The Knowledge Hub Universities

partnered with Coventry University

Machine learning Report

Full name: Ahmed Adel Alkaf

ID:aa2201459

https://github.com/1zr7/KH6001CMD_ML.git

Table of content

Contents

Table of figures

Abstract

Customer churn remains a major challenge for subscription-based businesses such as fitness centres. This project develops a supervised machine learning system to predict member churn using a real-world gym dataset. The approach includes exploratory data analysis, structured preprocessing, incorporation of unsupervised feature engineering, implementation of multiple classification algorithms, performance evaluation, hyperparameter tuning, and model explainability using SHAP. Results show that a Multi-Layer Perceptron (MLP) achieves the best accuracy 95.13%, demonstrating strong generalisation and practical applicability for early churn intervention.

1. Introduction

1.1 Problem statement

Gym membership churn, defined as customers discontinuing their membership, is a significant operational and financial concern. Predicting churn enables proactive retention strategies and optimises resource allocation. This study addresses the following question:

Can machine learning accurately predict whether a gym member will churn based on demographic, behavioural, and contractual characteristics?

The problem is formulated as a **binary classification task**, with the target variable:

- **Churn = 1** (customer left)
- **Churn = 0** (customer retained)

This makes classification algorithms an appropriate methodological choice

1.2 Existing Approaches

Existing work on gym membership churn prediction mainly focuses on behavioural indicators such as attendance frequency, contract duration, and additional spending. Gradient Boosting is widely used and has shown strong performance; for example, a Kaggle implementation by Magdy (2023) achieved 93% accuracy on a similar gym churn dataset.

In academic research, Aldosary and Alrashdan (2021) applied an MLP neural network incorporating the psychological concept of habit formation and achieved 92.1% accuracy. Their work highlights the importance of behavioural consistency and routine formation in predicting retention.

1.3 Contribution and Differences from Existing Work

This project extends existing work in two key ways:

1. **Broader Model Comparison:**
   Unlike prior approaches that focus mainly on a single model (Gradient Boosting or MLP), this study evaluates multiple algorithm families—Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, and MLP—under consistent preprocessing and cross-validation.
2. **Higher Predictive Performance:**
   The best-performing models in this project achieved **up to 95.25% accuracy** and **0.982 ROC-AUC**, outperforming public Kaggle implementations and matching or exceeding academic results.

2. Dataset Description and Pre-processing

The dataset used in this project was obtained from Kaggle (Vinueza, 2023) and contains operational and behavioural information about gym members, with the objective of predicting whether a customer will churn. The dataset consists of 4,000 observations and 14 variables, all of which are fully complete with no missing values, making it suitable for supervised machine learning tasks without requiring data imputation.

2.1 Dataset Overview

A summary of the dataset structure, obtained using df.info(), indicates that the data contains a mixture of integer and floating-point numerical features. All variables are numeric, which simplifies pre-processing and model integration. The dataset includes 13 predictor variables and 1 binary target variable (Churn), where:

- **Churn** = 1 indicates a customer who has terminated their membership.
- **Churn** = 0 represents an active customer.

Table 1 summarises the main features included in the dataset:

Feature Name	Description	type
gender	Indicates the client's gender (0 = male, 1 = female).	Binary
Near_Location	Whether the client lives near the gym (1 = near).	Binary
Partner	Whether the client enrolled through a corporate partnership program. (1=yes)	Binary
Promo_friends	Indicates if the client joined through a "refer-a-friend" promotion. (1 = yest)	Binary
Phone	Whether the client provided a valid phone number. (1 = yes)	Binary
Contract_period	Length of the membership contract in months. (1,6,12)	categorical
Group_visits	Indicates if the client participates in group classes. (1= yes)	Binary
Age	Age of the client in years.	Numerical (Continuous)
Avg_additional_charges_total	Average monthly spending on extra services (café, products, etc.).	Numerical (Continuous)
Month_to_end_contract	Number of months left until the current contract expires.	Numerical (Discrete)

Lifetime	Total duration the client has been a member (in months).	Numerical (Discrete)
Avg_class_frequency_total	Average number of classes attended weekly across membership history.	Numerical (Continuous)
Avg_class_frequency_current_month	Average class attendance during the current month.	Numerical (Continuous)
Churn	Target variable: 1 = churned, 0 = retained.	Binary / Target Variable

2.2 Descriptive Statistics

The descriptive statistics highlight several important characteristics of the dataset:

- **Binary variables** such as Near_Location, Partner, Promo_friends, and Phone all have means between 0.30 and 0.90, indicating varying levels of engagement and demographic patterns.
- **Age** ranges from 18 to 41 years, with a mean of approximately 29, suggesting a relatively young customer base.
- **Contract-related features**, such as Contract_period and Month_to_end_contract, span from 1 to 12 months, reflecting the gym's 3 subscription models.
- **Financial behaviour**, represented by Avg_additional_charges_total, shows a wide variance (mean ≈ 147, std ≈ 96), indicating different levels of optional spending.
- **Activity frequency features**, such as Avg_class_frequency_total and Avg_class_frequency_current_month, range from 0 to approximately 6 classes per week, providing useful signals about customer engagement.
- The target variable, **Churn**, has a mean of 0.265, indicating that 26.5% of customers churned, which introduces mild class imbalance but does not require resampling.

2.3 Dataset Suitability for Machine Learning

Overall, the dataset meets the coursework requirements by:

- Containing more than 1,000 samples (4,000 total).
- Including multiple numeric and categorical-like variables (binary indicators), which allow demonstration of feature engineering, scaling, model selection, and interpretability.
- Having no missing data, which enables a clean modelling process while still allowing for meaningful preprocessing work such as outlier detection, scaling, and feature transformation.

The dataset includes both behavioural and contractual attributes, which are commonly used in churn modelling literature and provide sufficient complexity for evaluating a wide range of machine learning algorithms, including tree-based models, linear models, and neural networks.

2.4 Limitations

A few limitations should be noted:

- All categorical features are encoded numerically, but they are binary, which restricts certain types of categorical interpretation.
- The dataset lacks explicit qualitative features such as motivation, satisfaction, or past complaints, which are often strong churn predictors in real-world scenarios.
- The dataset represents a single gym chain, meaning external validity may be limited.

