To receive full credit for this exam, your D2L submission should be two files:

1. Your solutions for Problems 1 through 3 in the file "Lastname_Firstname_Exam_3.pdf"
2. Your R code used to solve Problem 3 in the file "Lastname_Firstname_Exam_3.R"
    - You must include all non-code answers to Problems 1-3 on the PDF file. Do not expect me find the answers commented in your R code or to run the R code myself. If you do this, I will give you zero credit for that problem.

This exam contains three questions, all of which contain multiple parts. You must show all your work to receive full credit.

This exam is worth 134 points (13.4%) of your grade.

1. (18 Points) Short Answer

   (i) (2 Points) What is the difference between supervised and unsupervised learning?

   (ii) (2 Points) What is the difference between regression and classification? Give an example of a machine learning algorithm to tackle each problem.

   (iii) (2 Points) What is the irreducible error? How do we estimate the reducible error? What is one algorithm that we've discussed to estimate the test error?

(iv) (2 Points) What is the difference between parametric and non-parametric methods?

(v) (2 Points) Discuss the training and testing phases. When do we use our testing set? Can we fine tune our model based off the testing set results?

(vi) (2 Points) What is the majority vote? Discuss it in the context of decision trees and random forests.

(vii) (2 Points) What is the bias-variance tradeoff and why is it important in machine learning? What does generalization mean in machine learning?

(viii) (2 Points) Explain the k-Fold Cross Validation algorithm for $k = 5$.

(ix) (2 Points) Describe the decision tree and random forest algorithms. How does the random forest algorithm improve upon the decision tree? How could we use random forests for estimating causal effects?

2. (20 Points) Suppose we wish to estimate the model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ by ridge regression where we assume $\boldsymbol{\epsilon} \sim \mathbb{N}(\mathbf{0}, \sigma^2 I_n)$.

   (i) (10 Points) Show that the solution to the ridge regression estimator is given by

$$\widehat{\boldsymbol{\beta}}_R = \left(X'X + \lambda I_k\right)^{-1} X'\boldsymbol{y}.$$

(ii) (10 Points) The ridge regression estimator is biased and this bias increases as λ increases. What kind of bias exists and, given the formula for the bias in the lecture slides, why is this so? When would we be fine with this increase in bias from a machine learning perspective? What about from an econometrics perspective?

3. (96 Points Total) Please use the panel data set in the file ECON_418_518_Exam_3_Data.csv to answer the following questions. This data is from a famous study by David Card and Alan Krueger in 1993 titled *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*. The data set was used to evaluate the impact of New Jersey's minimum wage increase in April 1992 on employment in the fast-food industry by comparing it to Pennsylvania, which did not experience a minimum wage change. In the data set:

   - state 0 represents Pennsylvania
   - state 1 represents New Jersey
   - restaurant_id represents a specific fast-food restaurant
   - time_period is either February or November, where February is the time period before the minimum wage increase while November is the time period after the minimum wage increase.
   - total_emp represents the total employment level for a given fast-food restaurant.

   Please load in the data set to R as a data.table.

(i) (10 Points) Given the problem setup, which econometric model should we use to identify the effect of the minimum wage increase in New Jersey? State this econometric model. Does this model include state fixed effects? Why or why not?

(ii) (5 Points) Create two columns. The first being an indicator for if time_period is "Nov" while the other being an indicator for if the state is New Jersey. What is the mean total employment in each state for each time period?

(iii) (10 Points) Estimate a difference-in-difference (DiD) model by computing the difference in sample means. If you use lm(), you will not receive full credit. Provide the interpretation of this result. Provide the formal name for what exactly we are estimating and supplement this with a picture.

(iv) (10 Points) Now estimate the model using lm(). Construct a 95% confidence interval around the estimate of the average treatment effect of the treated (ATT). Show your work (compute the confidence interval by hand and use R to check your work). State the null and alternative hypothesis that the average treatment effect on those in New Jersey is five. Given your confidence interval, what can you say about this hypothesis test (Make sure your statement is precise and uses the terminology we discussed in class)? How about the hypothesis test of the ATT being zero?

(v) (10 Points) Suppose you had data for a few months prior to February 1992. Describe how you would test the fundamental assumption of the DiD estimator.

(vi) (10 Points) Suppose a study came out in April 1992 that stated if you eat fast-food your life span will be reduced by five years. What is this an example of? Would this effect your DiD estimate? If yes, clearly explain why and how your estimate would be impacted. If not, explain why this information is irrelevant.

(vii) (10 Points) Add restaurant fixed-effects to your DiD model and estimate it by using lm(). Does your DiD estimate change? Why might this be the case?

(viii) (10 Points) Do you trust your estimate of the average treatment effect on the New Jersey fast-food industry? Why or why not? If you do, state why. If you do not, state why and how you would alleviate your concerns.

(ix) (10 Points) Suppose we wanted to estimate the effect of the minimum wage increase non-parametrically to avoid making assumptions on the specification of our model. What is one way we discussed in class to do this? What is special or unique about these treatment effects?

(x) (11 Points) Create a GitHub repository following the same process you did for homework three. Name your repository in a such a way that describes your work for this problem. Add the code you used as well as a README.md file describing your work. Attach the link in your D2L submission.