STAT 380 Data Science Through Statistical Reasoning and Computation – Final Project Description (11/18/22)

*Due: By 11:59 PM on Monday, December 12th, 2022*

***Project Groups***

Students may work alone or in a group of up to four people. If you choose to work in a group, your group members will be asked to evaluate your performance/contribution(s) to the project. If you are a person who joins the group, but does not contribute, you should expect your final project grade to be reflective of this.

Groups should ***not*** collaborate with other groups or other people about this project.

***Project Overview***

The final project for this class involves an analysis of video game data that is related to the datasets used for the mini-projects. The final project will involve a combination of using data wrangling, data visualization, and modeling to answer a variety of research questions. In addition to applying modeling techniques that we have learned in class, you will be responsible for picking, researching, and implement one modeling technique that we have not covered.

The goal is to demonstrate that you are proficient at asking meaningful questions and answering them with results of an appropriate data analysis, that you are proficient in using R, and that you are able independently acquire knowledge/implement a method that you were not directly taught.

***Project Datasets***

This problem involves multiple datasets.

1. CODGames\_p1\_380.csv contains information about the results of an online, first-person shooter video game. (Games such as Call of Duty, Fortnite, Apex Legends, Battlefield has been immensely popular in the last 5 years.) The dataset contains information about a single player, who we will call Player 1. In the game, the player is part of a team trying to win an online match against other online competitors. Points are earned for various tasks such as eliminating enemy combatants, collecting items, capturing a location, etc. Each row represents the results of a single, randomly selected, online match in which Player 1 participated.
   * NOTE: When reading in this dataset, make sure the variables names are Map1, Map2, Choice, etc. instead of V1, V2, …
   * We have seen this issue before. To solve it, click the box for Yes for Heading in the Import Dataset interface.
   * This is not the same dataset used in the mini-projects and may require additional cleaning.
2. CODGames\_p2\_380.csv is similar to CODGames\_p1.csv but is for a different player, who we will call Player 2.
   * You will want to combine the information in this dataset with the information in CODGames\_p1\_380.csv.
3. CODGameModes.csv contains information about various types of games available in the game. The dataset contains 3 variables:
   * Mode – The game type
   * ScoreLimit – The maximum number of points that a team can score in a game mode. If a team reaches this limit, the game ends regardless of the time remaining.
   * TimeLimit – The maximum length of the game. If the time limit is reached before either team reaches the score limit, the game ends

***Project Deliverables***

There are several components to the final project:

* A written report that includes the details of your analysis, data visualizations, model comparisons, etc., and answers to the research questions. (One submission per group which includes all group member names.)
  + You have the option to do this in a word processing program (such as Microsoft Word, Google Docs, etc.) or as a Markdown document.
  + Either way, you should have explanations next to relevant plots/figures/tables, etc.
* A submission of R code. (One submission per group which includes all group member names.)
* An evaluation (score out of 10 and a description of the reasoning) for the performance of each teammate. (Each person in a group should submit individually.)

***Goals/Tasks***

NOTE: In the next several days, I will be completing the project. As I do so, I may make adjusted to this section. Although the research questions will not change, I may modify the information that I ask you to provide.

The goal of this project is to demonstrate proficiency in a) the ability to ask a question and answer it with data, b) the techniques we have covered in class, and c) and that you are able independently acquire knowledge/implement a method that you were not directly taught. To do this, you will complete the five tasks listed below.

***Task 1*** (Exploratory Analysis):

Relevant Information: There are a variety games types (GameType variable) within this dataset. The difference between the game types is that players have different objectives for the game. For instance, in the game type “Hardpoint”, teams earn points by capturing and defending a location. In “TDM” teams earn points by eliminating enemy opponents. Regardless of the game type, teams earn points and the game ends when either the time limit has expired or one of the teams reaches the score limit.

Research Question: Which game mode is most likely to reach the score limit?

Notes: This answer requires some data wrangling that may require knowledge that we have not covered. (Again, part of the skillset you are working to develop is learning how to answer questions you have not seen previously.) In particular, there is no distinction between HC – TDM and TDM, no difference between HC – Hardpoint and Hardpoint. Write code to clean the values in the GameType column to reflect this information. Additionally, you will have to perform an appropriate join in order to determine the score limit based on the game type before finding a way to decide if the limit was reached. (Although you could find a way to do this without using a join, you must demonstrate an appropriate join to receive full credit.) Finally, the answer should include a data visualization that depicts the proportion of matches that reach the score limit for each game type.

***Task 2*** (Inference):

Research Question: Suppose we wish to build an appropriate model for modeling the TotalXP variable. We wish to answer the following research questions:

a. Which predictors are associated with the TotalXP?

b. Of the predictors associated with the response, select one of the predictors and explain the relationship between the predictor and TotalXP. (The predictor you select is up to you, but it should be one of the predictors associated with TotalXP.)

Notes: As part of this, you should describe the variables you are considering for inclusion in the model, the type of model, the methodology you will use to determine which predictors are associated with the response (including any details/choices you will make), a depiction of the estimated model, and a description of the relationship between TotalXP and the predictor you have chosen. Be sure that you answer both Part a. and Part b.

***Task 3*** (Prediction):

Relevant Information: In this task, your goal is to compare a variety of classification methods. In particular, you should write your own research question that can be answered by comparing the effectiveness of various classification methodologies. To demonstrate your understanding of these methods, you should implement at least two classification methods from class, one of which must be random forest, and a third method that we ***will not*** cover in class. (The purpose of the using a method we did not cover is I want you to practice learning about a method and its implementation on your own. Basically, find a tutorial that explains the method and how to implement it.) You will then have to compare the results and decide which method was the most effective.

Research Question: Write your own question and be sure that the question and answer are clearly written in your report.

Notes: For this problem, you should provide a brief description of the methods that you will use. (A description is more than listing the name of the procedure.) You will implement and compare the effectiveness of these methods. As part of this process, you will have to make a number of decisions such as whether you will do any data wrangling (maybe you remove partial matches, maybe you create new variables, etc.), which methods will you use, how will you fairly compare the results between methods, which method is best etc. All of these decisions should be included in your report.

FIX THIS: REQUIRE RANDOM FOREST

NOTE1: You will make your life easier if you pick a response with a small number of levels.

NOTE2: When picking the third method (the one we will NOT cover in class), you should not pick decision trees, CART (classification and regression trees), or random forest since these methods will be covered in class.

NOTE 3: You are required to use at least 2 classification that we have covered in class. One of these two methods must be random forest which will be covered during the last two weeks of class.

***Important Dates*:**

* (2 Points) Milestone 1: In Canvas, by Friday, December 2nd at 11:59 PM, submit the name of all group members and evidence of progress towards any of the 3 tasks. This evidence could include your code, plots, part of your paper, etc.  ***It is recommended that you make a submission sooner***.
* (2 Points) Milestone 2: In Canvas, by Wednesday, December 7th at 11:59 PM, submit evidence of progress towards at least 2 of the 3 tasks.  ***It is recommended that you make a submission sooner***.
* (1 Point) Final Report: This is due by the end of the day (11:59 PM State College time) on Monday, December 12th. NOTE: Since this is the first day of finals week, I will consider extensions provided you can provide documentation of conflicts with other finals on the due date December 12th. This is mainly for people who are working individually.

Potential Score Breakdown:

5 Points - Milestones

15 Points – Task 1

15 Points – Task 2

25 Points – Task 3

10 Points – Overall quality of the report

Total (including points from Milestones): 70 Points