# The prediction of client subscribed a term deposit

Analyzed by Allison

# Contents

- Motivation & Goal
- Dataset Information
- Statistical Analysis
- Model
- Conclusion

# Motivation & Goal

## Motivation

1. Have a preliminary understanding of bank marketing dataset.
2. Through analysis to understand the impact of these variables on the predictive model.

## Goal

➢ Has the client subscribed a term deposit?

# Dataset Information

A. The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.
B. There are 41188 examples in this data.
C. There are 10 numeric variables and 10 categorical variables.
D. Data is ordered by date (from May 2008 to November 2010).
E. 4 attributes:
   **"Bank client"**
   **"Related with the last contact of the current campaign"**
   **"Others"**
   **"Social and economic context"**
F. The output variable Y is defined as **whether the client subscribed a term deposit**.

- **The complete information is presented in another powerpoint named "Data Information".*

# 4 Attributes

| Bank client | Related with the last contact of the current campaign | Others | Social and economic context |
|---|---|---|---|
| age | contact | campaign | emp.var.rate |
| job | month | pdays | cons.price.idx |
| marital | day_of_week | previous | cons.conf.idx |
| Education | duration | poutcome | euribor3m |
| default | | | nr.employed |
| housing | | | |
| loan | | | |

# Statistical Analysis

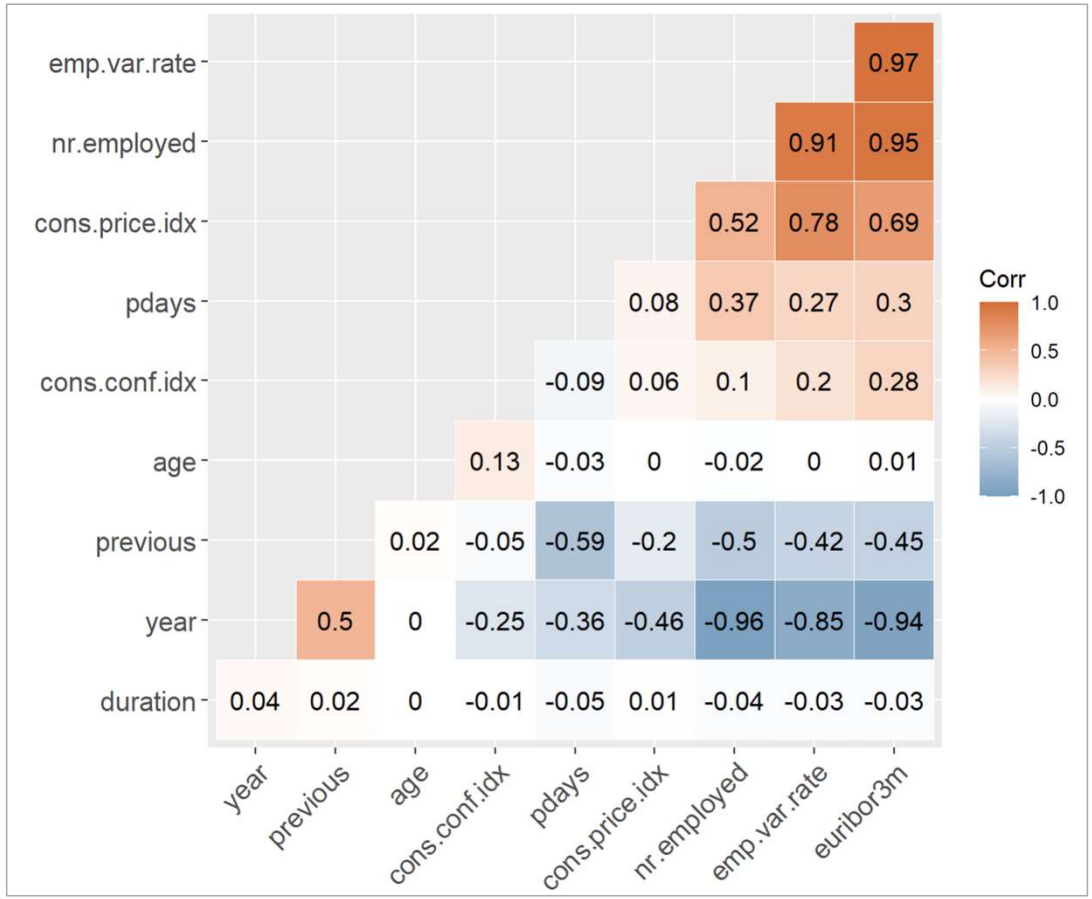
This part is that we want to know the disribution of data and relation between each variable and “y“. There are some method used:

| Descriptive Statistics |
| --- |
| • Measures of Central Tendency<br>• Measures of Dispersion<br>• Measures of Relative Position<br>• Graph |

| Hypothesis Testing |
| --- |
| • t test<br>• Chi-square test<br>• Fisher’s exact test |

# Correlation between continuous variables

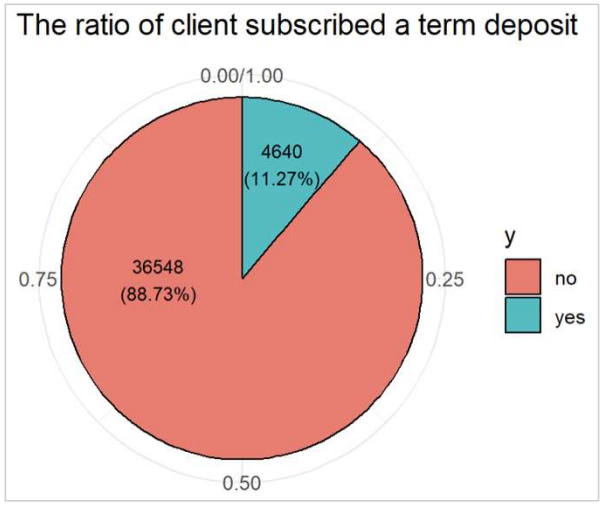
- Variables in “Social and economic context” are positively correlated with each other and negatively correlated with “Year” and “previous”.
- “age” and “duration” are less correlated with other variables.
- “pdays” are positively correlated with variables in “Social and economic context” and negatively correlated with other variables .

