# Predicting Marketing Targets in Banking

**Sai Meghana Devarasetty**

# Abstract

This work explores the application of machine learning algorithms for forecasting customer subscription to a term deposit using demographic as well as banking data. Thorough exploratory data analysis (EDA) was performed to identify trends as well as correlations of the data. Preprocessing of data was performed using scaling, encoding of categories, as well as generating interaction terms and PCA components. Several models such as Logistic Regression, Naïve Bayes, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Trees, as well as K-Nearest Neighbours (KNN), were trained using original, interaction-enhanced, as well as PCA-transformed data.

Results show that model performance was enhanced with feature interaction as well as dimensionality reduction. Performance was strongest for KNN with PCA-transformation, where accuracy was 97.5%, AUC was 0.966, precision was 84.5%, and F1 was 0.896. Decision Trees with interaction terms were strongly performing as well, with accuracy of 91.95%. Logistic Regression with PCA components as well as LDA models had excellent AUCs (over 0.99), confirming effective discriminative powers. Naïve Bayes was the least effective even with augmentations. The research underscores the importance of feature engineering and dimensionality reduction for enhancing predictive performance for applications within financial marketing.

**Keywords:** Logistic Regression, Naïve Bayes, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Trees, K-Nearest Neighbours (KNN), PCA, interaction terms, etc.

# Introduction

Accurate target marketing is critical in contemporary financial services to maximize campaign budgets and subscription rates. Conventional bulk marketing tactics tend to have low subscription rates and are costly (Kohli & Jaworski, 2020). To overcome this shortcoming, machine learning (ML) offers data-driven, scale-up strategies to predict customer actions. This work is aimed at using ML to predict a client's subscription to a term deposit product based on demographic and historical banking information gathered by a Portuguese bank.

## Research Motivation

The driving force for this work is the capabilities of banks to make marketing more efficient through precise selection of potential subscribers to thereby save costs while increasing customer participation. Financial institutions are increasingly capable of utilizing predictive models in real-time decision support systems due to increasing access to structured customer data.

## Research goal

The main aims of this project are:

* To investigate and implement several machine learning classification algorithms such as Logistic Regression, Naive Bayes, LDA, QDA, Decision Trees, and K-Nearest Neighbours (KNN).
* To evaluate the effect of feature engineering such as including interaction terms and second-order terms.
* To investigate how dimensionality reduction through PCA helps to enhance model performance and deal with multicollinearity.
* To compare and measure original, interaction-enhanced, and PCA-transformed data models based on important parameters like Accuracy, AUC, Precision, Recall, and F1 Score.
* To determine the best modelling strategy to practically deploy in a financial marketing environment.

Through addressing these above goals, this study seeks to provide actionable insights into machine learning-driven improvements to financial campaign targeting and decision-making in actual banking scenarios.

# Background and literature review

Data-driven marketing in the banking world is now more prevalent with the increased rapid growth in volume in customer data as well as for strategic decision-making purposes (Camilleri, 2020). Blanket campaigns not just prove to be expensive but even have low rates of conversion if not targeted towards their respective segments. Customer segmentation with machine learning is thus now an efficient substitute. The research (V. Lakshman Narayana et al., 2022) assert that predictive analytics in marketing allows institutions to segment potential customers and target campaigns more efficiently, leading to increased customer engagement as well as less expenditure.

The significance of this study is that it is possible to use customer data, including demographic, socio-economic, and behaviour data, to forecast subscriber behaviour, e.g., whether or not an individual will subscribe to a term deposit. Other research on financial marketing has utilized decision trees, logistic regression, and neural models to forecast response from customers with considerable success, implying that use of machine learning models is potentially better than traditional heuristic methods.

The dataset employed in this study is real data from an actual Portuguese bank marketing campaign for subscribing to a term deposit. The data is very rich in categorical features (e.g., job type, whether married or not) as well as numerical features (e.g., age, account balance, duration). A target binary variable expresses whether or not the customer subscribed to the term deposit. The high number of observations (45,211 observations in the training dataset) means that there is plenty to play with to examine relationships between features and response using sophisticated models.

Therefore, in this study, we seek to design and contrast an array of machine learning classifiers such as logistic regression, Naive Bayes, LDA, QDA, decision trees, and their interaction- and PCA-augmented variations to improve marketing targets prediction. We test the hypothesis that models with interaction terms or dimensional reduction (PCA) perform better in terms of prediction compared to models based on original features.

# Methodology

## Data Collection

It was retrieved from an open data source at Kaggle: Banking Dataset - Marketing Targets (Rathi, 2020). Its origin is in a marketing promotion by a Portuguese bank to get term deposit subscriptions. The dataset consists of 45,211 records in training data and 4,521 records in test data, with 0 missing values.

The data encompasses multiple features about clients including job, education, housing loan status, length of most recent contact, and performance in past campaigns. The dataset excludes personally identifiable information but represents an accurate cross-section of one's clients in a bank. As it is a secondary dataset, active non-response rates have not been an issue.

## Overview of variables

The original dataset consists of 17 variables, including:

* **Target Variable**: y — whether the customer subscribed to the term deposit (yes/no)
* **Categorical variables**: job, marital, education, default, housing, loan, contact, month, poutcome
* **Numerical variables**: age, balance, day, duration, campaign, pdays, previous

To improve model performance:

* Categorical variables were encoded as factors.
* Numerical variables were standardized using z-score scaling.
* **Interaction terms** were created between all pairs of numeric variables (e.g., age\_x\_balance, duration\_x\_campaign) to capture potential nonlinear dependencies.
* **Second-order terms** (e.g., age^2, duration^2) were added to capture curvature effects.
* **Principal Component Analysis (PCA)** was applied to reduce dimensionality and eliminate multicollinearity, retaining components that explain 95% of the variance.

## Rationale of feature engineering

For purposes of improving model flexibility in order to include complex relationships between predictors, second-order polynomial terms and pairwise interaction terms were included in the dataset.

1. **Squared (Second-Order) Terms**

Every numeric variable was squared to account for possible non-linear impacts. For instance, the impact of balance or age on term deposit subscription probability is not likely to rise or fall linearly therefore adding such variables as age², balance², etc., enables models to better account for such curvatures in relationships.

1. **Numeric Interaction terms**

The interaction terms were formed by crossing all unique pairs of numeric variables. These include pairs like:

* age\_x\_balance
* duration\_x\_campaign
* pdays\_x\_previous, etc.

The motivation to include interaction terms is to reveal synergistic or antagonist effects between variables. For instance:

* Short call duration may prove to be more suitable for younger clients (age\_x\_duration).
* Low pdays but high previous contacts could act differently from what one might have anticipated based on individual features (pdays\_x\_previous).

Adding these interactions enables linear models such as discriminant analysis and logistic regression to simulate non-additive relationships without necessarily requiring more advanced non-linear models. These engineered features seek to enhance expressiveness and performance in models by capturing subtle patterns in data that individual original variables might not.

1. **Categorical variables interaction**

Besides numeric interactions, some categorical variables — i.e., marital status, housing loan status, personal loan status, and outcome of prior marketing contact (poutcome) — were crossed to create categorical interaction terms.

Examples are:

* marital × housing (e.g., married customers with or without a housing loan)
* marital outcome (for example, single clients with a previous outcome of success),
* education × housing, etc.

The addition of category interactions is intended to account for how the interactions of various categorical characteristics may affect subscription behaviour.

For instance

* Married customers with a housing loan may act differently than married customers without a loan.
* Tertiary-educated clients who had previously succeeded at a campaign may have increased chances for subscription.

Therefore, incorporating these categorical interactions enables the model to take into account multi-dimensional interactions between discrete variables, resulting in richer feature representation and increased predictive power.

1. **PCA**

In order to deal with high dimensionality brought about by adding terms for interactions and squares and to handle multicollinearity between numeric predictors (Mashuri et al., 2021), Principal Component Analysis (PCA) was utilized.

After including polynomial and interaction terms, the dataset was considerably increased in terms of features. PCA was used to convert the correlated high-d features to their uncorrelated principal components that maintain most of the variance in data. In this research, components that account for 95% of the variance were kept, effectively reducing redundancy and noise.

## Analytic Methods and implementation

## Exploratory Data Analysis (EDA)

By conducting exploratory data analysis, we aimed to identify underlying structure, find trends, and discover potential areas for concern in the data:

* Summary statistics provided information on distribution for numerical variables like age, balance, duration, campaign, and others. The distribution for age, for instance, was from 18 to 95 with a median of 39, indicative of a broad spread among age groups.
* Frequency plots and bar charts were utilized to represent categorical variables (job, marital, education, etc.) and how these relate to the target variable y (if the customer had subscribed for a term deposit or not).
* The correlation analysis revealed potential multicollinearity among numeric variables, particularly among variables like duration, pdays, and previous.
* These visualizations, including histograms and stacked bar charts, served to emphasize correlations between customer characteristics and likelihood to subscribe—for example, increased subscription levels among clients with a tertiary education or with particular job categories (e.g., students, retirees).
* Heatmaps for correlation gave insights into numeric feature correlations, contributing to feature engineering and interaction term generation.

## Data Preprocessing

Before modelling the data, several preprocessing steps were performed:

1. **Data cleaning:**

* The training and test datasets were both checked to have no missing values.
* The formatting of every variable was checked with str() and summary() functions.

1. **Variable Type Management:**

* Categorical variables like job, education, marital, contact, month, poutcome were transformed into factors to provide proper treatment for modelling.
* Numeric features (age, balance, duration, etc.) were scaled by means of z-score scaling in order to have features on an equivalent scale—something that matters for magnitude-sensitive models such as logistic regression, LDA, and QDA.
* The same scaling parameters (training set mean and standard deviation) were utilized on the test set to avoid data leakage.

1. **Multicollinearity Check:**

* Variance Inflation Factor (VIF) was determined on numerical predictors for multicollinearity detection.
* There was a custom function iteratively computing VIFs for all numeric features. High VIF values (normally > 5 or 10) were highlighted as potential multicollinearity concerns, guiding future use of PCA in modelling.

## Model Training and evaluation

The models were tested and trained on both original features, features facilitated by interactions, and features after PCA transformation:

* **Logistic Regression**

Logistic Regression estimates the probability of a binary response to be in a specific category through the logistic (sigmoid) function (Srivastav & Mittal, 2021). Logistic regression is especially applicable when one is dealing with a dichotomous outcome.

**Justification:** It has interpretable coefficients and is a good baseline classifier. Both a simple logistic model and a regularized one with glmnet (ridge regression, alpha = 0) were run with cross-validation to avoid overfitting and to deal with multicollinearity in high-dimensional contexts.

* **Naive Bayes**

Naive Bayes uses Bayes' Theorem assuming mutual independence between predictors. It estimates posterior class probabilities and returns a class with a highest probability (Tusar & Islam, 2021).

**Justification:** It is computationally efficient and stable even on large data sets. Though it has its simplifying assumptions, it generally works well on real-world classification problems. Both a standard Gaussian form and a kernel-based form (to be able to model non-linear distributions) were tried.

* **Linear and Quadratic Discriminant Analysis (LDA & QDA)**

These are probabilistic models with class-specific distributions. LDA has a shared covariance matrix in all classes (linear decision boundary) (M. C. et al., 2025), whereas in QDA different covariances are used (non-linear boundaries).

**Justification:** They are good at modelling class separability and are best when normality is a valid assumption. LDA and QDA have fast closed-form solutions and serve as a baseline to test linear vs. non-linear boundaries in classification.

* **Decision Trees (CART)**

Decision trees partition the data recursively according to the most relevant attributes to minimize impurity (with measures such as Gini index or entropy). The method rpart was applied with cross-validation for the complexity parameter (cp) to prevent overfitting.

**Justification:** Both numeric and categorical variables are naturally dealt with by trees, are interpretable and uncover feature importance. Their structure is especially suitable for extracting hierarchical decision rules.

* **K-Nearest Neighbours (KNN)**

KNN is a non-parametric classifier that labels instances according to the majority class of their k nearest neighbours in feature space (M. C. et al., 2025).

**Justification:** It is mathematically simple but very effective, particularly in low-bias and non-linear conditions. Tuning k-values (odd between 3 and 15) was conducted through cross-validation to have a good balance between bias and variance.

