Instructions

- This homework assignment is worth 60 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: November 3, 2023 by 11:59 pm.**

Exercise 1

(4 points) What is the difference between simple linear and multiple linear regressions? Be specific.

Exercise 2

(6 points) A study was designed to compare Red Bull energy drink commercials. Each participant was shown the commercials, A and B, in random order and asked to select the better one. There were 140 women and 130 men who participated in the study. Commercial A was selected by 65 women and by 67 men. Find the odds of selecting Commercial A for the men. Do the same for women.

Exercise 3

(4 points) Select the correct statement. In binary (0-1) logistic regression:

(a) The target variable is continuous. That is, the target variable could be any number from a range of values.

(b) The target variable is divided into two equal subcategories.

(c) The target variable takes the values of 0 or 1.

(d) There is no target variable.

(e) All of the above.

(f) None of the above.

Exercise 4

(4 points) Select the correct statement. Logistic regression assumes:

(a) Linear relationship between the predictor variables and the target variable.

(b) Linear relationship between the predictor variables and the logit of the target variable.

(c) Linear relationship between the predictor variables.

(d) Linear relationship between the observations.

(e) All of the above.

(f) None of the above.

Exercise 5

Consider the Default.csv datafile presented in class. In this data set, we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit balance. **In Python**, answer the following:

(a) (3 points) Using the pandas library, read the csv file and called credit_default.

(b) (4 points) Using the where function from the numpy library, create a new variables in the credit_default data-frame called default_numb that takes the value of 1 when default = Yes and 0 when default = No.

(c) (5 points) Using the statsmodels.formula.api, build a logistic regression model in which default_numb is the target variable, balance and income are input variables.

(d) (4 points) Using the model from part (c), predict the likelihood of default of a customer with balance $= \$2,000$ and income $= \$70,000$.

(e) (5 points) Using the statsmodels.formula.api, build a logistic regression model in which default_numb is the target variable, balance, income and student are input variables. Notice that when the input variable is categorical, you need to included with C() in the model. That is, you need to enter C(student) as input in the logistic model.

(f) (5 points) Using the model from part (e), predict the likelihood of default of a customer with balance $= \$2,500$ and income $= \$20,000$ and student = Yes.

Exercise 6

Consider the Churn_Data.csv datafile. This datafile contains information related to a multinational bank. Assume that you work at a multinational bank that is aiming to increase it's market share in Europe. Recently, it has been noticed that the number of customers using the banking services has declined, and the bank is worried that existing customers have stopped using them as their main bank. This bank hired you as a consultant, you are tasked with finding out the reasons behind customer churn and to predict customer churn. The marketing team, in particular, is interested in your findings and want to better understand existing customer behavior and possibly predict customer churn. Your results will help the marketing team to use their budget wisely to target potential churners. **In Python**, answer the following:

(a) (3 points) Using the pandas library, read the csv file and create a data-frame called churn.

(b) (5 points) Using the DecisionTreeClassifier function, build a decision tree model, in which Churn is the target variable, Age, EstimatedSalary, CreditScore, Balance, and NumOfProducts are the input variables.

(c) (4 points) Using the model from part (b), predict the likelihood of churn of a customer with `Age = 50`, `EstimatedSalary = $100,000`, `CreditScore = 600`, `Balance = $100,000`, and `NumOfProducts = 2`.

(d) (4 points). Using the results from part (c), answer the following: assuming the marketing department of this multinational uses 40% as a reference to flag out customers. That is, a customer with likelihood of churn of 40% or greater would be flagged out as a customer who is likely to churn. Would the customer with the characteristics presented in part (c) be flagged out? Explain.