GTx ISYE6501x

SP23: Introduction to Analytics Modeling

Help joelquek

Course Progress Dates Discussion

Course / Final Quiz / Final Quiz

Previous ✓ ✓ Next

Verified Learners

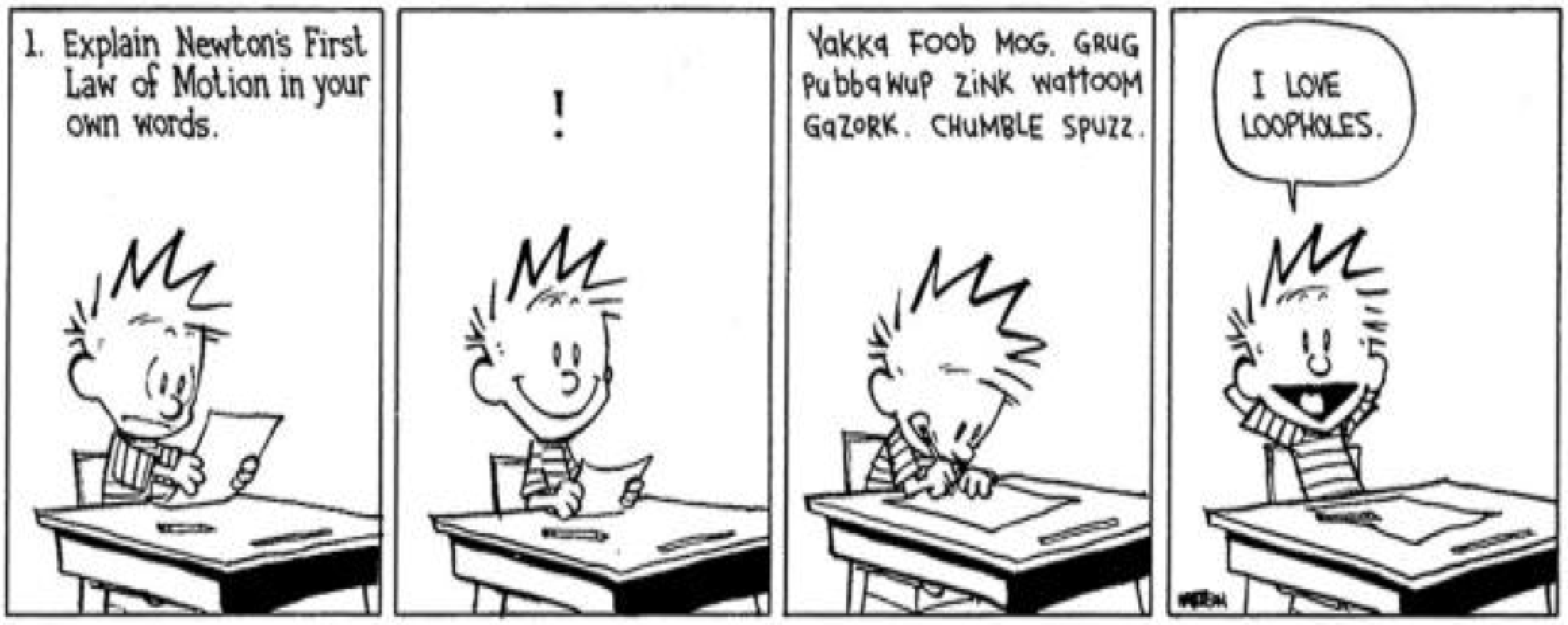
Bookmark this page

Final Quiz due May 1, 2023 14:00 +08 Completed

210 Minute Time Limit

Instructions

- **Work alone.** Do not collaborate with or copy from anyone else.
- You may use any of the following resources:
  - Two sheets (both sides) of handwritten (not photocopied or scanned) notes
  - Blank scratch paper and pen/pencil
- If any question seems ambiguous, use the most reasonable interpretation (i.e. don't be like Calvin):


1. Explain Newton's First Law of Motion in your own words.
!
Yakka Foob Mog. Grug Pubbawup Zink Wattoom Gazork. Chumble Spuzz.
I LOVE LOOPHOLES.

- **If you experience any technical issues (i.e. Math Processing Error), please save your current selected answers and refresh the page. If the issue persists, then please finish the exam and let the Instructors know about the issue in a private Piazza post afterwards.**
- Good Luck!

This is the beginning of the Final Quiz. Please make sure that you submit all your answers before the time runs out. Once you submit an answer to a question, you cannot change it. There is no overall Submit button.

After submitting all answers, please click the "End my Exam" button, above, before exiting from ProctorTrack to complete your exam.