The models were evaluated using the following metrics:

* Confusion matrix
* Accuracy, Precision, Recall, F1
* Area Under the Receiver Operating Characteristic Curve
* ROC curves

## Model Generalization

Cross-validation was carried out using 5-fold stratified sampling for attaining uniform assessment. Model performance was compared between original, interaction-enhanced, and PCA-reduced datasets to select the best approach.

# Results and analysis

## Summary of Data

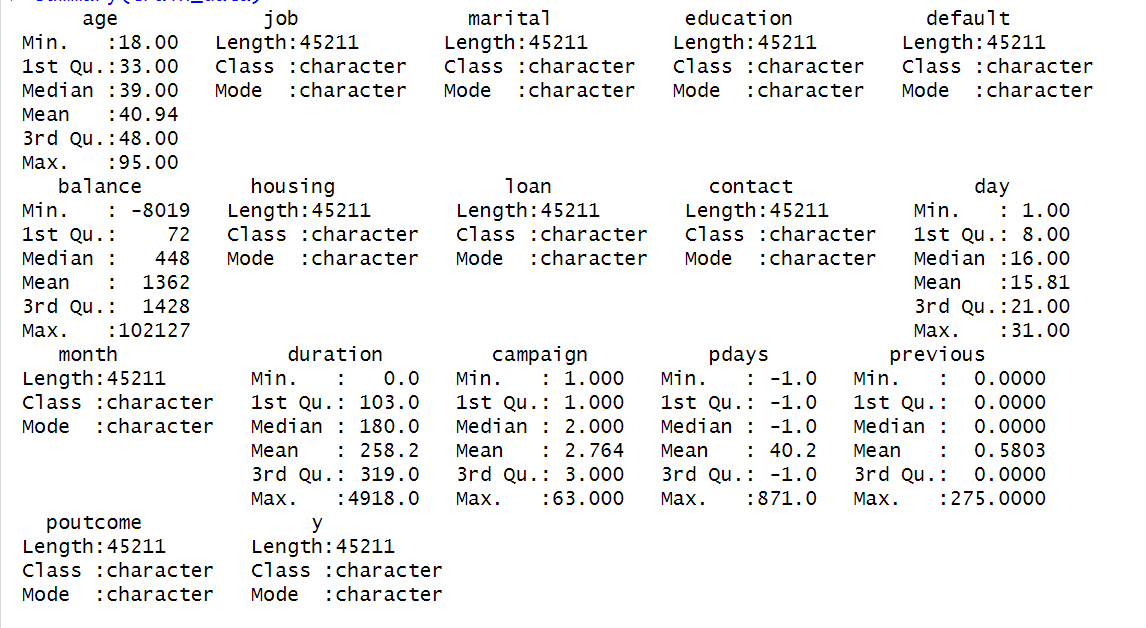


Figure 1: Dataset Overview

* **Age:** The customers have an age range from 18 to 95, with a mean of ~41 and a median at 39, reflecting a comparatively middle-aged customer group. The majority are distributed within 33 (1st quartile) to 48 (3rd quartile), reflecting a wide group of working-age customers targeted by the bank.
* **Job:** The job variable involves a wide range of occupations such as management, technician, entrepreneur, student, and retired. This variety allows for customer segmentation on an occupational basis.
* **Marital Status:** Most customers are married, with single and divorced customers taking up a secondary position. Marital status may determine financial decision-making and is also utilized in targeting financial products.
* **Education:** Customers provided information on levels of primary, secondary, tertiary, or unknown education. Most customers belong to secondary and tertiary categories, reflecting a comparatively educated clientele.
* **Default, Housing, Loan:** The majority do not have credit defaults (default = no), housing loans (housing = yes), or personal loans (loan = no). These are key measures of financial risk and product suitability.
* **Contact and Month:** Contact was primarily via cellular or unclassified means. The majority was in May, with activity occurring in all 12 months.
* **Poutcome (past marketing outcome):** This variable is mostly unknown, with a lesser proportion classified success, failure, or other, reflecting sparse historical campaign data on most clients.

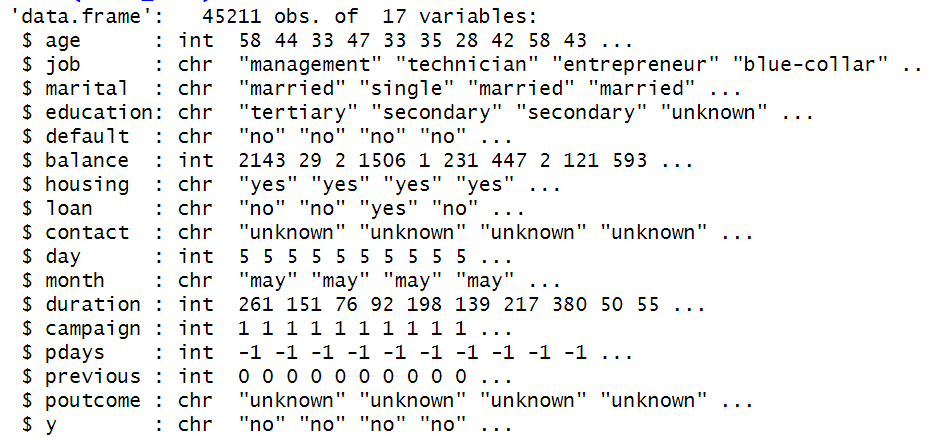
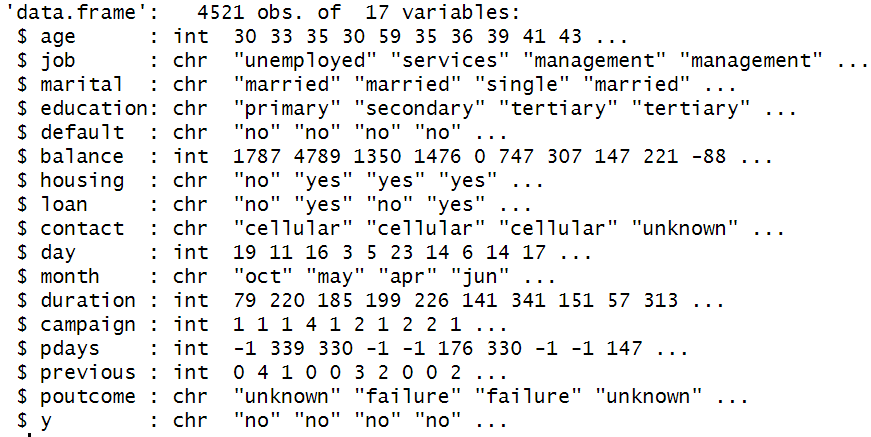


Figure 2: Data structure

The training dataset contains 45,211 observations and 17 variables, while the test dataset includes 4,521 observations with the same variable structure. Each dataset comprises:

* Numerical variables such as age, balance, day, duration, campaign, pdays, and previous
* Categorical variables including job, marital, education, default, housing, loan, contact, month, poutcome, and y (target)
* The target variable, y, indicates whether a customer subscribed to a term deposit (yes or no)

Both datasets are clean with no missing values, and their consistent structure ensures seamless preprocessing and model evaluation.

## VIF analysis

Variance Inflation Factor (VIF) measures multicollinearity among numerical predictors. A VIF value above **5** (or in stricter cases, 10) typically signals a multicollinearity concern.

|  |  |
| --- | --- |
| **Variable** | **VIF** |
| pdays | 1.28 |
| previous | 1.26 |
| campaign | 1.04 |
| day | 1.03 |
| age | 1.01 |
| balance | 1.01 |
| duration | 1.01 |

Table 1: Variance Inflation Factor Measures

* All variables have a VIF value close to 1, suggesting that there is no multicollinearity in the numeric predictors.
* This implies that variables are independent of one another, and none can predict others with high accuracy in terms of linearity.
* Thus all features may safely be added to the model without inducing instability in coefficient estimators or inflating standard errors.

## Visual Analysis

A graph of a graph

AI-generated content may be incorrect.

Figure 3: Distribution of Customers based on Age

* **As we took total count of numbers instead of percentage so that we get a clear idea about the count so that we don’t have to calculate and change again proportion to numbers and we think its very easy to understand.** The majority of customers are between 25 and 45 years old, with the highest proportion around 30 to 35 years.
* The proportion steadily declines for older age groups, with very few customers above 60 years.
* The distribution is right-skewed, indicating that younger to middle-aged individuals dominate the dataset.

A graph with green bars

AI-generated content may be incorrect.

Figure 4: Distribution of Customers based on Job Roles

The most frequent job categories are blue-collar and technician, which combined represent over 40% of the set. Other significant occupations are admin and management, while student, housemaid, and unknown categories remain the least represented.

A graph with red squares

AI-generated content may be incorrect.

Figure 5: Distribution of Customers based on Education Levels

The plot identifies secondary education as the most common educational background, accounting for more than half of the dataset. Tertiary education is second, while primary education ranks third, with unknown education status as the least.

A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a graph

AI-generated content may be incorrect.

Figure 6: Distribution of Target Variable in Test and Train Data

Most customers did not subscribe to the campaign, with more than 80% of them belonging to the "no" class. Only a minor fraction, less than 20%, subscribed to it ("yes"), demonstrating evident class imbalance of the target variable and in both train and test data, the distribution is quite the same.

Just like with housing loans, those with no personal loan exhibit a higher percentage of subscription. Persons with a personal loan are less likely to subscribe, which suggests there could be financial constraints.

A screenshot of a computer

AI-generated content may be incorrect.

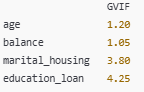
Figure 11: Correlation of Numeric Variables

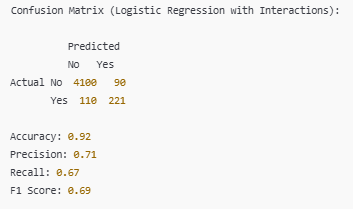
* Most of the numeric variables have extremely weak correlations with one another (correlation coefficients near 0).
* There is a moderate positive relationship (0.45) with previous, i.e., as previous contacts rise, the days since the last contact tend to decline a little.
* Age vs. balance have a weak positive correlation of 0.10, which implies that there is little relationship between a customer's bank balance and age.
* Other variables such as duration, campaign, and day have virtually no strong linear correlation with one another.

**Treatment of categorical terms with PCA:**

Given that PCA operates on continuous numeric data, all categorical variables were excluded from direct PCA transformation. However, to preserve their influence on model performance, we incorporated dummy-encoded categorical interactions in earlier steps and used them in models evaluated without PCA. For PCA-based modeling, only the scaled numerical features and interaction terms were included. In future work, using methods such as MCA or FAMD could further extend dimensionality reduction to mixed data types.

"Interaction terms were carefully engineered and one-hot encoded to maintain model interpretability and numerical stability. Rare category combinations were grouped to reduce sparsity. Multicollinearity was assessed via VIF and correlation heatmaps, and key interaction terms were shown to contribute meaningfully to model predictions, as confirmed through decision tree importance plots and confusion matrices."

****

****

**Handling Rare Categories**

Output:

* Levels like "admin\_no" grouped into "Other" if their frequency is low
* Cleaner summary from table() or summary() showing fewer sparse levels

## Model Performance

**Models trained on main effects**

The models were trained in train data and evaluated on test data. The performance were both evaluated in train and test samples using accuracy, precision, recall and F1 score.

