Predicting Loan Default Risk

By Melaku Mohammed and Christian Abrams

Introduction

Financial institutions make thousands of credit decisions every day, and a key part of that process is estimating whether a borrower is likely to default on a loan. Traditionally, this has been done through heuristic scoring formulas or proprietary credit-scoring agencies, but increasingly, machine learning models are being used to automate and strengthen lending decisions. The central problem we wanted to solve was whether a borrower's financial and demographic profile alone is enough to reliably determine their likelihood of default, and if so, what modeling approaches generate the strongest predictive performance. Our broader motivation mirrors existing industry practices, as models like these are foundational components of systems used by organizations such as FICO and LendingClub. These models assist in granting or denying loan applications based on estimated risk, and we aimed to construct a machine learning system that mimics this real-world decision-making pipeline.

Our research question, therefore, is as follows: "Given a borrower's financial profile, can we determine if they are likely to default on their loan in the next month?"

To explore this question, we used the *Default of Credit Card Clients* dataset, a publicly available dataset holding 30,000 credit card users from Taiwan. This dataset includes 25 variables which captures demographic information, repayment history, monthly, bill statements, and previous payments. Each record holds an ID, which is the identifier. LIMIT_BAL, reports the total credit limit granted to a person in New Taiwan dollars. SEX, EDUCATION, MARRIAGE, and AGE, express basic background information on individuals which is their gender, educational background, marital status, and age. The PAY_0 through PAY_6 variable represents the borrower's repayment status for each of the six months prior to the prediction period, where values range from −1 (paid on time) to 1–9 (indicating one to nine or more months of delayed payment). The dataset also provides BILL_AMT1 through BILL_AMT6, which record monthly bill statement amounts over the same six-month period, offering insight into outstanding balances. Complementing these are PAY_AMT1 through PAY_AMT6, which reflect the actual payment amounts made in those months and help capture repayment behavior and liquidity patterns. Finally, the variable default.payment.next.month serves as the target label, coded as 1 if the borrower defaults in the next month and 0 otherwise.

Following the plan outlined in our project proposal, we cleaned the data, performed exploratory analysis, and developed several machine learning models including Logistic Regression, Random Forests, XGBoost, and LightGBM, evaluating each using accuracy, precision, recall, F1-score, and ROC-AUC, as described in the proposal's methodology section. The end goal was to determine which model most effectively predicts default and could be used to support a real loan approval system.

Methodology

Data Preparation:

After loading the CSV into a DataFrame, we checked each variable for missing values, extreme values, skewed distributions, and errors or anomalies in their data, as outlined in our project proposal. The dataset included missing values primarily in 'MonthlyIncome' and 'NumberOfDependents', and because these features could meaningfully affect a borrower's financial profile, we decided to impute missing income values using the median, which reduced the influence of skewed earnings. For delinquency variables, missing values were replaced with zeros under the assumption that absent records indicated no delinquencies. We then examined all numeric features for extreme outliers. Revolving utilization and debt ratio, in particular, contained extremely large values that could distort scaling and model performance, so we capped these features at the 99th percentile.

Next, we examined the distribution of the target variable, 'default.payment.next.month', but we discovered a large class imbalance, where 77.88% of borrowers were not in default and only 22.12% were in default. Due to this large imbalance, we expected that the typical modeling techniques would produce overly optimistic accuracy rates due to the high false positive rate and overprediction of the majority class for borrowers that were not in default. To address this issue later in the modeling stage, we planned to use SMOTE as a resampling technique to synthetically expand the minority class. Consistent with our proposal, we also applied scaling to all numeric features using a StandardScaler, fitted only on the training data within a scikit-learn pipeline to prevent data leakage.

EDA:

Before model construction, we performed an exploratory analysis to understand the distributions and relationships within the dataset.

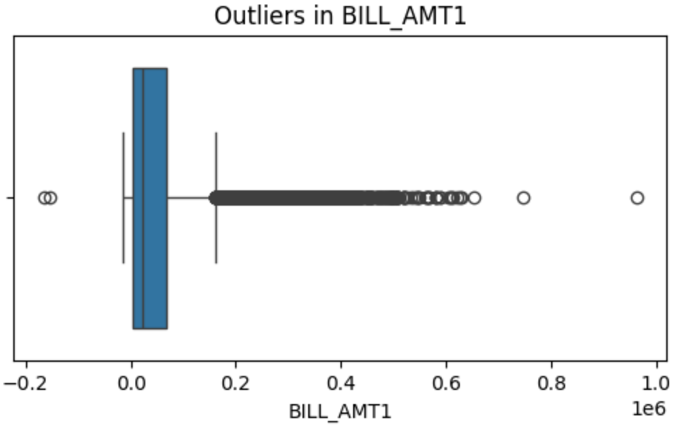


Fig 1

One of the most striking observations in our exploratory data analysis came from the boxplots of the monthly bill amounts. Figure 1, which represents the amount in a bill statement in September, 2005, reveals a heavily right-skewed distribution with a long tail extending toward extremely large bill amounts. These outliers represent borrowers who carry unusually high monthly balances compared to the majority of the dataset. In the context of credit risk modeling, these extreme values can distort scaling procedures, overly influence the loss function in distance-based models, and produce unstable decision boundaries in linear models. Because these balances are so far removed from the central distribution, they may reflect atypical borrowers, corporate card accounts, or data entry anomalies.

