Submission Details

Due: 11:59pm Friday of last week of finals.

The final draft requires two document:

1. Notebook - the code used to replicate the content of the report
2. Report - written in Microsoft Word, Latex, or alternative word processor

---

WARNING: There is a 50 point deduction if I cannot replicate the results in your report from your notebook using your .pkl file from Box.

There is a **double penalty** for mistakes made in previous FP assignments that you have not resolved.

---

1 Notebook [20 Points]

In your Jupyter Notebook, `Restart Kernel and Run All...` Then on Compass submit both

- the Jupyter notebook
- a PDF export

Hint: you may combine the FP2 and FP3 scripts after removing data wrangling and cleaning and unnecessary code. It should start with the .pkl file.

	Content	Full Points		No Points	
☐	Restarted kernel and ran all	2	yes	0	no
☐	Removed all non-essential code	2	yes	0	no
☐	Submitted Jupyter notebook **and** PDF export	2	yes	0	no
☐	Figure Setup				
	Increase size of figures, axis labels, title, and ticks	4	1 per item	0	none
☐	Print data's head, shape, and info	3	1 per	0	no
☐	Typos/grammatical errors	2	none to some	0	many
☐	Section headers for:				
	EDA, Inference, and Prediction	3	1 per header	0	missing
☐	Comments explaining code	2	enough	0	not enough

2 Report

There should be **no code** in the report. However, some figures and tables will be required. It should be written as a stand-alone from the notebook.

2.1 Format [10 Points]

Page limit: 8 pages. Nothing is graded after the eighth page.

Content	Full Points		No Points	
Report				
Reasonable font size and margins	5	yes	0	no
Typos/grammatical errors	5	none to some	0	many

2.2 Introduction [5 Points]

Provide a paragraph of containing your prediction problem, motivation, and context in whichever order suits you best. Improve upon what you wrote for the EDA assignment.

2.3 Data Wrangling and Cleaning [5 Points]

- Provide the name of the data set(s) and the source from where your data are retrieved [3 points]
- Describe any significant wrangling or cleaning steps: [2 points]
  - Did you have to merge multiple data sets?
  - Did you have to deal with missing data codes?
  - Did you have to convert numerical features to strings?
  - Did you have to do any aggregation?
  - etc.

2.4 EDA [15 Points]

Produce three figures in this section. Figures may have multiple plots if you wish. Your figures need not match those in FP2.

You may substitute either figure 2 xor figure 3 with a table if it helps motivate data decisions made (or not) in your project. *Tables must have a title and aptly named columns/rows.*

Figure 1: Produce a figure that best illustrates your label and (if applicable) any necessary transformations
Response 1: Provide a paragraph containing the interpretation (any transformation necessary?), description, significance, and/or takeaways of Figure 1.
Figure 2 & 3: Produce two additional figures that help motivate data decisions made (or not) in your project.
Response 2 & 3: Provide a paragraph containing the interpretation, description, significance, and/or takeaways of Figure (or table) 2 & 3.

Content	Full Points		No Points	
Figures(/optional one table)				
Titles and labels	9	3 per figure	0	no
Response				
Adequate & accurate descriptions	6	2 per	0	no

2.5 Inference [XX Points]

Using the final results from the Inference section of FP3, produce a table containing:

- Coefficient estimates for top three most significant coefficients
- Coefficient standard errors beneath estimates in parenthesis
- Coefficient significance levels
- Total number of features
- Total number of observations
- The R^2

Provide a paragraph or two with...

- the value of α used in regularization
- the number of features that were removed after regularizing

- an interpretation of the three reported coefficients
- a discussion of any other features of the output you find important for the reader to understand

Example:

	Life Expectancy
log GDP per Capita	$\hat{\beta}_1{}^{***}$
	($\hat{\beta}_1$ std. err)
log Population	$\hat{\beta}_2$
	($\hat{\beta}_2$ std. err)
Inflation Rate	$\hat{\beta}_3{}^{+}$
	($\hat{\beta}_3$ std. err)
Features	9
Obs.	12,345
R^2	0.490

Significance: $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Content	Full Points		No Points	
Table				
Coefficient estimates, standard errors, & significance	3	1 per coef	0	none
#Features, #observations, & R^2	3	1 per	0	none
Response				
α & #removed features	1	both	0	not
Interpretation	3	all correct	0	not

2.6 Prediction [15 Points]

For your individual (partner) project, describe the additional $\geq$ two ($\geq$ three) models you have chosen covering relevant hyperparameters, random states, early stopping, refitting, and any other training steps such that your reader can reproduce it without looking at your notebook.

Content	Full Points		No Points	
Model descriptions	15	7.5 (5) per	0	missing

2.7 Comparison [15 Points]

If your label is...

- Continuous
  - Provide a table of the MSE for all models
- Discrete
  - Provide a confusion matrix for all models (I recommend putting them side-by-side)
  - the accuracy of each model (also report the sensitivity or specificity if you are interest in them)

In both cases, discuss the results of the different models and why some performed well and others did not. *Possible* discussion points:

- What kind of data the respective models perform well on

- The relative flexibility vs. ease of interpretation of each model
- The relative bias/variance of each model

Content	Full Points		No Points	
Model performance diagnostics	9	4.5 (3) per	0	missing
Discussion	6	3 (2) per	0	missing

2.8 Conclusion [5 Points]

Conclude your paper by summarizing the following:

1. prediction problem [1 point]
2. data used to answer research question [1 point]
3. three models evaluated [1 point]
4. best performing model [1 point]
5. call to action [1 point]