**Birla Institute of Technology and Science, Pilani**

Work Integrated Learning Programmes Division Cluster Programme - M.Tech in Data Science and

Engg. I Semester 2019-20 (Makeup) **ANSWER-KEY**

|  |  |  |
| --- | --- | --- |
| Course Number: DSECL ZG523 | | **Answer-Key** |
| Course Name: | Introduction to Data Science |  |
| Exam. Mode: | Closed Book |  |
| Weightage: | 30% | Pages: 6 |
| Duration: | 90 minutes | Questions: 4 |
| Date: | 04/01/2020 2 – 3.30PM | Marks: 30 |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Instructions**

1. Questions are to be answered strictly in the order in which they appear in the question paper and in the page numbers specified in the answer-book.
2. **All answers must be directed to the question in short and simple paragraphs or bullet points; use visuals/diagrams wherever necessary; use last sheets of your answer-book for thinking and rough-work if necessary, and write legibly the final answer on the designated pages**.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Q1. Answer the following questions briefly: **(To be answered between page 3 to 6)** **[2 \* 4 = 8]**

1. Describe a scenario where Data Scientists needs to be a bearer of bad news.

Answer:

If Data Scientists analysis uncovers serious problems at the company, or paints a less-than-rosy picture of where the firm is headed, presenting that information to management can be uncomfortable. Although data scientists are almost never the cause of these problems, a bad manager might take their dissatisfaction out on you anyway. To some extent, this is a problem you may be able to mitigate with better communication and better expectation setting. In such scenario, sometimes have to be the bearer of bad news

Just Identification of Scenario – 1 mark

Proper Explanation of the scenario – 1 mark

1. List the four outcomes of the Data Science project step where research goal and project charter are defined.

Answer:

The outcome of research goal and project charter planning stage will be

1. A clear research goal
2. A good understanding of the context
3. Well defined deliverable
4. Plan of action with timelines

Each point = 0.5 marks, 0.5 \* 4 = 2 marks

1. In situations of highly sensitive topics like HIV Aids where people will not openly discuss and participate in surveys to share information about HIV Aids. Not all the victims will respond to the questions asked so researchers can contact people they know or volunteers to get in touch with the victims and collect information. What kind of sampling technique can be used for this purpose?

Answer:

Referral /Snowball Sampling

* Used in the situations where the population is completely unknown and rare
* Will take the help from the first element which we select for the population and ask him to recommend other elements who will fit the description of the sample needed.
* So this referral technique goes on, increasing the size of population like a snowball.
* Helps in situations where we do not have the access to sufficient people with the characteristics we are seeking. It starts with finding people to study.

Identification of the technique – 1 mark

Justification – 1 mark

1. Consider the following dataset where values are Not Missing At Random (NMAR). Describe two ways by which these missing values can be imputed.

|  |  |
| --- | --- |
| Gender | Age |
| Male | 42 |
| Male | NA |
| Male | 24 |
| Male | NA |
| Male | 36 |
| Male | 57 |
| Female | 32 |
| Female | NA |
| Female | NA |
| Female | 18 |
| Female | NA |
| Female | 23 |

Answer:

Two techniques that can be used

1. Imputed value can be 18 or 57 depending on other feature observation
2. Use any value except mean/median value

Identification of each of above mentioned technique – 1 mark each, 1 \* 2 = 2 Marks

Q2. You scored a cool data science-oriented support job with Next Media, owned by one of the major online advertiser. You’ve been working with engineers from Research division to design new data-mining-based techniques to better target display advertisements. Next Media is one of the top two ad exchanges, and delivers over 5 billion ad impressions each day. Your new techniques describe a browser-site-ad combination by various features, and then predict the probability that a user will click on the ad. (You have similar models for other sorts of conversions, too.) You are considering offering a premium targeting service. Based on a small pilot study, your preliminary evaluations show promise for the new techniques. You have a meeting scheduled with the Chief Marketing Officer (CMO) to discuss your project, and you would like to prepare for strategic-level questions regarding scaling up your project.

