CIS9660: Data Mining for Business Analytics

Exam 1

*March 11, 2021*

NAME (PRINT CLEARLY!!): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Baruch id (PRINT CLEARLY!!): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Instructions**

* This is a close-book exam.
* There are **Two** parts on the exam and a total of 100 points possible.
* The last three pages are blank. You can write on them if you need additional space.
* **Calculator Policy**: You can use a calculator that does not have the ability to communicate

with other electronic devices. (You are not allowed to use your smartphone’s calculator.)

(This page is for grading purpose.)

|  |  |  |
| --- | --- | --- |
| **Part** | **Points Possible** | **Points Assigned** |
| Part 1 | 40 |  |
| Part 2 | 60 |  |
| **Total** | **100** |  |

**Part 1: Multiple Choice (2 points each; 40 points total)**

Please write down the answers for Part 1 (Questions 1-21) in the following table. Choose only **one** answer for each question.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| 6 | 7 | 8 | 9 | 10 |
| 11 | 12 | 13 | 14 | 15 |
| 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 |  |  |

1. Which of the following about data mining is correct?
   1. When making business decisions, we should replace intuition with data mining
   2. Data is a complement to past experience
   3. The first step of data mining is data preparation
   4. The best way to evaluate your model is to test it in deployment
2. What could be the next step of model evaluation in the data mining process?
   1. Finalizing models
   2. Deployment
   3. Business understanding
   4. Business understanding or deployment
3. For which problem would data mining analysis be **least** appropriate:
   1. Predicting whether a customer will subscribe Netflix next year
   2. Predicting a patient’s possible future symptoms given the patient’s past history of symptoms
   3. Identifying the best performing salesperson
   4. Identifying groups of houses based on their house style, value, and location
4. Which of the following is the **least** likely to be supervised data mining tasks:
   1. Predict how Uber entry may influence city traffic
   2. Predicting whether a customer will return a product
   3. Find groups of customers who are more likely to cancel their orders
   4. Identifying groups of similar houses based on their house style, value, and location
5. Which of the following is **not** **true** about regressions:
6. Regression analysis is a statistical method that allows you to examine the relationship between variables
7. Explanatory variables are termed as the dependent variables and the variables to be explained are termed as the independent variables
8. The econometric problem is to estimate this slope—that is, to estimate the effect on Y of a unit change in X—using a sample of data on these two variables.
9. B and C
10. Which of the following is the **least** likely to be a potential implication of regressions:
    1. Discover distinct customer groups who are most likely to churn.
    2. Finding groups of similar firms based on profitability, growth rate, market size, products, etc.
    3. Predict which investment product would be popular among customers in a specific year.
    4. Detect whether a person fell in a video.
11. Linear regressions assume the linearity of the relationship between dependent and independent variables. Which of the following is **not** **true** about this assumption:
    1. In most cases, a linear relationship requires the dependent variable to be continuous.
    2. A simple method to check this assumption is to examine a scatter plot between the predictors and the target.
    3. If the dependent variable is a dummy, using linear regressions will be mathematically wrong.
    4. none
12. Which of the following is **not** **true** about the **Lack of multicollinearity** assumption of linear regressions:
    1. When this assumption is violated, the parameters will be non-identifiable.
    2. This assumption will be violated when having two or more control variables perfectly correlated with your independent variable of interest.
    3. This assumption will be violated if there is too little data available compared to the number of parameters to be estimated.
    4. none
13. Which of the following is **not** correct about outliers:
    1. Outlier is an observation that lies outside the overall pattern.
    2. Least squares regression is not resistant to the presence of outliers.
    3. An outlier can pull the fit of the line toward it.
    4. We should detect and remove outliers before estimating the parameters.
14. Which of the following is **true** about how to interpret the R2 and the adjusted R2 in practice:
    1. A high R2 or an adjusted R2 means that the regressors are a true cause of the dependent variable.
    2. A high R2 or an adjusted R2 means that you have the most appropriate set of regressors.
    3. An R2 or an adjusted R2 near 1 means that the regressors are good at predicting the values of the dependent variable in the sample.
    4. None
15. Which of the following is the last thing to be weighed when selecting variables:
    1. Omitted variable bias.
    2. Data availability and data quality
    3. R2 or an adjusted R2
    4. Personal experience
16. What is the major problem with using a linear model to estimate class probability:
17. It is only mathematically correct.
18. Violates at least one assumption of linear regressions.
19. The model will not give any results.
20. Both a and b
21. Which one of the following is correct:
22. Linear regressions estimate coefficient with MLE, while logit models use OLS.
23. Logit models can fit data better than linear regression.
24. The dependent variable for logit models can only be a dummy.
25. None of the above is correct.
26. Why correlation does not always imply causation:
27. There could be a variable that is not among the explanatory or response variables in a study, and yet may cause both the change in the dependent and independent variables
28. The dependent and independent variables are confoundedwhen their effects on a response variable cannot be distinguished from each other
29. Models may not fit the data well.
30. Except c
31. If the odds of one event increase, it means that:
32. This event has a higher probability to happen
33. This event has a lower probability to happen.
34. The probability could be either higher or lower.
35. The change in probability depends on the value of the independent variables.

