Predictive Marketing Analytics: Statistical Insights for Business Success

MATH 40028/50028: Statistical Learning

May 05, 2024

Introduction

Understanding customer behavior and predicting subscription patterns is paramount for financial institutions striving to optimize marketing strategies. By leveraging insights derived from past interactions, institutions can tailor their approaches effectively, thereby maximizing subscription rates and fostering long-term customer relationships.

Analyzing Telemarketing Campaign Data

The ‘Marketing_data.csv’ dataset, akin to bank.csv sourced from the UCI Machine Learning Repository, provides a comprehensive snapshot of customer interactions derived from telemarketing campaigns conducted by a Portuguese bank. With 11,162 observations and a diverse array of features, this dataset offers a rich repository of information crucial for understanding customer preferences and behaviors.

Dataset Composition and Focus

The dataset comprises 17 features, including target variables, demographic, financial, and campaign-specific variables, each offering valuable insights into customer engagement and subscription tendencies. The primary focus lies on the binary ‘deposit’ variable, indicating whether individuals subscribed to term deposits post-engagement with the bank’s telemarketing efforts.

Sampling Strategy and Potential Biases

The dataset comprises records from past campaigns, suggesting a target population of individuals contacted by the financial institution for marketing purposes. Sampling involves selecting 10% of examples from the larger dataset, ensuring representation while reducing computational load. However, the specifics of the sampling methodology are not provided, potentially introducing sampling bias.

Addressing Bias and Limitations

Various biases may arise, including selection bias from individuals responding to telemarketing calls and temporal bias due to the dataset’s chronological order. Additionally, self-reporting or response bias could impact data accuracy, influencing the dataset’s reliability.

Despite potential biases, the dataset offers valuable insights into client behavior and campaign performance. Analyzing relationships between client attributes and subscription outcomes enables the derivation of actionable insights to optimize future marketing strategies.

Prediction Problem:

The prediction problem is to develop the best classification model that predicts whether a customer will subscribe to a term deposit (‘yes’ or ‘no’) based on input features based like demographic, financial, and campaign-specific attributes.

Data Partitioning and Evaluation Strategy

The dataset will be split into 70% training and 30% test sets to ensure independence and prevent data leakage. This division precedes Exploratory Data Analysis (EDA) to maintain analysis integrity. Training data will inform machine learning models, while test data will assess model generalizability. Techniques like k-fold cross-validation and Bootstrapping will ensure accurate performance assessment using metrics such as Accuracy, Precision, Recall, F1 Score, and ROC-AUC, ensuring reliable predictive models for real-world applications.

Statistical learning strategies and methods

Data Cleaning: Upon inspection, no missing values and duplicates were found in the dataset, indicating data integrity.

Data Preprocessing: A new feature, ‘age_group,’ categorizes clients into age groups (‘Under 25’, ‘25-39’, ‘40-59’, ‘60+’). ‘month_part’ segments the last contact day into ‘start’, ‘middle’, and ‘end’. Interaction effects between “pdays” and “duration” were leveraged to create a new feature, capturing combined influence on customer engagement dynamics and deposit behavior. Categorical variables were converted into factors for improved modeling and analysis.

Exploratory Data Analysis (EDA) : Exploratory Data Analysis was performed on the training set to gain insights into the data distribution and correlations between features. Summary statistics are calculated for numerical variables, and graphical visualizations are created to visualize distributions and relationships. This helped in understanding the data characteristics and selecting appropriate features for modeling.

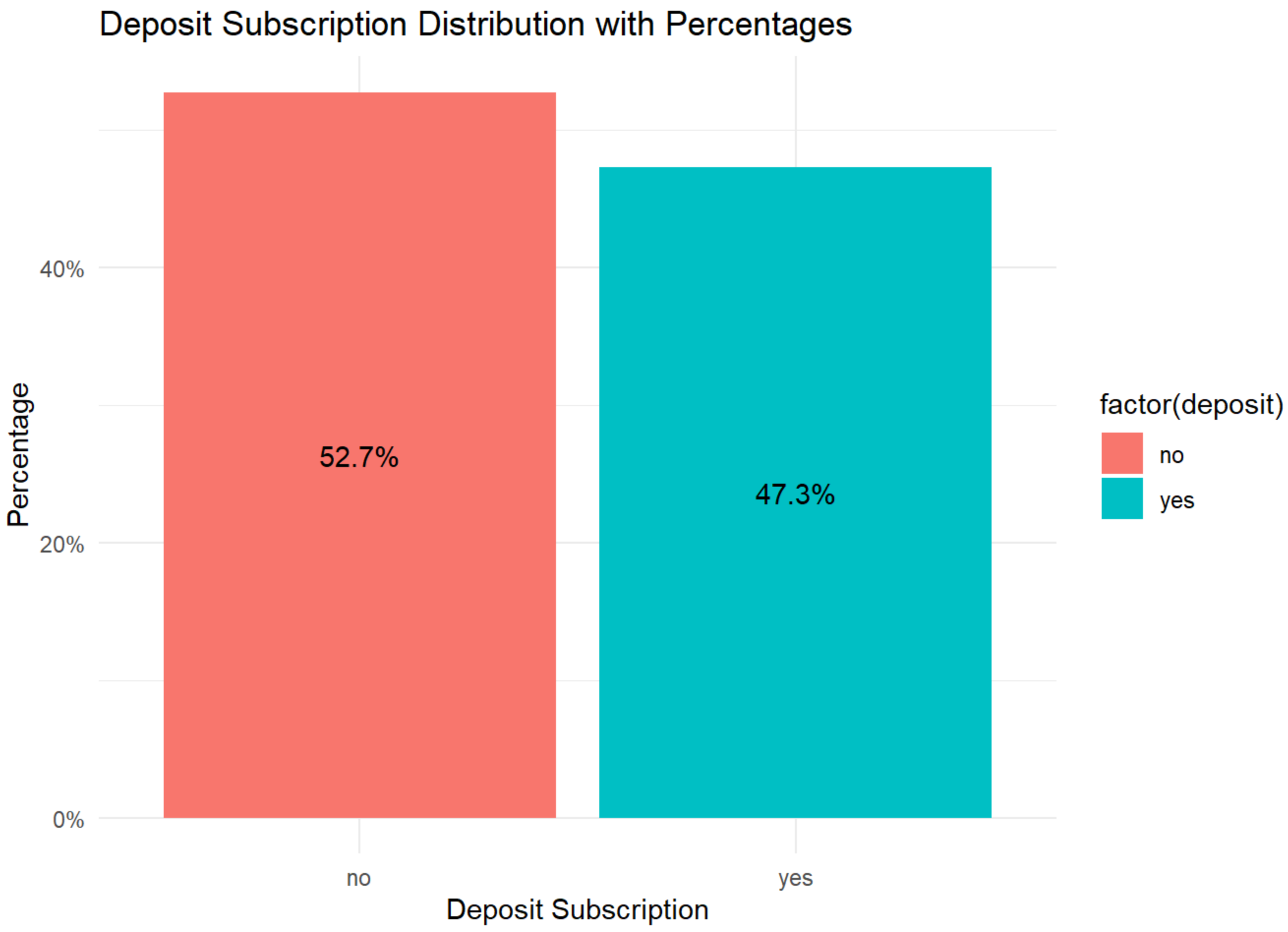
Correlation Analysis: Correlation analysis was conducted to examine relationships between numerical variables and identify potential predictors of client deposit behavior. The resulting correlation plot provided insights into the strength and direction of associations among variables.

Chi-square Test: Chi-square tests were employed to assess the association between categorical variables (e.g., job, marital status, education) and deposit outcomes.

Feature Selection The feature selection process involved logistic Lasso regression and random forest variable importance analysis. Lasso regression identified relevant predictors such as job type, housing status, and campaign-related variables. Random forest highlighted ‘duration’ as most influential, followed by ‘month’, ‘age’, and ‘balance’, providing valuable insights into deposit decision drivers. These methods collectively enhance model interpretability and inform marketing strategies.

