Submission [10 Points]

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

- the Jupyter notebook
- a PDF export

IF YOU HAVE CHANGED TOPICS SINCE FP2: Please upload new .pkl file to your box folder.

Content	Full Points		No Points	
Restarted kernel and ran all	4	yes	0	no
Submitted Jupyter notebook	3	yes	0	no
Submitted PDF export	3	yes	0	no

Format [10 Points]

Content	Full Points		No Points	
Removed all non-essential code	2	yes	0	no
Figure Setup				
Increase size of figures, axis labels, title, and ticks	4	1 per item	0	none
Typos/grammatical errors	1	none to some	0	many
Section headers for:				
Inference, Prediction, and Comparison	3	1 per header	0	missing

Inference [20 Points]

Use a reasonable train-test split. Using OLS, binomial logit, or multinomial logit appropriately...

1. Use cross-validation with LASSO to find the optimal α
2. Refit a regularized model on the full training data with optimal α
3. Refit a non-regularized model on the full training data using the features selected from step 2
   - Deal with dummy features accordingly
   - If using multinomial logit, remove a feature if 50% or more of the coefficients have been zeroed
4. Interpret the top three most significant marginal effects

Content	Full Points		No Points	
Reasonable train-test split	3	yes	0	no
LASSO CV optimal α	4	yes	0	no
Refit with optimal α	3	yes	0	no
Refit with selected features	4	yes	0	no
Correct interpretations	6	2 per feature	0	no

Prediction [50 Points]

Models	Classification	Regression
Naïve Bayes	✓	
KNN	✓	✓
SVM	✓	✓
Random Forest	✓	✓
AdaBoost or XGBoost	✓	✓
Neural Network	✓	✓

Print and store the inference model's performance (model 1). If performing classification, you may either tune or use the default threshold in prediction.

If an individual (partner) project, choose two (three) models from the table above.

- Create subsection headers for each chosen model
- Train the models dealing with hyperparameters, random states, early stopping, and/or refitting appropriately
- Print and store your models' performance (if you have a **classification problem**, also print a confusion matrix for 2 points of total performance points)
- 5% bonus if you train and test an additional model-based stacking ensemble on the inference and two (three) chosen models

Individual					Partner				
Content	Full Points		No Points		Content	Full Points		No Points	
Model 1: performance	4	yes	0	no	Model 1: performance	2	yes	0	no
Subsection headers	6	3 per	0	no	Subsection headers	6	2 per	0	no
Model 2: train	10	yes	0	no	Model 2: train	7	yes	0	no
Model 2: performance	10	yes	0	no	Model 2: performance	7	yes	0	no
Model 3: train	10	yes	0	no	Model 3: train	7	yes	0	no
Model 3: performance	10	yes	0	no	Model 3: performance	7	yes	0	no
					Model 4: train	7	yes	0	no
					Model 4: performance	7	yes	0	no
Bonus ensemble	5	yes	0	no	Bonus ensemble	5	yes	0	no

Comparison [10 Points]

In a markdown cell...

- Produce a table over model performance
- Compare all models relative relative flexibility and ease of interpretation
- Explicitly identify the best performing model according to the metric of your choosing

Content	Full Points		No Points	
All components in a markdown cell	3	yes	0	no
Complete table of model performance	3	yes	0	no
Comparisons over all models	3	yes	0	no
Best performing model	1	yes	0	no