# Analysis of Bank Term Deposit


GROUP 3

Yuhan Wang | Xinyao Fu | Wanjing Zhang | Peiqi Tang | Abhiraj Malappa | Vaishali Kelkar


# Business Scenario

Case:

- Marketing campaigns data collected from a Portuguese retail bank, from May 2008 to June 2013, total 41188 data points
- The marketing campaigns were based on phone calls
- Bank Targets customers of varied age, job, education etc by direct marketing to subscribe product-Term Deposit

Objective:

- Predict whether a client would subscribe to bank Term Deposit (Y) or not (N)

# Data Exploration

## Overview:

- Total 41188 observations, 21 attributes

### Numeric

- **Age**
- **Day** (last contact day)
- **Duration**(last contact duration)
- **Campaign**(number of contacts)
- **Pdays** (last contacted)
- **Previous**(contact before this campaign)

### Categorical

- **Job**
- **Marital**
- **Education**
- **Contact** (type)
- **Month** (last contact month)
- **Poutcome** (outcome of the previous marketing campaign)

### Binary

- **Default** (has credit in default?)
- **Housing** (housing loan?)
- **Loan** (personal loan?)
- **Y** (subscribed a term deposit? (binary: "yes","no")

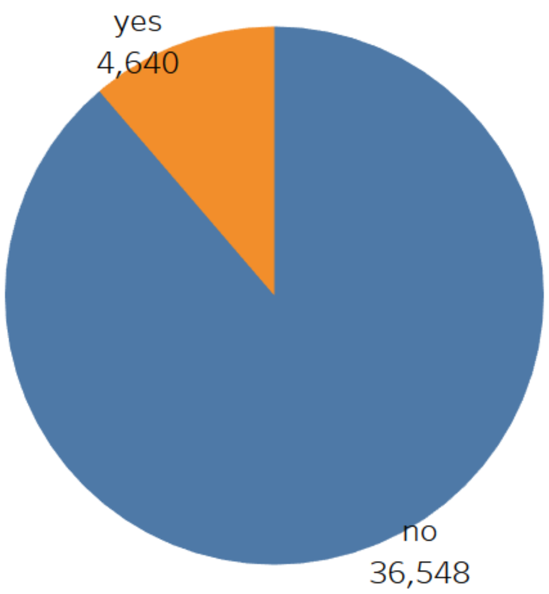
### Social and Economic attributes

- **emp.var.rate** (Employment Variation Rate)
- **cons.price.idx** (Customer Price Index)
- **cons.conf.idx** (Consumer confidence index)
- **Euribor3m** (euribor 3 month rate)
- **nr.employed** (number of employees – quarterly)

