Your name: _____

UCLA Anderson School of Management

Final Exam 2017, MGMTMFE 431.2
Data Analytics and Machine Learning

Prof. Lars A. Lochstoer

You are allowed one letter-sized cheat sheet (both sides) and a calculator. Please be clear and concise. Give your answer on the space allocated between questions. If you need extra space, use the opposite (empty) page or any of the three extra pages at the end of the exam. Questions are numbered 1 through 5. Most questions have subquestions. Make sure you answer them all. The last question is on page 12. Good luck!

1. Consider the following panel regression:

$$e_{i,t} = \mu_t + \delta_1 bm_{i,t-1} + \delta_2 prof_{i,t-1} + \varepsilon_{i,t},$$

where $i = 1, ..., N$ refers to firm and $t = 1, ..., T$ refers to time. The left hand side variable, $e_{i,t}$, is log return on equity. The right hand side variables are log book-to-market and profitability. The time-varying intercept is a *time fixed effect*.

(a) Assume that the cross-sectional average bm and $prof$ equals zero for each time t.

i. Give the analytical expresson for the regession estimate of the time fixed effect at each time t. That is, $\hat{\mu}_t = ?$

ii. Why might a reasearcher want to add time fixed effects (as opposed to a constant μ) to a panel regression?

(b) Assume $\delta_1 = -0.1$ and $\delta_2 = 0.2$. Assume the median analyst expectation of Apple and Google log ROE is 15% and 10%, respectively. Further, assume Apple bm and prof equal -0.5 and 0.5, respectively, while the corresponding numbers for Google are -0.3 and 0.5.

Given the median market expectation (as proxied by the median analyst forecast), what trade is suggested by your model? Explain why. (I am looking for a qualitative description of the trade, not exact number of shares or portfolio weights.)

(c) Assume the following error structure: $cov\left(\varepsilon_{i,t}, \varepsilon_{j,t+k}\right) = 0$ for all k and all $i \neq j$. Any other pairs of residuals are allowed to have non-zero covariance. What kind of standard errors would you apply to this case? A verbal description is sufficient.

2. Assume your research associate has deviced a stock-specific trading signal that she believes is a positive predictor of future cross-sectional stock returns. You want to see if the trading signal has marginal value above and beyond what is given by the firm log book-to-market ratios using Fama-MacBeth regressions.

   (a) Give the Fama-MacBeth regressions you will run. Use clear notation.

(b) Assume you get return data and log book-to-market using StockRetAcct_DT (the dataset used in several homework assignments) in R. Write psuedo-code that loads the data, prepares the variables needed, and runs the Fama-MacBeth regressions in the previous question (a).

(c) Using matrix algebra, give an expression for the portfolio weights at each time t for the portfolio that replicates the estimated Fama-Macbeth coefficient on the trading signal from the previous question. Take care the define any matrices used.

(d) Explain, using equations, how the t-statistic for the coefficients in the Fama-MacBeth are related to the Sharpe ratios on the coefficient-replicating portfolios

(e) Explain why adding industry dummies might improve the Sharpe ratios of your implied trading strategy.

3. You want to predict whether the 6% out-of-the-money 1-month S&P500 put option will end up out of the money or not at expiration. Let x_t be a monthly binary variable that is 1 if the option ended up in the money at expiration (end of month t) and zero otherwise. In addition, you have the options' implied volatility, IV_{t-1}, the implied volatility skew, $Skew_{t-1}$, bid-ask, ba_{t-1}, and open interest, oi_{t-1}. These variables corresponding to the option maturing at time t are known at time $t-1$ (end of month $t-1$). You decide to use a logistic regression.

(a) Give the logistic regression specification you will run. Take care to define all relevant variables.

(b) You decide to add an Elastic Net constraint to the model.

i. Explain clearly what an Elastic Net constraint is.

ii. Explain the K-fold cross-validation procedure often used to find the 'optimal' value of the constraint parameter 'lambda.'

(c) You decide to use an ROC-curve to show the results and to help devise a trading strategy based on this model. Explain:

   i. What an ROC curve is. In particular, what goes on each axis and how do you define these variables.

ii. Assume the curve is above the 45-degree line in the first half of the plot (i.e., when your horizontal axis variable is less than 0.5) and below the 45-degree otherwise. How does this inform you trading strategy? Explain clearly.

4. Textual analysis

   (a) Briefly explain what (i) "corpus", (ii) "stopwords", and (iii) "stemming" refers to in textual analysis.

   (b) Explain the main analysis in the paper "Lazy Prices." In particular, explain in detail one of the text-based metrics the researchers used as a trading-signal, as well as their main findings.

(c) Assume you have run LDA (Latent Dirichlet Allocation) on a corpus with two topics. Consider three documents A, B, and C:

A is 80% topic 1, 20% topic 2.

B is 50% topic 1, 50% topic 2.

C is 30% topic 1, 70% topic 2.

Construct a quantitative measure (feel free to take one you already know) of similarity and apply it to these documents. In particular which two documents are the most similar according to your measure? Which two documents are the least similar? Show your calculations.

5. Below is a data set of alpha's for different fund managers, as well as the percentage management fee each fund charges and the size (Net Asset Value) of the fund. You want to use a decision tree to predict alpha based on the management fee and fund size. Your tree is to have two terminal nodes.

Fund	Alpha (in %)	Fee (in %)	NAV ($ million)
A	-1.5	1.0	500
B	0.7	1.1	100
C	0.9	0.6	400

(a) Using Recursive Binary Splitting, create the decision tree. Draw the tree below. Give the intermediate node and its breakpoint, as well as the two terminal node values. Show your calculations.

(b) Give the qualitative intuition for the decision tree. I.e., what predicts mutual fund alpha? Does it make sense relative to what you know about mutual fund performance?

(c) Briefly explain how *boosting* works to improve the mean squared error of the decision tree's prediction error.

(d) Briefly explain the main difference between linear regression models and decision trees. Use equations to illustrate your logic.

(for extra space)

(for extra space)

(for extra space)