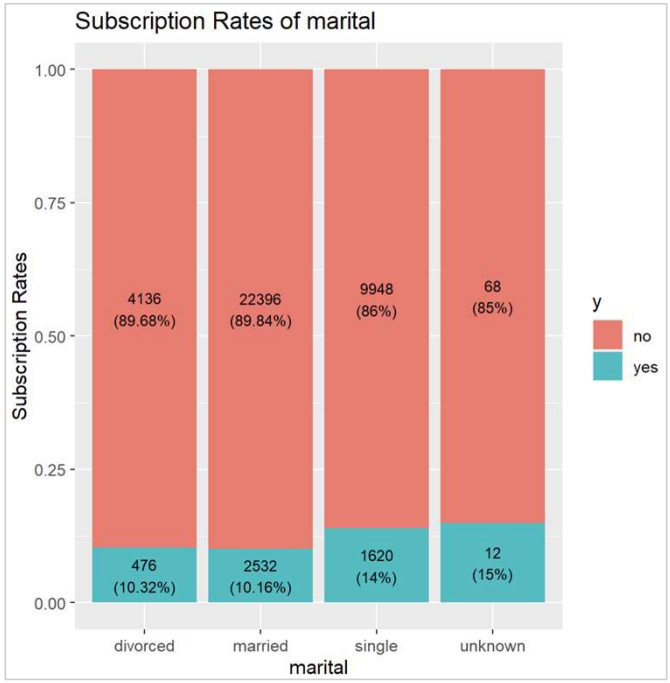


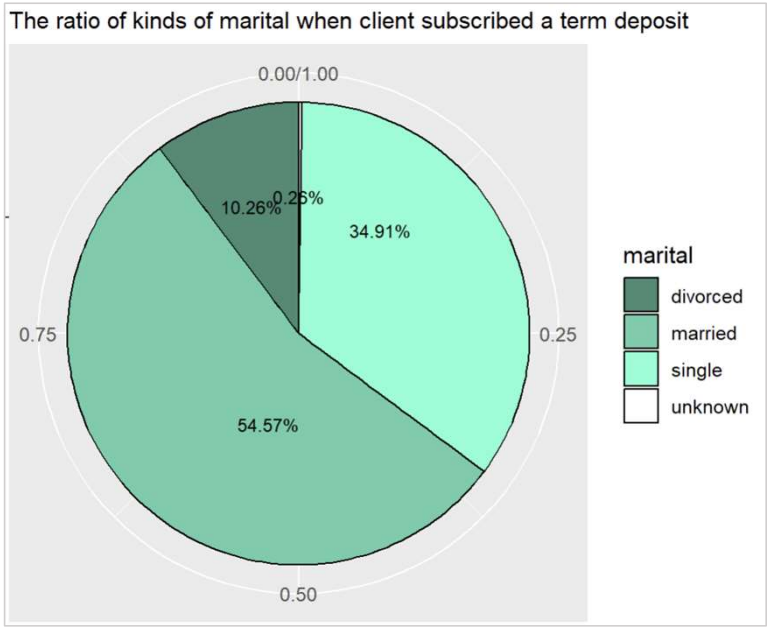
# Y & Marital

Among the 41,188 clients, 4640 of them subscribed the term deposit, which accounts for 11.27%.

- Single clients subscribed the term deposit is often than others.
- The least subscription rates of term deposit is married clients but which also has 2532 people. And for people who subscribed the term deposit, married takes over than 50%.
  → It is also important to pay attention to married clients.



