ST3189 Machine Learning Coursework

Name: Joline Chua Yong Jia
Student Number: 220456807

Total Pages: 10 Pages (Excluding Cover Page and References Page)

1. Introduction

Machine Learning is analysing and examining data and constructing algorithms and statistical models to improve the performance of systems on specific tasks. In this coursework, we will be covering two types: supervised learning is building and training a statistical model to predict an outcome using input variables, and unsupervised learning focuses on training a statistical model using un-defined data to find hidden patterns and correlations.

This project consists of the following three main tasks:

- **Unsupervised Learning**: which consists of identifying homogeneous population groups or dimension reduction techniques.
- **Regression**: which consists of continuous target variable(s).
- **Classification**: which consists of categorical target variable(s).

2. Mobile Device Usage Dataset

Unsupervised learning identifies structures in data without predefined labels. It is widely used for user segmentation, anomaly detection and market analysis. In this section, we will be exploring clustering techniques to analyse mobile device user behaviour and identify homogeneous population groups.

2.1 Substantive Issue

With the increasing smartphone penetration, understanding user behaviour is crucial for optimising application performance and personalising services. For telecom providers and app developers, understanding user types enables personalised service offerings, targeted marketing, and system optimisation. This section seeks to classify mobile users based on their engagement levels and device usage patterns.

2.2 Methodology

Principal Component Analysis (PCA), **Hierarchical Clustering** with **Manhattan (Average Linkage)** and **Euclidean (Ward Linkage)** distances are techniques used to understand the patterns and relationships between variables while clustering mobile device users. PCA was applied to reduce feature dimensions while retaining a percentage of the total variance, facilitating better cluster separation and visualisation.

2.3 Dataset and Variables

A total of 700 rows and 11 columns are included in the dataset that reflects mobile user behaviours based on various different device interaction metrics.

Variable	Data Type	Description
App Usage Time (min/day)	Float	Total time spent on mobile applications a day.
Screen On Time (h/day)	Float	Duration the screen remains active daily.
Battery Drain (mAh/day)	Float	Power consumption a day.
Number of Apps Installed	Integer	Total number of apps installed on the device.
Data Usage (MB/day)	Float	Mobile data consumption a day.

Table 1: Mobile Device Usage Dataset Description

There is no missing and duplicate data, the dataset was standardised using **StandardScaler** to ensure uniformity across numerical features.

2.4 Analysis

PCA is a dimensionality reduction technique to identify relationships amongst variables. From the dataset, we performed PCA and the first two principal components (PC1 and PC2) were selected as they accounted for majority (~97%) of the variance.

	Principal Component	Standard Deviation	Proportion of Variance
0	PC1	2.189778	0.957655
1	PC2	0.276717	0.015293
2	PC3	0.230717	0.010631
3	PC4	0.210745	0.008870
4	PC5	0.194450	0.007551

Table 2: Principal Components Analysis Results

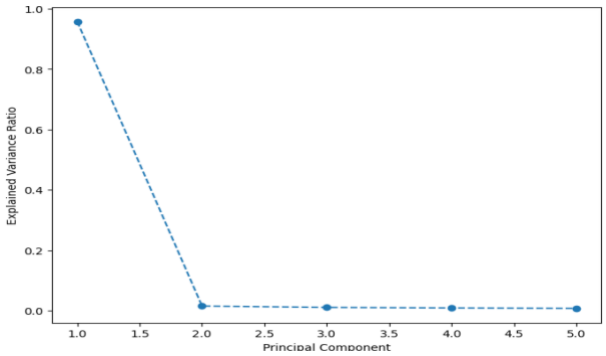


Figure 1: Scree Plot of PCA

Clustering in a PCA-reduced space helps to mitigate noise from correlated features and improves interpretability. In order to determine the most optimal number of clusters, we use the **Elbow Method** and **Silhouette Score** to obtain an ideal cluster count. The Elbow Method plot shows a distinct "elbow" at k = 4, while the Silhouette Scores for **Hierarchical** and **K-Means** clustering confirms that 4 clusters provide the best separation.