**Part 2: Linear Regressions (20 points total)**

A study published in Nature in 1993 (Rauscher,Shaw, and Ky, 1993) suggested that listening to Mozart for 10 to 15 minutes could temporarily raise your IQ by 8 or 9 points. That study made big news—and politicians and parents saw an easy way to make their children smarter. For a while, the state of Georgia even distributed classical music CDs to all infants in the state. Is there a Mozart effect? A review of dozens of studies found that students who take optional music or arts courses in high school do, in fact, have higher English and math test scores than those who don’t.

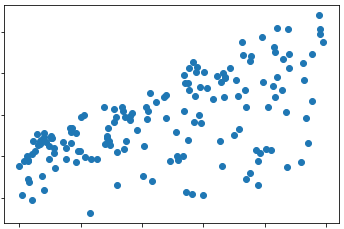
1. To answer this question, you want to test whether taking optional music or arts courses in high school leads to higher math test scores. If you start with a *simple* linear regression, what would be your dependent variable and what would be your independent variable (only list the most relevant variable)? Write down your simple linear regression model with specific variable names. (2 points)

DV:Math test score

Independent vaiable: whether taking optional music or arts courses

Math test score=B0+B1 whether taking optional music or arts courses

1. ~~Assume below is a scatter plot between your dependent variable and independent variable you list above. Does this suggest linearity? (1 point)~~

~~~~

Dependent variable

Independent variable

1. ~~Before running regressions, you need to select a sample of students. How will you draw the sample? Which assumption should the sample satisfy? (4 points)~~
2. Does your model have omitted variable bias? If yes, list three omitted variables and explain your reasons. (5 points)
3. Refine your simple linear model based on your answers to the above question. Write down a *multiple* regression model by including the omitted variables you list above. (2 points)

Math test score=B0+B1 whether taking optional music or arts courses

1. Re-evaluate your multiple regression model. Can you confidently say your model causally tests the “effect” of taking optional music or arts courses in high school on math test scores? Why do you think so? (6 points)

**Part 3: Logistic Regressions (20 points total)**

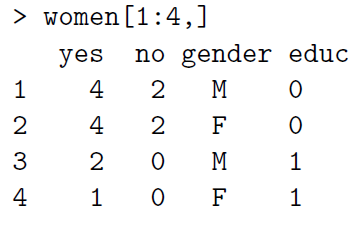
Attitude towards women's roles

In 1979, women's and men's attitudes toward women's familial roles were examined using the questionnaire \Do women belong in the home?" in a cross-sectional survey of US adults, cross-classified by respondent's gender and their formal education measured in years. Now you want to do analysis about how gender and education influence the answer.

**Key variables:**

* + yes: the number of people who responds yes
  + no: the number of people who responds no
  + gender: the gender, treated as categorical variable
  + educ: years of education, treated as continuous

**Sample data:**



1. What is the model you will use to answer the question? List as below (4 points)

Log odds of saying yes=B0+B1Male+B2Education

1. How should you change the structure of the data before estimating the above analysis in R? Show below the new data using the sample data (4 points)

Yes Male Education

1 1 0

1 1 0

1 1 0

1 1 0

0 1 0

0 1 0

……

1. Do you see any problem(s) if using your above model to answer the question in the problem statement? How can you change the model to fix the problem(s)?(6)

Omitted variable bias, list three of the them, how to solve the problem.

**Part 4: Solve Business Problems using Regressions (20 points total)**

Zoom Video Communications is rapidly emerging as the latest internet gold mine as millions of people flock to its conferencing service to see colleagues, friends and family while tethered to their homes during the pandemic. A CEO in a mid-sized technology company is considering about a permanent remote working plan. However, concerns have been raised that shifting to remote work may lead to a decline in productivity. Assume you are a data analyst in this company, can you do analysis using regressions to help the CEO make the right decision?

1. Work productivity could mean different things for different people. Based on your understanding, what could be the most appropriate variable to measure work productivity? Why do you think so? (4 points)
2. Based on your above answers, what would be your explanatory variable (only list the most important one)? Why do you think it is appropriate (4 points)?
3. Write down the regression model using the above two variables you list. (2 point)
4. How do you think you can refine your analysis so that you can provide more convincing results to the CEO for decision making? (3’ for one valid point, and 10’ in total) (note: answers with similar logic would only be counted as one valid point)

[BLANK PAGE]

*You can write here if you need additional space or as scratch papers.*

[BLANK PAGE]

*You can write here if you need additional space or as scratch papers.*

[BLANK PAGE]

*You can write here if you need additional space or as scratch papers.*