Assessment Task 3: Data Analytics Project For Marketing Campaign

Name: Sonukumari Rathore
Student Id: 25208188

Executive summary

This project looks into the fundamental factors that influence consumer subscription behaviour in marketing campaigns run by a financial institution. Using a structured collection of demographic, behavioural, and economic factors, the team created prediction and segmentation algorithms to discover high-potential subscribers and optimise marketing efforts.

A clean and model-ready dataset was created by rigorous data preparation, logical imputations, and feature engineering. Predictive modelling was performed using both parametric (Logistic and Bayesian Logistic Regression) and non-parametric (Random Forest) methods. The Random Forest model outperformed expectations (Accuracy = 0.89, Recall = 0.89, ROC-AUC = 0.95), demonstrating its ability to identify clients who are most likely to subscribe.

In addition to the predictive insights, K-Means clustering revealed four primary consumer segments—Young Professionals, Mid-Career Customers, Retirees, and High-Value Clients—which allowed for more focused and efficient marketing campaigns.

The findings generally support the use of data-driven marketing, emphasising the importance of call duration, past campaign performance, and contact scheduling as critical elements in campaign success. These information may help organisations optimise outreach techniques, resource allocation, and customer relationship management, eventually increasing total return on

Table of Contents

Introduction

Introduction

The success of a marketing effort in a saturated market is heavily reliant on reaching the appropriate audience at the right time with the correct message. In today's competitive financial and telecommunications environments, attracting and maintaining clients necessitates a smart, data-driven approach. Marketing initiatives may increase customer engagement and product subscriptions, but only when outreach activities are well-targeted and efficiently carried out.

This project examines data from several marketing campaigns launched by a financial institution to promote term deposit subscriptions. The primary goal is to assist the business in understanding what motivates consumer interaction and identifying characteristics that will increase future campaign success.

The dataset includes precise information about client demographics, contact history, economic indicators, and campaign outcomes. Subscribed is the result variable that tells if a customer accepted the offer after being contacted. This dataset, like many others in the real world, has inconsistencies such as missing values, "unknown" categories, and redundant or correlated characteristics, all of which were resolved throughout the preprocessing and data preparation steps.

Problem Statement

The purpose of this analysis is to discover and communicate the consumer segments, contact techniques, and campaign conditions that are most strongly associated with an increased likelihood of subscription. The analysis is separated into two major components.

⇒ Create predictive models based on demographic, behavioural, and economic factors to estimate customer likelihood of subscribing.

⇒ Extract actionable insights to inform marketing decisions, such as contact methods and audience segmentation.

Analytical Objective

This project converts marketing campaign data into actionable insights to help enhance strategic and operational decision-making. It uses predictive modelling and consumer segmentation to determine who is most likely to subscribe and why.

1. Model Training & Evaluation

Develop and compare predictive models to estimate customer subscription probability.

⇒ Parametric: Logistic and Bayesian Logistic Regression for interpretability.

⇒ Non-Parametric: Random Forest for higher predictive power and feature importance.

⇒ Evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

2. Customer Segmentation

⇒ Apply K-Means Clustering with PCA visualisation to identify customer groups such as *Young Professionals*, *Mid-Career*, *Retirees*, and *High-Value Customers*, enabling more targeted marketing strategies.

3. Feature Interpretation & Insights

Analyse feature importance and model coefficients to reveal drivers like call duration, previous campaign success, and economic factors. These insights guide data-driven decisions for future campaign optimisation.

Research Questions

⇒ How does communication type affect campaign success?
⇒ Which customer segments show the highest likelihood of subscribing?
⇒ Which factors most strongly influence subscription decisions, and how well can models predict them?

