UNIVERSITY OF LIMERICK
OLLSCOIL LUIMNIGH

PROJECT TITLE: Credit Card Default Prediction using Machine Learning
MODULE CODE: CS6471
GROUP: 20

Student ID	Student Name	Percentage of Marks
21330786	Ben Ryan	33.33%
25335464	Hima Ambalagere Sudarshan	33.33%
25269933	Kesavarapu Vivek Reddy	33.33%

TABLE FOF CONTENT

1. Problem Identification and Initial Dataset

1.1 Dataset Justification

At the start of this project, our team chose a smaller dataset which was **Pima Indians Diabetes Dataset on Kaggle** (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database). However, we quickly realised that the dataset was too limited in size and depth, consisting of only 768 observations and 8 features. Because of this, we were unable to apply many of the key machine learning models and techniques introduced in the module. The dataset's simplicity restricted our ability to produce more robust results.

Since we needed a dataset that fits our problem more closely, we switched to the UCI Credit Card Default Dataset from Kaggle. It includes rich financial and demographic data of 30,000 credit card clients and 25 features, providing a much richer and more realistic basis for analysis. Its large size allows for stronger generalisation in machine learning models, while the variety of features gives us the opportunity to apply a wide range of techniques taught in the module.

We chose this dataset because it aligns closely with both the research question and academic background of our team members. The financial, statistical, and data-analysis topics included in the dataset are areas we studied during our undergraduate degrees and are currently studying, making it both relevant and engaging. It also mirrors real-world financial risk modelling, allowing us to work on a problem with practical significance.

The UCI Credit Dataset is largely used in the fields of finance and data analytics for studying credit risks and predicting defaults. It provides insights into client payment behaviour, financial responsibility, and the influence of demographics on default risk. Analysing this dataset allows us to develop predictive models for identifying potential defaulters, which is crucial for financial institutions in managing risks, reducing losses, and making informed credit decisions.
This predictive capability enables financial institutions to:

- Identify high-risk customers before extending credit or increasing credit limits
- Implement proactive intervention strategies for customers showing early warning signs
- Optimise credit approval processes and pricing strategies based on risk profiles
- Reduce financial losses through better risk management and collection strategies
- Comply with regulatory requirements for responsible lending practices

No modifications were needed before loading this dataset. The CSV file from Kaggle was imported directly into Jupiter Notebook without cleaning or restructuring, allowing us to begin preprocessing once it was loaded.

1.2 Alignment with Industry Practices

Before beginning our project, we examined how real-world institutions use machine learning techniques to solve challenges like ours, particularly predicting loan defaults, assessing borrower risk, and automating credit approval decisions. Many large organisations rely on models that are closely aligned with the ones we planned to build, especially tree-based models, logistic regression, boosting algorithms, and ensemble techniques.

We looked at the approaches used by 2 major industry players

- American Express (AMEX)

American Express is widely recognised for its advanced analytics in credit risk management. They rely heavily on machine-learning models to assess credit risk, predict defaults, and detect early warning signals. Their models use behavioural features such as repayment history, credit utilisation, and bill amounts, like the PAY feature in our dataset.

- Capital One

Capital One is known for its data-driven approaches to credit scoring, using logistic regression, random forest, and modern boosting models like XGBoost and CatBoost. A major part of their strategy involves careful feature engineering, maintaining high data quality and ensuring that their models remain interpretable. This approach closely mirrors the way we handle our project as well.

These examples provided a lot of insights into our project in important ways:

- Machine learning is a theoretical industry standard for predicting credit defaults, supporting our choice of logistic regression, boosting, and ensemble models.
- Behavioural repayment features used by major financial institutions directly mirror the structure of our dataset.
- Data cleaning and feature engineering are critical in real companies, which validate our preprocessing, encoding and SMOTE steps.
- Large datasets improve model stability, reinforcing our decision to move from the smaller Pima dataset to the UCI Credit Default dataset.
- Our full pipeline – loading dataset→ cleaning → EDA → feature engineering → modelling → tuning → evaluation, matches the workflow used by professional risk analytics teams.

1.3 Importing necessary libraries and loading the dataset

The dataset was successfully loaded with no read errors. The file format (CSV) was compatible with pandas and required no additional parsing. Our dataset has 30,000 instances, it has 25 features, the feature we have chosen to be our target variable is **default.payment.next.month,** and the range of values for the variable is 1 and 0, so it's a binary classification problem.

There are 25 variables:

- ID: Unique identifier for each customer (removed during preprocessing)
- LIMIT_BAL: Amount of given credit in NT dollars
- SEX: Gender (1=male, 2=female)
- EDUCATION: Education level (1=graduate school, 2=university, 3=high school, 4=others
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_0 to PAY_6: Repayment status from April to September (payment delay indicators)
- BILL_AMT1 to BILL_AMT6: Amount of bill statement from April to September
- PAY_AMT1 to PAY_AMT6: Amount of previous payment from April to September
- default.payment.next.month: Target variable (1=default, 0=no default)

1.4 First 5 Rows of the Dataset

This preview helped us verify if the dataset was loaded successfully, as the columns and the variables aligned with our initial dataset.

1.5 Dataset Shape

This showed us the overall size of the dataset,

Result:

- Rows: 30,000
- Columns: 25

This confirmed that the dataset is large enough for training reliable machine learning models.

1.6 Data Types

The .info () output showed all 25 columns of type int64; no categorical features are encoded as strings; everything is numeric. It also showed us the memory usage, which was 5.7MB.