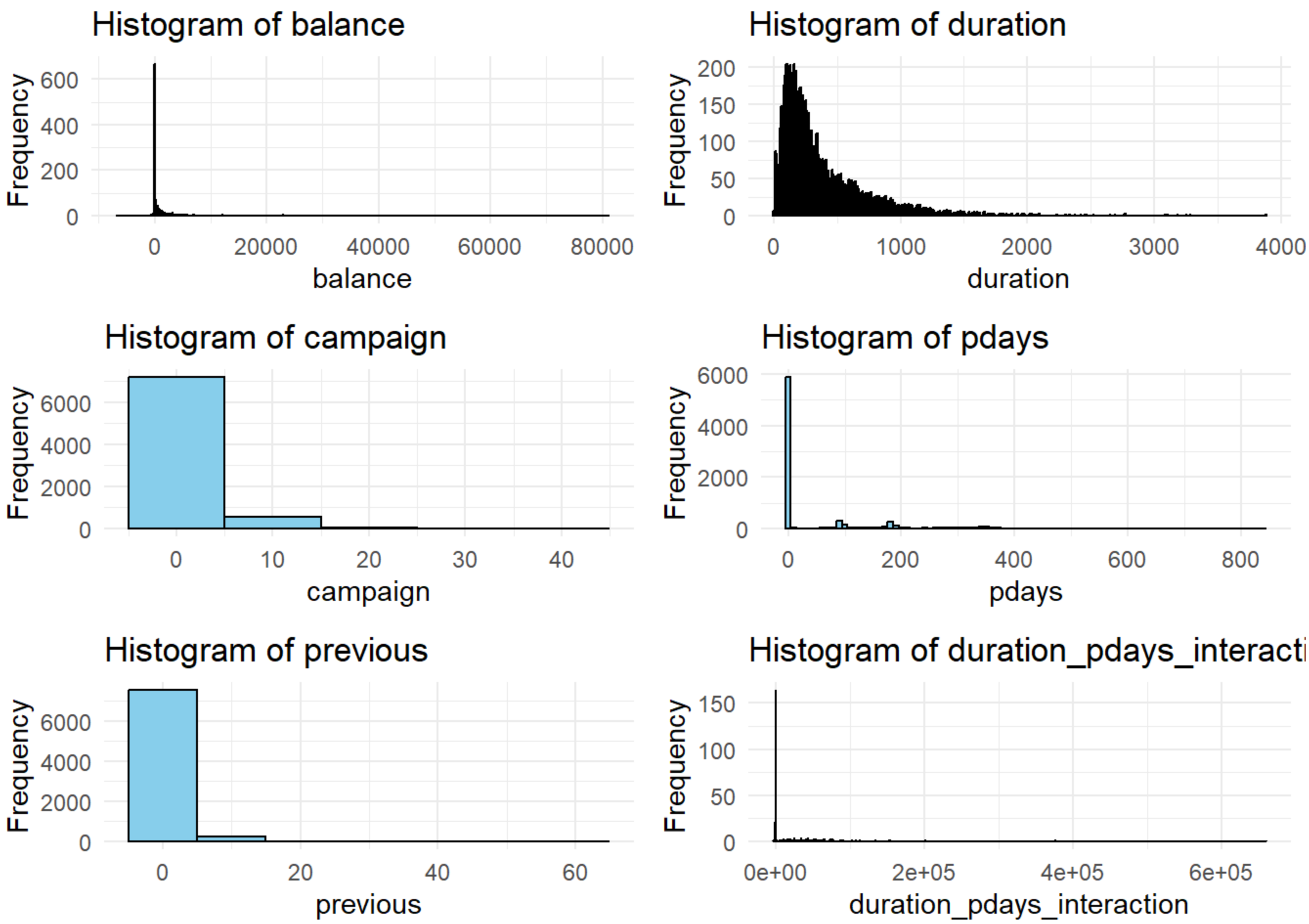
Exploratory Data Analysis

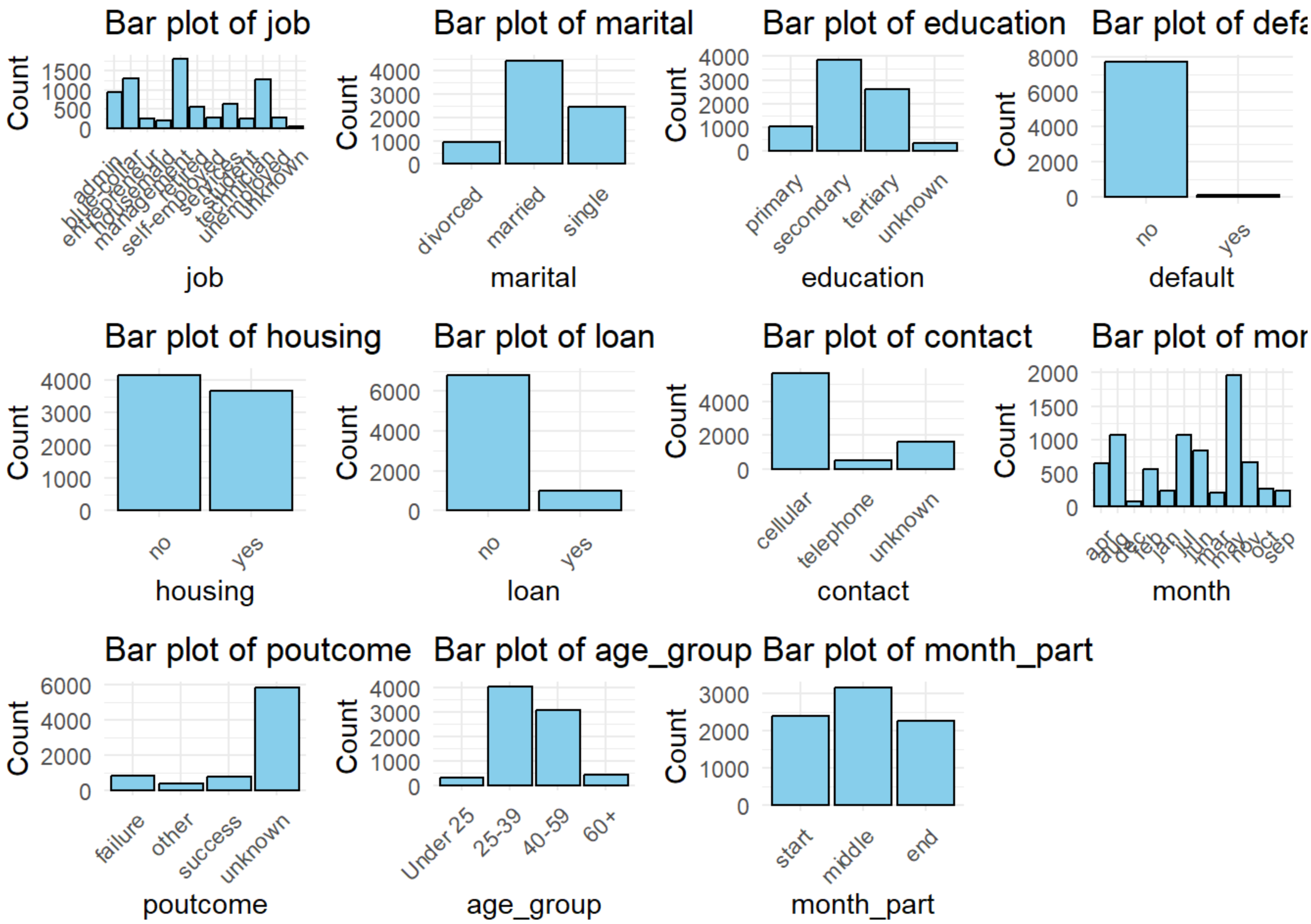
Distribution of term deposits - Target Variable



The graph depicts the distribution of the “deposit” variable in the training dataset. The majority of clients (52.7%) did not subscribe to a term deposit, while 47.3% did. This indicates a relatively balanced distribution.

Visualizing Numerical(Histograms) and Categorical Variables(Bar Plots)

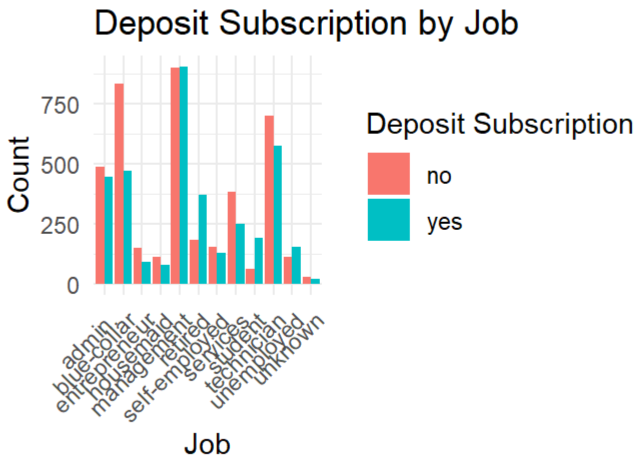


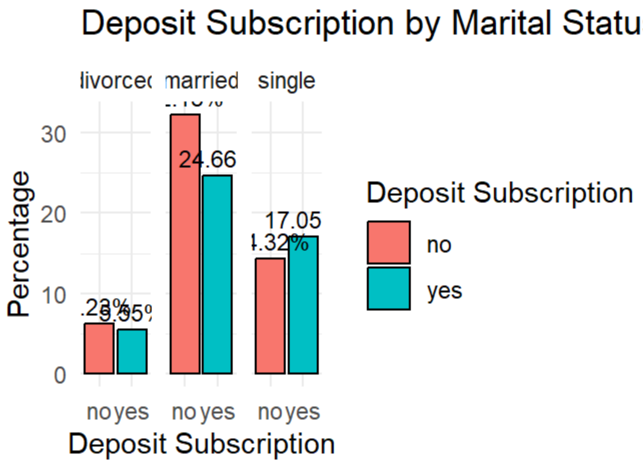


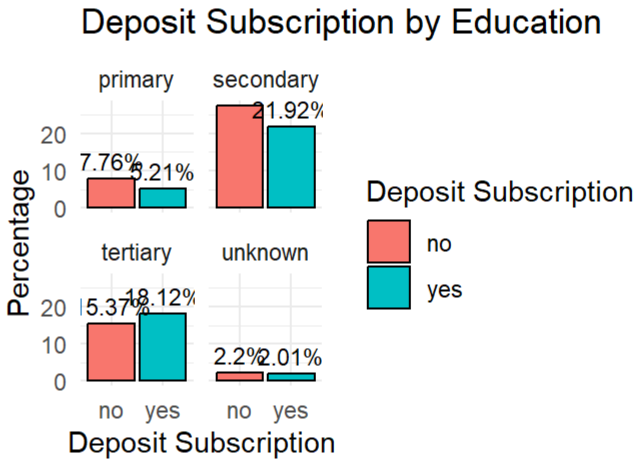
Findings from Visualizations:

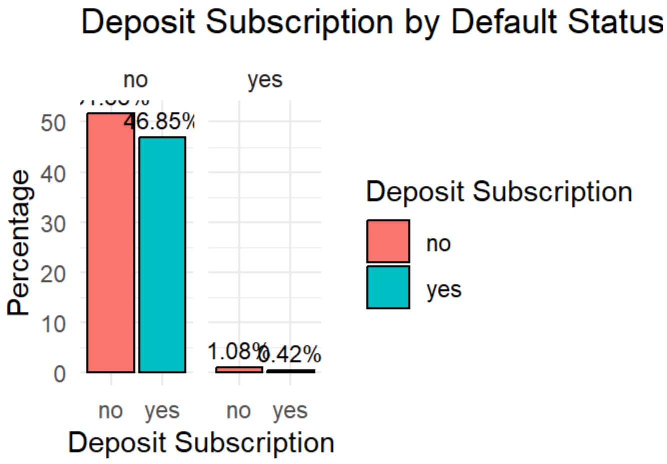
Analysis of numerical and categorical variables uncovers key insights into client demographics and behavior. “Balance” skews lower for most clients but with notable outliers having higher balances. “Duration” and “campaign” skew right, indicating shorter interactions and fewer contacts for most, with some outliers representing longer durations and higher contact frequencies. “Pdays” is heavily right-skewed, suggesting recent contact for many but with outliers indicating longer gaps since last contact. “Previous” distribution is right-skewed, indicating fewer previous contacts for most but some outliers with higher frequencies. Contact occurs consistently throughout the month, with a peak during the middle phase. Clients are balanced across age groups, with significant representation in 25-39 and 40-59 brackets. Common occupations include management, blue-collar, and technician roles. Most clients are married, followed by singles and divorced individuals, with varying education levels. Few defaults or personal loans are present, but housing loans are common. Cellular phones are the primary contact method. Previous campaign outcomes vary, with a significant portion categorized as “unknown.” Monthly contact patterns peak in May, followed by August and July. “Duration_pdays_interaction” shows significant skewness, suggesting potential outliers. These insights are crucial for refining marketing strategies and improving campaign effectiveness.

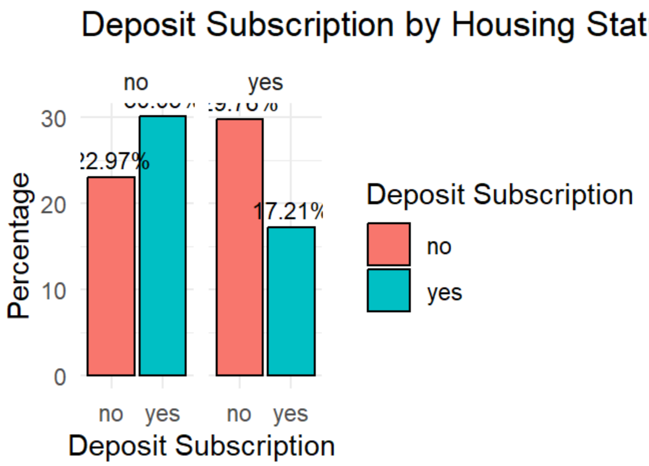
Demographic Insights and Seasonal Trends in Deposit Subscription Behavior