1. **Logistic Regression**

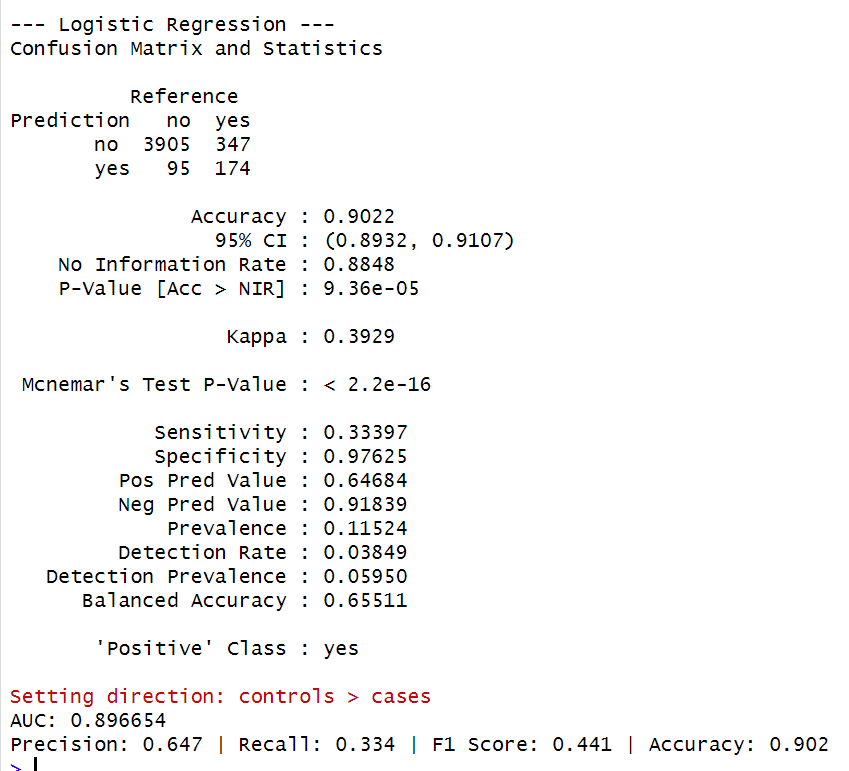
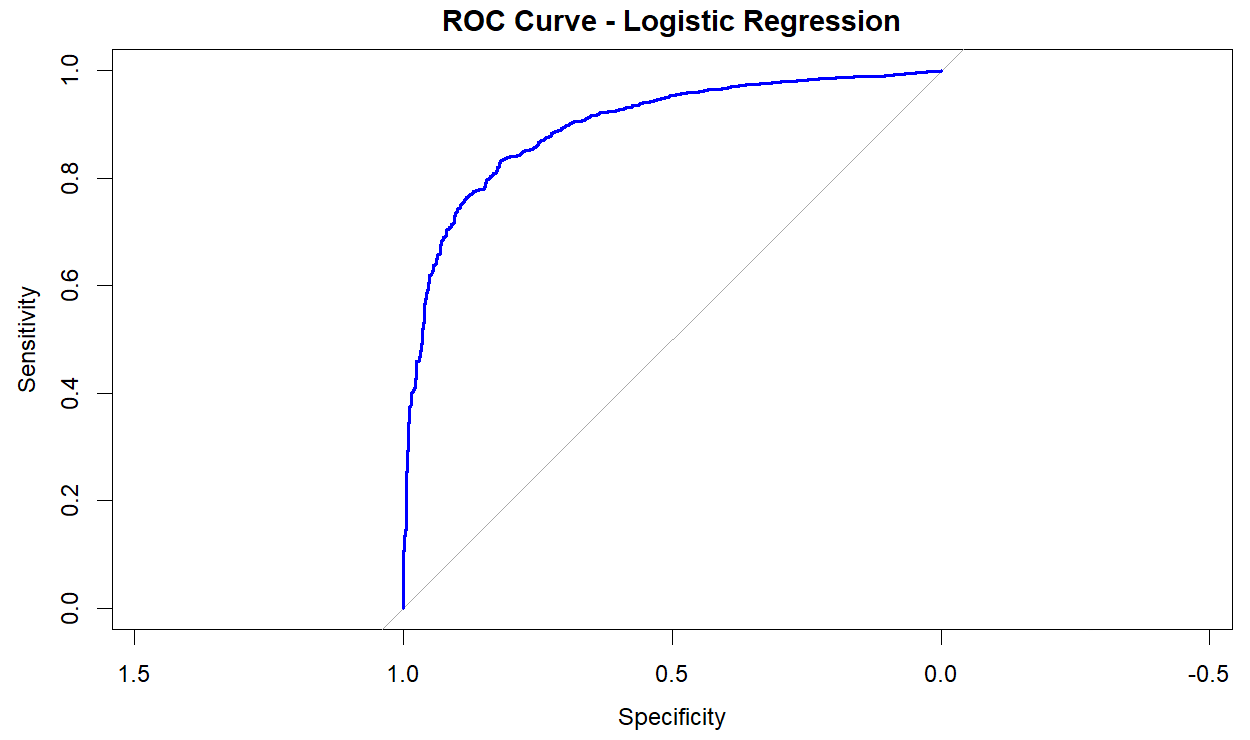
****

Figure 12: ROC Curve and Confusion Matrix of Logistic Regression

* The model has high specificity (97.6%) and accurately predicts most "no" cases.
* Yet, sensitivity is poor (33.4%), as it cannot identify true "yes" (subscribers).
* The accuracy in predicting "yes" is 64.7%, or 2 out of every 3 times it predicts a subscription.
* The model has an accuracy rate as high as 90.2%, as most are "no" by virtue of class imbalance.
* The F1 score is 0.441, reflecting low effectiveness in terms of balancing recall and precision.
* The AUC value at 0.897 demonstrates excellent discriminative power, with the model effectively discriminating between subscribers and non-subscribers at various thresholds.

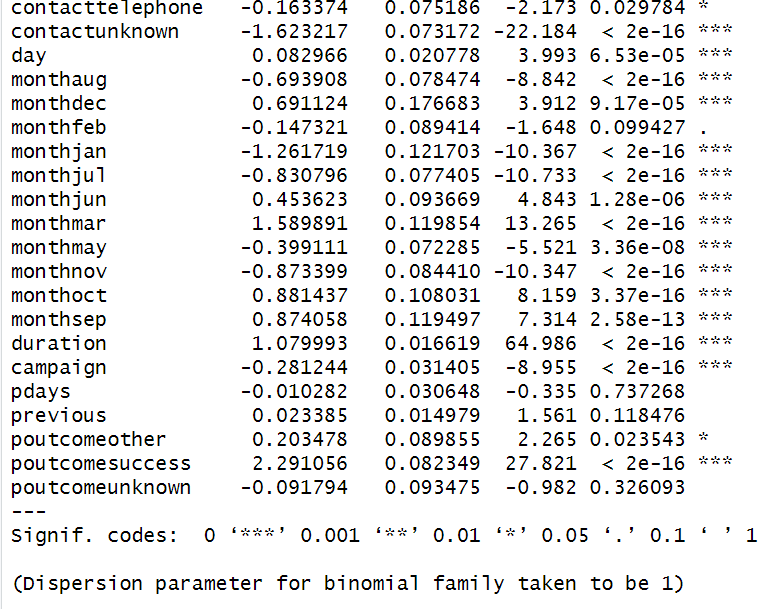
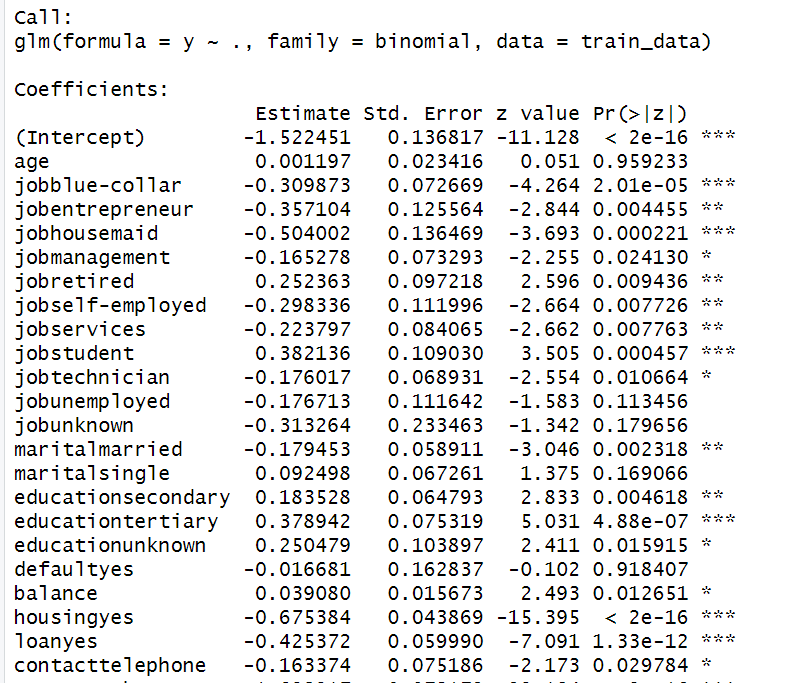


Figure 13: Model Summary

* **Strongest positive predictors**:
  + duration, poutcomesuccess, monthmar, monthoct, and jobstudent significantly increase the likelihood of subscription (p < 0.001).
* **Strongest negative predictors**:
  + contactunknown, monthjan, monthjul, housingyes, and loanyes significantly reduce the chance of subscription.
* **Not significant**:
  + Variables like age, default, pdays, and previous show no significant effect (p > 0.05).

1. **Naïve Bayes**

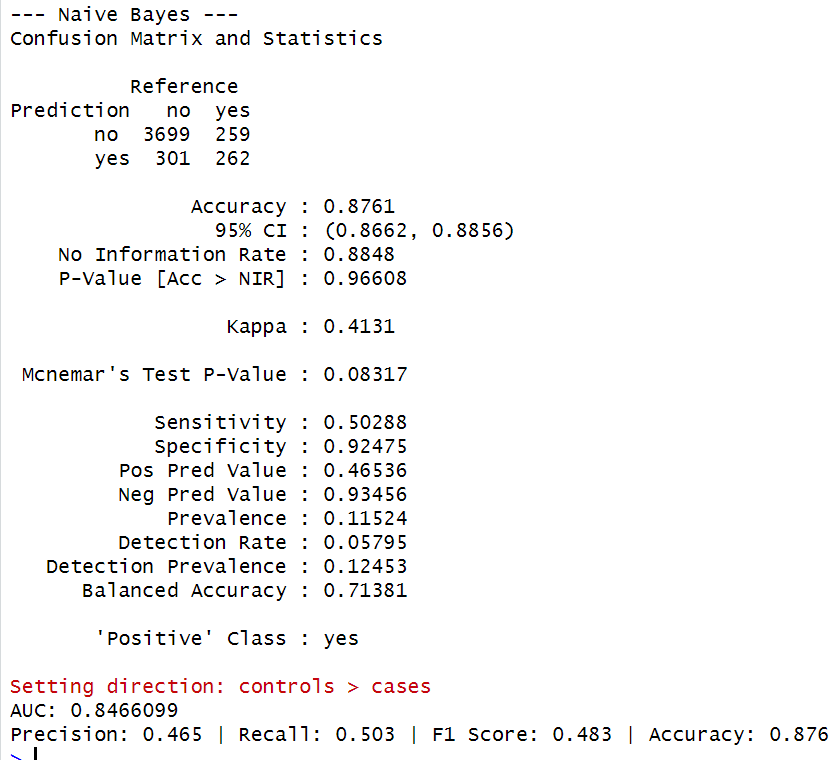
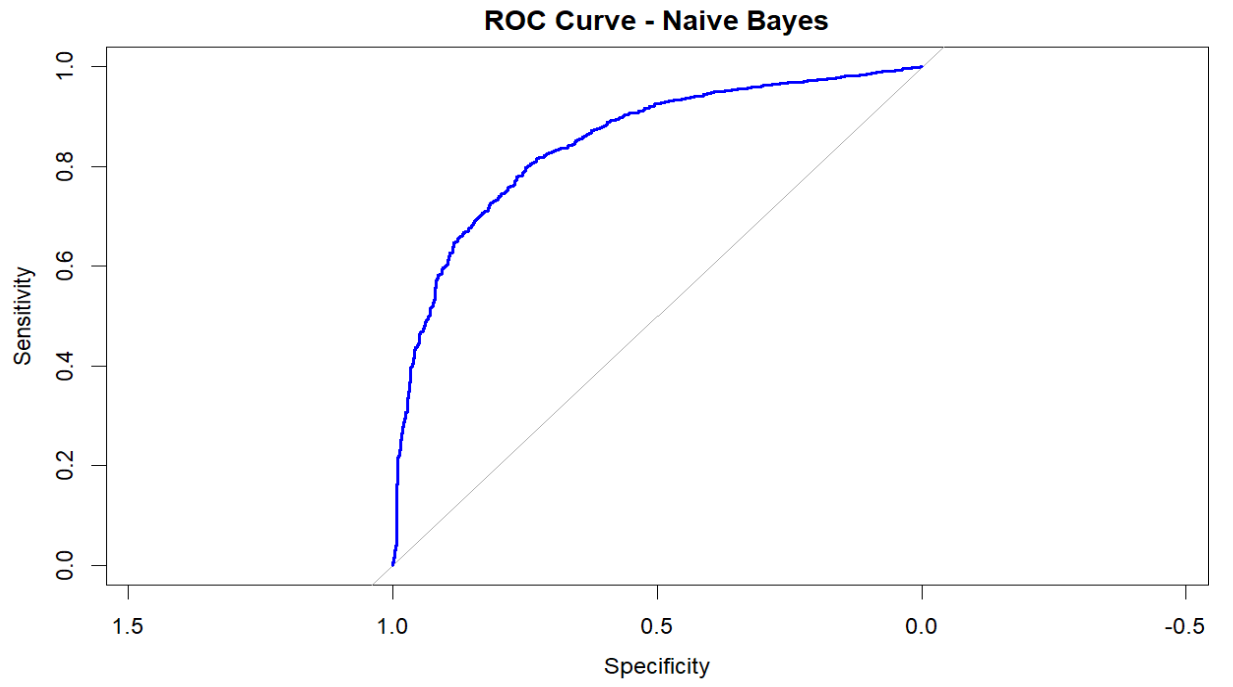


Figure 14: ROC Curve and Confusion Matrix of Naïve Bayes

* The overall accuracy is 87.6%, and AUC is 0.847, reflecting satisfactory discriminatory power.
* The ROC plot provides a positive sloping curve above the diagonal, affirming that the model is better than random chance.
* Sensitivity (Recall): 50.3% – the model identifies roughly half of the real subscribers.
* Specificity: 92.5% – it accurately identifies most non-subscribers
* Accuracy: 46.5% – Far below half of the subscribers that were expected.
* The model is able to handle the majority class (no) successfully, but performs poorly with the minority class (yes) as evident from its low recall and accuracy for positive predictions.