Information for Question 1

There are eight questions labeled "Question 1." Answer all eight questions. For each of the following eight questions, select the type of problem that the model is best suited for. For each question there may be more than one correct answer; you need only choose one. Each type of problem might be used zero, one, or more than one time in the eight questions.

Question 1

0.5/0.5 points (graded)

Select the type of problem that stepwise regression is best suited for. There may be more than one correct answer; you need only choose one.

○ Classification

○ Clustering

‹ Previous

Next ›

○ Experimental design

○ Prediction from feature data
✓

○ Prediction from time-series data

● Variable selection

✱

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 1

0.5/0.5 points (graded)
Select the type of problem that exponential smoothing is best suited for. There may be more than one correct answer; you need only choose one.

○ Classification

○ Clustering

○ Experimental design

○ Prediction from feature data

● Prediction from time-series data

○ Variable selection

✓

Submit You have used 1 of 1 attempt

Question 1

0.5/0.5 points (graded)
Select the type of problem that k-means is best suited for. There may be more than one correct answer; you need only choose one.

○ Classification

● Clustering

○ Experimental design

○ Prediction from feature data

○ Prediction from time-series data

○ Variable selection

✓

Submit You have used 1 of 1 attempt

Question 1

0.5/0.5 points (graded)
Select the type of problem that factorial design is best suited for. There may be more than one correct answer; you need only choose one.

○ Classification

○ Clustering

● Experimental design

○ Prediction from feature data

○ Prediction from time-series data

○ Variable selection

✓

Submit You have used 1 of 1 attempt

Question 1

0.5/0.5 points (graded)
Select the type of problem that GARCH is best suited for. There may be more than one correct answer; you need only choose one.

○ Classification

○ Clustering

○ Experimental design

○ Prediction from feature data

● Prediction from time-series data

○ Variable selection

✓

Submit You have used 1 of 1 attempt

Question 1

0.5/0.5 points (graded)
Select the type of problem that linear regression is best suited for. There may be more than one

correct answer; you need only choose one.

○ Classification

○ Clustering

○ Experimental design

● Prediction from feature data

○ Prediction from time-series data

○ Variable selection

✓

Submit You have used 1 of 1 attempt

Question 1

0.5/0.5 points (graded)
Select the type of problem that lasso regression is best suited for. There may be more than one correct answer; you need only choose one.

○ Classification

○ Clustering

○ Experimental design

○ Prediction from feature data
✓

○ Prediction from time-series data

● Variable selection

✱

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Question 1

0.5/0.5 points (graded)
Select the type of problem that a support vector machine is best suited for. There may be more than one correct answer; you need only choose one.

● Classification

○ Clustering

○ Experimental design

○ Prediction from feature data

○ Prediction from time-series data

○ Variable selection

✓

Submit You have used 1 of 1 attempt

Information for Question 2

There are eight questions labeled "Question 2." Answer all eight questions. For each of the following eight questions, select the type of analysis that the model is best suited for. For each question there may be more than one correct answer; you need only choose one. Each type of analysis might be used zero, one, or more than one time in the eight questions.

Question 2

0.625/0.625 points (graded)

Select the type of analysis that exponential smoothing is best suited for. There may be more than one correct answer; you need only choose one.

○ Using feature data to predict the amount of something two time periods in the future

○ Using feature data to predict the probability of something happening two time periods in the future

○ Using feature data to predict the whether or not something will happen two time periods in the future

● Using time-series data to predict the amount of something two time periods in the future

○ Using time-series data to predict the variance of something two time periods in the future

✓

Submit You have used 1 of 1 attempt

Question 2

0.625/0.625 points (graded)

Select the type of analysis that ARIMA is best suited for. There may be more than one correct answer; you need only choose one.

○ Using feature data to predict the amount of something two time periods in the future

○ Using feature data to predict the probability of something happening two time periods in the future

○ Using feature data to predict the whether or not something will happen two time periods in the future

● Using time-series data to predict the amount of something two time periods in the future

○ Using time-series data to predict the variance of something two time periods in the future

✓

Submit You have used 1 of 1 attempt

Question 2

0.625/0.625 points (graded)

Select the type of analysis that logistic regression is best suited for. There may be more than one correct answer; you need only choose one.

○ Using feature data to predict the amount of something two time periods in the future

● Using feature data to predict the probability of something happening two time periods in the future

○ Using feature data to predict the whether or not something will happen two time periods in the future

○ Using time-series data to predict the amount of something two time periods in the future

○ Using time-series data to predict the variance of something two time periods in the future

✓

Submit You have used 1 of 1 attempt

Question 2

0.0/0.625 points (graded)

Select the type of analysis that k-nearest-neighbor classification is best suited for. There may be more than one correct answer; you need only choose one.

