**Bank Marketing Analysis**

**1. Exploratory Data Analysis (EDA)**

I used the Bank Marketing dataset from the UCI repository, which includes client demographic, financial, and campaign-related attributes collected during a direct marketing campaign by a Portuguese bank. The primary goal of this project is to predict whether a client will subscribe to a term deposit, represented by the binary target variable deposit.

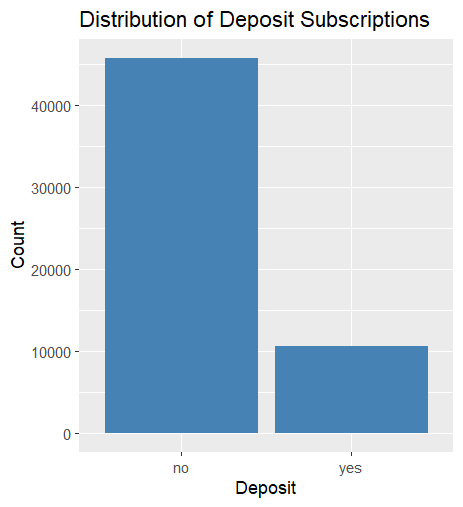
**Data Cleaning and Preparation**

The dataset contained several categorical variables with the value "unknown." These were converted to NA for easier handling. I then dropped two variables that were missing many values in that specific column: contact (34% missing) and poutcome (>99% missing), as their lack of information risked introducing noise without significant modeling value. The duration variable, though highly predictive, was excluded to avoid data leakage because it is only known after the outcome (subscription) occurs.

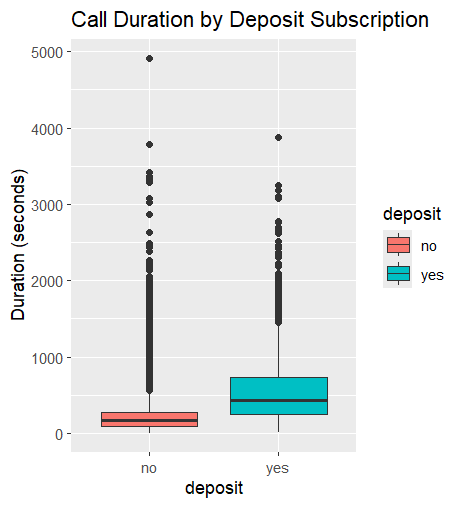
I converted categorical variables to factors and numeric variables were left as is for further processing. A basic check of missing values confirmed the remaining dataset was clean and ready for analysis.

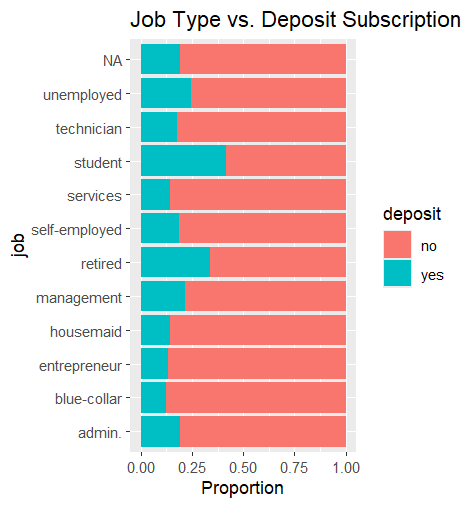
**Distribution Analysis**

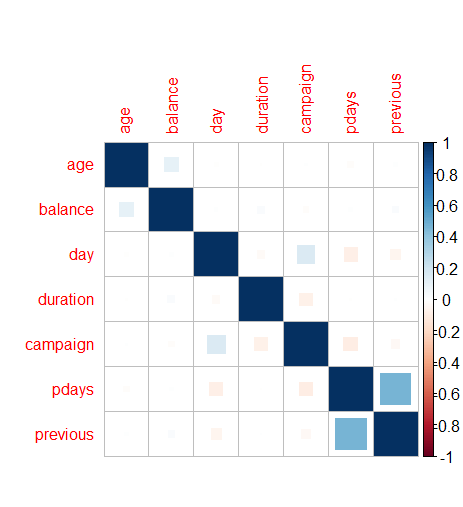
The target variable deposit was found to be imbalanced: approximately 81% of clients did not subscribe. This imbalance led me to focus on evaluation metrics beyond accuracy during the modeling phase.