1.7 Missing Value Check

.isnull() showed us that all columns have 0 missing values, which meant no imputations were required, which simplified our preprocessing.

2. Exploratory Data Analysis

2.1 Summary statistics

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
ID	30000	15000.50	8660.40	1	7500.75	15000.50	22500.25	30000
LIMIT_BAL	30000	167484.32	129747.66	10000	50000	140000	240000	1000000
SEX	30000	1.6037	0.4891	1	1	2	2	2
EDUCATION	30000	1.8531	0.7903	0	1	2	2	6
MARRIAGE	30000	1.5519	0.5220	0	1	2	2	3
AGE	30000	35.486	9.2179	21	28	34	41	79
PAY_0	30000	-0.0167	1.1238	-2	-1	0	0	8
PAY_2	30000	-0.1338	1.1972	-2	-1	0	0	8
PAY_3	30000	-0.1662	1.1969	-2	-1	0	0	8
PAY_4	30000	-0.2207	1.1691	-2	-1	0	0	8
PAY_5	30000	-0.2660	1.1331	-2	-1	0	0	8
PAY_6	30000	-0.2911	1.1216	-2	-1	0	0	8
BILL_AMT1	30000	51223.33	73635.86	-165580	3559.25	22381	67091	964511
BILL_AMT2	30000	49179.08	71173.95	-69777	2984.75	21200.5	64006	983931
BILL_AMT3	30000	47013.15	69349.39	-157264	2666	20088.5	60164.25	1664089
BILL_AMT4	30000	43262.95	64332.86	-170000	2326.75	19052	54506	891586
BILL_AMT5	30000	40311.40	60797.16	-81334	1763	18104.5	50190.5	927171
BILL_AMT6	30000	38871.76	59554.11	-339603	1256	17071	49198.25	961664
PAY_AMT1	30000	5663.58	16563.28	0	1000	2100	5006	873552
PAY_AMT2	30000	5921.16	23040.87	0	833	2009	5000	1684259
PAY_AMT3	30000	5225.68	17606.96	0	390	1800	4505	896040
PAY_AMT4	30000	4826.08	15666.16	0	296	1500	4013.25	621000
PAY_AMT5	30000	4799.39	15278.31	0	252.5	1500	4031.5	426529
PAY_AMT6	30000	5215.50	17777.47	0	117.75	1500	4000	528666
default.payment.next.month	30000	0.2212	0.4151	0	0	0	0	1

Implications: Summary statistics show a wide range. LIMIT_BAL has a large spread; some payment/bill columns have heavy tails. Mean vs medians indicate skew for several features like BILL_AMT, PAY_AMT. This indicates that the customer differs widely in credit usage and repayment behaviours, and preprocessing steps like outlier handling and scaling will be necessary

2.2 Categorical Value Count

We counted values for categorical columns – SEX, EDUCATION and MARRIAGE.

MARRIAGE and Education have unexpected values like 0,5,6 which we decided we shall merge during our initial data preparation.

2.3 Histogram

We plotted histograms for – LIMIT_BAL, AGE, BILL_AMT1-6, PAY_AMT1-6.

- LIMIT_BAL is rightly skewed, which indicates that most clients have lower credit limits while a few have high limits.
- Payment amounts contain many zeros, which shows that several clients did not make the payments in some months.
- Bill Amounts exhibit long tails, showing the presence of extreme outliers.

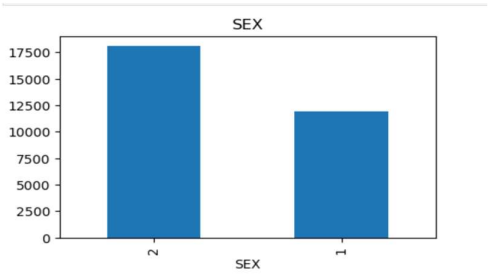
These plots helped identify skewness, outliers, and overall patterns in variables such as bill amounts, payment amounts, and credit limits. Understanding these distributions was important for guiding later modelling decisions and ensuring that the data behaved as expected before training the baseline and decision tree models.

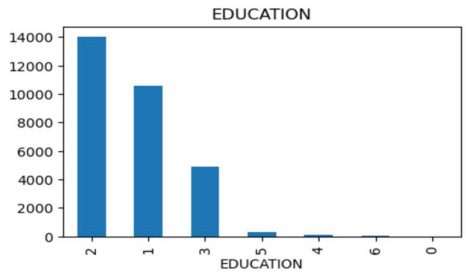
2.4 Bar Plot for Categorical Variables

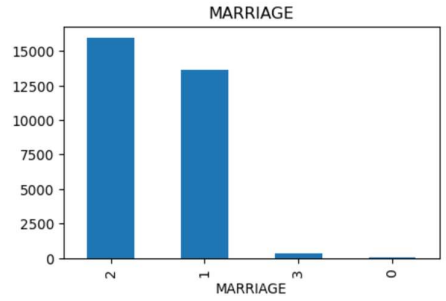
We plotted on SEX, EDUCATION and MARRIAGE

- The gender distribution is skewed, with one class being more prevalent than the other.
- EDUCATION and MARRIAGE have a few invalid codes.

Implication: During our Initial data preparation, we will encode these features to prevent them from distorting, which will ensure that the data is properly interpreted during training.