Methodology

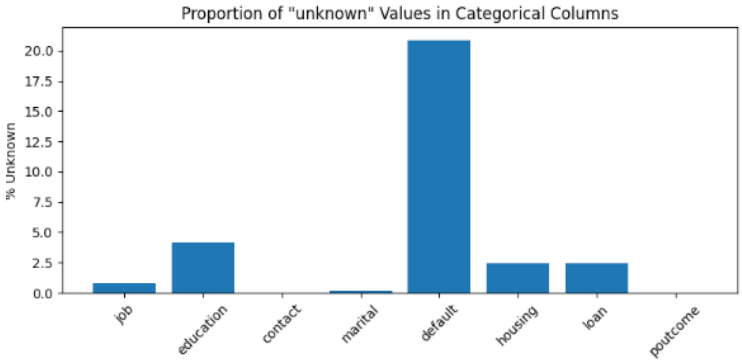
Methodology

Data Preparation

The source dataset contains missing values, outliers, and inconsistencies that could have influenced model results. As a result, the processes mentioned below were used to clean and modify the data.

Handling Unknown Values

This graph shows the fraction of "unknown" values across the dataset's categorical variables. It identifies data quality concerns, demonstrating that 'default' has the largest frequency of missing or undefined values (about 20%). Variables such as education, housing, and loan have modest unknown quantities. These discrepancies were addressed during data cleaning to guarantee consistent model performance.



Imputations Rules

Logical imputation rules were strategically implemented by using links between important variables for example, marital status was inferred from age, employment was estimated based on education level, and default risk was calculated using loan and property Ownership trends. Following these targeted imputations, all remaining unknown values were replaced with the most common (mode) value within each feature to ensure data consistency. This combined context-driven and statistical imputation technique conserved the dataset's natural linkages, decreased information loss, and resulted in a clean, coherent, and model-ready dataset ideal for precise predictive analysis.

Target Variable Transformation

The subscribed variable was cleaned and standardised by mapping all potential text variants (e.g., 'yes', 'y', '1') to binary values (1 for subscription, 0 for non-subscription). This resulted in a consistent framework for model training.

Feature Creation

	pdays	prev_contact	duration	age	campaign	duration_age
0	2	1	896	17	1	15232
1	4	1	92	17	3	1564
2	999	0	182	17	2	3094
3	999	0	498	17	2	8466
4	4	1	432	17	3	7344

In this stage, additional features were created to improve the model's capacity to detect hidden links in the data.

A binary variable, prev_contact, was established to identify whether a customer had been previously contacted (pdays < 999), offering insight on prior engagement influence.

Next, an interaction term duration_age was created by mulgtiplying call duration by age, which captures how response probability varies across age groups and call durations.

An extra interaction, duration_campaign, was introduced to better understand how the number

of contacts throughout a campaign relates to call length.

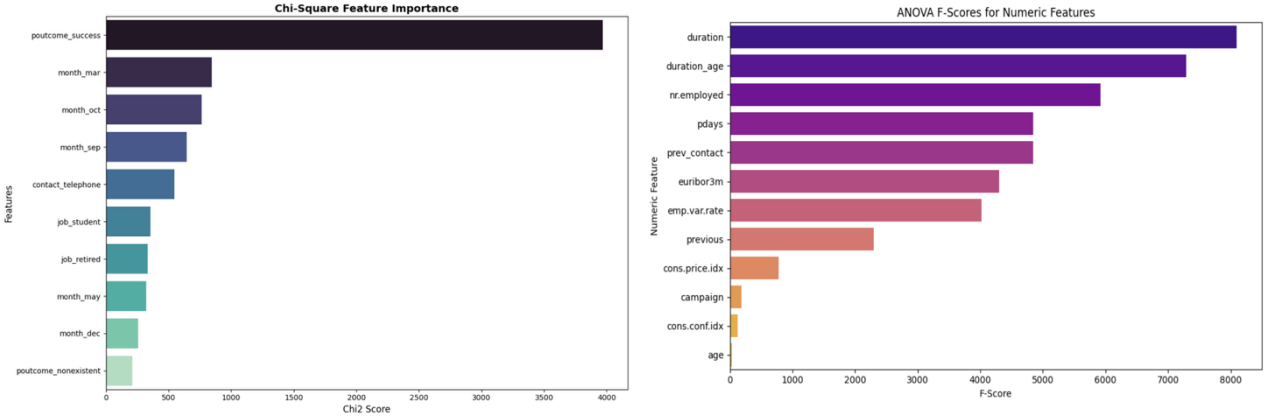
These manufactured characteristics add to the information by incorporating behavioural and contextual links, hence boosting model interpretability and predictive power.