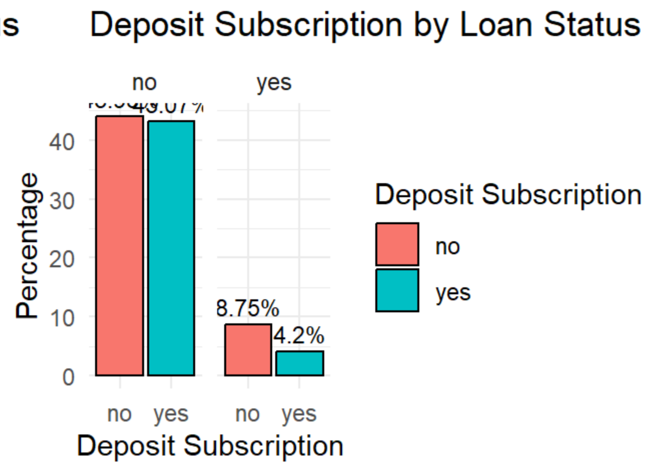

Deposit Subscription by Job

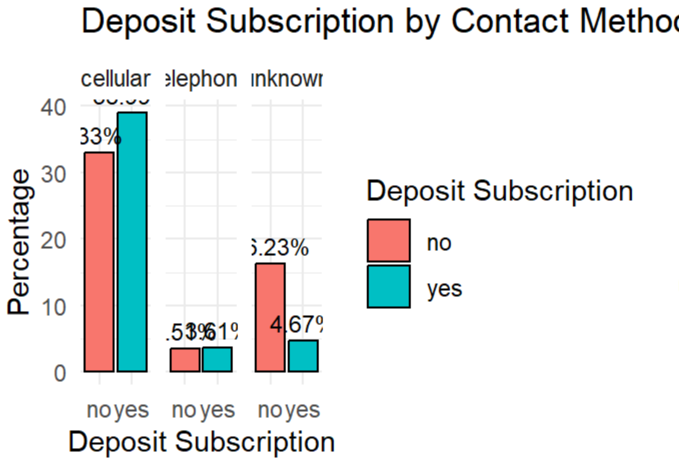

Deposit Subscription by Marital Status

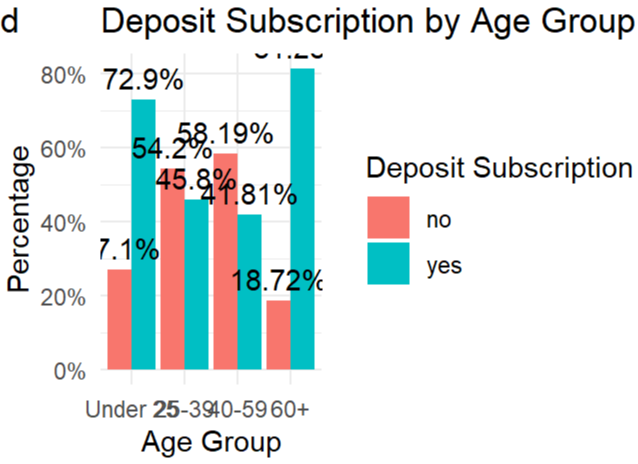

Deposit Subscription by Education

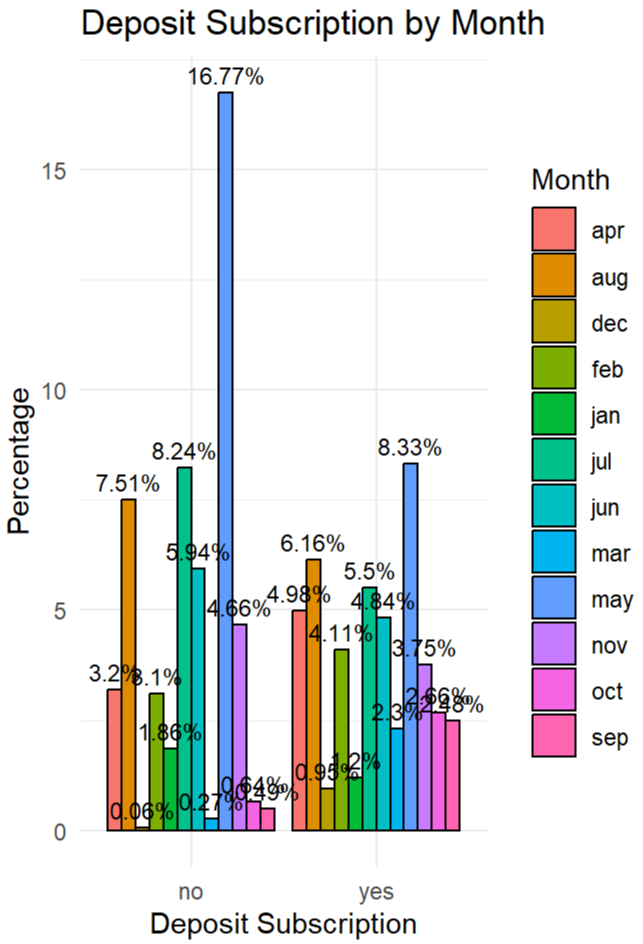

Deposit Subscription by Default Status

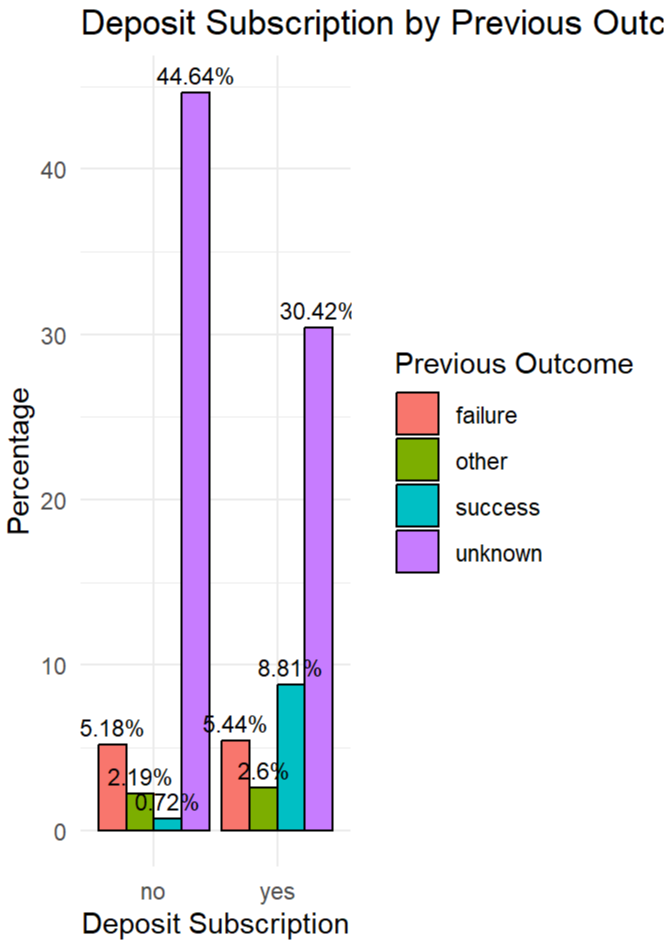

Deposit Subscription by Housing Status

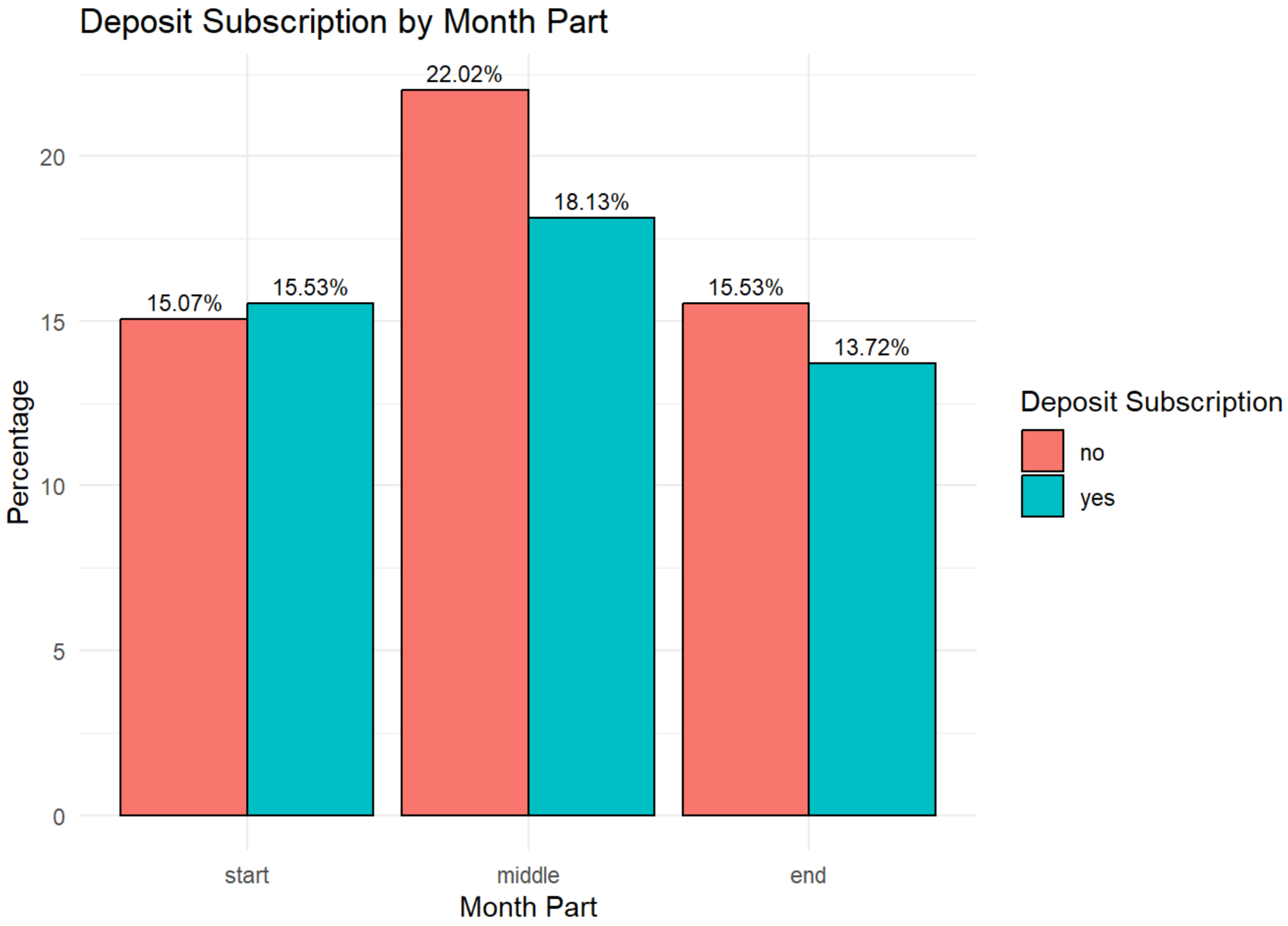

Deposit Subscription by Loan Status


Deposit Subscription by Contact Method


Deposit Subscription by Age Group

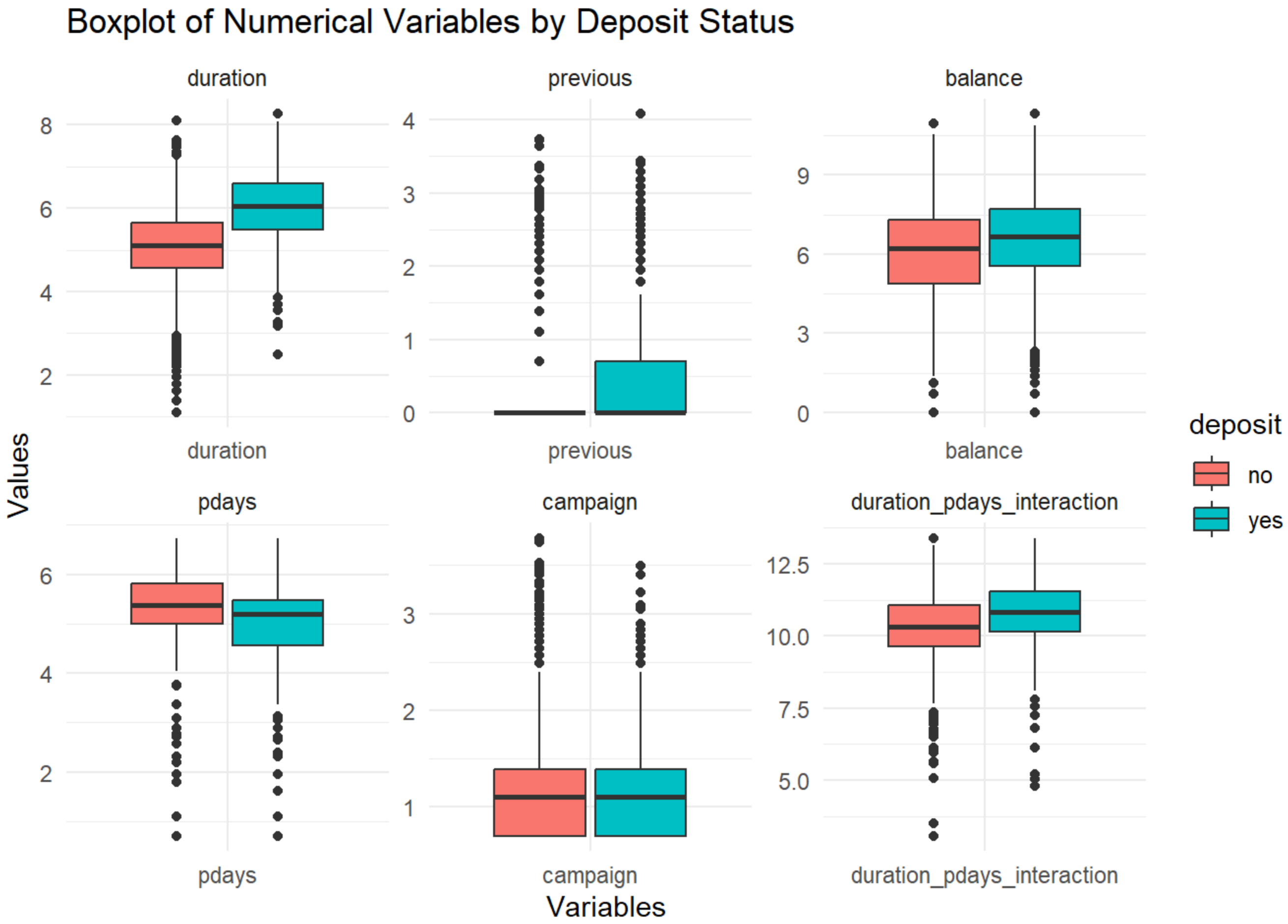

Deposit Subscription by Month


Deposit Subscription by Previous Outc



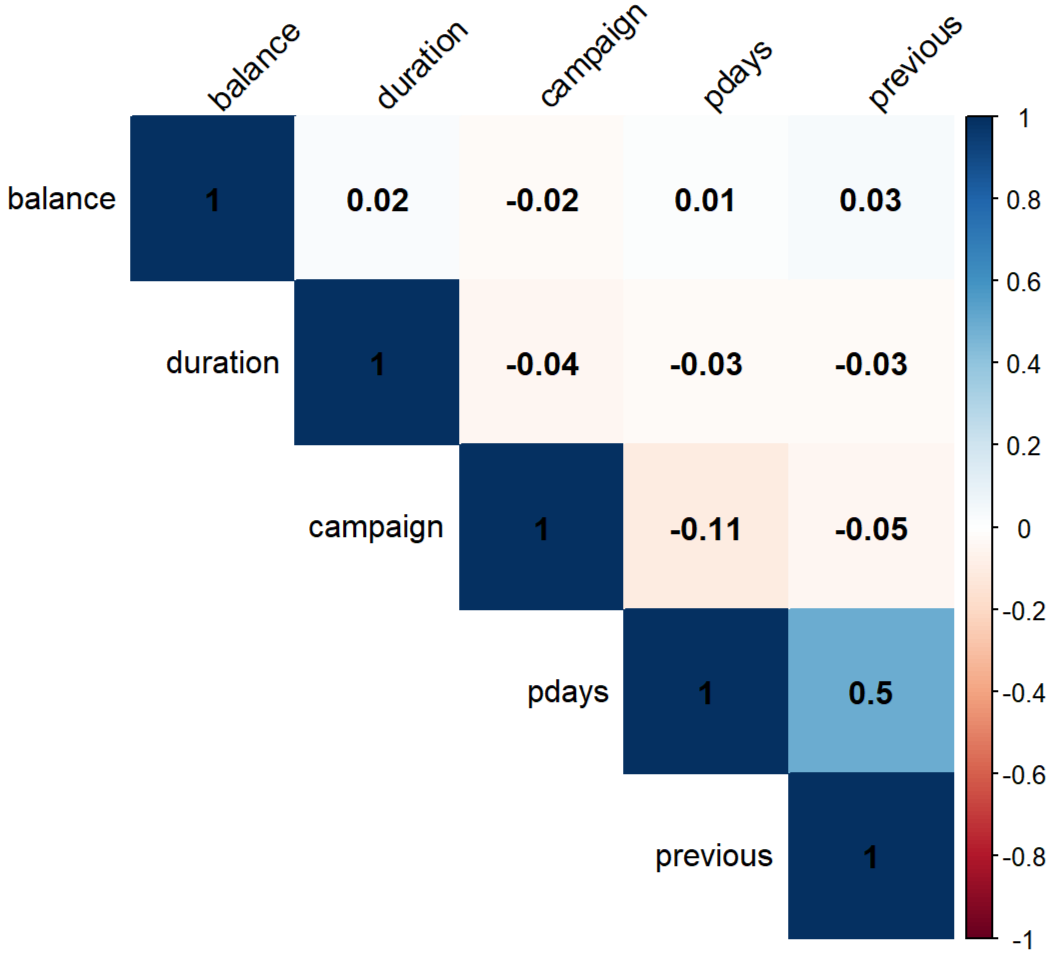
####**Key Insights:**

The analysis highlights diverse subscription patterns across demographic factors. Management and Technician roles show balanced outcomes, while Blue-collar and Services display lower subscription rates. Retired and Student categories exhibit higher subscription rates. Married individuals have the highest non-subscription proportion, followed by singles and divorced individuals. Tertiary-educated clients show the highest subscription rate, while those with unknown education levels show the lowest. Clients with no defaults have higher subscription rates, while those with defaults subscribe less. Those without housing or personal loans are more likely to subscribe. Cellular communication leads to higher subscription rates compared to other contact methods. Older clients are more likely to subscribe, with March, December, and October showing peak subscription rates. Previous successful outcomes positively influence deposit subscriptions. Contacts made during the middle of the month have higher subscription rates. These insights emphasize the significance of client characteristics and campaign timing in deposit subscription behavior.



Deposit interactions show longer durations and more previous contacts compared to non-deposit interactions, indicating sustained engagement. Those making deposits also tend to have higher median balances, hinting at a link between financial stability and deposit likelihood. Additionally, deposit interactions have longer gaps between contacts and a unique interaction effect between contact duration and days since the last contact, suggesting timing and engagement intensity play key roles in predicting deposit outcomes. These findings highlight the significance of consistent outreach efforts and financial stability in encouraging deposit behavior.

Correlation Analysis for Deposit Prediction



The correlation analysis provides insights into the interrelationships among various variables in the dataset. Notably, there are weak correlations observed between most pairs of variables, indicating that they tend to vary independently of each other. This suggests that factors such as account balance, duration of last contact, number of contacts during the campaign, days since the last contact from a previous campaign, and number of previous contacts do not strongly influence each other. However it reveals moderate positive correlation (0.497) between “pdays” (days since last contact) and “previous” (number of previous contacts), indicating clients contacted frequently in the past tend to have longer intervals between contacts. Additionally, a weak negative correlation (-0.105) exists between “pdays” and “campaign” (number of contacts in current campaign), suggesting recently contacted clients require fewer contacts.

Chi-Square Tests for Categorical Variables with the target variable - “deposit”.

```
##
##  Pearson's Chi-squared test
##
## data:  job_cross_tab
## X-squared = 278.88, df = 11, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test
##
## data:  marital_cross_tab
## X-squared = 76.039, df = 2, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test
##
## data:  education_cross_tab
## X-squared = 81.879, df = 3, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  default_cross_tab
## X-squared = 16.548, df = 1, p-value = 4.743e-05
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  housing_cross_tab
## X-squared = 312.3, df = 1, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  loan_cross_tab
## X-squared = 102.26, df = 1, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test
##
## data:  contact_cross_tab
## X-squared = 516.6, df = 2, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test
##
## data:  month_cross_tab
## X-squared = 709.76, df = 11, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test
##
## data:  poutcome_cross_tab
## X-squared = 729.38, df = 3, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test
##
## data:  age_cross_tab
## X-squared = 320.46, df = 3, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test
##
## data:  month_part_cross_tab
## X-squared = 21.889, df = 2, p-value = 1.766e-05
```