Despite these constraints, the dataset is highly appropriate for demonstrating the full lifecycle of machine learning model development .

3. Data Analysis and Pre-processing

This section presents the exploratory data analysis (EDA) and preprocessing steps performed prior to model training. The objective of these steps is to understand the underlying patterns within the dataset, detect potential issues such as imbalanced classes or outliers, and prepare the data for machine learning algorithms.

3.1 Missing Values Analysis

All 14 variables showed complete entries, with no missing values detected. This eliminates the need for imputation and ensures that model performance is not affected by incomplete observations.

3.2 Target Variable and Class Distribution

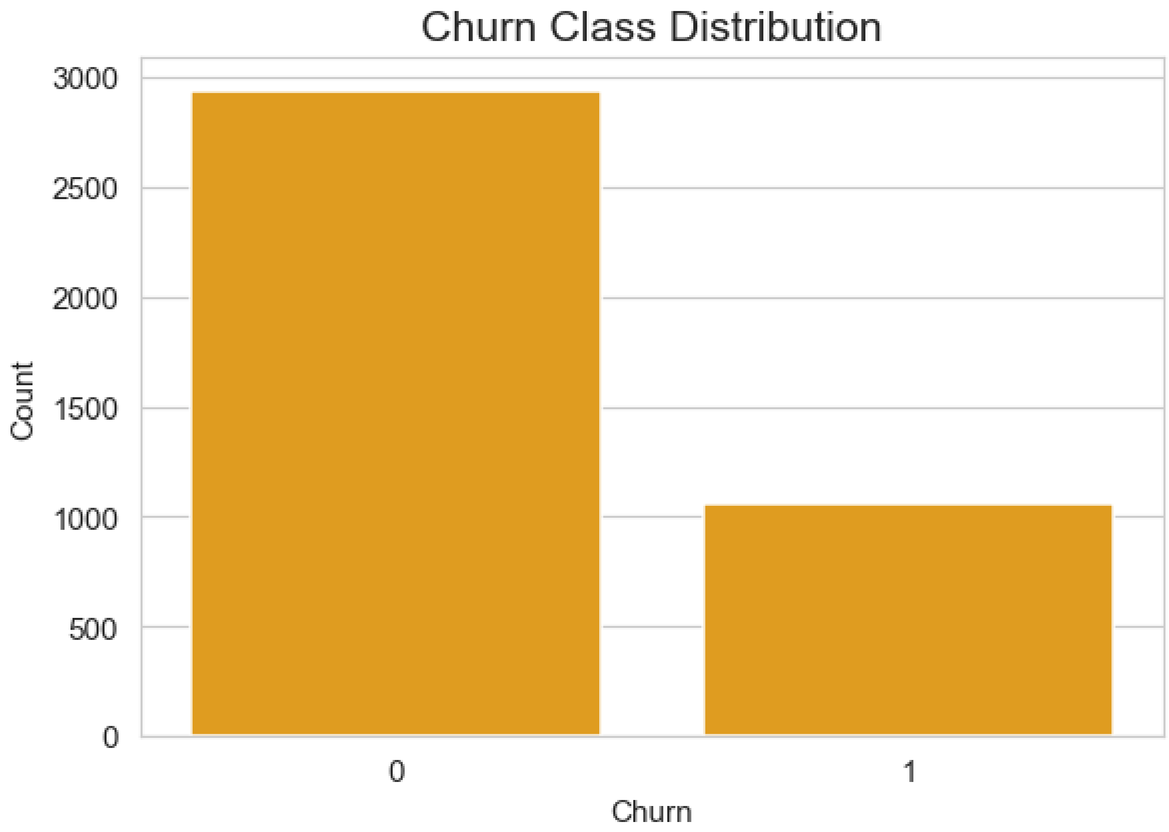


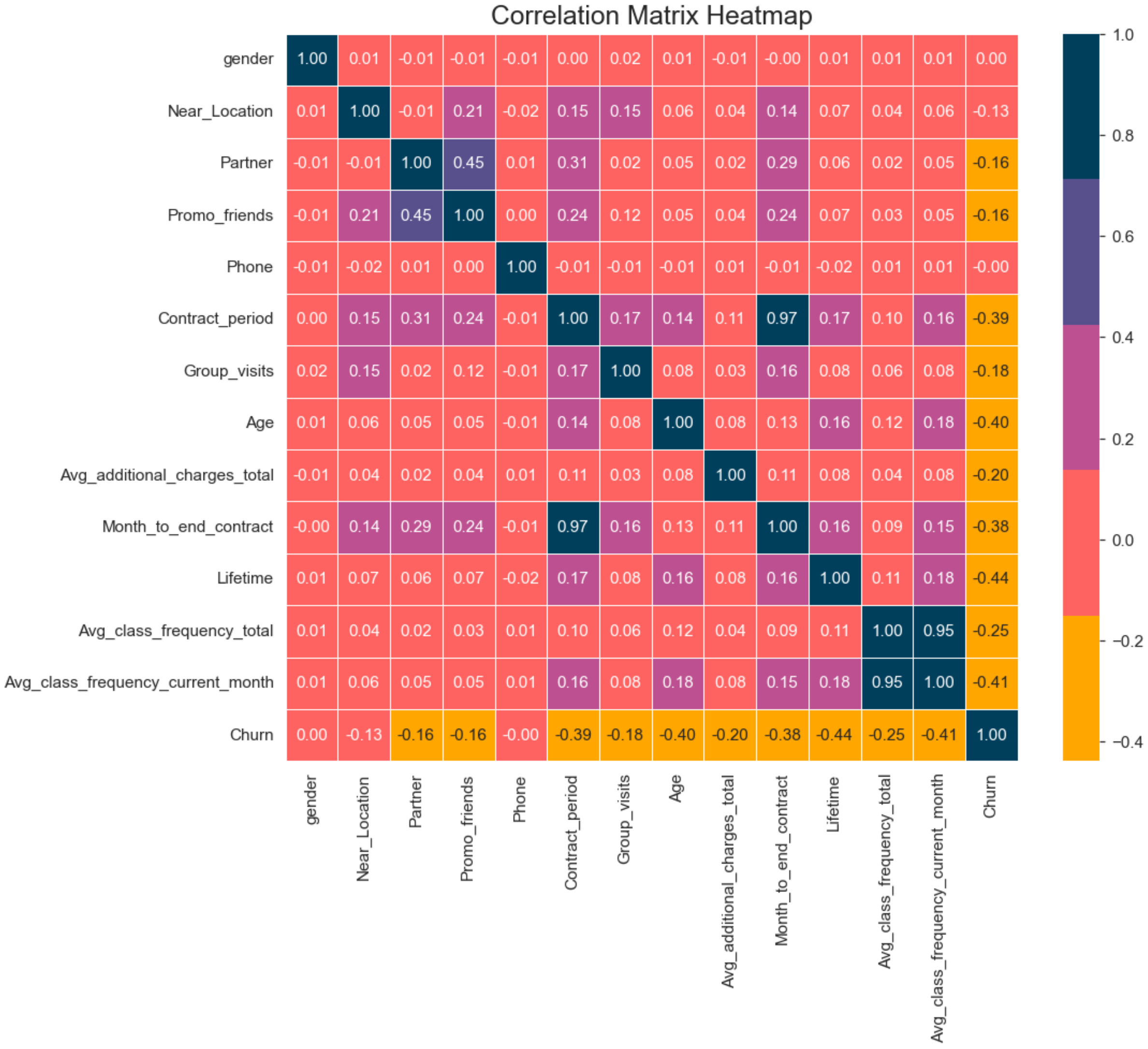
Figure 1 Class distribution analysis

Class distribution analysis revealed:

- 0 (Not Churned): 2,939 samples
- 1 (Churned): 1,061 samples

This corresponds to a churn rate of 26.5%, indicating a moderately imbalanced dataset. While not severe enough to require resampling techniques such as SMOTE, the imbalance is still relevant when selecting performance metrics (e.g., F1-score, ROC-AUC) and evaluating model bias.

3.3 Correlation Analysis



	gender	Near_Location	Partner	Promo_friends	Phone	Contract_period	Group_visits	Age	Avg_additional_charges_total	Month_to_end_contract	Lifetime	Avg_class_frequency_total	Avg_class_frequency_current_month	Churn
gender	1.00	0.01	-0.01	-0.01	-0.01	0.00	0.02	0.01	-0.01	-0.00	0.01	0.01	0.01	0.00
Near_Location	0.01	1.00	-0.01	0.21	-0.02	0.15	0.15	0.06	0.04	0.14	0.07	0.04	0.06	-0.13
Partner	-0.01	-0.01	1.00	0.45	0.01	0.31	0.02	0.05	0.02	0.29	0.06	0.02	0.05	-0.16