● Using feature data to predict the amount of something two time periods in the future

○ Using feature data to predict the probability of something happening two time periods in the future

○ Using feature data to predict the whether or not something will happen two time periods in the future ✓

○ Using time-series data to predict the amount of something two time periods in the future

○ Using time-series data to predict the variance of something two time periods in the future

✗

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 2

Question 2

0.625/0.625 points (graded)
Select the type of analysis that a support vector machine is best suited for. There may be more than one correct answer; you need only choose one.

- ○ Using feature data to predict the amount of something two time periods in the future
- ○ Using feature data to predict the probability of something happening two time periods in the future
- ● Using feature data to predict the whether or not something will happen two time periods in the future
- ○ Using time-series data to predict the amount of something two time periods in the future
- ○ Using time-series data to predict the variance of something two time periods in the future

✓

Submit You have used 1 of 1 attempt

Question 2

0.625/0.625 points (graded)
Select the type of analysis that k-nearest-neighbor regression is best suited for. There may be more than one correct answer; you need only choose one.

- ● Using feature data to predict the amount of something two time periods in the future
- ○ Using feature data to predict the probability of something happening two time periods in the future
- ○ Using feature data to predict the whether or not something will happen two time periods in the future
- ○ Using time-series data to predict the amount of something two time periods in the future
- ○ Using time-series data to predict the variance of something two time periods in the future

✓

Submit You have used 1 of 1 attempt

Question 2

0.625/0.625 points (graded)
Select the type of analysis that linear regression is best suited for. There may be more than one correct answer; you need only choose one.

- ● Using feature data to predict the amount of something two time periods in the future
- ○ Using feature data to predict the probability of something happening two time periods in the future
- ○ Using feature data to predict the whether or not something will happen two time periods in the future

○ Using time-series data to predict the amount of something two time periods in the future

○ Using time-series data to predict the variance of something two time periods in the future

✓

Submit You have used 1 of 1 attempt

Question 2

0.0/0.625 points (graded)
Select the type of analysis that a linear regression tree is best suited for. There may be more than one correct answer; you need only choose one.

○ Using feature data to predict the amount of something two time periods in the future
✓

○ Using feature data to predict the probability of something happening two time periods in the future

● Using feature data to predict the whether or not something will happen two time periods in the future

○ Using time-series data to predict the amount of something two time periods in the future

○ Using time-series data to predict the variance of something two time periods in the future

✗

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Question 3

3.0/4.0 points (graded)
Select all of the following that are examples of time-series data.

☐ Features of a kidney transplant recipient (age, height, weight, whether he/she has diabetes, etc.) that might affect survival after the transplant.

☐ Number of days each kidney transplant recipient was on the waitlist before surgery, for all patients in the last 20 years.

☑ Number of kidney transplants each year for the last 20 years.
✱

☐ Fraction of kidney transplant recipients still alive a year after transplant, in each of the last 20 years.
✓

✱

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Question 4

3.0/4.0 points (graded)
Select all of the following reasons that data should not be scaled until point outliers are removed.

□ If data is scaled first, the range of data after outliers are removed will be wider than intended.

☑ If data is scaled first, the range of data after outliers are removed will be narrower than intended.
✱

☑ Point outliers would appear to be valid data if not removed before scaling.

□ Valid data would appear to be outliers if data is scaled first.

✱

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Question 5

4.0/4.0 points (graded)
Select all of the following situations in which using a variable selection approach like lasso or stepwise regression would be important.

☑ It is too costly to create a model with a large number of variables.

☑ There are too few data points to avoid overfitting if all variables are included.

□ Time-series data is being used.

☑ There are fewer data points than variables.

✓

Submit You have used 1 of 1 attempt

Information for Question 6

There are four questions labeled "Question 6." Answer all four questions. For each of the following four questions, select the type of model that the software package is best suited for analyzing. Each type of model might be used zero, one, or more than one time in the four questions.

Question 6

1.0/1.0 point (graded)
Which type of model is ARENA best suited for?

Discrete-event simulation ✓

Discrete-event simulation ✓

Submit You have used 1 of 1 attempt

Question 6

1.0/1.0 point (graded)
Which type of model is R best suited for?

Linear regression ✓

Submit You have used 1 of 1 attempt

Question 6

1.0/1.0 point (graded)
Which type of model is SimPy best suited for?

Discrete-event simulation ✓

Submit You have used 1 of 1 attempt

Question 6

1.0/1.0 point (graded)
Which type of model is PuLP best suited for?