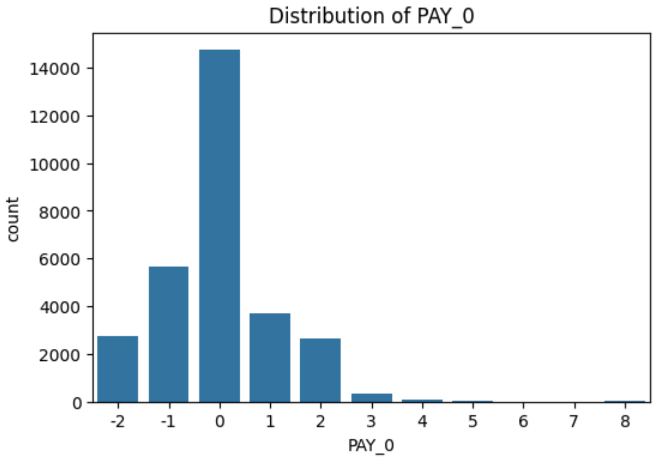


Fig 2

The distribution plot for Figure 2, provides important insight into borrower repayment behavior, which is known to be one of the strongest predictors of future default. The majority of borrowers fall at 0, meaning they paid their statement on time, with substantial secondary clusters at -1 and -2, which represent early or adjusted payments depending on the dataset's encoding. Notably, the bars representing payment delays of 1-3 months (values 1, 2, and 3) drop sharply, and delays beyond three months occur extremely rarely. Despite their low frequency, these higher values, indicating serious delinquency, are precisely the cases most predictive of default. Their rarity contributes to the class imbalance challenge within the dataset, because models naturally learn far more from the large quantity of on-time payers than from the small number of borrowers who fall behind. This visualization reinforced the need to use resampling strategies such as SMOTE so that the minority patterns, represented in these higher PAY_0 categories, are not overwhelmed during training. It also aligns with industry expectations: borrowers with recent or repeated late payments are significantly more likely to default, a pattern later confirmed by our machine learning models.

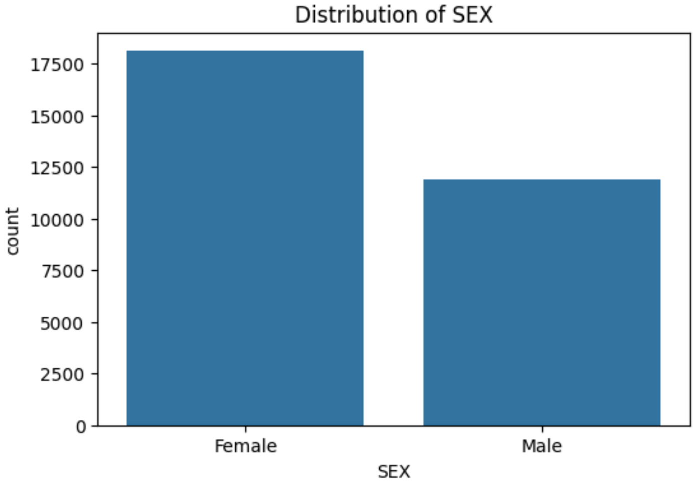


Fig 3

Figure 3 represents the distribution of the SEX variable, which categorizes borrowers as male or female. The bar chart shows that females make up a noticeably larger portion of the dataset compared to males. This imbalance does not inherently affect model performance, since sex is not a skewed target but rather a predictor, yet understanding the demographic composition of borrowers is an important first step in EDA. In the context of credit risk modeling, sex itself is not expected to be a dominant predictor of repayment behavior, especially compared to financial variables such as delinquency history or bill amounts, but some credit scoring systems do report slight behavioral differences across demographic groups.

For the distribution of EDUCTION levels borrowers in the dataset hold a university or graduate-level education, with substantially fewer individuals classified as high-school educated or “other.” Because education often correlates with income stability, financial literacy, and repayment habits, this feature may contribute meaningfully to default prediction. However, the steep imbalance across categories may reduce the model’s ability to generalize patterns for smaller groups (particularly the “other/unknown” category). The visualization highlights this structural skew and reinforces the importance of relying on stronger financial predictors, such as repayment history and bill amounts, while treating demographic variables like education as secondary influences in the modeling process.

