Bank Marketing Data Set

Introduction:

The Bank marketing dataset has 41188 examples with 20 inputs and 1 output variable. This data is from Portuguese Banking institution. It has numerical as well as categorical attributes and response attribute y denotes client subscribed to term deposit or not (yes or no). The goal is to build models that can predict if client will subscribe to term deposit or not. Since response variable is binary, different classification models will be used incrementally till it gives model with best accuracy.

Dataset:

The dataset is downloaded from UCI Machine Learning Repository and is related to direct marketing campaigns of a Portuguese Banking institution. These campaigns were based on phone calls. Often, more than one calls were done to the same client to access if their product "term deposit" will be subscribed (yes) or not subscribed(no). This dataset is available at http://archive.ics.uci.edu/ml/datasets/Bank+Marketing There were 4 datasets in it from which bank-additional-full.csv is used that has all examples (41188) and 20 inputs ordered by date (from May 2008 to November 2010). There are 20 input variables and 1 output variable (desired target). The dataset had different types of client data like age, job, martial, education, default, housing, loan, contact, month, day_of_week, duration, campaign, pdays, previous, poutcome, em.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr. Employed and one output variable y that denotes if client subscribed to term deposit or not. These dataset attributes denote customer data, socio-economic data, telemarketing data and some other data. Some attributes are numerical, and some are categorical. The dataset was loaded in R Studio and checked for any missing values using is.na function and found that it didn't have any missing values. So, we have a clean dataset.

Attribute Information:

1. **age** – Client Age- (numeric)
2. **job** – Type of Job - (categorical)
   ('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3. **marital** - Client's marital status - (categorical)
   (divorced, married, single, unknown, note: divorced means divorced or widowed)
4. **education** - Client's education - (categorical)
   (basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown)
5. **default** - has credit in default? - (categorical)
   (no, yes, unknown)
6. **housing** - Has housing loan? - (categorical)
   (no, yes, unknown)

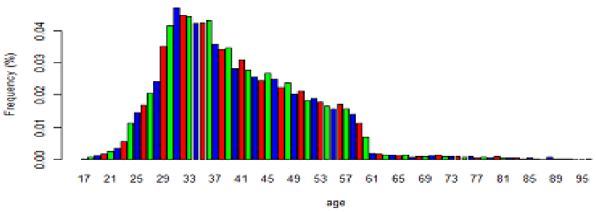
7. **loan** - has personal loan? - (categorical)
   (no, yes, unknown')
8. **contact** – last contact month of year - (categorical)
   (cellular, telephone)
9. **month** - Month of last contact with client - (categorical)
   (January - December)
10. **day_of_week** - last contact day of the week - (categorical)
    (Monday - Friday)
11. **duration** - last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no').
12. **campaign**: number of contacts performed during this campaign and for this client (numeric)
13. **pdays -** number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means clients were not previously contacted)
14. **previous** - Number of client contacts performed before this campaign **-** (numeric)
15. **poutcome** - outcome of the previous marketing campaign - (categorical)
    (failure, nonexistent, success)
16. **emp.var.rate** - Quarterly employment variation rate - (numeric)
17. **cons.price.idx** - Monthly consumer price index - (numeric)
18. **cons.conf.idx** - Monthly consumer confidence index - (numeric)
19. **euribor3m** - Daily euribor 3-month rate - (numeric)
20. **nr.employed** - Quarterly number of employees - (numeric)

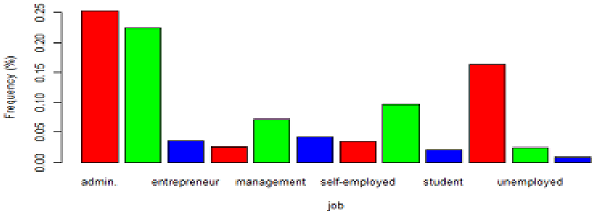
Output variable (desired target) –

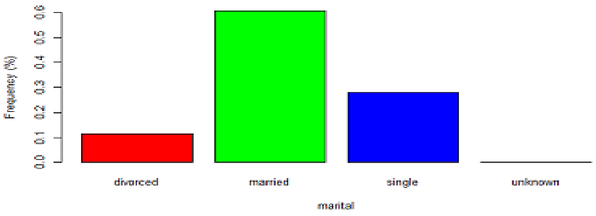
21. **Term Deposit -** has the client subscribed a term deposit? - (binary: 'yes','no')