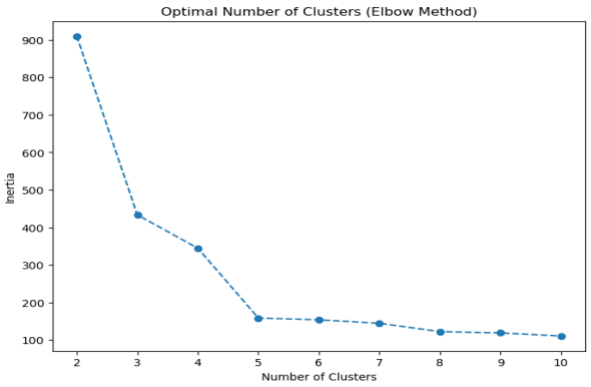


Figure 2: Elbow Method for Optimal Clusters

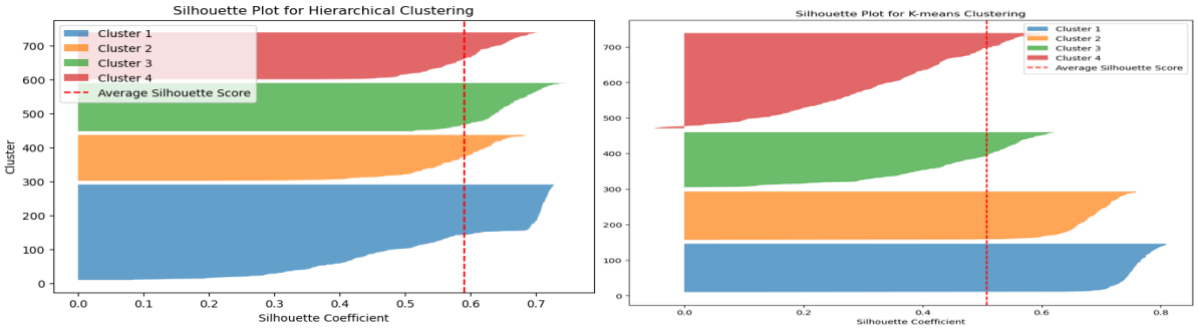


Figure 3: Silhouette Score for Clustering

Silhouette score ranges from -1 to 1, where the higher value, the clearer the cluster separation. The silhouette score for hierarchical clustering is higher (0.5906) than for K-Means (0.5079), suggesting more distinct clusters in the hierarchical approach.

The hierarchical clusters were analysed based on key behavioural features. Standardized feature averages highlight differences between groups.

Hierarchical Cluster	App Usage Time (min/day)	Screen On Time (hours/day)	Battery Drain (mAh/day)	Number of Apps Installed	Data Usage (MB/day)
Cluster 1	-0.9807	-0.9722	-1.0362	-1.0302	-0.9351
Cluster 2	1.5264	1.5791	1.4365	1.4325	1.6329
Cluster 3	-0.2018	-0.1033	-0.0123	-0.0253	-0.1683
Cluster 4	0.7038	0.5337	0.7094	0.7146	0.4726

Figure 4: Hierarchical Cluster Characteristics

Hierarchical Cluster	Age
Cluster 1	38.44
Cluster 2	38.18
Cluster 3	38.68
Cluster 4	38.68

Figure 5: Hierarchical Cluster Demographics

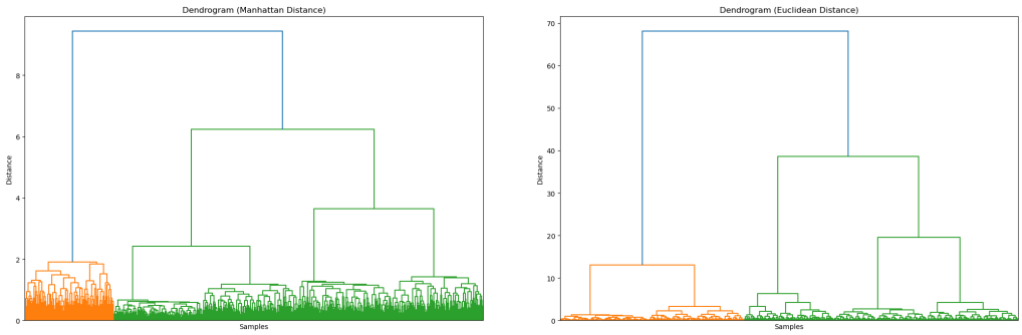


Figure 6: Dendrogram Visualisations

Manhattan linkage (0.7956) preserved the pairwise distances from the original data better than Ward (Euclidean) linkage (0.7353), the cophenetic correlation coefficients suggest Manhattan linkage is a better approach for clustering.