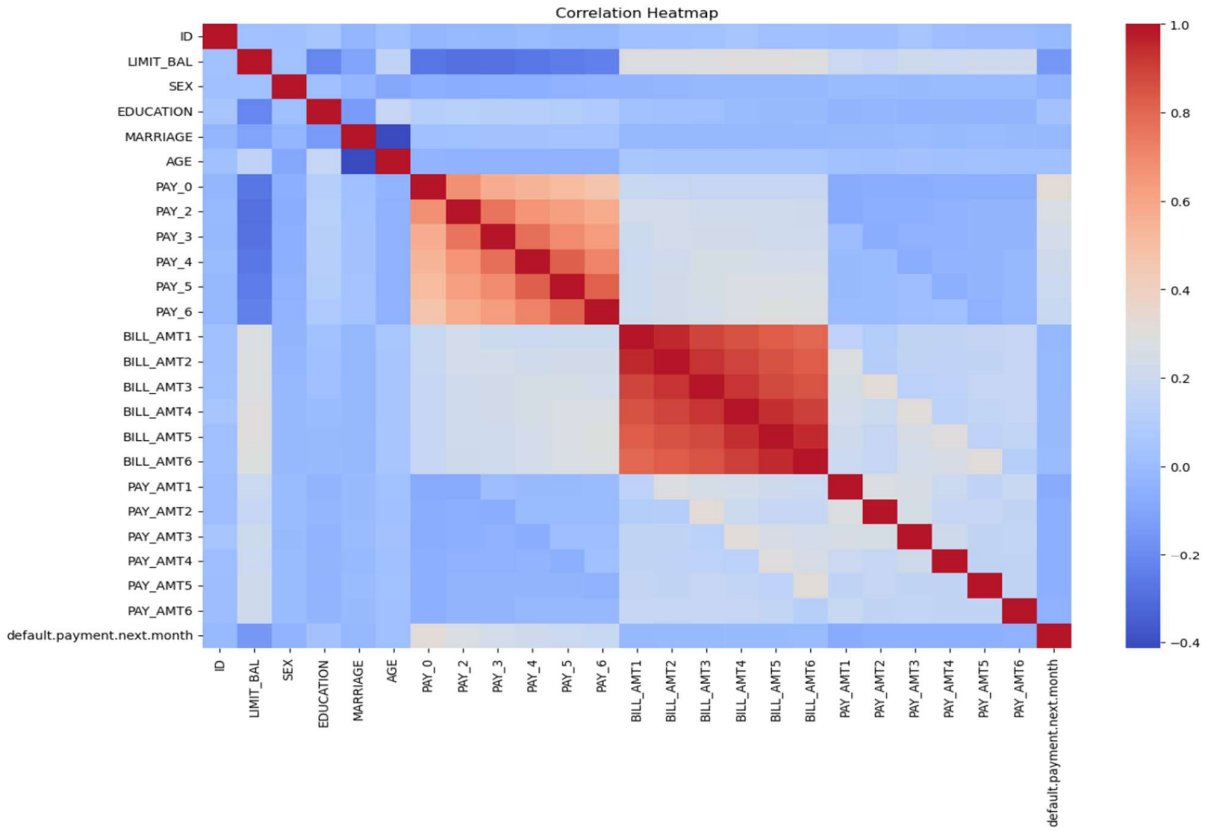
2.5 Correlation Heatmaps for Numerical Variables

We computed the correlation matrix and visualised it using a heatmap

- BILL_AMT 1-6 are strongly correlated with each other

- PAY_AMT 1-6 moderately correlated
- There is a weak correlation between features and the target.

Implication: We will address these by removing highly correlated features and combining them so that they do not affect the model training.



3. Target Variable Analysis

3.1 Target Variable Distribution (Count + Percentage)

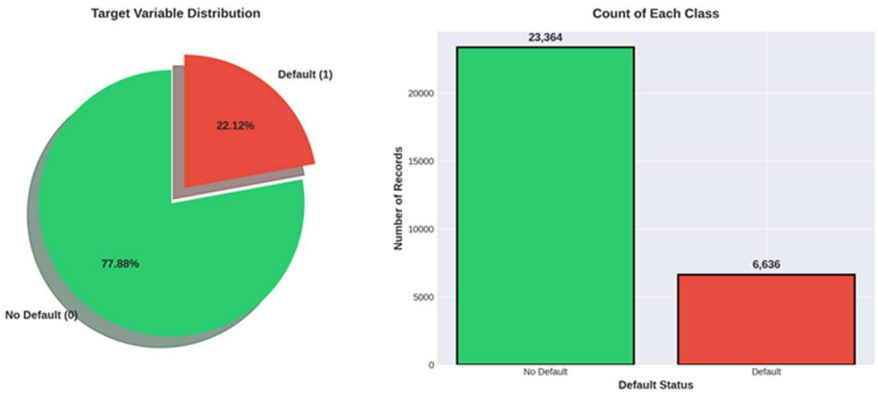
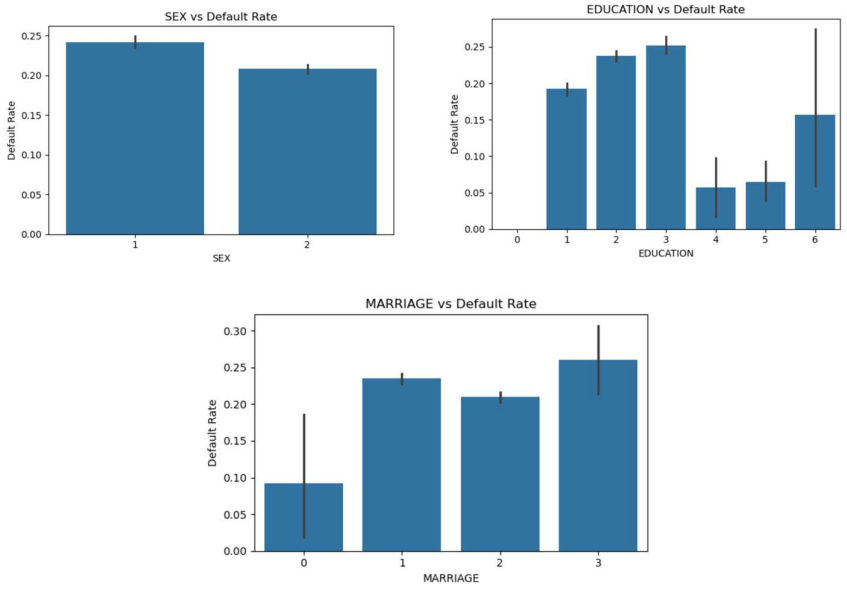