The results of Pearson's Chi-squared tests reveal crucial insights into the relationship between various categorical variables and deposit subscription status. Each variable examined demonstrates a significant association with the decision to subscribe to a deposit, highlighting the diverse factors influencing clients' financial choices. Age Groups, Job categories, marital status, and education levels all exhibit distinct patterns, suggesting that demographic factors play a pivotal role in deposit decisions. Additionally, the presence of default status, housing loans, and personal loans significantly impacts subscription rates, reflecting the influence of financial circumstances. Moreover, the choice of contact method, timing of outreach, days of a month and outcomes of previous campaigns emerge as key determinants, underscoring the importance of effective marketing strategies in promoting deposit subscriptions. In summary, these findings provide valuable insights for banking professionals, emphasizing the multifaceted nature of client behavior and the need for tailored approaches to attract deposit subscriptions.

Statistical Learning Methods and Term Deposit Prediction: Applicability Analysis

Exploring the nuanced applicability of statistical learning methods to predicting term deposit subscriptions is crucial for informed decision-making in banking and finance. The below are the methods employed in this predictive analysis.

Random Forest: Random forest is a robust ensemble learning method that is well-suited for the prediction problem of classifying term deposit subscriptions. This method combines multiple decision trees to make predictions, effectively capturing non-linear relationships and interactions between demographic, financial, and campaign-specific attributes. Random forest can handle both numerical and categorical data, making it suitable for diverse datasets. Moreover, random forest is less sensitive to outliers and does not require feature scaling, simplifying the preprocessing steps. By leveraging the predictive capabilities of multiple decision trees, random forest can provide accurate predictions of term deposit subscriptions while reducing the risk of overfitting. Therefore, random forest is a powerful tool for addressing the prediction problem at hand.

Support Vector Machines (SVM): Support Vector Machines (SVM) offer a versatile approach for the prediction problem of term deposit subscription classification. SVMs aim to find the optimal hyperplane that separates classes within the feature space, making them effective for both linear and non-linear classification tasks. In the context of predicting term deposit subscriptions, SVMs can capture complex relationships between demographic, financial, and campaign-specific attributes and the likelihood of subscription. This method is particularly useful when there is a clear margin of separation between subscribing and non-subscribing customers. Additionally, SVMs can handle high-dimensional data and non-linear relationships through the use of kernel functions, providing flexibility in modeling various aspects of the prediction problem.

Logistic Regression: Logistic regression is a suitable approach for the prediction problem of classifying whether a customer will subscribe to a term deposit. This method assumes a linear relationship between the input features, such as demographic, financial, and campaign-specific attributes, and the log-odds of the binary outcome variable. Since the outcome variable is binary (‘yes’ or ‘no’), logistic regression can effectively model the probability of a customer subscribing to a term deposit based on these attributes. Moreover, logistic regression provides interpretable results, allowing stakeholders to understand the impact of each predictor on the likelihood of subscription. Therefore, logistic regression is a practical choice for this prediction problem, providing insights into the factors influencing term deposit subscriptions.

Decision Trees Decision trees offer a superior approach for predicting term deposit subscriptions due to their ability to capture complex decision-making processes inherent in the prediction problem. With input features such as demographic, financial, and campaign-specific attributes, decision trees can efficiently partition the feature space to identify key predictors influencing subscription behavior. This method’s interpretability is particularly advantageous for understanding the underlying factors driving customer decisions, enabling stakeholders to gain actionable insights into customer behavior. Moreover, decision trees excel in handling both numerical and categorical data, making them suitable for diverse datasets commonly encountered in banking and finance. Overall, decision trees represent a robust and interpretable approach for predicting term deposit subscriptions, offering valuable insights into customer behavior and aiding in informed decision-making processes.