Linear programming (optimization) ✓

Submit You have used 1 of 1 attempt

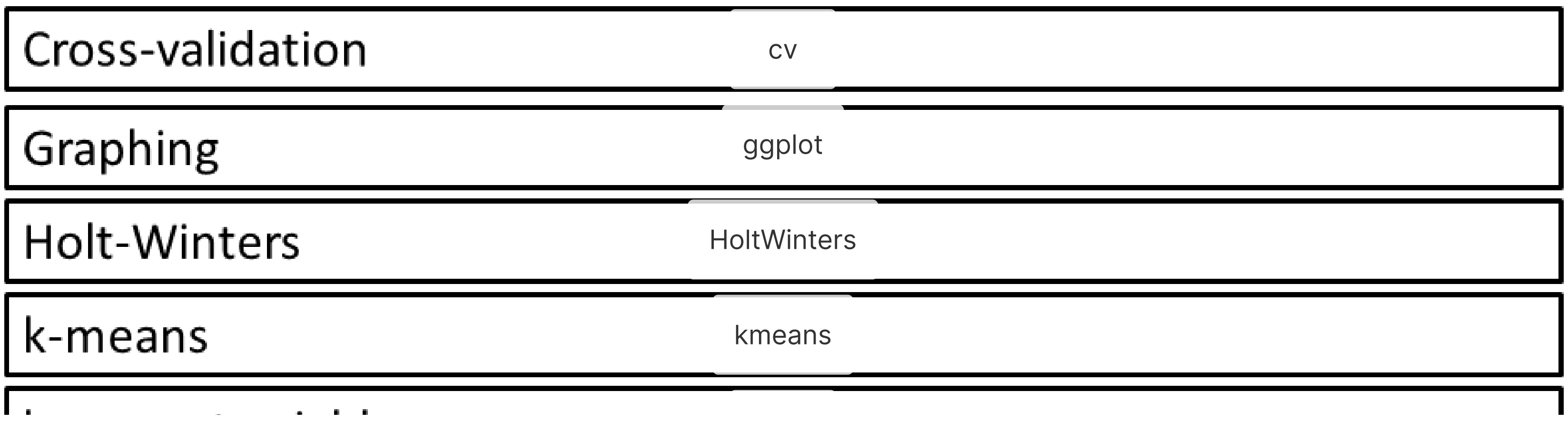
Question 7

6.5/7 point (graded)

⌨ Keyboard Help

For each of the analytics tasks listed below, drag to it the R function(s) that do it. If there is a function that does not do any of the tasks below, then don't drag it anywhere; all other functions should be used.

FrF2



Cross-validation cv

Graphing ggplot

Holt-Winters HoltWinters

k-means kmeans

k-nearest neighbor	kknn
Linear regression	glm lm
Make predictions from models	predict
PCA	prcomp
Random forest	randomForest
Scale data	scale
Support vector machine	ksvm
Train various models	train

Submit You have used 1 of 1 attempts.

Reset Show Answer

FEEDBACK

✓ Correctly placed 12 items

✗ Misplaced 1 item

✱ Final attempt was used, highest score is 6.5

Good work! You have completed this drag and drop problem.

Question 8

2.0100000000000002/3.0 points (graded)

The following process was followed to predict sales of a product each month for the next three years:

1. Split past sales data randomly into three sets: training, validation, and test.
2. Build 20 different models using the training data.
3. Evaluate all 20 models on the validation data.
4. Select the model that performed best on the validation data.
5. Evaluate the selected model on the test data.
6. Use the selected model to predict monthly sales for the next three years based on real-time data, and observe its true performance.

Select all of the following that are true.

☑ Every model's expected performance on **training data** will be better than its expected performance on the **validation data**, because the model fits partly to random patterns in the training data.
✱

☑ The selected model's expected performance on **test data** will be worse than its expected performance on the **validation data**, because there is a selection bias: the selected model is more likely to have better-than-average performance on random patterns in the validation data.
✱

☑ The selected model's expected performance on **test data** must be better than its observed performance on **real-time data**, because the training data and test data were taken from the same population, but the real-time data might be different.

✱

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 9

4.0/4.0 points (graded)

A negative correlation has been observed between selfishness and income (more-selfish people have lower income, and vice versa -- this is the opposite of what most people expect).

Based on that observed negative correlation, select all of the following statements about the direction of causality between selfishness and income that are true.

- ☐ Selfishness causes lower income: Selfishness is a negative people skill, which in turn leads to lower income.
- ☐ Lower income causes selfishness, and higher income causes non-selfishness: People with less money are less likely to give to others, and people with more money are more likely to give to others.
- ☐ Both selfishness and lower income are positively correlated with another factor, which causes both.
- ☑ Can't tell without more analysis.