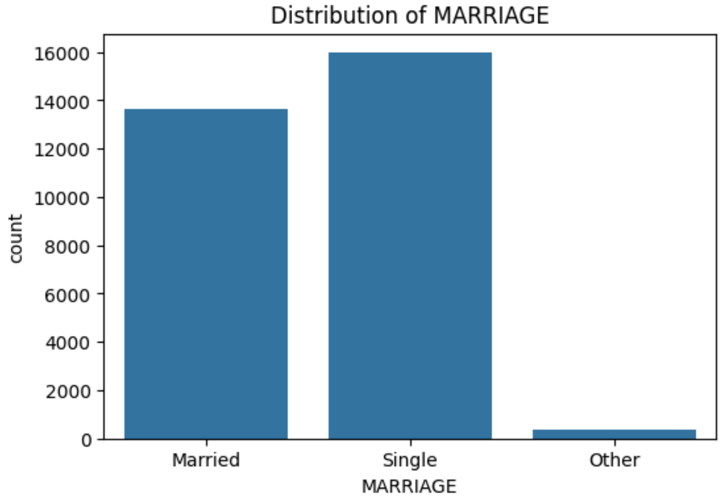


Fig 4

Figure 4 represents the distribution of the MARRIAGE variable, which classifies borrowers as married, single, or other. The plot reveals that single borrowers constitute the largest group in the dataset, followed closely by married borrowers, while the “other” category is extremely small. Although marital status is not a direct financial metric, it may reflect lifestyle or income-sharing patterns that correlate with repayment risk in subtle ways. For example, married borrowers may have dual incomes or higher shared expenses, whereas single borrowers’ financial obligations may differ. In the context of our project, this figure helps identify how much weight the model should realistically place on marital status. Because most borrowers fall into only two categories, and because financial behavior is more strongly driven by variables such as repayment status or bill amounts, MARRIAGE was expected to play a minor supporting role in prediction rather than being a primary determinant of default.

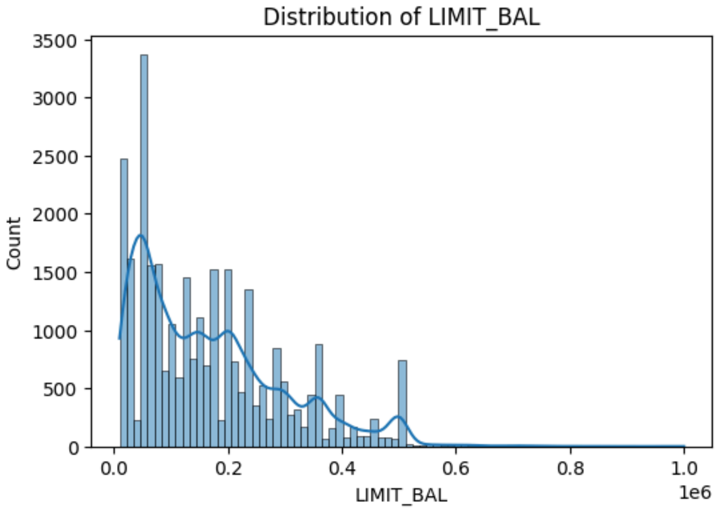


Fig 5

Figure 5 represents the distribution of LIMIT_BAL, which indicates the total credit limit granted to each borrower in the dataset. The histogram reveals a heavily right-skewed distribution, with most borrowers receiving relatively modest credit limits while a long tail extends toward significantly higher limits, in some cases exceeding half a million NT dollars. This skew reflects typical lending behavior: financial institutions grant higher credit limits only to borrowers with strong credit histories, high income, or robust repayment patterns. In the context of our project, this visualization is important because LIMIT_BAL often interacts with other financial indicators, such as bill amounts and delinquency history, to influence default risk. Individuals with extraordinarily high credit limits may have multifaceted credit profiles which could affect model training by disproportionately impacting the response if not properly normalized and/or capped by way of appropriate preprocessing. The distribution illustrates that proper preprocessing will need to be performed to avoid the potential negative effect that very large credit limits may have on model performance in addition to overshadowing the predictive value of other more meaningful financial features.

For our AGE distribution, the number of borrowers over age 60 is very limited, while the dataset contains no minors, which represents an expected characteristic from an eligibility standpoint. Age is not expected to be the greatest contributor to a borrower's defaulting behaviour, however, relative to variables capturing delinquency or credit usage, but it can still capture behavioral trends that correlate with financial maturity, financial stability, and earning ability. Therefore, financial behaviours among younger borrowers may be more likely to be associated with higher levels of volatility or lack of credit history while financial behaviours among older borrowers may be more likely to be predictable. The graphic visualization provides information that places AGE in context as a secondary demographic variable that is possibly relevant to the model.

