ISE-529 Predictive Analytics

Final Examination – August 11, 2022

Instructions

- You are to complete the exam by typing your answers into this PowerPoint as indicated.
- You will have 120 minutes to complete the exam and submit it to GradeScope (in the same manner as done for homework assignments). Late submissions will be penalized.
- The exam is open-book / open-notes. You may consult any resource except another person.
- Good luck!

Problem 1

You are evaluating three different linear model types:

- Traditional linear regression (with OLS fitting)
- Ridge regression model
- Polynomial regression model

You have fit each model three times (with different training partitions) and then used the model to predict the response variable of a test observation whose true value for the response variable is 36. The results of these three tests for each of the three model types is shown on the following page

Problem 1

Prediction on Observation Whose True Value = 35

	Model A	Model B	Model C
Fit using training partition 1	37.4	35.1	36.4
Fit using training partition 2	37.2	32.7	34.9
Fit using training partition 3	37.6	38.5	35.7

Which model type is most likely to be:

- Model A: XXX
- Model B: XXX
- Model C: XXX

Explain your answers briefly and in your own words:

- XXX

Problem 2

Lift

- An income tax agency for a state government is responsible for reviewing and auditing returns to detect fraud. Every year, approximately 200,000 returns are filed and 10% of them are estimated to contain some fraud in some form.
- This year, the agency is using a new classification model to help identify income tax returns to be audited. The model calculated the predicted probability of a return containing fraud based on several predictor values.
- The agency selected the 1000 returns with the highest predicted probability of fraud and audited them. 220 of these audited returns were found to contain fraud.

Problem 2

- What is the lift for this model at the 5% level?
  - XXX

- For possible partial credit, describe below how you got your answer:
  - XXX

Problem 3

ROC Curve Calculations

- Continuing the scenario, from problem 2 (audit prediction classification model), below are 5 observations from this model that contain the classification model's predicted probability (for fraud) and whether or not the observation (tax return) did actually contain fraud:

OBSERVATION	PREDICTED PROBABILITY	ACTUAL FRAUD?
1	8%	N
2	32%	N
3	46%	Y
4	65%	N
5	82%	Y

Problem 3A

ROC Curve Calculations

- Using 4 different classification thresholds (.2, .4, .6, .8), complete the table below:

Threshold	True Positives	False Positives	True Negatives	False Negatives
.2				
.4				
.6				
.8				

Problem 3B

ROC Curve Calculations

- Using the same 4 classification thresholds (.2, .4, .6, .8), complete the table below:

Threshold	Sensitivity	Specificity	True Positive Rate	False Positive Rate
.2				
.4				
.6				
.8				

Problem 3C

ROC Curve Calculations

- Based on the “ROC separation” criteria, which classification threshold should be chosen?
  - XXX

Problem 4

Restaurant Prediction Model

A large restaurant is open for six hours per day (5:00PM – 11:00PM). It has developed a model to predict the number of customers that enter the restaurant during each hour that they are open. The model uses as a predictor the hour that it is open and codes this information as follows:

Hour	Hour code
5:00 – 6:00	0
6:00 – 7:00	1
7:00 – 8:00	2
8:00 – 9:00	3
9:00 – 10:00	4
10:00 – 11:00	5

Problem 4A

Restaurant Prediction Model

- The data scientist building the model has decided to use Poisson Regression. Give two reasons why Poisson regression is preferable to linear regression for this modeling purpose:
  - XXX
  - XXX

Problem 4B

Restaurant Prediction Model

- The model developer has decided to treat the hour predictor as a category and not a measure. Do you think this is the best decision? Explain your answer.
  - XXX

Problem 4C

The modeler encodes the hour in binary “dummy” variables and drops the 0 hour (5:00PM – 6:00PM) so that it serves as the default value (if all other dummy variables are 0, we know that this observation was for the 5:00PM – 6:00PM hour. After fitting the model, we get the following coefficients:

B0 (intercept): 2

B1 (hour 1 coefficient): 2.5

B2: (hour 2 coefficient): 3

B3: (hour 3 coefficient): 3.2

B4: (hour 4 coefficient): 2.4

B5: (hour 5 coefficient): 1

Problem 4C

- What is the expected average number of customers during hour 3 (8:00PM – 9:00PM)?
  - XXX
- What is the expected total number of customers between 6:00PM and 9:00PM
  - XXX
- How many more customers are expected during hour 2 (7:00PM – 8:00PM than hour 0 (5:00PM – 6:00PM) as a factor of the hour 0 customer level?
  - XXX

Problem 5

Classification Tree Model

- A hotel chain is creating a decision tree model to predict whether or not a customer is likely to be a return visitor based on two categorical predictors:
  - X1: Customer survey (was the customer's overall satisfaction with the visit is Low, Medium, or High)
  - X2: Discount offered (did the hotel offer the customer a discount on their next visit: yes/or

Problem 5

Classification Tree Model

- 10 observations of these two predictors and the customer response (return visit = yes/no) are shown below:

Satisfaction	Discount?	Return visit?
High	Yes	Yes
Medium	No	No
Medium	Yes	Yes
Low	No	No
High	No	Yes
Medium	Yes	Yes
High	No	Yes
Low	No	No
Medium	Yes	Yes
High	No	Yes

Problem 5

Classification Tree Model

- Complete the table below showing the candidate cut-points for the first branch in the binary decision tree (add rows as needed):

Predictor	Cutpoint	Misclassification Rate (for both nodes combined after the split

Problem 5

Classification Tree Model

- Based on your results on the previous page, where would you make the first decision tree cut?
  - XXX