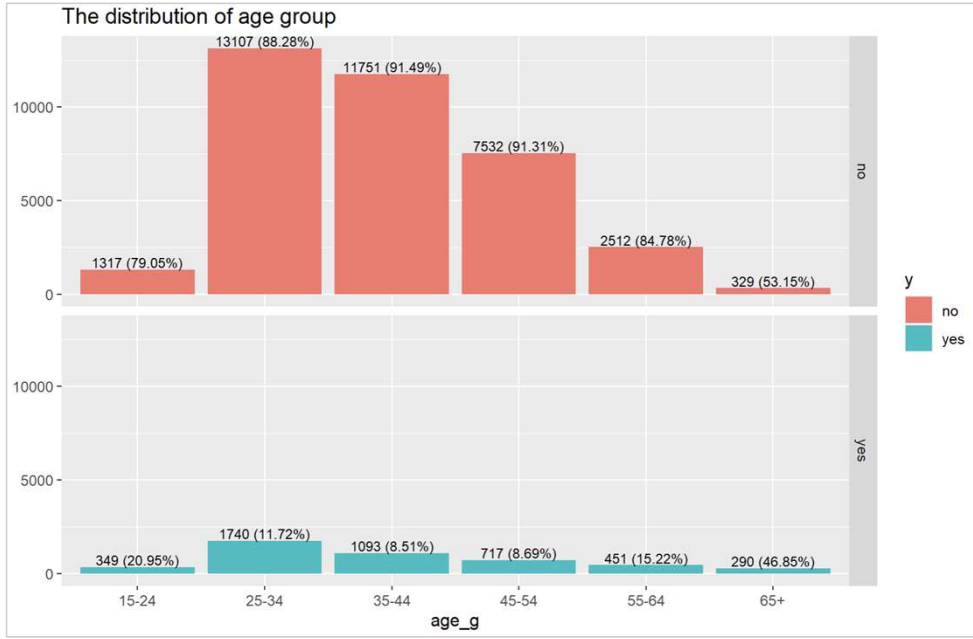



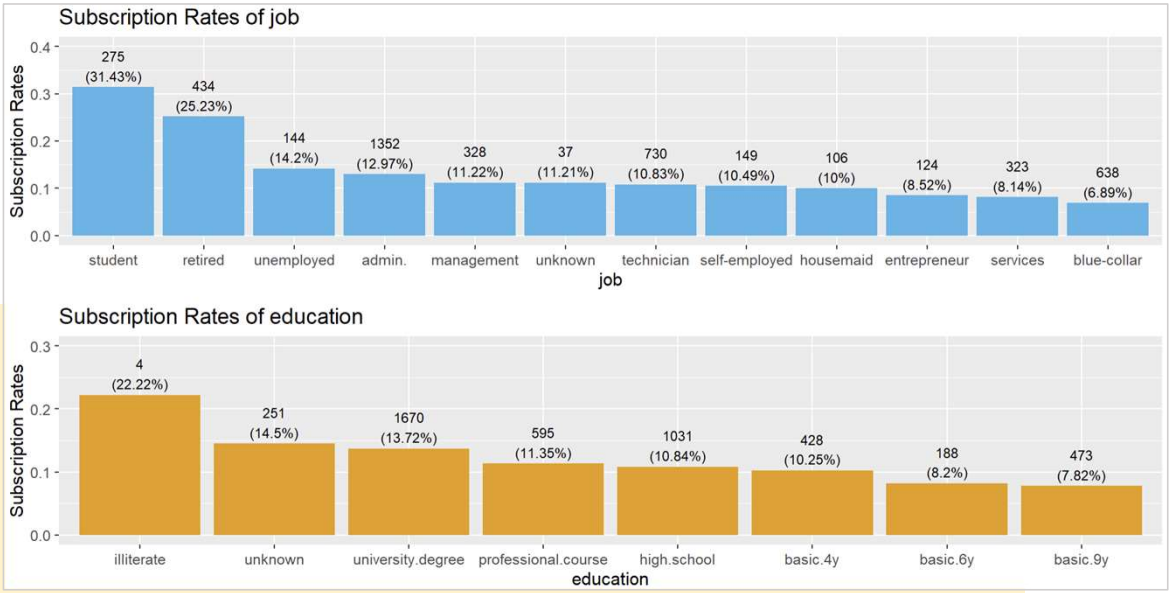
# age

| | y | n | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | IQR | stdev | var |
|---|---|---|---|---|---|---|---|---|---|---|---|
| age | Yes | 4640 | 0.617 | 4.217 | 7.483 | 9.22 | 12.354 | 69.983 | 8.137 | 6.686 | 44.705 |
| | No | 36548 | 0 | 1.583 | 2.725 | 3.681 | 4.65 | 81.967 | 3.067 | 3.452 | 11.914 |


- Most bank customers are between the ages of 25 and 60.
- "25-34" is the most numbers of who subscribed term deposits but the highest proportion is "65+".



# Job & education

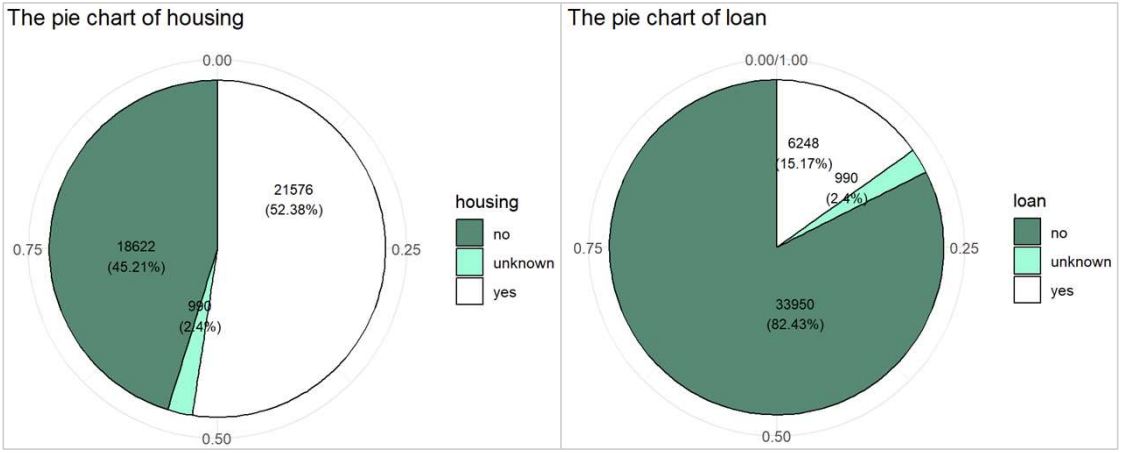


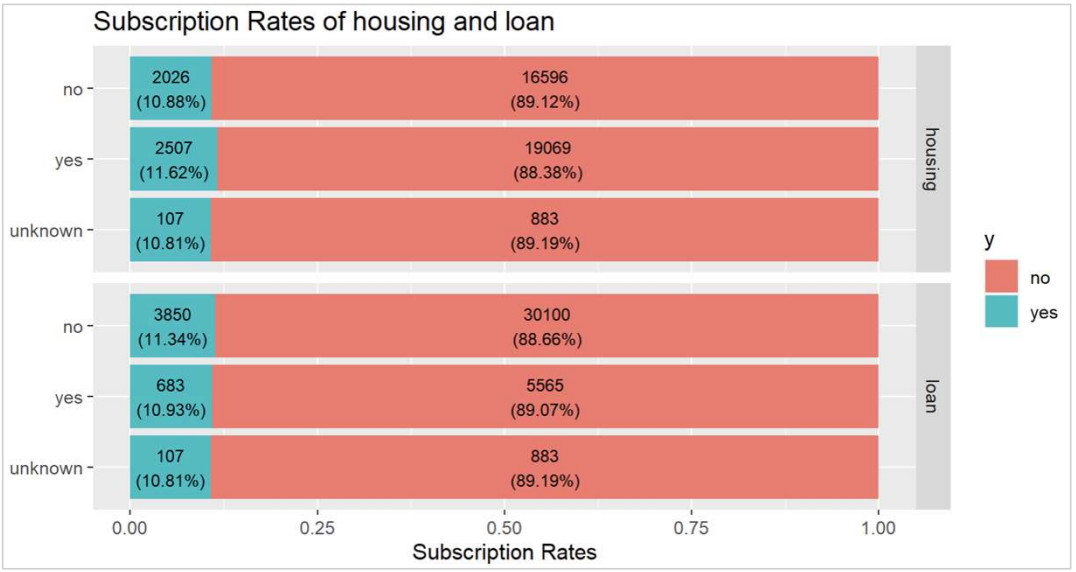

- The subscription rate in job and education means that the rate of subscribed term deposits in the same category. For example, there is 31.43% of all student subscribed term deposits.
- For job, the top 3 subscription rates of term deposit are students, retired people and unemployed.
- The subscription rate of blue-collar is only 6.89%.

  → Indicating low demand for this service.
- For education, the most subscription rates of term deposit is illiterate but which just has 4 people.
- The second is unknown which means we can not ignore these clients.
- Surprisingly, basic.4y is higher than basic.6y and basic.9y.