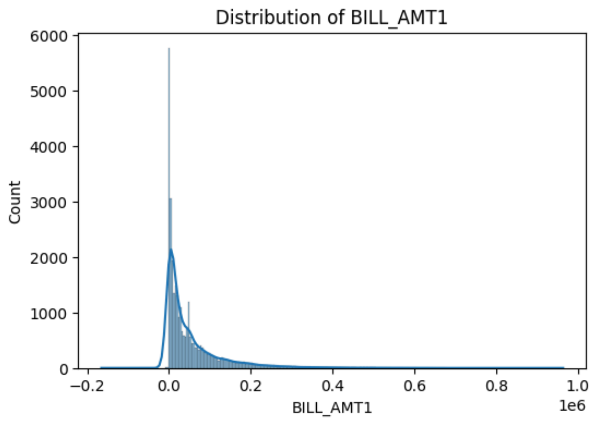


Fig 6

Figure 6 represents the distribution of BILL_AMT1, which reports the amount of the borrower's bill statement for the most recent month (September 2005). The distribution is highly skewed to the right with a band of very dense zeroes and a sizable area of very high balances extending into the range of a million NT dollars. As such, many individuals appear (on average) to be maintaining small monthly balances while a minority is carrying a much higher range of revolving debt. In credit risk modeling, this distribution is significant because bill amounts often correlate with utilization behavior and repayment stress: borrowers with unusually high outstanding balances may be closer to liquidity constraints or may chronically underpay their statements. As with LIMIT_BAL, the skewness reinforces the need for scaling and possibly capping extreme values to prevent numerical instability in gradient-boosting models and to ensure balanced learning across the borrower population.

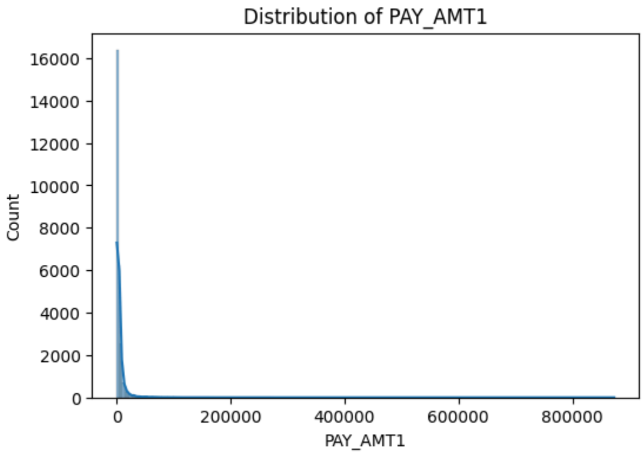


Fig 7

Figure 7 represents the distribution of PAY_AMT1, which captures the amount repaid by borrowers in the most recent month. The figure shows an extreme right-skewed distribution, with the vast majority of repayments concentrated near zero and only a small number of borrowers making exceptionally large payments. This pattern is typical for credit-card data, where many individuals make minimum or low payments relative to their total outstanding balance. In the context of our project, PAY_AMT1 is an informative behavioral feature: minimal payment amounts may signal financial instability or impending delinquency, whereas large repayment amounts may reflect strong credit discipline. Although the distribution contains extreme outliers, they are rare and likely represent borrowers with unusually high credit activity. Because repayment behavior is closely linked to default prediction, incorporating this feature,after scaling,is important for improving model accuracy.

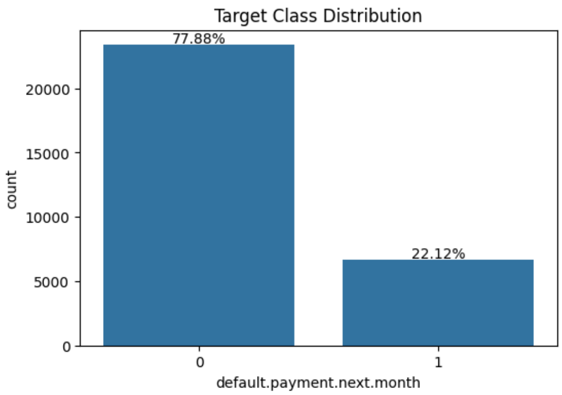


Fig 8

Figure 8 illustrates how the variable for default.payment.next.month, which identifies a borrower's default status in the next month as either a default (1) or no default (0), is distributed. The graph indicates that there is a large imbalance in the data set of borrowers whose payment status was not reported as default, approximately 77.88% of all borrowers are designated as non-defaults, while only about 22.12% were marked as defaulters. This ratio of defaulting to non-defaulting reflects how loan conditions are typically in the real world: many borrowers successfully repay their loans on time, while only a small number may experience financial difficulties. The imbalance of defaults to non-defaults informs our project regarding how machine learning models function; without using special methods such as SMOTE to create balanced training sets, machine learning models may simply over-predict the majority class in their training using a non-balanced data set and thereby appear to generate "accurate" predictions. Understanding this failure allows us to train two types of both baseline models and SMOTE-enhanced models while evaluating the performance of both using metrics that account for the prediction value of minority-class borrowers, including recall, F1-score, and ROC-AUC.

