1. **Exploring Differences and correlation in Two Variables**

Between the dependent y variable and Default variable.

The information in Figure 9 provides the results of a Welch Two Sample t-test, which is a statistical test used to determine if there is a significant difference between the means of two groups. The variable being tested is y, and the data is divided into two groups based on whether the client has credit in default or not.

The p-value is very small (7.561e-10 which is less than 0.001), which suggests strong evidence against the null hypothesis that the true difference in means between the two groups is zero. Therefore, we can conclude that the difference in means between the two groups is likely to be real and not just due to chance. The 95% confidence interval also provides an estimate of the range of values within which the true difference in means between the two groups lies with 95% confidence. The interval is (0.03707787, 0.07123692), which suggests that the mean value of y for the "no" group is higher than the "yes" group.

In the Figure 10, the correlation coefficient between two variables (denoted as "default" and "y") was calculated using Pearson's product-moment correlation. The output of the analysis indicates that the correlation coefficient is -0.0224, which suggests a weak negative linear relationship between the two variables.

The t-value of -4.768 and degrees of freedom (df) of 45209 were calculated to test whether the correlation coefficient is significantly different from 0. The p-value of 1.866e-06 indicates that the correlation coefficient is significantly different from 0 at the alpha level of 0.05, meaning that we reject the null hypothesis of no correlation between the two variables. The alternative hypothesis states that the true correlation is not equal to 0, and the 95% confidence interval for the correlation

coefficient is (-0.0316, -0.0132). This means we are 95% confident that the true population correlation lies between these two values.

In summary, this output suggests that there is a weak negative linear relationship between the two variables, and the correlation is statistically significant.

2. Logistic regression

2.1.Likelihood ratio (deviance) test

The below table is the combination of results when fitting model which vary from model M1 to model M11, each pair of models are nested.

To find the optimal model in statistics and machine learning, we need to compare different models and choose the one that performs the best on the data. A common approach to model selection is based on comparing their deviance, which is a measure of the goodness of fit. The deviance is defined as the difference between the log-likelihood of the fitted model and the log-likelihood of the null model. To compare models, we can calculate the deviance of each model and then compute the difference in deviance between the models. This difference is approximately chi-squared distributed, and we can use this distribution to calculate a p-value. If the p-value is less than the significance level, we reject the null hypothesis and choose the more complex model with the smaller deviance. Comparing deviances provides a statistical framework for choosing the best model, but other factors such as interpretability and complexity should also be considered. In the case, the AIC also be considered, it states that the smaller AIC, the better model. Although the model M1 has the lowest AIC, there is not much difference among the other models.

Table 2: Comparisons among models with AIC, Deviance, and df

Name of models	AIC	Deviance	Degrees of freedom
M1	28355	28305.26	24
M2	28528	28500.26	13
M3	28630	28605.82	11
M4	28674	28655.81	8
M5	28677	28661.38	7
M6	29102	29088.46	6
M7	29102	29088.46	6
M8	29102	29088.46	6
M9	29102	29088.46	6
M10	29262	29250.24	5
M11	29981	29972.8	3

The motivation of testing the deviance and likelihood ratio is for testing whether a reduced model is true versus whether the full model is true.

The p-value that is different from the others is the one corresponding to the chi-squared value of 5.57 and degrees of freedom of 1. The p-value calculated for this chi-squared value and degrees of freedom is approximately 0.018, which is greater than the usual significance level of 0.05. This indicates that the null hypothesis cannot be rejected at the 0.05 significance level.

In contrast, for all the other chi-squared values and degrees of freedom, the p-values are much smaller than 0.05, indicating strong evidence against the null hypothesis.

Table 3: Comparisons among models with χ^2 distribution, df, and p-value

States of hypotheses	χ^2 distribution	df	p-value	Conclusion

H_0: Model M1 fits. H_1: Model M2 fits.	195	11	2.71e-35	Accept H_1
H_0: Model M2 fits. H_1: Model M3 fits.	105.56	2	8.73e-24	Accept H_1
H_0: Model M3 fits. H_1: Model M4 fits.	49.99	3	2.23e-10	Accept H_1
H_0: Model M4 fits. H_1: Model M5 fits.	5.57	1	0.018201	Accept H_0
H_0: Model M5 fits. H_1: Model M6 fits.	427.08	1	1.30e-95	Accept H_1
H_0: Model M6 fits. H_1: Model M7 fits.	0	0	NaN	Accept H_1
H_0: Model M7 fits. H_1: Model M8 fits.	0	0	NaN	Accept H_1
H_0: Model M8 fits. H_1: Model M9 fits.	0	0	NaN	Accept H_1
H_0: Model M9 fits. H_1: Model M10 fits.	161.78	1	8.91e-37	Accept H_1
H_0: Model M10 fits. H_1: Model M11 fits.	722.56	2	1.20e-157	Accept H_1

2.2.Linear regression models

The formula of the model M4 is:

$y = \beta 0 + \beta 1$housing-yes + $\beta 2$loan-yes + $\beta 3$contact-telephone + $\beta 4$contact-unknown + $\beta 5$poutcome-other + $\beta 6$poutcome-success + $\beta 7$*poutcome-unknown

where y is the response variable (binary outcome), and the remaining variables are binary indicators for housing ownership (housingyes), personal loan status (loanyes), contact method (telephone and unknown), and outcome of the previous marketing campaign (other, success, and unknown).

