Ted Rogers School of Information Technology Management

ITM 618: Business Intelligence and Analytics (Fall 2021)

Final Course Project

Deadline: Thursday December 2, 2021

Marketing for the Banking System

This final course project could be done as a **group** or **individually**.

The problem for this course project is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The **classification goal** is to predict whether the client subscribes a term deposit or not. The target class is the last attribute (*subscribed*) and has two values (*yes* and *no*).

The training set is in *trainset.csv* and the test set is in *testset.csv*. The dataset contains subscribed (yes) and unsubscribed (no) customers.

Attribute Information:

1. **age** (numeric)
2. **job**: type of job (categorical: 'admin.','blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired','self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
3. **marital**: marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4. **education**: (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
5. **housing**: has housing loan? (categorical: 'no','yes','unknown')
6. **loan**: has personal loan? (categorical: 'no','yes','unknown')
7. **contact**: contact communication type (categorical: 'cellular','telephone')
8. **month**: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
9. **day_of_week**: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
10. **duration**: last contact duration, in seconds (numeric).
11. **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
12. **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
13. **poutcome**: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
14. **nr.employed**: number of employees - quarterly indicator (numeric)
15. **Target Attribute**: Subscribed - has the client subscribed a term deposit? (binary: 'yes','no')

Steps

The project involves the following steps:

1. **Data exploration**: try to know data and represents statistics for the important features among the features and the target attribute.
2. **Preprocessing the data**. The goal of this step is to extract features from records in the training set and use these features to test data sets. Note that the data have "**unknown**" values need to be cleaned.
3. **Use classification-learning methods** provided by **R** to learn models from the set of training examples. You can use any of the classification methods (decision tree, KNN, …) for this purpose. You should build at least two different models.
4. **Test the learned models** on the test set and report the testing results.

What to hand in:

1- Your programs (**.R files**) for this project.
2- A **report** that contains: (you can find the report template in D2L.)
   a) The objective of the project (you may use an introduction section to describe it).
   b) What learning methods (in RStudio) you use to learn the models, and the testing results of the learned models on the test data (such as classification accuracy or ROC curve).
   c) Any discussion and conclusion that you find during the project.
3- Overview **Slides**
   a) The slides should present an overview of the project. You can have up to 10 slides. You can find the slide template in D2L.

How to hand in:

1) **Overview Slides**
2) **Project Report**
3) **R program**

You must upload your slides (as a PDF file), R program (as a text file) and your report (as a PDF file) to D2L before the deadline.

Marking Scheme (25 points)

Your project mark will consist of the following components:

- Your slides (5 points)
- Clearness and organization of your report (5 points)
- Soundness and correctness of your solution (as described in your report and implemented in your programs if any)
  - Data preprocessing and cleaning (6 points)
  - Modeling (6 points)
  - Evaluation (3 points)

Technical resources:

You can use the classification packages that was discussed in lab sessions and also use the following links:

- https://www.r-bloggers.com/cheat-sheet-for-prediction-and-classification-models-in-r/
- KNN: https://stat.ethz.ch/R-manual/R-devel/library/class/html/knn.html
- Decision Tree: https://www.r-bloggers.com/using-decision-trees-to-predict-infant-birth-weights/
- Data preprocessing: https://www.r-bloggers.com/preparing-the-data-for-modelling-with-r/