One-Hot Encoding

One-hot encoding was used to transform category information into numerical formats appropriate for machine learning models. Each category inside a feature (e.g., employment type, marital status, education, and contact month) was converted into a binary column with values of 0 or 1. This ensures that the model can read categorical data while avoiding ordinal bias. One-hot encoding increased the model's interpretability and prediction performance by clearly depicting each category.

Feature Selection

Two statistical tests were used to identify the most influential predictors:



The combined findings of the Chi-Square and ANOVA studies show that both category and numerical factors play an important influence in determining customer subscription behaviour.

Previous campaign success, contact month, and method of communication identified as the biggest categorical predictors of consumer reaction, emphasising the relevance of past involvement and timeliness in marketing results.

Call duration and economic stability parameters such as employment fluctuation rate and staff count had the greatest statistical significance, demonstrating that both customer engagement depth and external market circumstances impact conversion likelihood.

These findings jointly led feature selection for predictive modelling, ensuring that only the most statistically important variables were maintained to construct accurate and understandable models.

Outlier Detection and Treatment

	count	mean	std	min	25%	50%	75%	max
duration	41168.0	254.406115	235.232414	11.00000	102.000	180.000	319.000	1271.330
duration_age	41168.0	10147.833609	9813.154819	414.00000	3871.000	6956.000	12692.000	53460.000
nr.employed	41168.0	5167.052308	72.230165	4963.60000	5099.100	5191.000	5228.100	5228.100
pdays	41168.0	962.509862	186.815733	3.00000	999.000	999.000	999.000	999.000
prev_contact	41168.0	0.036752	0.188154	0.00000	0.000	0.000	0.000	1.000
euribor3m	41168.0	3.621520	1.734135	0.65768	1.344	4.857	4.961	4.968
emp.var.rate	41168.0	0.081937	1.570960	-3.40000	-1.800	1.100	1.400	1.400

Outlier treatment was used on numeric variables such as age, duration, campaign, pdays, prior, cons.conf.idx, euribor3m, nr.employed, duration_age, and emp.var.rate to lessen the impact of extreme values on model performance.

A percentile capping approach (1st-99th percentile) was used, with any data points lying below or above these thresholds replaced. This winsorisation method maintains the general data distribution while avoiding outliers from distorting statistical correlations.

Following treatment, summary statistics revealed that numeric variables were now within tolerable ranges, resulting in a more stable and dependable dataset for model training.

Model Development

Data Splitting

To maintain class balance, the dataset was split into 80% training and 20% testing sets using train_test_split with target variable stratification ('subscribed'). The training set was used to fit the model, while the test set confirmed performance on previously unknown data, ensuring fair evaluation and reducing overfitting.

1. **Parametric Model: Logistic Regression**

   Logistic Regression was chosen as the major parametric model because it establishes a clear and understandable link between customer attributes and subscription likelihood.

   ⇒ **The Maximum chance Estimation (MLE)** method determines the most likely set of parameters that match the data, so determining which factors greatly enhance or decrease the chance of subscription. Variables such as prior campaign success or call duration might be immediately evaluated as positive or negative influences.

   ⇒ **The Bayesian Logistic Regression** version builds on this model by including previous information and uncertainty estimation. This not only eliminates overfitting but also measures the model's confidence in each parameter estimate. It is especially useful in marketing datasets where class imbalance and unpredictability are common ensuring steady and dependable forecasts.

