ST3189 Assessed Coursework Project

By Md Fahim Faisal Khan

Student number: 220456427

1. Unsupervised Learning: Segmenting Customers Using Behavioural and Demographic Clustering

Link to dataset: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

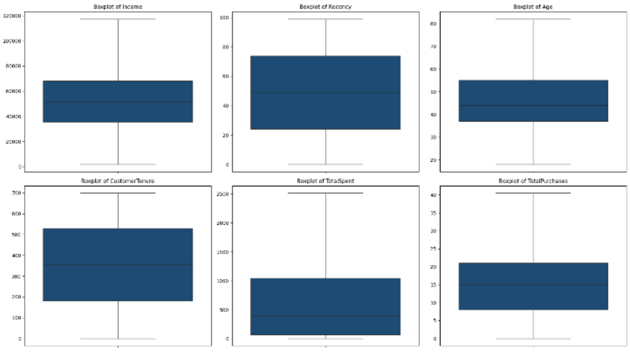
1.1 Introduction

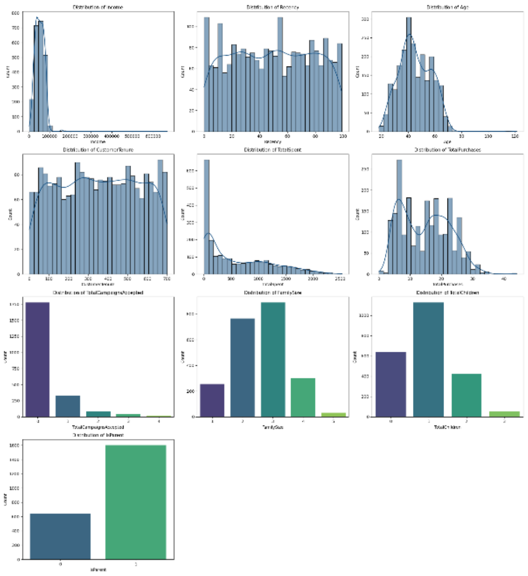
Customer segmentation is an important technique in data-driven marketing because it allows organisations to adapt engagement based on specific behavioural and demographic characteristics. This study investigates an anonymized marketing dataset to identify underlying client categories that lack established labels. The analysis is led by three main questions:

1. What consumer segments exist in the data?
2. How do these segments differ in expenditure, behaviour, and marketing responses?
3. Which focused strategies are advised for each group?

To answer these questions, we used unsupervised learning techniques such as KMeans and Agglomerative (Hierarchical) Clustering, together with PCA for visualization, Silhouette Scores for validation, and cluster profiling to extract actionable insights.

1.2 Data preparation and Explanatory Data Analysis (EDA)

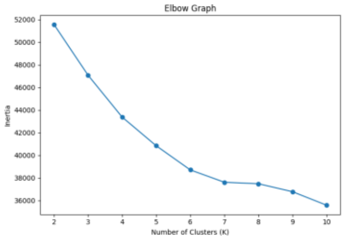


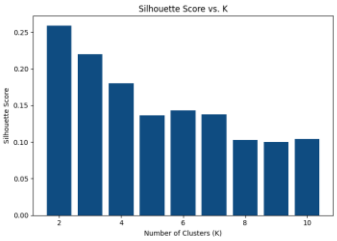


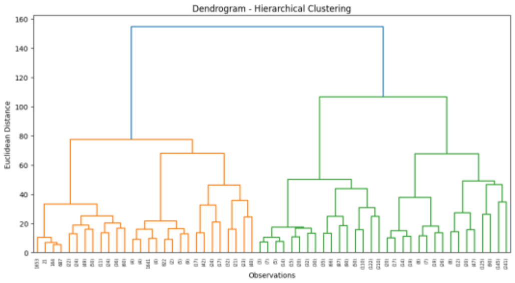
Exploratory Data Analysis - Key Features

The dataset underwent extensive preprocessing to verify its appropriateness for clustering. Given the right-skewed distribution of the Income variable, missing values (1.07%) were imputed using the median. Irrelevant or constant columns (ID, Z_CostContact, Z_Revenue) and redundant characteristics (Year_Birth, after transformation) have been eliminated. New variables were created to improve interpretability, such as Age (derived from Year_Birth), CustomerTenure (derived from Dt_Customer), IsParent (based on Kidhome and Teenhome), and high-level aggregations like TotalSpent, TotalPurchases, and TotalCampaignsAccepted. Categorical variables with excessive granularity were condensed into fewer, more relevant categories and then encoded. Extreme outliers were also capped at the first and 99th percentiles to reduce undue influence. These processes produced a clean, structured dataset ready for effective unsupervised learning, as indicated by the boxplots above. The distribution of our key features is also shown above.