In summary, each method possesses unique strengths and assumptions, with their applicability to the prediction problem contingent upon various factors such as data characteristics, relationship complexity, and computational constraints. By leveraging a combination of methods and conducting comprehensive performance evaluations, insights into the most suitable approach for the given dataset and prediction objective can be gleaned.

Predictive analysis and results

Feature Selection with Logistic Lasso Regression

The Lasso regression with logistic regression method was employed focusing on feature selection for enhanced model interpretability. Through cross-validation, an optimal lambda value of 0.0011 was determined, guiding the selection of relevant features such as job type, housing status, month, campaign, pdays, previous, age_groups and previous campaign outcomes.

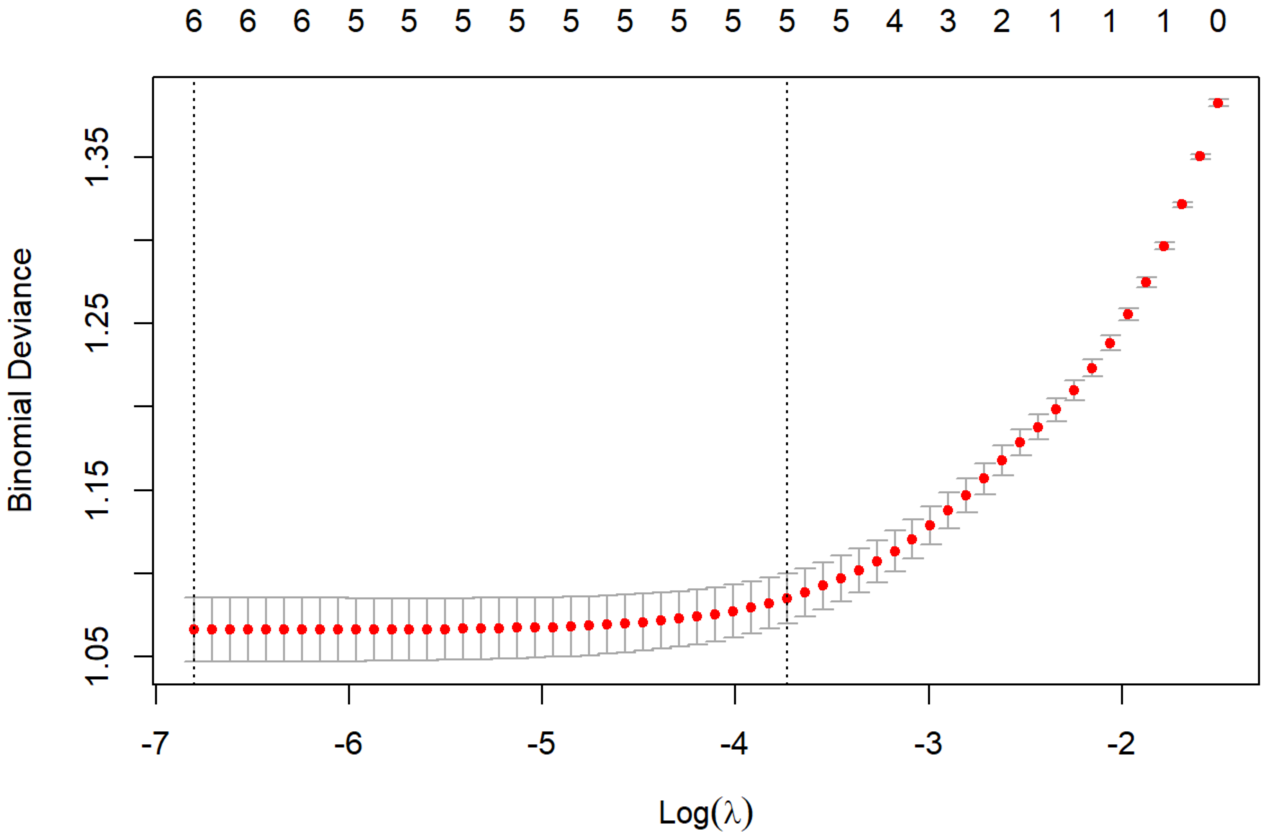
```
## Lasso Logistic Regression for feature selection

# Create a matrix for predictors in the training data
X_train <- as.matrix(train_data[, -which(names(train_data) == "deposit")])

# Create a response vector
y_train <- ifelse(train_data$deposit == "yes", 1, 0)  # Convert "deposit" to binary: 1 for "yes", 0 for "no"

# Perform Lasso logistic regression with cross-validation
lasso_cv <- cv.glmnet(X_train, y_train, alpha = 1, family = "binomial")

plot(lasso_cv)
```



```
# Get the optimal lambda value
optimal_lambda <- lasso_cv$lambda.min

print(paste("Optimal lambda value for Lasso Regression :", optimal_lambda))
```

```
## [1] "Optimal lambda value for Lasso Regression : 0.00110858157639337"
```

```
# Fit the final Lasso logistic regression model with the optimal lambda value
lasso_model <- glmnet(X_train, y_train, alpha = 1, lambda = optimal_lambda, family = "binomial")

# Get selected feature indices
selected_indices <- which(coef(lasso_model) != 0)

# Print selected features
cat("Selected Features:", "\n")
```

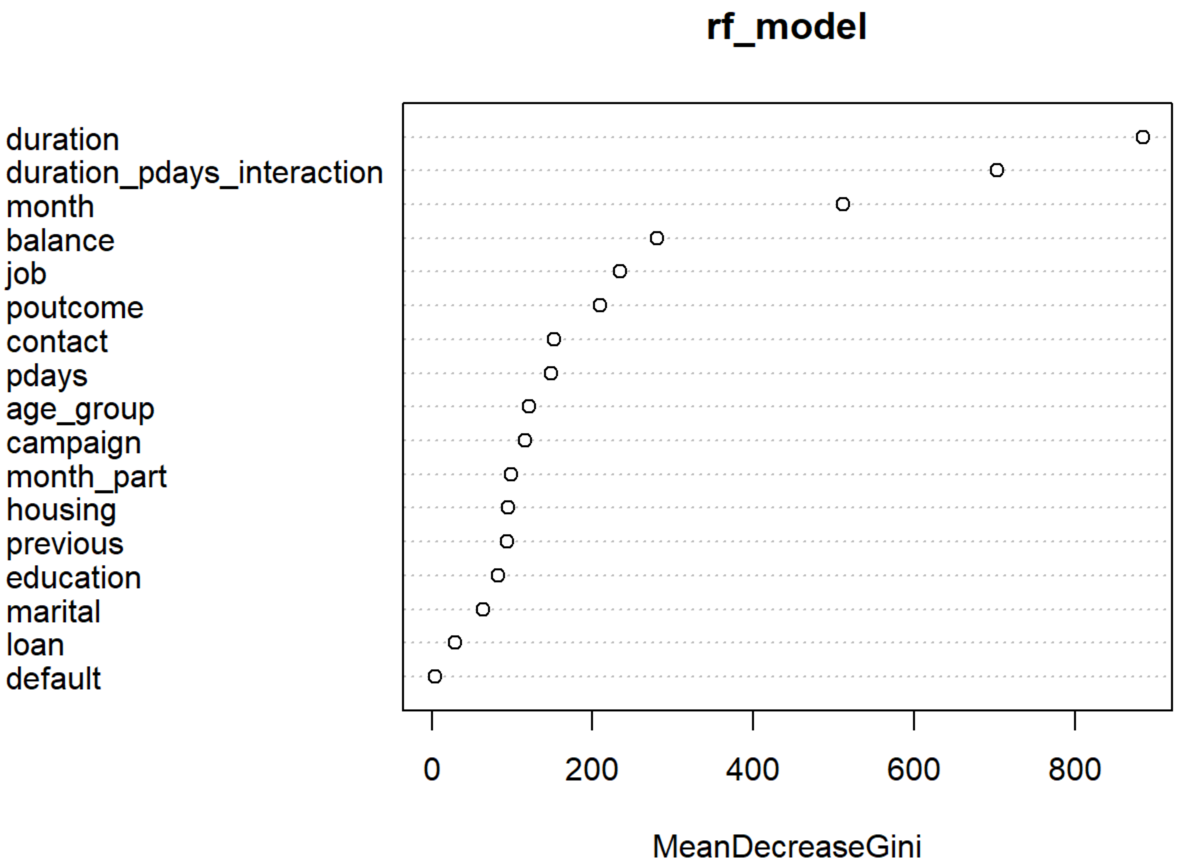
```
## Selected Features:
```

```
cat(names(train_data)[-which(names(train_data) == "deposit")][selected_indices], "\n")
```

```
## job housing campaign pdays previous poutcome age_group
```

Random Forest with selected features from Variable Importance

The variable importance analysis of the random model highlights key predictors influencing customer deposit behavior. ‘Duration’ emerges as the most influential factor, indicating the importance of the last contact duration during the campaign. The interaction between ‘pdays’ and ‘duration’ suggests varying client responses based on contact timing. Additionally, ‘month’ demonstrates notable importance, reflecting seasonal trends or campaign effectiveness. ‘Age’ and ‘day’ also play significant roles, along with ‘balance’, ‘pdays’, and ‘poutcome’, emphasizing their relevance in predicting deposit outcomes. These insights inform strategic decision-making for future marketing and engagement initiatives.



Performance Analysis of Random Forest Models on test data with Variable Selection Techniques

The evaluation metrics provide valuable insights into the performance of the Random Forest models trained on different sets of variables.

- **All Variables Model**: This model, trained with all available predictor variables, demonstrates solid performance across all metrics. It achieves high precision 89.91%, recall 81.80%, and F1 score 85.66%, indicating a good balance between identifying positive instances (deposits) and minimizing false positives. The ROC AUC of 91.92% suggests strong discrimination ability, while the test accuracy of 85.66% indicates overall model effectiveness.

- **Top 10 Variables Model**: Focused on the top 10 predictors based on variable importance, this model maintains strong performance, albeit slightly lower than the all-variables model. It still achieves commendable precision 88.16%, recall 80.71%, and F1 score 84.27%. The ROC AUC of 91.12% indicates good discrimination ability, and the test accuracy of 84.23% reflects its overall effectiveness. **Despite the reduced feature set, the model maintains robustness, indicating substantial predictive power in the selected predictors. This streamlined approach simplifies the model and enhances interpretability without compromising predictive accuracy, albeit slightly lower than the model with all predictors.**

- **Model Selected by Lasso**: In contrast, the model selected by Lasso regularization exhibits notably different performance characteristics. While it achieves high recall 88.70%, indicating its ability to capture a large proportion of positive instances, its precision 67.05% and F1 score 76.37% are comparatively lower. The ROC AUC of 72.96% suggests weaker discrimination ability, and the test accuracy of 71.27% indicates reduced overall effectiveness compared to the other models.