2. **Non-Parametric Model: Random Forest Classifier**

   The Random Forest Classifier was employed as a non-parametric model to detect complicated, non-linear interactions that Logistic Regression can overlook.
   It provides robust, high-accuracy predictions by combining many decision trees, preventing overfitting and successfully managing high-dimensional data. Its feature significance output identifies important variables such as call duration, past campaign performance, and employment rate, transforming model findings into usable, data-driven marketing insights that aid in prioritising high-impact customer groups and optimising campaign tactics.

Model Evaluation

Model performance was assessed using the following metrics:

⇒ Accuracy, Precision, Recall, F1-Score, and ROC-AUC, ensuring both predictive power and class discrimination.
⇒ Visual Diagnostics: ROC and Precision–Recall curves were used to assess performance consistency.
⇒ Feature Importance & Coefficient Analysis: Identified key drivers influencing subscription behaviour.

Customer Segmentation (Clustering)

K-Means clustering was used as an unsupervised learning approach to group customers with similar behavioural and demographic characteristics. The objective was to identify various customer segments that may inform focused marketing tactics and customised outreach.

Before clustering, the selected features (both categorical and numerical) were standardised with StandardScaler to ensure that all variables received equal weight during distance-based computations. Only numeric characteristics were scaled, leaving dummy-coded category features binary (0/1).

Two complimentary strategies were used to find the ideal number of clusters:

- ⇒ The Elbow Method plots the Within-Cluster Sum of Squares (WCSS) to determine when more clusters result in declining benefits.
- ⇒ The Silhouette Score assesses how well each observation aligns with its allocated cluster relative to others.

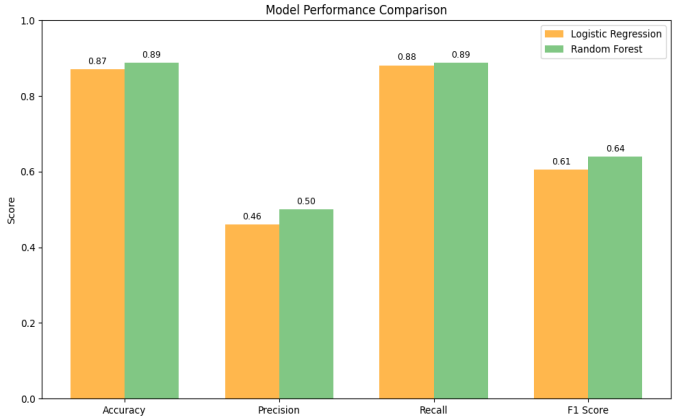
Both measurements indicated that k = 4 gave a well-balanced solution with compact and well-separated clusters. Once the optimal k was determined, the final K-Means model (k=4) was fitted to the standardised data, yielding four different customer groups.

To facilitate comprehension, major Component Analysis (PCA) was employed to compress high-dimensional data into two major components for cluster separation visualisation. This offered a better knowledge of group boundaries and overlap.

Result

Results

Model Performance Summary



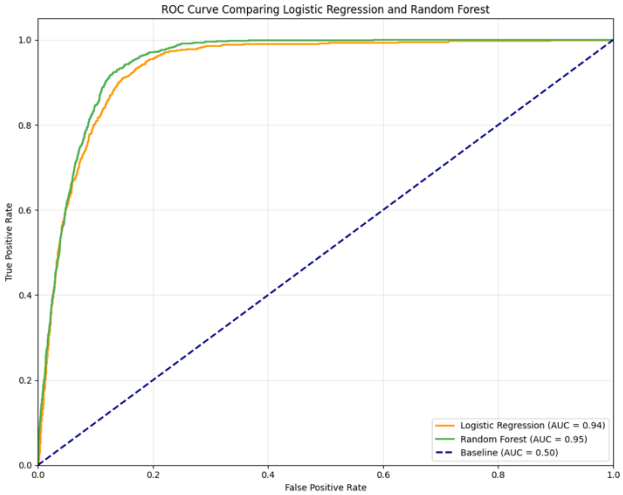
Both the Logistic Regression and Random Forest models performed well in predicting which clients were most likely to subscribe, so directly assisting profit-maximizing marketing strategies. The Random Forest model scored slightly higher on all evaluation criteria, Accuracy (0.89), Precision (0.50), Recall (0.89), and F1-Score (0.64) than Logistic Regression, which scored Accuracy (0.87) and F1-Score (0.61).