Key numeric variables were explored visually and statistically. For example, clients who subscribed to a deposit generally had longer call durations and fewer previous contacts. Although duration was excluded from the modeling phase, it helped reveal behavioral patterns during the exploratory phase.



Job roles such as "student," "retired," and "management" were associated with higher deposit subscription rates. Summary statistics and correlation matrices revealed relatively weak inter-variable correlations, suggesting limited multicollinearity and supporting the use of models that can independently assess variable influence.



**EDA Conclusions**

The dataset is well-structured with both categorical and numeric features, making it suitable for classification modeling. I decided to explore logistic regression for its interpretability, random forest for its flexibility and predictive power, and k-means clustering for customer segmentation. Class imbalance, while not extreme, was taken into consideration during model evaluation.

**2. Model Development, Validation, and Optimization**

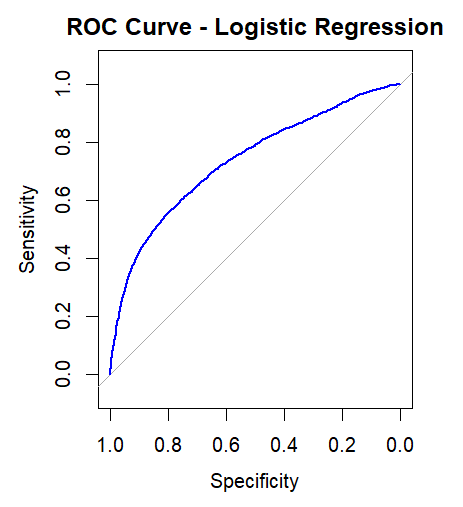
**Model 1: Logistic Regression**

I began with logistic regression to establish a simple, interpretable baseline. Using all features except duration, contact, and poutcome, I fit a logistic model to the training set (70/30 split).

**Performance**:

* Accuracy: 82.5%
* AUC: 0.7341
* Sensitivity: 97.7%
* Specificity: 16.4%

Despite high accuracy and sensitivity, the model performed poorly in detecting actual subscribers (low specificity and moderate AUC). This was expected due to class imbalance and the linear nature of the model.

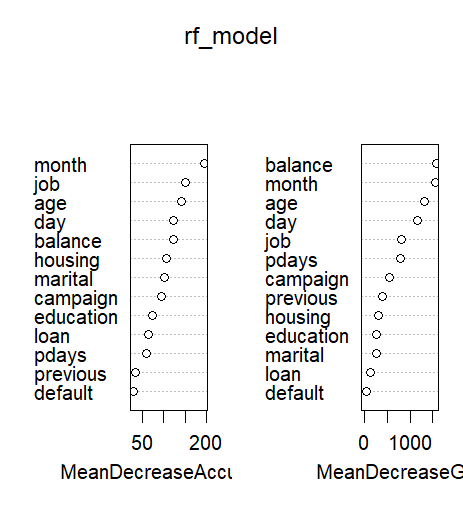
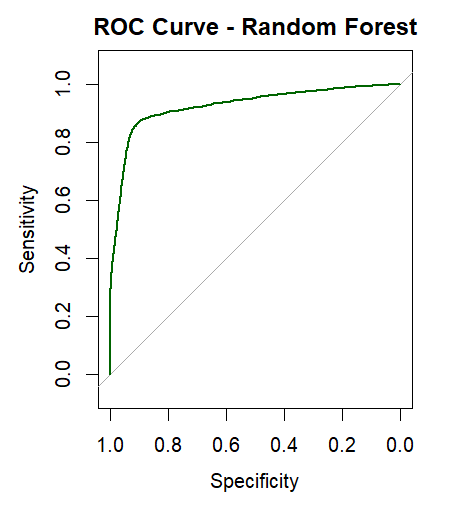


**Model 2: Random Forest**

To address limitations of logistic regression, I trained a Random Forest classifier using the same predictors. The model was evaluated on the same test set.