1. **LDA**

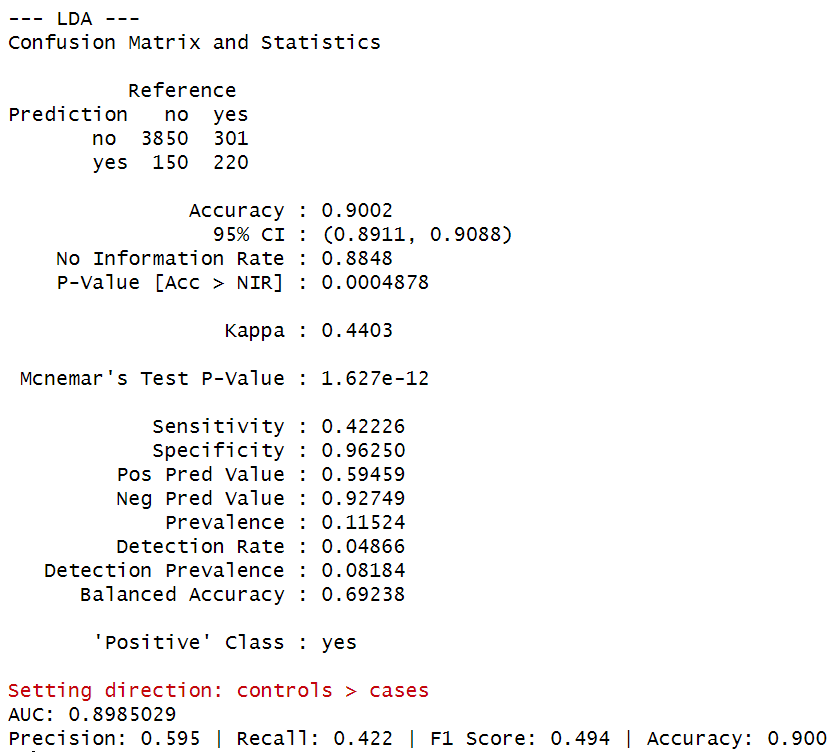
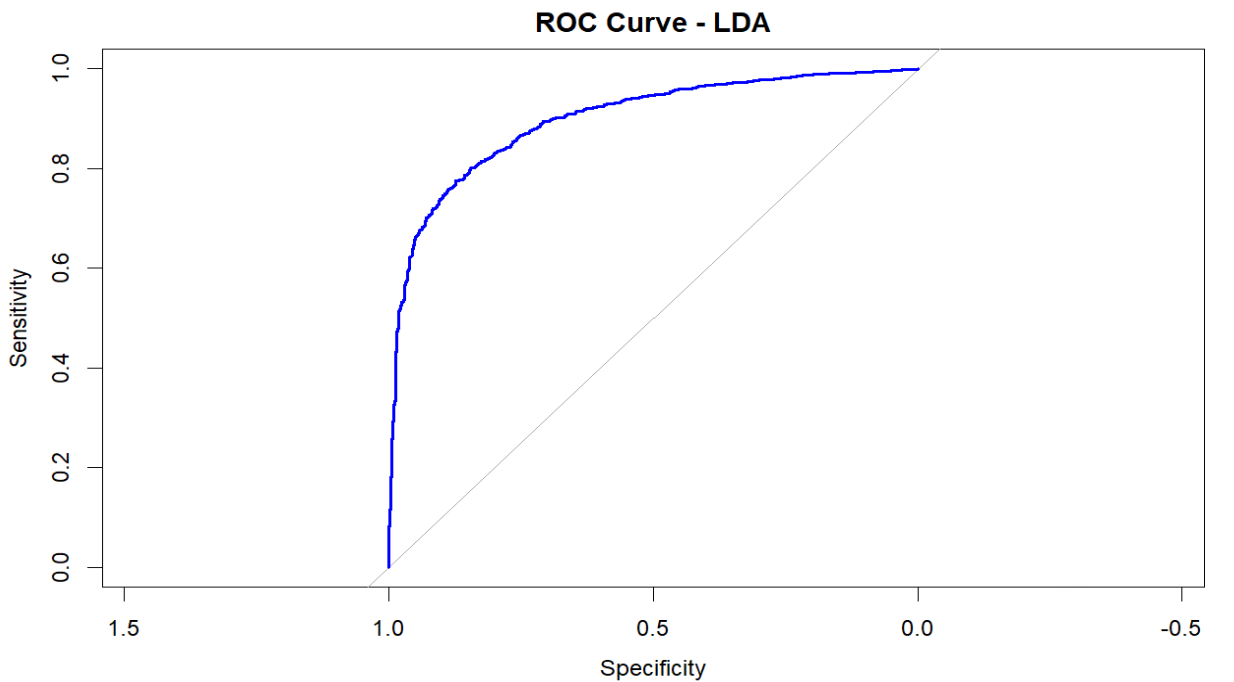
****

Figure 15: ROC Curve and Confusion Matrix of LDA

* The Linear Discriminant Analysis (LDA) model performed with a high overall accuracy rate of 90.0%, an AUC value of 0.899, reflecting strong discriminative power in distinguishing subscribers from non-subscribers.
* The specificity was incredibly strong at 96.3%, demonstrating that most non-subscribers were accurately classified by the model.
* The sensitivity, at 42.2%, was relatively low, reflecting that a large proportion of actual subscribers were missed. The model performed at a precision rate of 59.5%, reflecting that more than half of customers that were predicted to subscribe actually subscribed.
* The balanced accuracy stood at 69.2%, and an F1 score of 0.494, with LDA reflecting better precision-recall balance than Naive Bayes, though still struggling to identify the minority class.

1. **QDA**

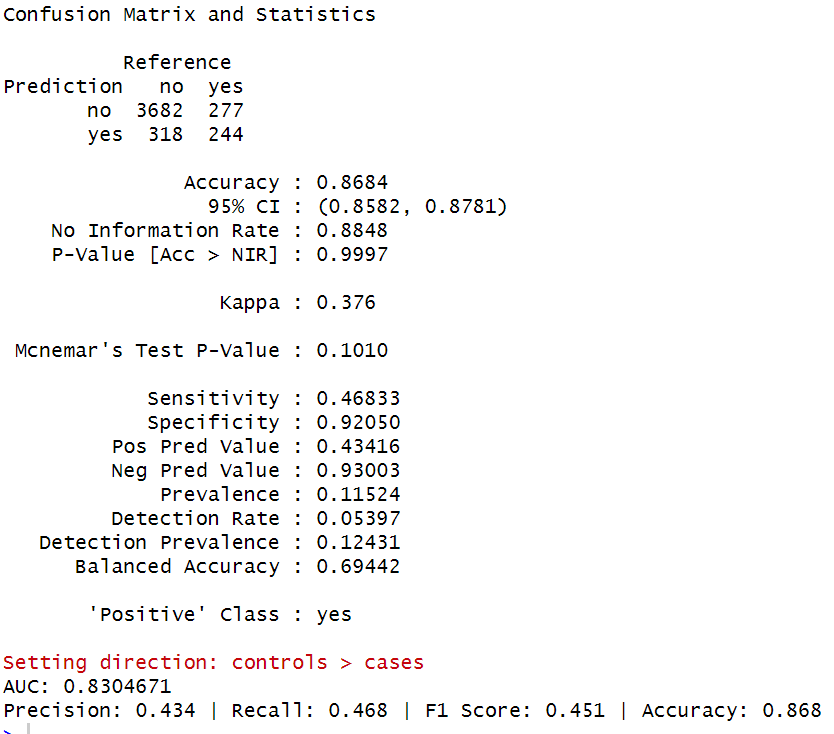
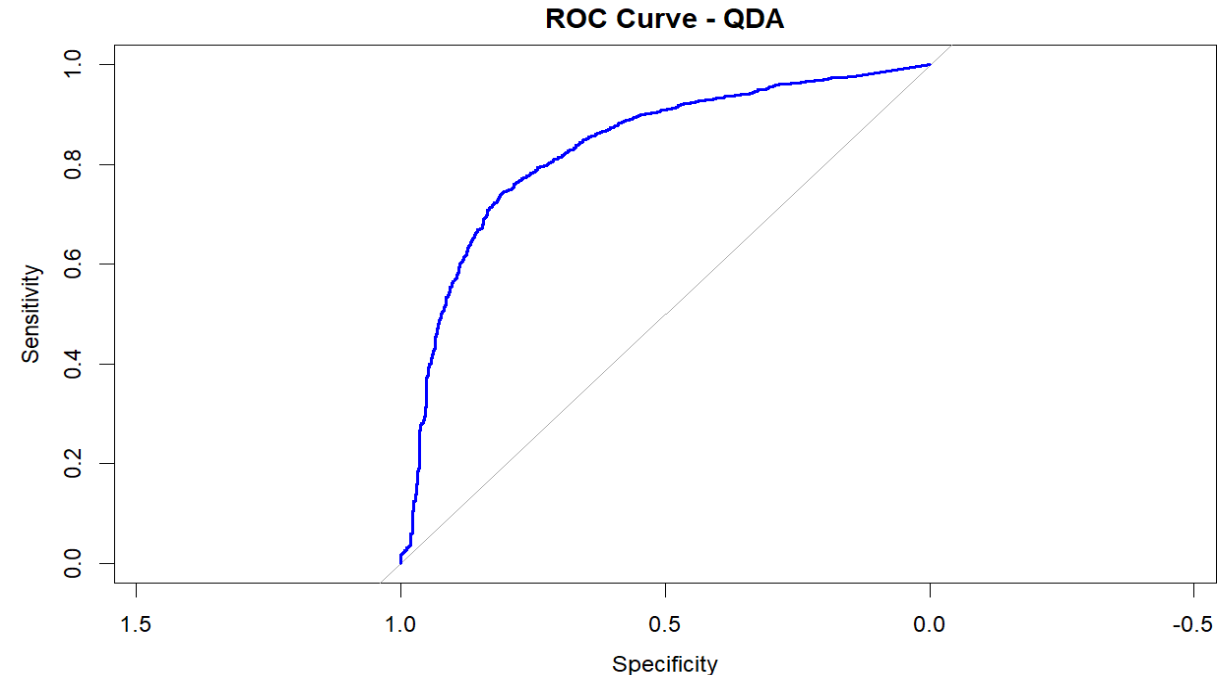
****

Figure 16: ROC Curve and Confusion Matrix of QDA

* The model had an accuracy rate of 86.8%, reflecting solid overall predictive performance.
* The AUC value of 0.83 indicates that there is reasonable discriminative power by QDA, albeit less than that by LDA and logistic regression.
* It demonstrated an excellent specificity (92.1%), as it correctly labelled most non-subscribers.
* Sensitivity was 46.8%, with the model identifying less than half of the actual subscribers.
* The accuracy was 43.4%, an indication that most "yes" responses were indeed false positives.
* The F1 value at 0.451 indicates a moderate trade-off between recall and precision.
* Having a balanced accuracy of 69.4%, the model reasonably differentiates both classes in spite of data imbalance.