2.5 Results

RQ1: How can homogeneous user groups (clusters) be identified based on mobile usage patterns, and what distinguishes heavy users from moderate or low users?

Using hierarchical clustering, four distinct user groups were identified based on mobile usage behaviours. These groups were characterised by app usage time, screen-on duration, battery consumption, number of installed apps and data usage. The clusters distinguished heavy users, who showed high app usage and data consumption, from moderate and low users, who showed significantly lower engagement levels.

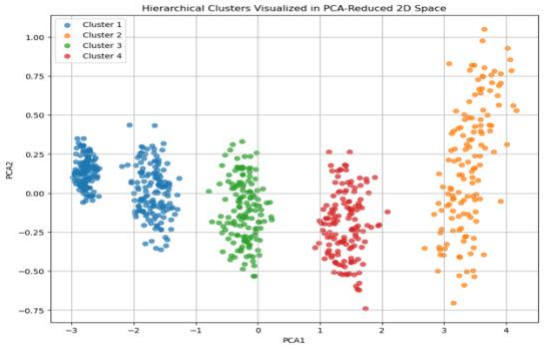


Figure 7: PCA Visualisation (Hierarchical Clustering)

Cluster 1 (blue) are the low engagement users who exhibit the lowest app usage, screen time and battery drain. Cluster 2 (yellow) are the high engagement users who display the highest app usage, data consumption and screen time. They are likely to be frequent social media users or gamers leading to higher battery drain. Cluster 3 (green) are the moderate users that have balanced engagement, with a moderate app usage and screen-on time. Cluster 4 (red) are the average users with steady engagement, between moderate and high engagement.

Cluster (Hierarchical / K-Means)	Description	Age Range
Hierarchical 1 / K-Means 2	Low app usage, low engagement users	~38.18 years
Hierarchical 2 / K-Means 1	Heavy users, high app usage & data consumption	~38.41 years
Hierarchical 3 / K-Means 0	Moderate-to-light users, balanced usage	~38.80 years
Hierarchical 4 / K-Means 3	Average users, steady engagement	~38.34 years

Figure 8: Cluster Descriptions

RQ2: Which mobile usage features contribute most to behavioural variations, and can PCA effectively reduce dimensionality while preserving key insights?

PC1 alone explained ~96% of the total variance, PCA loadings showed that App Usage Time, Battery Drain and Number of Apps Installed contributed most significantly to PC1.

	PC1	PC2
App Usage Time (min/day)	0.448728	-0.100431
Screen On Time (hours/day)	0.447153	0.104249
Battery Drain (mAh/day)	0.448287	-0.427462
Number of Apps Installed	0.448227	-0.376664
Data Usage (MB/day)	0.443654	0.808979

Figure 9: PCA Loadings for PC1 and PC2

By reducing five behavioural features to just two principal components, PCA facilitated effective clustering, enhanced visualisation and maintained meaningful patterns. This supports the use of PCA in behavioural segmentation, especially when dealing with correlated variables.

3. World Happiness Dataset

Regression is a type of supervised machine learning technique used to model the relationship between a continuous dependent outcome variable and one or more independent predictor variables. In this section, we apply regression models to help quantify the impact of various socio-economic factors on national happiness levels to predict the target variable.

3.1 Substantive Issue

With happiness and well-being increasingly acknowledged as essential indicators of a country's development, happiness scores can become insightful information for governments and organisations in shaping public policies for countries' social development. This section seeks to explore how different features influence the happiness score across countries.