This improvement has significant implications for the economy. A higher recall rate indicates that the Random Forest model identifies a greater proportion of potential subscribers, which boosts campaign conversion rates and income possibilities. Higher precision indicates that fewer non-responsive clients are approached, lowering operational costs such as call time, personnel, and advertising spend. Together, this results in a more efficient allocation of marketing dollars, with each consumer encounter having a higher chance of conversion.

Although Logistic Regression demonstrated significantly lower predictive strength, its interpretability provides economic insight into marginal effects, assisting managers in determining which variables, such as call time or employment stability, have the greatest influence on customer decisions. This insight informs price strategies, channel prioritisation, and client targeting, resulting in increased long-term customer value.

To summarise, Random Forest maximises predictive profitability, but Logistic Regression improves strategic and policy-level decision-making. Combining the two models enables businesses to balance short-term campaign efficiency with long-term strategic insights, resulting in more sustained marketing performance and a higher total return on investment (ROI).

Model Comparison



The ROC Curve evaluates how well each model distinguishes between clients who are likely and unlikely to subscribe. Both models show good predictive power, with the Random Forest model having a slightly higher AUC (0.95) than Logistic Regression (0.94). This suggests that Random Forest has a somewhat better ability to accurately identify clients who are really interested in subscribing.

From an economic and marketing standpoint, this performance translates into more precise targeting, resulting in decreased outreach to uninterested customers (false positives) and more potential subscribers (true positives).

This means that marketing campaigns driven by the Random Forest model are more likely to target responding customers, resulting in increased subscription rates and a higher return on

investment (ROI).

Although Logistic Regression performed well, its interpretability provides useful insights into why customers subscribe, allowing managers to tailor strategies to the most significant aspects. In conclusion, while both models are useful, Random Forest is a more dependable and cost-effective method for estimating subscription likelihood and optimising campaign effectiveness.

Estimation Method (Bayesian Method)

The Bayesian Logistic Regression model has an accuracy of 86.6% and a ROC-AUC of 0.93, demonstrating a great capacity to differentiate between customers who are likely and unlikely to subscribe.

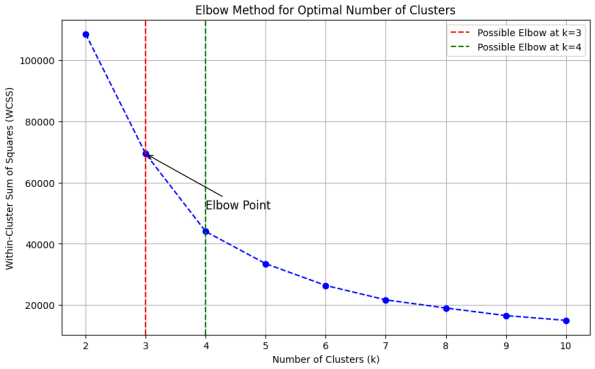
```
Evaluation Metrics for Bayesian Logistic Regression:
Accuracy: 0.8655574447413165
ROC-AUC Score: 0.932914294032307
Confusion Matrix:
 [[6353  954]
 [ 153  774]]
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.87      0.92      7307
           1       0.45      0.83      0.58       927

    accuracy                           0.87      8234
   macro avg       0.71      0.85      0.75      8234
weighted avg       0.92      0.87      0.88      8234
```

The recall of 0.83 for the subscribed class (1) indicates that the model successfully detects most potential subscribers, making it ideal for marketing applications where missing an interested consumer is more expensive than contacting a non-subscriber.

Although the accuracy (0.45) for subscribers is modest, which means that some false positives occur, this trade-off is acceptable in marketing scenarios because greater outreach can still result in meaningful conversions.