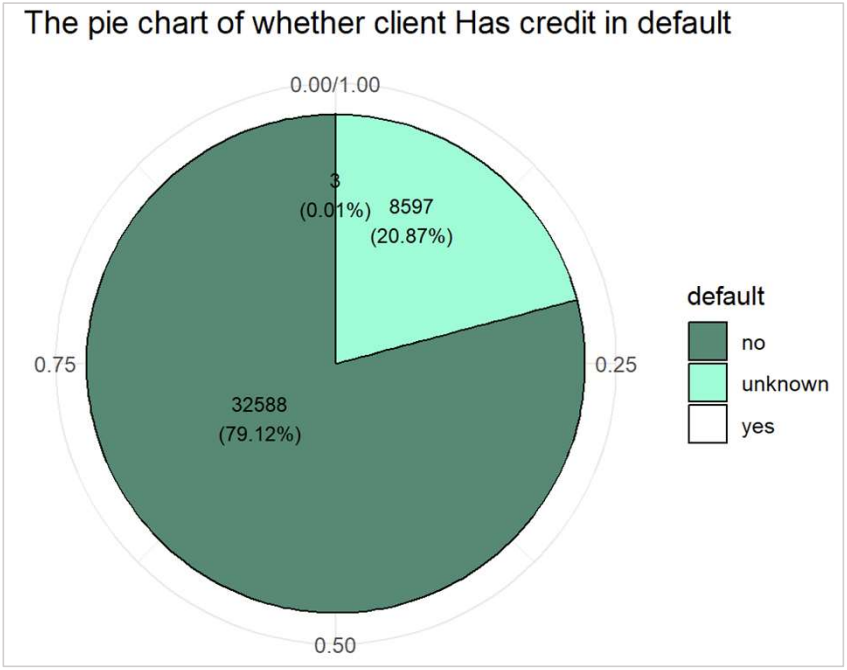
# Housing & Loan

- There are over than 50% people who have housing loan, and about 15% people who have personal loan.
- In housing and loan, subscription rates of each class are around 9:1.
  → Regardless of whether a client has “housing” or “loan”, it will not have much impact on subscribing to the term deposit.

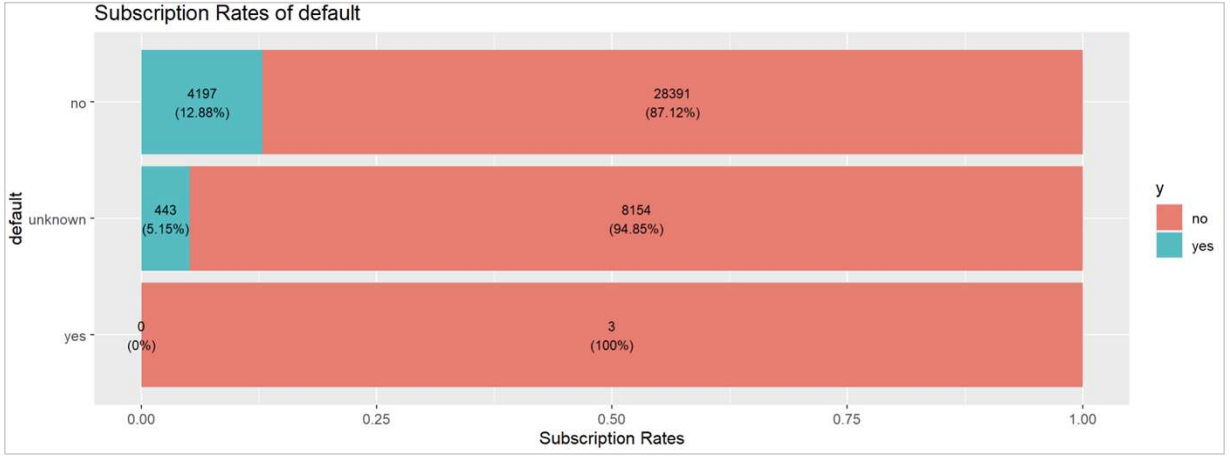


# Default

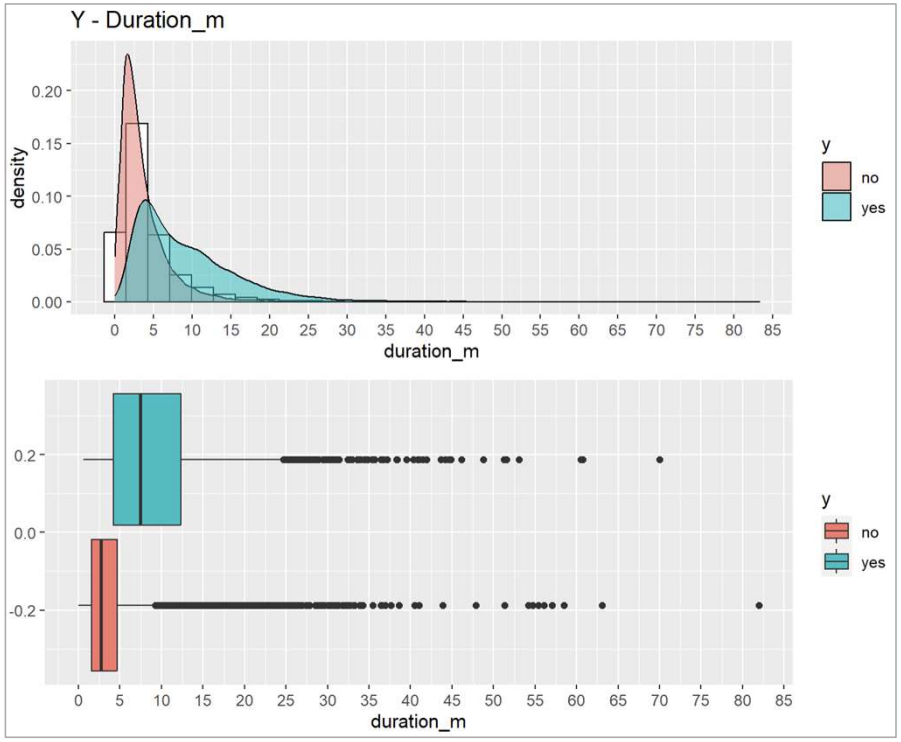



- Unknown is around 21%, and there is 5.15% of unknown subscribed the term deposit.
  → It needs to impute the missing value.
- There are only 3 client has credit in default.
- The subscription rate of that clients does not have credit in default takes 12.88%.