✓

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 10

4.0/4.0 points (graded)

Select all of the following situations where imputing missing data is probably better than including a "data missing" binary variable.

- ☐ 50% of the data points have missing values for this variable, and you believe that points with missing data have a different distribution of values from points where data is present.
- ☐ 50% of the data points have missing values for this variable, and you cannot build a good predictive model for the missing data.
- ☑ 2% of the data points have missing values, and you can build a good predictive model for the missing data.
- ☐ 2% of the data points have missing values, and you cannot build a good predictive model for the missing data.

✓

Submit You have used 1 of 1 attempt

Information for Question 11

Information for Question 11

There are four questions labeled "Question 11." Answer all four questions. For each of the following four questions, select the model that is more directly appropriate. Assume you have a relevant set of predictor data to use. Each type of model might be used zero, one, or more than one time in the four questions.

Question 11

1.0/1.0 point (graded)

Which model is more directly appropriate to estimate the likelihood that a specific apple tree will produce more than 30 apples this year?

Logistic regression ✓

Submit You have used 1 of 1 attempt

Question 11

1.0/1.0 point (graded)

Which model is more directly appropriate to forecast the number of hot dogs that will be sold at a baseball game?

Linear regression ✓

Submit You have used 1 of 1 attempt

Question 11

1.0/1.0 point (graded)

Which model is more directly appropriate to estimate the probability that a specific online auction will have a winning bid above $86?

Logistic regression ✓

Submit You have used 1 of 1 attempt

Question 11

1.0/1.0 point (graded)

Which model is more directly appropriate to predict the price of a house a year from now?

Linear regression ✓

Submit You have used 1 of 1 attempt

Question 12

2.0100000000000002/3.0 points (graded)

Select all of the following situations where a supervised learning model (like classification) is more directly appropriate than an unsupervised learning model (like clustering).

☑ For each data point, the response is known

☑ For each data point, the response is known.
✱

□ For each data point, the response is not known but an expert has provided an estimate of the response.
✓

□ For each data point, the response is not known and there is no expert estimate.

✱

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Question 13

4.0/4.0 points (graded)
A hospital has collected data on how long hip replacement surgery patients have required before regaining nearly-full motion without pain, as well as attributes of each patient (age, height, weight, pre-surgery range of motion, other medical conditions, etc.). Now, the hospital wants to use that data to predict recovery time for a new patient.

Select all of the following situations where a classification model is more directly appropriate than a linear regression model.

□ The hospital wants to estimate the amount of time it will take for the new patient to regain nearly-full motion without pain.

☑ The hospital wants to predict whether or not the new patient will regain nearly-full motion without pain in six months or less.

☑ The hospital wants to predict whether or not the new patient will regain nearly-full motion without pain in six months or less, if he loses 10 pounds (4.5kg) before the surgery.

✓

Submit You have used 1 of 1 attempt

Information for Question 14

There are four questions labeled "Question 14." Answer all four questions. For each of the following four questions, select the model that is more directly appropriate. Assume you have a relevant set of predictor data to use. Each type of model might be used zero, one, or more than one time in the four questions.

Question 14

1.0/1.0 point (graded)
Given the expected performance of thousands of stocks, and the covariances between them, find an investment portfolio with the best mix of expected return and low risk.

Which model is more directly appropriate?

Optimization ⌄ ✓

Submit You have used 1 of 1 attempt

Question 14

1.0/1.0 point (graded)

Given distances and current and predicted travel speeds on each road, find the quickest way to drive from your current location to Georgia Tech if there are no unexpected delays.

Which model is more directly appropriate?

Optimization ✓

Submit You have used 1 of 1 attempt

Question 14

1.0/1.0 point (graded)

Given the weights and volumes of thousands of proposed scientific experiments that could be sent into space on the next private rocket launch, the amount of money each lab has offered to pay for its experiment to be included, and the capacity of the rocket, find the set of experiments that will maximize the income of the company launching the rocket.

Which model is more directly appropriate?

Optimization ✓

Submit You have used 1 of 1 attempt

Question 14

1.0/1.0 point (graded)

Given the rates of people moving from room to room in a museum, times and routes to walk from one room to another, and capacities of rooms and hallways and doorways, find the maximum number of people the museum should allow inside so that congestion is unlikely.

Which model is more directly appropriate?

Simulation ✓

Submit You have used 1 of 1 attempt

Questions 15a-f

15.0/18.0 points (graded)

A medical practice that focuses on hip and knee replacement surgery would like to increase its income by scheduling more surgeries per day. This description is simplified from its real complexity; if you're a medical expert, please do not rely on your expertise to fill in all that extra complexity (you'll end up making the questions more complex than I intended).