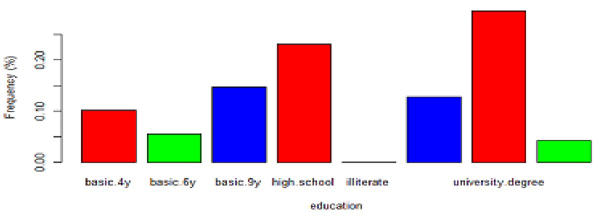
Model Selection and Validation:

Since, dataset is clean and response variable is either yes or no based on the client if he subscribed to the term deposit or not. So, classification models like K-nearest-neighbors (KNN), Classification and regression trees (CART) and C5.0 will be used here. After implementing all these models accuracies will be compared using confusion matrix to determine best model for this dataset. Dataset is first analyzed using bar plots and histograms to understand frequency distribution of the variables. Given below frequency bar plots of some attributes in the data set.









Some attributes needed transformation to numeric class for fitting the models. Using as.numeric function that transformation was done. Then data is split in 8:2 ratio. 80% data is used for training the model and 20% for testing the model. After having training and testing dataset, we can now fit models. KNN, CART & C5.0 models were incrementally used to find best model.

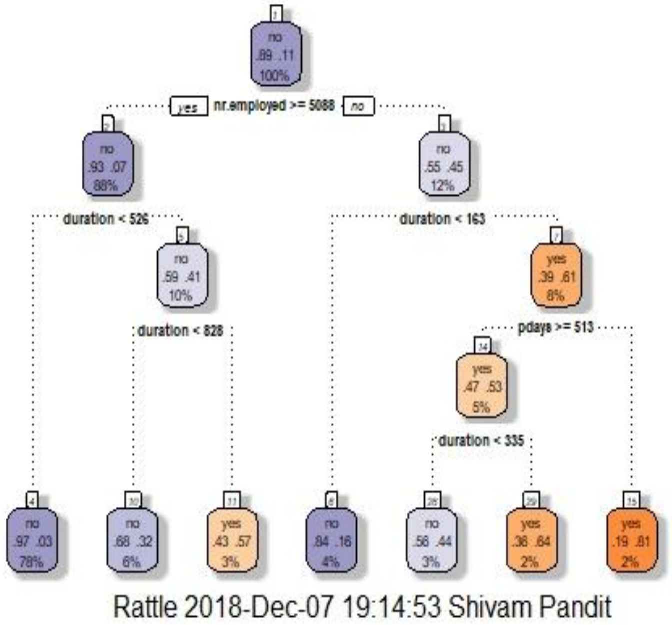
Results:

For Classification and Regression Trees (CART):

Below is the results and confusion matrix for CART
.

- ✓ Accuracy: 0.9134
- ✓ Sensitivity: 0.9629
- ✓ Specificity: 0.5222
- ✓ Misclassification Rate: 1 – Accuracy = 0.0866

```
=======================================
                  predicted default
actual default       no     yes   Total
---------------------------------------
no                 7042     271    7313
                  0.855   0.033
---------------------------------------
yes                 442     483     925
                  0.054   0.059
---------------------------------------
Total              7484     754    8238
=======================================
```



In CART, we got accuracy of 91.34% with misclassification rate of 8.66%

For K- nearest neighbors (KNN):

Below is the results and confusion matrix for KNN.

- ✓ Accuracy: 0.8971
- ✓ Sensitivity: 0.9732
- ✓ Specificity: 0.2951
- ✓ Misclassification Rate: 1 – Accuracy = 0.1387

```
=======================================
                  predicted default
actual default       no     yes   Total
---------------------------------------
no                 7591     305    7896
                  0.801   0.032
---------------------------------------
yes                1009     568    1577
                  0.107   0.060
---------------------------------------
Total              8600     873    9473
=======================================
```

In KNN, we got accuracy of 86.13% with misclassification rate of 13.87%

For C5.0:

Below is results and confusion matrix got for Random Forest.

- Accuracy: 0.9065
- Sensitivity: 0.9573
- Specificity: 0.5049
- Misclassification Rate: 1 – Accuracy = 0.935

```
=======================================
                  predicted default
actual default       no     yes   Total
---------------------------------------
no                 7001     312    7313
                  0.850   0.038
---------------------------------------
yes                 458     467     925
                  0.056   0.057
---------------------------------------
Total              7459     779    8238
=======================================
```

In C5.0, we got accuracy of 90.65% with misclassification rate of 9.35%

Conclusion:

After running multiple models on the dataset, CART is found to give best accuracy of 91.34% with misclassification rate of 8.66%.

References:

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014