# Data Exploration

## Overview:

- Class imbalance - 88.7% no and 11.3% yes

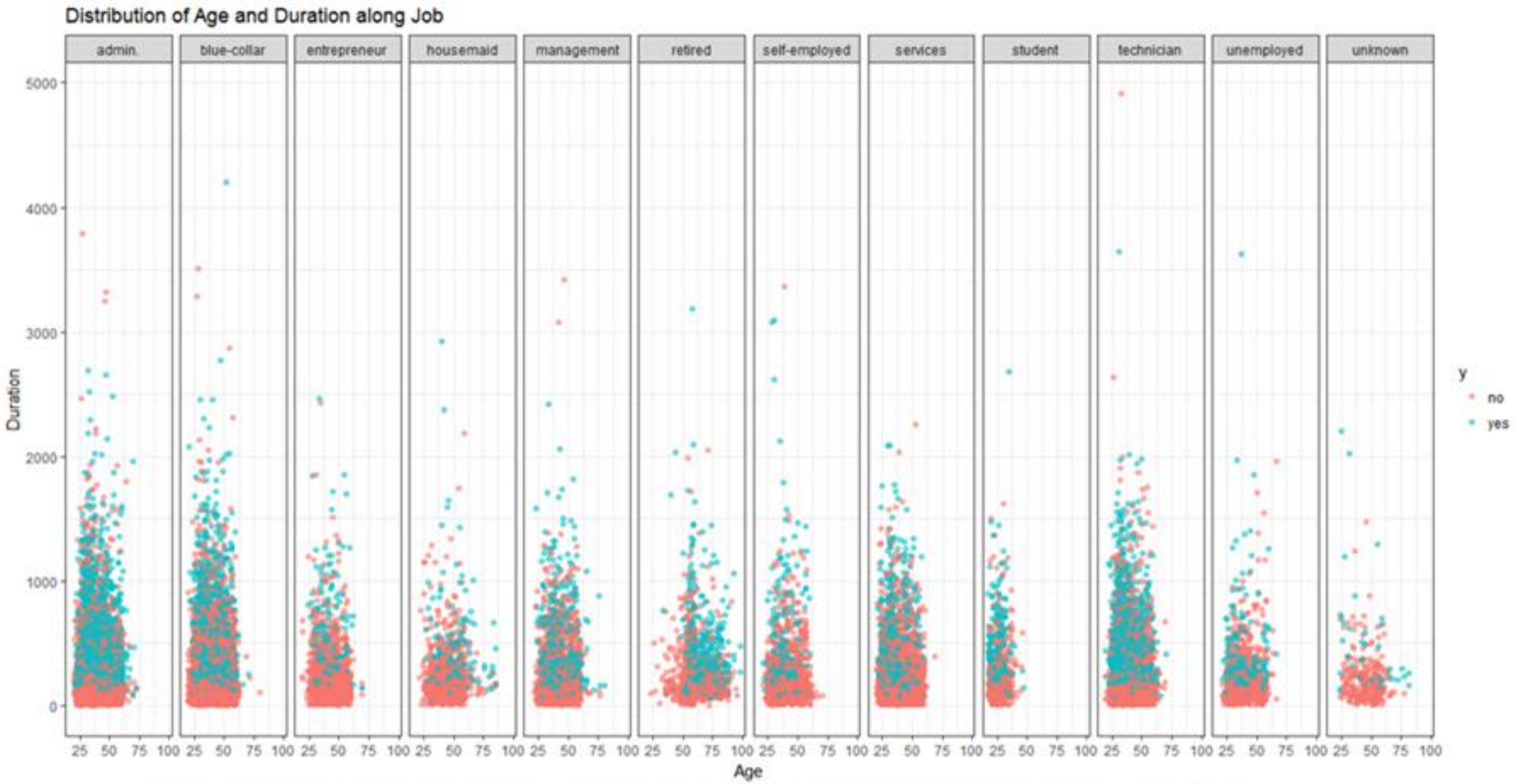



- Missing data - 26% of the records have some missing/unknown attribute values

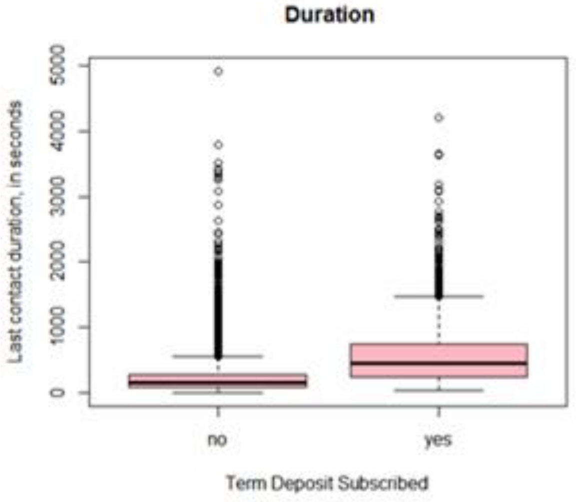
```
> summary(complete.cases(df1))
   Mode   FALSE    TRUE    NA'
logical   10700   30488
```

# Data Exploration: Key Findings

- ★ The lower duration group dominated by ‘NO’ (Red dots)
- ★ The Jobs-Retired, Housemaid, Unemployed and Unknown have lowest Duration and same for NOT subscribing to Term Deposit
- ★ The Student and Retired Job type show distinct age groups.

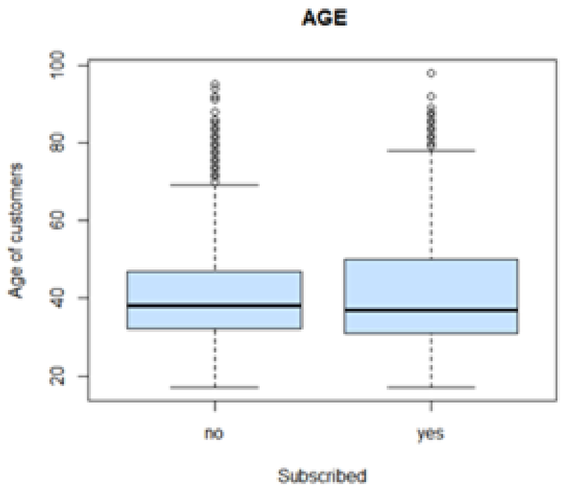


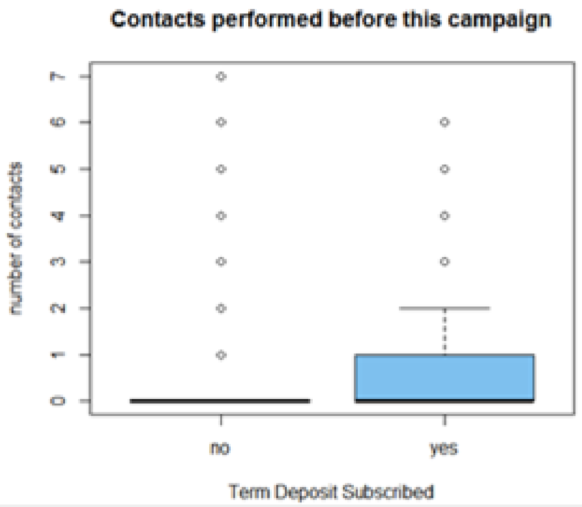

The Scatter plot distribution of Duration of contact (Y axis) against Age (X axis), for each Job type