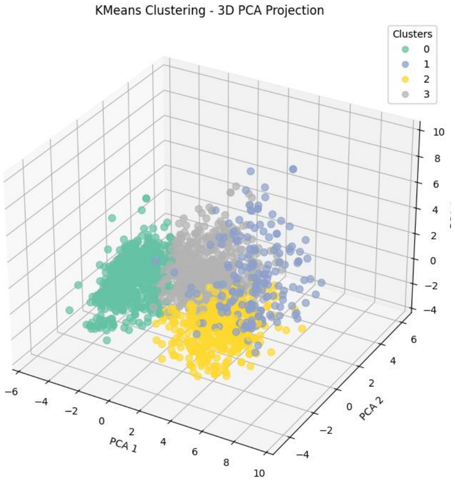
1.3 Clustering Preparation and Optimal K Selection

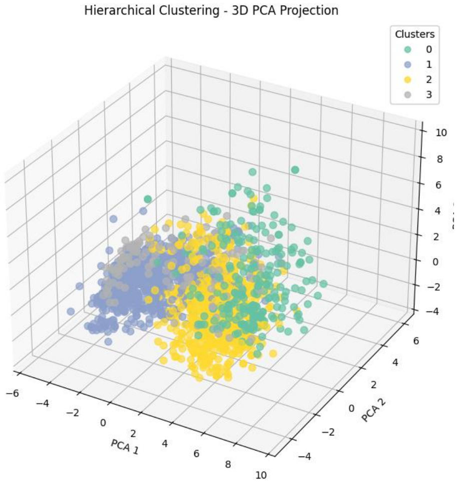






To expedite the procedure, all numerical and encoded category features were programmatically picked based on their data types before clustering. These attributes were subsequently standardized to provide an equal contribution to distance computations. To establish the appropriate number of clusters, KMeans clustering was tested with a variety of K values (2-10). As shown in the Elbow Plot, inertia decreases dramatically up to K = 4 before plateauing, indicating decreasing returns after this point. Simultaneously, the Silhouette Score, which assesses cluster cohesion and separation, peaked at K = 2 before stabilizing around K = 4. This balance of interpretability and structure led to the decision to use K = 4 as the ideal cluster count. A dendrogram was constructed for hierarchical clustering, confirming the existence of four unique cluster groupings based on linkage distance.





3D PCA plots were used to visualise cluster separation for both KMeans and Agglomerative clustering. While the Agglomerative model achieved a slightly higher silhouette score (0.196 vs. 0.181), both methods showed similar structural separation in reduced space. We will now proceed to profiling our respective clusters.

1.3.1 Combined Cluster Profiling Summary

To interpret the clustering results, average feature values were used to construct customer profiles across both KMeans and Hierarchical models. Across both methods, four clear customer personas emerged. KMeans identified distinct groups such as high-income wine buyers, price-sensitive parents, and digitally engaged middle-income households. In contrast, Hierarchical clustering revealed a similar structure, though with slightly more overlap between segments and nuanced differences in channel behaviour and campaign responsiveness. While cluster interpretations were largely consistent, KMeans offered clearer separation and interpretability, especially for premium and budget-conscious segments. These insights enable actionable marketing strategies:

- High-income, high-spending clusters can be targeted with premium bundles, loyalty rewards, and early product access.
- Digitally active middle-tier customers respond well to email campaigns and online-exclusive deals.
- Price-sensitive family segments benefit from discount packs, flexible payment plans, and promotions tied to family needs.
- Campaign-responsive groups can be nurtured through personalised follow-ups and tiered incentive programmes.

This profiling reinforces the strategic value of clustering in enabling personalised, data-driven customer engagement.

KMeans Cluster Profile (Selected Features)

Feature	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Income	35,294.29	81,029.78	72,243.75	57,966.96
TotalSpent	98.23	1,613.36	1,237.16	752.11
TotalPurchases	7.88	20.82	20.29	21.58
CustomerTenure	317.96	353.61	354.55	415.97
IsParent	0.88	0.16	0.17	0.98
TotalCampaignsAccepted	0.08	2.02	0.16	0.24

Hierarchical Cluster Profile (Selected Features)

Feature	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Income	78,131.89	39,596.41	67,182.89	44,780.66
TotalSpent	1,489.88	196.30	1,080.99	529.89
TotalPurchases	20.94	10.47	21.23	14.12
CustomerTenure	336.82	336.96	389.40	367.19
IsParent	0.22	0.90	0.53	0.82
TotalCampaignsAccepted	1.82	0.00	0.14	1.05

1.4 Conclusion

This research successfully identified four unique consumer segments using unsupervised learning, based on demographic characteristics, purchasing behaviour, and campaign receptivity. KMeans and Hierarchical clustering both created significant categories, but KMeans provided more interpretable results for downstream analysis. The generated profiles enabled actionable marketing recommendations suited to each sector, ranging from premium loyalty schemes to discount-driven engagement tactics. Overall, this study highlights the importance of clustering in revealing latent consumer structure and facilitating data-driven decision-making.

