Final Exam.

This exam is due **Monday, Dec. 16 at 5 pm sharp** – it cannot be late given the rules on grading. It is open book / note / internet, but you may not speak to any humans about it. This should take 2-3 hours max.

1. You are the head of human resources for a major tech firm and want to improve your hiring practices. The goal is to predict employee performance after 1 year based on variables known at the time of the hire. Define your target variable, the set of variables **X** that you would gather, and the type of model you would run. Examine challenges to inferences you might have and whether or not the assumptions of your model will hold given the data you would be able to gather.
2. Explain the logic behind feature selection – why do it? Why not (instead) simply throw all of the variables **X** at estimating $f(\mathbf{X})$? For a polynomial regression (termwise linear), what assumptions might be "improved" if you perform feature selection? How might feature selection go wrong? And last, if you do not perform feature selection, how else can you avoid some of the problems feature selection is meant to address?
3. Imagine someone hires you to predict the price of a stock tomorrow and you have (reasonably) unlimited data – is this a hard problem or not? Why? Imagine someone else hires you to predict the score of each football game next week. Is this a hard problem or not? Why? Which of these two problems would be more challenging?
4. Attached is a dataset. You should estimate the best model you can using the data provided. Turn in a) your code, b) a short essay (1 page or so) on how confident you are in your results / whether or not you think your model is a good one and c) your final model $f(\mathbf{X})$. For estimating this, I'm assuming you will either use a simple OLS regression, a polynomial regression, or a random forest. In the first two cases, you can simply write down f(X) in the form y = a + b(x1) … In the second case, turn in the decision tree. Choose the right modelling approach – don't pick one and run with it!
   These data are by country and represent how well political parties do in terms of securing **cabinet seats** in a coalition government. If a party receives 0 seats you can assume they are not in the coalition government; more than 0 indicates they are members. All parties prefer more seats to fewer. See: https://en.m.wikipedia.org/wiki/Coalition_government

   Target variable (Y) is cabinet_proportion – the percentage of cabinet seats a party gets.

   Variables **X** in the dataset:

party	party name
seats	raw number of seats in parliament the party won in the last election
sq_cabinet	is the party a member of the status quo (i.e., prior) ruling cabinet
sq_pm	is the party the prior prime minister
election_year	year of election
banzhaf	measure of power of party
shapley	different power measure
splus	different power measure
country	name of country
cabinet_name	name of cabinet
caretaker	is this a caretaker government
cabinet_party	is the party in the cabinet
prime_minister	is the party the new prime minister of the cabinet
left_rightx	party ideology
left_righty	party ideology
cabinet_seats	number of seats the party received for joining coalition
total_cabinet_size	total seats in cabinet
party_name_english	party name
country_id	country id
election_id	election id
seats_share	proportion of seats in parliament the party has
enpp	effective number of parties in system
mingov	minority government or not
bicameral	bicamerial system or not
miw_proportion	different power measure
cabinet_proportion	Y
seats_proportion	proportion of seats in parliament the party has

country_dummy1	country==AUT
country_dummy2	country==BEL
country_dummy3	country==DEU
country_dummy4	country==DNK
country_dummy5	country==FIN
country_dummy6	country==IRL
country_dummy7	country==ISL
country_dummy8	country==ITA
country_dummy9	country==LUX
country_dummy10	country==NLD
country_dummy11	country==NOR
country_dummy12	country==PRT
country_dummy13	country==SWE