# duration_m

| | y | n | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | IQR | stdev | var |
|---|---|---|---|---|---|---|---|---|---|---|---|
| duration_m | Yes | 4640 | 0.617 | 4.217 | 7.483 | 9.22 | 12.354 | 69.983 | 8.137 | 6.686 | 44.705 |
| | No | 36548 | 0 | 1.583 | 2.725 | 3.681 | 4.65 | 81.967 | 3.067 | 3.452 | 11.914 |

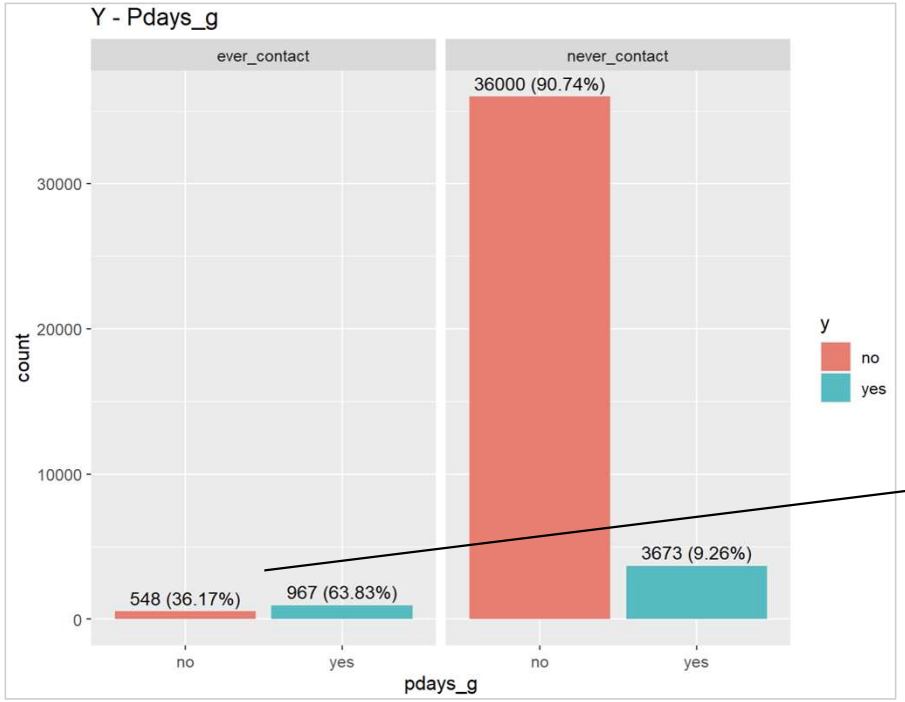



- For convenience of observation, we convert the unit of duration from seconds to minutes which named "duration_m".
- 50% of customers who subscribed a term deposit are concentrated around 5 to 13 minutes. However, 50% of customers **not** subscribed a term deposit are concentrated around 1.5 to 5 minutes.

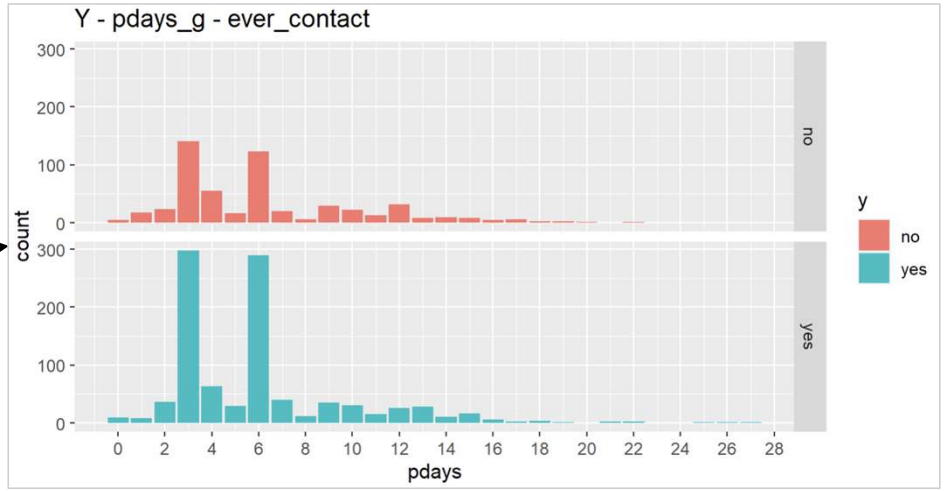
  → The length of the phone call significantly affected the willingness to subscribe to term deposits.

# pdays_g

| | | pdays_g | |
|---|---|---|---|
| | | never_contact (0-27) | ever_contact (999) |
| Y | yes | 3673 | 967 |
| | no | 36000 | 548 |

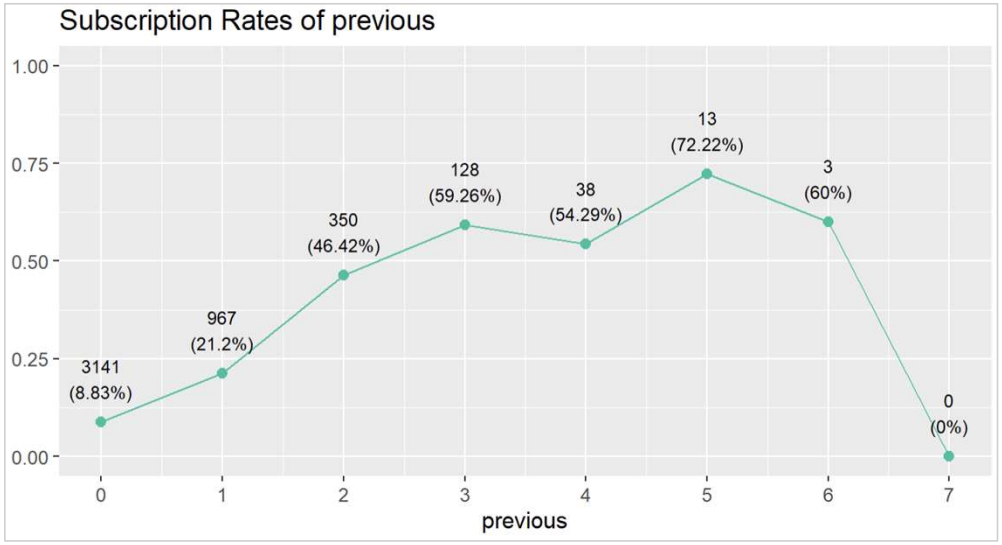



- According to whether clients have been contacted before, "pdays" is divided into 2 groups and a new variable "pdays_g" is formed.
- In "ever_contact", it was about 64% client subscribed a term deposit. However, it was just about 10% clients in "never_contact".
  → Clients contacted at the last activity have a higher acceptance of this service.
- In "ever_contact", nearly 600 clients subscripted term deposits were concentrated 3 days or 6 days after the last activity contact.