2. Regression Analysis: Estimating Life Expectancy Using Socioeconomic and Health Indicators

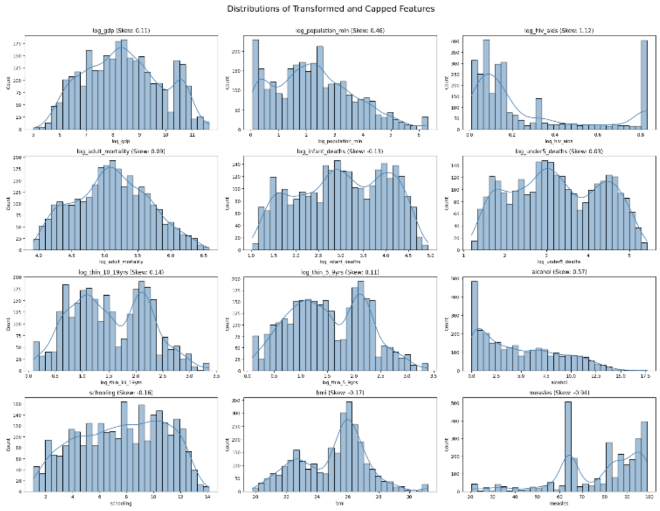
Link to dataset: https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated

Introduction

Life expectancy is an important indicator of a country's health and development, impacted by variables such as healthcare quality, economic status, disease prevalence, and education. This project uses supervised machine learning techniques to model and forecast life expectancy using a World Health Organization (WHO) dataset that includes a variety of health, demographic, and socioeconomic characteristics across nations and time periods. To verify robustness and interpretability, multiple regression models were tested, including Multiple Linear Regression, Ridge, Lasso, and Random Forest Regressor. R^2 and RMSE were used to evaluate the models' explanatory power and prediction accuracy. To streamline experimentation, preprocessing techniques like as imputation, encoding, scaling, and feature transformation were integrated as part of a pipeline.

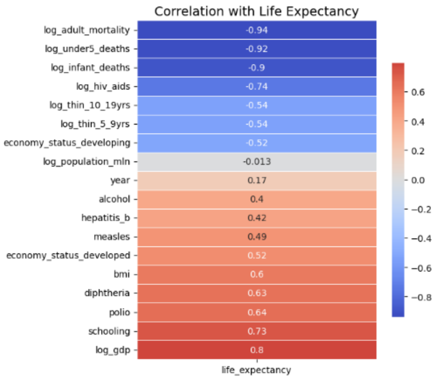
Research Questions

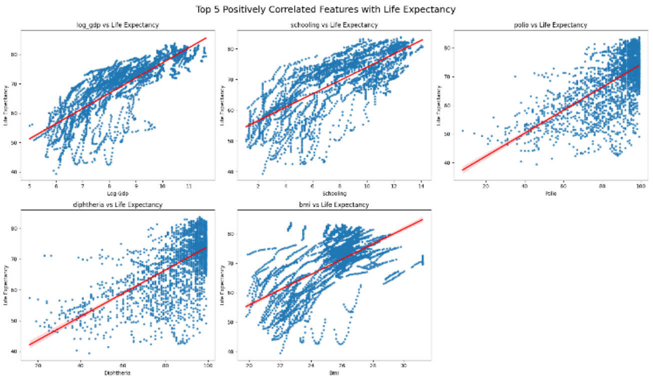
1. Which features are most influential in determining a country's life expectancy?
2. How accurately can life expectancy be predicted using available socioeconomic and health indicators?
3. Which regression model provides the best balance of accuracy and generalisability across countries and regions?

