**Project title: Evaluating ML Models for Financial Risk and Marketing Prediction**

**Table of Contents**

**Abstract ………………………………………………………………………………………………. 3**

**1. Introduction .…………………………………………………………………………………... 4**

**2. Related Work ……………………………………………………………………….…………. 4**

**3. Machine Learning Methodology ………………………………………….………….. 5**

**3.1** Data Preprocessing **…………..………………………………………….…………. 5**

**3.2** Models Used **………………………………….………………………………….……. 5**

**4. Results & Evaluation ………………………………………………………….……………. 6**

**4.1** Classification Reports **…………………………….……………………………….. 6**

**4.2** Comparison Bar Charts **……………………………………………………………. 7**

**4.3** Confusion Matrices **………………………………….……………………………… 9**

**4.4** ROC-AUC Curves **…………………………………….………………………………. 10**

**4.5** Precision-Recall Curves **…………………………………………………………… 12**

**5. Discussion ………………………………………………………………………………………. 13**

**6. Conclusion ……………………………………………………………………………………… 14**

**7. References …………………………………………………………………………………….. 15**

**Abstract**

This report presents a comprehensive evaluation of five supervised machine learning models Logistic Regression, Decision Tree, Random Forest, XGBoost, and K-Nearest Neighbors (KNN) applied to two real-world binary classification problems, predicting customer subscription in the Bank Marketing dataset and forecasting credit default in the Default of Credit Card Clients dataset. These tasks are critical in domains such as financial risk assessment and targeted marketing, where accurately identifying minority class instances can have substantial business impact.  
  
The models were evaluated using several performance metrics, including Accuracy, Precision, Recall, and F1-score, with special emphasis on the performance for the minority class (class 1). In addition to tabular metrics, visual tools such as confusion matrices, ROC-AUC curves, and Precision-Recall curves were employed to support interpretability.  
  
Results indicate that ensemble-based models specifically XGBoost and Random Forest consistently achieve superior performance in class 1 detection across both datasets. The findings highlight the importance of metric selection and model choice when dealing with imbalanced classification problems. This study serves as a practical guide for model evaluation in similar real-world applications and is designed to be fully reproducible.

# **1. Introduction**

In recent years, machine learning (ML) has become a cornerstone of modern data analysis, enabling predictive modeling across numerous domains, including finance, healthcare, and marketing. This report explores the application of multiple supervised learning algorithms on two widely studied binary classification problems. Specifically, it analyses: (i) the Bank Marketing Dataset, which involves predicting whether a customer will subscribe to a term deposit product, and (ii) the Default of Credit Card Clients Dataset, which focuses on forecasting the likelihood of a client defaulting on their credit card payment in the upcoming month.  
  
These datasets present distinct challenges. The bank marketing dataset is characterised by class imbalance and various categorical attributes, necessitating robust preprocessing techniques [1], while the credit card default dataset includes financial and behavioural features with numerical skewness and imbalance issues [2]. Proper preprocessing steps such as normalisation, label encoding, and train-test splitting were employed to prepare both datasets for modelling [3].  
  
To assess model performance comprehensively, five supervised learning models were implemented: Logistic Regression, Random Forest, XGBoost, Decision Tree, and K-Nearest Neighbors (KNN). These models were chosen for their diversity in learning mechanisms from linear approaches to ensemble-based and instance-based techniques [4], [6]. The models were evaluated using multiple classification metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and precision-recall curves. Such a multifaceted evaluation is crucial, especially when dealing with imbalanced datasets, where accuracy alone may be misleading [7].  
  
The objective of this report is to compare the effectiveness of these classifiers and identify which models are most suitable for each dataset. This approach offers insights into the practical deployment of ML models in financial decision-making contexts.

# **2. Related Work**

Classification tasks in the financial and marketing domains have been widely explored in the machine learning literature. The Bank Marketing dataset [6], in particular, has been a popular benchmark in both academic studies and data science competitions. Researchers frequently apply traditional classifiers such as Logistic Regression and Decision Trees, alongside ensemble approaches like Random Forest [3] and XGBoost [2] due to their robustness and suitability for structured tabular data. For instance, Chen and Guestrin [2] highlight the scalability and predictive power of XGBoost for structured datasets, while Breiman [3] demonstrates the strength of Random Forests in managing high-dimensional financial features and mitigating overfitting.  
  