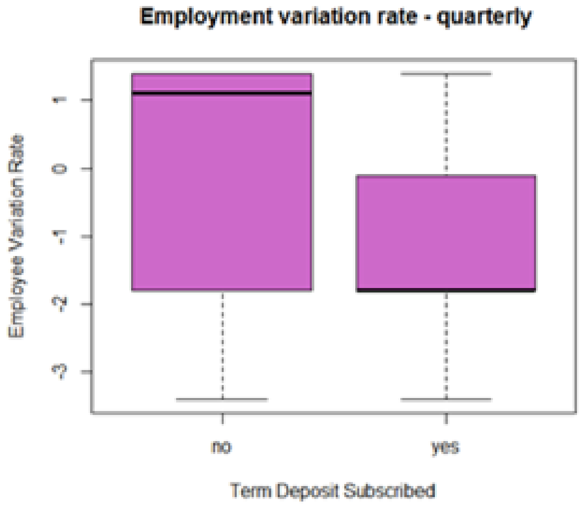



Box plot for Term Deposit (X axis) against Duration (Y axis)

# Data Exploration




Similar distribution of target variable, with Median- ~40.

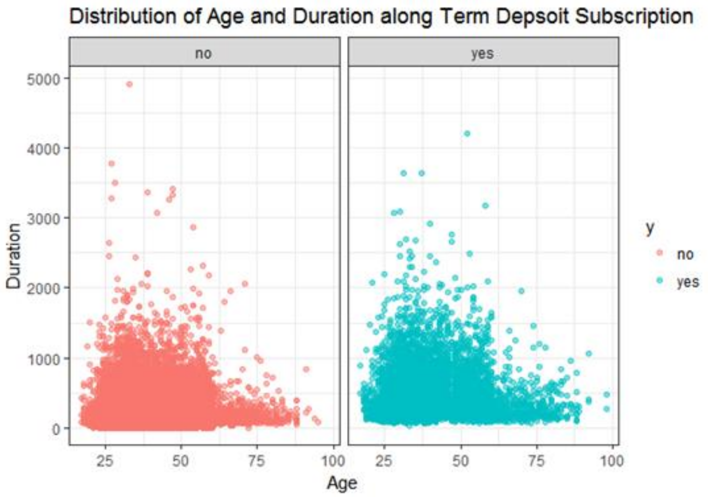
- ★ For more Business = Target age group is 35-50.

Unequal distribution, 0 Contact = No Subscription, More no. of Contact is likely to cause subscription of Term Deposit.

- ★ This attribute can be exploited for more business

Unequal distribution- Median and range for NOT subscribed higher

- ★ Varied Employment less likely to attract customers for 'yes'