# previous

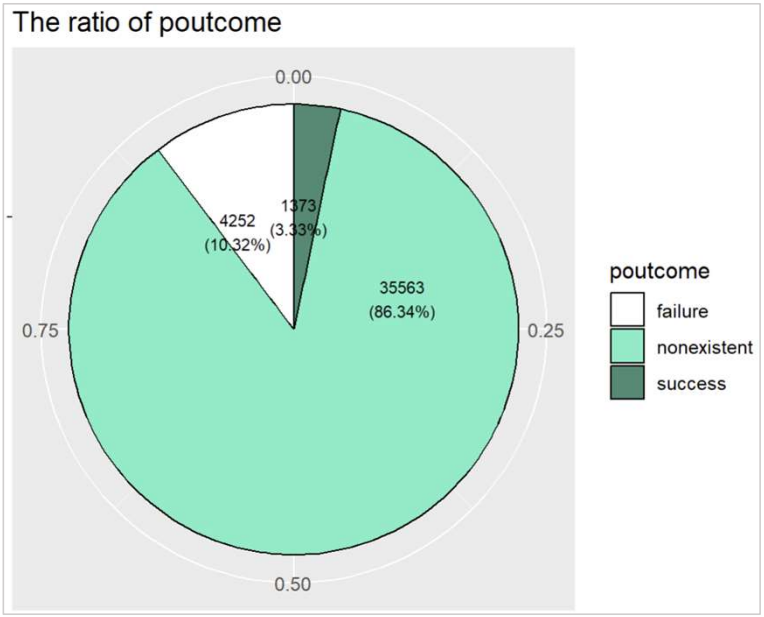
| previous | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| no | 32422 | 3594 | 404 | 88 | 32 | 5 | 2 | 1 |
| yes | 3141 | 967 | 350 | 128 | 38 | 13 | 3 | 0 |

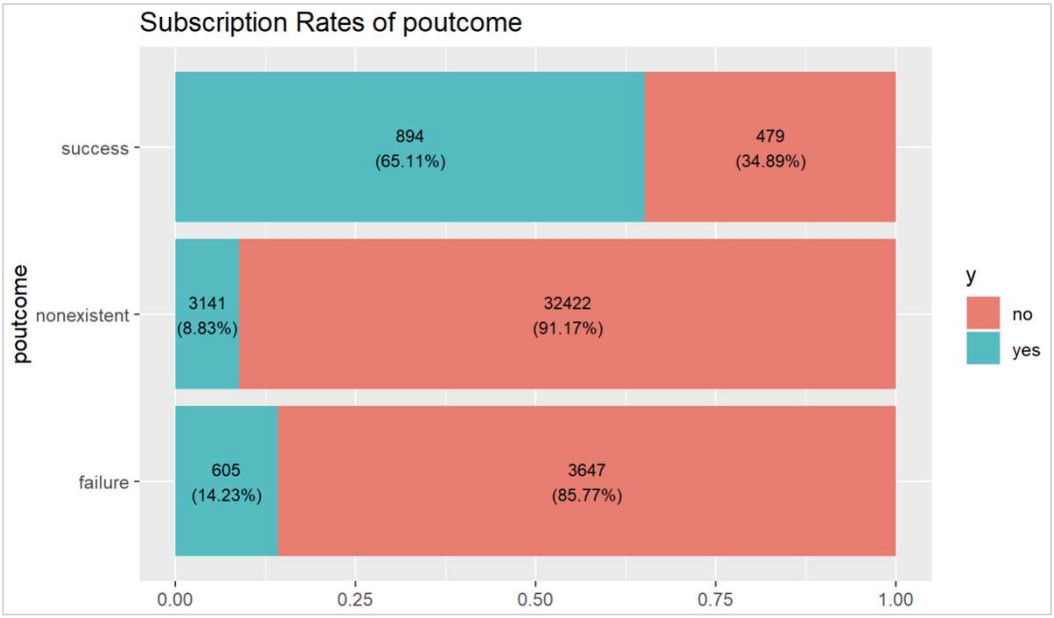



- The number of times to exclude contacting a customer is 7 times. The more times the clients are contacted before the advertising campaign, the higher the subscription rate of term deposit are.

# poutcome

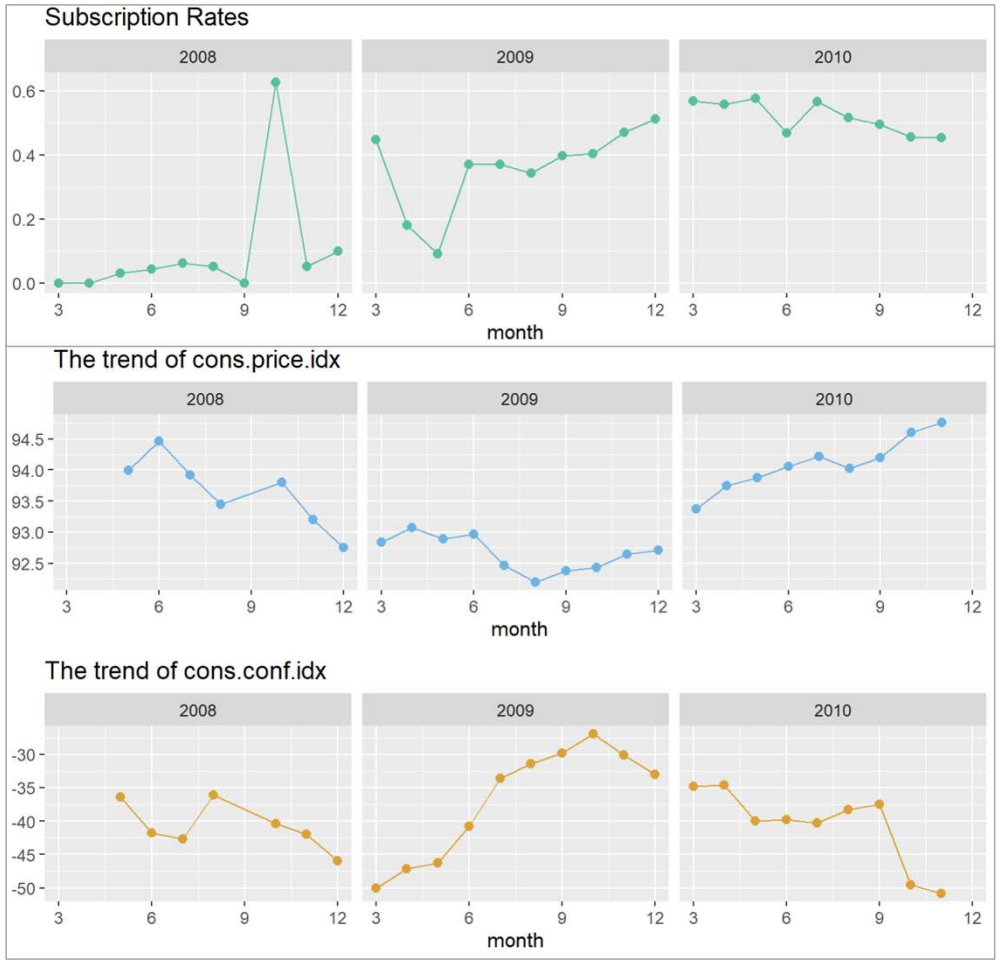
- In this data, there are 5625 people joined previous marketing campaign and the result of success only 3.33%.
- The subscription rate for a term deposit is as high as 65% among the relevant clients who had achieved success in previous marketing campaigns and the subscription rate of “failure” is higher than “nonexistent”.
  → There is a positive relationship between subscription rate and participation of previous marketing campaigns