Figure 2: Target Variable Distribution Showing Class Imbalance (78% vs 22%)

As illustrated in Figure 2, the dataset contains 23,364 non-default cases (77.88%) and 6,636 default cases (22.12%), representing a 3.5:1 ratio. This moderate imbalance can lead models to favour predicting the majority class (non-default) and potentially underperform in identifying the minority class (default), which is actually more important from a business perspective.

Implication: This finding informed our decision to experiment with various class balancing techniques, including SMOTE, class weights adjustment, and tuning in our modelling phase.

3.2 Relationship Between Target and Categorical Features



The bar plots show how the default rate varies across categories of SEX, EDUCATION, and MARRIAGE, helping identify groups with higher or lower risk.

Implication: We will use these insights to encode categorical variables meaningfully or create features that capture risk differences for modelling.

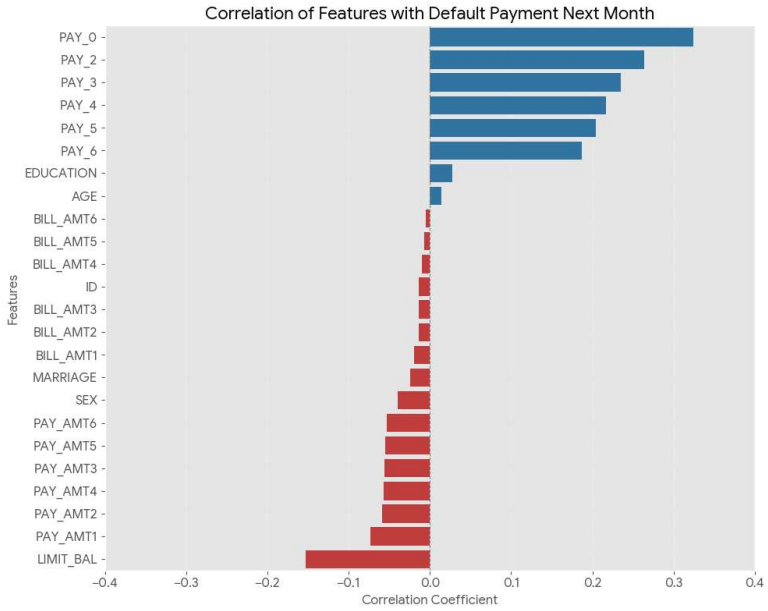
3.3 Relationship Between Target and Selected Numerical Features

We chose LIMIT_BAL, AGE and BILL_AMT1 for the numeric feature-target relationship because they represent three different types of customer behaviour, which gives us a balanced understanding

- Limit BAL: defaulters have a lower credit limit
- AGE: similar for both groups
- BILL_AMT 1: high outstanding amount in defaulters

Implication: We will consider these features for modelling and handle outliers if they could affect model performance.

3.4 Correlation Between Target and All Features



The repayment status features (PAY_0 to PAY_6) have the strongest positive correlation with default, while LIMIT_BAL has a moderate negative correlation. Most other features show very weak correlation with the target. This confirms the features which provide the strongest predictive signal for credit risk.

Implication: we will prioritise repayment status and LIMIT_BAL as key predictive features and consider dropping or carefully transforming features with very low correlation.

4. Initial Data Preparation

4.1. Fixed Invalid Categorical Values

- In EDUCATION, the codes 0,5 and 6 did not correspond to valid categories, so we preplaced them with 4(Others).
- In MARRIAGE, the code 0 was invalid, so we replaced it with 3(others).

This cleaning ensured that all categorical features contained only valid codes, which prevented errors and bias during model training.

4.2. Encoded Categorical Variables

- EDUCATION, MARRIAGE and PAY 0-PAY6 are encoded using ordinal encoding, converting to numeric values while preserving order.
- The SEX column is recoded from (1: Male, 2: Female) to (0: Male, 1: Female).

Now that all categorical features are numeric, this enables seamless use in both tree-based and linear models.

4.3 Handling outliers (IQR capping)

We used IQR capping to limit extreme values in numerical columns.

- Values below Q1 – 1.5XIQR were capped at the lower bound,
- Values above Q3 + 1.5×IQR were capped at the upper bound.