In summary, the all-variables and top-10-variables models perform similarly well, with robust precision-recall balance and high overall accuracy. Conversely, the Lasso-selected model, while excelling in recall, sacrifices precision and overall model effectiveness, indicating potential overfitting or selection bias in the feature selection process. Therefore, for this dataset, leveraging the broader set of variables or focusing on the top predictors appears to yield more robust and effective Random Forest models.**While feature selection techniques like Lasso offer model interpretability and reduced complexity, the emphasis on accuracy prompts the utilization of all variables in the predictive model.**

```
## [1] "Evaluation Metrics of Random Forest Models"
```

```
##                     Precision Recall    F1_Score  ROC_AUC   Test_Accuracy
## All Variables       0.8964218 0.8146035 0.8535565 0.9191116 0.8536877
## Top 10 Variables    0.8809227 0.8060468 0.8418231 0.9104246 0.8414452
## Selected by Lasso   0.6712624 0.8887621 0.7648503 0.7294032 0.7139445
```

Performance Evaluation on test data Using Resampling Techniques - Random Forest

In evaluating model performance through bootstrapping and 5-fold cross-validation, we observe consistent and robust results across the models:

- **Baseline (All Variables)**: The model trained on all available predictor variables demonstrates stable and reliable performance across both bootstrapping and 5-fold cross-validation. With an accuracy of 85.67% in the baseline setting, it achieves high precision (89.81%), recall (81.92%), and F1 score (85.68%). The ROC AUC of 91.94% underscores its strong discriminatory capability.

- **Bootstrapping**: Employing bootstrapping for model evaluation yields results comparable to the baseline model. The bootstrapped model maintains a high accuracy of 85.31% along with commendable precision (89.32%), recall (81.70%), and F1 score (85.34%). The ROC AUC of 91.59% indicates robust discrimination, consistent with the baseline performance.

- **5-Fold Cross-Validation**: Similarly, utilizing 5-fold cross-validation for evaluation yields performance metrics closely aligned with the baseline and bootstrapping approaches. With an accuracy of 85.79%, the cross-validated model demonstrates strong precision (89.78%), recall (82.20%), and F1 score (85.82%). The ROC AUC of 91.67% further validates its discriminatory power.

In summary, both bootstrapping and 5-fold cross-validation techniques provide reliable estimates of model performance, consistently reaffirming the efficacy of the baseline model trained on all variables. These resampling methods offer robust assessments of model generalization and can guide decision-making in deploying the model for real-world applications.

```
## [1] "Evaluation Metrcis of the Random Forest Models: "
```

```
##                       Model  Accuracy Precision    Recall  F1_Score   AUC_ROC
## 1 Baseline (All Variables) 0.8536877 0.8949343 0.8163149 0.8538186 0.9192464
## 2            Bootstrapping 0.8526515 0.8923642 0.8170850 0.8530521 0.9156781
## 3  5-Fold Cross-Validation 0.8542849 0.8975487 0.8146035 0.8540670 0.9174927
```

Random Forest Model Performance on test data with Bootstrapping (Confidence Intervals)

The Random Forest model, incorporating all predictors with bootstrapping, demonstrates robust performance on the test data. It achieves an average accuracy of 85.30% (CI: [84.81%, 85.77%]), indicating the proportion of correctly classified instances. The model exhibits strong predictive capability with an average precision of 89.34% (CI: [88.60%, 89.92%]) and recall of 81.68% (CI: [80.48%, 82.57%]). This balance between precision and recall is further reflected in the model's F1 score of 85.34% (CI: [84.80%, 85.83%]). Moreover, the average AUC of 91.57% (CI: [91.33%, 91.87%]) underscores the model's effective discrimination between positive and negative cases. Overall, these metrics reaffirm the reliability and consistency of the Random Forest model in predictive tasks, strengthened by the application of bootstrapping techniques.

```
## [1] "Evaluation Metrics of Random Forest Model with bootstrapping are: "
## [1] "Average Accuracy: 0.852747088683189"
## [1] "Confidence Interval (Accuracy): [ 0.847260376231711 ,  0.857128993729471 ]"
## [1] "Average Precision: 0.892829652364044"
## [1] "Confidence Interval (Precision): [ 0.884232075495888 ,  0.899563303528492 ]"
## [1] "Average Recall: 0.816748431260696"
## [1] "Confidence Interval (Recall): [ 0.807458642327439 ,  0.825171135196806 ]"
## [1] "Average F1 Score: 0.85308110254738"
## [1] "Confidence Interval (F1 Score): [ 0.848205068728892 ,  0.857401489425082 ]"
## [1] "Average AUC: 0.915956062432179"
## [1] "Confidence Interval (AUC): [ 0.91339452363796 ,  0.918233417792913 ]"
```

Random Forest Model Performance with 5-Fold Cross-Validation on held-out fold and Test Data

The Random Forest model, evaluated using 5-fold cross-validation on the held-out fold, demonstrates an average accuracy of 85.68%, with an average precision of 90.28% and recall of 81.63%. The F1 score, a harmonic mean of precision and recall, stands at 85.74%. These metrics indicate the model's consistent performance across different subsets of the validation data. On the test data, the Random Forest model maintains its effectiveness with an accuracy of 85.58%. It achieves a precision of 89.64% and recall of 81.92%, resulting in an F1 score of 85.60% and an AUC of 91.91%. These results confirm the model's reliability and generalization ability, as it performs consistently well on unseen data, aligning closely with the performance observed during cross-validation.

```
## [1] "Average Evaluation Metrics of Random Forest(5-Fold Cross-Validation) on the held-out fol
d:"
## [1] "Average Accuracy: 0.853961528725661"
## [1] "Average Precision: 0.897686772453933"
## [1] "Average Recall: 0.815907041426123"
## [1] "Average F1 Score: 0.854734404999799"
## [1] "Evaluation Metrics of Random Forest(5-Fold Cross-Validation) on Test Data:"
## [1] "Test Accuracy: 0.857270827112571"
## [1] "Precision: 0.896208825357365"
## [1] "Recall: 0.822589845978323"
## [1] "F1 Score: 0.857822724568709"
## [1] "ROC AUC for Test Data: 0.919106629951948"
```

Significant Predictor Variables in Logistic Regression Model

The logistic regression model reveals several significant predictor variables beyond housing status and call duration that contribute to predicting the outcome. Factors such as job type, education status, housing and personal loan status, mode and month of contact, duration of call, number of contacts performed during the current campaign, Age of the clients and previous campaign outcomes demonstrate notable coefficients, indicating their impact on the likelihood of a positive outcome.

```
##
## Call:
## glm(formula = formula, family = "binomial", data = train_data)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 3.130e-01  3.180e-01   0.984 0.324982
## jobblue-collar             -2.718e-01  1.268e-01  -2.144 0.032048 *
## jobentrepreneur            -2.865e-01  2.087e-01  -1.373 0.169792
## jobhousemaid               -4.859e-01  2.282e-01  -2.129 0.033256 *
## jobmanagement              -2.931e-01  1.292e-01  -2.269 0.023291 *
## jobretired                 -3.240e-01  1.923e-01  -1.685 0.091998 .
## jobself-employed           -4.653e-01  1.951e-01  -2.385 0.017075 *
## jobservices                -2.571e-01  1.457e-01  -1.764 0.077673 .
## jobstudent                  2.546e-01  2.201e-01   1.157 0.247202
## jobtechnician              -1.728e-01  1.199e-01  -1.441 0.149690
## jobunemployed              -1.004e-01  1.959e-01  -0.512 0.608356
## jobunknown                 -3.602e-01  4.192e-01  -0.859 0.390194
## maritalmarried             -1.574e-01  1.028e-01  -1.531 0.125744
## maritalsingle               4.051e-02  1.166e-01   0.347 0.728316
## educationsecondary          1.983e-01  1.133e-01   1.750 0.080130 .
## educationtertiary           5.391e-01  1.337e-01   4.030 5.57e-05 ***
## educationunknown           -6.100e-03  1.892e-01  -0.032 0.974281
## defaultyes                 -1.333e-01  2.746e-01  -0.486 0.627229
## balance                     1.831e-05  1.078e-05   1.698 0.089453 .
## housingyes                 -6.150e-01  7.481e-02  -8.220  < 2e-16 ***
## loanyes                    -5.158e-01  1.020e-01  -5.055 4.30e-07 ***
## contacttelephone           -9.494e-02  1.295e-01  -0.733 0.463318
## contactunknown             -1.645e+00  1.182e-01 -13.922  < 2e-16 ***
## monthaug                   -7.516e-01  1.342e-01  -5.602 2.11e-08 ***
## monthdec                    1.719e+00  5.035e-01   3.414 0.000641 ***
## monthfeb                   -1.437e-01  1.523e-01  -0.944 0.345315
## monthjan                   -1.258e+00  2.040e-01  -6.166 7.00e-10 ***
## monthjul                   -9.460e-01  1.371e-01  -6.900 5.21e-12 ***
## monthjun                    4.538e-01  1.608e-01   2.821 0.004785 **
## monthmar                    2.092e+00  2.699e-01   7.749 9.25e-15 ***
## monthmay                   -5.644e-01  1.283e-01  -4.400 1.08e-05 ***
## monthnov                   -8.069e-01  1.466e-01  -5.506 3.68e-08 ***
## monthoct                    8.975e-01  2.154e-01   4.167 3.09e-05 ***
## monthsep                    7.702e-01  2.322e-01   3.317 0.000909 ***
## duration                    5.633e-03  1.609e-04  34.999  < 2e-16 ***
## campaign                   -8.687e-02  1.588e-02  -5.471 4.48e-08 ***
## pdays                       4.678e-04  6.613e-04   0.707 0.479311
## previous                    1.565e-02  1.660e-02   0.943 0.345719
## poutcomeother              -1.291e-01  1.585e-01  -0.814 0.415405
## poutcomesuccess             2.212e+00  1.761e-01  12.557  < 2e-16 ***
## poutcomeunknown            -5.116e-01  1.652e-01  -3.097 0.001953 **
## duration_pdays_interaction -3.683e-06  1.359e-06  -2.710 0.006725 **
## age_group25-39             -1.094e+00  1.804e-01  -6.065 1.32e-09 ***
## age_group40-59             -1.091e+00  1.904e-01  -5.732 9.95e-09 ***
## age_group60+                1.306e-01  2.743e-01   0.476 0.634011
## month_partmiddle           -4.656e-02  8.035e-02  -0.579 0.562306
## month_partend               1.482e-01  9.091e-02   1.630 0.103019
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10807.8  on 7812  degrees of freedom
## Residual deviance:  6280.3  on 7766  degrees of freedom
## AIC: 6374.3
##
## Number of Fisher Scoring iterations: 6
```

Performance Evaluation on test data Using Resampling Techniques - Logistic Regression

In evaluating the logistic regression models through bootstrapping and 5-fold cross-validation, we analyze the following results:

- **Baseline (All Variables)**: The logistic regression model achieved an accuracy of 82.17%, with precision and recall rates at 81.45% and 85.40%, respectively. The F1 score, reflecting a balanced measure of precision and recall, stands at 83.38%. Notably, the model exhibits strong discriminatory capability, as evidenced by an AUC-ROC value of 90.31%.

- **Bootstrapping**: Employing bootstrapping for evaluation yielded results consistent with the baseline model. The bootstrapped logistic regression model maintained a high accuracy of 82.08%, with commendable precision (81.44%) and recall (85.19%). The F1 score remained balanced at 83.27%, while the AUC-ROC value slightly decreased to 90.17%.

- **5-Fold Cross-Validation**: Utilizing 5-fold cross-validation for evaluation yielded performance metrics closely aligned with the baseline and bootstrapping approaches. With an accuracy of 82.14%, the cross-validated logistic regression model demonstrates strong precision (80.88%) and recall (85.23%). The F1 score remains consistent at 83.00%, while the AUC-ROC of 90.22% further validates its discriminatory power.

