BANK MARKETING DATA MINING

Moro et al., 2014

CONTENTS

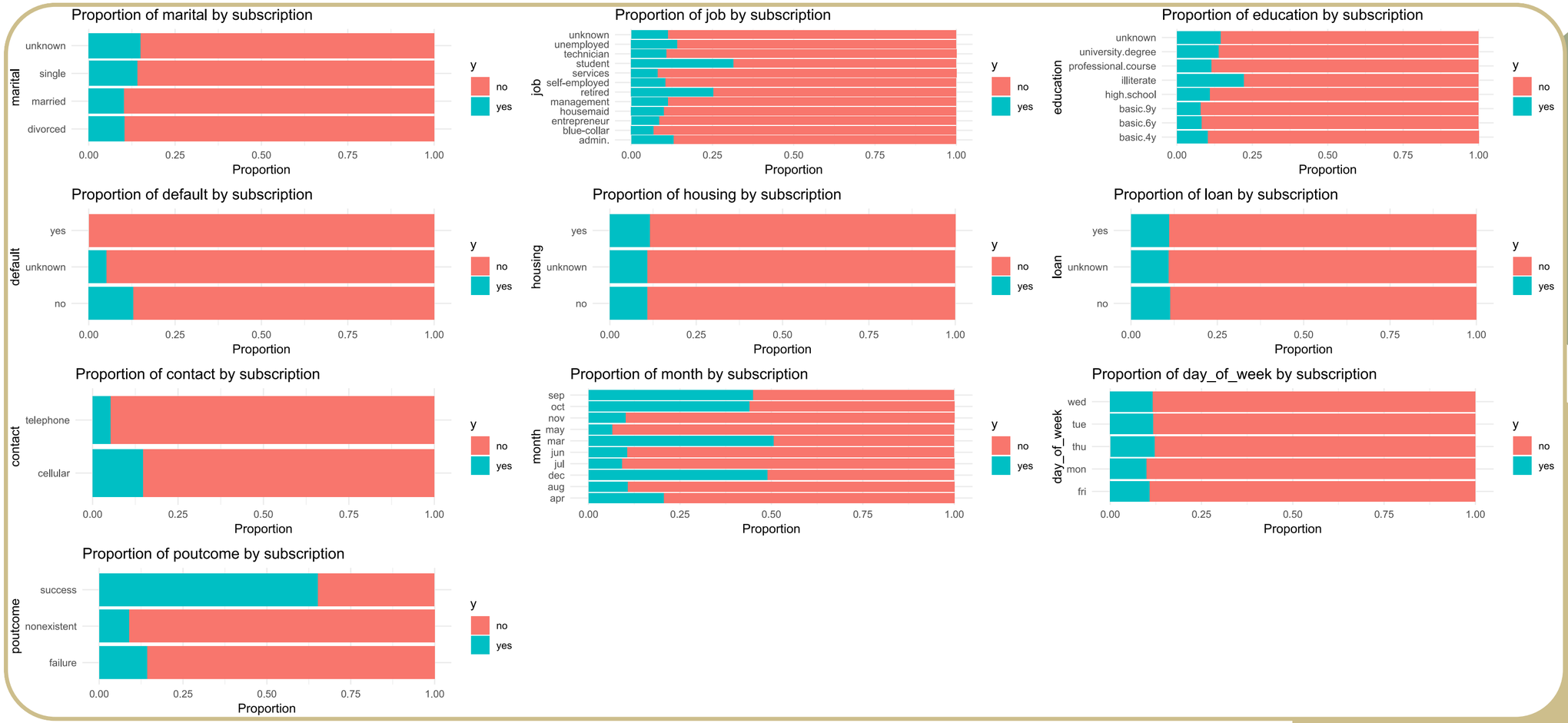
INTRODUCTION

- Two datasets were provided, one contains only 10% of the other one's data.
- The data pertains to a Portuguese bank's direct marketing initiatives that were based on phone calls.
- There are 21 attributes, 10 are numerical and 11 categorial.
- The classification goal is to predict the client's probability of subscribing (yes/no) to a term deposit (the 21st attribute: y).

```
> str(bank_full)
'data.frame':   41188 obs. of  21 variables:
 $ age           : int  56 57 37 40 56 45 59 41 24 25 ...
 $ job           : chr  "housemaid" "services" "services" "admin." ...
 $ marital       : chr  "married" "married" "married" "married" ...
 $ education     : chr  "basic.4y" "high.school" "high.school" "basic.6y" ...
 $ default       : chr  "no" "unknown" "no" "no" ...
 $ housing       : chr  "no" "no" "yes" "no" ...
 $ loan          : chr  "no" "no" "no" "no" ...
 $ contact       : chr  "telephone" "telephone" "telephone" "telephone" ...
 $ month         : chr  "may" "may" "may" "may" ...
 $ day_of_week   : chr  "mon" "mon" "mon" "mon" ...
 $ duration      : int  261 149 226 151 307 198 139 217 380 50 ...
 $ campaign      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays         : int  999 999 999 999 999 999 999 999 999 999 ...
 $ previous      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome      : chr  "nonexistent" "nonexistent" "nonexistent" "nonexistent" ...
 $ emp.var.rate  : num  1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
 $ cons.price.idx: num  94 94 94 94 94 ...
 $ cons.conf.idx : num  -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...
 $ euribor3m     : num  4.86 4.86 4.86 4.86 4.86 ...
 $ nr.employed   : num  5191 5191 5191 5191 5191 ...
 $ y             : chr  "no" "no" "no" "no" ...
```

