Notes:

1. Answer all questions
2. You will need to submit the code, report and a slide deck summarizing your answers.
3. Note: No data should be submitted. We will use the standard data set as is..
4. You will be randomly given a problem to present in the class on July 8th
5. Deadline: July 7th 11.59 pm

Problem 1

The data for credit card defaults are provided here

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Attribute Information:

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:
X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender (1 = male; 2 = female).
X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status (1 = married; 2 = single; 3 = others).
X5: Age (year).
X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

Goals:

1. Use POWER BI to explore the data. Summarize your observations
2. Clean and pre-process the data if needed
3. Use Logistic regression, Neural Network and Classification trees to build classification models
4. Summarize the performance metrics
    a. Overall Error
    b. Confusion matrix
    c. ROC curves
    d. Lift charts
5. Which model would you choose? Discuss

Copyright 2016 QuantUniversity LLC.
Cannot be reproduced/used without explicit written permission from QuantUniversity

Problem 2

The dataset represents a set of possible advertisements on Internet pages. The attributes encode the geometry of the image (if available) as well as phrases occurring in the URL, the image's URL and alt text, the anchor text, and words occurring near the anchor text. There are two class labels: advertisement ("ad") and not advertisement ("nonad").

Data:

http://archive.ics.uci.edu/ml/machine-learning-databases/internet_ads/

Goals:

1. Use POWER BI to explore the data. Summarize your observations
2. Clean and pre-process the data (Note: 28% of the continuous data is missing). Don't just delete records. Think of a strategy to fill data
3. Use Logistic regression, Neural Network and Classification trees to build classification models
4. Summarize the performance metrics
    a. Overall Error
    b. Confusion matrix
    c. ROC curves
    d. Lift charts
5. Which model would you choose? Discuss

Problem 3

Predicting hourly power generation up to 48 hours ahead at 7 wind farms

This problem mimics the operation 48-hour ahead prediction of hourly power generation at 7 wind farms, based on historical measurements and additional wind forecast information (48-hour ahead predictions of wind speed and direction at the sites). The data is available for period ranging from the 1st hour of 2009/7/1 to the 12th hour of 2012/6/28.

The period between 2009/7/1 and 2010/12/31 is a model identification and training period, while the remainder of the dataset, that is, from 2011/1/1 to 2012/6/28, is there for the evaluation. The training period is there to be used for designing and estimating models permitting to predicting wind power generation at lead times from 1 to 48 hours ahead, based on past power observations and/or available meteorological wind forecasts for that period. Over the evaluation part, it is aimed at mimicking real operational conditions. For that, a number of 48-hour periods with missing power observations where defined. All these power observations are to be predicted. These periods are defined as following. The first period with missing observations is that from 2011/1/1 at 01:00 until 2011/1/3 at 00:00. The second period with missing observations is that from 2011/1/4 at 13:00 until 2011/1/6 at 12:00. Note that to be consistent, only

Copyright 2016 QuantUniversity LLC.
Cannot be reproduced/used without explicit written permission from QuantUniversity

the meteorological forecasts for that period that would actually be available in practice are given. These two periods then repeats every 7 days until the end of the dataset. In between periods with missing data, power observations are available for updating the models.

Data:

https://www.kaggle.com/c/GEF2012-wind-forecasting/data

Goals:

1. Use POWER BI to explore the data. Summarize your observations for each wind generation unit
2. Clean and pre-process the data
3. Use Regression, Neural Network and Regression trees to build prediction models
4. Summarize the performance metrics
    a. MAE
    b. RMS
    c. MAPE
5. Which model would you choose? Discuss
6. Submit the results in the format benchmark.csv

Copyright 2016 QuantUniversity LLC.
Cannot be reproduced/used without explicit written permission from QuantUniversity