3.2 Methodology

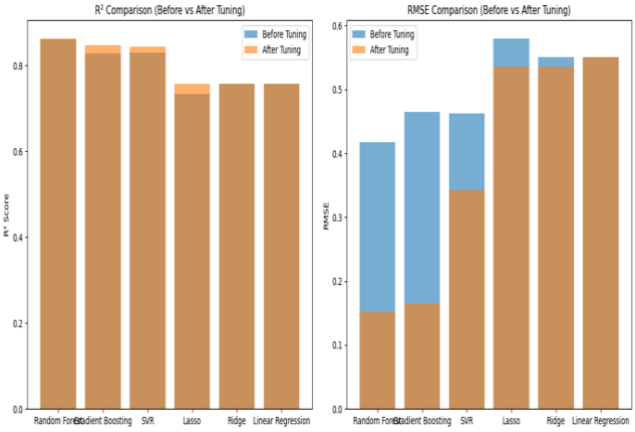


Figure 9: R^2 and RMSE Comparison Bar Charts

To model and evaluate the predictors of happiness, some regression techniques were applied: **Linear Regression**, **Lasso Regression**, **Ridge Regression**, **Support Vector Regression (SVR)**, **Random Forest Regressor** and **Gradient Boosting Regression**. Each model was trained using scaled training data using **StandardScaler** and evaluated on a test set. Hyperparameter tuning was performed using **GridSearchCV** for all models except Linear Regression to optimise their performance. The dataset was split into **training (80%)** and **test (20%)** sets. The models were compared using these performance metrics: $\boldsymbol{R^2}$, which measures the proportion of the variance in the happiness score that is explained by the model, **Root Mean Squared Error (RMSE)**, which measures the accuracy of a model's predictions and **Mean Absolute Error (MAE)** which represents the average value of errors by subtracting between the predicted and actual values. **Residual Analysis** and **Quantile-Quantile (Q-Q) plots** were also used to validate model assumptions and evaluate the normality of residuals.

3.3 Dataset and Variables

A total of 137 rows and 9 variables are included in this dataset, sourced from World Happiness Report 2024. It captures national-level indicators with independent predictor variables and "Life Ladder", the dependent target variable reflecting the happiness score.

Variable	Data Type	Description
Life Ladder	Float	Subjective well-being score (0-10) to the Cantril ladder question: "On which step of the ladder would you say you personally feel you stand at this time?"
Log GDP per Capita	Float	Economic prosperity in terms of purchasing power parity, adjusted for population size and logged.
Social Support	Float	National average of responses to: "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"

Healthy Life Expectancy	Float	Life expectancy at birth, measured in years.
Freedom to Make Life Choices	Float	National average of binary responses to: “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”
Generosity	Float	Donations as a proxy for generosity.
Perceptions of Corruption	Float	Average of responses to two questions: “Is corruption widespread in government?” and “Is corruption widespread in business?”
Country Name	String	Country
Positive Affect	Float	Average of responses to: “Did you experience laughter, enjoyment, or doing something interesting a lot of the day yesterday?”
Negative Affect	Float	Average of responses to: “Did you experience worry, sadness, or anger a lot of the day yesterday?”

Table 2: World Happiness Dataset Description

3.4 Analysis

After preprocessing and scaling the dataset, all six models were trained and tested. Model performance was evaluated based on **R²**, **RMSE** and **MAE**.

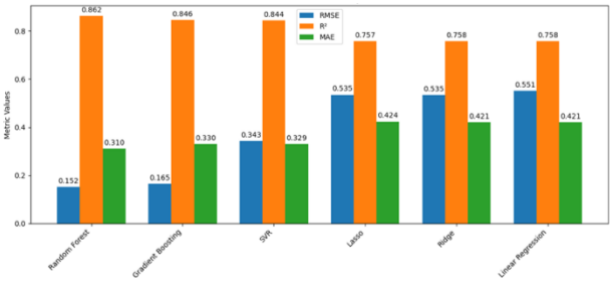


Figure 10: Model Performance Overview

Bar plots for R², RMSE and MAE show that the random forest model achieved the highest R² (0.862), lowest RMSE (0.152) and lowest MAE (0.310), indicating the best overall performance.

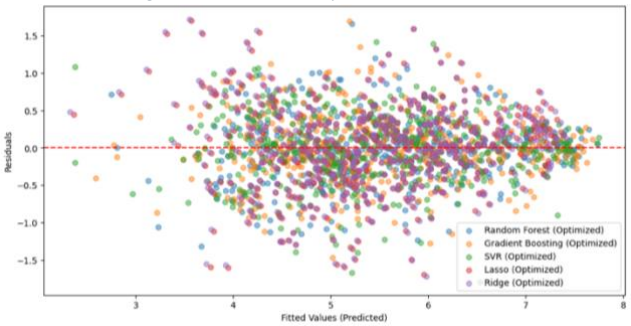


Figure 11: Residuals vs Fitted Plot Comparison

Residuals vs Fitted plots were used to examine homoscedasticity and bias in the models. Most residuals were randomly scattered, suggesting no major bias, though linear models showed more spread.