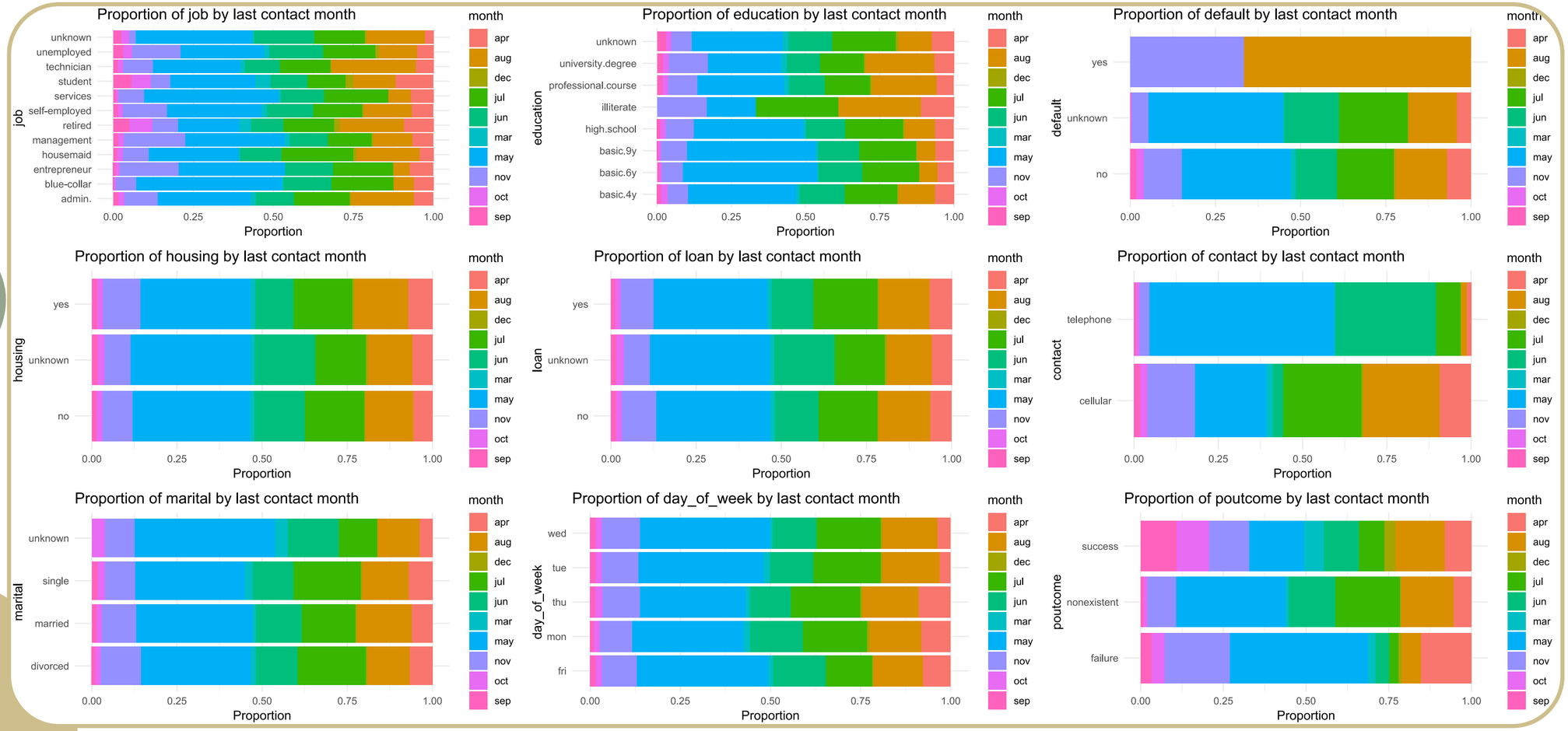
EXPLORATORY DATA ANALYSIS

– RELATIONSHIPS BETWEEN THE CATEGORICAL VARIABLES & THE TARGET VARIABLE Y



References: OAIC n.d.

EXPLORATORY DATA ANALYSIS
– EXAMPLE RELATIONSHIP BETWEEN CATEGORICAL VARIABLES

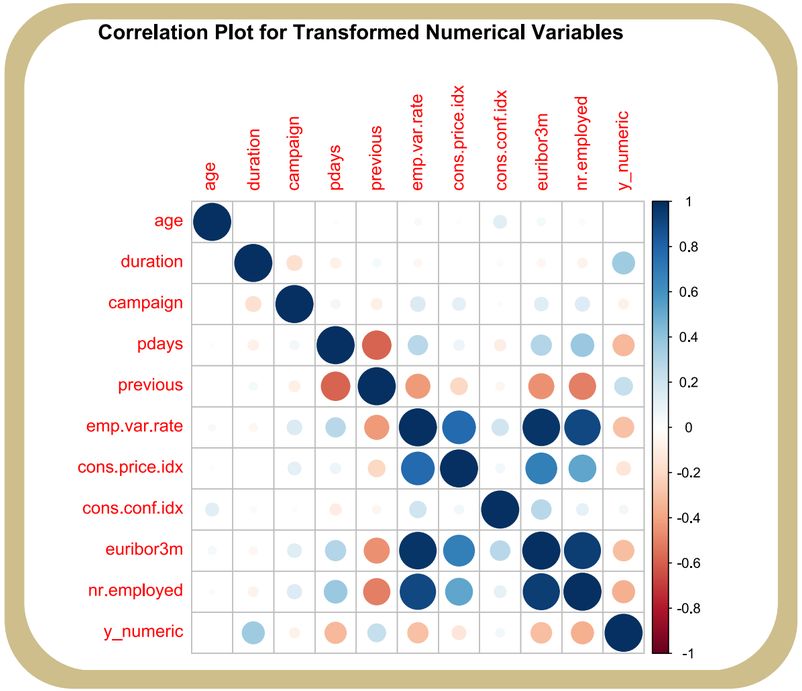


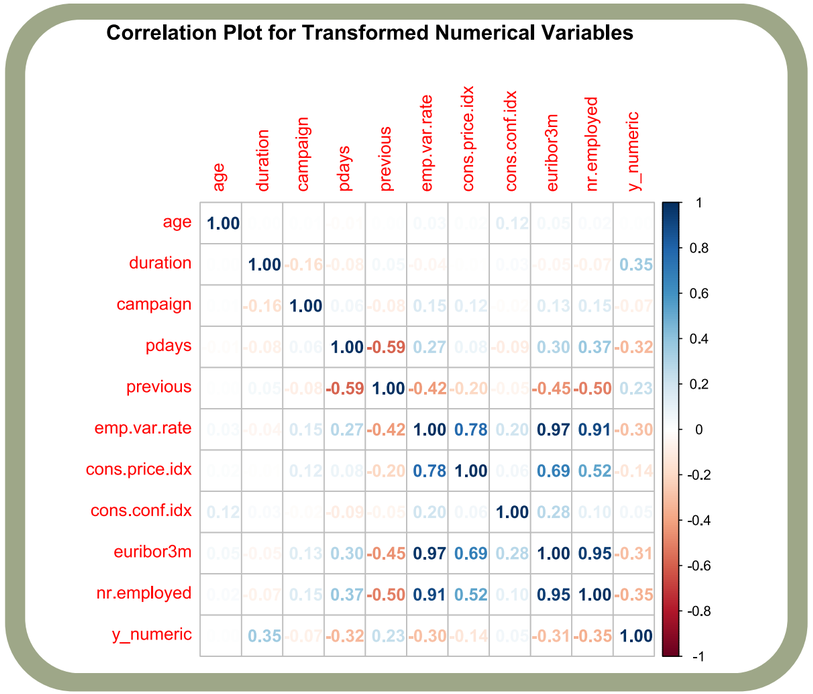
References: Westpac n.d.; Intelligent Banking Solutions n.d.