The logistic regression model M4 indicates that housing ownership, personal loan status, contact method, and outcome of the previous campaign have a significant association with the binary outcome. The intercept is -1.41, and the coefficients for housingyes, loanyes, contacttelephone, contactunknown, poutcomeother, poutcomesuccess, and poutcomeunknown are -0.67, -0.57, -

0.17, -1.01, 0.30, 2.29, and -0.24, respectively. The residual deviance of 28661 is smaller than the null deviance of 32631, indicating a good fit.

The final model formula is:

$$y = -1.41 - 0.67\text{housingyes} - 0.57\text{loanyes} - 0.17\text{contacttelephone} - 1.01\text{contactunknown} + 0.30\text{poutcomeother} + 2.29\text{poutcomesuccess} - 0.24\text{poutcomeunknown}.$$

2.3.Visualization of model

From the plot in Figure 11, we can see that the predictor variables df$housingyes, df$loanyes, df$contactunknown, df$poutcomeother, and df$poutcomesuccess all have negative coefficients, which suggests that they are associated with a lower probability of the outcome variable df$y being 1 (who say “yes”). The predictor variable df$contacttelephone has a smaller negative coefficient, which suggests that it has a weaker negative association with the outcome variable. The predictor variable df$poutcomeunknown has a positive coefficient, which suggests that it is associated with a higher probability of the outcome variable df$y being 1 (who say “yes”). The intercept has a large negative coefficient, which suggests that the probability of the outcome variable df$y being 1 (who say “yes”) is low overall.

3. Ensemble Methods

3.1.Model 0: A Single Classification Tree

In the Figure 12, the Conditional Inference Tree algorithm was used to classify samples into two classes: 'no' and 'yes'. The model was trained on a dataset containing 22,605 samples and 16 predictors, and evaluated using 10-fold cross-validation with one repetition. The final value of the tuning parameter criterion chosen by the model was 0.8811111, resulting in an accuracy of 90.05%

and a moderate agreement between the predicted and actual classes. These results suggest that the model is effective at classifying samples with a high degree of accuracy, and may be useful for predicting whether a sample belongs to one of the two classes.

ROC curve: The output in the Figure 15 displays the probabilities of each observation belonging to the "no" or "yes" class. These probabilities are then used to plot a Receiver Operating Characteristic (ROC) curve, which summarizes the performance of the model. The Area Under the Curve (AUC) is 0.8918, indicating that the model has a good ability to distinguish between the two classes.

3.2.Model 1: Bagging of Classification Tree

Figure 16 shows the result of applying the Bagged CART algorithm on a dataset with 22605 samples and 16 predictors, where the dependent variable "y" indicates whether the client subscribed to a term deposit or not. The algorithm was evaluated using 10-fold cross-validation repeated once, and the summary of sample sizes shows that each fold had almost the same number of samples. The accuracy of the model is 0.8998, which means that it correctly predicted 89.98% of the cases, while the Kappa statistic is 0.4643, indicating a moderate agreement between the predictions and the actual values. Overall, the Bagged CART algorithm appears to perform reasonably well in predicting the outcome of term deposit subscriptions.

The plot in Figure 17 displays a bar chart of the variables on the x-axis, with their importance scores on the y-axis. The height of each bar represents the importance of the corresponding predictor in the model, based on the mean decrease in accuracy (MDA) metric. In the case, “Duration” has the greater its influence on the predictions made by the model. Followed by two variables “Age” and “Balance”.

ROC curve

The information in Figure 20shows the predicted probabilities of the classes 'no' and 'yes' for 6 observations. The area under the ROC curve for the model is 0.9058, indicating good performance in distinguishing between the two classes.

3.3.Model 2: Random Forest for classification trees

In Figure 21, the Model 2 use a Random Forest algorithm for a classification task, where the goal is to predict if a client has subscribed to a term deposit or not. The algorithm was trained on a dataset containing 22,605 samples, with 16 predictors and 2 classes ('yes' or 'no'). A 10-fold cross-validation was used for resampling. The algorithm was tuned using different values of mtry (the number of variables randomly sampled at each split) and the accuracy metric was used to select the best model, which was achieved with mtry=22 and achieved an accuracy of 0.9048. This means that the Random Forest model correctly predicted the outcome of 90.48% of the test samples.

ROC curve: In the Figure 24, the ROC curve illustrates the AUC of 0.9271 suggests that the Model 2 has good predictive power and is capable of distinguishing between the two classes.