Promo_friends	-0.01	0.21	0.45	1.00	0.00	0.24	0.12	0.05	0.04	0.24	0.07	0.03	0.05	-0.16
Phone	-0.01	-0.02	0.01	0.00	1.00	-0.01	-0.01	-0.01	0.01	-0.01	-0.02	0.01	0.01	-0.00
Contract_period	0.00	0.15	0.31	0.24	-0.01	1.00	0.17	0.14	0.11	0.97	0.17	0.10	0.16	-0.39
Group_visits	0.02	0.15	0.02	0.12	-0.01	0.17	1.00	0.08	0.03	0.16	0.08	0.06	0.08	-0.18
Age	0.01	0.06	0.05	0.05	-0.01	0.14	0.08	1.00	0.08	0.13	0.16	0.12	0.18	-0.40
Avg_additional_charges_total	-0.01	0.04	0.02	0.04	0.01	0.11	0.03	0.08	1.00	0.11	0.08	0.04	0.08	-0.20
Month_to_end_contract	-0.00	0.14	0.29	0.24	-0.01	0.97	0.16	0.13	0.11	1.00	0.16	0.09	0.15	-0.38
Lifetime	0.01	0.07	0.06	0.07	-0.02	0.17	0.08	0.16	0.08	0.16	1.00	0.11	0.18	-0.44
Avg_class_frequency_total	0.01	0.04	0.02	0.03	0.01	0.10	0.06	0.12	0.04	0.09	0.11	1.00	0.95	-0.25
Avg_class_frequency_current_month	0.01	0.06	0.05	0.05	0.01	0.16	0.08	0.18	0.08	0.15	0.18	0.95	1.00	-0.41
Churn	0.00	-0.13	-0.16	-0.16	-0.00	-0.39	-0.18	-0.40	-0.20	-0.38	-0.44	-0.25	-0.41	1.00

Figure 2 correlation matrix Heatmap

A Pearson correlation heatmap was computed to explore linear relationships among the features. Key insights include:

- Strong positive correlation between
  - Avg_class_frequency_total and
  - Avg_class_frequency_current_month
    This is expected, as weekly activity in the current month contributes to the overall average.
- Moderate correlation between financial spending (Avg_additional_charges_total) and engagement frequency features, suggesting that frequent gym users may purchase more additional services.
- Negative correlations with Churn were observed for visit frequency and lifetime features, indicating that active, long-term clients are less likely to churn.