In summary, both bootstrapping and 5-fold cross-validation techniques provide reliable estimates of logistic regression model performance, consistently reaffirming the efficacy of the baseline model. These resampling methods offer robust assessments of model generalization, guiding decisions in real-world applications.

```
## [1] "Evaluation Metrics for Logistic Regression Models:"
```

```
##                      Model  Accuracy Precision    Recall  F1_Score   AUC_ROC
## 1 Baseline (All Variables) 0.8217378 0.8144723 0.8539646 0.8337510 0.9031499
## 2            Bootstrapping 0.8208002 0.8144883 0.8516942 0.8326538 0.9015355
## 3  5-Fold Cross-Validation 0.8361076 0.8206583 0.8806334 0.8495887 0.9103838
```

Logistic Regression Model Performance on test data with Bootstrapping (Confidence Intervals)

The logistic regression model, evaluated with bootstrapping on the test data, demonstrates an average accuracy of 82.01% (CI: [81.40%, 82.44%]). The average precision is 81.32% (CI: [80.29%, 82.09%]), and the average recall is 85.23% (CI: [84.37%, 86.23%]). The F1 score, a harmonic mean of precision and recall, stands at 83.22% (CI: [82.72%, 83.63%]). Moreover, the average AUC, a measure of the model’s ability to distinguish between positive and negative cases, is 90.14% (CI: [89.71%, 90.63%]). These results indicate the model’s reliability and consistency in predictive performance, as well as its ability to generalize effectively to unseen data, reinforced by bootstrapping techniques.

```
## [1] "Evaluation Metrics of Logistic Regression with bootstrapping on Test data:"
## [1] "Average Accuracy: 0.820692744102717"
## [1] "Confidence Interval (Accuracy): [ 0.813944461033144 ,  0.826813974320693 ]"
## [1] "Average Precision: 0.814366434890997"
## [1] "Confidence Interval (Precision): [ 0.802265714695752 ,  0.823857699006249 ]"
## [1] "Average Recall: 0.851637193382772"
## [1] "Confidence Interval (Recall): [ 0.843397033656589 ,  0.859669138619509 ]"
## [1] "Average F1 Score: 0.832564115778082"
## [1] "Confidence Interval (F1 Score): [ 0.827130913558594 ,  0.837865852537437 ]"
## [1] "Average AUC: 0.901545449476515"
## [1] "Confidence Interval (AUC): [ 0.899203879993766 ,  0.903458374973372 ]"
```

Logistic Regression Model Performance with 5-Fold Cross-Validation on held-out fold and Test Data

The logistic regression model, evaluated using 5-fold cross-validation on the held-out fold, demonstrates an average accuracy of 82.72%, with average precision and recall at 82.23% and 85.77%, respectively. The F1 score, a balanced measure of precision and recall, stands at 83.74%, with an AUC of 90.57%. On the test data, the logistic regression model maintains its effectiveness with an accuracy of 81.91%, achieving a precision of 81.49% and recall of 84.65%, resulting in an F1 score of 82.91% and an AUC of 90.22%. These results indicate the model’s consistency in performance across different subsets of data and its ability to generalize well to unseen data. While logistic regression provides a reasonable baseline for classification tasks, its performance metrics are slightly lower compared to more complex models like Random Forest.

```
## [1] "Evaluation Metrics of Logistic Regression with 5-Fold Cross-Validation on held-out fold:"
```

```
## [1] "Average Accuracy: 0.828747205503714"
```

```
## [1] "Average Precision: 0.822824194725152"
```

```
## [1] "Average Recall: 0.860842456775666"
```

```
## [1] "Average F1 Score: 0.841269222718357"
```

```
## [1] "Average AUC: 0.905608859252164"
```

```
## [1] "Evaluation Metrics of Logistic Regression with 5-Fold Cross-Validation on Test data:"
## [1] "Test Accuracy: 0.822932218572708"
## [1] "Precision: 0.818331503841932"
## [1] "Recall: 0.850541928123217"
## [1] "F1 Score: 0.834125874125874"
## [1] "ROC AUC: 0.902897574798376"
```

Performance Evaluation on test data Using Resampling Techniques - Support Vector Machine

In assessing model performance through bootstrapping and 5-fold cross-validation, we analyze the results of Support Vector Machine (SVM) models:

- **SVM Baseline**: The baseline SVM model achieves an accuracy of 83.67%, with a precision of 84.40% and recall of 81.88%. It demonstrates a balanced F1 score of 83.12% and a strong discriminatory capability with an AUC-ROC value of 91.11%.

- **SVM Bootstrapping**: Employing bootstrapping for evaluation yields results comparable to the baseline SVM model. The bootstrapped SVM model maintains a high accuracy of 83.26% along with commendable precision (83.98%), recall (84.04%), and F1 score (84.01%). The AUC-ROC value of 90.96% indicates robust discrimination, consistent with the baseline performance.

- **SVM 5-Fold Cross-Validation**: Utilizing 5-fold cross-validation for evaluation yields performance metrics closely aligned with the baseline and bootstrapping approaches. With an accuracy of 83.34%, the cross-validated SVM model demonstrates strong precision (84.16%), recall (83.97%), and F1 score (84.06%). The AUC-ROC of 90.96% further validates its discriminatory power.

In summary, both bootstrapping and 5-fold cross-validation techniques provide reliable estimates of SVM model performance, consistently reaffirming the efficacy of the baseline model. These resampling methods offer robust assessments of model generalization, guiding decisions in real-world applications.

```
## [1] "Evaulation Metrics of Support Vector Machine:"
```

Model <chr>	Accuracy <dbl>	Precision <dbl>	Recall <dbl>	F1_Score <dbl>	AUC_ROC <dbl>
SVM Baseline	0.8366677	0.8439850	0.8188450	0.8312249	0.9110980
SVM Bootstrap	0.8332039	0.8411383	0.8400456	0.8405725	0.9094437
SVM 5-fold CV	0.8327859	0.8406625	0.8397034	0.8401826	0.9096104

3 rows

** Support Vector Machine Model Performance on test data with Bootstrapping (Confidence Intervals)**

The bootstrapping process, involving 100 resampled iterations of the training data, provides a robust evaluation framework for our SVM model. The average accuracy of 83.28% underscores its consistent performance, with confidence intervals between 82.80% and 83.90%. Precision, averaging 84.14%, assures us of the model’s ability to accurately identify positive instances, with confidence bounds between 83.31% and 85.02%. Additionally, the model’s average recall of 83.88% indicates its proficiency in capturing most positive cases, with confidence intervals ranging from 83.14% to 84.46%. The harmonic mean of precision and recall, reflected in the F1 score averaging 84.01%, ensures a balanced performance in classification tasks, with confidence bounds between 83.61% and 84.50%. Furthermore, the impressive average AUC of 90.95% with confidence bounds between 90.80% and 91.05% signifies strong discriminatory power, confirming the model’s efficacy in distinguishing between positive and negative instances. These results, evaluated on the test data, validate the SVM model’s reliability and suitability for real-world applications.

```
## [1] "Evaluation Metrics of Support Vector Machine(SVM) with bootstrapping on Test data:"
## [1] "Average Accuracy: 0.833341295909227"
## [1] "Confidence Interval (Accuracy): [ 0.83009853687668 ,  0.83742908330845 ]"
## [1] "Average Precision: 0.841917163348647"
## [1] "Confidence Interval (Precision): [ 0.832844880062394 ,  0.847071519563239 ]"
## [1] "Average Recall: 0.839212778094695"
## [1] "Confidence Interval (Recall): [ 0.833556759840274 ,  0.845849971477467 ]"
## [1] "Average F1 Score: 0.840551286081567"
## [1] "Confidence Interval (F1 Score): [ 0.837126913088314 ,  0.844294143592257 ]"
## [1] "Average AUC: 0.909750073987013"
## [1] "Confidence Interval (AUC): [ 0.907901090075445 ,  0.911420084366642 ]"
```

Support Vector Machine Model Performance with 5-Fold Cross-Validation on held-out fold and Test Data

The Support Vector Machine (SVM) model, evaluated using 5-fold cross-validation on the held-out fold, demonstrates an average accuracy of 83.59%, with average precision and recall at 84.30% and 84.63%, respectively. The F1 score, a balanced measure of precision and recall, stands at 84.47%. On the test data, the SVM model maintains its effectiveness with an accuracy of 83.34%. It achieves a precision of 84.16% and recall of 83.97%, resulting in an F1 score of 84.03% with an impressive average AUC of 90.96%. These results indicate the model's reliability and consistency in predictive performance across different subsets of data. The SVM model shows promising performance in accurately classifying instances into their respective categories, suggesting its suitability for the classification task at hand.

```
## [1] "Evaluation Metrics of Support Vector Machine(SVM) with 5-fold Cross-Validation on held-out
fold:"
## [1] "Average Evaluation Metrics on Validation Data:"
## [1] "Average Accuracy: 0.835851472471191"
## [1] "Average Precision: 0.843202511229604"
## [1] "Average Recall: 0.846088183407292"
## [1] "Average F1 Score: 0.844613252433919"
## [1] "Evaluation Metrics on Test Data:"
## [1] "Test Accuracy: 0.833681696028665"
## [1] "Precision: 0.842105263157895"
## [1] "Recall: 0.83970336565887"
## [1] "F1 Score: 0.840902599257355"
## [1] "ROC AUC for Test Data: 0.909610377912837"
```

Performance Evaluation on test data Using Resampling Techniques - Decision Trees

In evaluating the Decision Tree models through bootstrapping and 5-fold cross-validation, we observe the following:

- **Baseline Model**: The Decision Tree model trained on all variables demonstrates a stable accuracy of 80.95%, with precision and recall at 80.86% and 78.63%, respectively. The F1 score, a balanced measure of precision and recall, stands at 79.73%, with an AUC-ROC of 83.82%.

- **5-Fold Cross-Validation**: The Decision Tree model evaluated using 5-fold cross-validation exhibits an accuracy of 80.11%, achieving a higher precision of 82.53% compared to the baseline. The recall and F1 score are consistent with the baseline, indicating robust performance, with an AUC-ROC of 82.27%.

- **Bootstrapping**: Employing bootstrapping for model evaluation yields results comparable to the baseline. The bootstrapped Decision Tree model maintains an accuracy of 80.52% and precision of 81.83%. Notably, the model's recall improves to 80.83%, resulting in a balanced F1 score of 81.28%. The AUC-ROC of 83.46% indicates robust discriminatory capability.