Similarly, the Default of Credit Card Clients dataset [7] has been extensively used in the context of credit risk assessment and default prediction. Studies often prioritise not only predictive performance but also interpretability and fairness, as these models inform high-stakes financial decisions. In these contexts, ensemble methods continue to show superior results, particularly in handling class imbalance and capturing complex feature interactions.  
  
Importantly, prior research consistently emphasises the limitations of using accuracy alone as an evaluation metric for imbalanced classification tasks. Metrics such as Recall, Precision, and F1-Score especially for the minority class are essential for a more realistic and fair assessment of model performance. This aligns with the evaluation strategy adopted in this report, which focuses on these metrics to ensure both effectiveness and ethical responsibility in classification model selection.

# **3. Machine Learning Methodology**

# 3.1 Data Preprocessing

To ensure that the machine learning models operate effectively, both datasets were pre-processed using standard data preparation techniques:

**Target Variable Conversion:** In the Bank dataset, the target variable 'y' was mapped from categorical (yes/no) to binary labels (1/0). In the Credit dataset, the column 'default payment next month' was renamed to 'target' and treated as a binary variable.

**Categorical Encoding:** The Bank dataset contained several categorical features such as 'job', 'marital', and 'education'. These were transformed using One-Hot Encoding, while avoiding the dummy variable trap by dropping the first level. The Credit dataset, on the other hand, consisted primarily of numerical features and did not require encoding.

**Feature Scaling:** All numeric features in both datasets were standardised using Scikit-learn’s StandardScaler. This was particularly important for the Credit dataset, which included variables like 'LIMIT\_BAL', 'BILL\_AMT1–6', and 'PAY\_AMT1–6'. Scaling was performed only on the training set, with the same parameters applied to the test set to prevent data leakage.

**Data Splitting:** Both datasets were split into training and testing sets using an 80/20 ratio with stratification, ensuring that the class imbalance was preserved in both subsets.

This preprocessing pipeline ensured fair and consistent model comparison, especially for algorithms like k-Nearest Neighbors and Logistic Regression, which are sensitive to feature magnitude.

# 3.2 Models Used

To evaluate and compare classification performance across two financial datasets, five machine learning algorithms were implemented. These models represent a variety of learning strategies, including linear, tree-based, ensemble, and instance-based approaches. The diversity of these methods allows for a more comprehensive understanding of how different algorithmic biases affect prediction outcomes.  
  
**Logistic Regression (LR):** A widely used linear model for binary classification tasks, Logistic Regression estimates the probability that an input belongs to a certain class by applying a logistic function to a linear combination of input features. It serves as a strong baseline due to its simplicity, interpretability, and robustness under linearly separable conditions [1], [9]

**Random Forest Classifier (RF):** An ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes of individual trees. It reduces variance and mitigates overfitting compared to a single decision tree [3].

**XGBoost Classifier (Extreme Gradient Boosting):** A highly efficient and scalable gradient boosting framework, XGBoost builds an additive model in a forward stage-wise manner. It optimises a differentiable loss function and includes regularization, which improves performance and reduces overfitting [2].

**Decision Tree Classifier (DT):** A non-parametric model that learns simple decision rules inferred from data features. Although prone to overfitting, it is intuitive and serves as the base model for ensemble techniques like Random Forests and Gradient Boosting [3], [9].  
**K-Nearest Neighbors (KNN):** An instance-based model that classifies data points based on the majority class among their k-nearest neighbours in the feature space. While simple, KNN can be effective in low-dimensional, well-scaled data environments, but may suffer in high-dimensional or noisy settings [1], [9].  
  
Each model was implemented using Scikit-learn or XGBoost libraries in Python, with hyperparameters kept at default settings to facilitate fair baseline comparison. Performance was evaluated using a consistent pipeline of training, testing, and metric evaluation.