EXPLORATORY DATA ANALYSIS - CORRELATION BETWEEN NUMERICAL VARIABLES AND TARGET VARIABLE Y

- Numerical attributes that have little to no correlations with the target variable y:
  1. Age (0.00)
  2. Campaign (0.07)
  3. Consumer Confidence Index (0.05)

- Numerical attributes with very high/low correlations:
  1. Euribor 3-Month Rate and Employment Variation Rate (0.97)
  2. Euribor 3-Month Rate and Number of Employees (0.95)
  3. Employment Variation Rate and Number of Employees (0.91)





EXPLORATORY DATA ANALYSIS
- FACTORS AFFECTING THE ATTRIBUTES' RELATIONSHIPS

Demographics

Variables like age, marital status, and job category represent different life stages and responsibilities.

Financial Stability

Higher levels of education may correlate with a better understanding of financial products.

Risk Tolerance

Risk tolerance can influence a client's decision to invest in a term deposit. Clients with more stable jobs or higher education might have a higher risk tolerance.

Economic Conditions

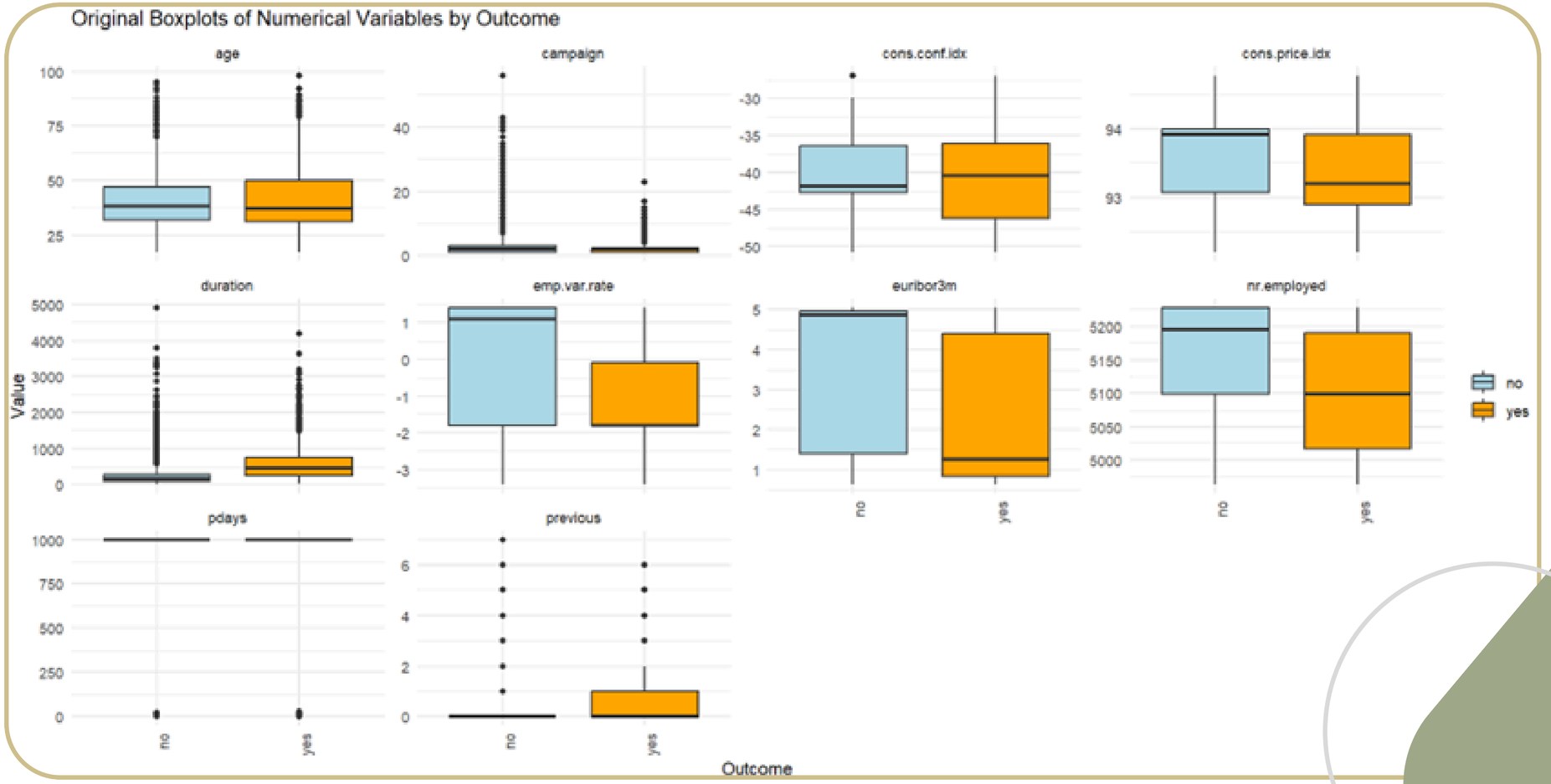
Interest rates (reflected in variables like euribor3m) and employment rates can affect the attractiveness of term deposits

Marketing Effectiveness

The number of contacts in a campaign can have a diminishing return effect; too many contacts may lead to annoyance and a lower likelihood of subscription.