3.5 Outlier Detection

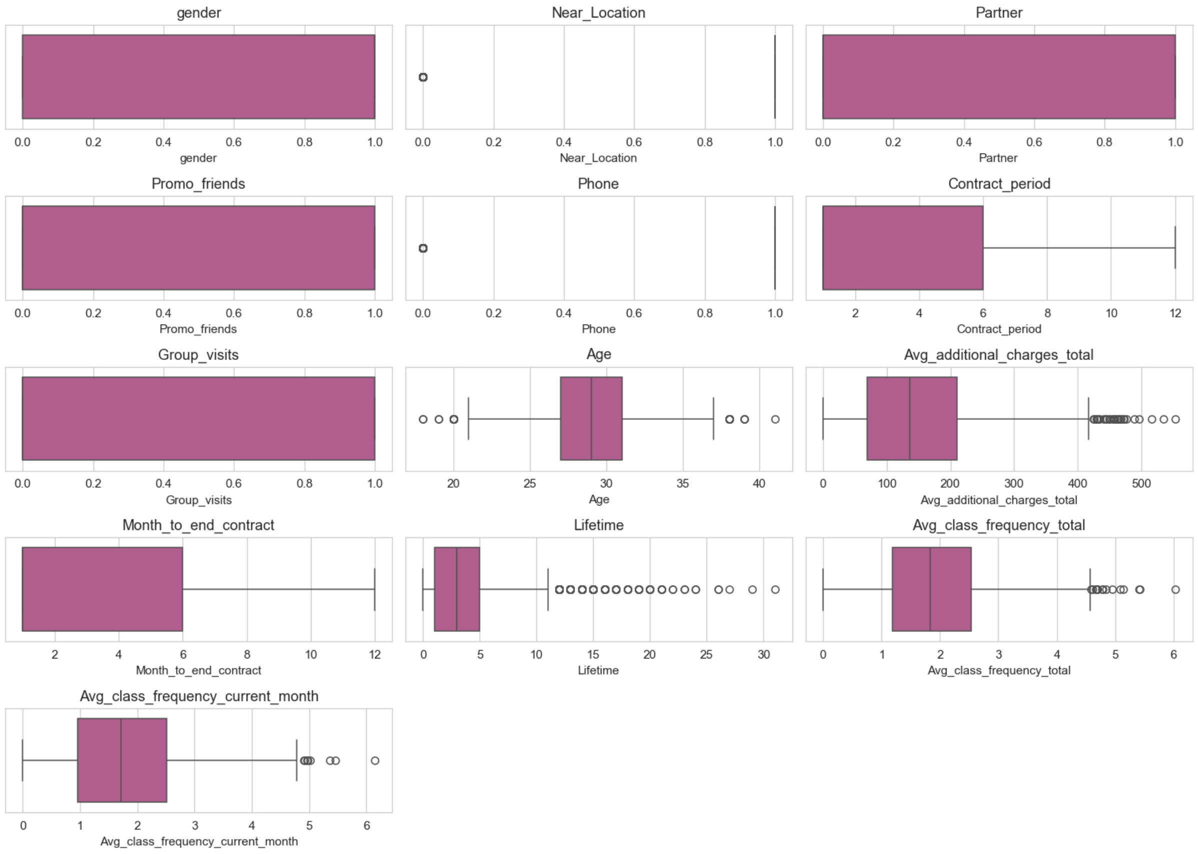


Figure 3 outlier detection

Boxplots were generated for continuous variables to identify extreme observations. Several features exhibited outliers, particularly:

- Avg_additional_charges_total (high-spending customers)
- Activity metrics (very frequent class attendees)

These outliers were retained in the dataset because:

1. They represent valid customer behaviour rather than data-entry errors.
2. Many machine learning models used (e.g., tree-based algorithms) are robust to outliers.
3. Removing such observations might hide important patterns relevant to churn prediction.

3.7 Unsupervised Feature Engineering Using K-Means

K-Means clustering was applied to the scaled numerical features to identify latent customer segments. The optimal number of clusters was determined using elbow and silhouette analysis, resulting in k = 2. The two clusters contained 2,194 members and 1,806 members, respectively.

