Final Exam

Due Dec 16 at 11:59pm **Points** 120 **Questions** 24
Available Dec 1 at 5pm - Dec 17 at 12:01am **Time Limit** 240 Minutes
Allowed Attempts 2

Instructions

Complete the exam individually and without third party help. You may use any course materials, books, or internet resources (except asking people or asking on question-and-answer forums such as stackoverflow or quora). You have two attempts to pass the final exam with a minimum of 60%. The highest score counts towards your final grade. The estimated time to complete the test is 3 hours; however, 4 hours is allotted. Once you start the exam, the time starts running. Don't start the exam because you are curious and want a quick peek -- the timer will start running, and I cannot reset it.

Late submissions will not be accepted. Please note that this must be completed by the last day of exams, which is a Saturday, by 11:59pm Eastern.

Take the Quiz Again

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	101 minutes	102.5 out of 120

Correct answers are hidden.

Score for this attempt: **102.5** out of 120
Submitted Dec 14 at 1:38pm
This attempt took 101 minutes.

Question 1 **4.5 / 4.5 pts**

kNN requires specification of a value for *k*. What is a reasonable initial value for *k* for a training data set of *n* cases?

- ◉ √n
- ○ 1
- ○ any value works
- ○ 3
- ○ n

Question 2

4.5 / 4.5 pts

You are performing an in-depth analysis of the mtcars dataset, investigating the relationship between a car's miles per gallon (mpg) and various attributes. To build an accurate linear regression model, you consider including the variables 'weight' (wt) and 'cylinders' (cyl) as predictors. After fitting the model, you obtain the following coefficient values: β0 = 36.02, β1 = -3.72, and β2 = -2.25. Which interpretation of these coefficients is correct?

- ◉ An increase of one unit in weight leads to a decrease of 3.72 units in mpg, holding the number of cylinders constant. The same applies to the cylinders variable with a coefficient of -2.25.
- ○

An increase of one unit in weight leads to a decrease of 36.02 units in mpg, regardless of the number of cylinders. The same applies to the cylinders variable with a coefficient of -3.72.

The model intercept is 36.02, representing the expected mpg when weight and cylinders are both zero. The coefficients β1 and β2 are not interpretable in this context. D) Weight and cylinders have a positive linear relationship with mpg. The coefficients β1 and β2 indicate the increase in mpg for every one-unit increase in weight and cylinders, respectively.

None of the above

Question 3

4.5 / 4.5 pts

Martin Fowler's taxonomy of NoSQL databases refers to key-value and document databases as _____ databases.

Aggregate oriented

Question 4

4.5 / 4.5 pts

The mean grade for last semester was 89.3 with a standard error of 6.72. What is the upper bound of the 95% confidence interval for the mean grade?

102.5

Question 5

4.5 / 4.5 pts

Sandeep found that there's a correlation between starting salary for data scientists and the ranking of the university from which they earned their PhD. He would like to create a model that allows him to forecast the approximate salary he should be getting for his clients when he places them in a position. What approach would you recommend he choose?

- Build a regression model using ranking as the dependent variable
- Calculate the correlation coefficient R and see if it is above 0.6
- (selected) Build a regression model using ranking as the independent variable
- Construct a weighted average model

Question 6

4.5 / 4.5 pts

Which of the following statements about feature normalization are correct? Check all that apply.

- [ ] min-max normalized values are always lower than z-score standardized values
- [x] min-max normalized values are in [0,1]
- [ ] z-score standardized values cannot be negative
- [x] z-score standardized values are computed as (X-μ)/σ

Question 7

4.5 / 4.5 pts

You are performing a linear regression analysis using the mtcars dataset to predict a car's miles per gallon (mpg) based on its horsepower (hp). After fitting the linear regression model in R, you have obtained the following statistics:

- Total Sum of Squares (SST): 1126.05
- Residual Sum of Squares (SSE): 307.88

Calculate the coefficient of determination (R-squared) for this linear regression model using R code. Use two decimal values.

0.73

Question 8

4.5 / 4.5 pts

Match one of the following components or concepts of multiple linear regression with the most appropriate role or interpretation in analyzing the relationship between CO2 concentrations (CO2) and temperature (Temp) in the CO2 dataset, which contains measurements of CO2 concentrations and temperature over time.

Components/Concepts: Multicollinearity best matches which definition below for the CO2 dataset?

Roles/Interpretations:

1. Assessing the proportion of variance in CO2 concentrations that is explained by the linear combination of temperature and other predictor variables, adjusted for the number of predictors.
2. Identifying whether there are high correlations among the predictor variables, which could affect the stability and interpretability of regression coefficients.
3. Evaluating the magnitude and direction of the effect of a one-unit change in temperature on CO2 concentrations while keeping other predictors constant.
4. Introducing a product term to capture the joint influence of temperature and another predictor on CO2 concentrations, accounting for their interaction effect.

Match the Components/Concepts with the Roles/Interpretations:
A) Coefficient of Determination (R-squared) - B) Multicollinearity - C) Adjusted R-squared - D) Interaction Term

- 3
- 1
- 4

- (•) 2

Partial

Question 9

2.25 / 4.5 pts

A line graph of amusement park visitors over the past eleven months shows a weekly downward trend, *i.e.*, the traffic decreases generally week over week. Srujan has developed a forecasting model using a four-week weighted moving average with the most recent time period having a weight of 6 and the prior three weeks have weights of 3, 2, and 1. Which of the statements below are most likely accurate? Check all that apply.