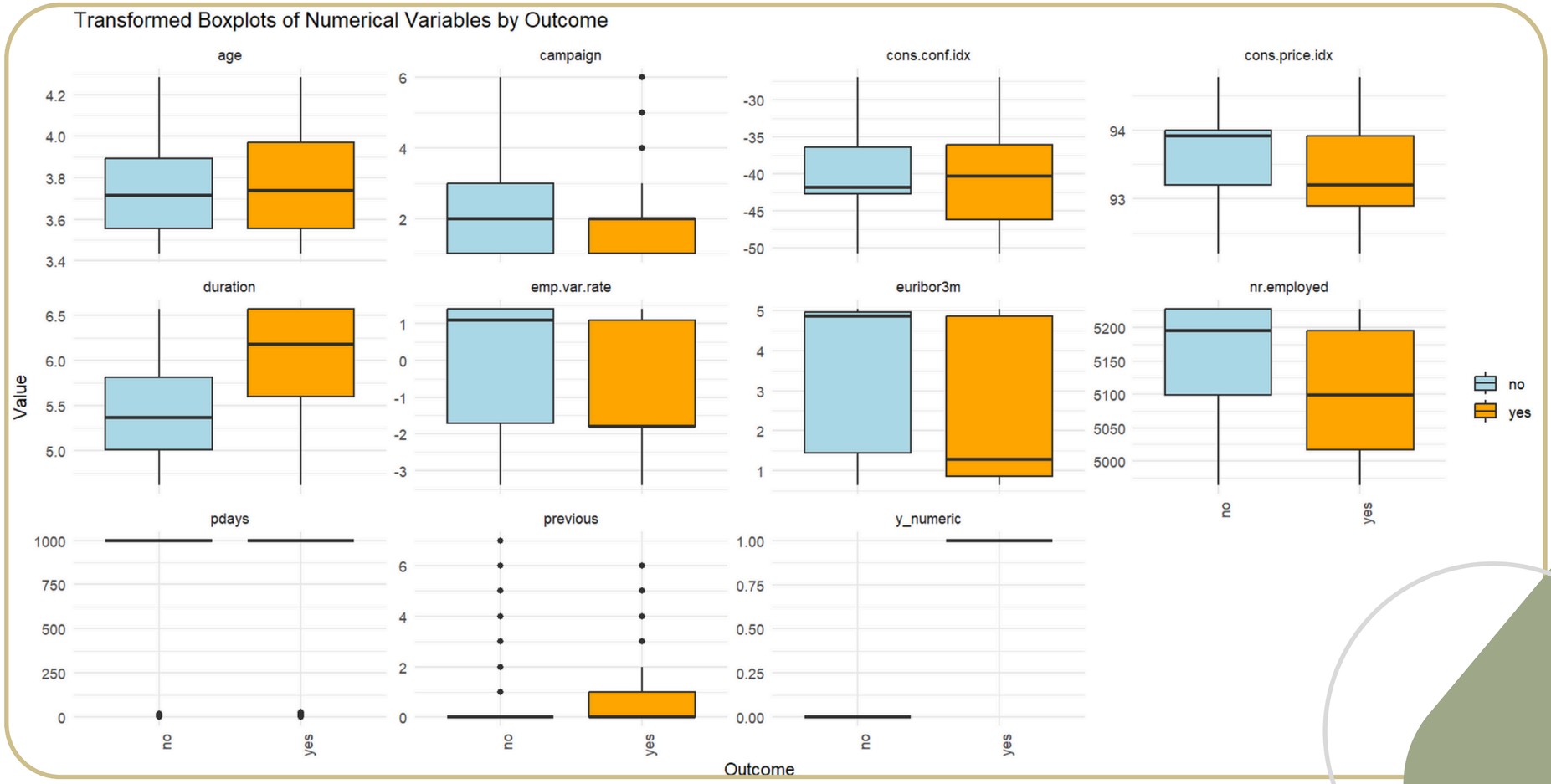
References: Chen 2020; Sembiring & Leon 2021; Xu et al. 2022

EXPLORATORY DATA ANALYSIS
– OUTLIERS BEFORE LOGARITHMIC TRANSFORMATION



References: Bharathi 2020; Toolify.ai 2024

EXPLORATORY DATA ANALYSIS
- OUTLIERS AFTER LOGARITHMIC TRANSFORMATION

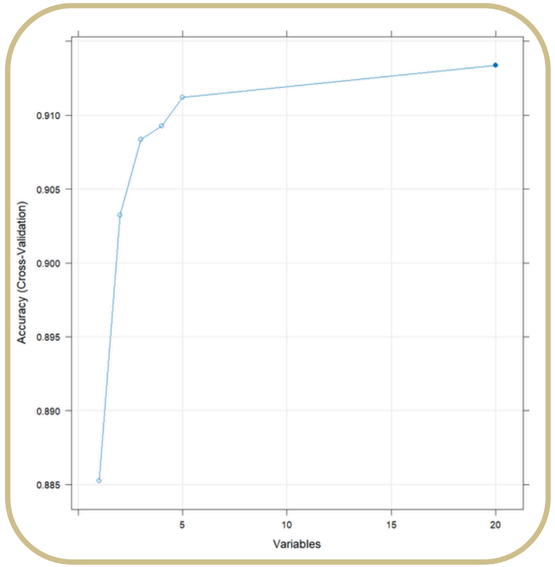


References: Bharathi 2020; Toolify.ai 2024

METHODS: ATTRIBUTE SELECTION

- RFE method
  It iteratively trains model
  Remove the least important features
  Select the optimal feature subsets
- Number of variables increases within 5 - accuracy significantly improves
  Number of variables increases over 20- accuracy slows down
  It may indicate that the increase in attributes has less significant impact on the model improvement
- The top 5 attributes are selected: **duration, euribor3m, nr.employed, pdays, month**

```
> print(rfe_results)

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD KappaSD Selected
         1   0.8853 0.2268   0.005448 0.04077
         2   0.9032 0.4613   0.007837 0.05342
         3   0.9084 0.4702   0.008336 0.04533
         4   0.9093 0.5130   0.007427 0.02948
         5   0.9112 0.5145   0.007480 0.03803
        20   0.9134 0.5383   0.005661 0.02507        *

The top 5 variables (out of 20):
   duration, euribor3m, nr.employed, pdays, month
```



METHODS: DATA PREPROCESSING

- 80% of training sets, 20% of testing sets
- **One-Hot Encoding Method**

  Split the categorical variable 'month' into different categories
- **Feature Scaling Method**

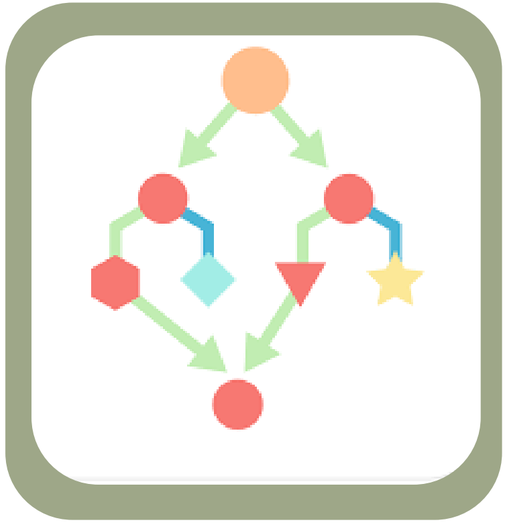
  Standardizing the features

```
> print(str(train_data_scaled))
'data.frame':   32951 obs. of  15 variables:
 $ duration    : num  0.436 -0.173 0.28 -0.158 0.613 ...
 $ euribor3m   : num  0.716 0.716 0.716 0.716 0.716 ...
 $ nr.employed: num  0.335 0.335 0.335 0.335 0.335 ...
 $ pdays       : num  0.197 0.197 0.197 0.197 0.197 ...
 $ monthapr    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ monthaug    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ monthdec    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ monthjul    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ monthjun    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ monthmar    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ monthmay    : num  1 1 1 1 1 1 1 1 1 1 ...
 $ monthnov    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ monthoct    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ monthsep    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ y           : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
NULL
```