**Performance**:

* Accuracy: 89.3%
* AUC: 0.9238
* Sensitivity: 96.9%
* Specificity: 55.8%
* Balanced Accuracy: 76.4%

Random Forest provided significantly better results and offered insights into variable importance: balance, month, and age were top predictors. The model also demonstrated improved ability to classify minority ("yes") cases.

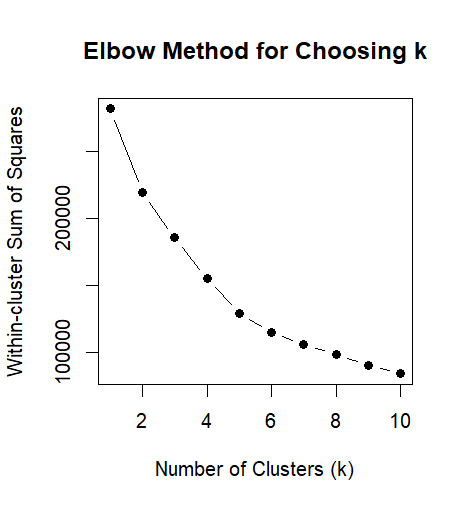
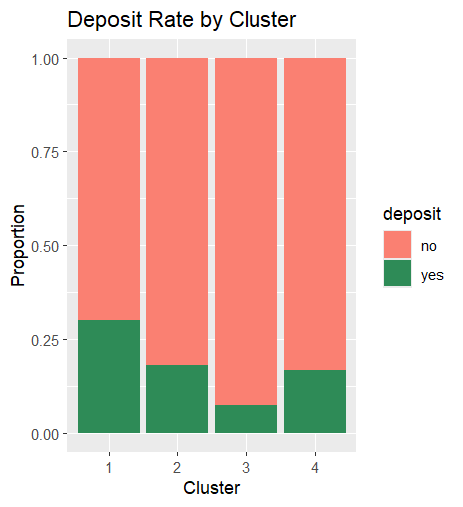
**Model 3: K-Means Clustering**

I performed unsupervised K-means clustering to segment customers based on standardized numeric features: age, balance, campaign, pdays, and previous. The elbow method suggested 3-4 clusters as optimal.

**Cluster Insights**:

* **Cluster 1**: Moderate age and balance, highest deposit rate (30%)
* **Cluster 2**: Older, wealthier clients, moderate deposit rate (18%)
* **Cluster 3**: Overexposed to campaigns (avg. 14 contacts), lowest deposit rate (7.5%)
* **Cluster 4**: Youngest group, low balance, low deposit rate (16.6%)

These clusters helped contextualize customer behavior and provided useful segmentation for campaign targeting.



**Model Comparison Summary**

| **Model** | **Accuracy** | **AUC** | **Strengths** | **Weaknesses** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 82.5% | 0.7341 | Interpretable, baseline model | Poor minority class detection |
| Random Forest | 89.3% | 0.9238 | Strong performance, good balance | Less interpretable |
| K-Means Clustering | N/A | N/A | Customer segmentation | Not predictive, needs context |

**3. Decisions**

Based on model performance and interpretability, Random Forest was the most effective predictive model. It significantly improved classification of deposit subscribers without sacrificing overall accuracy. Logistic regression served as a valuable baseline but fell short in identifying the minority class.

K-Means clustering added prescriptive value by segmenting clients into behaviorally distinct groups. For example, Cluster 1 had the highest subscription rate and could be targeted in future campaigns. Cluster 3, despite heavy contact frequency, had poor conversion, suggesting inefficiencies in targeting.

These insights can be used to:

* Prioritize marketing efforts toward high-likelihood clusters
* Tailor communication strategies per cluster profile
* Improve overall campaign efficiency and customer experience

In conclusion, the combination of predictive (classification) and prescriptive (clustering) approaches offers a comprehensive framework for strategic decision-making in customer targeting.