- [ ] When a downward trend is observed, the weight for the most recent time period must be smaller than all others.
- [x] The forecast does not take trend into account.
- [ ] The forecast is never lower than the lowest of number of visitors for the past four weeks.
- [ ] The statistical significance is likely above 0.05.
- [ ] The forecast has a large MAD.

Question 10

4.5 / 4.5 pts

One of the main drawbacks of relational (SQL) databases is that they _____.

- cannot store network or graph data
- use vendor specific forms of SQL dialects
- don't scale well to distributed use across clusters
- cannot be used for analytics
- manage concurrency poorly
- require joins of tables

Question 11

4.5 / 4.5 pts

What is the *slope* of the simple linear regression model for forecasting sales for the next year based on the **attached data set (https://northeastern.instructure.com/courses/157427/files/22232139/preview)** ? Enter the value rounded to the nearest whole number.

178,295

Question 12

4.5 / 4.5 pts

Which of these measures can be used to establish fit of a model? Check all that apply.

- [x] Mean Absolute Deviation
- [x] R^2 or Adjusted R^2
- [x] Mean Squared Error
- [x] Mean Absolute Percent Error
- [ ] Tracking Signal

Partial

Question 13

1.5 / 4.5 pts

Which of the following tasks are done during the data preparation phase in CRISP-DM? Check all that apply.

- [x] construct data
- [x] select data
- [x] verify data quality
- [x] explore data
- [x] clean data

Question 14

4.5 / 4.5 pts

Given vectors v=<3,9,1,-4> and w=<-4,-5,0.5,8>, what is their Euclidean Distance? Enter the value below rounded to two digits after the decimal point. Do not enter any additional text, *e.g.*, enter 4.34.

19.73

Question 15

4.5 / 4.5 pts

Given the **attached data set (https://northeastern.instructure.com/courses/157427/files/22232227/preview)**, what is the mean squared error for a simple linear regression model that is capable to predicting the value for the next time period?

13,383,171,520

Question 16

4.5 / 4.5 pts

To predict a nominal (categorical, class) feature based on numeric features, which of the following models generally works best?

- prediction interval

any of the above would work

multiple regression

exponential smoothing

none of the above are appropriate as the prediction is for a nominal feature

k nearest neighbor (kNN)

Partial

Question 17

2.25 / 4.5 pts

Which of the following tasks are done during the data understanding phase in CRISP-DM? Check all that apply.

- [x] verify data quality
- [ ] construct data
- [x] explore data
- [x] select data
- [ ] clean data

Question 18

4.5 / 4.5 pts

A correlation between years of education and average salary has been assessed with an *R=0.87*. Which statement about the correlation is **FALSE**?

- As years of education increases, salary increases generally as well
- The correlation is weak as R is close to 1
- The correlation is reasonably strong and indicates a predictive relationship between the two variables
- A regression model would have good predictive power

Incorrect

Question 19

0 / 5 pts

Given the multiple regression output below, what is the upper bound for the 95% confidence interval of the predicted systolic blood pressure for a male, 41 years, weighing 174lbs? Round to the nearest whole number, *i.e.*, no decimals. Enter the number only.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.99
R Square	0.98
Adjusted R Square	0.97
Standard Error	2.32
Observations	11

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	1813.916268	906.9581	168.7646	2.87357E-07
Residual	8	42.99282278	5.374103		
Total	10	1856.909091			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	30.99410295	11.94378039	2.594999	0.031865
Age	0.861414686	0.248231411	3.470208	0.00844
Weight	0.334859197	0.130668274	2.562666	0.033508

135

Incorrect

Question 20

0 / 5 pts

Jim Watson created a **multiple regression predictive model** useful for forecasting employee retention. He observed an Adjusted R^2 of 0.44. Which statements about his model would be correct?

- All of the above statements are correct.
- None of the above statements are correct.
- A multiple regression model is not evaluated based on Adjusted R^2 but rather on Multiple R.

The Adjusted R^2 is adequate and indicates sufficient statistical significance.

The Adjusted R^2 is not less than 0.05 and therefore the model is not statistically significant.

The Adjusted R^2 is too low and shows that model is a poor fit for the data.

Question 21

5 / 5 pts

When should you retire a model?

When a new model has been created that is superior to the existing model.

Neither (a) nor (b).

When a change in business conditions invalidates the model's assumptions.

Both (a) and (b).

Question 22

4 / 4 pts

A new forecasting model has been constructed using linear trendline regression. There is a concern that the model generally underpredicts. This condition can be detected by its _____.

- Adjusted R^2
- mean squared error (MSE)
- mean absolute deviation (MAD)
- *p*-value based on the *F*-statistic
- bias
- R^2

Question 23

10 / 10 pts

Kai wants to ensure that his multiple regression model predicting the quarter mile acceleration is properly constructed. He builds a correlation matrix to inspect potential collinearity. Based on the built-in dataset **mtcars** in R, construct a full correlation matrix. The strongest correlation is observed to be between disp and cyl with an R = 0.90. Enter the names of the columns (*e.g.*, qsec) and the coefficient of correlation rounded to two digits.

Answer 1:

disp

Answer 2:

cyl

Answer 3:

0.90

Question 24

10 / 10 pts

Which of the following are ethical concerns that ought to be heeded by data scientists and ought to be communicated to business leaders when deploying a machine learning or data mining model? Check all that apply.

- [x] Bias in machine learning models is common and must be either avoided or properly communicated to the model's users
- [x] The privacy of personal data must be maintained.
- [x] Machine learning models are susceptible to intentional subversion.
- [x] Algorithms can contain and perpetuate biases of their creators.

Quiz Score: **102.5** out of 120