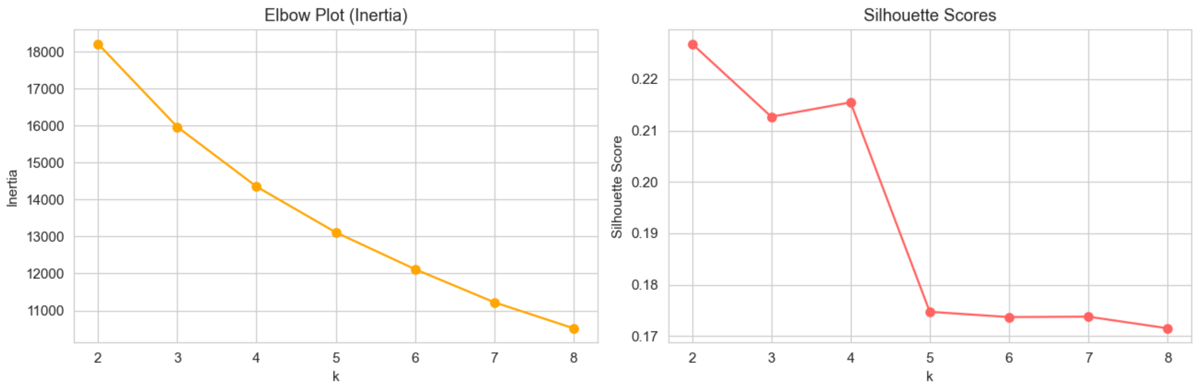


Figure 4 Kmean clustring plots

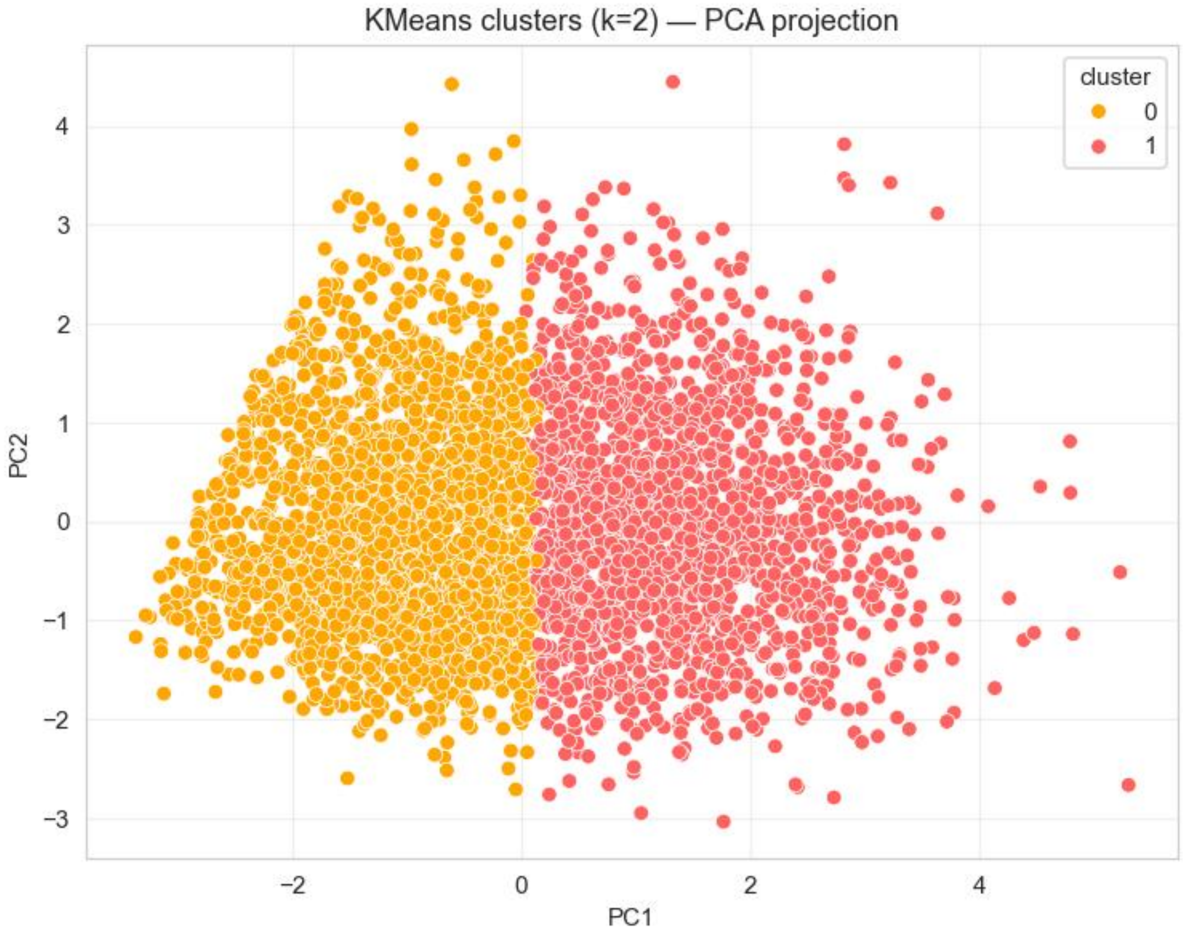


Figure 5 PCA pojection

Cluster profiles showed clear behavioural differences:

- **Cluster 0** represented younger members with lower spending, shorter contract periods, lower lifetime, and lower class attendance.

- **Cluster 1** consisted of older members with higher additional charges, longer membership duration, and higher class frequency.

These patterns indicate distinct engagement levels across the customer base. The cluster label (kmeans_cluster) was added as a new categorical feature and used in all supervised models to capture segment-level behavioural structure.

3.7 Categorical Feature Analysis

Although all variables are numerically encoded, several features function as binary categorical features, including:

- gender
- Near_Location
- Partner
- Promo_friends
- Phone
- Group_visits

This confirms that:

- The dataset does not contain classical multi-category string features.
- Most features are binary indicators.
- **Contract_period** behaves as a low-cardinality discrete feature, but still represents a meaningful numerical progression (1,6,12 months).

4. Data Pre-processing Pipeline

4.1 Feature Categorisation

Based on the dataset structure, the features were grouped as follows:

- **Numerical Features:**
  Age, Avg_additional_charges_total, Month_to_end_contract, Lifetime, Avg_class_frequency_total, Avg_class_frequency_current_month
- **Binary Categorical Features:**
  gender, Near_Location, Partner, Promo_friends, Phone, Group_visits
- **Multi-class Categorical Feature:**
  Contract_period

All variables were assigned to one of these categories, and no missing values were present.

4.2 Train–Test Split

The dataset was split into training and testing sets using stratified sampling to preserve the original churn ratio (73.5% non-churn, 26.5% churn).

- **Training**: 3,200 samples
- **Testing**: 800 samples

Stratification ensures reliable evaluation on an imbalanced dataset.

4.3 Preprocessing Pipeline

A unified preprocessing pipeline was implemented using **scikit-learn's ColumnTransformer**, with three transformations:

1. **Standard Scaling** for numerical features
   Ensures comparable feature ranges, benefiting models like logistic regression, neural networks, and KNN.
2. **One-Hot Encoding** for the multi-class feature
   Prevents the model from treating contract duration as a numeric scale.
3. **Passthrough** for binary categorical features
   These are already encoded as 0/1 and require no additional transformation.

The pipeline is applied inside each model to avoid data leakage during cross-validation and hyperparameter tuning.

5. Machine Learning models

5.1 Baseline Models

Two baseline classifiers were implemented: Logistic Regression and a Decision Tree. Logistic Regression provides a strong linear benchmark, while the Decision Tree offers an interpretable non-linear alternative.