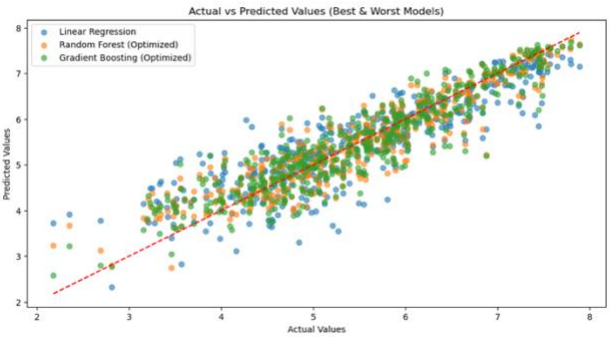
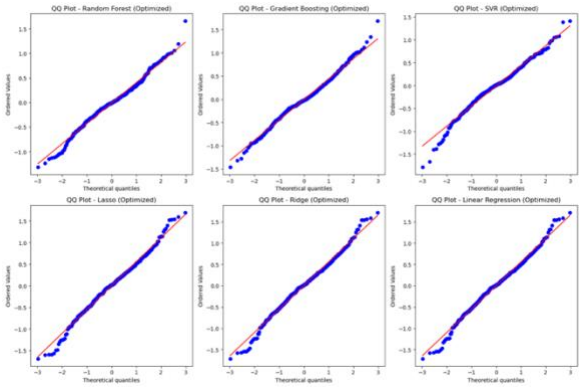


Figure 12: Actual vs Predicted Values (Best and Worst model) Plot Comparison

Figure 12 compares actual and predicted happiness scores. Random Forest and Gradient Boosting models showed predictions tightly clustered along the 45° line, reflecting strong predictive capability. While other models had slightly greater deviation, particularly at the extremes of the happiness scale.



Q-Q plots are used to assess normality of residuals. All 6 models exhibited approximately normal residuals, especially the ensemble methods, which produced tighter distributions along the red diagonal line.

Figure 13: Q-Q Plots Comparison

Random forest model outperformed the others as it is better able to capture complex and non-linear interactions between variables.

3.5 Results

RQ1: Can happiness scores be accurately predicted using GDP, social support, health and freedom metrics?

Yes, the Model Performance Bar Chart shows that the results show that happiness scores, represented by "Life Ladder" can be accurately predicted using socio-economic indicators.

Model	RMSE	R^2	MAE
Random Forest	0.152397	0.862179	0.310389
Gradient Boosting	0.164728	0.846467	0.330485
SVR	0.342590	0.844068	0.329454
Lasso	0.535335	0.757347	0.423651
Ridge	0.535071	0.758212	0.420970
Linear Regression	0.551198	0.758217	0.420959

Figure 14: Model Performance Comparison Table

The best performing model, Random Forest, achieved a R^2 value of 0.86, which represents that 86% of the variance in happiness scores is explained by the independent variables. This highlights a strong relationship between the predictor and target variables.

RQ2: How do different regression models compare in predicting happiness scores?

The models showed varied performance with Random Forest performing the best with the lowest RMSE (0.152), highest R^2 (0.862) and lowest MAE (0.310). Gradient Boosting and SVR also performed well, with strong predictive power. While Lasso, Ridge and Linear Regression had poorer performances in comparison, with higher error metrics and lower R^2.

While the six models performed well generally, there are some limitations to keep in mind. This dataset only includes 137 observations which may limit the generalisability of results over different years and across different countries. Furthermore, happiness is inherently subjective and things like cultural differences may influence their interpretations and results. Hence, these factors may create noise to the model as it may not be captured fully.

RQ3: Which feature(s) has (have) the strongest influence on happiness scores?

Using feature importance from the two best performing models, Random Forest and Gradient Boosting, “Log GDP per capita” emerged as the most dominant predictor of happiness, highlighting the strong relationship between national prosperity and subjective well-being. This reinforces the fact the economic and social infrastructure of a society plays key roles in shaping a nation’s well-being. This is followed by “Social support” and “Healthy life expectancy at birth” also had moderate importance. Whereas variables such as “Generosity” and “Perceptions of corruption” seem to be less significant in affecting happiness scores.