Call duration is higher for the age groups below 60

# Data Preparation

## Problems with the data:

- Missing Data
- Class Imbalance
- Possibility of Target Leakage

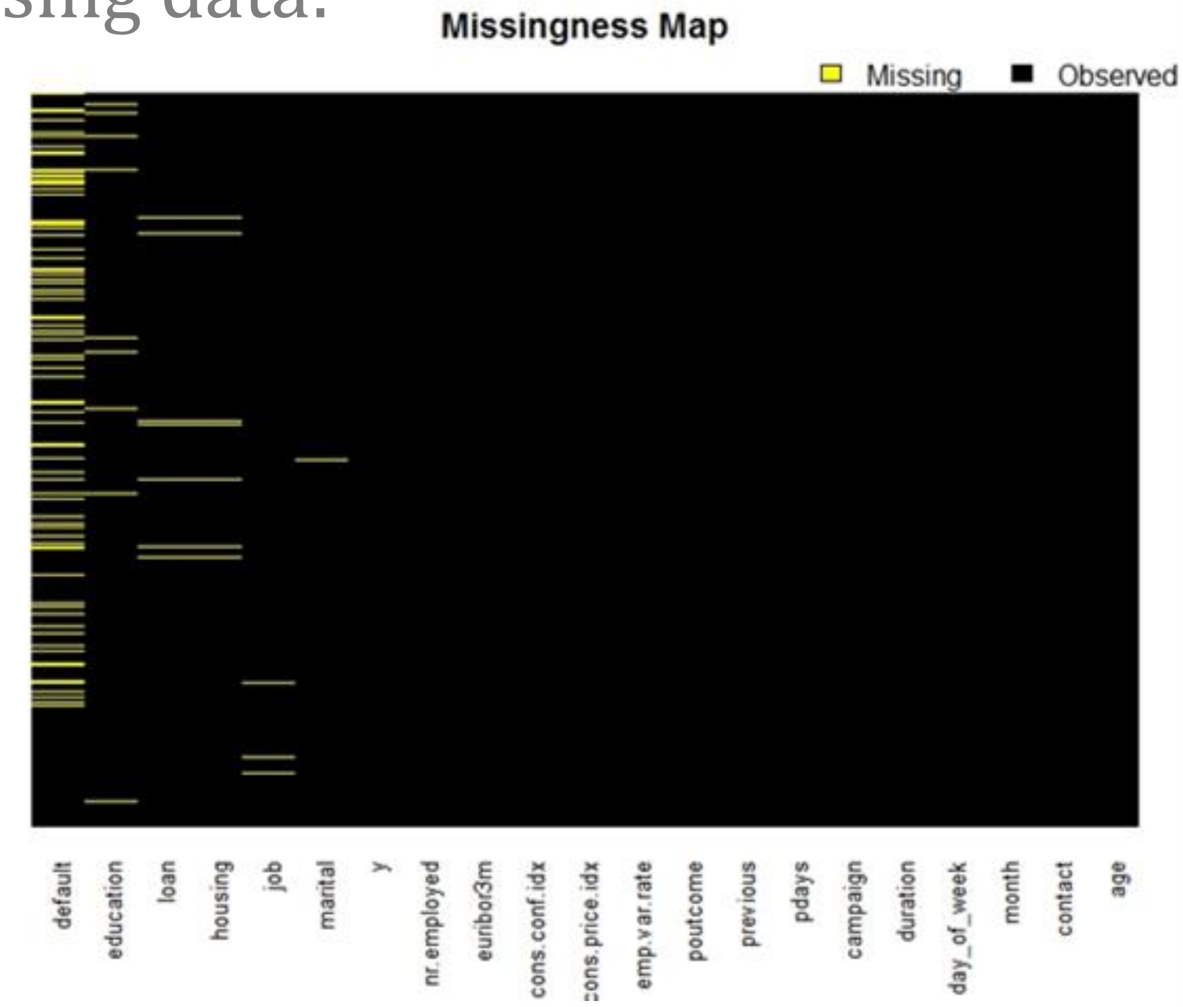
# Data Preparation

## Handling Missing data:

- ‘Default’ attribute is highly imbalanced and has very high percentage of missing data

| Default | |
|---|---|
| no | 32,588 |
| unknown | 8,597 |
| yes | 3 |

- NMAR - Housing and Loan attributes
- Technique tried - MICE, kNN
- MAR - <2%, deleted



# Data Preparation

## Handling Class imbalance:

- Error on minority prediction is really high
- Techniques we tried to overcome - **SMOTE** and cost sensitive learning
- **Synthetically created minority** observations and **undersampled majority class** to create balanced data
- Used the balanced data for training model

```
Error matrix for the Random Forest model on bank_knn.csv

       Predicted
Actual    no  yes Error
   no   0.87 0.02  0.02
   yes  0.08 0.03  0.70

Overall error: 10%, Averaged class error: 36%
```

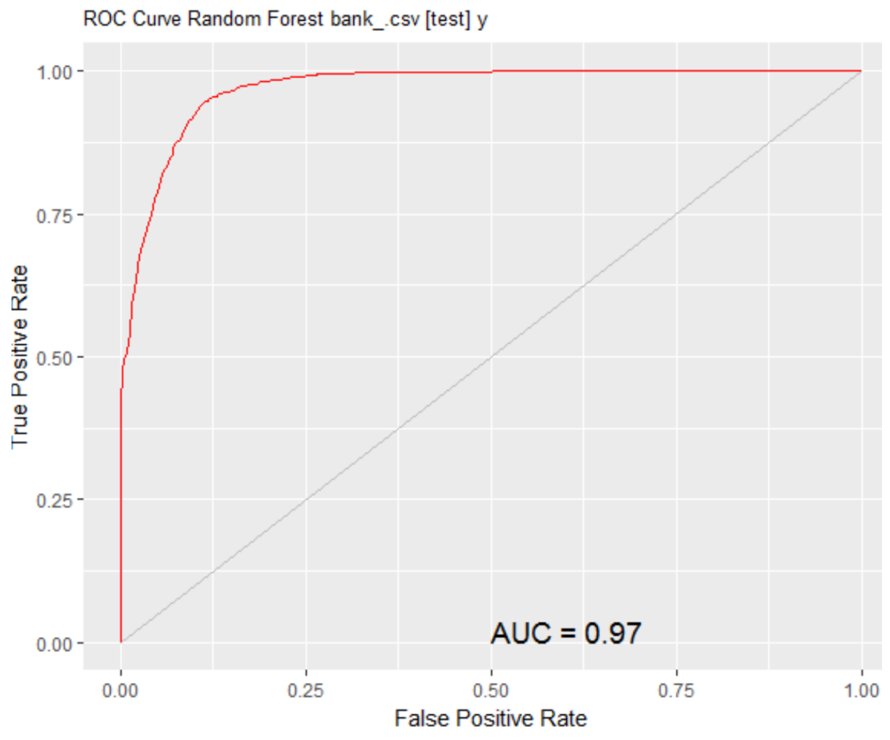
```
smotedDataTraining<-SMOTE(y ~ ., trainingDS, perc.over =100,perc.under=200)
> table(smotedDataTraining$y)

  no  yes
6568 6568
```