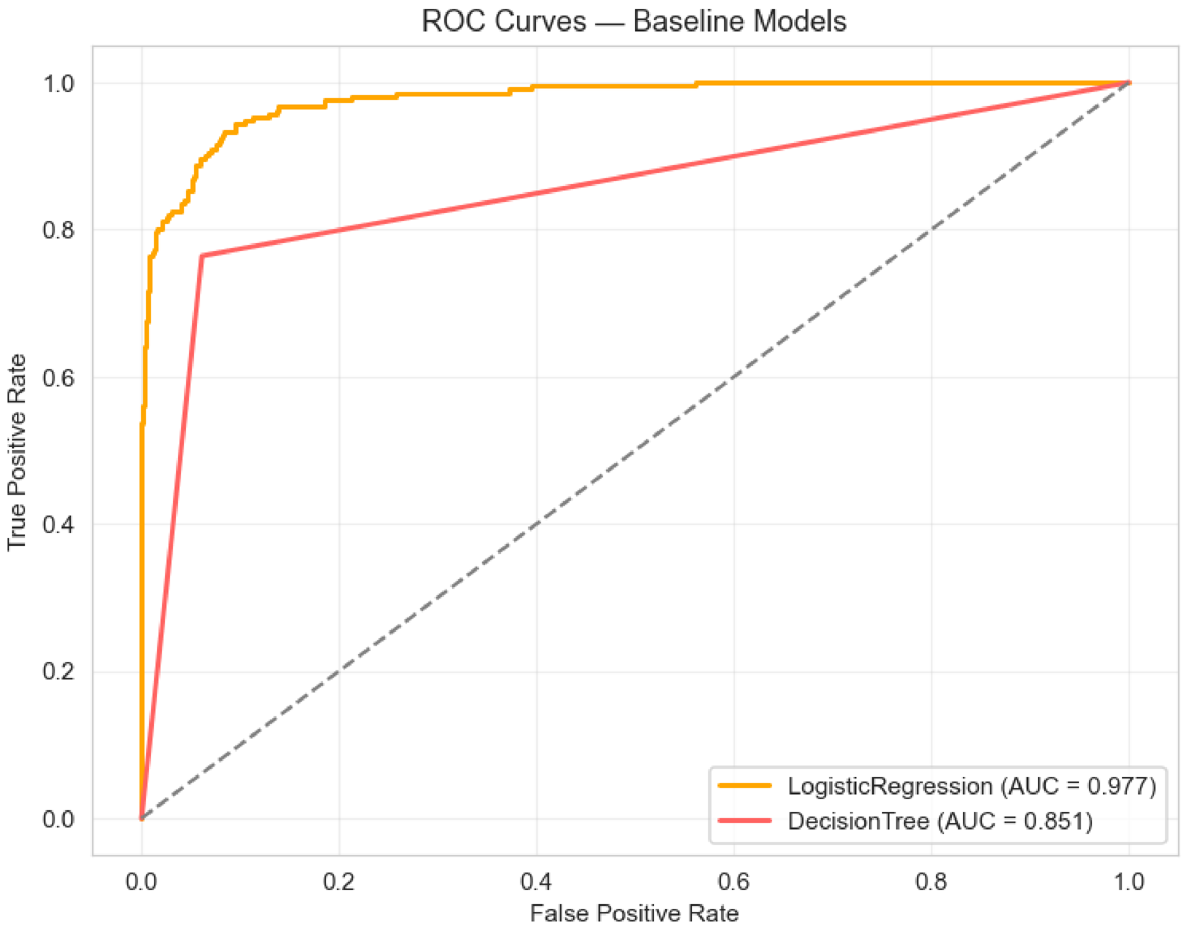


Figure 6 ROC Curves- baseline models

Logistic Regression achieved substantially higher discriminative performance, suggesting that the dataset is well separated under a linear decision boundary.

Confusion Matrix:

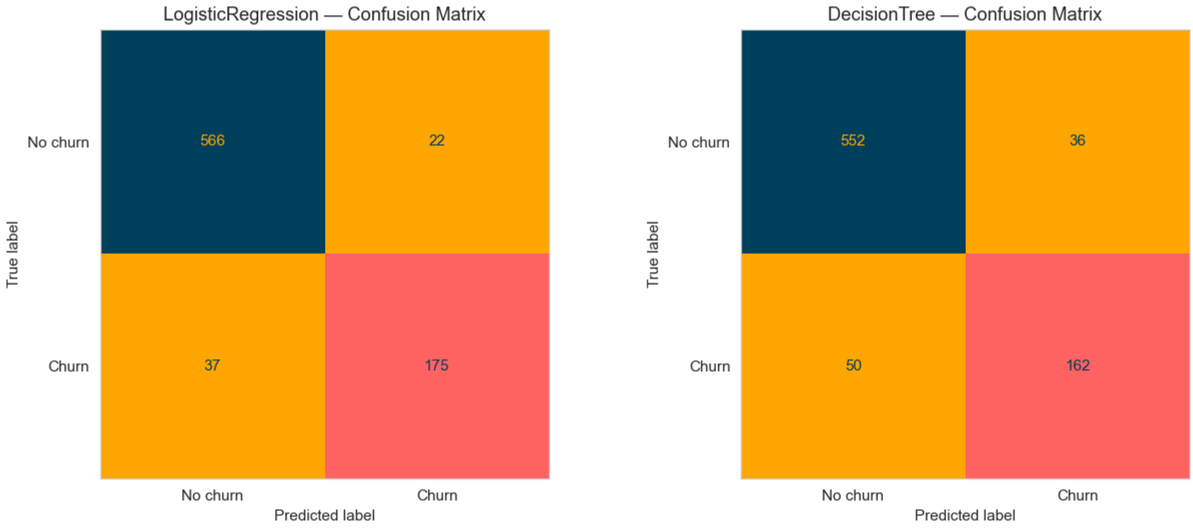


Figure 7 confusion matrix baseline models

Logistic Regression performed strongly on both cross-validation and test evaluation, indicating good generalisation.

5.2 Advanced Models

Three advanced models were evaluated: Random Forest, Gradient Boosting, and a Multi-Layer Perceptron (MLP). These models capture complex, non-linear relationships and typically outperform simpler baselines on structured data.