References: Data Tricks 2019

METHODS: CLASSIFICATION MODELS

Logistic Regression Model



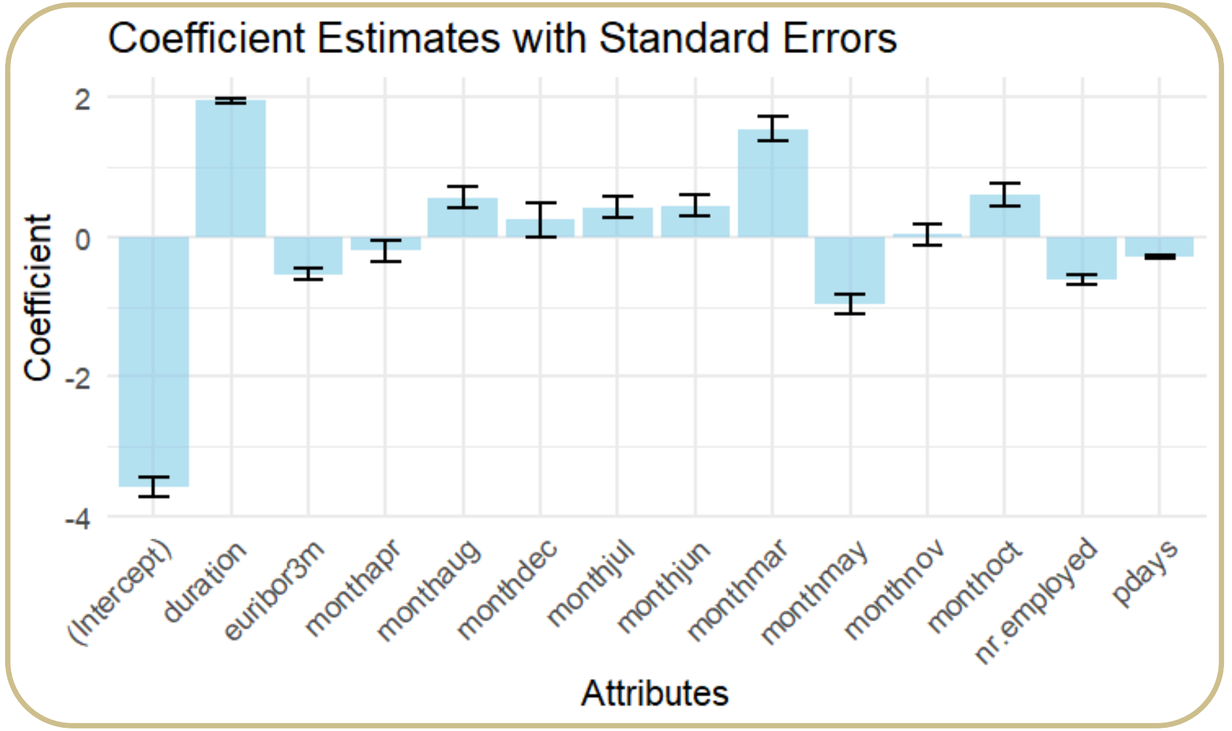
Random Forest Model

Hyperparameter Tuning

RESULTS: LOGISTIC REGRESSION MODEL

- the impact of independent variables on ‘y’

- Select the attributes of certain months to see if they have a greater influences on the response variable ‘y’
- The result shows that the selected attributes of month does not impact the response variable much

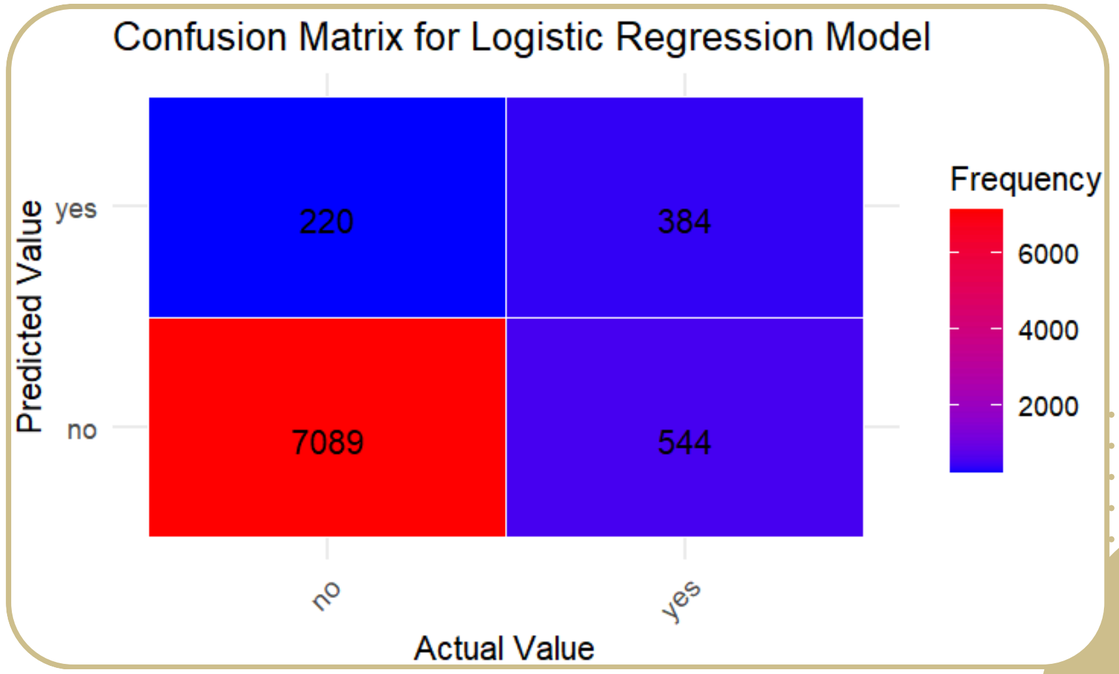


RESULTS: LOGISTIC REGRESSION MODEL

-Imbalanced performance due to the imbalanced data

- The model performs well in predicting 'No' but poorly in predicting 'Yes', which is due to the relatively low number of 'Yes' samples, leading the model to learn more from the 'No' samples.