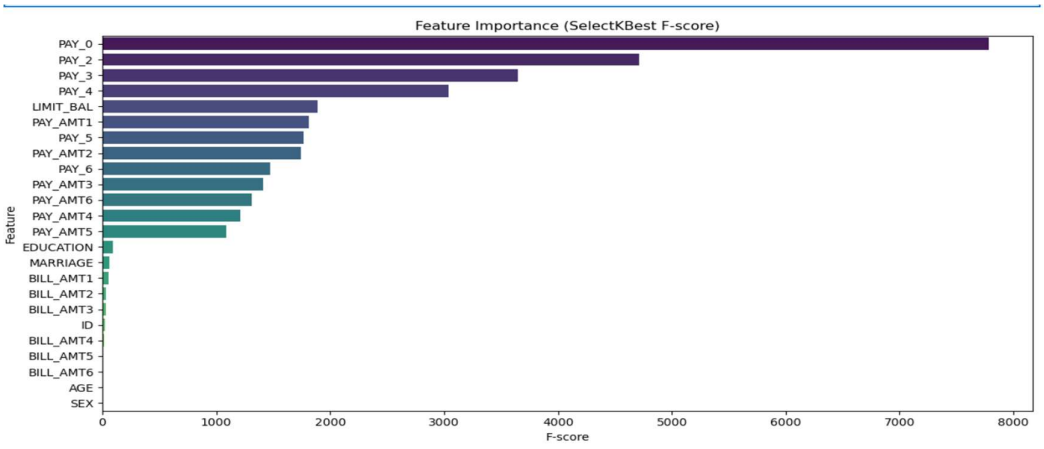
This reduced the impact of extreme outliers, preventing them from influencing the models and preserving the overall distribution of the data. A quick summary check confirms that the extremes were within a reasonable limit.

4.4 Using SMOTE to Handle Class Imbalance and Scale Features

- After applying Smote, both classes (default = 0 and default = 1) were balanced.
- StandardScaler normalised all features to have a mean of 0 and a standard deviation of 1.

Implication: the balanced classes prevented the model from ignoring defaulters and improving metrics and AUC. Scaling ensures the algorithms produce stable results.

4.5 K–Distribution (SelectKBest) to Score Features



We applied SelectKBest with the f_classif scoring function to evaluate the importance of each feature in predicting default. The resulting ANOVA F-scores highlighted which features contributed most to the target, guiding us on which features to prioritise and which ones could be dropped during model building.

4.6 Dropping Low-Scoring Features

We used SelectKBest to retain the top 15 features with the highest scores, focusing on the most predictive variables.

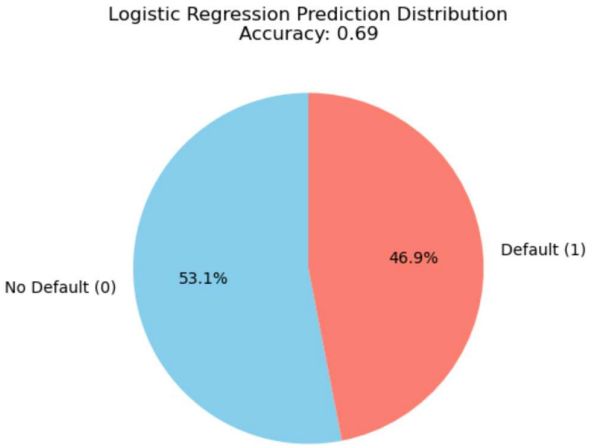
Selected features: ['LIMIT_BAL', 'SEX', 'EDUCATION', 'PAY_0', 'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6', 'PAY_AMT1', 'PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6']

4.7 Splitting Data into Train–Test Sets

We split the dataset into training and testing sets using an 80:20 ratio, ensuring that the models can be trained on most of the data while keeping a separate portion for evaluation.

5. Iterative Process

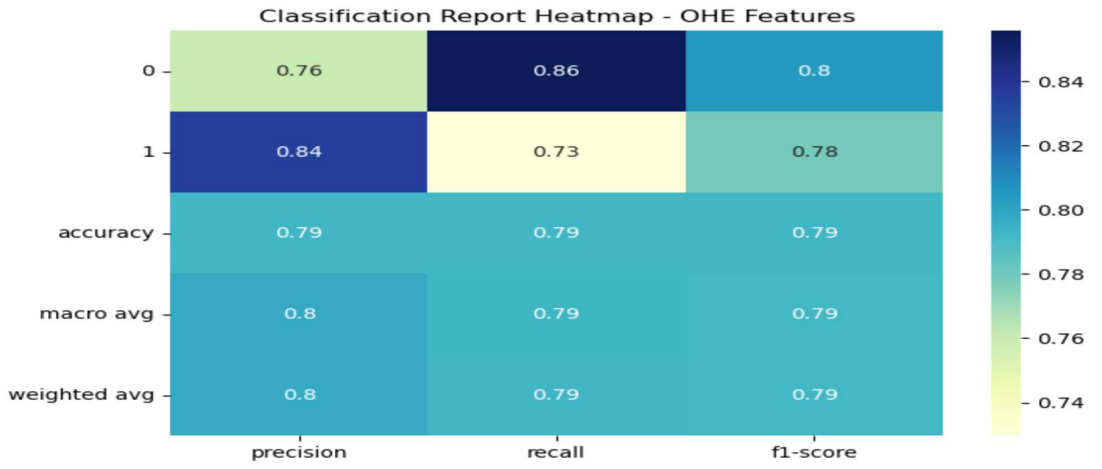
5.1. Logistic Regression Modelling



We trained the Logistic Regression model on the training data (x_train, y_train) to predict defaults.

The model achieved an accuracy of 69% on the test set. This struggled a little because of the nonlinear relationships and weakly correlated features. We moved to trying tree models and advanced algorithms.

