Portfolio: Linear Models

Note: you can work solo, or with one other person on this project

Objectives:

- Perform data cleaning and data exploration on medium-sized data sets
- Perform machine learning using linear models
- Evaluate model performance

Turn in:

- Print your Rmd notebooks to pdf, upload to your portfolio, and create a link to it on your index page (if you have a partner, both should upload all pdfs to your portfolio)
- Upload your Rmd print-to-pdfs to eLearning (only one person needs to upload to eLearning)

Instructions:

1. Create an Rmd notebook, name it Regression, with appropriate headings for your name(s) and date. Throughout the notebook use headings to indicate important steps of the assignment and use text blocks to explain what happens in each code block. Write a paragraph explaining in general terms how linear regression works, and what are its strengths and weaknesses.
2. Select a data set from the web that has at least 10K rows and has a target column suitable for regression. List the source of your data in markdown. Perform the following steps with interspersed commentary and code blocks:
    a. Divide into 80/20 train/test
    b. Use at least 5 R functions for data exploration, using the training data
    c. Create at least 2 informative graphs, using the training data
    d. Build a simple linear regression model (one predictor) and output the summary. Write a thorough explanation of the information in the model summary.
    e. Plot the residuals and write a thorough explanation of what the residual plot tells you. Use this source to help you: https://data.library.virginia.edu/diagnostic-plots/
    f. Build a multiple linear regression model (multiple predictors), output the summary and residual plots.
    g. Build a third linear regression model using a different combination of predictors, interaction effects, polynomial regression, or any combination to try to improve the results. Output the summary and residual plots.
    h. Write a paragraph or more comparing the results. Indicate which model is better and why you think that is the case.
    i. Using your 3 models, predict and evaluate on the test data using metrics correlation and mse. Compare the results and indicate why you think these results happened.

3. Create a second Rmd notebook, name it Classification, with appropriate headings for your name and date. Throughout the notebook use headings to indicate important steps of the assignment and use text blocks to explain what happens in each code block. Write a paragraph explaining in general terms how linear models for classification work, and what are the strengths and weaknesses of these linear models.
4. Select a data set from the web that has at least 10K rows and has a target column suitable for classification. List the source of your data in markdown. Perform the following steps with interspersed commentary and code blocks:
    a. Divide into 80/20 train/test
    b. Use at least 5 R functions for data exploration, using the training data
    c. Create at least 2 informative graphs, using the training data
    d. Build a logistic regression model and output the summary. Write a thorough explanation of the information in the model summary.
    e. Build a naïve Bayes model and output what the model learned. Write a thorough explanation of the data.
    f. Using these two classification models models, predict and evaluate on the test data using all of the classification metrics discussed in class. Compare the results and indicate why you think these results happened.
    g. Write a paragraph listing the strengths and weaknesses of Naïve Bayes and Logistic Regression.
    h. Write a paragraph listing the benefits, drawbacks of each of the classification metrics used, and briefly describe what each metric tells you.
5. Upload the print-to-pdfs for both notebooks, and create a link to them on your index page.

Grading Rubric:

Element	Points
Step 1 Regression notebook	70
Step 2 Classification notebook	70
Step 3 Create links to the document and code on the index page	10
Total	150

Grading Rubric:

- 90 and above for exceptional work
- 80-89 for good work
- 70-79 for average work
- below 70 for low quality work

Caution: All course work is run through plagiarism detection software comparing students' work as well as work from previous semesters and other sources.

Caution: All course work is run through plagiarism detection software comparing students' work as well as work from previous semesters and other sources.