# cons.price.idx & cons.conf.idx

- In 2008, the trends of both indicators were downward. They were different from the subscription rate which had remained around 5% and suddenly exceeded 60% in October. (Financial crisis of 2007–2008)
- In 2009, the overall trend of consumer price index was downward but others began to rise after June.
  → **weak** consumer spending power
- In 2010, only the trends of consumer confidence index was downward.
  → People became increasingly less optimistic about current and future economic conditions.



# Hypothesis Testing 1

- Compare the test results of “age” and “age_g”, both are statistically significant.
- According to the test results of 12 variables, "housing" and "loan" are not statistically significant
  → It means that except for “housing” and “loan”, there are differences of other variables between the two groups with subscription and fixed deposit and without subscription and fixed deposit.

**# Y – Each Variable**

| ***# Bank Client Attributes*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Variables** | age | age_g | job | maritial | eduation | default | housing | loan |
| **Test** | t test | Chisq | Chisq | Chisq | Fisher | Fisher | Chisq | Chisq |
| **P-value** | 1.805e-06 | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 | 0.0004998 | < 2.2e-16 | 0.05829 | 0.5787 |

| ***# Related with The Last Contact of The Current Campaign Attributes*** | | | | |
|---|---|---|---|---|
| **Variables** | contact | month | day_of_week | duration |
| **Test** | Chisq | Chisq | Chisq | t test |
| **P-value** | < 2.2e-16 | < 2.2e-16 | 2.958e-05 | < 2.2e-16 |

# Hypothesis Testing 2

- Compare the test results of "pdays" and "pdays_g", both are statistically significant.
- According to the test results of 10 variables, all of them are statistically significant.
  → It means that there are differences of all variables between the two groups with subscription and fixed deposit and without subscription and fixed deposit.

| # Y - Each Variable | | | | | |
|---|---|---|---|---|---|
| ***# Other Attributes*** | | | | | |
| **Variables** | campaign | pdays | pdays_g | previous | poutcome |
| **Test** | t test | t test | Chisq | t test | Chisq |
| **P-value** | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 |
| ***# Social and Economic Context Attributes*** | | | | | |
| **Variables** | emp.var.rate | cons.price.idx | cons.conf.idx | euribor3m | nr.employed |
| **Test** | | | t test | | |
| **P-value** | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 |

# Model

## Data Preprocessing

- Create New Variables
- Check & Impute Missing Value

## Training Model

- Training Data & Testing Data
- Feature Selection
- Model Performance

# Data Preprocessing

**According to the some data information and previous analysis , we do some preprocessing.**

## Create New Variables

- year (because original data is ordered by date)
- age_g (split age into  groups)
- pdays_g (split pdays into 2 groups: “ever_contact”, “never_contact”)
- duration_m (convert the unit of “duration” from second to minute)

## Check & Impute Missing Value

- There are 6 variables which totally have 12718 unknown value.
- Use package “mice” to impute missing value.

**Missing Value counting**

```
          variable unknown
1          default    8597
2        education    1731
3          housing     990
4             loan     990
5              job     330
6          marital      80
7              age       0
8          contact       0
9             year       0
10           month       0
11     day_of_week       0
12        duration       0
13        campaign       0
14           pdays       0
15        previous       0
16        poutcome       0
17    emp.var.rate       0
18  cons.price.idx       0
19   cons.conf.idx       0
20       euribor3m       0
21     nr.employed       0
22               y       0
> sum(dt0 == "unknown")
[1] 12718
```

# Training Model

## Training Data & Testing Data

- The ratio of training data and test data is 8:2.

## Feature Selection

- Features are selected based on the previous analysis and using LR model with stepwise selection.

## Model Performance

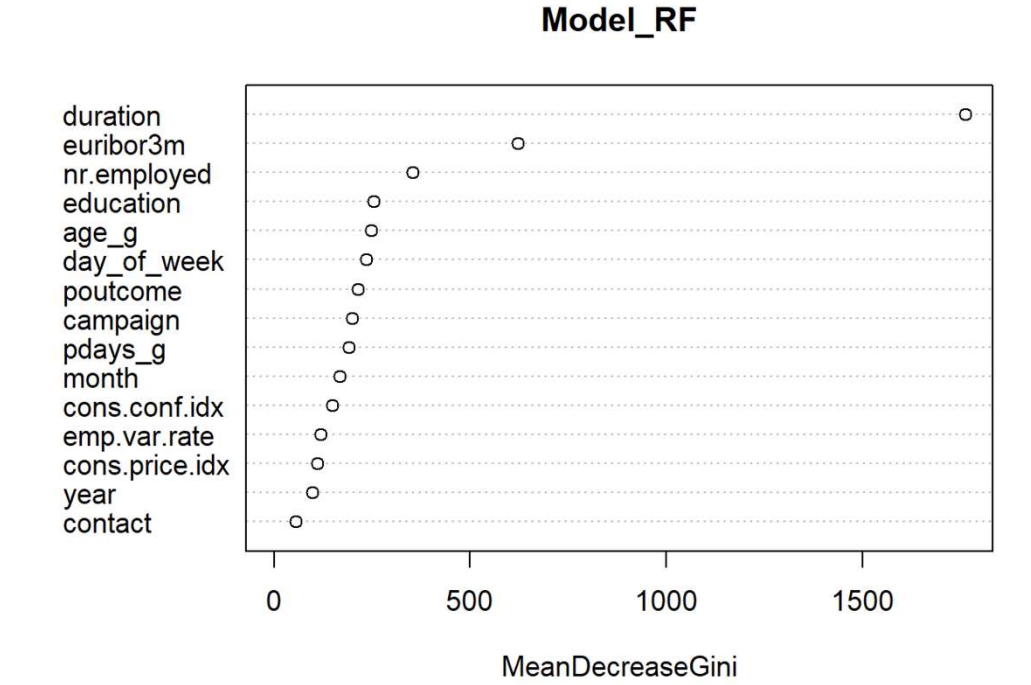
- Model: LR、RF、 DT、 SVM
- Validation Index:
  Accuracy、 Sensitivity、Specificity、PPV、NPV、ROC、AUC