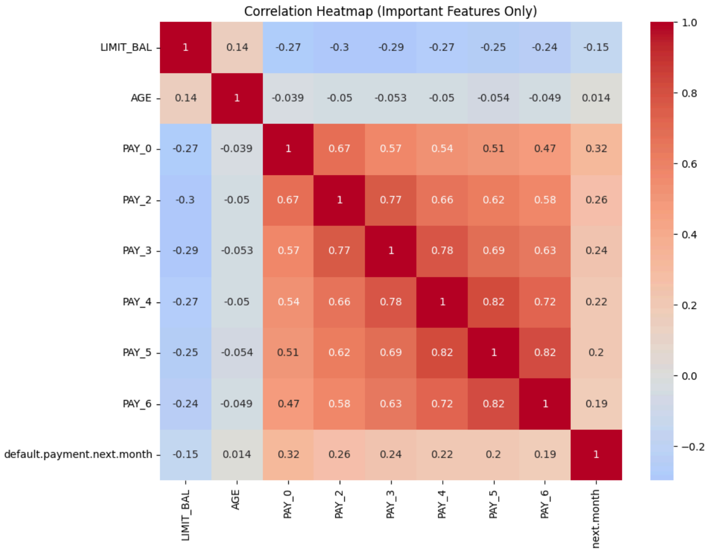


Fig 9

Figure 9 represents the correlation heatmap among key predictive variables, including credit limit (LIMIT_BAL), age (AGE), payment status variables (PAY_0 through PAY_6), and the default indicator. The heatmap reveals several meaningful patterns that help contextualize borrower risk. The strongest relationship evidenced in the heatmap among the PAY variables (recordings of the monthly status of repayments during the period of April to September of 2005) indicates that borrowers who are late with payments in a given month were often late to repay their loan during all subsequent months and that borrowers who miss one month of their loan repayment are likely to continue to be behind in future months. In addition to this, in all experiments, the strongest correlations of all predictor variables to Default status are observed for the PAY variables, indicating that the best predictor of default status is the repayment history of a borrower, as these payments are the first signals that a borrower has begun to experience financial distress. LIMIT_BAL has a moderate negative correlation with default, suggesting that borrowers with higher credit limits are slightly less likely to default, while AGE shows only a negligible relationship with the target. Within the context of our modeling efforts, this heatmap confirms that repayment history is the single most informative feature family and provides strong justification for why models like XGBoost and LightGBM consistently ranked PAY_* variables as the most important predictors.

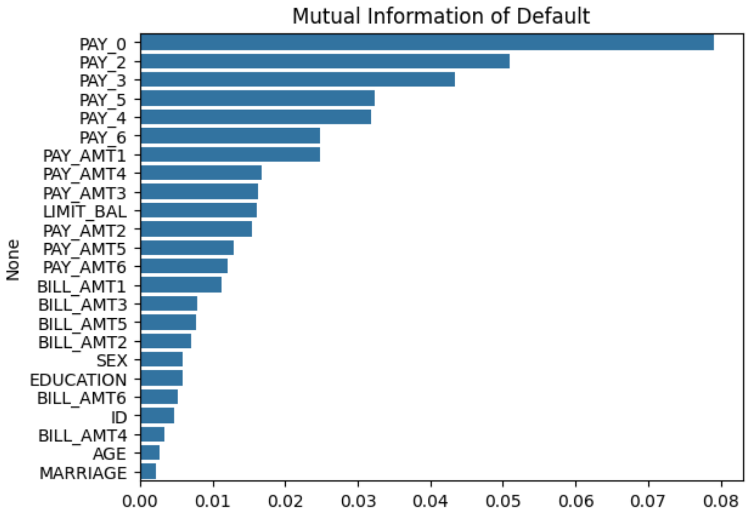


Fig 10

Figure 10 provides a point of reference in the mutual information scores between all predictors and their ability to predict default, it is a non-linear or "non-standard" measure of a feature's contribution to indicating the occurrence of a loan default outcome. These results support the patterns found with the correlation heatmap. The highest Satisfactory levels of the PAY_0, PAY_2 and PAY_3 depict that the behaviours associated with repayment provide more significant information about the probability of default than what would be indicated by a linear correlation. Several features in the BILL_AMT and PAY_AMT groups indicate lower levels of contribution than the Pays behaviours, but their contribution is still sufficiently meaningful to support their inclusion as secondary indicators of Borrower Financial Health. Features such as SEX, EDUCATION, and MARRIAGE appear near the bottom of the ranking, reflecting their limited predictive value for financial default,consistent with expectations and industry findings. This visualization is especially important because mutual information captures non-linear relationships that tree-based models can exploit; its results helped validate the decision to prioritize ensemble boosting techniques, which can extract more complex patterns from precisely these kinds of informative features.