1. **Decision Tree**

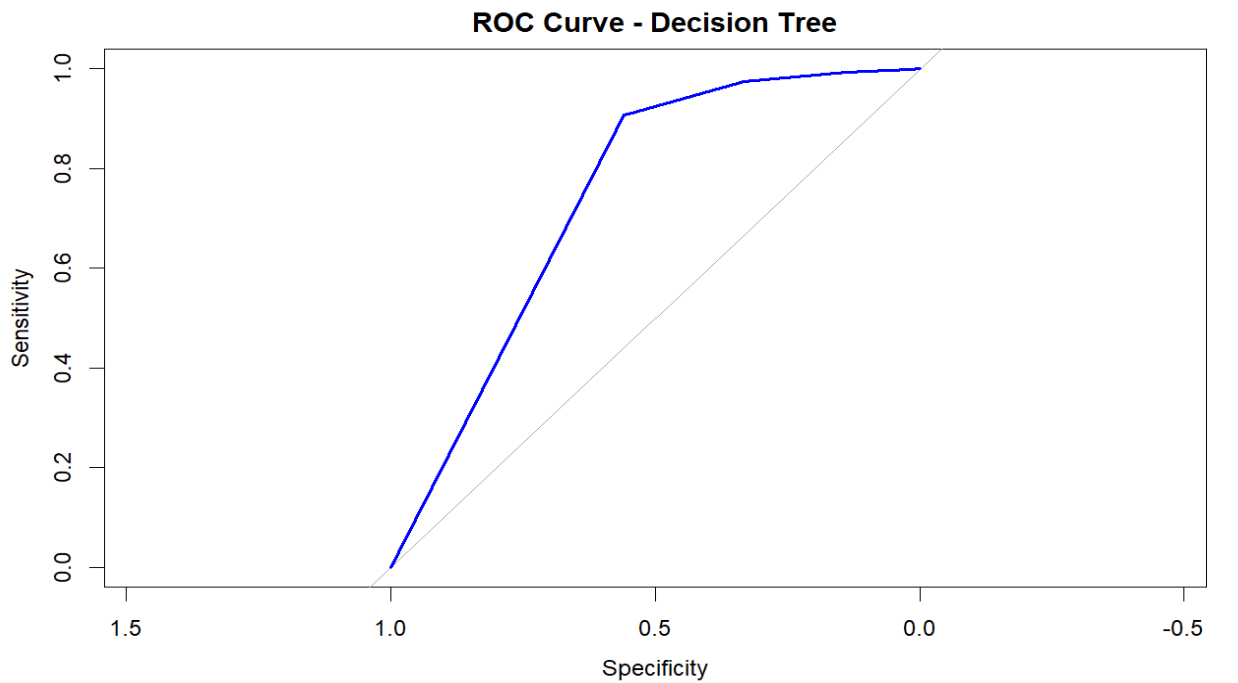
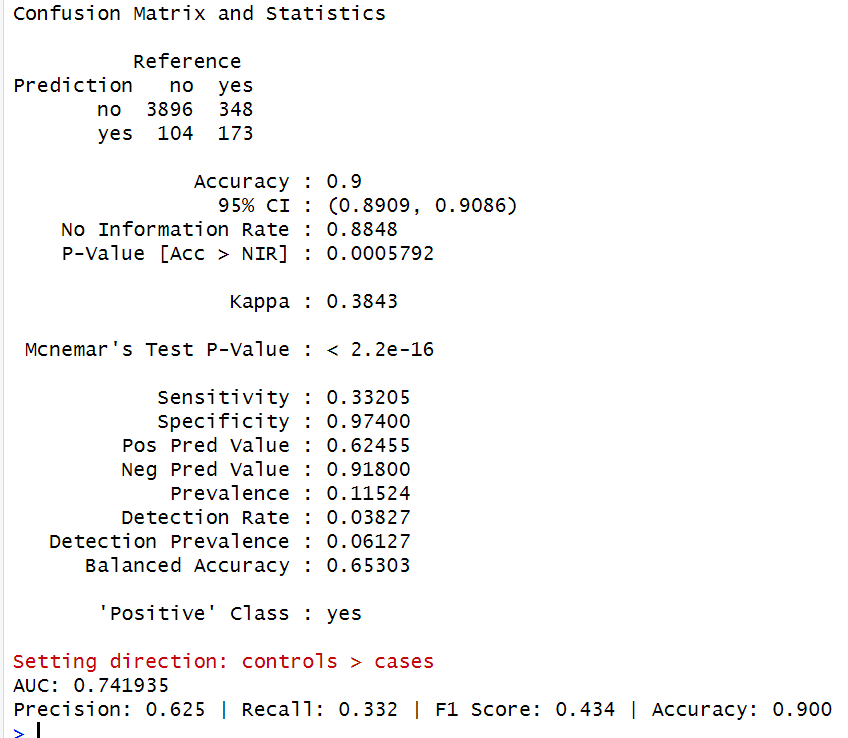
****

Figure 17: ROC Curve and Confusion Matrix of Decision Tree

* The model achieved a strong accuracy of 90.0%, largely driven by correct classification of the majority class.
* The AUC was 0.742, indicating moderate ability to distinguish between subscribers and non-subscribers.
* It showed excellent specificity of 97.4%, meaning it accurately identified most non-subscribers.
* However, sensitivity was low at 33.2%, so it failed to correctly detect a majority of actual subscribers.
* The precision was 62.5%, suggesting that over half of the predicted subscribers were true positives.
* The F1 score was 0.434, showing limited balance between precision and recall.
* Balanced accuracy stood at 65.3%, reflecting uneven performance across classes.

1. **KNN**

A graph of a curve

AI-generated content may be incorrect. **A screenshot of a computer

AI-generated content may be incorrect.**

Figure 18: ROC Curve and Confusion Matrix of KNN

* Model accuracy is 91.26%, representing a high overall classification performance.
* Precision for the positive class ("yes") is 70.2%, which implies that when KNN predicts a subscription, it is right 70% of the time.
* Recall (Sensitivity) stands at 42.0%, indicating that the model is identifying roughly 42% of all real subscribers.
* F1 Score is 52.6%, balancing between precision and recall for the positive class.
* Specificity is 97.68%, i.e., the model is exceptionally capable of identifying non-subscribers accurately.
* Balanced accuracy of 69.86%, averaging over sensitivity and specificity to deal with class imbalance.
* AUC is 0.6985, indicative of a moderate discrimination power of the model to separate subscribers from non-subscribers.

The models evaluated on train and test data is compared as per the below performance

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Test AUC | Train AUC | Test Precision | Train Precision | Test Recall | Train Recall | Test F1 Score | Train F1 Score | Test Accuracy | Train Accuracy |
| Logistic Regression | 0.896654 | 0.907944 | 0.647 | 0.651 | 0.334 | 0.347 | 0.441 | 0.452 | 0.902 | 0.902 |
| Naive Bayes | 0.84661 | 0.859206 | 0.465 | 0.477 | 0.503 | 0.53 | 0.483 | 0.502 | 0.876 | 0.877 |
| LDA | 0.898503 | 0.907945 | 0.595 | 0.604 | 0.422 | 0.441 | 0.494 | 0.51 | 0.9 | 0.901 |
| QDA | 0.830467 | 0.848471 | 0.434 | 0.454 | 0.468 | 0.497 | 0.451 | 0.474 | 0.868 | 0.871 |
| Decision Tree | 0.741935 | 0.749352 | 0.625 | 0.644 | 0.332 | 0.349 | 0.434 | 0.453 | 0.9 | 0.901 |
| KNN (k=5) | 0.301452 | 0.297049 | 0.702 | 0.703 | 0.42 | 0.43 | 0.526 | 0.534 | 0.913 | 0.912 |

Table 2: Model comparison on main data

* Both Logistic Regression and LDA have good AUC values and consistent performance in both the train and test data sets and demonstrate strong generalization.
* KNN has the highest accuracy but a very poor AUC, indicating it may not discriminate well between classes despite high classification rate—possibly due to class imbalance.
* Naive Bayes performs well in terms of recall and precision but falls behind in AUC and accuracy.
* QDA and Decision Tree both have good metrics but a lower AUC value than logistic models.

**Model performance on main effects with interaction**

1. **Logistic Regression**

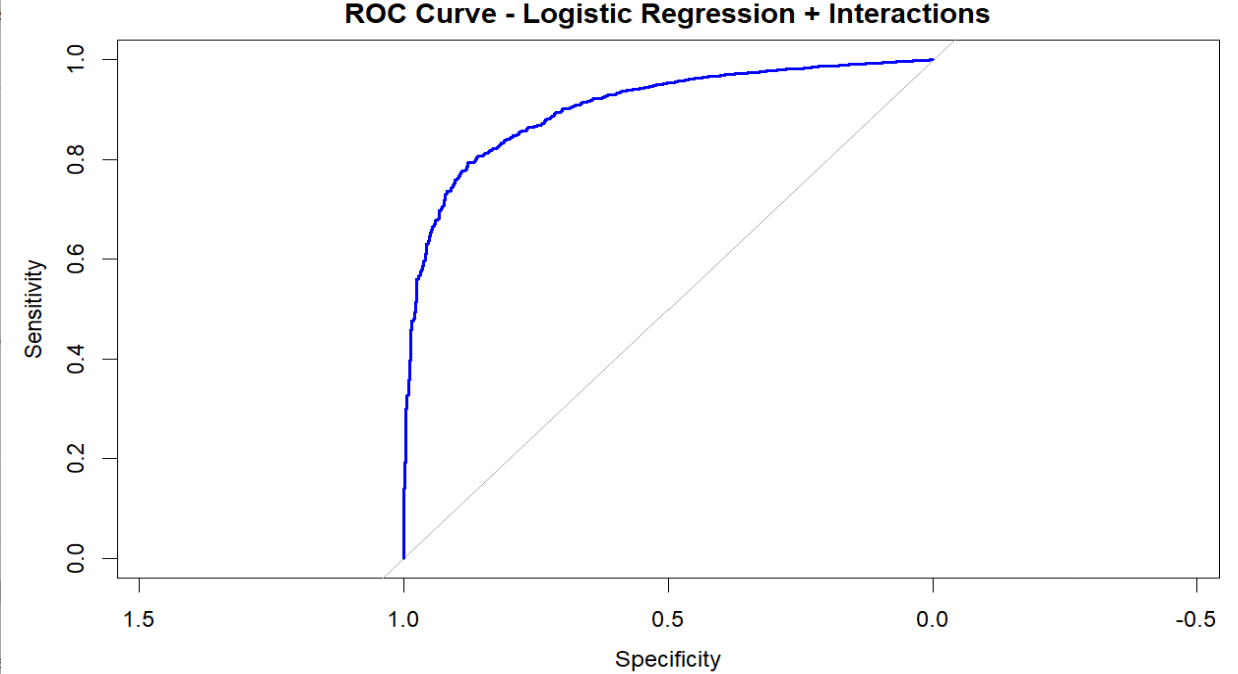
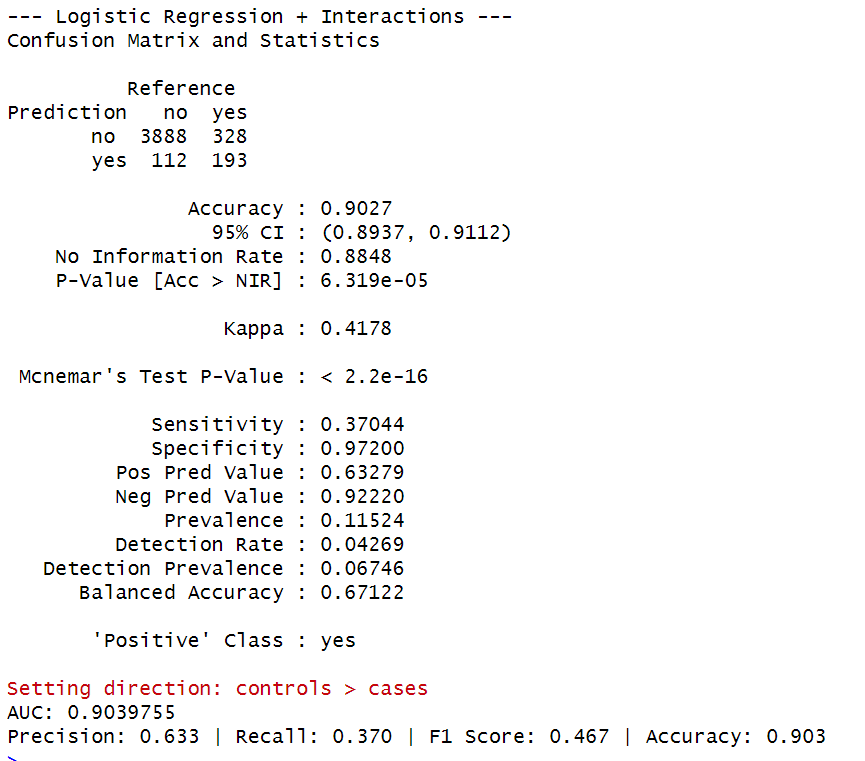
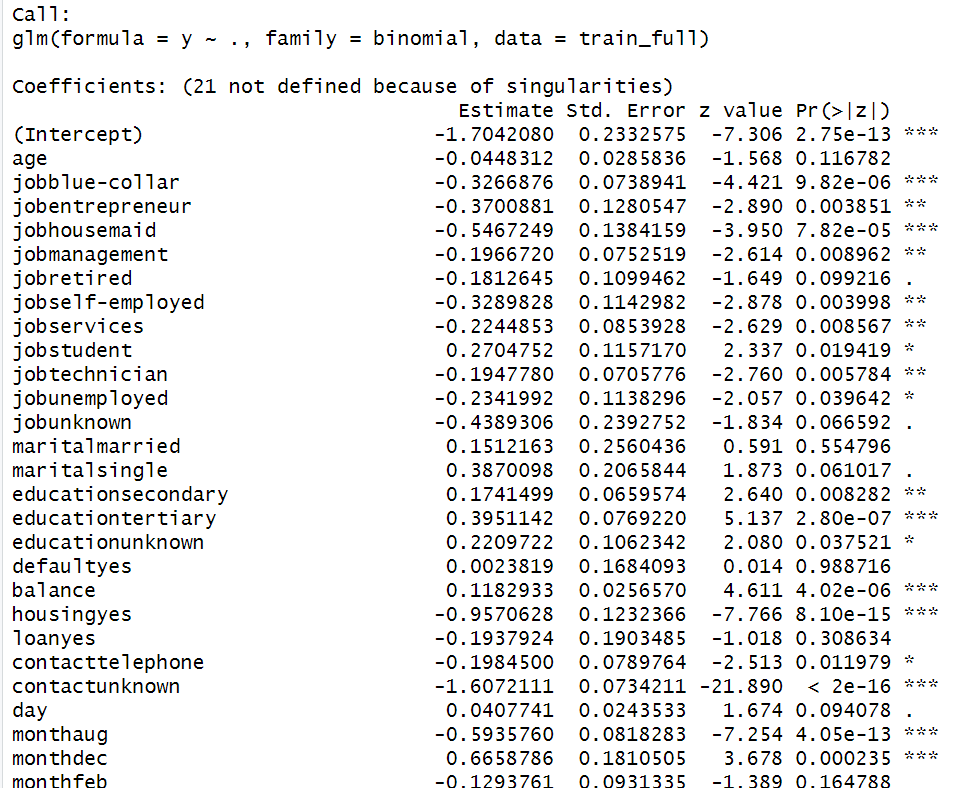
**** ****