Economically, this result suggests that the model may optimise campaign resource allocation by prioritising clients with high estimated subscription probability. By precisely evaluating leads, the company may eliminate wasted calls and focus promotional spending on high-value prospects, hence increasing total return on marketing investment (ROMI).

Customer Segmentation (Clustering)



The K-Means clustering method was used to identify separate client groups based on behavioural, demographic, and economic characteristics. Using the Elbow Method, four clusters (k=4) were chosen as the best segmentation points, balancing model simplicity and explanatory power. This judgement was reinforced by the silhouette score, which peaked at k=4, suggesting well-separated and internally coherent clusters.

⇒ Young Professionals are often early-career professionals with moderate call duration and participation levels.

⇒ Mid-Career Segment: Stable income earners with variable responses to advertising.

⇒ Retirees: older customers with minimal economic activity but stable subscription patterns.

⇒ High-Value Customers: Highly engaged and financially active clientele with high subscription potential.

These clusters offer actionable insights for focused marketing campaigns. Personalised outreach to high-value and mid-career groups, for example, can raise campaign ROI, and engagement techniques targeted to younger

clients can improve long-term retention.

Business Questions

1. **How does communication type affect campaign success?**

Both the Chi-Square test and the model coefficients revealed that communication method is a statistically significant factor influencing consumer reaction.

Customers contacted using cellular channels had a greater subscription rate than those reached via regular phone calls. The Random Forest feature significance verified this tendency, implying that current and flexible communication options result in higher engagement and conversion.

Economically speaking, dedicating marketing funds to mobile-based campaigns can improve cost efficiency and response rates, particularly among younger and mid-career client clusters identified through segmentation.

2. **Which customer segments show the highest likelihood of subscribing?**

The K-Means clustering (k=4) identified four distinct consumer segments: young professionals, mid-career customers, retirees, and high-value customers.

The High-Value and Retiree clusters had the highest subscription likelihood, as evidenced by significant past campaign performance and extended engagement durations.

Young Professionals, on the other hand, shown future conversion potential but are now less committed, most likely owing to a lack of financial security or a shorter relationship history.

From a corporate standpoint, this segmentation allows for targeted advertising focussing retention efforts on loyal, high-value groups while customising awareness initiatives for younger demographics.

3. **Which factors most strongly influence subscription decisions, and how well can models predict them?**

Feature selection methods (Chi-Square and ANOVA) and predictive modelling consistently highlighted call duration, previous campaign outcome, employment rate, and contact timing as the most influential predictors of subscription.

The Random Forest model achieved the highest predictive accuracy (AUC = 0.95), outperforming Logistic Regression (AUC = 0.94) and Bayesian Logistic Regression (AUC = 0.93), indicating strong model generalisation and reliability.

This shows that campaign success can be predicted with high confidence using these key variables.

From a strategic view, optimising engagement quality (longer calls), maintaining contact with previously successful customers, and targeting periods of economic stability will likely yield the highest returns on marketing investment.

Conclusion

Conclusion

This project effectively converted a large-scale marketing campaign dataset into relevant business analytics, improving both strategic and operational decision-making. The investigation identified the important behavioural and contextual elements driving customer subscriptions through rigorous data cleaning, feature engineering, and modelling.

The Random Forest Classifier emerged as the most successful prediction model, with an AUC of 0.95 and the ability to recognise non-linear correlations between campaign variables. Its better memory and precision show a significant capacity to differentiate potential subscribers while increasing marketing efficiency. In contrast, Logistic and Bayesian Logistic Regression provided interpretability by demonstrating how variables such as past campaign performance, call duration, and economic stability indicators (e.g., employment rate and euribor3m) had a direct impact on consumer decisions. Together, these models form a well-balanced framework, with Random Forest for high-performance targeting and Logistic Regression for policy transparency.

The K-Means segmentation enhanced the findings by identifying four unique consumer clusters: young professionals, mid-career customers, retirees, and high-value clients. This segmentation enables diverse marketing techniques, such as cultivating high-value and retiree consumers for quick ROI, as well as developing engagement programs for younger professionals to encourage long-term commitment.