# Data Preparation

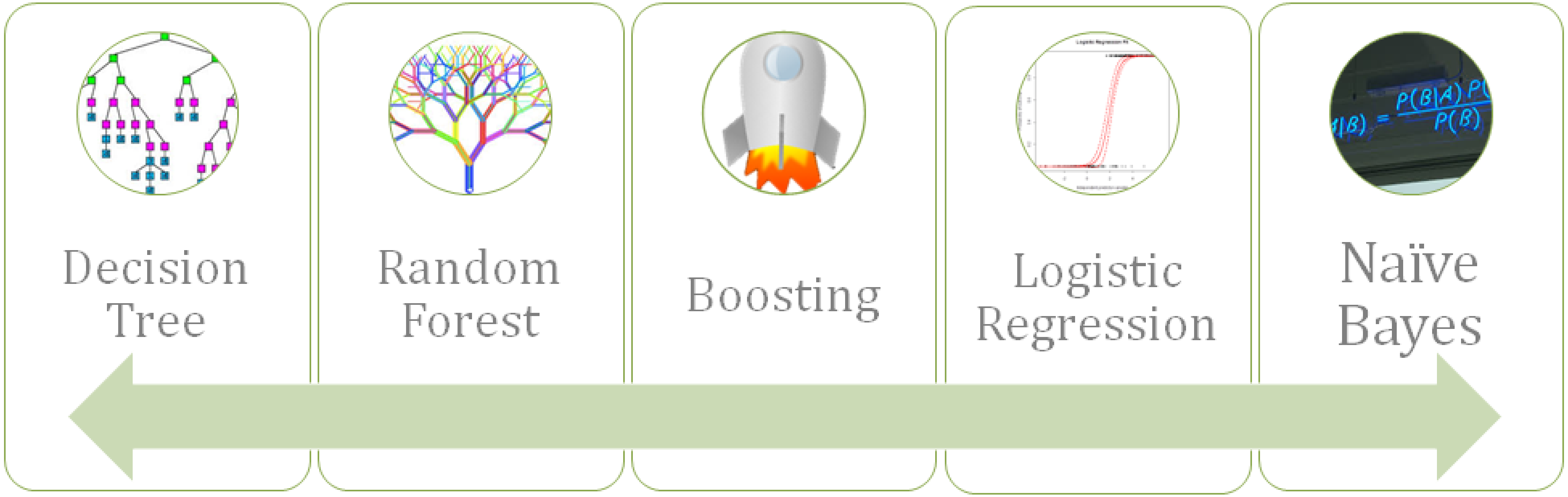
## Dropping variables

- Attribute call duration highly affects the output target.
- If duration is 0 then target value is always be ‘no’ since the call is still not made
- Dropped the variable




Accuracy without treating target leakage

# Data Modelling



# Model Building

## Random Forest

Implementation -

- Data split - 70% training, 30% test
- Built training model over smoted data
- No. of trees -200, eliminating overfitting
- Tested the model on imbalanced data
- Attributes euribor3m, age, job, nr.employed provide the highest information gain