3.4.Model 3: Random Forest with Boosting

In the Figure 25, Cross-validation was used to evaluate the Stochastic Gradient Boosting model's performance with different values of the tuning parameters 'interaction.depth' and 'n.trees' on the given dataset. The final model with n.trees = 150, interaction.depth = 3, shrinkage = 0.1, and n.minobsinnode = 10 achieved an accuracy of 90.69% and kappa value of 0.455. However, it is important to split the data into separate training and testing sets to ensure the model's accuracy and generalizability.

ROC curve: The ROC curve for Model 3 in the Figure 28 shows the true positive rate plotted against the false positive rate for predicting "Yes". With an area under the curve of 0.9207, the model has good discriminatory power.

3.5.Model comparisons

By accuracy

Conclusion, the following table is the combination of all the models in the ensemble methods, with the accuracy score for each training and test dataset in each model respectively.

Based on the accuracy and kappa values for both the training and test data, the Random Forest model for classification tree (Model 2) is the final chosen model. It has the highest accuracy and kappa values on the test data, indicating better performance in predicting the target variable.

Table 4: Model comparisons by accuracy and Kappa points

Model name	The training data	The test data
Model 0: A Single Classification Tree	Accuracy: 0.9048 Kappa: 0.4495	Accuracy: 0.9014 Kappa: 0.4145
Model 1: Bagging of Classification Tree	Accuracy: 0.9985 Kappa: 0.9928	Accuracy: 0.816 Kappa: 0.0095
Model 2: Random Forest for classification tree	Accuracy: 1 Kappa: 1	Accuracy: 0.9059 Kappa: 0.4829
Model 3: Random Forest with Boosting	Accuracy: 0.9132 Kappa: 0.4927	Accuracy: 0.9029 Kappa: 0.4377

By ROC curve

In the Figure 29, the ROC curves show the performance of four different classification models. The Random Forest model had the highest AUC value of 0.9271, indicating better performance in distinguishing between true positives and false positives. The Bagged Trees and Gradient Boosting Machine models also had relatively high AUC values, while the Classification Tree model had the lowest AUC value.

By F-beta score

Based on the F-beta scores provided, Model 2 and Model 1 outperformed the other models. Model 2, which is the Random Forest for classification tree, and Model 1, which is the Bagging of Classification Tree, are both ensemble methods that combine the predictions of multiple base models.

Table 5: Model comparisons by F-beta scores

Model name	F-beta score
Model 0: A Single Classification Tree	0.9607993
Model 1: Bagging of Classification Tree	0.9997499
Model 2: Random Forest for classification tree	1
Model 3: Random Forest with Boosting	0.9666278

1.1. Model evaluation and deployment

1.1.1. Model evaluation

A model accuracy of 0.9516489 (or approximately 95.2%) indicates that the Random Forest model is able to correctly predict the target variable for the new dataset with high accuracy. This means

that the model is able to generalize well to new, unseen data, and is a good indication that the model has learned the underlying patterns in the training data.

1.1.2. Model deployment

The columns in the Figure 30 represent the variables used in the model, and the values in each column represent the feature importance of that variable in the model, measured by mean decrease accuracy and mean decrease Gini.

The mean decrease accuracy tells you how much the accuracy of the model decreases when a particular variable is not included, while the mean decrease Gini measures the total reduction of the Gini impurity that is achieved by a variable when it is included in the model.

Based on the table, some variables have a high mean decrease accuracy and mean decrease Gini, such as 'duration', 'age', and 'housingyes', which suggests that these variables are important in predicting the outcome of the model. On the other hand, some variables have a low mean decrease accuracy and mean decrease Gini, such as 'jobself-employed', 'educationunknown', and 'defaultyes', which suggests that these variables may not be very important in predicting the outcome of the model.

II. Discussion and Conclusion

2.1.Data analysis and variability

Finding

The analysis of age, balance, and duration revealed that the mean age was around 41 years, with some outliers indicated by a moderate degree of variability. The balance data had a high degree of relative variability, with a wide range of balances and a skewed distribution. The duration data had a moderate degree of variability and suspected outliers. Therefore, further investigation is

recommended to understand the source of the variability and skewness in the balance data and to determine the reasons for the suspected outliers in the duration data.

Recommendation

Based on the analysis of age, balance, and duration, it is recommended that the organization takes steps to investigate the sources of variability and skewness in the balance data and the reasons for suspected outliers in the duration data. This investigation can help the organization better understand the factors influencing customer behavior and inform decision-making.

2.2.Logistic Regression Analysis

Findings

Some predictor variables such as housing loan, personal loan, and unknown contact communication type are associated with lower probability of clients subscribing to term deposit. The impact of telephone contact communication is negative but weaker. However, an unknown result of the previous marketing campaign has a positive coefficient, indicating a higher probability of subscription. A large negative coefficient of the intercept suggests a low overall probability of the outcome variable.

Recommendations

It is crucial to take immediate action to improve the predictors with negative coefficients and explore potential strategies to optimize the model's predictive power. Further investigation of the positive association of the client group who has unknown engagemnt in the previous campaign outcome whether they subscribed a term deposit is recommended. In-depth data analysis, including interactions between predictors, can provide profound insights into their relationships. Prioritizing

these actions can ensure the model's accuracy and effectiveness in making informed decisions for the organization.