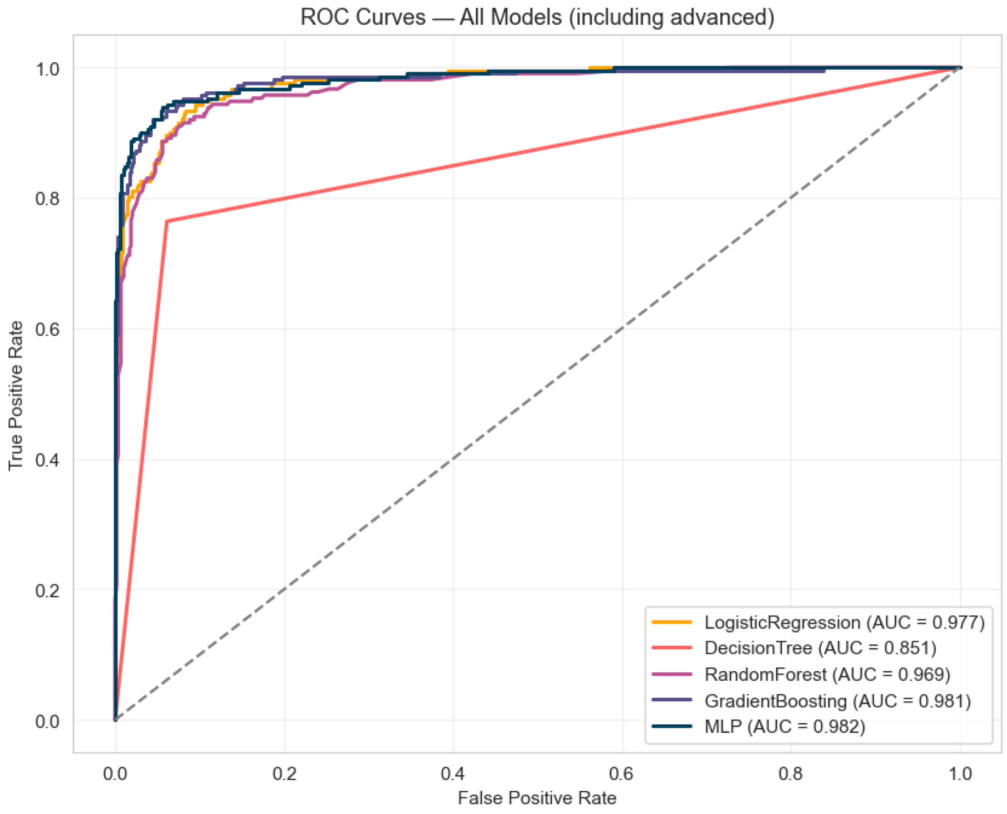


Figure 8 all models ROC Curve

The MLP achieved the highest cross-validated ROC-AUC, indicating superior ability to learn interactions between behavioural and contract-related features.

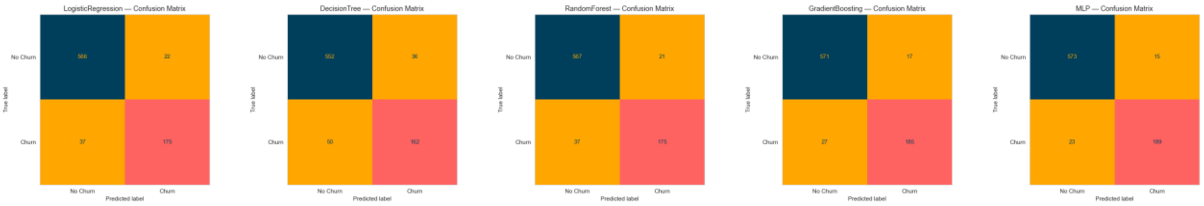


Figure 9 all models confusion matrix

5.3 Test Set Evaluation

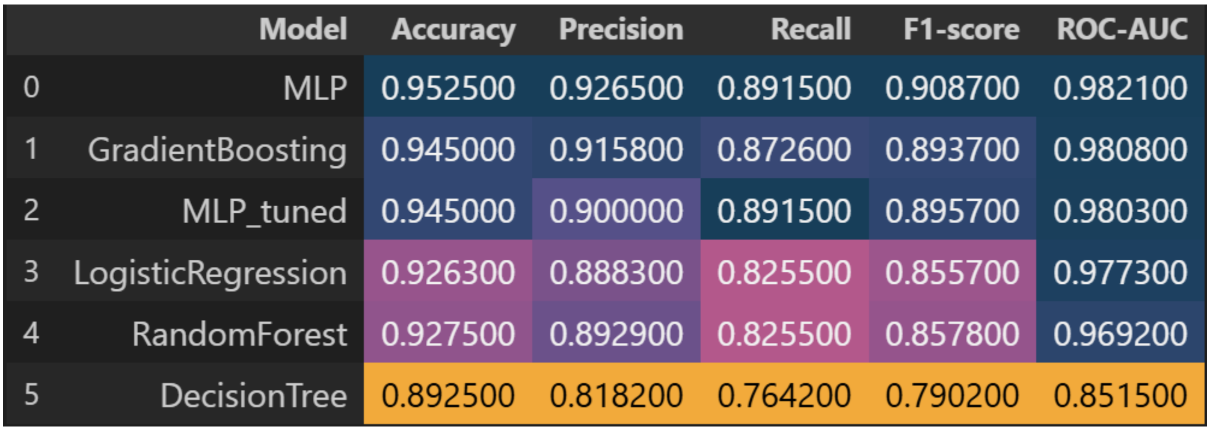
Gradient Boosting produced the strongest test-set performance among all tree-based models, correctly identifying more churners (higher TP) compared to Logistic Regression and Random Forest.

MLP (Neural Network)

The MLP achieved the highest cross-validated ROC-AUC (0.9856), confirming that neural networks can model non-linear churn-related patterns effectively. Its final test performance was consistent with the boosted methods and showed no signs of overfitting.

6. Model Performance Comparison

A consolidated comparison of all models is shown in Table 6. The MLP achieved the best overall performance across most metrics, followed closely by Gradient Boosting. Logistic Regression and Random Forest performed similarly, while the Decision Tree underperformed relative to the other models.



	Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
0	MLP	0.952500	0.926500	0.891500	0.908700	0.982100
1	GradientBoosting	0.945000	0.915800	0.872600	0.893700	0.980800
2	MLP_tuned	0.945000	0.900000	0.891500	0.895700	0.980300
3	LogisticRegression	0.926300	0.888300	0.825500	0.855700	0.977300
4	RandomForest	0.927500	0.892900	0.825500	0.857800	0.969200
5	DecisionTree	0.892500	0.818200	0.764200	0.790200	0.851500

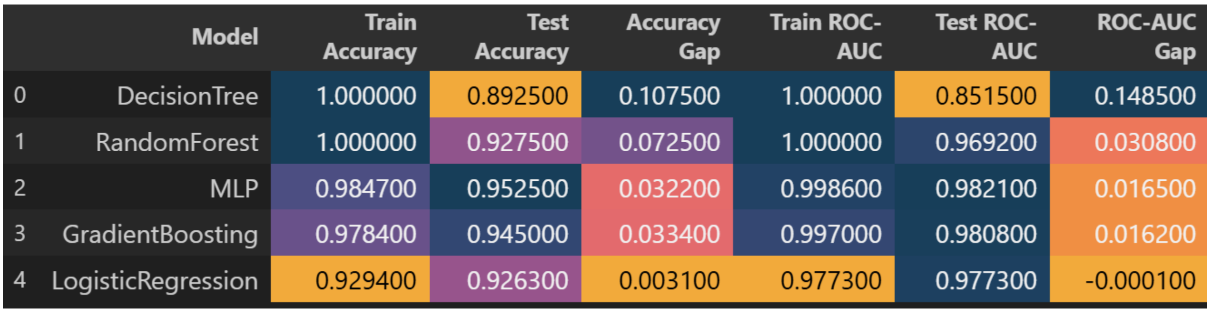
Figure 10 performance comparrison of all models

6.1 Summary

- **MLP** achieved the highest Accuracy, F1-score, and ROC-AUC, making it the strongest model overall.
- **Gradient Boosting** delivered competitive performance with slightly lower recall.
- **Logistic Regression** performed well as a simple baseline.
- **Decision Tree** showed the weakest results, confirming its tendency to overfit.

