MAST 6251: HW2

Dream Job

Due Friday, 2/10 at 11:59pm

1 Background

Typically, my students have an interest in movies or sports analytics. This homework gives you a chance to practice your analysis skills in these areas.

2 Data: Movies or NFL (you pick one)

2.1 The Movies Dataset:

Get the data here: https://www.kaggle.com/rounakbanik/the-movies-dataset. These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

2.2 NFL Play Data:

Get the data here: https://www.nflfastr.com/. The package contains NFL play-by-play data back to 1999.

3 Objectives and Deliverables

- For this homework, **run a logistic regression on one of the two datasets**. As with the last homework, the goal is the same: say something interesting about this data. The main goal is to keep giving you practice on analyzing real datasets and then using the model for some decision or some takeaway.
  - For the movies dataset, the outcome of interest is whether or not a movie is considered "good" or "bad" based on the movie's rating. It's up to you on how you want to decide the rating cutoff for

a "good" movie (e.g., a rating of greater than or equal to 4 out of 5). Some explanatory variables to consider: Who was in the movie? What was the budget? Who produced it? When was it produced? How many other movies were released at the same time? How long was the movie? Etc.

  - For the NFL data, the outcome should be whether the home team won. This dataset has both information from within a game (play by play action) but feel free to summarize at the game level if that is easier. Some explanatory variables to consider: Who is playing who? How late in the season are we? Which season is it? Does a winning streak matter? Etc.

- Part of the purpose of this homework is to get comfortable obtaining, cleaning, and prepping data for an analysis. In the real world, 90% of your time will be spent getting the data ready and the other 10% running the model and summarizing the results.

- Make new variables if needed. For example, in the movie data is this Pixar's first or tenth movie? Does that matter? Or for the NFL data, how many matchups have these teams already had this season? Does it matter? Think about how the original variables can be combined or manipulated to create new variables.

- Think carefully about what potential confounds might be explaining your results. As we started to talk about in class, a common problem with observational data is that it is passive and there may be other variables present that are driving the relationship that comes through in the data. For instance, suppose you could set up any experiment of your choice (even if it might not be possible). How might you design it and what would you be interested in testing? **Please make sure to dedicate at least a paragraph or two to exploring these issues.**

- Think carefully about what your coefficient estimates mean from your model. Calculate marginal effects (the change in the probability that y = 1 given some change in X) and show the impacts visually.

- As with the last homework, complete the analysis in R Markdown and then compile to a PDF output. **There is a 4 page limit on this assignment.**

- Work in groups of up to five students. Make sure all names and SMU IDs are included in the PDF output.