2.3.Ensemble Method Analysis Approach

Findings

Age is the most significant factor that influences the client's decision to subscribe to the bank's term deposit.

Retired, student, and unemployed clients are more likely to subscribe to the bank's term deposit.

Clients with higher balances in their account are more likely to subscribe to the bank's term deposit.

Clients who have not taken a loan and those who do not have a housing loan are more likely to subscribe to the bank's term deposit.

Clients who were previously contacted via telephone and those who were contacted via an unknown mode of contact are more likely to subscribe to the bank's term deposit.

The months of March, September, October, and December are good months for marketing campaigns as they have the highest impact on clients' subscription decisions.

The duration of the call is a critical factor that influences clients' subscription decisions.

Recommendations

- Focus on older clients and those with higher account balances.
- Develop marketing campaigns targeting retired, student, and unemployed clients.
- Identify and target clients who have not taken loans and those who have a housing loan.

- Develop marketing campaigns that focus on contacting clients via telephone or an unknown mode of contact.
- Focus marketing campaigns during the months of March, September, October, and December.
- Train sales representatives to extend the duration of the call without being intrusive or irritating to clients.
- Develop products and services that cater to the specific needs of retired, student, and unemployed clients.
- Provide customized services to clients with higher balances in their account.

These recommendations can help the bank increase its subscription rates by identifying the key factors that influence clients' decisions to subscribe to the bank's term deposit. By focusing on these factors and developing targeted marketing campaigns and services, the bank can increase its customer base and improve customer satisfaction.

2.4.Conclusion

In conclusion, this study analyzed data related to a Portuguese banking institution's remote direct marketing campaigns and identified several factors that influence clients' decisions to subscribe to the bank's term deposit. Logistic regression and ensemble method analyses were performed, revealing age, account balance, employment status, loan history, contact mode, and duration of call as significant factors. Based on these findings, recommendations were provided to increase subscription rates through targeted marketing campaigns and services.

However, further investigation is required to understand the source of variability and skewness in balance data and reasons for suspected outliers in duration data. Additionally, deep neural network

models can be deployed to explore more information and optimize the model's accuracy and effectiveness in decision-making.

This study highlights the importance of data analysis in understanding customer behavior and making informed decisions. With the increasing availability of data, businesses can leverage data analysis techniques to identify key factors that influence customer behavior and develop targeted marketing campaigns and services to improve their business outcomes.

For limitations, the use of the dataset since 2012 and the age of the respondents being 10 years older can limit the accuracy and generalizability of the above findings. Outdated data, changing demographics, cohort effects, biases, and limited scope can all affect the validity of the results. The dataset may not reflect current trends, behaviors, or attitudes of the population, and the age difference between the respondents in the dataset and the current population could lead to cohort effects that are not captured in the study. Furthermore, the dataset may be biased towards certain groups or demographics, which could affect the study's results.

Appendix

```
        Welch Two Sample t-test

data:  df$y by df$default
t = 6.2235, df = 866.81, p-value = 7.561e-10
alternative hypothesis: true difference in means between group no and group yes is not equal to 0
95 percent confidence interval:
 0.03707787 0.07123692
sample estimates:
 mean in group no mean in group yes
         1.117961          1.063804
```

Figure 9: T-test between the dependent y variable and Default variable

```
        Pearson's product-moment correlation

data:  df$default and df$y
t = -4.768, df = 45209, p-value = 1.866e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03163025 -0.01320388
sample estimates:
        cor
-0.02241897
```

Figure 10: Pearson Correlation testing between the dependent y variable and Default variable