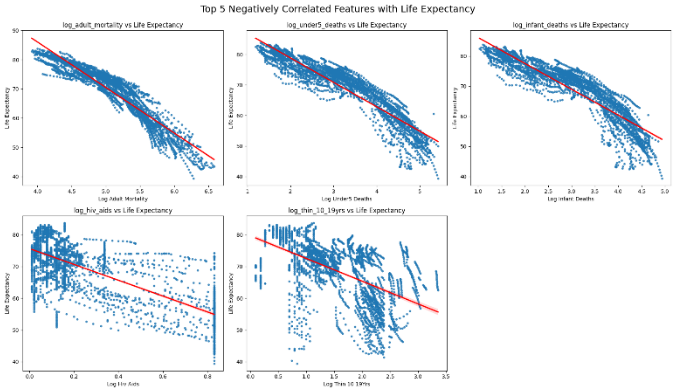


2.1 Methodology: Data Cleaning and Exploratory Data Analysis (EDA)

The dataset was examined for missing values, which were absent from all features. To decrease skewness and stabilize variance, numerical features with high skewness were log-transformed, including GDP, population_mln, HIV_aids, and adult_mortality. Features with moderate outliers but less skewed distributions, such as polio, diphtheria, and measles, were capped with the IQR approach to reduce the impact of extreme values while maintaining the underlying distribution. This procedure helped normalize the data, as shown in the histograms, where the distributions of log-transformed features (e.g., GDP, HIV_aids) are now more symmetrical and features like Schooling and BMI remain within a reasonable range, with no significant outliers. The distributions of converted and capped features (shown above) reveal that log transformation and capping were successfully implemented to reduce skewness while keeping the overall data structure. The heatmap depicts the connections between characteristics and life expectancy, highlighting significant positive factors such as log_gdp, education, and polio, as well as negative influences such as log_adult_mortality and log_under5_deaths. The scatter plots highlight these correlations, with the red regression lines indicating that higher GDP, schooling, and polio vaccination rates correlate with longer life expectancy, whereas adult mortality and newborn mortality indicate the inverse. These visualizations aid in determining which factors have the most impact on life expectancy, directing subsequent study and model creation for prediction.







2.2 Data Preprocessing and Model Evaluation

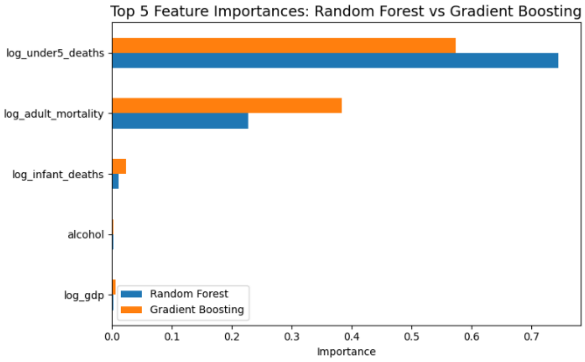
For data preparation, we used One-Hot Encoding for categorical variables and Standard Scaling for numerical features to guarantee that all data was on the same scale. To guarantee fair evaluation, the dataset was divided into training and test sets, each having a test size of 0.3. We used Cross-Validation (CV) to evaluate model performance across multiple data subsets and reduce overfitting. Model evaluation measures included Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R^2 score, and CV R^2 to assess prediction accuracy and consistency. These processes ensured a thorough and reliable evaluation of the models' capacity to predict life expectancy.

Model	**MAE**	**RMSE**	**R^2 (Test)**	**CV R^2 (5-fold)**
Random Forest	0.4011	0.5532	0.9964	0.9964

Gradient Boosting	0.6564	0.8273	0.9920	0.9913
MLR	1.1564	1.5570	0.9711	0.9709
Ridge Regression	1.1805	1.5904	0.9705	0.9707
Lasso Regression	1.3330	1.8169	0.9615	0.9618

The Random Forest model delivered the best performance, with the lowest MAE and RMSE values, and the highest R^2 score on both test data and cross-validation (CV). This indicates that Random Forest was the most accurate model for predicting life expectancy in this dataset.

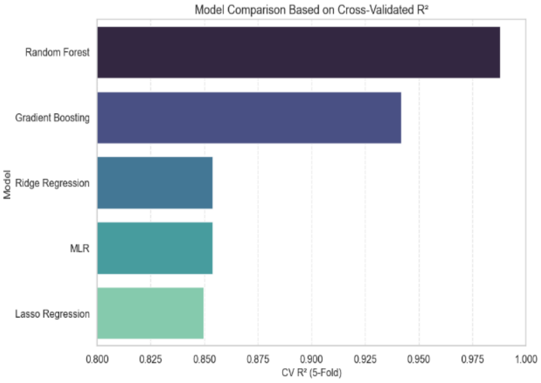
The Gradient Boosting model performed well as well, but with slightly higher MAE and RMSE compared to Random Forest. However, it still maintained impressive R^2 values, indicating strong predictive power, albeit with slightly less precision than Random Forest. The Multiple Linear Regression (MLR) model, despite being the simplest model, showed lower accuracy with higher MAE and RMSE. Its R^2 value was also lower, suggesting that linear models are less effective for this dataset compared to more complex models like Random Forest and Gradient Boosting. The Ridge Regression, which is a regularized version of MLR, showed better performance than MLR, but still fell behind Random Forest and Gradient Boosting. It was slightly more stable in terms of coefficient estimation but did not provide as strong predictions as the ensemble models. Lastly, the Lasso Regression, with L1 regularization that tends to zero out less important features, performed the worst. It had the highest MAE and RMSE among all models, suggesting that the Lasso method was not as effective for this dataset.