# Feature Selection

- The features selected after using LR model with stepwise are put into the RF model, and Mean Decrease Gini is calculated to know the importance of features.
  → "Duration" is most important than other features.

**Final Selected Features**

| | | |
|---|---|---|
| age_g | day_of_week | emp.var.rate |
| education | duration | cons.price.idx |
| contact | campaign | cons.conf.idx |
| year | pdays_g | euribor3m |
| month | poutcome | nr.employed |



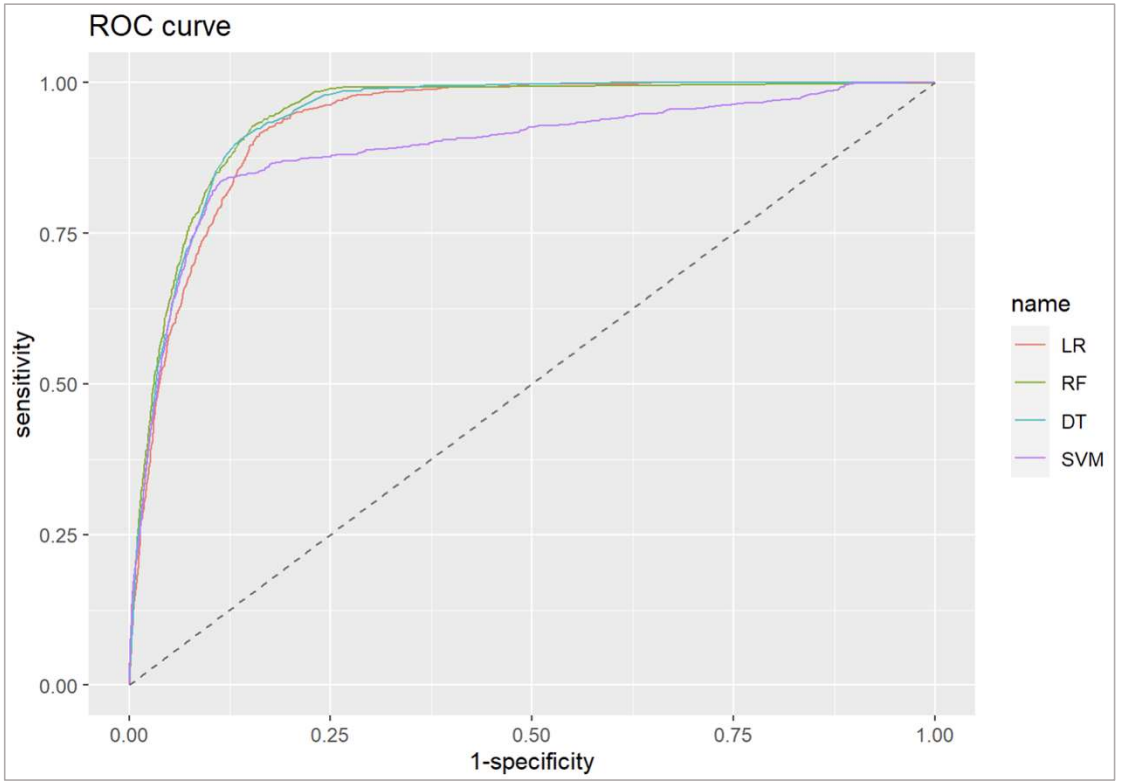
# Model Performance

- According to the table, when the performance of other indicators is very close, the sensitivity of Decision Tree is higher.
  → This means that among people who actually subscribe to term deposits, decision tree model is more accurate in predicting those who subscribe to term deposits.

| Model | Accuracy | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|
| Logistic Regression | 0.9069 | 0.30411 | 0.98092 | 0.66184 | 0.91986 |
| Random Forest | 0.9170 | 0.47614 | 0.97124 | 0.67031 | 0.93778 |
| Decision Tree | 0.9156 | 0.54828 | 0.96075 | 0.63171 | 0.94541 |
| SVM | 0.9114 | 0.36515 | 0.97847 | 0.67556 | 0.92620 |

# Model Performance

- According to the ROC curve, it is obvious that the performance of random forest and decision tree are better than others.
- Though the performances of RF and DT are very close, considering the previous model indicators, the DT will be chosen overall.



# Conclusion

- Base on the importance calculated by the RF model, “duration" is the most important feature. “euribor3m” is unexpectedly 2nd important feature because it has nothing to do with the subscription rate of term deposit intuitively.
- Although the AIC decreased when the automatic model selection method is used, only the sensitivity of DT is over than 0.5.
- Since data imbalance will affect sensitivity and fitting results.
  → Using k-fold cross validation to make fitting of model better.

# References


- UCI Machine Learning, Bank Marketing, url: https://archive.ics.uci.edu/dataset/222/bank+marketing
- JianKai Wang, 隨機森林 (Random Forest), url: https://rpubs.com/jiankaiwang/rf
- Kendal Wong, Chapter 24: Decision Trees, url: https://ademos.people.uic.edu/Chapter24.html#34_decision_trees_with_all_variables
- Skydome20, R筆記 – (14)Support Vector Machine/Regression(支持向量機SVM), url: https://rpubs.com/skydome20/R-Note14-SVM-SVR
- R Documentation, Plot a ROC curve with ggplot2, url: https://rdrr.io/cran/pROC/man/ggroc.html

# Thanks