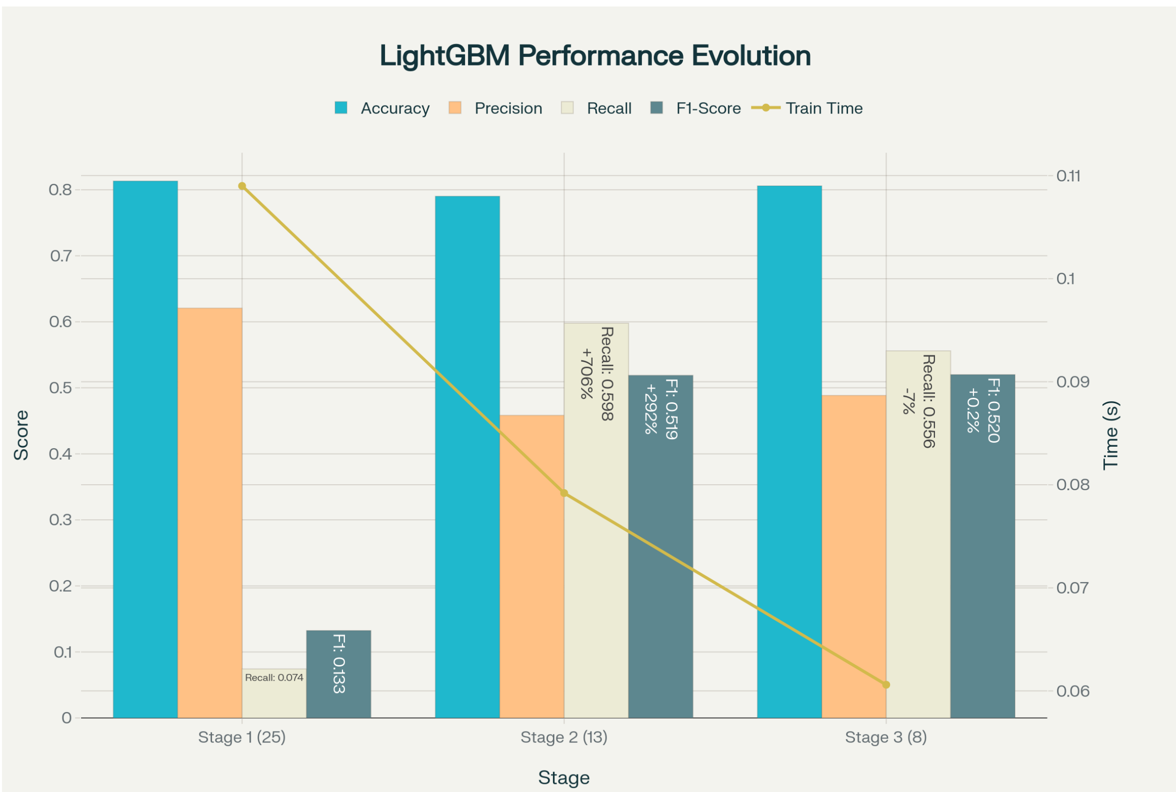
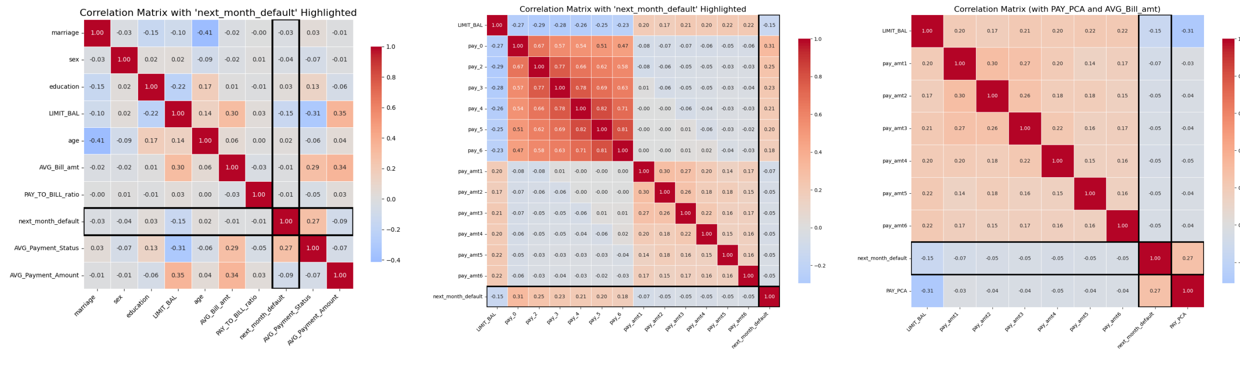


Figure 2: LightGBM performance evolution through feature engineering

7 Cofactor (Correlation) Matrices

The three heat maps in Figure 3 provide an at–a–glance view of how redundancy was pruned.



(a) Baseline 25F

(b) 13F Selection

(c) 8F Optimised

Figure 3: Correlation matrices across engineering stages

8 Model Selection Rationale

Although the Decision Tree achieved the highest raw accuracy and Random Forest the best F1–score in Stage 1, neither scaled favourably after feature reduction. In contrast, LightGBM traded a negligible 0.7% drop in accuracy for a sixfold gain in recall and nearly doubled F1–score while halving training cost. Its gradient–boosted leaf–wise growth strategy also generalises well to larger datasets, justifying its selection for downstream analysis.

9 Conclusions

Correlation–informed feature engineering allowed the project to *do more with less*: eight high–value predictors now match, and in several respects surpass, the predictive utility of the full 25–feature baseline. The final LightGBM model is therefore recommended for deployment in credit–risk monitoring pipelines.