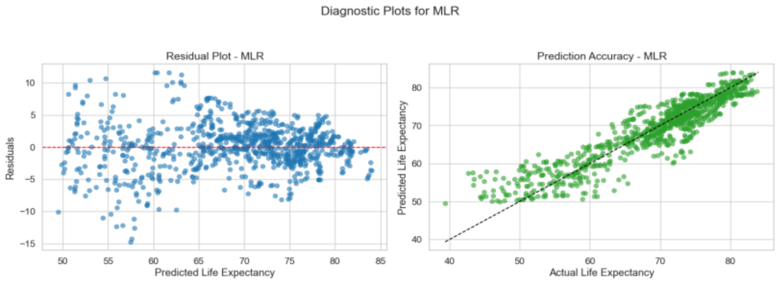
The bar chart compares feature importances from the Random Forest and Gradient Boosting models. It shows that variables such as log_under5_deaths and log_adult_mortality dominate as the most important predictors of life expectancy. These variables are directly associated with death and hence they capture much of the variance in the target variable. To explore the relationships among the remaining features and reveal additional drivers of life expectancy, we excluded these top predictors in the next modelling phase. We rerun the model and get the follow results.

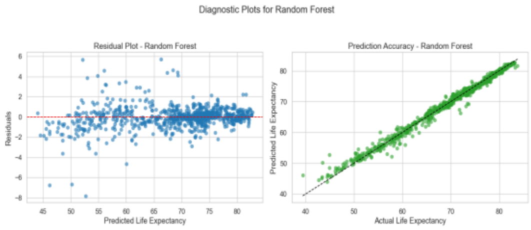
Model Evaluation Results:

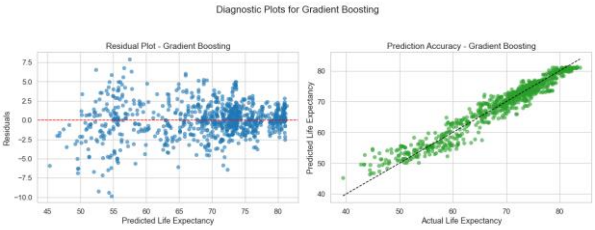
Model	**MAE**	**RMSE**	**R^2 (Test)**	**CV R^2 (5-fold)**
Random Forest	0.6668	1.0721	0.9866	0.9881
Gradient Boosting	1.6402	2.2240	0.9423	0.9418
Ridge Regression	2.7043	3.6510	0.8446	0.8539
MLR	2.7043	3.6510	0.8446	0.8539
Lasso Regression	2.7460	3.7073	0.8398	0.8499

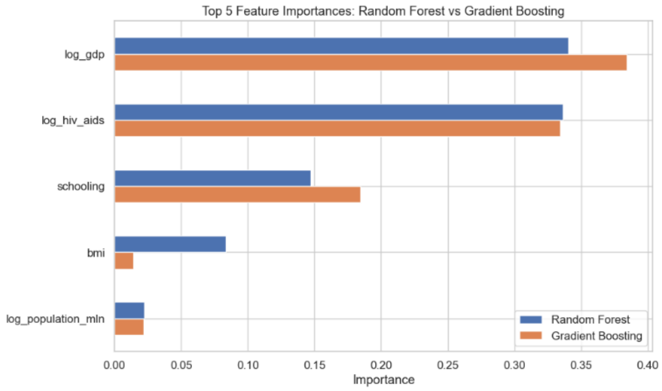


After refining our feature set by removing dominant predictors, model performances were reevaluated. Random Forest remained the top performer with the lowest MAE and RMSE, followed closely by Gradient Boosting. To assess model reliability and residual behaviour, diagnostic plots were generated for the top two models alongside our baseline Multiple Linear Regression (MLR). The residual plot for MLR reveals noticeable heteroscedasticity and wider variance, indicating fewer stable predictions. In contrast, Random Forest and Gradient Boosting exhibit tighter residual distributions around zero and higher alignment in predicted vs. actual values, highlighting better model fit and generalisation.









The final feature importance chart, post-removal of dominant variables, reveals new influential predictors: log_gdp, log_hiv_aids, and schooling consistently rank highest across both Random Forest and Gradient Boosting. This suggests that economic strength, disease burden, and access to education remain key determinants of life expectancy, even after excluding overly dominant features.