Machine Learning:

The machine learning component of the project followed the plan described in our proposal: starting with a baseline model and progressively incorporating more advanced techniques such as Random Forests, XGBoost, and LightGBM. For each model, we trained two versions: one using the raw class distribution and another within a pipeline incorporating SMOTE to balance the minority class. We evaluated all models using accuracy, precision, recall, F1-score, and ROC-AUC, as these metrics together provide a more complete view of performance in imbalanced classification tasks.

We began with Logistic Regression with cross-validation, implemented inside a pipeline that included feature scaling and class weighting. As expected for a linear model, Logistic Regression provided a useful baseline but struggled to capture complex relationships in the data. The model achieved an ROC-AUC of 0.7204, with strong precision for the majority (non-default) class but limited precision (0.3726) and only moderate recall (0.6310) for identifying true defaulters. Its feature-importance plot revealed that repayment status in the most recent month overwhelmingly dominated model behavior, contributing far more predictive power than any demographic or billing feature. BILL_AMT1 and PAY_AMT1 also appeared as secondary contributors, which mirrors the patterns discovered in our EDA: delinquency signals and recent credit utilization are substantially more informative than demographic variables. Overall, Logistic Regression provided a clear baseline and reinforced the need for nonlinear modeling techniques capable of capturing interactions between multiple repayment and bill-amount variables.

Our next model, a Random Forest Classifier, incorporated a grid search over the number of trees. The final version of this model achieved a sizable improvement in ROC-AUC using a score of 0.7771. The ROC-AUC score of 0.7771 exemplifies the ability of this tree ensemble model to capture non-linear relationships and interactions between features. The Random Forest model produced an increased level of Precision (0.6748) and Recall (0.3766) on the minority default class as compared to Logistic Regression, although the recall levels still remain limited. The feature-importance results again placed PAY_0 and LIMIT_BAL among the top predictors, with several PAY variables and BILL_AMT1 following closely. This ranking aligns strongly with both the correlation heatmap and the mutual information plot from EDA: repayment history is consistently the clearest early-warning signal of financial distress, while a borrower's credit limit and bill amounts provide additional context about credit exposure. Unlike Logistic Regression, however, Random Forest distributed importance more evenly across several repayment months (PAY_2–PAY_6), illustrating that late-payment patterns across time,not just the most recent repayment status,contribute meaningfully to default prediction.

Building on these insights, we trained an XGBoost classifier, tuning n_estimators, max_depth, and learning_rate. XGBoost achieved a significantly stronger ROC-AUC of 0.7923, the highest among all baseline models. Although precision for the default class remained below what we would ultimately hope for (0.6973), the model exhibited strong precision for non-defaults and generally improved F1-scores compared to prior models. The permutation-importance plot showed an even more pronounced hierarchy than Random Forests: once again, PAY_0 dominated as the single most important variable, followed by LIMIT_BAL, BILL_AMT1, and PAY_2. This reflects XGBoost's ability to partition the feature space around the strongest signal,in this case, repayment history,and then refine predictions using subtler bill-amount and payment-amount variations. The model's results strongly supported our earlier conclusion from mutual information analysis that delinquency variables contain nonlinear relationships that boosting models are particularly well-equipped to exploit.

Finally, we implemented LightGBM, tuning tree depth, learning rate, and the number of boosting iterations. LightGBM performed comparably to the XGBoost model, attaining an ROC/AUC of 0.7922 and achieved very balanced results in precision, recall, and F1 among the baseline models tested. Although the precision of the default class (0.7055) and F1 (0.4878) were the best of all baseline models, the efficiency and stability of the LightGBM algorithm remain intact when utilized on large moderately imbalanced tabular data sets during the training phase. LightGBM's feature-importance plot closely resembled that of XGBoost, with PAY_0 overwhelmingly dominant and LIMIT_BAL, BILL_AMT1, PAY_2, and PAY_AMT1 forming the next tier of predictive variables. This consistency across boosting models emphasizes that the dataset contains a clear hierarchy of predictive signals: short-term delinquency behavior is the single strongest determinant of default risk, while credit-limits, bill statements, and payment magnitudes form secondary but important predictors. Demographic features,including SEX, EDUCATION, and MARRIAGE,consistently appeared near the bottom of importance rankings, supporting the conclusion that behavioral financial variables are far more predictive than demographic categories in default-risk modeling.