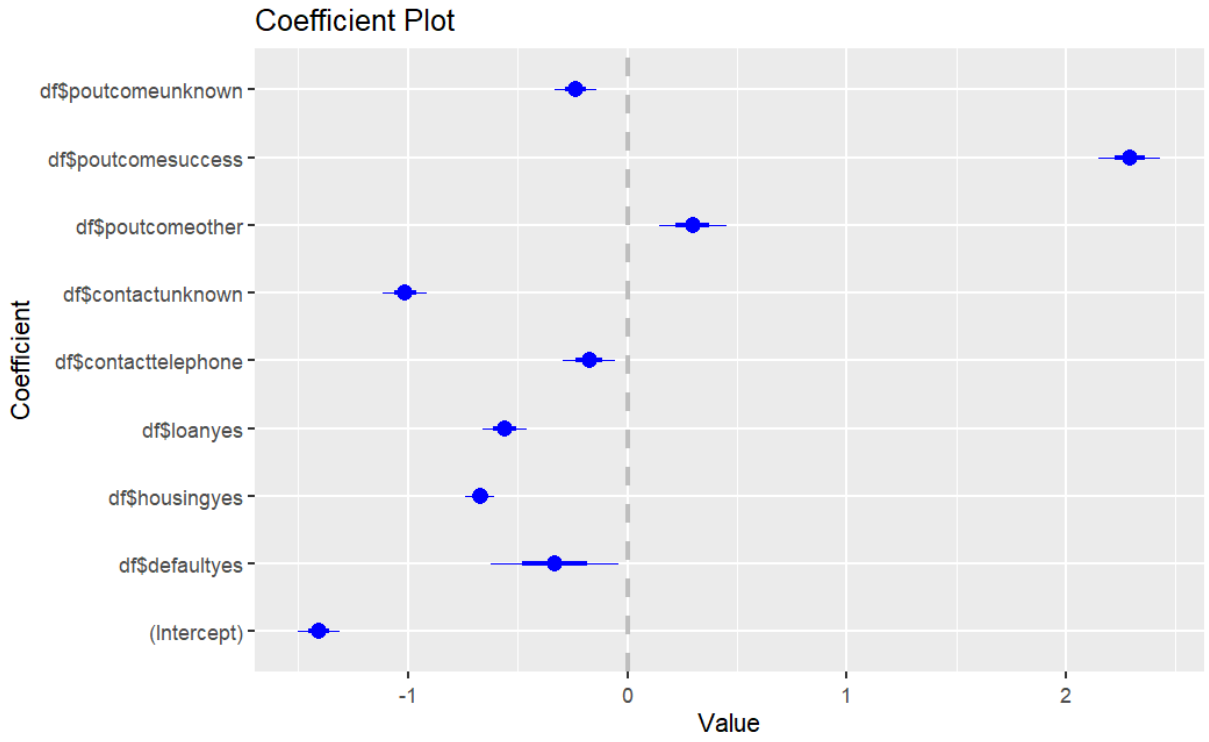

Figure 11: The coefficient plot for logistic regression model

```
Conditional Inference Tree

22605 samples
   16 predictor
    2 classes: 'no', 'yes'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 20344, 20345, 20344, 20345, 20343, 20345, ...
Resampling results across tuning parameters:

  mincriterion  Accuracy   Kappa
  0.0100000     0.8992703  0.4635379
  0.1188889     0.8997129  0.4609373
  0.2277778     0.8995801  0.4657933
  0.3366667     0.9001550  0.4616572
  0.4455556     0.8996683  0.4548385
  0.5544444     0.8995356  0.4441280
  0.6633333     0.8996682  0.4369598
  0.7722222     0.8997124  0.4313896
  0.8811111     0.9004639  0.4319240
  0.9900000     0.8999777  0.4322156

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mincriterion = 0.8811111.
```

Figure 12: Conditional Inference Tree of model 0

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  19382   540
       yes  1613  1070

               Accuracy : 0.9048
                 95% CI : (0.9009, 0.9086)
    No Information Rate : 0.9288
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4495

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9232
            Specificity : 0.6646
         Pos Pred Value : 0.9729
         Neg Pred Value : 0.3988
             Prevalence : 0.9288
         Detection Rate : 0.8574
   Detection Prevalence : 0.8813
      Balanced Accuracy : 0.7939

       'Positive' Class : no
```

Figure 13: Confusion matrix and statistics for the training data of model 0

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  19408   592
       yes  1637   969

               Accuracy : 0.9014
                 95% CI : (0.8974, 0.9053)
    No Information Rate : 0.9309
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4145

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9222
            Specificity : 0.6208
         Pos Pred Value : 0.9704
         Neg Pred Value : 0.3718
             Prevalence : 0.9309
         Detection Rate : 0.8585
   Detection Prevalence : 0.8847
      Balanced Accuracy : 0.7715

       'Positive' Class : no
```

Figure 14: Confusion matrix and statistics for the test data of model 0

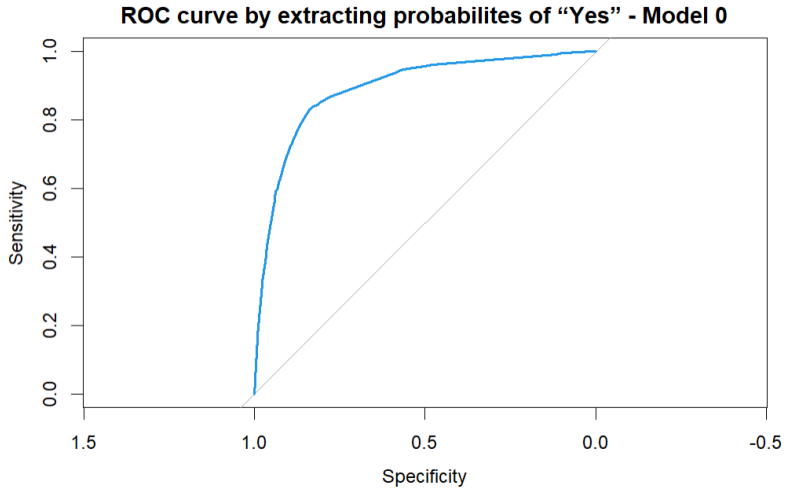

Figure 15: ROC curve of model 0

```
Bagged CART

22605 samples
   16 predictor
    2 classes: 'no', 'yes'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 20345, 20344, 20345, 20345, 20345, 20345, ...
Resampling results:

  Accuracy   Kappa
  0.8998006  0.4642924
```