2.3 Conclusion

This study answered our initial research questions by identifying significant socioeconomic and health-related variables of life expectancy. Our preprocessing pipeline—which included transformation, scaling, and outlier treatment—ensured model stability. The Random Forest model has the highest accuracy, with a R^2 of 0.987 and negligible prediction errors. GDP, HIV prevalence, and education were found to be the most influential variables, reinforcing their relevance in shaping global health outcomes. These insights not only improve interpretability but also provide a data-driven framework for public health policy aimed at education access, disease prevention, and economic development.

3. Classification Analysis: Predicting Term Deposit Subscriptions Using Machine Learning

3.1 Introduction

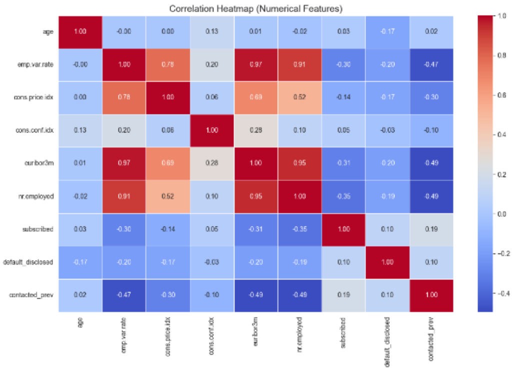
Link to dataset: UCI Bank Marketing Dataset

In the banking industry, predicting client responses to marketing campaigns is crucial for increasing outreach efficiency and conversion rates. This work focuses on determining if a client will subscribe to a term deposit based on demographic, financial, and interaction data. Using the Bank Marketing dataset from the UCI repository, we deploy four models—Logistic Regression, Decision Tree (base and pruned), Random Forest, and XGBoost—to evaluate their performance in handling real-world marketing data and identifying significant predictive factors.

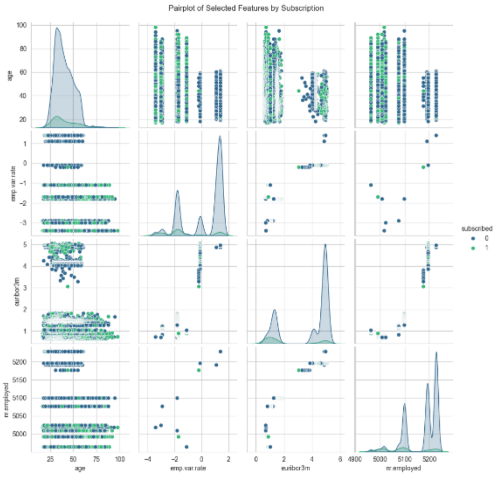
Research Questions

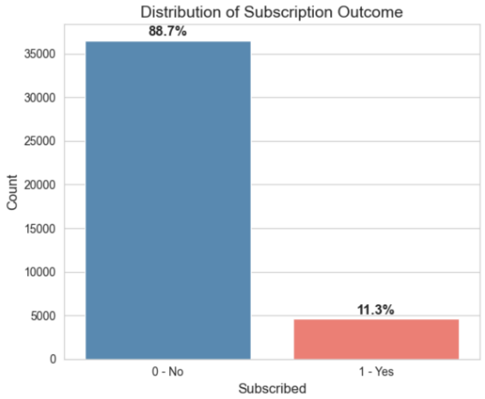
1. Can we predict term deposit subscription outcomes using available customer and campaign features?
2. Which features are most predictive of a client's likelihood to subscribe?
3. How do different classification models compare in terms of predictive power and interpretability?
4. What practical marketing insights can be derived from the model results?

3.2 Dataset Overview and Preprocessing



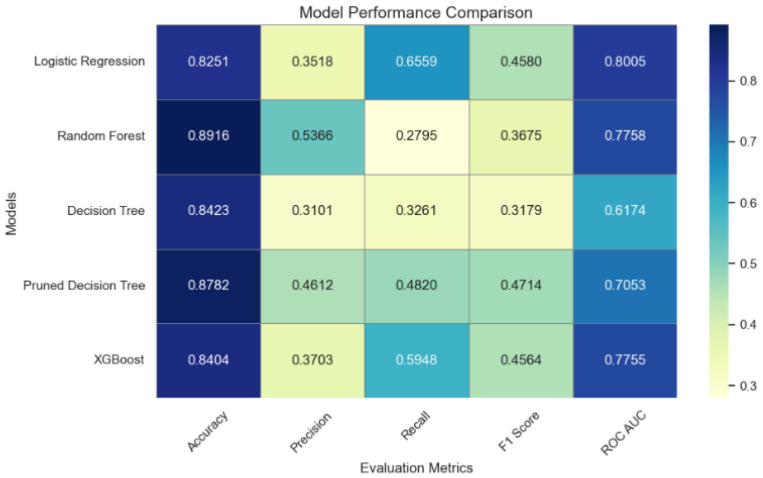
Prior to modelling, substantial data preparation and exploratory analysis were carried out to assure quality and insight. The mode was used to impute "unknown" values in employment and marital, and the sparse default category was removed in favour of a binary default_disclosed feature. Low-frequency job titles were aggregated, education levels were divided into tiers, and new features like contact_timing, campaign_group, and contacted_prev were developed. The duration column was removed to prevent leaking, and numerical outliers were dealt with using capping to reduce skew without altering distributions. A column transformer was used to apply one-hot encoding and standard scaling. To investigate feature correlations with the target, a correlation heatmap revealed that euribor3m, nr.employed, and emp.var.rate had the strongest (negative) associations with subscribed, while a pairplot revealed distinct patterns in these features across subscription outcomes. These visualisations helped to inform feature selection and confirmed the underlying structure of the categorization data.