After evaluating the baseline models on the imbalanced dataset, we extended our analysis by incorporating SMOTE into the training pipelines. The goal of this stage was to determine how artificially balancing the minority class would affect prediction quality,particularly recall for defaulters, which was consistently the weakest metric across all baseline models. SMOTE is a synthetic minority oversampling technique that improves on the minority class by allowing models to observe more patterns associated with default and reducing the propensity of these models to overwhelmingly predict the majority class. All four models we tested, Logistic Regression, Random Forests, XGBoost and LightGBM were retrained using the same parameter grids and evaluation metric measures. This is necessary in order to allow for a fair comparison against the baseline results.

The Logistic Regression model trained with SMOTE produced expected results for a linear classifier. The ROC-AUC dropped slightly to 0.698, but the recall of a defaulting borrower increased significantly from 0.631 to 0.574 which indicates a stronger ability to locate minority class borrowers. The precision of a default borrower was still not that high (0.364), however, the minority class F1 score for the SMOTE-enhanced Logistic Regression model increased dramatically from 0.469 to 0.445. Feature importance rankings continue to be similar to those of the baseline model. PAY_0 is still the most important feature by a significant margin, with BILL_AMT1 and MARRIAGE following in importance as the second and third most important features, respectively. Because SMOTE creates synthetic defaulters by interpolating existing minority patterns, the model became more sensitive to variations in delinquency behavior, which explains the expanded importance of behavioral features and reduced emphasis on demographic variables.

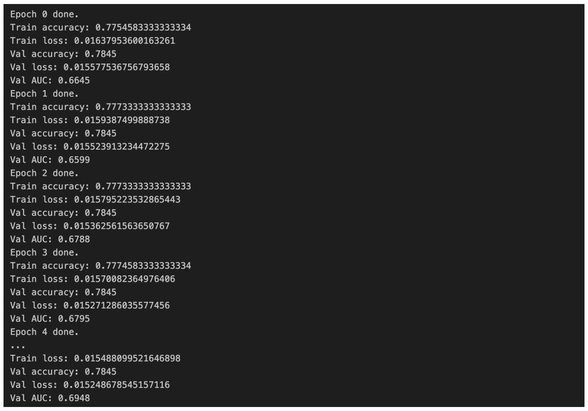
The Random Forest using SMOTE significantly increased performance of the minority class. The ROC decreased marginally (AUC = 0.758), but precision for normal cases increased significantly to 0.534 and recall increased to 0.494 which resulted in a F1-score that is much higher than the baseline (F1 = 0.513). This change in performance shows that when using SMOTE trained ensemble models, there is a trade-off between increased accuracy in classifying true defaulters at the expense of classifier confidence (overpredicting non-defaulters). Feature-importance rankings demonstrated a slight rebalancing compared to the baseline results. PAY_0 remained the most influential feature, but PAY_2, PAY_AMT1, and LIMIT_BAL increased in importance, indicating that synthetic sampling allowed the model to pick up more subtle distinctions among delinquent borrowers. The increased contribution of these features aligns with SMOTE's goal: to promote learning from minority-class structure.

Similarly, the XGBoost + SMOTE model provides improved performance across the board for the minority class by reaching an ROC-AUC score of 0.762 and having a default Precision of 0.490, and Recall at 0.545 for this model, which is a very significant improvement from the Baseline model which had only a Recall of 0.369. In addition, the F1-score increased to 0.516, indicating a more balanced approach towards detecting defaulting borrowers, while attempting to minimise false positives. Feature importance shifts closely mirrored this trend. While PAY_0 still dominated, PAY_2, BILL_AMT1, and LIMIT_BAL all gained relative importance, suggesting that boosting algorithms became increasingly sensitive to broader behavioral patterns once additional synthetic minority-class samples were introduced. Compared to

Random Forests, XGBoost showed a more refined gradient in importance across repayment and bill variables, which fits with its iterative boosting mechanism for reducing residual error.

Finally, the LightGBM + SMOTE model delivered the strongest performance of all SMOTE-enhanced classifiers. It achieved the highest ROC-AUC of the group (0.765) while also producing the best balance between precision (0.500) and recall (0.528) for defaulting borrowers. Notably, its minority-class F1-score (0.514) surpassed that of the other models, reinforcing LightGBM's efficiency and stability on structured tabular data. Feature importance distributions remained similar to the non-SMOTE version: PAY_0 stayed dominant, but secondary predictors such as BILL_AMT1, PAY_2, MARRIAGE, and LIMIT_BAL increased in influence. This indicates that LightGBM,more so than XGBoost,benefited from the more consistent minority-class structure introduced through SMOTE, allowing it to learn nuanced relationships across financial indicators and demographic attributes.