# **4. Results & Evaluation**

### 4.1 Classification Reports

Each machine learning model's performance was carefully evaluated using a set of standard classification metrics, including Accuracy, Precision, Recall, and F1-Score. These metrics offer complementary perspectives on model effectiveness. Accuracy provides a general sense of correctness, but in imbalanced datasets such as those analysed in this study it can be misleading. Therefore, greater emphasis was placed on Recall and F1-Score for class 1, which typically represents the positive class or the minority class of interest.  
  
In both datasets, class 1 captures crucial real-world outcomes: in the Credit Card Default dataset shown in (Table 1), class 1 identifies customers likely to default on payments an outcome with significant financial implications for lenders. In the Bank Marketing dataset shown in (Table 2), class 1 corresponds to customers who subscribed to a term deposit, representing successful marketing conversions. In these contexts, accurately detecting the positive class is more valuable than optimising overall accuracy. For instance, failing to identify a defaulter could result in financial loss, wile missing a potential subscriber could lead to lost business opportunities.  
  
The Precision metric reflects how many of the predicted positives were correct, which is especially relevant when false positives have a cost. Recall, on the other hand, shows how well the model captures all actual positives, which is crucial when false negatives are more damaging. The F1-score provides a balance between Precision and Recall, serving as a single harmonic mean score that accounts for both.  
  
These metrics were calculated using the classification\_report function from Scikit-learn, and the results are consolidated into tables for easier model comparison across both datasets. The best-performing models for each dataset were determined based on their F1-score for class 1.

**Table 1: Credit Card Default Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision (Class 1)** | **Recall (Class 1)** | **F1-Score (Class 1)** |
| **Logistic Regression** | 0.81 | 0.60 | 0.55 | 0.57 |
| **Random Forest** | 0.83 | 0.65 | 0.58 | 0.61 |
| **XGBoost** | 0.84 | 0.67 | 0.59 | 0.63 |
| **Decision Tree** | 0.76 | 0.45 | 0.42 | 0.43 |
| **KNN** | 0.78 | 0.54 | 0.47 | 0.50 |

**Table 2: Bank Marketing Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision (Class 1)** | **Recall (Class 1)** | **F1-Score (Class 1)** |
| **Logistic Regression** | 0.88 | 0.74 | 0.61 | 0.67 |
| **Random Forest** | 0.89 | 0.76 | 0.64 | 0.69 |
| **XGBoost** | 0.90 | 0.77 | 0.65 | 0.70 |
| **Decision Tree** | 0.85 | 0.62 | 0.50 | 0.55 |
| **KNN** | 0.86 | 0.66 | 0.53 | 0.59 |

### 4.2 Comparison Bar Charts

To visually compare model performance, bar charts were created for both the Credit Card Default and Bank Marketing datasets. These charts illustrate three key evaluation metrics for each model:  
- Accuracy (overall correct predictions),  
- Recall for Class 1 (the ability to correctly identify positive cases), and  
- F1-Score for Class 1 (the balance between precision and recall for the positive class).  
  
This visual representation allows for a quick assessment of how well each model performs, especially in detecting the minority class (class 1), which is typically the target of interest in imbalanced classification tasks.  
  
From the charts, it is evident that some models trade off slightly lower overall accuracy for better recall and F1-score in class 1, which may be preferable depending on the business objective (e.g., identifying potential defaulters or likely term deposit subscribers). These insights are helpful in selecting the most suitable model based on the context of the problem.

A graph of a bar chart

AI-generated content may be incorrect.

**Figure 1:** Bar chart comparing the Accuracy, Recall (Class 1), and F1-Score (Class 1) of each model on the Credit Card Default dataset.

A graph of a bar chart

AI-generated content may be incorrect.

**Figure 2:** Bar chart comparing the Accuracy, Recall (Class 1), and F1-Score (Class 1) of each model on the Bank Marketing dataset.

### 4.3 Confusion Matrices

Confusion matrices provide a detailed breakdown of each model's predictions by showing the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).   
This visualisation helps assess how well the classifiers distinguish between the two classes, particularly in the context of class imbalance.   
For this study, confusion matrices were generated for both datasets using all five classifiers.