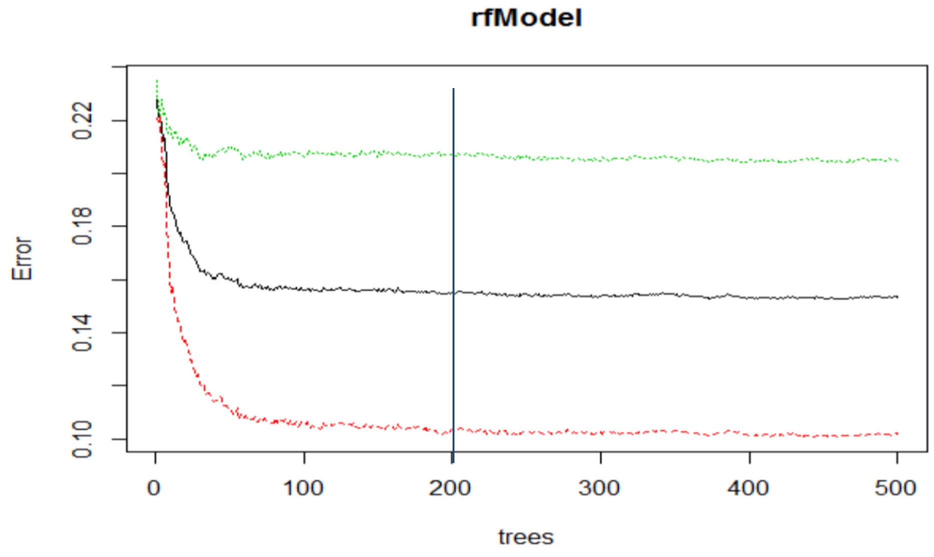
Results -

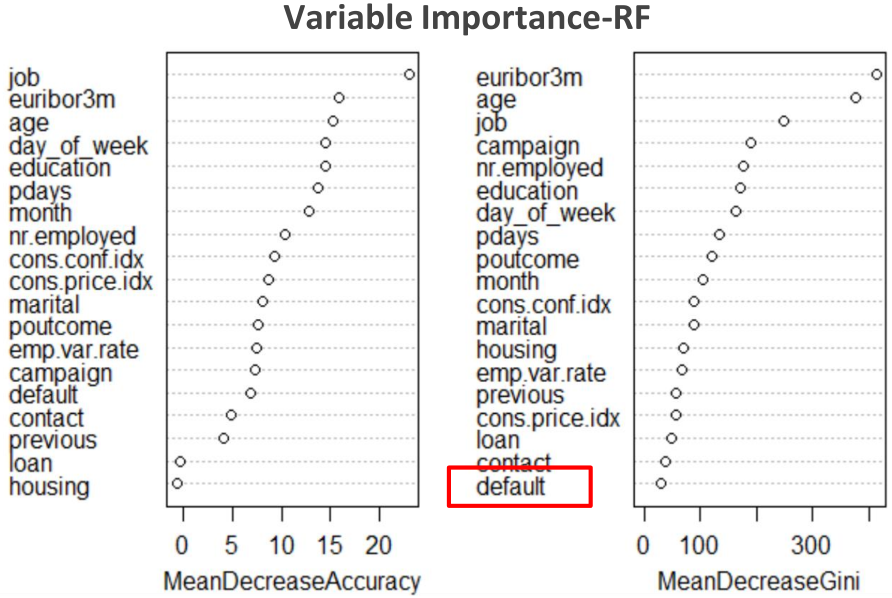
- AUC - 0.81; F1 measure - 0.59

```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  9363  862
       yes  368  884

               Accuracy : 0.8928
```