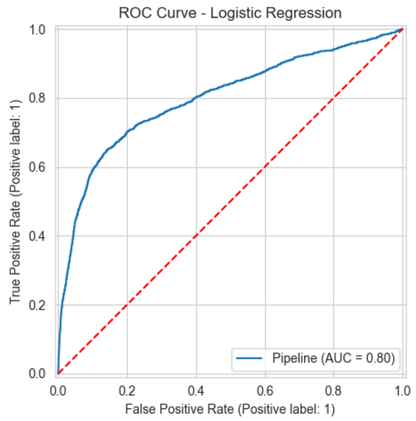


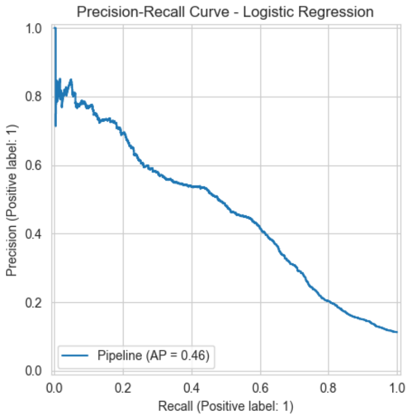
The target variable is highly imbalanced, with only 11.3% of clients subscribing. This imbalance will be addressed later using appropriate model-based strategies.

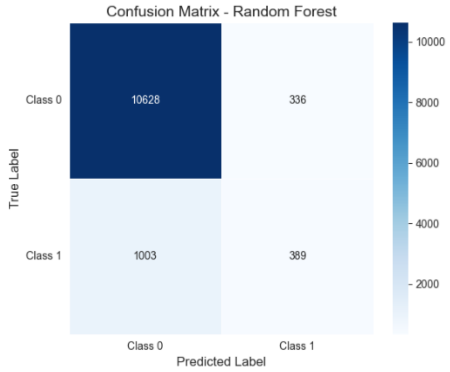
3.3 Model Evaluation Results:

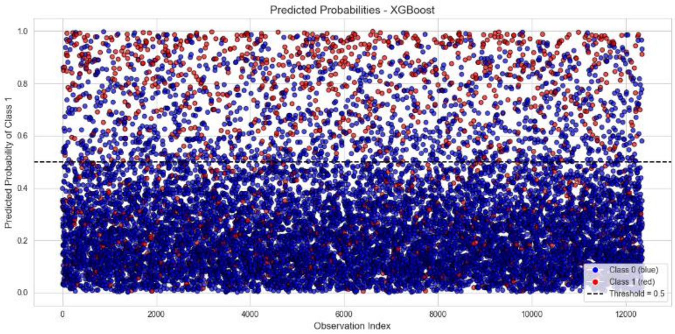


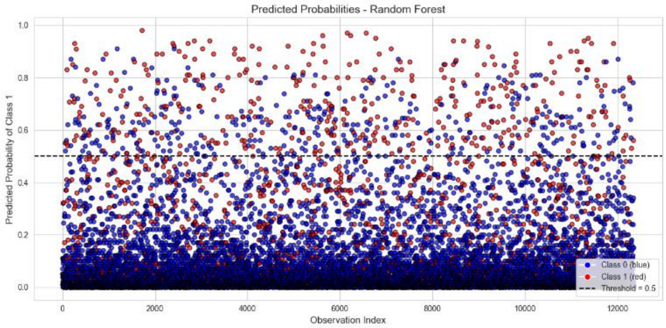
Variations in model performance across metrics were indicative of precision and recall trade-offs. Despite its lower precision, Logistic Regression produced the highest recall (0.66) and ROC AUC (0.80), demonstrating its effectiveness in acquiring subscribers. When finding as many positives as possible is crucial, Random Forest's low recall (0.28) restricts its application, while having the best accuracy (0.89) and precision (0.54). In comparison to its unpruned cousin, the Pruned Decision Tree offered a strong balance, increasing both recall (0.48) and F1 score (0.47). With the second-highest recall (0.59) and a competitive F1 score (0.46), XGBoost regularly outperformed logistic regression, making it a viable option for lead targeting with fewer false positives.