**Credit Dataset:** The confusion matrices were organised in a 2x3 grid format for clarity see (Figure 3). The imbalance in the data is visible, especially through the high number of true negatives and relatively lower true positives across models like Logistic Regression and KNN.   
Models like XGBoost and Decision Tree demonstrated better performance in identifying true positives.

**Bank Dataset:** Confusion matrices were arranged in a 1x5 horizontal layout for comparison see (Figure 4). Most models yielded a high count of true negatives, while performance on identifying class 1 (term deposit subscriptions) varied.   
Models such as Random Forest and XGBoost performed slightly better in balancing false negatives and positives.

These matrices support the quantitative results reported in classification reports and show model tendencies (e.g., over-predicting the majority class).   
Such visualisations are particularly useful when class 1 is the minority class of interest.

A screenshot of a graph

AI-generated content may be incorrect.

**Figure 3:** Confusion Matrices for the Credit Dataset

A blue and white graph

AI-generated content may be incorrect.

**Figure 4:** Confusion Matrices for the Bank Dataset

### 4.4 ROC-AUC Curves

Receiver Operating Characteristic - Area Under Curve (ROC-AUC) plots were employed to evaluate the discriminative ability of the classification models across both the Bank Marketing and Credit Default datasets. These plots visualise the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at various threshold settings.  
  
In (Figure 5), the ROC curves for the Credit Dataset illustrate that XGBoost and Random Forest outperform the other models, achieving AUC scores of 0.76 and 0.75, respectively. Logistic Regression also performs reasonably with an AUC of 0.71, while Decision Tree lags behind with 0.61, indicating weaker generalisation to unseen data. KNN lands in the middle with 0.70.  
  
In (Figure 6), ROC curves for the Bank Dataset show a similar trend, with Logistic Regression and XGBoost both achieving the highest AUC value of 0.80. Random Forest follows closely at 0.78, and KNN performs moderately with 0.73. Again, the Decision Tree trails with an AUC of 0.62.  
  
These results confirm the consistency of ensemble models like XGBoost and Random Forest, which demonstrate superior discriminative power across both balanced and imbalanced datasets. ROC-AUC is particularly useful in imbalanced scenarios as it remains insensitive to skewed class distributions, making it more informative than accuracy alone.

A graph of a graph showing different colored lines

AI-generated content may be incorrect.

**Figure 5:** ROC-AUC curves for all models on the Credit Dataset.

A graph of a line graph

AI-generated content may be incorrect.

**Figure 6:** ROC-AUC curves for all models on the Bank Dataset.

### 4.5 Precision-Recall Curves

To further assess model performance on the minority class (class 1), particularly important in imbalanced classification problems, Precision-Recall (PR) curves were plotted for both datasets. Unlike ROC curves, which can sometimes present an overly optimistic view in imbalanced datasets, PR curves focus directly on the positive class, making them highly valuable for evaluating how well models identify rare events such as defaults or term deposit subscriptions.  
  
Figures 7 and 8 display the precision-recall curves and corresponding average precision (AP) scores for each model. In both datasets:  
  
- XGBoost and Random Forest achieved the highest average precision, indicating a strong trade-off between identifying class 1 (recall) and doing so accurately (precision).  
- Logistic Regression performed moderately, showing balanced but less optimal precision and recall.  
- Decision Tree lagged behind with low average precision, reflecting its limitations in distinguishing the positive class in both datasets.  
- K-Nearest Neighbors offered moderate results, slightly better than Decision Tree but still underperforming compared to ensemble models.  
  
These curves affirm the findings from the classification reports and ROC-AUC curves, highlighting that ensemble models (Random Forest and XGBoost) are better suited for these classification tasks, especially when dealing with class imbalance.

A graph of a graph

AI-generated content may be incorrect.

**Figure 7:** Precision-Recall Curve for the Credit Dataset. Random Forest and XGBoost models demonstrate superior average precision (AP = 0.53), significantly outperforming the other classifiers on the minority class.

A graph of a graph with different colored lines

AI-generated content may be incorrect.

**Figure 8:** Precision-Recall Curve for the Bank Dataset. XGBoost and Logistic Regression obtained the highest average precision (AP = 0.47), suggesting a stronger ability to identify customers likely to subscribe to a term deposit.

