Business Analytics
Sample Final Exam

Solution

Please write your name: _____

Instructions:

1. Write your name above on this Exam paper.

2. Answer all questions in the indicated spaces on this Exam paper. Use the back of the preceding page if you need more space, and clearly indicate where your work can be found.

3. Please show all your work on this Exam paper. Your grade will depend on the clarity of your answers, the detailed reasoning that you used and the correctness of your work and answers.

4. If the answer to a sub-question depends on the answer to a previous one, you will get full credit if you give the right reasoning based on an incorrect previous response. Therefore do not skip sub-questions because you could not answer the previous ones.

5. The exam is open book, open notes. You may use calculators but no computer or mobile phone is allowed.

6. Total possible points = ?

7. Exam duration: 3 hours.

8. Copy and sign the Honor Code: "I will not lie, cheat or steal to gain an academic advantage, or tolerate those who do".

Good Luck!

1. (?? points total) True or False. If the answer is True, you need to indicate this and the question is worth 1 point. If the answer is False, you need to indicate this and give the correct statement. In this case, the question is worth 3 points.

(a) A linear regression model with cross terms helps capture the influence of one covariate on the coefficient of another.

True.

(b) A more complex model will always increase the prediction accuracy.

False. Too complex model may result in over-fitting.

(c) Linear optimization models are more difficult to solve than non-linear models.

False. Linear models are easier than non-linear models.

(d) The conditional ignorability assumption is stronger than randomized experiment.

False. Randomized experiment is stronger than conditional ignorability.

(e) For a logistic regression classification model, we can control the false-positive rate and false-negative simultaneously.

False. If we increase one of them, we have to decrease the other one.

(f) A valid instrumental variable should be uncorrelated with the treatment variable of interest.

False. A valid instrumental variable should be correlated with the treatment variable of interest, and uncorrelated with the potential outcome.

(g) The solution to an optimization model may not be unique.

True.

(h) The parallel trend assumption is the key to the identification strategy of Diff-in-Diff analysis.

True.

2. Short Answer (21 points total)

(a) (7 points) For a classification model, the out-of-sample AUC is 0.76. How can we interpret this number? What is the benefit of AUC over overall accuracy as a measure of model performance?

AUC=0.76 means for two data points, one with positive outcome and the other with negative outcome, the model can predict which is which correctly for 76% of the time. AUC performs more robustly when the data sample is imbalanced.

(b) (7 points) Explain the reason why a well-randomized experiment, there is no need to include the covariates in the regression to estimate the causal effect of the treatment.

For a well-randomized experiment, all the covariates are balanced in the treatment group and the control group. They should not have significant impact on the experiment outcome.

(c) (7 points) Explain in your own words what is bias-variance trade-off. What will happen to the bias and variance of a model if the model complexity increases.

If the model is too simple, it cannot capture the true model-generating process and will have a huge bias. If the model is too complex, it will over-fit the training set and result in a huge variance.

3 (10 points) We have the following training set:

X	1	2	3	4
Y	0	2	1	3

The linear regression model trained on this data sample is

$$Y \approx -0.5 + 0.8X$$

The testing set is:

X	0	5
Y	-1	4

What are the in-sample R^2 and out-of-sample R^2 for this model?

4. (10 points) (a) Assume that a fitted logistic regression model is

$$\mathbb{P}(Y_i = 1|X_i) \approx \frac{\exp(1.5 + 0.7X_{i1} - 1.5X_{i2} + 0.9X_{i3})}{1 + \exp(1.5 + 0.7X_{i1} - 1.5X_{i2} + 0.9X_{i3})}$$

The model predicts that $Y_i = 1$ if $\mathbb{P}(Y_i = 1|X_i) \geq 0.5$ and predicts $Y_i = 0$ if $\mathbb{P}(Y_i = 1|X_i) < 0.5$. Under what condition of $X_i = (X_{i1}, X_{i2}, X_{i3})$ will the model predict $Y_i = 1$? In particular, if $X_i = (1, 2, 2)$, what is the prediction of Y_i with the model?

(b) The confusion matrix of the classification model with the classification threshold $t = 0.5$ is given by the following:

	Predicted $Y = 1$	Predicted $Y = 0$
Actual $Y = 1$	20	10
Actual $Y = 0$	15	25

What is the false positive rate? If we want to decrease the false positive error rate to 0.25, how should we adjust the value of t? If we make this adjustment of t, what will happen to the false negative error rate?

(c) Consider a classification problem with dependent variable $Y \in \{0, 1, 2, 3, 4\}$. We define $p_k = \mathbb{P}(Y = k)$ as the probability that $Y = k$ ($k = 0, 1, 2, 3, 4$), where $\sum_{k=0}^{4} p_k = 1$. The probability distribution of Y is summarized in the following table:

k	0	1	2	3	4
p_k	0.3	0.15	0.1	0.25	0.2

We denote the 0-1 loss of predicting that $\hat{Y} = k$ as $l_k = \mathbb{P}(\hat{Y} \neq Y | \hat{Y} = k)$. What is the value of l_k for $k = 0, 1, 2, 3, 4$?

4. (10 points) Problem Set 4, Q1(c).

5. (5 points) Problem Set 5, Q1(b).

6. (15 points) Problem Set 6, Q1.

7. (15 points) Problem Set 7, Q1.

8. (10 points) Problem Set 8, Q1.

9. (15 points) Problem Set 9, Q2 (Just formulate the problem as an LP, no need to solve it).

10. (15 points) Problem Set 10, Q2 (Just formulate the problem as an IP and discuss whether the cost will increase or decrease, no need to solve it).

11. (20 points) Problem Set 11, Q2.

12. (10 points) Problem Set 12, Q1(a,c). For part (c), just formulate the problem as a non-linear program, no need to solve it.

13. (10 points total) Design a simulation method to sample a random variable X with the following distribution:

$$\mathbb{P}[X = k] = \begin{cases} 0.2 & k = 3 \\ 0.1 & k = 6 \\ 0.25 & k = 7 \\ 0.15 & k = 10 \\ 0.3 & k = 12 \end{cases}$$