MGSC 310 Portfolio Project

Purpose & Overview

When I graduated college, it took me *eight* months of daily job applications before I was hired to a job with a livable wage.[1] And even when I did enter the workforce, my understanding of how to navigate that environment was severely lacking. The purpose of this course is to ensure that *you* all avoid that fate—that you enter the workforce as prepared as possible, with the skills, resources, and *bona fides*, to pursue your careers. One way to prepare you for the workforce is to ensure you leave with a ***project portfolio*** that you can present to potential employers and cite during job interviews as work experience.

Instructions

Here's how it'll work: each portfolio project will consist of groups of 2-4.[2] Each group will then select a real-world dataset (or datasets) and estimate a series of predictive models against this dataset (or datasets) *to answer a business problem or problems*. **Each group must identify a business case their models are attempting to predict.** These business problems must have a clear business value associated with them.

Each group must estimate one predictive model per student in the group. Clearly indicate on your slides estimated which model. Each student must present the model they created.

In general, I want you to show me that you can apply the skills learned in this class. During your presentations, you will focus ***not only*** on the *technical* features of your model, but the **purpose** behind it. Your goal is to convince me that your proof of concept has merit or that further work is worthwhile to provide value to the business. During your presentation treat the audience as coworkers at your business who are deciding on whether to implement your proposed model.

Due Dates

Tuesday, October 17th :

- Students must have selected their group and written each participant's name into portfolio project presentation Google Sheet **here**. ←
    - Come up with your group company name. That's the company you work at!
    - Write each member's name, one per column.
    - Narrow down the fields you're interested in in the final column. This will help us find useful datasets for you!
- Students must also have also emailed me to arrange to meet (as a group) in a 15-minute conversation about what they are thinking about working on, which data they wish to use, and what business question they are attempting to answer.
    - Meetings will occur between October 17th and November 15th. Meetings will be either over Zoom or on class days, in Beckman 307G (Tuesdays/Thursdays

[1] Let this be a lesson: be careful about *what* jobs you apply to and *how*. Ask me more if you are curious.

[2] Special dispensation will be given to those who wish to be in a group of one, but they must reach out to me and there must be room in time slots I have allocated.

before class).

Tuesday, November 14th:

A) **Students must upload to Canvas an outline of their project.** This outline should:

1. Identify the dataset or datasets you will use.
    a. Assistance identifying a dataset and project will have been addressed during our meetings earlier in the semester.
2. State in no uncertain terms the business question you will seek to answer and **why prediction an answer to that question should bring value to the business**.
3. Declare the statistical/predictive methods you will use to analyze your question of interest (one per group member). (e.g. random forest, linear regression, etc).
4. The target variable you are trying to predict, and what independent variables you will use to predict it.
5. Include the names of the students in the group.
6. The outline should be **professional** in nature. Not simply bullet points, but a PDF or word doc that has clearly had thought put into it. The outline should be no more than 1 page.
    a. This type of open-ended outline is exactly what businesses will ask of you. Come up with quality standards you believe are defensible and hold yourself to them.

B) **Students must also select a presentation slot HERE ←**

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.

Wednesday, November 29th:

- **Students must upload a compiled Jupyter Notebook to Canvas with summary statistics (means, min, max)** and at least *three* Altair plots of interest over the finalized dataset. The finalized dataset should be ready for modeling and analysis.
- Both your raw and *feature* dataset should be accessible to me. I suggest dropping both datasets into a Google Drive folder and sharing it to me. This should be done *by* midnight of the 30th.

Nov 30th, Dec 5th, Dec 7th:

A **15-20 minute** presentation in class. Presentations should include:

- The motivation/business value of the predictive model
- A description of raw data and summary of data cleaning/feature engineering
    - Do not just say “this was the min/max/average of each variable”, explain why it matters
- 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

Grades

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 consideration will be given if:

1) Groups with multiple presenters work off each other's results cohesively.
2) You upload your code and replication files to GitHub with a useful README.md file.
3) You build a dashboard to show your results using something like Streamlit
4) You do something particularly creative during feature transformation, use a novel dataset, or create a particularly powerful or notable statistical model.

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/