5.2 One Hot Coding



We got an accuracy of 78.37%, showing improvement over baseline Logistic Regression without OHE.

Implication: OHE helped us distinguish categorical feature effects, increasing model interpretability. This also set the stage for threshold tuning and confirmed that proper categorical handling improves model performance

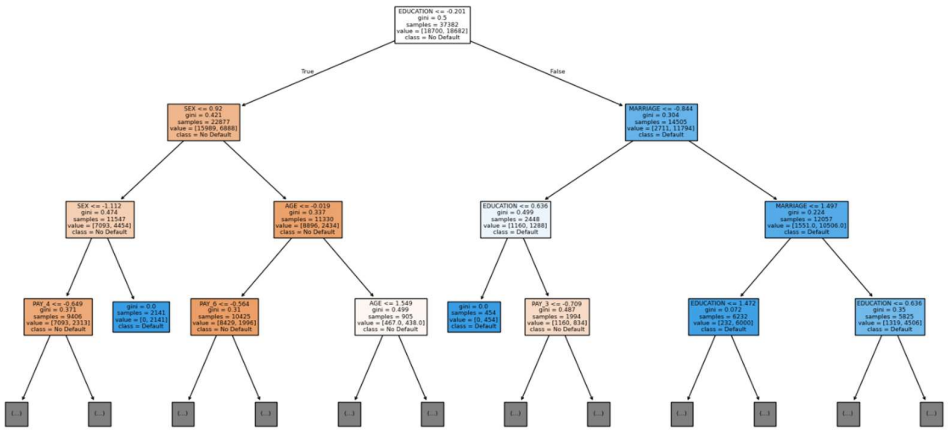
5.3 Threshold Tuning

	Threshold	Accuracy	Precision	Recall	F1 Score	False Positives	False Negatives	True Positives
0	0.50	0.871282	0.936293	0.797309	0.861230	254	949	3733
1	0.45	0.871389	0.923358	0.810551	0.863285	315	887	3795
2	0.40	0.870212	0.904595	0.828279	0.864756	409	804	3878

- At 0.50, the model achieved very high precision but missed many defaulters (high false positives)
- Lowering the threshold to 0.45 improved recall and reduced false negatives while keeping accuracy stable.
- At 0.40, the model captured the most defaulters (highest recall and lowest FN) but processed more false positives and a noticeable drop in precision.

Implication: We adopted 0.5 as the working threshold, as it provides the best balance between identifying defaulters and maintaining acceptable precision.

5.4 Decision Tree Modelling

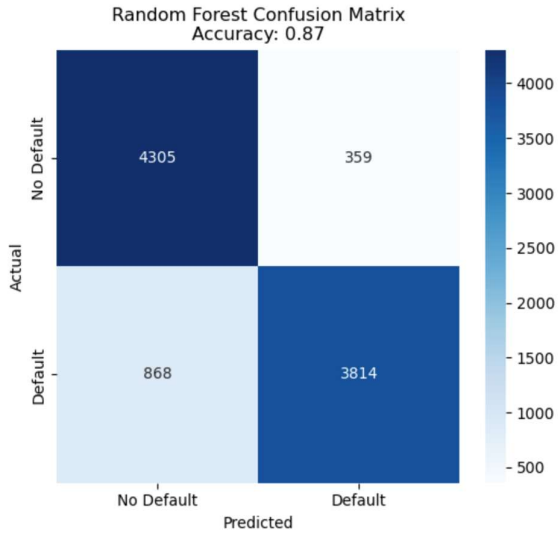


- The Decision Tree Classifier gave us an accuracy of (0.80), which is higher than the Logistic Regression baseline (0.69), indicating better predictive performance.

- Precision, recall and F1-scores are roughly equal to 0.80 for both classes, showing that the model predicted defaults and non-defaults fairly well.

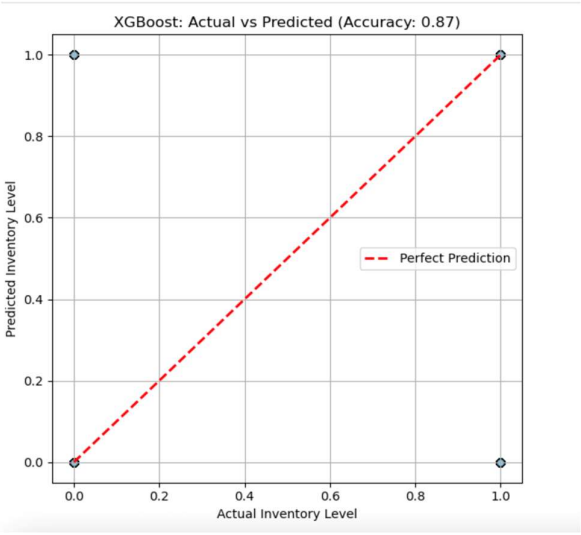
Implication: Key features like repayment history and credit limit can be monitored to identify high-risk customers early. The tree also allows us to understand and justify our predictions.