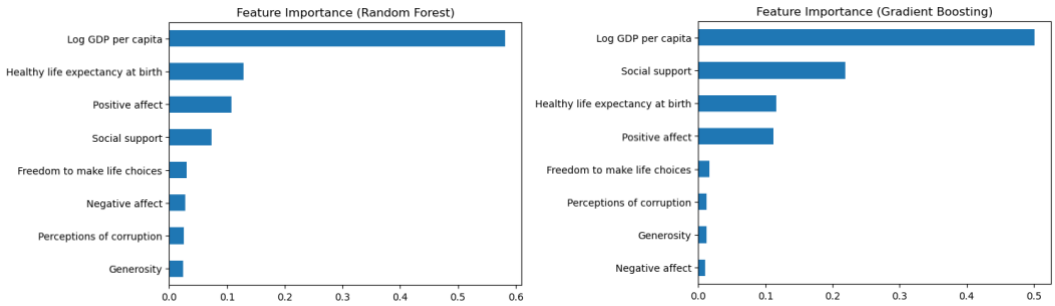


Figure 15: Feature Importance Bar Chart (Random Forest and Gradient Boosting models)

4. Credit Card Fraud Detection Dataset

Classification is a supervised learning technique used to predict categorical labels by mapping input features to predefined classes. In this section, we apply classification models to detect fraudulent credit card transactions, comparing multiple approaches to determine the most effective fraud detection method.

4.1 Substantive Issue

Given the rapid growth of online payments, identifying fraud in real-time is crucial for preventing financial losses. Traditional fraud detection systems often are unable to fully adapt to ever chaning fraudulent techniques, leading to undetected cases or high rates of false positives. Identifying fraudulent transactions accurately is crucial for financial institutions to minimise financial losses and safeguard and protect its consumers. This section seeks to explore various classification techniques to build models capable of distinguishing fraudulent transactions from legitimate ones using historical transaction data.

4.2 Methodology

Logistic Regression, **Decision Tree**, **XGBoost**, **Gradient Boosting**, **Ridge Classifier** and **Lasso Regression** are classification techniques used to analyse the dataset and distinguish fraudulent transactions from non-fraudulent ones. These models were trained to recognise patterns in transaction data and predict fraudulent behaviour based on historical records.

To address the class imbalance in this dataset as seen from fraudulent transactions accounting for only 0.17%, **Synthetic Minority Oversampling Technique (SMOTE)** was applied to ensure the models were trained on a balanced representation of fraudulent and non-fraudulent transactions. **StandardScaler** was used to normalise numerical features to ensure equal weighting across models, improving the performance of models sensitive to feature scaling. The models were trained on an **80/20 split** of the data.

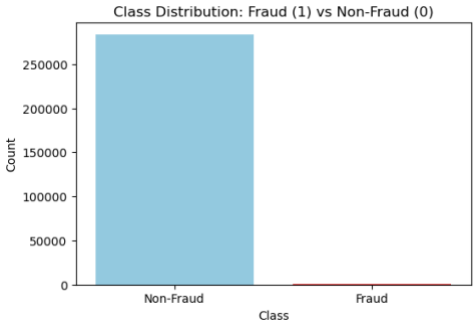


Figure 16: Distribution of Fraudulent (1) and Non-Fraudulent (0) Transactions

4.3 Dataset and Variables

The dataset consists of anonymised transactions, with 284,807 records and 31 columns.

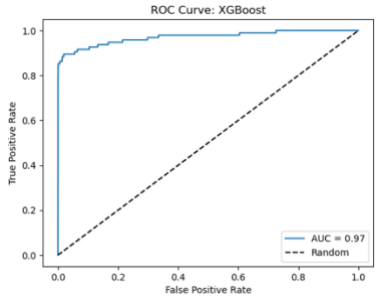
Variable	Data Type	Description
Time	Numeric	Time (in seconds) between an individual transaction and first transaction in the dataset.
Amount	Numeric	Transaction amount.
Class	Integer	1 for Fraud, and 0 otherwise.
V1, V2, … ,V27, V28	Integer	Principal components obtained from PCA.

Table 3: Credit Card Fraud Detection Dataset Description

4.4 Analysis

To classify fraudulent transactions, six machine learning models were trained and evaluated. The dataset was split into 80% training and 20% testing. Each model was evaluated and

assessed based on key classification metrics: **Accuracy**, **Precision**, **Recall**, **F1-score** and **Area Under the Receiver Operating Characteristic (ROC) Curve** (AUC). The ROC Curve Comparison provides further insight into model performance by visualising the trade-off between the true positive rate (fraudulent) and the false positive rate (non-fraudulent). AUC values range from 0.5 (random guess) to 1 (perfect classification).