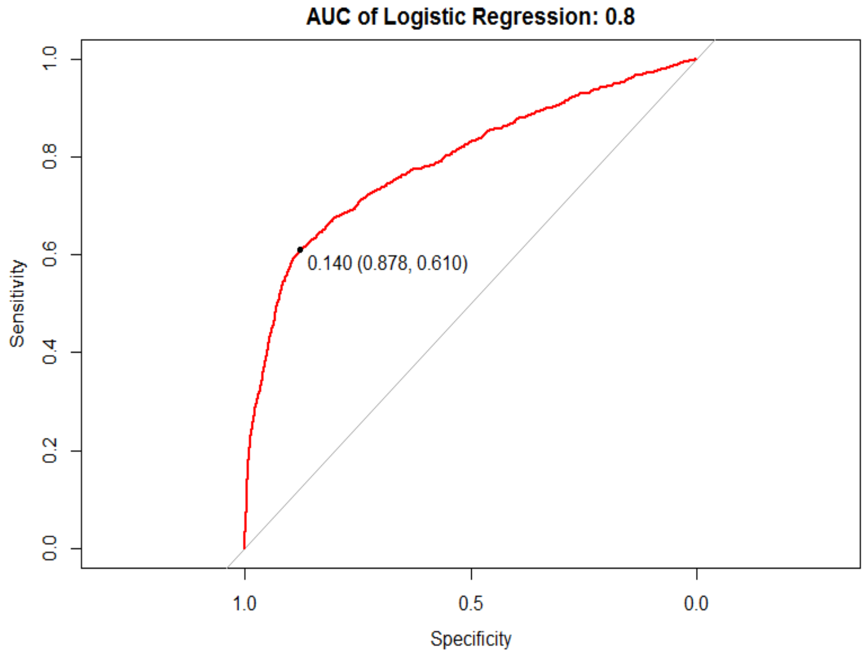



# Model Building

## Logistic Regression

```
Coefficients:
                       Estimate   Std. Error z value Pr(>|z|)
(Intercept)        -226.5262570   33.6341096  -6.735 1.64e-11 ***

day_of_weektue        0.0655693    0.0577084   1.136 0.255866
day_of_weekwed        0.1578269    0.0573340   2.753 0.005909 **
campaign             -0.0440027    0.0092673  -4.748 2.05e-06 ***
pdays                -0.0011001    0.0002010  -5.474 4.41e-08 ***
previous             -0.0698795    0.0560122  -1.248 0.212186
poutcomenonexistent   0.4482422    0.0867991   5.164 2.42e-07 ***
poutcomesuccess       0.7975636    0.1965989   4.057 4.97e-05 ***
emp.var.rate         -1.4664054    0.1249431 -11.737  < 2e-16 ***
cons.price.idx        2.0571816    0.2217528   9.277  < 2e-16 ***
cons.conf.idx         0.0286120    0.0070324   4.069 4.73e-05 ***
euribor3m             0.2057426    0.1154538   1.782 0.074744 .
nr.employed           0.0064281    0.0027376   2.348 0.018868 *
defaultunknown       -0.2417188    0.0578827  -4.176 2.97e-05 ***
defaultyes           -8.6330627  113.4654990  -0.076 0.939351
```




**Confusion Matrix**

```
test.prediction    no   yes
             No 10829  1078
             Yes  135   314
```

Few variables were dropped from model, but did not resulted in improved AUC

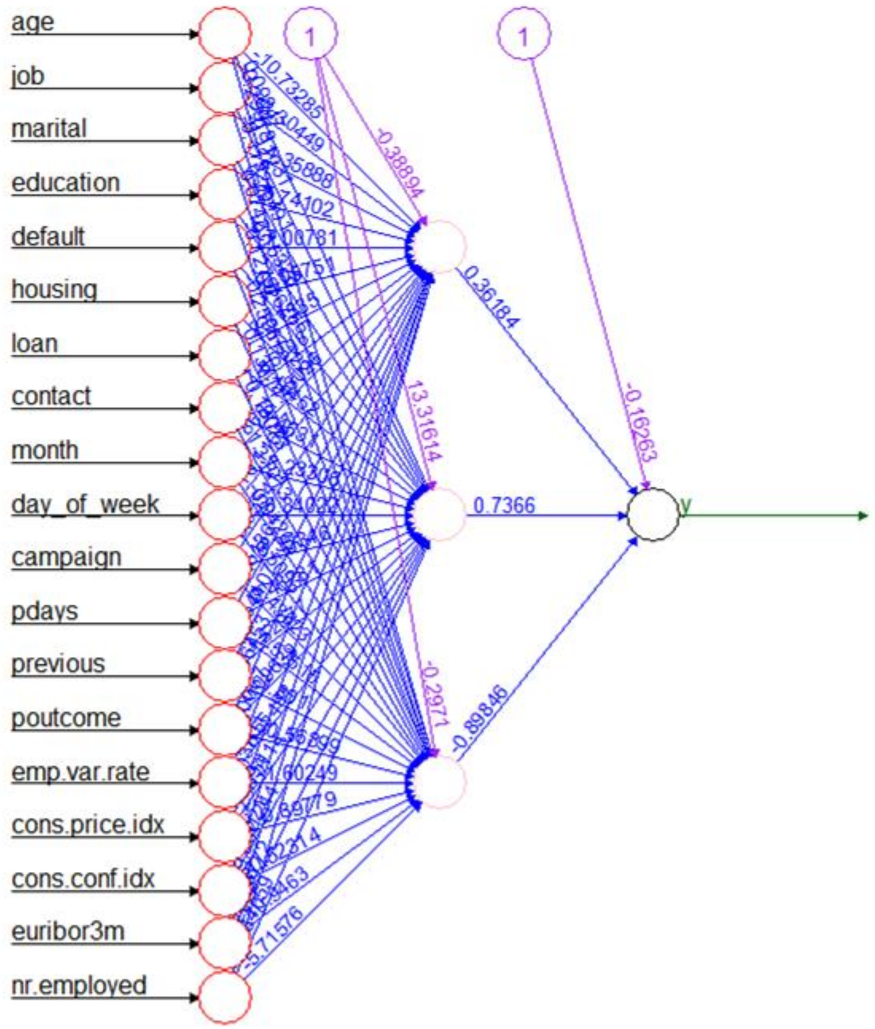
# Model Building

## Neural Network

- Data Split 70% training and 30% Test
- Data normalized using the min-max method and scale the data in the interval [0,1]