In summary, both bootstrapping and 5-fold cross-validation techniques provide reliable estimates of model performance, consistently reaffirming the efficacy of the Decision Tree model. These resampling methods offer robust assessments of model generalization and can guide decision-making in deploying the model for real-world applications.

```
##                          Model  Accuracy Precision    Recall  F1_Score   AUC_ROC
## 1      Decision Tree Baseline 0.8094954 0.8086340 0.7863409 0.7973316 0.8381915
## 2     Decision Tree 5-fold CV 0.8080024 0.8107503 0.8260125 0.8183103 0.8378092
## 3 Decision Tree Bootstrapping 0.8050015 0.8200096 0.8053109 0.8120034 0.8336540
```

Decision Tree Model Performance on test data with Bootstrapping (Confidence Intervals)

The Decision Tree model, enhanced with bootstrapping, showcases robust performance across various evaluation metrics. On average, it achieves an accuracy of approximately 80.5%, with a 95% confidence interval ranging from 78.96% to 82.61%. Precision, indicating the proportion of correctly predicted positive cases, averages at 82.10%, with a confidence interval spanning from 78.06% to 86.33%. Recall, representing the model's ability to capture all positive cases, stands at 80.49%, with a confidence interval ranging from 75.93% to 85.48%. The F1 score, a harmonic mean of precision and recall, demonstrates strong performance, averaging at 81.23%, with a confidence interval between 79.41% and 83.00%. Furthermore, the model exhibits impressive discrimination ability, as reflected by an average AUC of 83.51%, with a confidence interval of 81.11% to 87.21%. These results underscore the effectiveness and reliability of the Decision Tree model in predicting term deposit subscriptions when augmented with bootstrapping.

```
## [1] "Evaluation Metrics of Decision Tree Model with bootstrapping are: "
## [1] "Average Accuracy: 0.806906539265452"
## [1] "Confidence Interval (Accuracy): [ 0.789787996416841 ,  0.825537473872798 ]"
## [1] "Average Precision: 0.822068854594891"
## [1] "Confidence Interval (Precision): [ 0.795895435877053 ,  0.868527015437393 ]"
## [1] "Average Recall: 0.806525955504849"
## [1] "Confidence Interval (Recall): [ 0.762364517969196 ,  0.845821448944666 ]"
## [1] "Average F1 Score: 0.813794562959496"
## [1] "Confidence Interval (F1 Score): [ 0.794744525547445 ,  0.831905232755021 ]"
## [1] "Average AUC: 0.83656089560753"
## [1] "Confidence Interval (AUC): [ 0.812055054743247 ,  0.875957537704787 ]"
```

Decision Tree Model Performance with 5-Fold Cross-Validation on held-out fold and Test Data

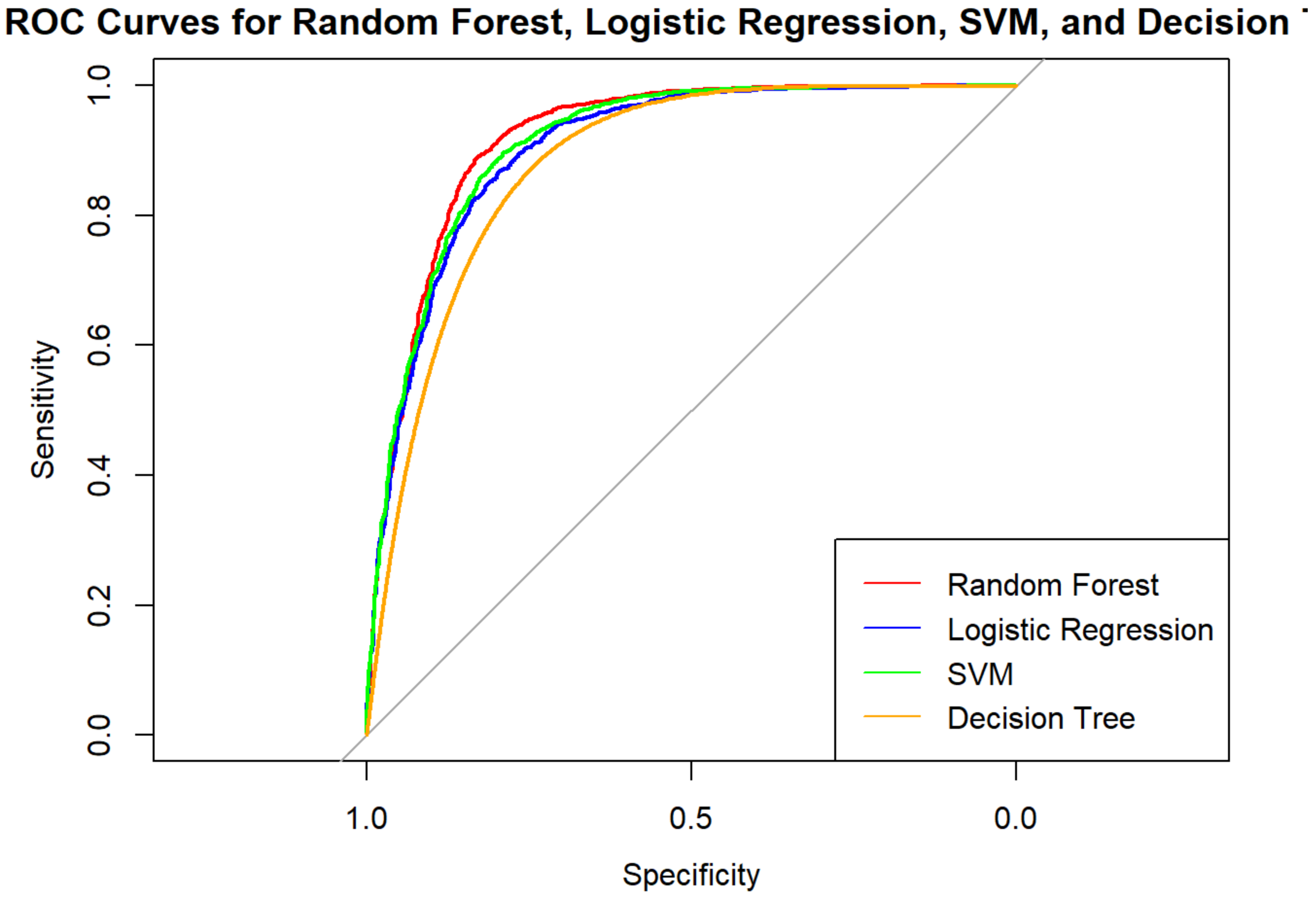
The Decision Tree model demonstrates consistent performance across both the held-out set and test data sets. During 5-fold cross-validation, the model achieved an average accuracy of approximately 81.6%, with an average precision of 82.7% and recall of 82.5%. This suggests its ability to accurately classify term deposit subscriptions while maintaining a high proportion of true positive predictions. On the test data, the model yielded similar results, achieving an accuracy of 80.9%, with precision and recall rates of 81.0% and 83.1%, respectively. The F1 score, a balanced measure of precision and recall, stood at approximately 82.0%, indicating robust performance. Additionally, the ROC AUC score on the test data was calculated at 0.84, signifying strong discrimination ability in distinguishing between positive and negative instances. Overall, these findings underscore the reliability and consistency of the Decision Tree model in predicting term deposit subscriptions.

```
## [1] "Average Evaluation Metrics of Decision Tree (5-Fold Cross-Validation) on the held-out fol
d:"
## [1] "Average Accuracy: 0.811978097866557"
## [1] "Average Precision: 0.826835554591305"
## [1] "Average Recall: 0.814374443930781"
## [1] "Average F1 Score: 0.820128136296045"
## [1] "Evaluation Metrics of Decision Tree (5-Fold Cross-Validation) on Test Data:"
## [1] "Test Accuracy: 0.809495371752762"
## [1] "Precision: 0.810239287701725"
## [1] "Recall: 0.830576155162578"
## [1] "F1 Score: 0.820281690140845"
## [1] "ROC AUC for Test Data: 0.838191456965288"
```

Performance Evaluation of Statistical Learning Methods

ROC Curve Analysis

The ROC curve analysis provides insights into the predictive performance of different models for term deposit subscriptions. Random forest emerges as the top-performing model, exhibiting superior discriminative power with a steep curve, leading to higher true positive rates and lower false positive rates. Following closely, logistic regression demonstrates respectable performance, with the support vector machine (SVM) also showing commendable results, slightly surpassing logistic regression. Despite not achieving the highest ROC AUC among the models, the decision tree still performs impressively, showcasing meaningful discriminative power while offering simplicity and interpretability. Overall, the findings suggest that random forest is the optimal choice for term deposit subscription prediction, with SVM presenting a viable alternative for achieving competitive performance.



Comparative Analysis of Model Performance on Test Data

The comparison of model performance on the test data reveals insightful trends. Random Forest emerges as the top-performing model with an accuracy of 85.76%. It demonstrates strong precision (89.88%) and recall (82.03%), resulting in a balanced F1 score of 85.77%. Logistic Regression follows closely with an accuracy of 82.17%. Although its precision (81.45%) is slightly lower than Random Forest, it compensates with a higher recall (85.40%), leading to a respectable F1 score of 83.38%. Support Vector Machine (SVM) exhibits competitive performance, achieving an accuracy of 83.67%. With precision, recall, and F1 score values of 85.39%, 83.02%, and 84.18% respectively, SVM demonstrates robust predictive capabilities. Decision Tree, while not the top performer in terms of accuracy, still shows promising results with an accuracy of 80.95% and balanced precision, recall, and F1 score values of 81.02%, 83.06%, and 82.03% respectively. Overall, while Random Forest excels in overall accuracy and F1 score, Logistic Regression, SVM, and Decision Tree offer strong alternatives with balanced precision and recall rates.

```
## [1] "Evaluation Metrics on Test Data:"
```

```
##                    Model  Accuracy Precision    Recall  F1_Score
## 1          Random Forest 0.8569722 0.8976280 0.8203080 0.8572280
## 2    Logistic Regression 0.8217378 0.8144723 0.8539646 0.8337510
## 3 Support Vector Machine 0.8366677 0.8538732 0.8300057 0.8417703
## 4          Decision Tree 0.8094954 0.8102393 0.8305762 0.8202817
```

Conclusion : Elevating Marketing Strategies for Optimal Results

Strategic Solutions for Enhanced Marketing Campaigns

In the pursuit of optimizing subscription rates, strategic insights gleaned from demographic profiles, financial behaviors, campaign nuances, contact methodologies, and engagement dynamics unveil pathways to success:

Demographic Insights: Retirees and students have higher subscription rates, while married individuals subscribe less. Tertiary-educated clients are more likely to subscribe.

Financial Status: Clients with no defaults or personal loans are more likely to subscribe. Absence of housing loans correlates with higher subscription rates.

Campaign Specifics: Successful past campaigns increase subscription likelihood. Timing of campaigns, particularly in March, December, and October, affects subscription rates.

Contact Method and Timing: Cellular communication and specific months like March, December, and October influence subscription decisions.

Engagement Dynamics: Longer and more frequent contacts lead to higher subscription rates, emphasizing the importance of consistent outreach.

By integrating these strategic insights into marketing initiatives, financial institutions can orchestrate targeted and impactful campaigns, thereby maximizing subscription rates and nurturing enduring customer relationships.

Scope and Generalizability

The predictive analysis undertaken unveils significant insights into the determinants of customer subscription behavior regarding term deposits. Leveraging advanced statistical learning methods like Random Forest, Logistic Regression, Support Vector Machines and Decision tree, alongside meticulous feature engineering and comprehensive performance evaluation, a deep comprehension of the predictive task and model performance has been achieved. This analysis extends its implications beyond the confines of the dataset, offering insights applicable to a spectrum of marketing strategies and customer engagement initiatives within financial institutions. Through the integration of cross-validation and bootstrapping techniques, the models' performance metrics are reinforced and proven robust across various subsets of data, enhancing their suitability for real-world applications. This ensures the reliability of the identified predictive patterns, which can be extrapolated to similar datasets and contexts. Overall, the analysis provides a solid foundation for optimizing marketing strategies and fostering enduring customer relationships in the financial sector.

Limitations and Possibilities for Improvement

Acknowledging inherent limitations is crucial. The dataset's sampling methodology and biases, including sampling and temporal bias, may impede the generalizability of findings. Relying solely on telecommunication-based marketing data might neglect other impactful factors influencing customer behavior. Despite endeavors to address feature engineering and model selection, unexplored variables or interactions may yet exist, potentially enhancing predictive performance. To surmount these limitations and enrich the analysis, various avenues for improvement can be explored. Integrating additional datasets or information sources, such as transaction data or socioeconomic indicators, could furnish a more comprehensive understanding of customer behavior. Furthermore, delving into ensemble models or deep learning architectures might unveil intricate data patterns, leading to augmented predictive efficacy. In conclusion, while the analysis yields invaluable insights, continuous refinement and innovation are imperative for surmounting limitations and ensuring the real-world applicability of findings. By persistently enhancing methodologies and integrating diverse data sources, financial institutions can refine marketing strategies and cultivate enduring customer relationships with efficacy.