MGSC 310

Prof. Jonathan Hersh

Final Project Instructions

General instructions: Your final project should take a real-world data set and estimate a series of predictive models against this dataset. You should identify a business use-case for the prediction that has a clear business value associated with it. You must estimate one predictive model per student in your group. Groups can consist of 2-4 students. Clearly indicate on the presentation slides which student estimated which model, and that student must present that model.

In general, I want you to show me you can apply the skills learned in this class. You should treat this like a proof of concept, where you are presenting to counterparts at a business who is deciding on whether to implement your proposed analytics model.

November 15th – Due: students must upload to Canvas an outline of their project and sign up for a presentation slot (1:00 class presentation signup**,** 4:00 class presentation signup. Note if it says you don't have access to the Google sheet, sign in with your Chapman account. Please **don't** request access.**).** This outline should include:

a) identify a dataset you will use
b) the names of the students who will be part of your group (at least two, up to four)
c) the methods you will use to analyze your question of interest (one per group member)
d) the outcome you are trying to predict, and what variables you will use to predict it
e) motivation to your project -- as in the business or practical management use case of such a prediction

If you have difficulties identifying a dataset or project please reach out to the TA or the professor. If you have a dataset from an internship, consulting opportunity or job you have been meaning to analyze, you are welcome to use this dataset. After reviewing your projects I may suggest alterations to the project.

November 29th – Due: students must upload a compiled RMarkdown document to Canvas with summary statistics (means, min, max) and at least three plots of interest over the finalized dataset. The finalized dataset should be ready for modeling and analysis.

Nov 29th, **Dec 1st**, **Dec 8th Due: a 15-20 minute** presentation in class. Sections of the presentation should include:

- Motivation/business value of the predictive model
- Description of raw data and summary of data cleaning/feature engineering
- At least two summary plots to describe the dataset to be analyzed
- One predictive model estimated per team member
- Comparison of performance for each of the models
- Conclusion regarding whether the model should be implemented to attain the business objective identified

Final grades will be assigned based on a combination of accurately applying the skills we've learned in class, overall presentation quality and aesthetics, inventiveness/creativity of the project, and appropriately identifying the business value of the predictive model.

Extra credit will be given if: 1) You upload your code and replication files to github; 2) If you build a dashboard to show your results using Tableau, Python Dash, or R Shiny; 3) If you do something particularly creative with regard to feature transformation or use of a novel dataset.

My own recommendation for presentations:

- You should be able to summarize each slide in a sentence or two
- Cut any slide that doesn't add to the story you are telling
- Don't put too much text on each slide
- Make sure labels and titles on figures are readable and in big enough fonts
- No more than two figures per slide (with some rare exceptions)

Useful sites to find datasets:

- Kaggle: https://www.kaggle.com/datasets
- Kaggle: https://www.kaggle.com/annavictoria/ml-friendly-public-datasets
- FiveThirtyEight https://data.fivethirtyeight.com/
- TidyTuesday: https://github.com/rfordatascience/tidytuesday
- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.php
- Careers @ UW: https://careers.uw.edu/blog/2021/10/05/21-places-to-find-free-datasets-for-data-science-projects-shared-article-from-dataquest/
- https://towardsdatascience.com/26-datasets-for-your-data-science-projects-658601590a4c
- https://piktochart.com/blog/100-data-sets/

Interesting Sports Datasets

Sport	Package Name	Link
Baseball	Baseballr	https://billpetti.github.io/baseballr/
NCAA Football	cfbfastR	https://saiemgilani.github.io/cfbfastR/
NCAA Basketball	basketballr	https://github.com/lbenz730/ncaahoopR
Basketball	nbastatR	https://www.rdocumentation.org/packages/nbastatR/versions/0.1.10131
Football	NflFastR	https://www.nflfastr.com/
F1 Racing	formul1data	https://github.com/arkraieski/formula1data