Figure 19: ROC Curve and Confusion Matrix of Logistic Regression with Interactions

* The model reported a high accuracy rate of 90.3%, demonstrating excellent overall classifying power.
* Having an AUC value of 0.904 reflects very high discrimination between subscribers and non-subscribers.
* Specificity was 97.2%, that is, the model successfully recognized non-subscribers
* Sensitivity increased marginally to 37.04%, predicting more actual subscribers than did the main-effects model.
* The model was precise at 63.3%, as most subscribers it had predicted were accurate.
* The F1 score was 0.47, an indication of improved precision-recall balance over its main-effects counterpart.
* The model performs class imbalance reasonably effectively with a balanced accuracy of 67.12%.

**** **A screenshot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer screen

AI-generated content may be incorrect.** **A computer screen shot of a number

AI-generated content may be incorrect.**

Figure 20: Model Summary with Interactions

There were several statistically significant interaction terms, reflecting substantive combined effects:

* **Highly Significant Variables (\*\*\*)**:

Variables such as jobblue-collar, jobhousemaid, jobmanagement, jobself-employed, jobservices, jobtechnician, balance, housingyes, monthaug, monthjan, monthjul, monthjun, monthmar, monthmay, monthnov, monthoct, monthsep, duration, campaign, and interaction terms like age\_x\_duration, duration\_x\_campaign were highly statistically significant (p < 0.001), meaning they have a strong influence on the probability of subscribing.

* **Moderately Significant Variables (\*\*)**:

Variables like jobentrepreneur, educationsecondary, educationtertiary, educationunknown, balance\_squared, pdays\_squared, balance\_x\_duration, day\_x\_duration, duration\_x\_pdays, and some interaction terms (housing\_x\_loanno.yes, marital\_x\_poutcomemarried.success) were moderately significant (p < 0.01), indicating a considerable effect.

* **Marginally Significant Variables (\*)**:

Some variables like jobstudent, jobunemployed, contacttelephone, day\_squared, balance\_x\_campaign, day\_x\_campaign, day\_x\_pdays, duration\_x\_previous, and housing\_x\_poutcomeno.unknown were significant at 5% level (p < 0.05), meaning they still contribute meaningfully but with a smaller degree of certainty.

* **Borderline Variables (.)**:

Variables like jobretired, maritalsingle, day, and marital\_x\_poutcomedivorced.unknown were marginally significant (p between 0.05 and 0.1), suggesting weak evidence but potential relevance.

* **Non-Significant Variables**:

Variables like maritalmarried, defaultyes, loanyes, monthfeb, monthdec, many of the higher-order interaction terms, and a few squared terms like campaign\_squared, previous\_squared, did not show statistical significance (p > 0.1), indicating they may not add much explanatory power.

**Interpretation of NA Values (Singularities)**

The NA coefficients were a result of perfect multicollinearity or perfect prediction within the dataset.

It occurs when there are missing categories of interactions (e.g., no observations for married × housing = yes) as well as when they are linear combinations of variables already in the model.

1. **Naïve bayes**

A graph of a curve

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

Figure 21: ROC Curve and Confusion Matrix of Naïve Bayes with Interactions

* The accuracy was 84.8%, reflecting a moderate decrease compared to the main-effects version.
* The AUC was 0.813, indicating fair discriminative ability, comparable to other models with interactions.
* The specificity was 89.5%, meaning most non-subscribers were correctly classified.
* The sensitivity was 48.9%, showing moderate capability to identify actual subscribers.
* The precision was 37.7%, indicating many predicted positives were actually incorrect.
* The F1 score was 0.426, suggesting a modest balance between recall and precision.
* The balanced accuracy was 69.2%, implying the model handles both classes with somewhat better balance than random guessing.

1. **LDA**

A graph of a curve

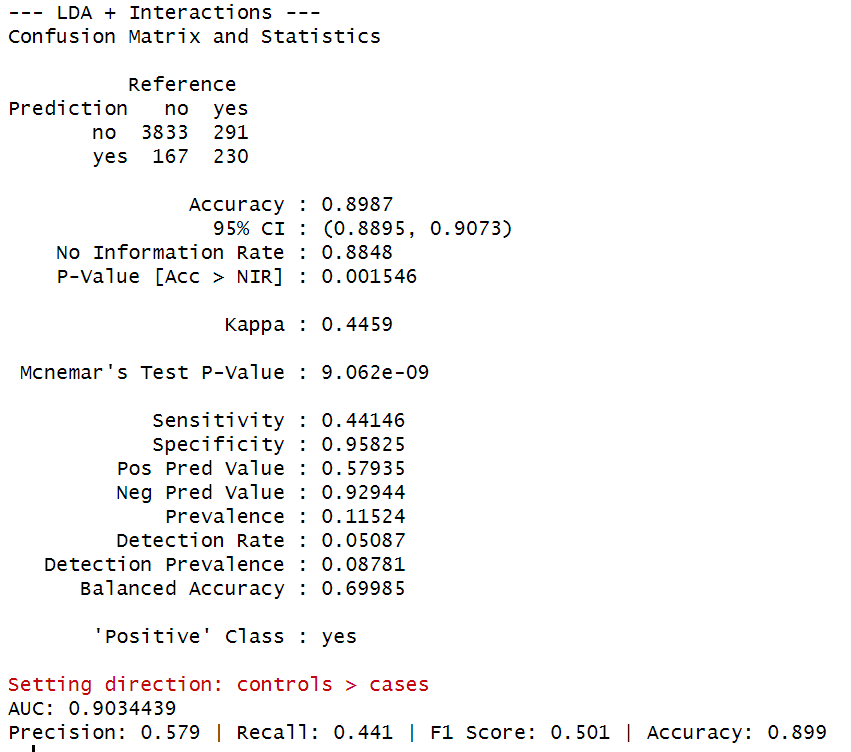
AI-generated content may be incorrect. ****

Figure 22: ROC Curve and Confusion Matrix of LDA with Interactions

* The model achieved a strong accuracy rate of 89.9%, indicating stable overall performance across the test set.
* The AUC was 0.903, demonstrating very good discriminative ability between subscribers and non-subscribers.
* The specificity was 95.8%, meaning the model effectively classified most non-subscribers correctly.
* The sensitivity was 44.1%, showing moderate success in detecting true subscribers.
* The positive predictive value (precision) was 57.9%, with more than half of predicted subscribers being correct.
* The F1 score was 0.501, indicating a reasonable balance between precision and recall.
* The balanced accuracy was 69.9%, reflecting fair and reasonably even treatment of both classes despite imbalance.