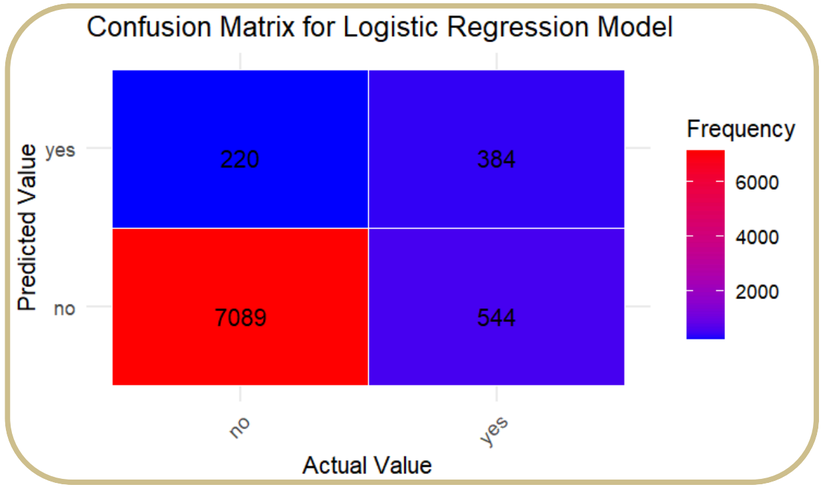
```
               Accuracy : 0.9072
                 95% CI : (0.9008, 0.9134)
    No Information Rate : 0.8873
    P-Value [Acc > NIR] : 2.435e-09

                  Kappa : 0.4527

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9699
            Specificity : 0.4138
         Pos Pred Value : 0.9287
         Neg Pred Value : 0.6358
             Prevalence : 0.8873
         Detection Rate : 0.8606
   Detection Prevalence : 0.9267
      Balanced Accuracy : 0.6918

       'Positive' Class : no
```

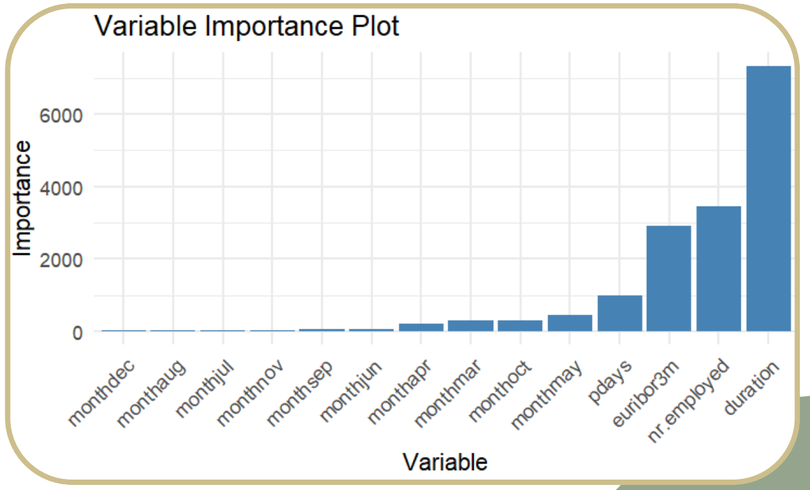


RESULTS: RANDOM FOREST MODEL

- Based on the result of Logistic Regression Model, using oversampling method to balance the dataset
- The result shows that the performance for 'yes' samples has been greatly improved

```
               Accuracy : 0.8391
                 95% CI : (0.831, 0.847
    No Information Rate : 0.8873
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4899

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

            Sensitivity : 0.8254
            Specificity : 0.9472
         Pos Pred Value : 0.9919
         Neg Pred Value : 0.4079
             Prevalence : 0.8873
         Detection Rate : 0.7324
   Detection Prevalence : 0.7384
      Balanced Accuracy : 0.8863

       'Positive' Class : no
```





References: Bobbitt 2020

Results: Hyperparameters tuning

Logistic regression model

Factor: alpha(α), lambda (λ)