Currently, each hip surgery is scheduled for t_h hours and each knee surgery is scheduled for t_k hours, where t_h is the average plus two standard deviations of the length of a hip replacement surgery, and t_k is the same for knee replacement surgery. As a result, there is often a lot of time between surgeries where doctors are "idle" (not doing surgery). On the other hand, surgeries rarely start late.

The medical practice would like to decrease that idle time by fitting more surgeries into each day, while keeping a low probability of surgeries starting late due to a doctor or surgery room being used longer than expected by the previous surgery.

a. Select all of the models/approaches the practice could use to predict the time a specific patient's surgery will take, based on demographic and medical information about the patient, and the doctor's past surgery lengths.

- ☐ ARIMA
- ☐ Exponential smoothing
- ☑ Linear regression
- ☐ Queuing
- ☑ Ridge regression

✓

b. Select all of the models/approaches the practice could use to schedule as many surgeries as possible (with no concern about late starts), given predictions from part a., and information about doctor and surgery room availability.

- ☐ Elastic net
- ☐ k-means
- ☐ Linear regression
- ☑ Optimization
- ☐ Principal component analysis

✓

Now, suppose the practice still wants to do the same thing as in part b, but also have no more than 15% of surgeries start late because a doctor or surgery room is used longer than expected by the previous surgery.

c. Select all of the models/approaches the practice could use to schedule as many surgeries as possible while having no more than 15% of surgeries start late, given predictions from part a. and the distribution of surgery length, and information about doctor and surgery room availability.

- ☐ Louvain algorithm
- ☐ Random linear regression forest
- ☐ Ridge regression
- ☑ Simulation
- ☑ Stochastic optimization

✓

Over time, surgery length might decrease due to improved technology and/or doctors having more experience.

d. Select all of the following models/approaches the practice could use to determine whether there has been a big-enough change that they should re-fit the model in part a. on more-recent data.

☐ A/B testing using surgery time as the response and each doctor as the two options

☑ CUSUM on the differences between predicted and actual surgery times
✱

☑ Logistic regression to estimate the probability that a surgery will take more than t_h or t_k hours

☐ Markov chain, with the number of surgery patients waiting as the states

☐ The same method as in part a., and see if the new model gives significantly different results from the old model
✓

✱

The practice is also considering expanding, by hiring new doctors to meet more demand.

e. Select all of the models/approaches the practice could use to predict future demand for knee and hip replacement surgeries, based on past demand each year.

☑ ARIMA
✱

☑ Exponential smoothing
✱

☐ Fractional factorial design

☑ GARCH

☐ Multi-armed bandit

✱

If the practice hires enough new doctors to meet predicted demand, they also need to know how many new surgery rooms to build.

f. Select all of the reasons that the results obtained using a simulation model might be incorrect.

☑ It is unclear whether the new doctors would attract patients with different average characteristics (e.g., younger doctors might attract younger patients).
✱

☑ Running two replications of the model could lead to very different conclusions, even using the same input data.

☑ The practice has no data on the new doctors' surgery speed.
✱

☑ The simulation cannot capture the inherent variability in surgery times.

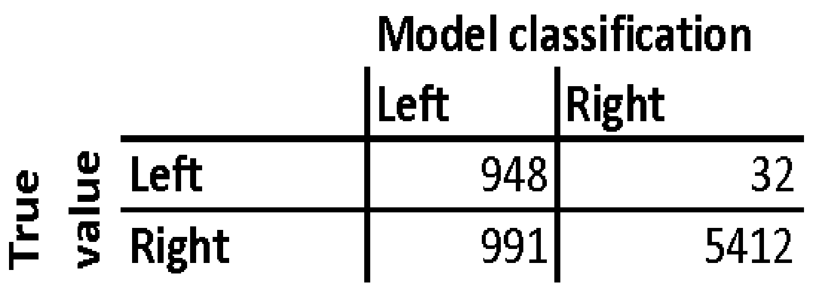
☐ There is no way to determine using a simulation model how many surgery rooms are needed.

✱

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Information for Questions 16a,b



		Model classification	
		Left	Right
True value	Left	948	32
	Right	991	5412

Figure 2. Confusion matrix (Sensitivity 96.7%, Specificity 84.5%)

A support vector machine model has been created to predict whether a person is right-handed or left-handed, based on the person's genetic profile. The figure above shows a confusion matrix of the model's performance on a test data set that it was not trained on.