Figure 16: Bagged CART for model 1

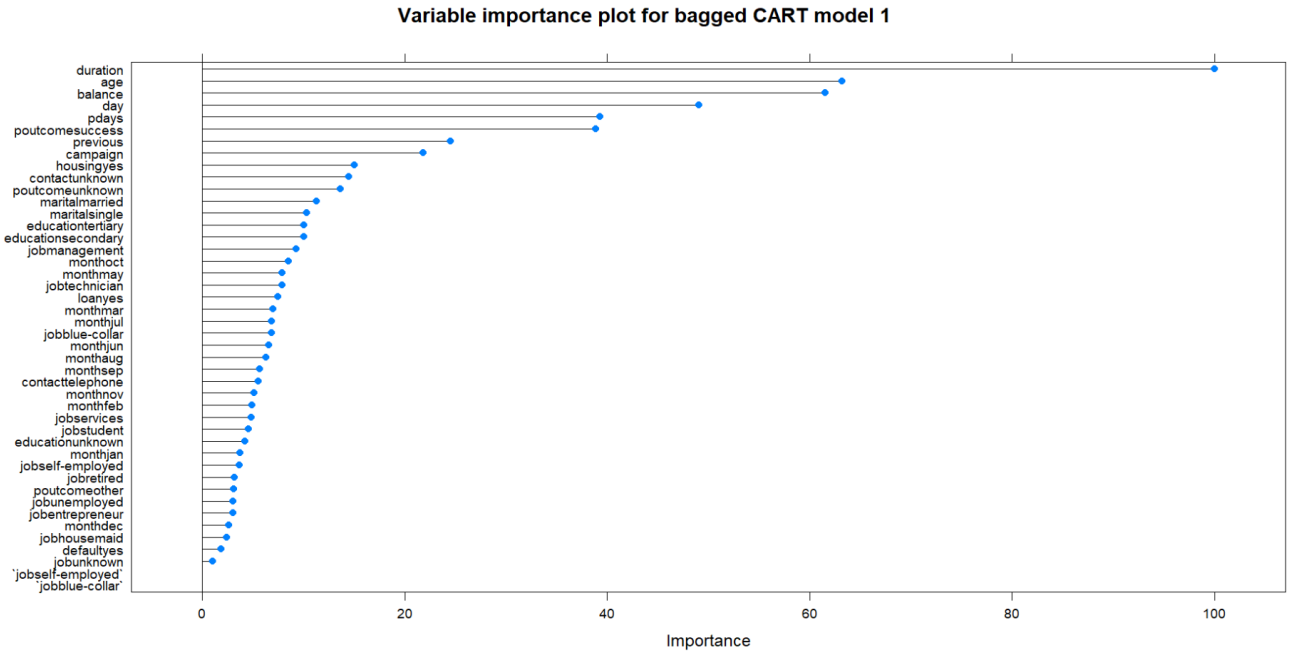

Figure 17: Variable importance plot for bagged CART model 1

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  19916     6
       yes    28  2655

               Accuracy : 0.9985
                 95% CI : (0.9979, 0.999)
    No Information Rate : 0.8823
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9928

 Mcnemar's Test P-Value : 0.0003164

            Sensitivity : 0.9986
            Specificity : 0.9977
         Pos Pred Value : 0.9997
         Neg Pred Value : 0.9896
             Prevalence : 0.8823
         Detection Rate : 0.8810
   Detection Prevalence : 0.8813
      Balanced Accuracy : 0.9982

       'Positive' Class : no
```

Figure 18: Confusion matrix and statistics for the training data of model 1

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  18190  1732
       yes  2427   256

               Accuracy : 0.816
                 95% CI : (0.8109, 0.821)
    No Information Rate : 0.9121
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0095

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.88228
            Specificity : 0.12877
         Pos Pred Value : 0.91306
         Neg Pred Value : 0.09542
             Prevalence : 0.91205
         Detection Rate : 0.80469
   Detection Prevalence : 0.88131
      Balanced Accuracy : 0.50553

       'Positive' Class : no
```

Figure 19: Confusion matrix and statistics for the test data of model 1

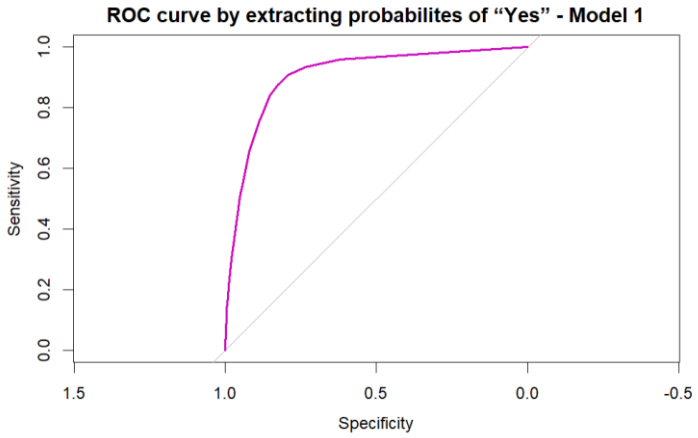

Figure 20: ROC curve of model 1

```
Random Forest

22605 samples
   16 predictor
    2 classes: 'no', 'yes'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 20345, 20345, 20344, 20344, 20345, 20345, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.8892723  0.1422070
  22    0.9048439  0.4809699
  42    0.9044461  0.4851000

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 22.
```