From a business standpoint, three core insights stand out:

- ⇒ Prior campaign success is the strongest predictor of future subscription re-engaging these customers ensures higher conversion at lower cost.

- ⇒ Timing matters campaigns in March, September, and October consistently yield better performance, guiding optimal scheduling.

- ⇒ Contact method and engagement depth cellular communication and longer call durations significantly enhance response rates, underlining the importance of personal, real-time outreach.

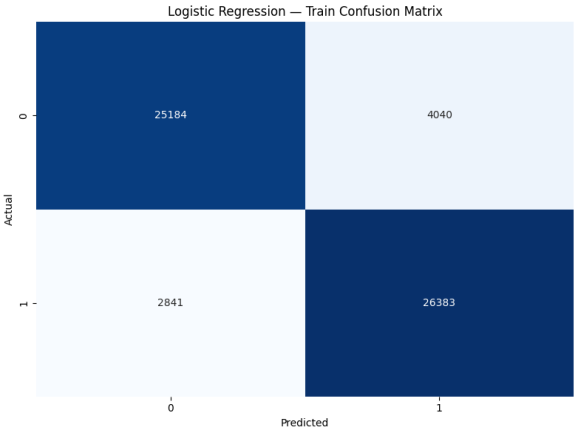
Overall, this study shows that data-driven marketing may significantly enhance efficiency and ROI when combined with sophisticated predictive analytics and segmentation. The findings not only guide resource allocation and targeting strategy but also provide a reproducible analytical foundation for ongoing campaign improvement. To maintain a competitive edge in client acquisition and retention, the organisation should implement an integrated marketing analytics pipeline that includes model-driven forecasts, real-time monitoring, and periodic re-segmentation.

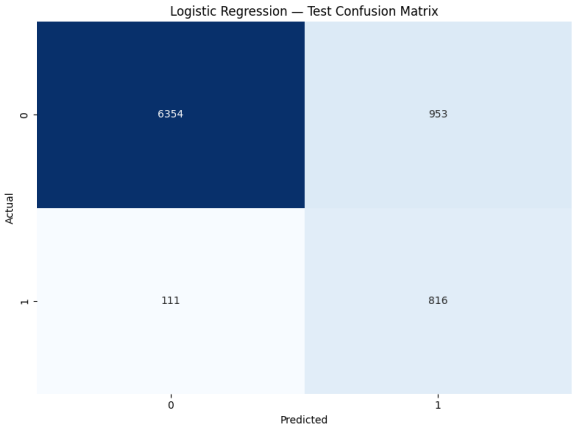
References

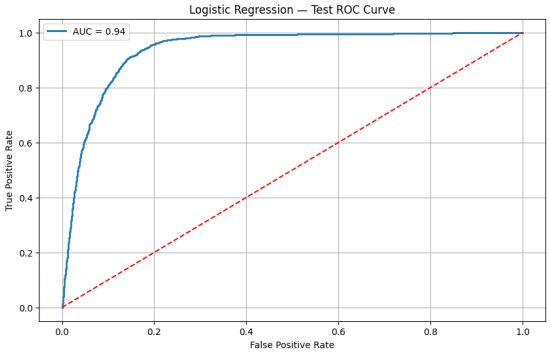
1. Bank Marketing Data Set. (2014). *UCI Machine Learning Repository*. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
2. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
3. Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
4. Kotler, P., & Keller, K. L. (2016). *Marketing management* (15th ed.). Pearson Education.
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
6. Song, P. (2024, July 1). Predicting marketing campaign success with machine learning. *ML Journey*. https://mljourney.com/predicting-marketing-campaign-success-with-machine-learning/
7. Simpson, D. (2024, September 22). Predictive modeling in marketing: The ultimate guide. *Lift AI Insights*. https://www.lift-ai.com/blog/predictive-modeling-marketing
8. Wang, G. (2025, February 7). Customer segmentation in digital marketing using a Q-learning based differential evolution algorithm integrated with K-means clustering. *PLOS ONE*, 20(2). https://doi.org/10.1371/journal.pone.0318519
9. Haixiang, S. (2025, June 30). Machine-learning models for customer-behavior analytics: A review. *Theory and Practice of Science and Technology*, 6(6), 24-29. https://doi.org/10.47297/taposatWSP2633-456905.20250606