More Information for Question 16a

There are four questions labeled "Question 16a." Answer all four questions. For each of the following four questions, select the calculation that is most appropriate to support or refute the statement. Each calculation might be used zero, one, or more than one time in the four questions.

Question 16a

1.0/1.0 point (graded)

Which calculation is most appropriate to support or refute the statement "If someone is left-handed, then the model is very likely to predict the person to be left-handed"?

948/(948+32) = 96.7% ✓

Submit You have used 1 of 1 attempt

Question 16a

1.0/1.0 point (graded)

Which calculation is most appropriate to support or refute the statement "If someone is right-handed, then the model is very likely to predict the person to be right-handed"?

5412/(5412+991) = 84.5% ✓

Submit You have used 1 of 1 attempt

Question 16a

1.0/1.0 point (graded)

Which calculation is most appropriate to support or refute the statement "If the model predicts someone to be right-handed, then the person is very likely to be right-handed"?

5412/(5412+32) = 99.4% ✓

Submit You have used 1 of 1 attempt

Question 16a

1.0/1.0 point (graded)

Which calculation is most appropriate to support or refute the statement "If the model predicts someone to be left-handed, then the person is very likely to be left-handed"?

948/(948+991) = 48.9% ✓

Submit You have used 1 of 1 attempt

Question 16b

2.0/2.0 points (graded)

Select all of the following ways that it is reasonable to use this model.

□ Use the model's classification when it predicts left-handedness, but remain undecided when it predicts right-handedness

☑ Use the model's classification when it predicts right-handedness, but remain undecided when it predicts left-handedness

✓

Submit You have used 1 of 1 attempt

Questions 17abcde

8.616/12.0 points (graded)

Every morning at 9:30am, the manager of a fast-food restaurant determines how many hamburgers to pre-make so they are ready to be immediately given to customers at lunchtime. If not enough are pre-made, customers will have to wait a long time in line, and might go to the competing fast-food restaurant next door instead. If too many are pre-made, some will spoil before they can be used. This description is simplified from its real complexity; if you're an expert in the restaurant industry, please do not rely on your expertise to fill in all the extra complexity (you'll end up making the questions below more difficult than I intended).

a. The manager has come up with the following incorrect idea:

GIVEN the past fraction of days π_k that k hamburgers were ordered at lunch, the past overall distributions of (a) arrival rates of customers, (b) the time it takes to give a customer a pre-made hamburger, (c) the time it takes to make and give the customer a newly-made hamburger, (d) customers leaving because of lines, and (e) hamburger spoilage, USE a simulation model TO find the best tradeoff between the expected number of customers served each day and the cost of spoiled hamburgers.

Select all of the statements below that show a reason why the manager's idea is wrong.

☑ The data doesn't include attributes of days (holidays, etc.).

☐ Simulation models can't be used to evaluate tradeoffs.

☑ The simulation model doesn't account for seasonality.

☐ The simulation model doesn't account for random variation.

✓

b. The manager has come up with another incorrect idea:

GIVEN past data on the probability p_{kn} that n hamburgers will be ordered at lunch today if k were ordered yesterday, USE a Markov chain model TO find the steady-state probabilities π_k that there are k hamburgers ordered at lunch in a day. Then, GIVEN those probabilities and the relative costs of losing customers or having hamburgers spoil, USE an optimization model TO determine the most cost-effective number of hamburgers to order each day.

Select all of the statements below that show a reason why the manager's idea is wrong.

☑ The number of hamburgers sold day to day isn't memoryless.
✱

☐ The Markov chain model doesn't account for seasonality.
✓

☐ The Markov chain model doesn't account for attributes of days (holidays, etc.).
✓

☑ The Markov chain model can't link one day to the next.

✱

c. Select all of the possible paths below that could reasonably lead to a good solution.

☐ Predict the lunchtime demand for each day. Find the best tradeoff between the cost of lost customers and the cost of spoiled food, as a function of the probability customers will leave for different line lengths. Then, based on an expert estimate of the probability customers will leave for different line lengths, determine the best number of hamburgers to pre-make.
✓

☐ Predict the lunchtime demand for each day. Estimate the wait time for each customer. Then find the best tradeoff between decreased wait time and the time it takes to pre-make some food.

☑ Predict the lunchtime demand for each day. Estimate the probability of customers leaving because the line is too long, based on demand and service rates. Then find the best tradeoff between opportunity cost of lost customers and cost of spoiled food.
✱

✱

d. Select a set of models from the list below, to create a solution that the manager can put together to determine how many hamburgers to pre-make each day.

☑ GIVEN daily data on sales, weather, holidays, day of week, and number of people who left because the line was too long, USE a random forest model TO estimate demand.