Random Forest model

Factor: mtry, ntree, nodesize

Results: Logistic Regression

Alpha	Lambda	Accuracy
0	0.001	0.9085
0.5	0.001	0.9105
1	0.001	0.9103

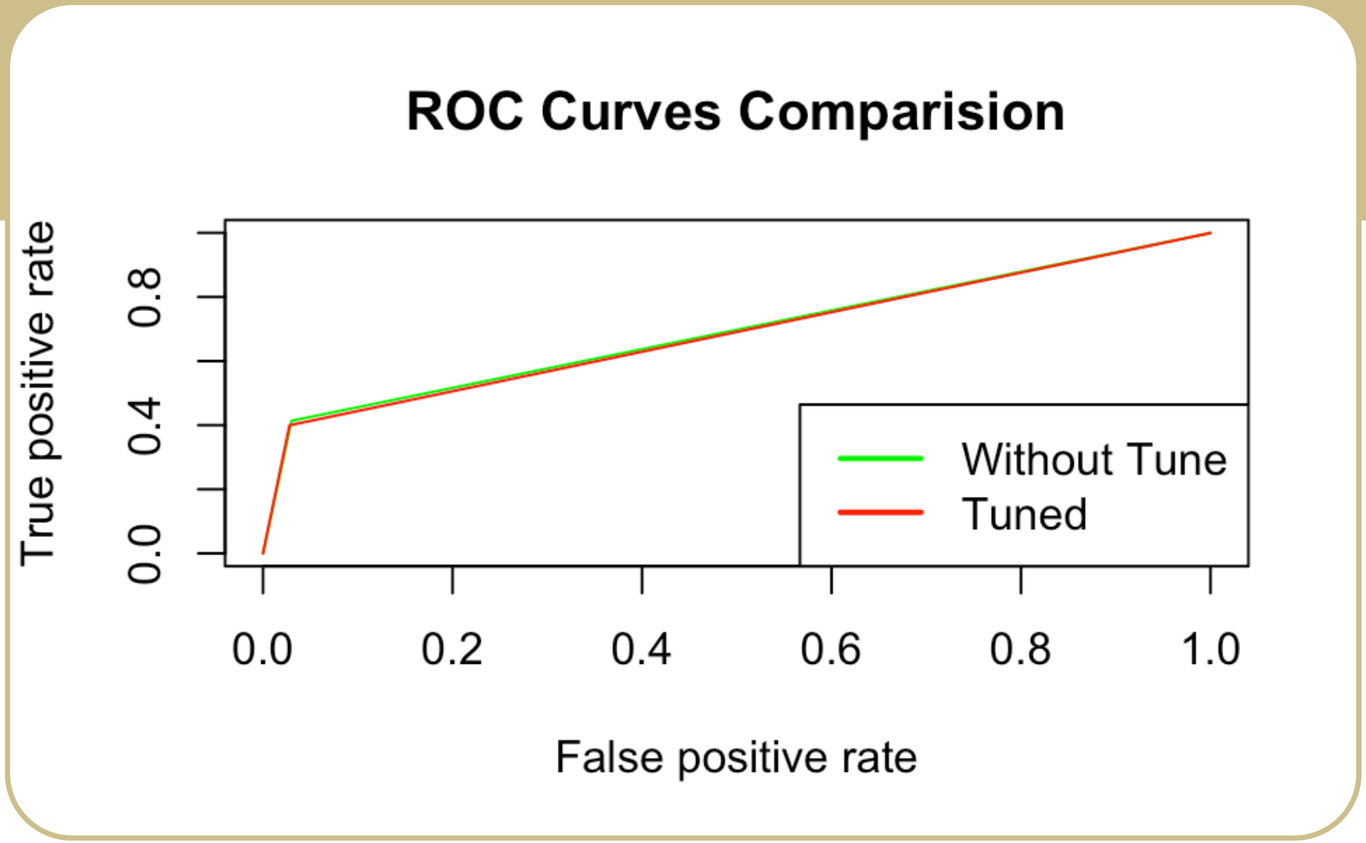
Regularization

Alpha : Type of regularization

Lambda : strength of regularization

References: Favilla 2023

Results: ROC curve before & after tuning of Logistic regression model



Results: Random Forest

mtry : number of variables to be sampled as split criteria at each node

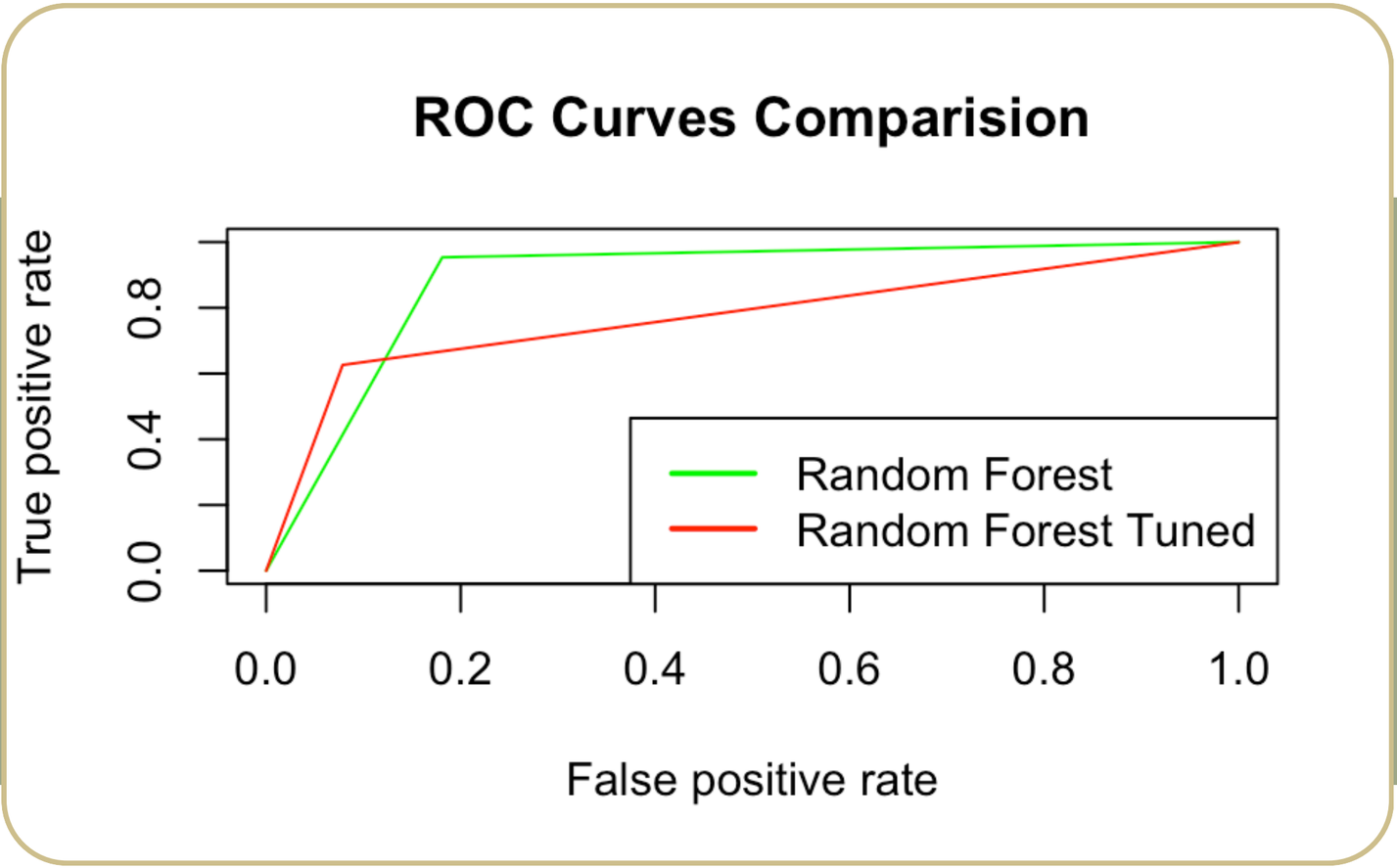
ntree : number of trees

nodesize : minimum size of terminal nodes

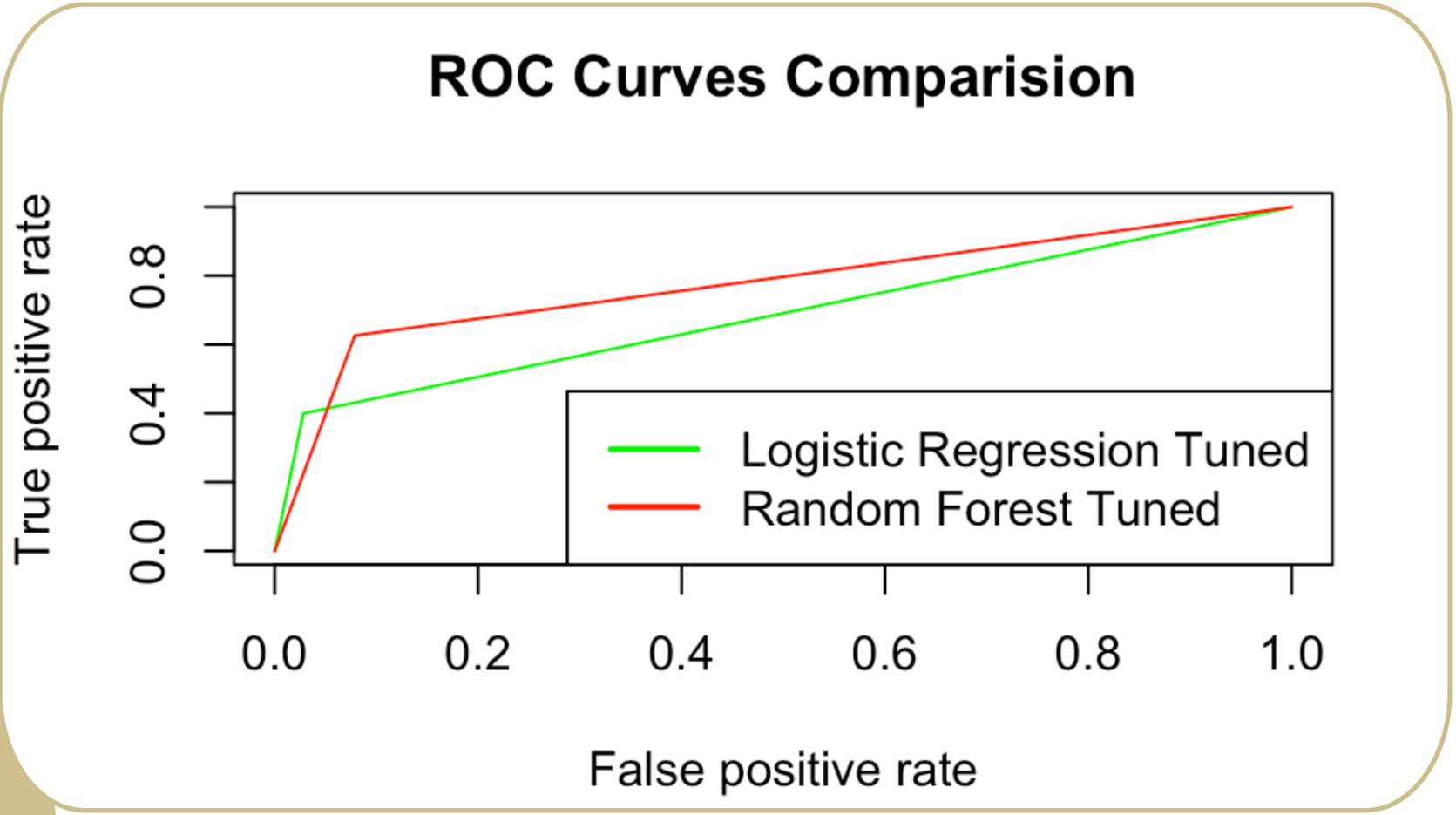
mtry	ntree	nodesize	accuracy
4	1000	1	0.8422
7	1000	1	0.8586
15	1000	5	0.8881

References: RDocumentation n.d.

Results: ROC curve before & after tuning of Random Forest model



Discussion: Comparison of Logistic Regression & Random Forest



Discussion: Strengths & Weaknesses of the Models

Logistic Regression

Strength:

- Higher Accuracy
- Higher Sensitivity compared to random forest, indicating better detection of positive cases.
- Interpretability

Weakness:

- Low Specificity
- Limited for complex relationships

Random Forest

Strength:

- Balanced Accuracy, indicating good performance across both classes
- Robust to overfitting
- Potentially Higher AUC

Weakness:

- Reduced interpretability
- Lower Sensitivity
- Large model sizes

References: Kumar 2024

References

1. Bharathi 2020, *Difference between Log Transformation and Standardization*, Stack Exchange, viewed 5 May 2024, <https://stats.stackexchange.com/questions/483187/difference-between-log-transformation-and-standardization>.
2. Bobbitt, Z 2020, *How to Build Random Forests in R (Step-by-Step)*, Statology, viewed 5 May 2024, <https://www.statology.org/random-forest-in-r/>.
3. Chen, K 2020, The effects of marketing on commercial banks' operating businesses and profitability: evidence from US bank holding companies, *International Journal of Bank Marketing*, 38(5), pp. 1059-1079.
4. Data Tricks 2019, *One-hot encoding in R: three simple methods*, Data Tricks, viewed 5 May 2024, <https://datatricks.co.uk/one-hot-encoding-in-r-three-simple-methods>.
5. Favilla, V 2023, *Regularization in Logistic Regression*, Medium, viewed 5 May 2024, <https://medium.com/@vincefav/regularization-in-logistic-regression-14b50d7cc31>.