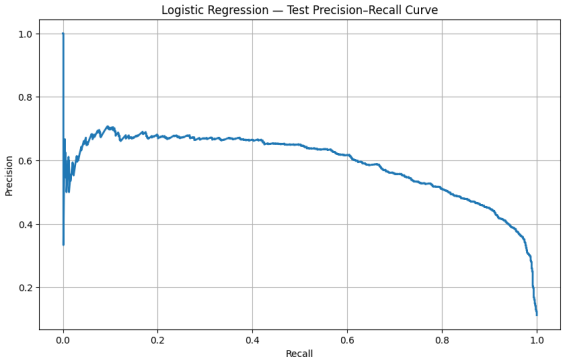
Appendices

1. Logistic Regression

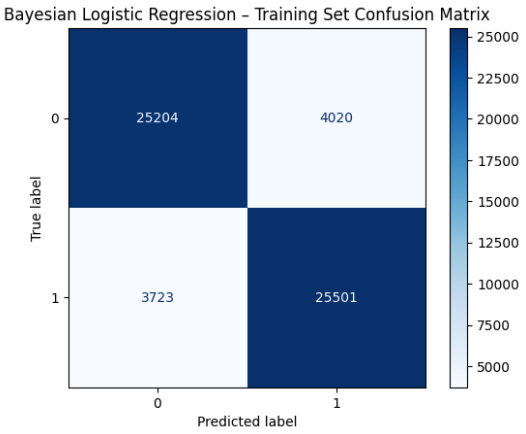


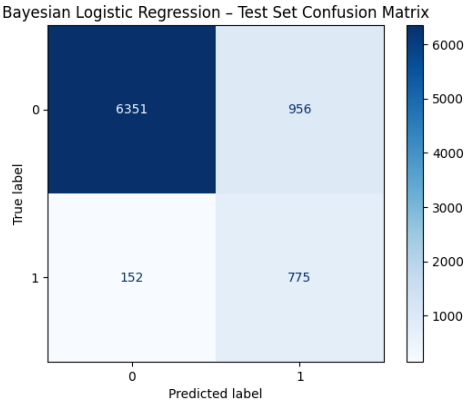


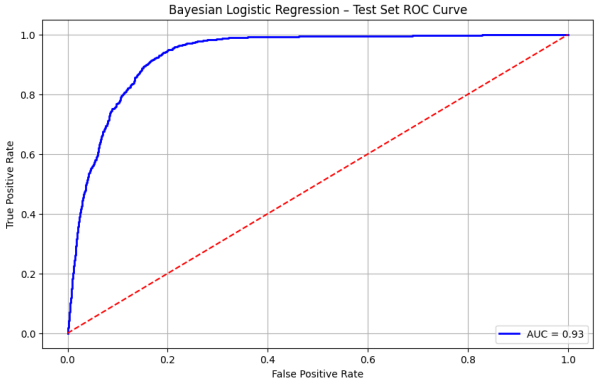


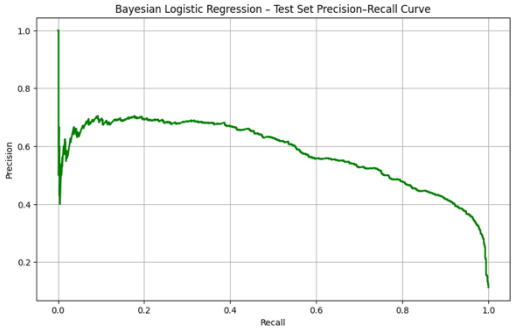


2. Bayesian Logistic Regression









3. Random Forest Regression