□ GIVEN daily data on sales, weather, holidays, day of week, and number of people who left because the line was too long, USE a clustering model TO estimate demand.

□ GIVEN past data on customers who leave based on time of day and line length, USE an optimization model TO estimate the probability of a customer leaving because the line is too long.

☑ GIVEN past data on customers who leave based on time of day and line length, USE logistic regression TO estimate the probability of a customer leaving because the line is too long.

☑ GIVEN the estimated demand, estimated probability of leaving because the line is too long, cost of spoiled food, and estimated cost of lost customers, USE optimization TO find the number of hamburgers to pre-make that minimizes overall costs.

□ GIVEN an estimate of demand, the probability of a customer leaving because the line is too long, past arrival and service rates, cost of spoiled food, and estimated cost of lost customers, USE a support vector machine model TO find the number of hamburgers to pre-make that minimizes overall costs.

✓

e. Select all of the following complexities that are not accounted for in any of the models in part d.

☑ When the company runs a national hamburger sale, demand will be higher.
✱

☑ The number of hamburgers sold on holidays might be different from the number sold on regular days.

☑ Pre-made hamburgers do not taste as good as freshly-made ones, so selling more pre-made hamburgers one day might decrease demand on future days.
✱

✱

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Questions 18a-d

5.025/6.0 points (graded)

In the United States in 2015, the overall population of 19-24-year-olds (about 27 million people) was approximately 49% women and 51% men. In the US college population of 19-24-year-olds (about 12 million people), 57% of college students were women and 43% were men.

a. To test whether this discrepancy is significant, an analyst wants to use a binomial distribution. What would be an appropriate test?

○ Find the probability of 49% or more "yes" answers from a binomial distribution with n=27,000,000 and p=0.57.

○ Find the probability of 57% or more "yes" answers from a binomial distribution with n=27,000,000 and p=0.57.

○ Find the probability of 49% or more "yes" answers from a binomial distribution with n=12,000,000 and p=0.49.

● Find the probability of 57% or more "yes" answers from a binomial distribution with n=12,000,000 and p=0.49.

✓

b. Select all of the approaches below that might help determine whether there has been a signficiant change in the fraction of college students who are men and who are women over the past 50 years.

□ Classification with each year as a data point, using fraction of college students who are women as the response and the year as the predictor

☑ CUSUM on the fraction of college students who are men, with each year as a data point
✱

☑ Exponential smoothing on the fraction of college students who are women, with each year as a data point
✱

□ Logistic regression with each year as a data point, using fraction of college students who are men as the response and the year as the predictor
✓

✱

One suggested explanation for the discrepancy is that there is a difference between girls' and boys' high school grades, partly due to boys' higher frequency of misbehavior.

c. What nonparametric test could be used to check whether girls' and boys' median high school grades are significantly different?

○ Paired-sample signed rank test

○ McNemar's test

● Two-sample unpaired rank test (Mann-Whitney)

○ One-sample signed rank test

✓

A logistic regression model shows that high school GPA is a significant predictor of whether a person will go to college.

d. Select all of the statements below that could be a causal relationship between high school GPA and college attendance. Base your answer only on the information above and the timing involved. [For the purpose of this question, do not think about whether the statements are true; assume they are true (even if you don't believe them) and determine whether, if true, the statement shows a causal relationship.]

☑ Many colleges are less likely to admit students with a lower high school GPA.
✱

□ Most community colleges will admit any high school graduate.

☑ The same factors that cause boys to have lower high school GPAs might also make them less likely to want to attend college.

☑ High school students who get higher GPAs do so because they are more serious about school, and therefore are more likely to want to attend college.

☑ Colleges believe that a higher high school GPA is a sign that a student is taking school seriously, and colleges prefer to admit serious students.
✱

✱

Data for this question was taken from http://nces.ed.gov/fastfacts/display.asp?id=372, http://nces.ed.gov/pubs2015/2015073.pdf, https://www.census.gov/popest/data/national/asrh/2015/2015-nat-af.html, http://www.apa.org/news/press/releases/2014/04/girls-grades.aspx, and http://economics.yale.edu/sites/default/files/fortin-121108.pdf.

This discrepancy is getting more and more attention in education (and in education analytics); if you have any thoughts about it, please let me know!

Submit You have used 1 of 1 attempt

edX

About
Affiliates
edX for Business
Open edX
Careers
News

Legal

Terms of Service & Honor Code
Privacy Policy
Accessibility Policy
Trademark Policy
Sitemap
Cookie Policy
Do Not Sell My Personal Information

Connect

Blog
Contact Us
Help Center
Security
Media Kit