1. **QDA**

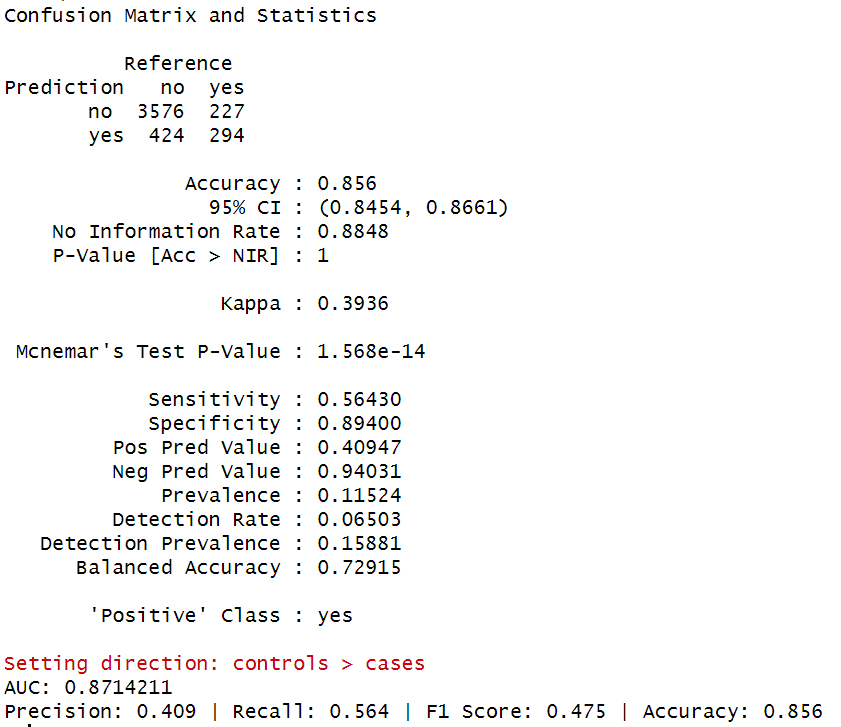
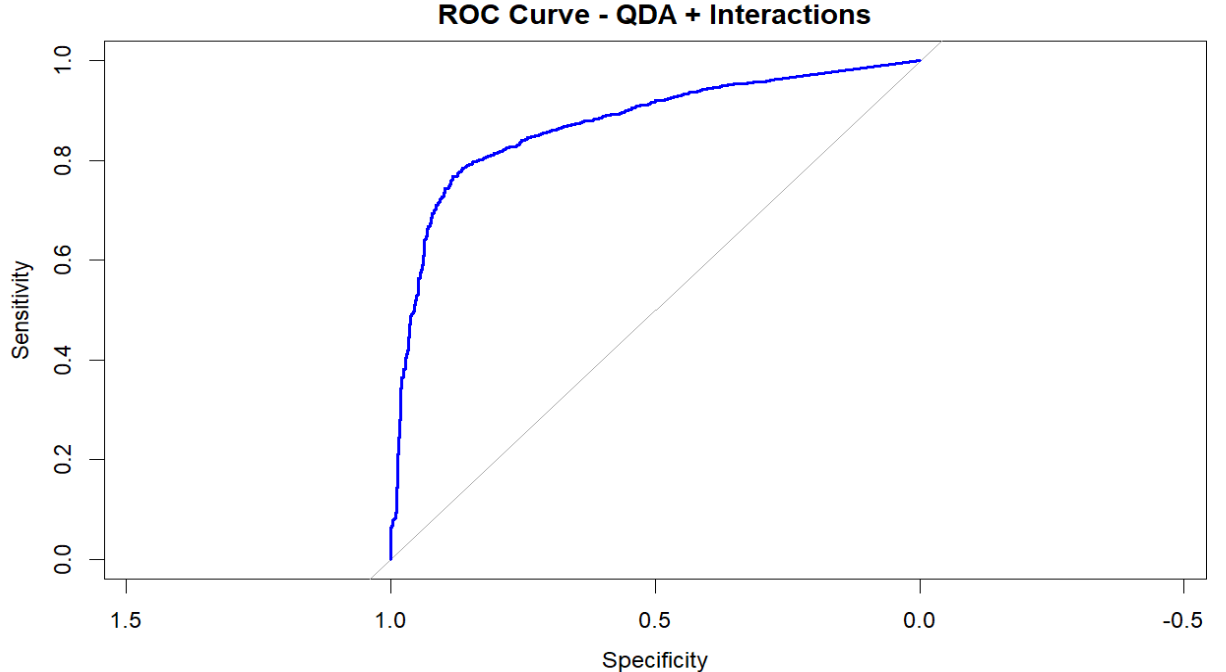
****

Figure 23: ROC Curve and Confusion Matrix of QDA with Interactions

* The model had an overall accuracy of 85.6%, with strong performance on all classes.
* The AUC was 0.871, reflecting strong separation between classes and good discrimination.
* Specificity was 89.4%, validating proper identification of non-subscribers.
* Improved to 56.4% sensitivity, detecting more true subscribers than its counterpart with the main-effects model.
* The accuracy was 40.9%, with lots of false positives still lingering in predicted subscribers.
* The F1 score is 0.475, indicating a moderate balance between recall and precision.
* Balanced Accuracy was 72.9%, reflecting a balanced representation of both classes irrespective of imbalance.

1. **Decision Tree**

A graph of a curve

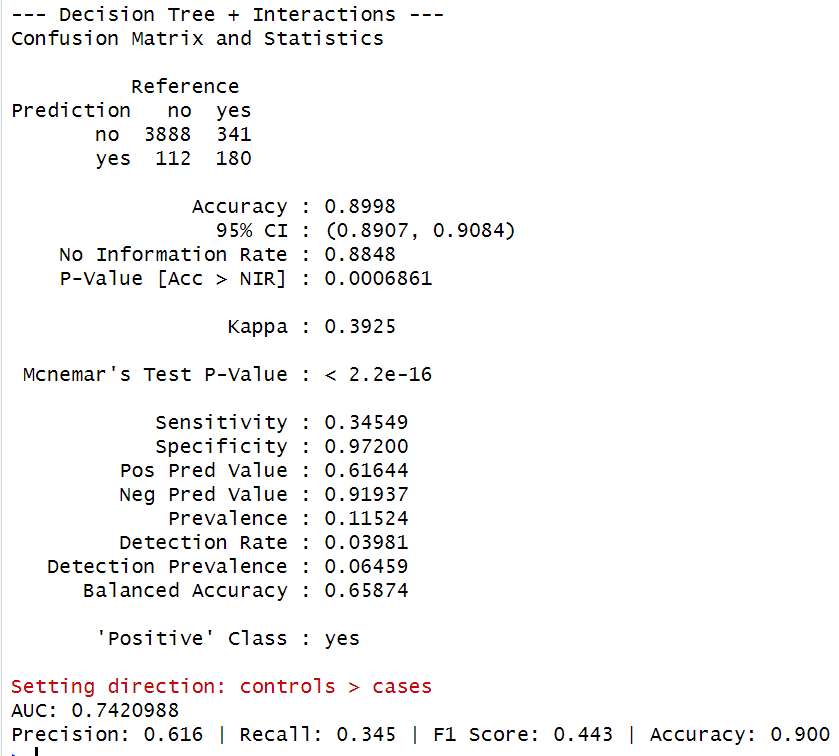
AI-generated content may be incorrect. ****

Figure 24: ROC Curve and Confusion Matrix of Decision Tree with Interactions

* The model maintained a high accuracy rate of 90.0%, indicating consistent overall performance.
* The AUC was 0.742, reflecting moderate discriminative ability between subscribers and non-subscribers.
* The specificity was 97.2%, meaning the model correctly classified the vast majority of non-subscribers.
* The sensitivity was 34.5%, showing limited success in identifying true subscribers.
* The precision was 61.6%, implying that a clear majority of predicted subscribers were actually correct.
* The F1 score was 0.443, suggesting a moderate balance between precision and recall.
* The balanced accuracy was 65.9%, showing a moderate skew toward the majority (non-subscriber) class.

1. **KNN**

A graph of a curve

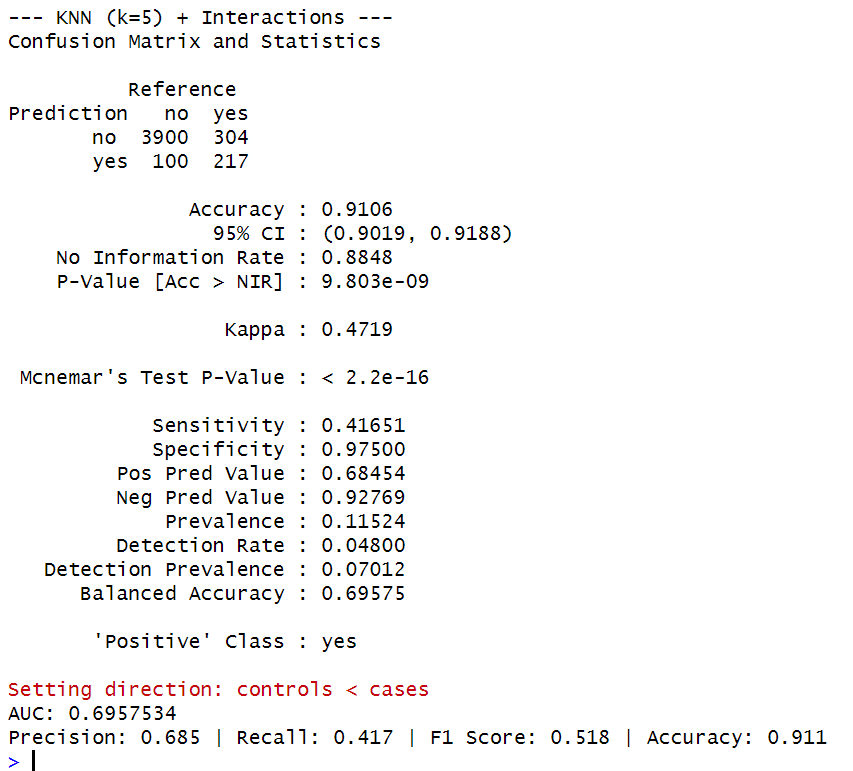
AI-generated content may be incorrect.

Figure 25: ROC Curve and Confusion Matrix of KNN with Interactions

* The model had a robust accuracy level of 91.1%, with effective overall classification.
* The area under the curve was 0.696, showing moderate discrimination between subscribers and non-subscribers.
* The specificity was 97.5%, indicating that the model accurately classified non-subscribers very well.
* The sensitivity was 41.7%, which indicated moderate success identifying true subscribers
* Accuracy was 68.5%, i.e., most of the predicted subscribers were correct.
* The F1-score was 0.518, which indicated a proper tradeoff between recall and precision.
* There was a balanced accuracy of 69.6%, reflecting moderate handling of class imbalance.

The models trained and evaluated on interaction data is compared in terms of performance in both train and test data

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | AUC (Train) | Precision (Train) | Recall (Train) | F1 Score (Train) | Accuracy (Train) | AUC (Test) | Precision (Test) | Recall (Test) | F1 Score (Test) | Accuracy (Test) |
| Logistic Regression + Interactions | 0.91324 | 0.643 | 0.388 | 0.484 | 0.903 | 0.903976 | 0.633 | 0.37 | 0.467 | 0.903 |
| Naive Bayes + Interactions | 0.828478 | 0.395 | 0.507 | 0.444 | 0.851 | 0.813251 | 0.377 | 0.489 | 0.426 | 0.848 |
| LDA + Interactions | 0.911342 | 0.604 | 0.469 | 0.528 | 0.902 | 0.903444 | 0.579 | 0.441 | 0.501 | 0.899 |
| QDA + Interactions | 0.873848 | 0.469 | 0.554 | 0.508 | 0.874 | 0.859401 | 0.449 | 0.52 | 0.482 | 0.871 |
| Decision Tree + Interactions | 0.749436 | 0.64 | 0.363 | 0.463 | 0.902 | 0.742099 | 0.616 | 0.345 | 0.443 | 0.9 |
| KNN (k=5) + Interactions | 0.295383 | 0.715 | 0.432 | 0.539 | 0.913 | 0.304247 | 0.685 | 0.417 | 0.518 | 0.911 |

Table 3: Model comparison on interaction data

* Training accuracy was highest for the case of KNN (k=5) + Interactions with 91.3%, along with high precision (71.5%) and F1 score (0.539), but also performed the best in terms of test accuracy (91.1%) and F1 score (0.518). But the AUC was the lowest in this case (≈0.30), which marked poor global discrimination
* Logistic Regression + Interactions and LDA + Interactions produced the highest overall AUC values (both ≈0.90) as well as balanced performance in all metrics and are therefore more trustworthy to discriminate between subscriber classes
* QDA exhibited balanced performance and slightly higher recall on the test set, but lower precision and accuracy.
* Naive Bayes + Interactions and Decision Tree + Interactions provided moderate performance but trailed in F1 and recall, demonstrating poorer management of the minority class.