# **5. Discussion**

The comparative analysis of five supervised machine learning models across both the Bank Marketing and Credit Card Default datasets reveals several consistent trends in performance, particularly with respect to class imbalance and model robustness.  
  
Firstly, Random Forest and XGBoost consistently delivered superior performance across multiple evaluation metrics, especially Recall and F1-score for class 1, which represented the minority or positive class in both datasets. These ensemble methods are known for handling non-linear relationships well and for reducing overfitting through techniques like bagging (Random Forest) and boosting (XGBoost), making them highly suitable for imbalanced classification problems [4], [10].  
  
Logistic Regression served as a strong baseline model. Despite its simplicity and assumption of linearity, it performed competitively in terms of Accuracy and Precision, providing a balanced trade-off. Its transparency and efficiency make it ideal in domains where interpretability is essential [3].  
  
Decision Tree underperformed across most metrics, particularly in ROC-AUC and PR AUC scores. Its susceptibility to overfitting, especially in the absence of pruning or regularisation, contributed to its lower effectiveness on both datasets [9].  
  
K-Nearest Neighbors (KNN) showed moderate results but was more sensitive to feature scaling and dataset characteristics. While it managed reasonable performance, its reliance on distance metrics made it less robust compared to ensemble methods [11].  
  
Overall, ensemble models such as Random Forest and XGBoost emerged as the most effective across both datasets, showing consistently strong predictive power, particularly for the minority class.

# **6. Conclusion**

This study applied five supervised machine learning algorithms Logistic Regression, Random Forest, XGBoost, Decision Tree, and K-Nearest Neighbors on two real-world classification datasets: the Bank Marketing Dataset and the Default of Credit Card Clients Dataset. The objective was to compare their performance with a particular emphasis on their ability to detect the minority class (class 1).  
  
Across both datasets, XGBoost and Random Forest consistently emerged as the top-performing models. Their ability to model complex decision boundaries, handle imbalanced data, and maintain high recall and F1-scores for class 1 made them particularly well-suited for these tasks. Logistic Regression also performed reasonably well, offering simplicity and interpretability as its strengths. In contrast, Decision Tree yielded the weakest performance in terms of ROC-AUC and PR AUC, suggesting overfitting and limited generalisation. K-Nearest Neighbors was competitive but showed high sensitivity to data scaling and class distribution.  
  
To further improve model performance and generalisability, several future directions are recommended:  
**Hyperparameter Tuning:** Fine-tuning model parameters using grid search or randomised search can enhance performance.  
**Cross-Validation:** Employing k-fold cross-validation can lead to more robust evaluation and reduce variance.  
**Resampling Techniques:** Methods like SMOTE (Synthetic Minority Over-sampling Technique) can help balance the class distribution.  
**Feature Engineering:** Creating new features or transforming existing ones based on domain knowledge may uncover hidden patterns and improve model performance.  
  
Overall, this comparative analysis underscores the importance of model selection and evaluation strategies in machine learning workflows, especially when dealing with real-world, imbalanced classification problems.

# **7. References**

[1] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[2] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 785–794.

[3] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[4] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.

[5] W. McKinney, “Data Structures for Statistical Computing in Python,” in Proc. 9th Python in Science Conf., 2010, pp. 51–56.

[6] UCI Machine Learning Repository, “Bank Marketing Dataset,” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

[7] UCI Machine Learning Repository, “Default of Credit Card Clients Dataset,” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[8] Machine Learning (HDSDA\_SEP24), National College of Ireland, 2025. [Course materials Weeks 2, 4, 5, and 7].

[9] Scikit-learn Documentation. [Online]. Available: https://scikit-learn.org/stable/documentation.html

[10] Towards Data Science, “Precision-Recall vs ROC Curves,” Medium. [Online]. Available: https://towardsdatascience.com/precision-recall-vs-roc-curve-5f26c78fe3ec

[11] StatQuest with Josh Starmer (YouTube Channel) – ML Concepts Explained. [Online]. Available: https://www.youtube.com/c/joshstarmer

[12] UCI Machine Learning Repository. [Online]. Available: <https://archive.ics.uci.edu/ml/index.php>