7.Model Generalisation and Overfitting Diagnostics

Overfitting was evaluated by comparing each model's training and test performance. Figure 9 summarises the accuracy and ROC-AUC gaps.



	Model	Train Accuracy	Test Accuracy	Accuracy Gap	Train ROC-AUC	Test ROC-AUC	ROC-AUC Gap
0	DecisionTree	1.000000	0.892500	0.107500	1.000000	0.851500	0.148500
1	RandomForest	1.000000	0.927500	0.072500	1.000000	0.969200	0.030800
2	MLP	0.984700	0.952500	0.032200	0.998600	0.982100	0.016500
3	GradientBoosting	0.978400	0.945000	0.033400	0.997000	0.980800	0.016200
4	LogisticRegression	0.929400	0.926300	0.003100	0.977300	0.977300	-0.000100

Figure 11 Overfitting Diagnostic Table

7.2 learning Curve Interpretation

The learning-curve behaviour is consistent with the overfitting diagnostics:

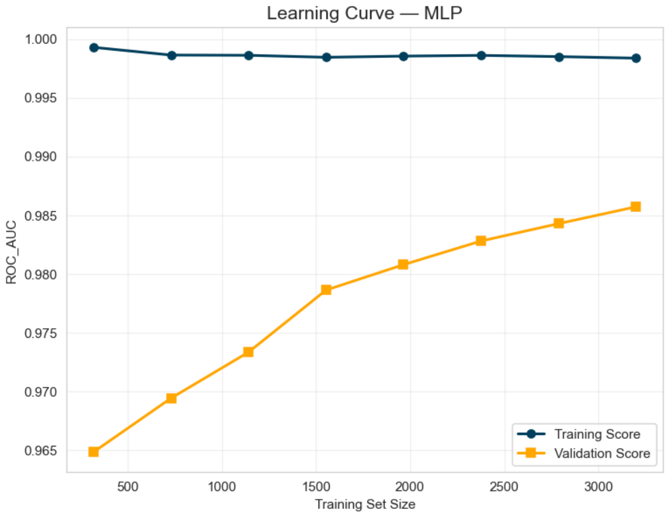


Figure 12 learn curve MLP

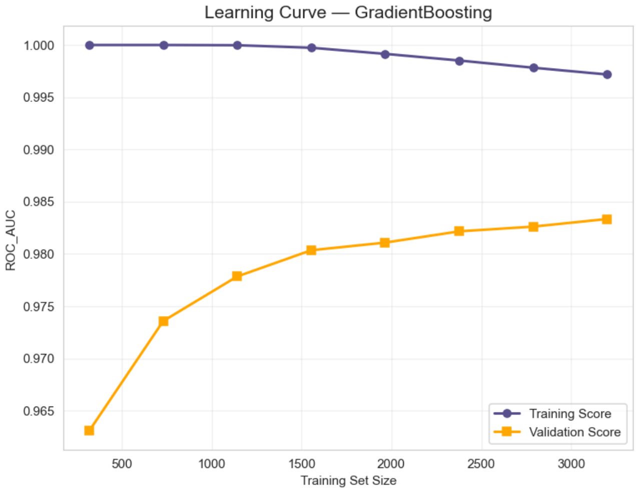


Figure 13 learning curve gradent boosting

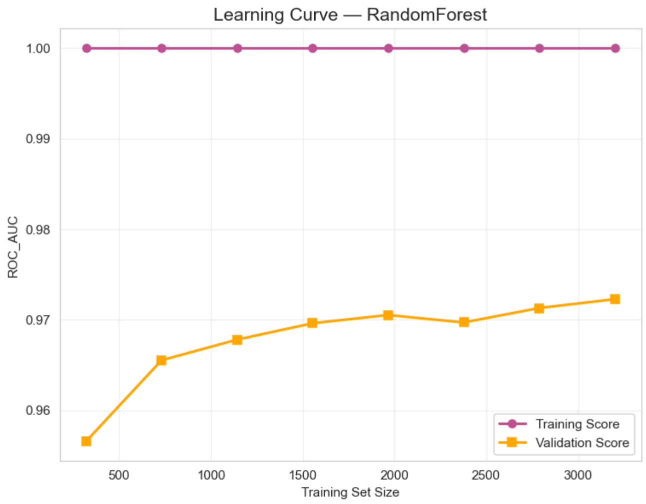


Figure 14 learning curve random forest

- **Random Forest** fits the training data perfectly but performs much worse on the test set, confirming high variance and typical overfitting behaviour.
- **Gradient Boosting** display small train–test gaps, suggesting good generalisation. Their curves flatten smoothly, showing that additional data would produce only minor gains.
- The **MLP** improves as training size increases, which aligns with its slightly wider but still controlled performance gap.

references

Aldosary, M., & Alrashdan, A. (2021, March). Churn Prediction for Gym Members Using Artificial Neural Networks Assisted with The Psychological Concept of Habit Formation in The Fitness Industry. In 11th Annual International Conference on Industrial Engineering and Operations Management, https://doi.org/10.46254/AN11.20210720.

Magdy, S. (2023). Gym classification (93% acc) [Computer software]. Kaggle. https://www.kaggle.com/code/shamsmagdy12/gym-classification-93-acc

Vinueza, A. (2023). Gym customers features and churn [Dataset]. Kaggle. https://www.kaggle.com/datasets/adrianvinueza/gym-customers-features-and-churn/data