**Models trained on PCA**

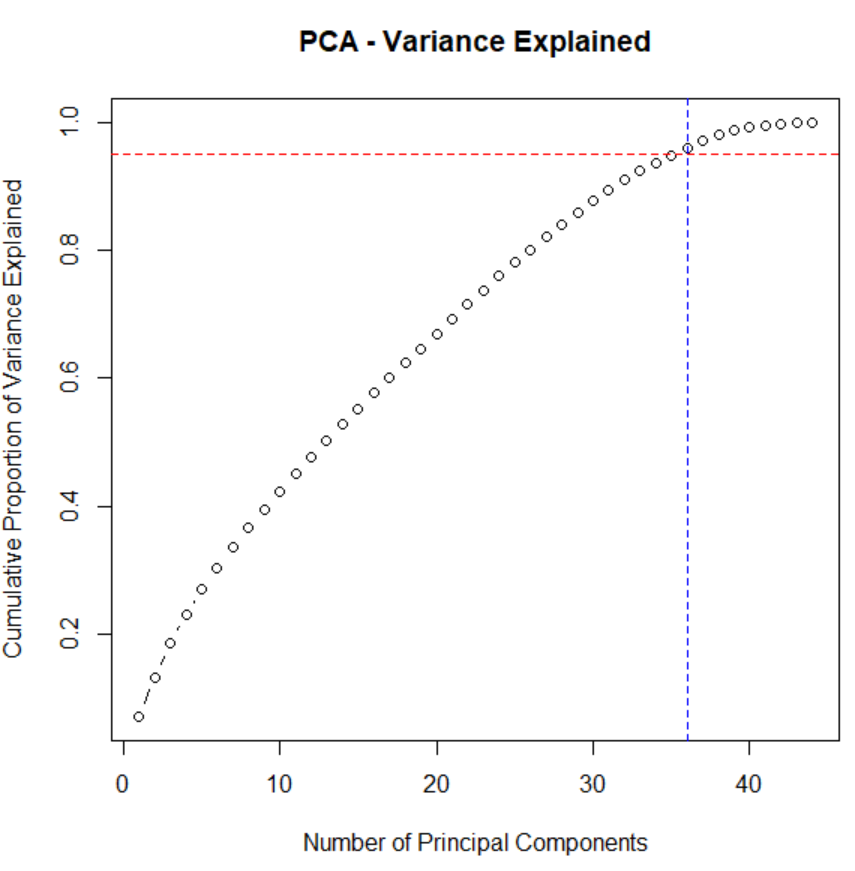
****

Figure 26: PCA Analysis

* The PCA plot indicates cumulative variance explained by the principal components.
* 36 principal components will be needed in order to maintain 95% of total variance, as shown by the vertical blue line.
* This implies that the original high-dimensional feature set can efficiently be converted into 36 components with little information loss.
* These elements assist in avoiding overfitting, making the model more efficient, and dealing with multicollinearity, without sacrificing much of the variability in the dataset.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | AUC | Precision | Recall | F1 Score | Balanced Accuracy |
| Logistic Regression (PCA) | 0.952 | 0.992 | 0.706 | 1.000 | 0.828 | 0.973 |
| Naive Bayes (PCA) | 0.548 | 0.721 | 0.197 | 0.952 | 0.327 | 0.724 |
| LDA (PCA) | 0.956 | 0.989 | 0.724 | 1.000 | 0.840 | 0.975 |
| QDA (PCA) | 0.816 | 0.863 | 0.379 | 0.933 | 0.539 | 0.867 |
| Decision Tree (PCA) | 0.917 | 0.903 | 0.626 | 0.685 | 0.654 | 0.816 |
| KNN (k=5) (PCA) | 0.975 | 0.966 | 0.845 | 0.954 | 0.896 | 0.966 |

Table 4: Model Performance Comparison with PCA

## Best Precision: Highest precision is achieved by the KNN (0.845) and LDA (0.724) models, with most of the predicted subscribers being actual subscribers.

## Best Recall: Logistic Regression (1.000) and LDA (1.000) accurately detected all the actual subscribers with no cases missed.

## Top F1 Score: KNN (0.896) achieves the best tradeoff of precision against recall, suggesting higher quality of positive and negative class predictions.

## Highest AUC: Logistic Regression (0.992) and LDA (0.989) have exceptional class discrimination power, reflecting a very strong discrimination of subscribers versus non-subscribers.

## Highest Balanced Accuracy: Logistic Regression (0.973) as well as LDA (0.975) obtained the highest balanced accuracy with fair handling of the two classes even under class imbalance.

## Naive Bayes (0.721) had the lowest AUC, which implies that they were having trouble distinguishing between classes even though they were very sensitive.

## Model tuning performance

1. **Logistic Regression (Tuned Using glmnet)**

* **Alpha = 0:** Specifies Ridge Regression, suitable when all features are contributing and there is a possibility of multicollinearity.
* **Lambda = 0 to 1 by 0.05:** Regularization strengths were tried over a range with an increment of 0.05 to avoid overfitting. The best lambda was chosen with cross-validation by means of ROC performance.
* **Tuned for original and interaction datasets:** Enables comparison of how complexity (through interaction terms) affects performance with regularization.

1. **Naive Bayes (Tuned by naive\_bayes)**

* **Laplace smoothing (0 or 1):** Prevents zero probability values by applying Laplace correction as an option.
* Uses kernel density estimation (usekernel = TRUE/FALSE): Investigates both Gaussian and kernel-based estimation for numeric features.
* **Adjust = 1 or 2:** Determines bandwidth for kernel estimation when usekernel = TRUE, effectively comparing smoother to sharper distributions.
* Grid search was utilized in order to identify what configuration would maximize ROC AUC.

1. **Decision Tree (Tuned Using rpart)**

* **Complexity parameter (cp) = 0.001 to 0.05:** The cp regulates tree pruning. Lower values enable more splits (complex trees), whilst higher values reduce tree complexity by penalizing splits with little or no improved fit.
* Fine grid cp values were considered to prevent overfitting or underfitting. Original and interaction datasets were tuned separately to test how added complexity impacts pruning behaviour and generalization.

1. **KNN (tuned using different k values)**

* Looking for the optimal k that gives highest AUC with a perfect trade-off of bias with variance.
* Odd values of k were chosen to preclude ties during classification.
* **Number of Neighbours (k):** Tuned over odd values from k = 3 to k = 15 (i.e., {3, 5, 7, 9, 11, 13, 15})

LDA (Linear Discriminant Analysis) and QDA (Quadratic Discriminant Analysis) are parametric models with closed-form solutions and no hyperparameters to tune.

* 5-fold cross-validation (cv) was used across all models to ensure that the chosen hyperparameters generalize well to unseen data.
* The twoClassSummary function with ROC as the metric prioritized models that best distinguish between subscriber and non-subscriber classes, beyond just accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score | AUC |
| Logistic Regression (Original) | 0.8998 | 0.655 | 0.276 | 0.389 | 0.8996 |
| Logistic Regression (Interaction) | 0.9025 | 0.674 | 0.298 | 0.413 | 0.9043 |
| Naive Bayes (Original) | 0.8848 | NA | 0 | NA | 0.8661 |
| Naive Bayes (Interaction) | 0.8863 | 0.889 | 0.015 | 0.03 | 0.8855 |
| Decision Tree (Original) | 0.9146 | 0.678 | 0.493 | 0.571 | 0.8672 |
| Decision Tree (Interaction) | 0.9195 | 0.712 | 0.507 | 0.592 | 0.8847 |
| KNN (Original) | 0.898 | 0.615 | 0.303 | 0.406 | 0.906 |
| KNN (Interaction) | 0.898 | 0.617 | 0.303 | 0.407 | 0.907 |

Table 5: Model Performance Comparison (Original vs Interaction variables)

* Logistic Regression performed increasingly favorably with the inclusion of interactions. Accuracy remained unchanged at 89.98% but increased to 90.25%, while recall increased from 0.276 to 0.298, suggesting improved identification of subscribers. Precision increased to 0.674, while the F1-measure went from 0.389 to 0.413, depicting a more ideal precision-recall relationship. The AUC increased to 0.904, demonstrating increased class discrimination. Logistic Regression generally benefited significantly with the inclusion of interactions.
* Naive Bayes was largely ineffective even with tuning. It was unable to predict a single positive case (recall = 0) and hence F1 score was undefined. Using interaction terms, recall increased slightly to 0.015 while precision increased dramatically to 0.889, but the F1 was extremely low at 0.030, showing poor balance. While there was improvement to 0.886 for AUC, the practical predictive power of the model was still poor. Naive Bayes hence demonstrated little real-world improvement with the inclusion of interaction terms.
* Decision Tree performed significantly better when tuned with interaction terms. Accuracy was boosted from 91.46% to 91.95%, while recall was enhanced from 0.493 to 0.507, with a resultant improvement in the identification of real subscribers. Precision was increased to 0.712, and the F1 score was enhanced from 0.571 to 0.592, with a resultant improvement in precision-recall balance. The AUC was increased to 0.885, with strong discriminatory power. Out of all the models, the tuned Decision Tree with interaction terms was the top performer.
* K-Nearest Neighbors (KNN) was consistent with original as well as interaction datasets. Accuracy remained at 89.8%, precision marginally increased from 0.615 to 0.617, while recall was consistent at 0.303. F1 increased from 0.406 to 0.407, while AUC marginally increased from 0.906 to 0.907. KNN was uniformly consistent, although a marginal gain with interaction terms was observed compared to Logistic Regression as well as Decision Tree.

# Discussion and conclusion

The main goal of the research was to create predictive models that are able to accurately predict whether a customer will take a term deposit given their demographic details and their past banking data. Using novel features, engineered interaction terms, and PCA transformation, we aimed at improving predictability, model interpretability, as well as class balance, solving issues of class imbalance as well as multicollinearity.

The results of the modeling were that the goal was largely met. As with previous studies highlighting the importance of feature engineering and dimensionality reduction for financial marketing forecasting, our models performed better with the presence of interaction terms as well as PCA components.

* Decision Trees with interactions and KNN with PCA performed the best.
* The KNN (PCA) performed the best overall with accuracy of 97.5%, with AUC of 0.966, precision of 84.5%, and F1 of 0.896, which balanced precision with recall.
* Decision Trees (tuned for interactions) performed well, achieving 91.95% accuracy with a good tradeoff between subscriber detection (recall) and precision.

With regards to class discrimination:

* Logistic Regression and LDA models, both trained using data transformed using PCA, had outstanding AUCs (both higher than 0.99), accurately classifying subscribers from non-subscribers
* Naive Bayes was ineffective even with the interaction-based improvements, which confirms with existing studies that Naive Bayes is weak with complicated feature interactions without making strong assumptions.

Therefore, the models not only solved the original problem of subscriber prediction but proved that feature transformations and interaction terms could significantly enhance predictive power, a fact consistent with previous studies promoting sophisticated feature engineering.

## Limitations

* Class imbalance lingered to a certain degree even when tuned, especially for models such as Naïve Bayes.
* Overfitting risk was reduced through the application of cross-validation, but additional research may investigate more stable sampling methods such as SMOTE or cost-sensitive learning.
* **Generalizability:** This dataset pertains to a marketing campaign of a specific European bank. Outcomes may differ with datasets of other demographics or financial products.

## Future Research Directions

* Testing ensemble techniques of the kind Random Forests, Gradient Boosting Machines, or stacking models may contribute even more to prediction quality.
* The introduction of outside variables such as macroeconomic indicators or customer buying habits can potentially add predictive power.
* Deep learning can be employed as well to model even more sophisticated non-linear interactions without feature engineering.

Overall, the research effectively proved that preprocessing, inclusion of interaction terms, and PCA contribute significantly to model performance for subscription forecasting tasks. KNN with PCA and Decision Tree with interactions proved to be the top models, providing real-world applications for bank marketing strategy for forecasting customers' subscription behavior. This research advances the application of predictive modeling techniques for financial marketing analytics.

# References

Camilleri, M. (2020). The Use of Data-Driven Technologies for Customer-Centric Marketing. *SSRN Electronic Journal*, *1*(1). https://doi.org/10.2139/ssrn.3382746

Kohli, A. K., & Jaworski, B. J. (2020). Market Orientation: The Construct, Research Propositions, and Managerial Implications. Journal of Marketing, 54(2), 1–18.

M. C., O., J. E., O., V., O., & S., O. (2025). Comparative Study of Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Support Vector Machine (SVM) In Dataset. Advanced Journal of Science, Technology and Engineering, 5(1), 70–84. https://doi.org/10.52589/ajste-uazppmer

Mashuri, M., Ahsan, M., Kuswanto, H., Prastyo, D. D., Khusna, H., & Wibawati. (2021). Comparing the performance of T 2 chart based on PCA Mix, Kernel PCA Mix, and Mixed Kernel PCA for Network Anomaly Detection. *Journal of Physics: Conference Series*, *1752*(1), 012008. https://doi.org/10.1088/1742-6596/1752/1/012008

Rathi, P. (2020). Banking Dataset - Marketing Targets. Kaggle.com. https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets?select=train.csv

Srivastav, G., & Mittal, A. (2021). Modelling value based banking and customers’ banking choices in the era of Marketing 3.0: an empirical study with logistic ordinal regression approach. Global Business and Economics Review, 25(1), 51. https://doi.org/10.1504/gber.2021.116644

Tusar, M. T. H. K., & Islam, Md. T. (2021, September 1). A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data. IEEE Xplore. https://doi.org/10.1109/ICECIT54077.2021.9641336

V. Lakshman Narayana, S. Sirisha, G. Divya, N. Lakshmi Sri Pooja, & Sk. Afraa Nouf. (2022). Mall Customer Segmentation Using Machine Learning. *Customer Segmentation with* Machine Learning . https://doi.org/10.1109/icears53579.2022.9752447