XGBoost achieves the highest AUC of 0.971, indicating the strongest ability to differentiate fraudulent from non-fraudulent transactions among all classifiers by handling feature interactions and imbalanced data through tree boosting.

Figure 17: ROC Curve for XGBoost Model

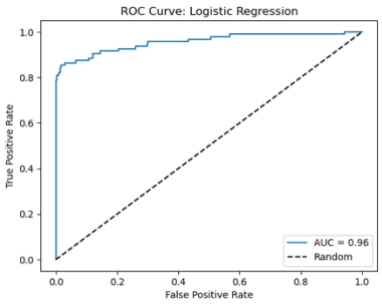


Figure 18: ROC Curve for Logistic Regression Model

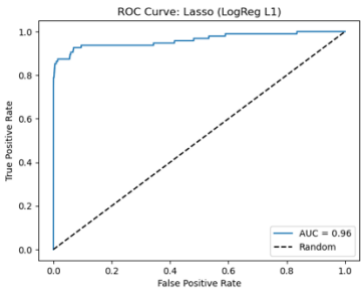


Figure 19: ROC Curve for Lasso Regression Model

Lasso Regression and Logistic Regression perform similarly, with AUC values of 0.962 and 0.956 respectively, both showing strong performance in fraud classification.

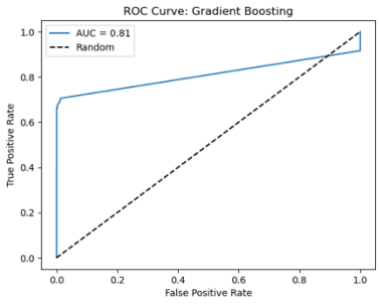


Figure 20: ROC Curve for Gradient Boosting Model

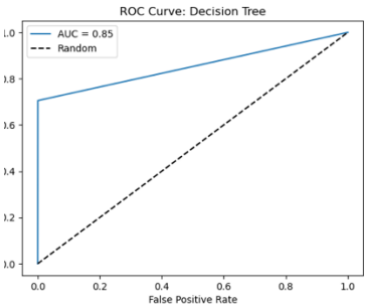


Figure 21: ROC Curve for Decision Tree Model

Gradient Boosting and Decision Tree show significantly lower AUC scores of 0.809 and 0.852 respectively, indicating weaker ability in identifying fraud cases.

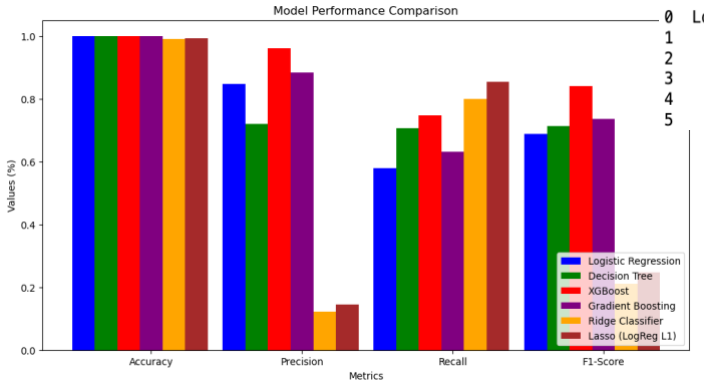


Figure 22:Model Performance Comparison

	Model	Accuracy	Precision	Recall	F1-Score	AUC
0	Logistic Regression	0.999119	0.846154	0.578947	0.687500	0.956194
1	Decision Tree	0.999048	0.720430	0.705263	0.712766	0.852402
2	XGBoost	0.999524	0.959459	0.747368	0.840237	0.970946
3	Gradient Boosting	0.999242	0.882353	0.631579	0.736196	0.808903
4	Ridge Classifier	0.989990	0.121600	0.800000	0.211111	NaN
5	Lasso (LogReg L1)	0.991312	0.144643	0.852632	0.247328	0.961518

Figure 23: Performance Metrics Table

While accuracy provides a general measure of correctness, precision minimises false positives and recall ensures fraudulent transactions are correctly identified.

