IS 407: Assignment 5

Due by Sunday, December 5, 2021 11:59PM

Complete the homework assignment in Python.

Submit both your .ipynb syntax file and a HTML of your homework assignment.

Note: Make sure to check for missing values. You may use any variables of your choosing, but make sure to explain why you chose these variables.

1. Read in **bank-additional-full.csv** using pandas. The data contains bank marketing data, where rows represent banking clients and the columns represent different bank campaigning information.

A) Create a **logistic regression** for explaining the odds ratios of term deposit.

- Provide the odds ratios, odd ratio 95% confidence intervals, and p-values.
- Explain what variables are contributing to the outcome and what variables are preventing or less likely for the outcome to occur.

B) Create a **logistic regression** to predict the outcome of term deposit.

- Provide the classification report. Report the performance metrics and explain what the metrics means in predicting the outcome.
- Provide the confusion matrix. Use it to calculate the specificity and the negative predictive value. Explain what these metrics mean in predicting the outcome.

C) Create a **Decision Tree** to predict the outcome of term deposit.

- Provide the Decision Tree flowchart. What do you learn from the association rules?
- Provide the classification report. Report the performance metrics and explain what the metrics means in predicting the outcome.

- Provide the confusion matrix. Use it to calculate the specificity and the negative predictive value. Explain what these metrics mean in predicting the outcome.

D) Create use **Feature selection to choose the top 30% variables of importance** for the decision tree predicting term deposit.

- Rerun the model and provide the classification report. Report the performance metrics and explain what the metrics means in predicting the outcome.

E) **Compare your different models from parts B through D**. How do these models perform? Would you feel confident using them for a real-world research question? If so, which model would you use and why?

Data Dictionary

Bank client data:

- Age (numeric)
- Job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
- Marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed)
- Education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
- Default: has credit in default? (categorical: 'no', 'yes', 'unknown')
- Housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
- Loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

Related with the last contact of the current campaign:

- Contact: contact communication type (categorical: 'cellular','telephone')
- Month: last contact month of year (categorical: 'jan', 'feb', 'mar', …, 'nov', 'dec')
- Day*of*week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- Duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if

duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

- Campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- 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)
- Previous: number of contacts performed before this campaign and for this client (numeric)
- Poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Social and economic context attributes

- Emp.var.rate: employment variation rate - quarterly indicator (numeric)
- Cons.price.idx: consumer price index - monthly indicator (numeric)
- Cons.conf.idx: consumer confidence index - monthly indicator (numeric)
- Euribor3m: euribor 3 month rate - daily indicator (numeric)
- Nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

- y - has the client subscribed a term deposit? (binary: 'yes', 'no')