Rubric - final report

The goal of this assignment is to practice your writing skills and to document your own work so you can easily catch up with it at some later point.

Your report should be no more than 2000 words (excluding references); we will deduct 1 point for each additional 50 words. Please feel free to use your mid-semester report but revise it if necessary based on the feedback you received.

Please submit your pdf to gradescope by 11:59pm on December 3rd!

Formatting requirements:

1. Your report should start with the title, your name, your affiliation, and a link to your Github repository.
2. Make sure all of your work has been pushed to your repository and that your repository is well-organized.
3. Your text should be concise and clear. No code should be included in the report, and the reader should be able to understand your methods and results without looking at your code.
4. All figures and tables should have captions. Be aware of your overall word count as you include captions.
5. All axes should be labeled and the visualization type should fit the data you plot.
6. Add a references section to cite publications, data sources, any previous work that you mention in your report. References do not count towards the total word count.

The report will be graded on a scale of 50 points.

Report sections

Please include the following sections in your report and do not deviate from this structure.

Introduction
This should be similar in structure to the mid-semester report. In your introduction, make sure to motivate your problem and explain your dataset. **5 points**

EDA
This should be largely the same as the EDA you presented in your midterm presentation. But feel free to update your figures and generate new ones if necessary. **5 points**

Methods
In this section, please explain the data preprocessing and ML pipeline you developed. Make sure to discuss which (un)supervised ML models you used, what parameters you tune and the

values you try. What metric do you use to evaluate your models' performance and why? Measure uncertainties due to splitting and due to non-deterministic ML methods you use (e.g., random forest). In general, explain what considerations went into each step of the pipeline. **15 points**

Results
Discuss how your scores compare to a baseline model (in classification: what is the score if your prediction is the most populous class; in regression: use the R2 score which returns 0 for a constant model that predicts the mean of y, and 1 for a perfect regressor). Calculate global and/or local feature importances and discuss your findings. Translate your results and model interpretations in the context of the problem. That is, how does your machine learning results fits into a business/human/academic context. **10 points**

Outlook
The outlook is the place to describe what else you could do to improve the model or the interpretability, and what are the weak spots of your modeling approach. How would you improve this model? What additional techniques could you have used? What additional data could you collect to improve model performance? **5 points**

References
List the publications, data sources, and any previous work you found. **5 points**

Github repo

Additionally, **5 points** will be given for your github repository. The first thing people inspect in your repo is the readme file. The readme file should give an overview of the project, it states what python version and package versions were used to develop the code so others can run it locally and reproduce your results. There should be a licence file to let people know what they can and cannot do with your code. Github offers a couple of licence options, check those out and decide what's best for you. The repository should have the following directory structure:

```
.
├── data/
├── figures/
├── results/
├── reports/
├── src/
├── .gitignore
├── LICENSE
└── README.md
```

All (raw and preprocessed) data files are in /data, all your generated figures are in /figures, your results (predictions, saved models, etc) should be in \results, the pdf version of your reports

(midterm and final) should be in \reports, and all of your source codes (ipython notebooks or python files) should be in \src.