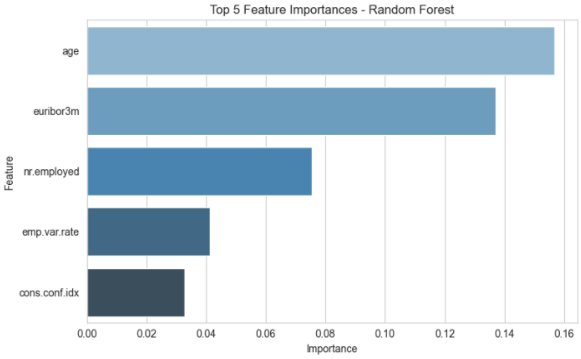


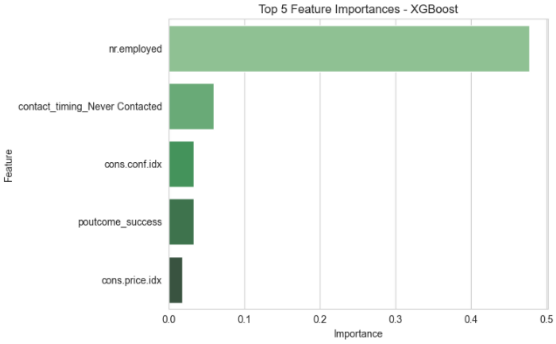


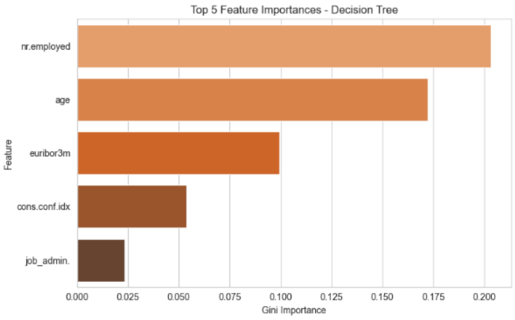




Various strategies were used across models to overcome the dataset's notable class imbalance (~11% positive class). The class_weight="balanced" argument was used to rebalance the loss function and penalize incorrect minority class classification for Random Forest, Decision Tree, and Logistic Regression. To help XGBoost concentrate on the underrepresented class, the scale_pos_weight option was set to indicate the ratio of negative to positive cases. The assessment metrics and anticipated probability distributions showed that these modifications enhanced recall and precision-recall trade-offs.

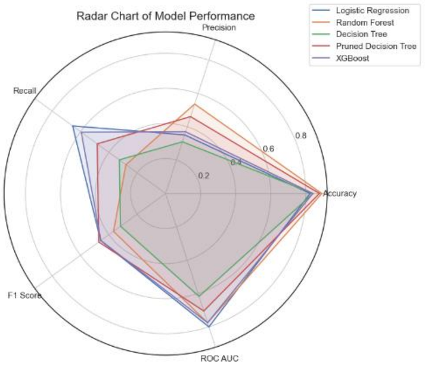






The feature importance plots for Random Forest, XGBoost, and Decision Tree models show a high level of consistency in finding significant subscription predictors. Across all three models, nr.employed, age, euribor3m, and cons.conf.idx consistently rank as top drivers. Notably, nr.employed has the highest relevance in both XGBoost and Decision Tree, demonstrating a dominant influence on prediction. Random Forest also promotes age and euribor3m, implying that these variables capture important patterns related to economic sentiment and client demographics. The inclusion of features such as contact_timing_Never Contacted and poutcome_success in XGBoost emphasizes its sensitivity to interaction history. Overall, the models consistently emphasise a combination of economic indicators and personal qualities, highlighting the importance of both variables in predicting term deposit subscription.

3.4 Conclusion



This classification analysis shown that a combination of demographic, economic, and campaign-related characteristics can accurately predict client term deposit subscriptions. Logistic Regression, Random Forest, and XGBoost all performed well, with XGBoost offering an outstanding balance of precision and recall. Across models, feature importance analysis consistently revealed nr.employed, age, euribor3m, and cons.conf.idx as key predictors, highlighting the significance of macroeconomic conditions and consumer traits. Models effectively handled the skewed target distribution by employing class weighting and imbalance-sensitive tuning. From a strategic standpoint, the findings suggest that banks can optimize marketing efforts by targeting segments aligned with these influential features, particularly clients contacted during good economic times or with prior engagement history and thus converting predictive insights into actionable results.