| TRUE NEGATIVES | FALSE NEGATIVES | FALSE POSITIVE | TRUE POSITIVE |
|---|---|---|---|
| 8999 | 135 | 869 | 294 |

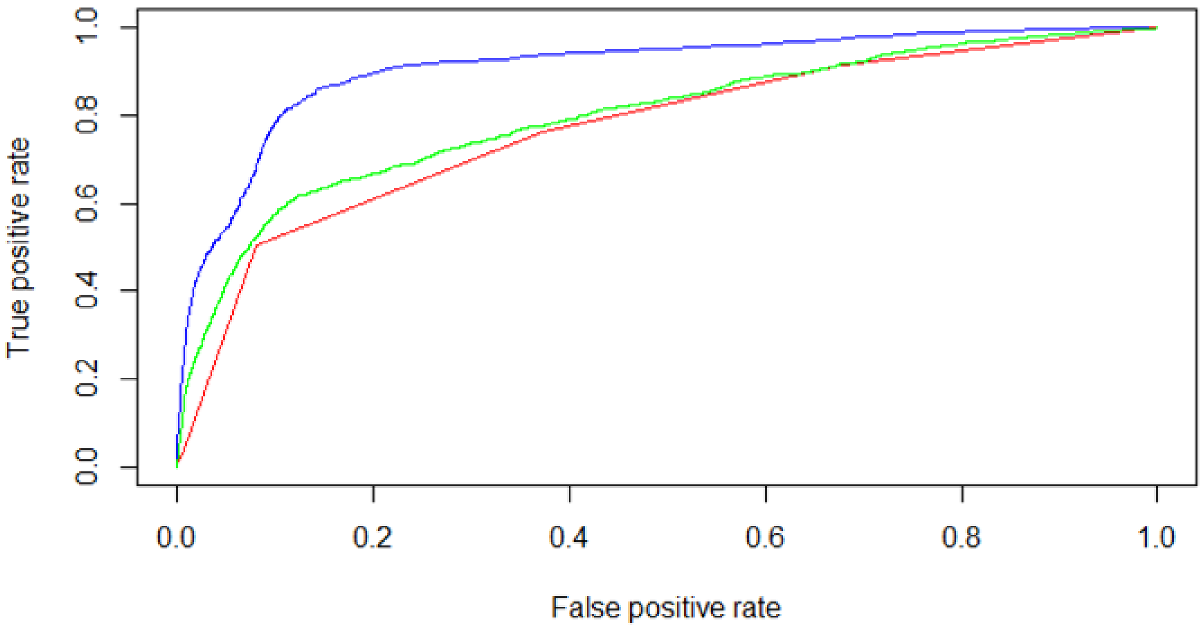
| | |
|---|---|
| ACCURACY | 0.902496 |
| RECALL | 0.685315 |
| PRECISION | 0.252794 |
| F1 | 0.369347 |




- The accuracy is higher, though the Precision is lower due to lesser TP and hIgher FP

# Model Comparison

## Performance Measures




| Models | Decision Tree | Random Forest (RF) | Boosting | Naive Bayes | Logistic Regression | Neural Network |
|---|---|---|---|---|---|---|
| Recall | 0.6 | 0.71 | 0.58 | 0.66 | 0.70 | 0.68 |
| Precision | 0.36 | 0.51 | 0.41 | 0.27 | 0.23 | 0.25 |
| F1 | 0.45 | 0.59 | 0.48 | 0.39 | 0.34 | 0.37 |
| Accuracy | 0.84 | 0.89 | 0.86 | 0.77 | 0.90 | 0.90 |
| AUC | 0.74 | 0.81 | 0.74 | 0.72 | 0.80 | 0.69 |

- Best overall accuracy is for Logistic regression model, excellent in predicting the ‘no’ class
- Positive Predictive Value (PPV) or Precision is highest for Random Forest Model.
- The best F1 measure we could achieve is 0.59, for Random forest hence we choose this model

# Improvements

- Time-series nature of social and economic attributes could be considered to improve the prediction
- Only 5 attributes speak about a customer. More data related to a customer behaviour can improve the prediction
- We could try separating new customers and existing customers

# Conclusion & Suggestions

- Random forest is the best predicting model with a F1 measure of 0.59. Bank can better manage available resources by concentrating on potential customers predicted by this model
- Influential attributes with actionable insight - (pdays, poutcome, day_of_week)
- Months March, September, October are high on conversion. Bank officials can look into the cause of better performance of these months and can try implementing the same in other months

# Thank you !