While it wasn't mentioned in our initial proposal, we decided that trying as many models as possible to widen our comparison bubble should be one of our goals, which is why we decided to implement a binary classification neural network. Using a sequential model with 24 inputs and 2 outputs, as well as using other techniques like dropout layers, ReLu activations, and batch normalization, we were able to train a model on our dataset. The reasoning for not using a neural network in the first place was due to the idea that things like random forests, Boosted Decision Trees (xgboost and lightgbm), and even the most basic logistic regression models are all more powerful and easier implemented tools when it comes to binary classification.



```
Epoch 0 done.
Train accuracy: 0.7754583333333334
Train loss: 0.01637953600163261
Val accuracy: 0.7845
Val loss: 0.015577536756793658
Val AUC: 0.6645
Epoch 1 done.
Train accuracy: 0.7773333333333333
Train loss: 0.0159387499888738
Val accuracy: 0.7845
Val loss: 0.015523913234472275
Val AUC: 0.6599
Epoch 2 done.
Train accuracy: 0.7773333333333333
Train loss: 0.015795223532865443
Val accuracy: 0.7845
Val loss: 0.015362561563650767
Val AUC: 0.6788
Epoch 3 done.
Train accuracy: 0.7774583333333334
Train loss: 0.01570082364976406
Val accuracy: 0.7845
Val loss: 0.015271286035577456
Val AUC: 0.6795
Epoch 4 done.
...
Train loss: 0.015488099521646898
Val accuracy: 0.7845
Val loss: 0.015248678545157116
Val AUC: 0.6948
```

Fig 11

The weak AUC score for our neural network proves that theory and drives home the idea that sometimes less is more.

Results

Our results revealed distinct trends in how varying models reacted both to the SMOTE training pipeline and to the imbalanced dataset. For example, although the standard logistic regression method achieved a good level of overall accuracy, the recall of defaulted borrowers was very low. This demonstrated that simply applying linear mechanisms was not sufficient to account for the complex nature of financial interactions involved in borrower defaults. The random forest was able to slightly outperform the logistic regression approach by achieving higher than the average recall and F1-score, but had an ongoing tendency to underpredict defaults. In contrast, both the XGBoost and LightGBM approaches produced the highest ROC-AUC scores (area under the receiver operating characteristic curve) and produced better balanced predictions over both classes than the previous two models. The use of XGBoost and LightGBM to create boosted models enabled these algorithms to use the subtle interaction between repayment history, bill amounts, and credit limits highlighted as the highest-ranked predictors of default.

In addition, the introduction of the SMOTE technique significantly increased the level of performance for the minority class in all models. Specifically, recall for the defaulted borrowers increased dramatically, which indicated that the SMOTE resampling technique was successfully reducing the level of imbalance that had previously been introduced to the baseline models. Although ROC-AUC was slightly decreased for certain models, the increase in recall, primarily with the XGBoost-SMOTE and LightGBM-SMOTE models, represented a significant improvement for real-world lending applications where correctly identifying at-risk borrowers will be of utmost importance in decision-making. Feature-importance rankings remained stable overall, with delinquency variables dominating, but SMOTE allowed models to extract additional predictive signals from bill amounts and payment magnitudes.

The best overall model was the LightGBM with SMOTE model. This model demonstrated the most balanced trade-off between recall, precision, and AUC, and therefore it offers the best option as the basis for a real-world system that predicts credit risk.

Conclusion

The results of this study support the notion that it is possible to predict whether a borrower will default on a loan in the future based on financial and demographic characteristics by employing state-of-the-art machine learning techniques. The predictions for all models followed a consistent trend. Delinquency history is the single best predictor of future defaults, whereas demographic characteristics are not nearly as important in predicting whether a borrower will default.

The results of the experiments also demonstrated that class imbalance will significantly inhibit the performance of all models unless explicitly addressed. By using SMOTE, all models learned from the data differently, allowing each model to significantly increase recall for defaulting borrowers and identify deeper trends related to repayment patterns, amount of bills, and amount of credit limit.

Of all methods evaluated, LightGBM trained with SMOTE produced the best balance of reliable predictive ability. In addition to producing very high recall rates for defaulting borrowers, LightGBM also produced good precision rates and achieved the highest ROC-AUC among all of the models enhanced with SMOTE. This balance makes LightGBM with SMOTE particularly appropriate for many real-world lending environments in which the consequences of failing to detect a true defaulter are far greater than the inconvenience of an insignificant increase in the number of false positives.

The research highlights how careful data preparation combined with resampling and boosting algorithms can create a robust model for assessing creditworthiness. The results support the role of data-driven decision-making within financial institutions and offer a practical base for future efforts related to data analytics in finance.

Opportunities for extending the current research include implementing cost-sensitive learning algorithms, developing new methods for identifying which variables influence models most significantly, and using the current model in conjunction with real-time loan decision-making processes.