XGBoost model has the highest precision and F1-score, making it the most balanced model in terms of identifying fraudulent transactions while minimising false positives. Decision

Tree and Gradient Boosting models show lower precision, indicating a higher rate of false positives.

Ridge Classifier and Lasso Regression models performed poorly in precision and F1-score, suggesting they are not optimal choices for fraud detection despite high accuracies. Recall is highest for Lasso Regression and Ridge Classifier models, suggesting that these models identify more fraudulent transactions but at the cost of precision, leading to a higher number of false positives.

4.5 Results

Based on the comparison plot of the ROC curves below, XGBoost model (green) outperforms the other 5 models as the ROC curve is the closest to the top-left corner, indicating high true positive rates and low false positives. XGBoost model is the best performing classification model also due to its high accuracy, precision and recall scores.

RQ1: Can machine learning models accurately classify transactions as fraudulent and non-fraudulent?

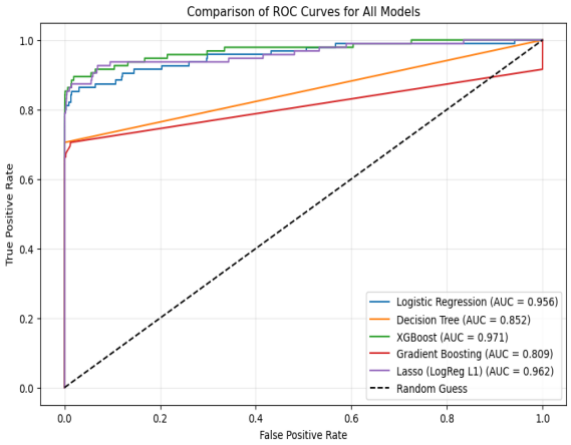


Figure 24: Combined ROC Curves of All Models

The results showed that XGBoost outperformed all other models, achieving an AUC score of 0.97, making it the most effective classifier for fraud detection. Logistic Regression and Lasso Regression also demonstrated strong performance, with AUC values around 0.96. In contrast, Decision Tree and Gradient Boosting had lower AUC scores (0.81–0.85), indicating weaker predictive power. Ridge Classifier performed poorly and lacked an AUC score due to its nature.

RQ2: Which features contribute the most to detecting fraud, and how can this knowledge enhance fraud prevention strategies?

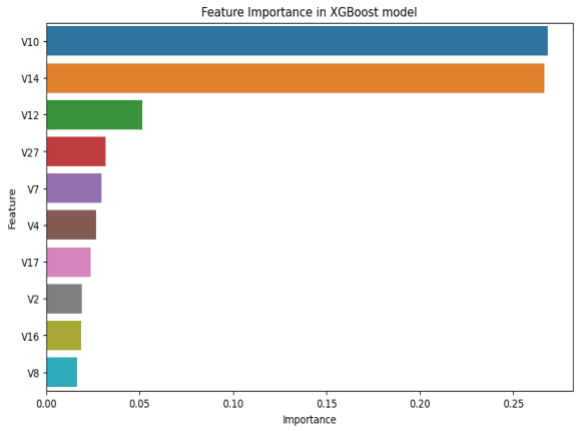


Figure 25: Feature Importance Plot for XG Model

Feature importance analysis using XGBoost model showed that V10 and V14 were the most influential features in fraud detection. Other key predictors included V12, V27 and V7, while the transaction amount was not a major factor in classification.

By understanding the most relevant features, financial institutions can enhance fraud detection mechanisms by monitoring these signals more closely and focus on transactions with high-risk indicators by developing custom alert systems for example. This insight can also aid in targeting specific transaction behaviours and refining fraud prevention strategies, such as improving real-time monitoring, risk scoring and anomaly detection.

5. References

Part 1 Unsupervised Learning Dataset: Mobile Device Usage and User Behaviour Dataset
https://www.kaggle.com/datasets/valakhorasani/mobile-device-usage-and-user-behavior-dataset?resource=download

Part 2 Regression Dataset: World Happiness 2024 Dataset
https://worldhappiness.report/data/

Part 3 Classification Dataset: Credit Card Fraud Dataset
https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data

Backup Files of All 3 Datasets and 3 Jupyter Notebooks
https://drive.google.com/drive/folders/1hoRwybnFxigoY9bnGZjEKrIMdrjli6sK?usp=drive_link