Figure 21: Random Forest for classification trees

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  19922     0
       yes     0  2683

               Accuracy : 1
                 95% CI : (0.9998, 1)
    No Information Rate : 0.8813
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.8813
         Detection Rate : 0.8813
   Detection Prevalence : 0.8813
      Balanced Accuracy : 1.0000

       'Positive' Class : no
```

Figure 22: Confusion matrix and statistics for the training data of Model 2

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  19258   742
       yes  1386  1220

               Accuracy : 0.9059
                 95% CI : (0.902, 0.9096)
    No Information Rate : 0.9132
    P-Value [Acc > NIR] : 0.9999

                  Kappa : 0.4829

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9329
            Specificity : 0.6218
         Pos Pred Value : 0.9629
         Neg Pred Value : 0.4682
             Prevalence : 0.9132
         Detection Rate : 0.8519
   Detection Prevalence : 0.8847
      Balanced Accuracy : 0.7773

       'Positive' Class : no
```

Figure 23: Confusion matrix and statistics for the test data of Model 2

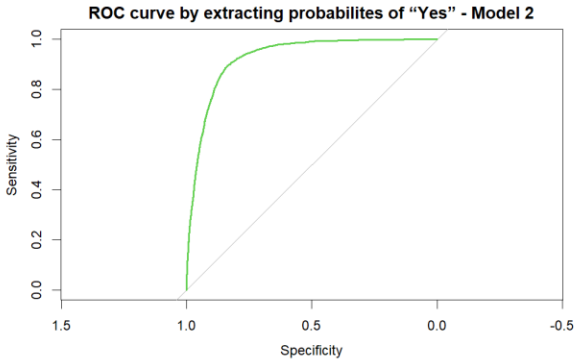

Figure 24: ROC curve of model 2

```
Stochastic Gradient Boosting

22605 samples
   16 predictor
    2 classes: 'no', 'yes'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 20345, 20344, 20345, 20345, 20345, 20344, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  Accuracy   Kappa
  1                   50      0.8945366  0.2153765
  1                  100      0.9007746  0.3539289
  1                  150      0.9036943  0.4001394
  2                   50      0.9033402  0.3921302
  2                  100      0.9049771  0.4255749
  2                  150      0.9060830  0.4429355
  3                   50      0.9043575  0.4113243
  3                  100      0.9060387  0.4424123
  3                  150      0.9069235  0.4552792

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter
 'n.minobsinnode' was held constant at a value of 10
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 150, interaction.depth = 3, shrinkage = 0.1
 and n.minobsinnode = 10.
```

Figure 25: Model 3 – Random Forest with Boosting

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  19500   490
       yes  1473  1142

               Accuracy : 0.9132
                 95% CI : (0.9094, 0.9168)
    No Information Rate : 0.9278
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4927

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9298
            Specificity : 0.6998
         Pos Pred Value : 0.9755
         Neg Pred Value : 0.4367
             Prevalence : 0.9278
         Detection Rate : 0.8626
   Detection Prevalence : 0.8843
      Balanced Accuracy : 0.8148

       'Positive' Class : no
```

Figure 26: Confusion matrix and statistics for the training data of model 3

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  19368   564
       yes  1630  1044

               Accuracy : 0.9029
                 95% CI : (0.899, 0.9068)
    No Information Rate : 0.9289
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4377

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9224
            Specificity : 0.6493
         Pos Pred Value : 0.9717
         Neg Pred Value : 0.3904
             Prevalence : 0.9289
         Detection Rate : 0.8568
   Detection Prevalence : 0.8817
      Balanced Accuracy : 0.7858

       'Positive' Class : no
```