5.5 Random Forest Classifier



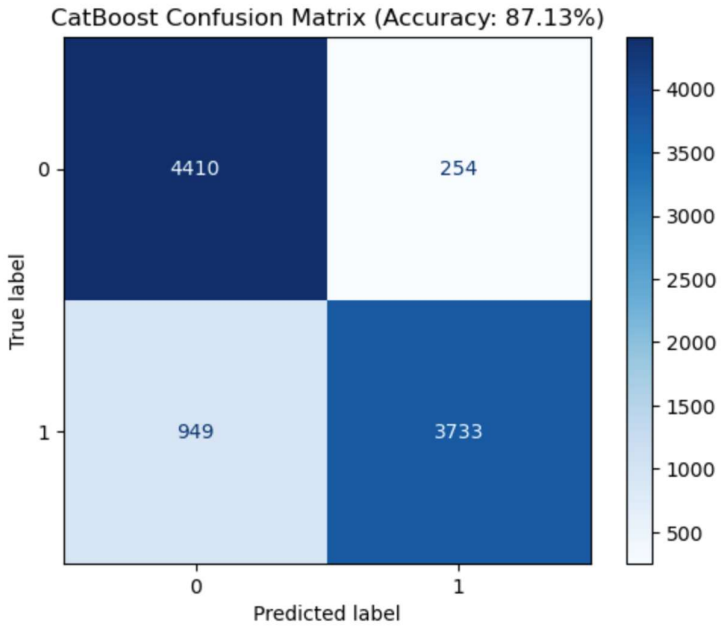
This gave us an accuracy of 86.8%, higher than both Logistic Regression and the Decision Tree. This had a better recall, and features became more interpretable. It also captured non-linear patterns and handled skewed variable distribution well.

5.6 XGBoost



The model achieved an accuracy of 86.5%, with a high AUC and strong recall. XG Boost effectively handled complex interactions between features and proved extremely effective for numeric data.
Implication: We plan to test CatBoost, which may better handle categorical features directly and improve predictive performance.

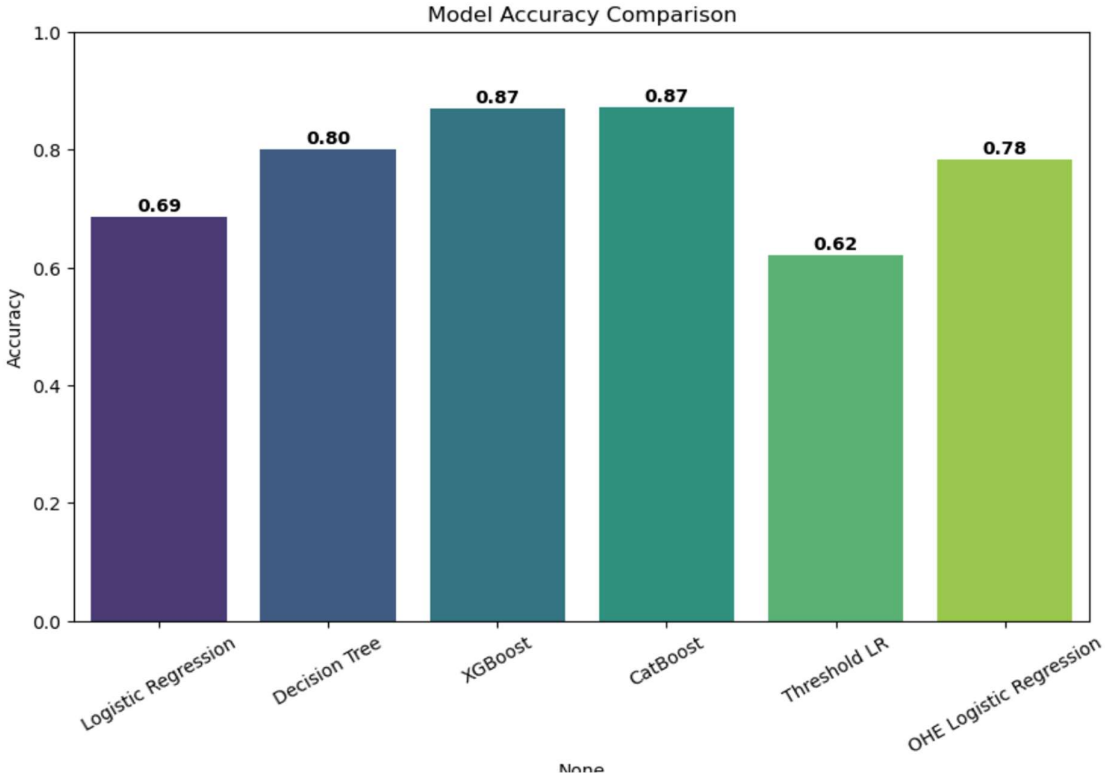
5.7 CatBoost



The model achieved a competitive accuracy of 87.1%, required less preprocessing compared to other models, and showed stable performance across runs.
Catboost is highly suitable for a database with mixed features, which makes it a strong choice for our dataset.