6. Intelligent Banking Solutions n.d., *Principles of Bank Collections*, Intelligent Banking Solutions, viewed 5 May 2024, <https://www.ibshome.com/principles-of-bank-collections/#:~:text=For%20money%20collection%2C%20the%20primary,the%20terms%20of%20the%20agreement>.
7. Kumar, S 2024, *Balancing Act: The Pros and Cons of Machine Learning Algorithms*, LinkedIn, viewed 5 May 2024, <https://www.linkedin.com/pulse/balancing-act-pros-cons-machine-learning-algorithms-mba-ms-phd-aty6c/>.
8. Moro, S, Cortez, P & Rita, P 2014, A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, pp.22-31.
9. OAIC n.d., *Defaults*, OAIC, viewed 5 May 2024, <https://www.oaic.gov.au/privacy/your-privacy-rights/credit-reporting/repayment-history-and-defaults#:~:text=A%20credit%20provider%20can%20list%20a%20default%20on%20your%20credit,overdue%20payment%20and%20requesting%20payment>.
10. RDocumentation n.d., *random_forest_parameters: Hyperparameter optimisation or parameter tuning for Random Forest by grid search*, RDocumentation, viewed 5 May 2024, <https://www.rdocumentation.org/packages/scorecardModelUtils/versions/0.0.1.0/topics/random_forest_parameters>.
11. Sembiring, MRP & Leon, FM 2021, The Influence of Demographics Factor on Pension Planning and Financial Literacy of Private Employee, *Business and Entrepreneurial Review*, 21(1), pp.131-152.
12. Toolify.ai 2024, *Transform Outliers and Skewed Data with Log Transformation*, Toolify.ai, viewed 5 May 2024, <https://www.toolify.ai/ai-news/transform-outliers-and-skewed-data-with-log-transformation-1564967#:~:text=Deleting%20outliers%20can%20also%20have,the%20performance%20of%20predictive%20models>.
13. Westpac n.d., *WHAT'S THE DIFFERENCE BETWEEN A TERM DEPOSIT AND A SAVINGS ACCOUNT?*, Westpac, viewed 5 May 2024, <https://www.westpac.com.au/personal-banking/bank-accounts/term-deposit/savings-vs-term-deposit/#:~:text=What%20is%20a%20term%20deposit,on%20your%20money%20will%20be>.
14. Xu, S, Yang, Z, Ali, ST, Li, Y & Cui, J 2022, Does financial literacy affect household financial behavior? The role of limited attention, *Frontiers in psychology*, 13, pp. 1-23.

THANK YOU

Any Questions?

By : F2F Group3