Figure 27: Confusion matrix and statistics for the test data of model 3

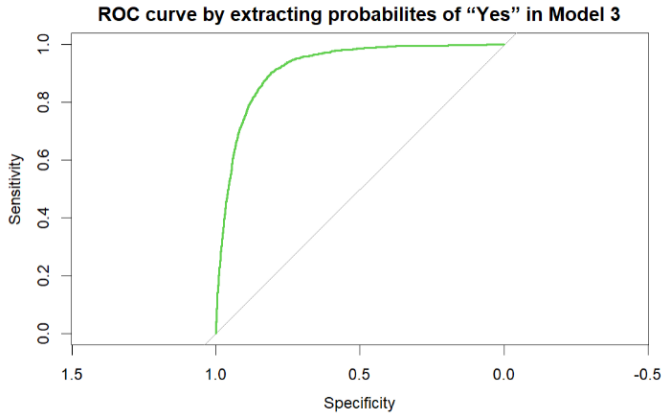

Figure 28: ROC curve of model 3

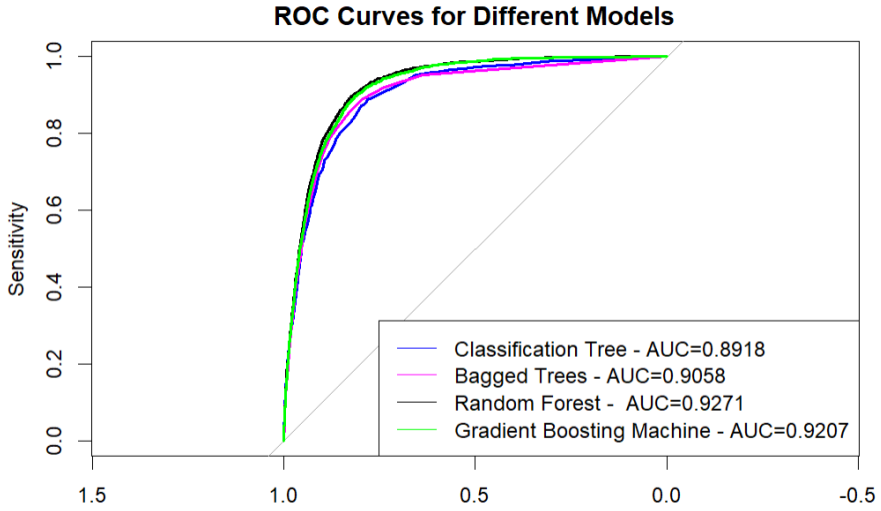

Figure 29: Model comparisons by ROC curve

	no	yes	MeanDecreaseAccuracy	MeanDecreaseGini
age	57.4643171	14.3805835	60.9400225	437.603130
jobblue-collar	2.1590730	5.3399611	4.8024687	41.357758
jobentrepreneur	0.9561197	8.6095768	6.6522698	15.461197
jobhousemaid	6.2351247	-1.0186021	4.9481362	14.158164
jobmanagement	4.9948924	0.1945127	4.6399608	46.300194
jobretired	8.8388366	-7.1850619	3.7395259	16.329241
jobself-employed	0.8677612	0.1882197	0.8254282	22.933455
jobservices	3.9125416	-0.6556188	3.0161092	29.702964
jobstudent	14.0250402	-3.4687190	12.2818048	20.218864
jobtechnician	4.5874391	1.7075598	4.8903479	49.063615
jobunemployed	-3.1539500	-4.7800999	-5.1139283	17.390388
jobunknown	7.1999287	-0.4106658	6.5935910	4.901365
maritalmarried	13.6961417	7.0685543	16.7651088	54.148990
maritalsingle	15.6615687	2.3616914	16.3998588	41.880275
educationsecondary	2.0301692	0.7593623	2.1502013	47.586469
educationtertiary	5.4380951	4.7285944	7.4257131	48.859204
educationunknown	-2.2667042	-1.6036641	-2.8314233	19.825643
defaultyes	-1.7277749	-0.5999884	-1.8431794	6.011791
balance	7.4376159	7.7367782	10.8438501	488.901121
housingyes	43.7278286	18.7427464	51.7002616	97.272955
loanyes	1.6827743	13.0239017	9.6983285	39.093208
contacttelephone	8.7120714	-1.9270935	6.5296989	25.943929
contactunknown	43.3743576	13.1175379	46.0442373	73.385751
day	77.5518317	2.0813475	76.0544884	378.286435
monthaug	33.3683376	-4.8883660	32.7890299	40.452975
monthdec	27.9727512	14.7366511	30.6042324	20.807449
monthfeb	49.6987120	14.1695779	50.4381078	40.119316
monthjan	22.8520689	-2.2723152	22.2181555	22.491094
monthjul	26.3812946	2.4553363	26.9898152	39.268820
monthjun	29.3276528	-4.9330181	29.0593355	53.492390
monthmar	55.2630090	59.5628579	75.2686888	79.236717
monthmay	15.6460010	2.5116598	15.8046682	43.398304
monthnov	24.8109156	-0.8011573	24.5794174	39.359421
monthoct	54.7524901	29.4627642	60.2561456	47.730843
monthsep	38.2915583	24.6881098	44.6641601	40.513669
duration	131.5360628	240.2395750	244.0270887	1317.810393
campaign	24.4273883	8.2368518	25.4530558	161.683846
pdays	27.0440439	33.7084972	37.6041761	211.155577
previous	10.5994318	8.5373831	11.2870316	70.506280
poutcomeother	4.3706132	3.6532817	5.0189378	14.720475
poutcomesuccess	16.0055577	104.7213231	45.9140678	310.328054
poutcomeunknown	11.3873657	7.7724754	11.5716347	28.327141

Figure 30: Deployment of Random Forest Model – Model 3