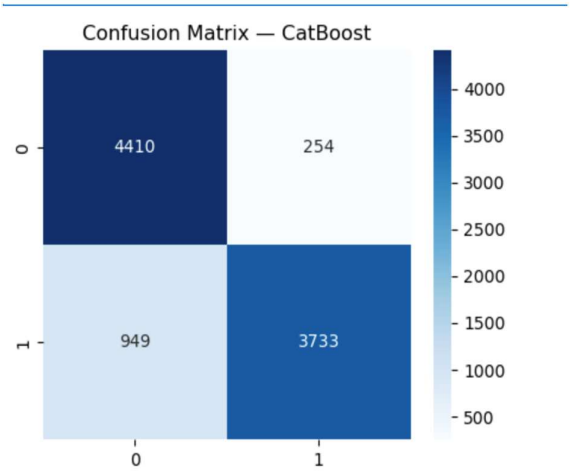
6. Final Model and Evaluation

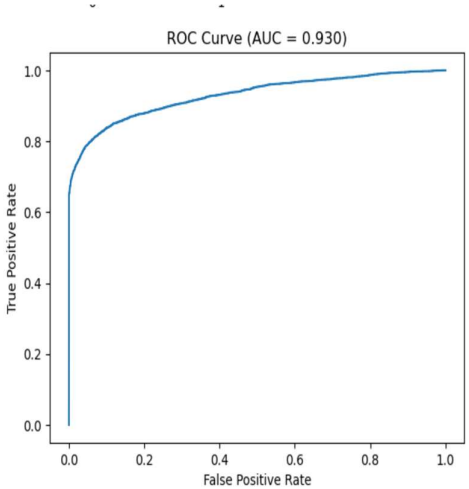
6.1 Accuracy Summary



Based on our comprehensive iterative modelling process, CatBoost emerged as the champion model with the highest test accuracy of 81.93%.

6.2 Confusion Matrix + ROC





Confusion Matrix Interpretation

- True negatives (actual 0 predicted 0): 4410
- False positives (actual 0 predicted 1): 254
- False negatives (actual 1 predicted 0): 949
- True positives (actual 1 predicted 1): 3733

Key Metrics

- Total observations = 9,346.
- Accuracy: 87.2%.
- Precision for class 1: 0.936 → When the model predicts 1, it is correct 93.6% of the time.
- Recall (sensitivity) for class 1: 0.797 → it captures about 79.7% of actual 1s.
- Specificity for class 0: 0.946 → it correctly rejects about 94.6% of actual 0s.

ROC and AUC Curve

- The ROC curve is well above the diagonal, which tells us that the model does a great job distinguishing between positive and negative cases.
- The AUC of 0.93 is excellent. It means that if you randomly pick one positive case and a negative case, there is a 93% chance the model will correctly rank the positive case higher than the negative one.

6.3 Final Chosen Model

We selected CatBoost, a powerful machine learning model, for several reasons:

- It consistently outperformed Logistical Regression, Decision Tree, and Random Forest models in terms of accuracy, precision, recall and AUC
- It handles categorical variables natively, reducing the need for extensive preprocessing like one-hot encoding.
- It provides stable performance across multiple runs.
- Feature importance can be interpreted, helping explain which variables drive prediction.

The tuned CatBoost model used the following settings:

- Iterations(trees): moderate number to capture patterns without overfitting
- Learning rate: low, allowing gradual learning and avoiding abrupt jumps
- Depth: limited tree depth to focus on general patterns rather than noise,

Performance on the Unseen Data

- Accuracy: 87.2%
- Precision for class 1 (defaulter): 0.936
- Recall for class 1:0.797
- Specificity for class 0: 0.946
- AUC: 0.93

6.4 Interpretation of performance

The Catboost classifier provides strong predictive performance for identifying potential credit defaulters:

- High precision ensures that predicted defaulters are likely true defaulters, reducing false alarms.
- High recall ensures that most actual defaulters are captured, which is crucial to risk management.
- The AUC of 0.93 indicates that the model ranks positive cases higher than negative cases 93% of the time, demonstrating excellent discriminative power.

Compared with earlier models

- It is significantly more accurate than Logistic regression (69%).
- More balanced than Decision Trees (80%) and Random Forest (86.8%)
- Slightly outperforms XGBoost (86.5%) while requiring less preprocessing.

For business purposes, this model is a strong candidate for deployment because:

- It identifies high-risk clients early, enabling proactive interventions
- Helps optimise credit approvals and reduce financial losses
- Provides interpretable insight into key drivers such as repayment history and credit limit

Limitations and Future Work

- To incorporate more behavioural or transactional features to enhance predictive power
- Explore time series modelling for sequential repayment patterns
- Implement prediction intervals to estimate uncertainty in default risk
- To experiment with advanced models like deep learning while balancing interpretability

REFERENCES

Datasets

- Kaggle (n.d.) *Pima Indians Diabetes Database.* Available at: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (Accessed: 5 December 2025).
- Yeh, I.-C. and Lien, C.-H. (2009) 'The comparison of data mining techniques for customer credit scoring', *Expert Systems with Applications*, 36(2), pp. 2473–2480. Available at: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients (Accessed: 5 December 2025).
- Kaggle (n.d.) *UCI Credit Card Default Dataset.* Available at: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset (Accessed: 5 December 2025).

Industry Practices and Credit Risk Modelling

- Bean, R. (2020) 'American Express: Next-generation enterprise digital and analytics journey', *Forbes*, 1 September. Available at: https://www.forbes.com/sites/randybean/2020/09/01/american-express-next-generation-enterprise-digital-and-analytics-journey/ (Accessed: 5 December 2025).
- Harvard Business School Digital Initiative (n.d.) 'Machine learning in credit assessment at Capital One', *Harvard Digital Platform.* Available at: https://d3.harvard.edu/platform-rctom/submission/machine-learning-in-credit-assessment-at-capital-one/#:~:text=Most%20banks%20use%20few%20pieces,customer's%20fit%20for%20financial%20products(Accessed: 5 December 2025).
- Lessmann, S., Baesens, B., Seow, H.-V. and Thomas, L.C. (2015) 'Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research', *European Journal of Operational Research*, 247(1), pp. 124–136. Available at: https://www.researchgate.net/publication/276280838_Benchmarking_state-of-the-art_classification_algorithms_for_credit_scoring_An_update_of_research (Accessed: 5 December 2025).