1. In framing the problem of targeting on-line advertisements as a supervised data mining problem, what will be an instance? What will be the target variable? What will be the features?
2. What investment do you see as being most important to the success of this venture? Why? Be precise. What implications might this have that the CMO should be aware of?
3. You would like to be prepared to field questions about whether this capability has the promise to yield sustained competitive advantage for Next Media over "Adverts” who are also designing methods for targeting display ads. List the top three reasons that it might, for this specific case. Be specific. **(To be answered between page 7 to 10)** **[3 + 2 + 3 = 8]**

Answer:

1. The details are -

* An instance will be an impression opportunity: a “slot” on a page, plus a browser visiting the page, plus any information about the moment of the impression (time of day, day of week, etc.).
* The target variable will be some quantity of interest to the advertiser, such as clicking on an ad or buying a product.
* The features will be feature of the page, such as its topic, and features of the browser, such at what can be gleaned from the IP address (such as geographic location), what can be seen in the http request (they might not know that), and data we have gathered on the pages that this browser has visited in the past.

Each of above point i.e. instance, target and features – 1 marks each, 1 \* 3 = 3 Marks

1. We’ll need to invest in data to train and to evaluate the new techniques at a scale larger than the pilot study. This will be an investment, because we can’t just show the ad to those who we think are most likely to respond. To train and evaluate the model, we have to target more-or-less randomly. The implications are that the ads served for the data gathering purpose may not be as profitable as if we use our current best practices.

Proper explanation – 2 marks

Consider partial marking as well

1. The top three reasons
2. We have a much larger volume of data than the ad networks, across many advertisers and many publishers. This is a complementary resource that will be hard to replicate, and we believe that more data will lead to more accurate targeting models. Furthermore, this can create a virtuous cycle, wherein we create better targeting, which brings us more customers, thereby increasing the amount of data we have.
3. We have tremendous analytical/technical capability via Research. We probably have much more sophisticated data mining/machine learning expertise. These experts value being at the research lab, and the freedom and prestige that it affords them. This will be very costly/impossible for most ad networks to replicate. What would it take to get the very best machine learning expertise to choose an ad network over Yahoo! Research?
4. Research Org is patenting machines. We presumably can create an intellectual property fortress that can sustain competitive advantage.
5. Competitors will not be able to tell by looking at our results exactly what has led to our success (causal ambiguity, plus trade secrets). Data mining-based products are attractive in this respect, because how the work exactly is not outwardly apparent.

Explanation of each of above points (any three) – 1 mark each

Consider partial marking as well

Q3. Consider the following dataset for answering the following feature subset selection related questions.

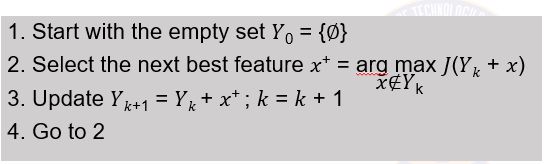
|  |  |  |
| --- | --- | --- |
| Feature1 (F1) | Feature2(F2) | Feature3(F3) |
| A | C | X |
| A |  | Y |
|  | C | Z |
| B | C | Q |

1. Describe the steps used in Sequential Forward Selection (SFS) algorithm for feature selection.
2. What the two stopping criteria’s that are most commonly used in SFS algorithm.
3. Draw a lattice structure that will represent the different models that will be generated when the above mentioned dataset is used.

**(To be answered between page 11 to 14)** **[3 + 2 + 3 = 8]**

Answer**:**

1. The steps is SFS



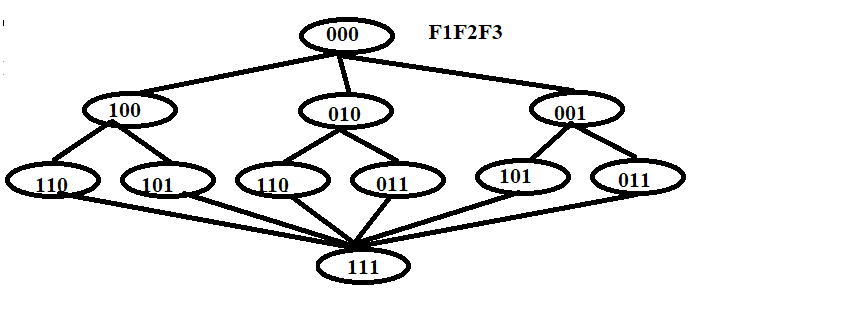
Each step – 1 mark, 1 \* 3 = 3 marks

1. Commonly used Stop criterions
2. Increase / Decrease in Predictive accuracy
3. Predefined number of features is reached

Each criteria – 1 mark, 1 \* 2 = 2 marks

1. The lattice structure – assume features as F1, F2 and F3.

1 indicates the feature selected, 0 represents it’s not selected.



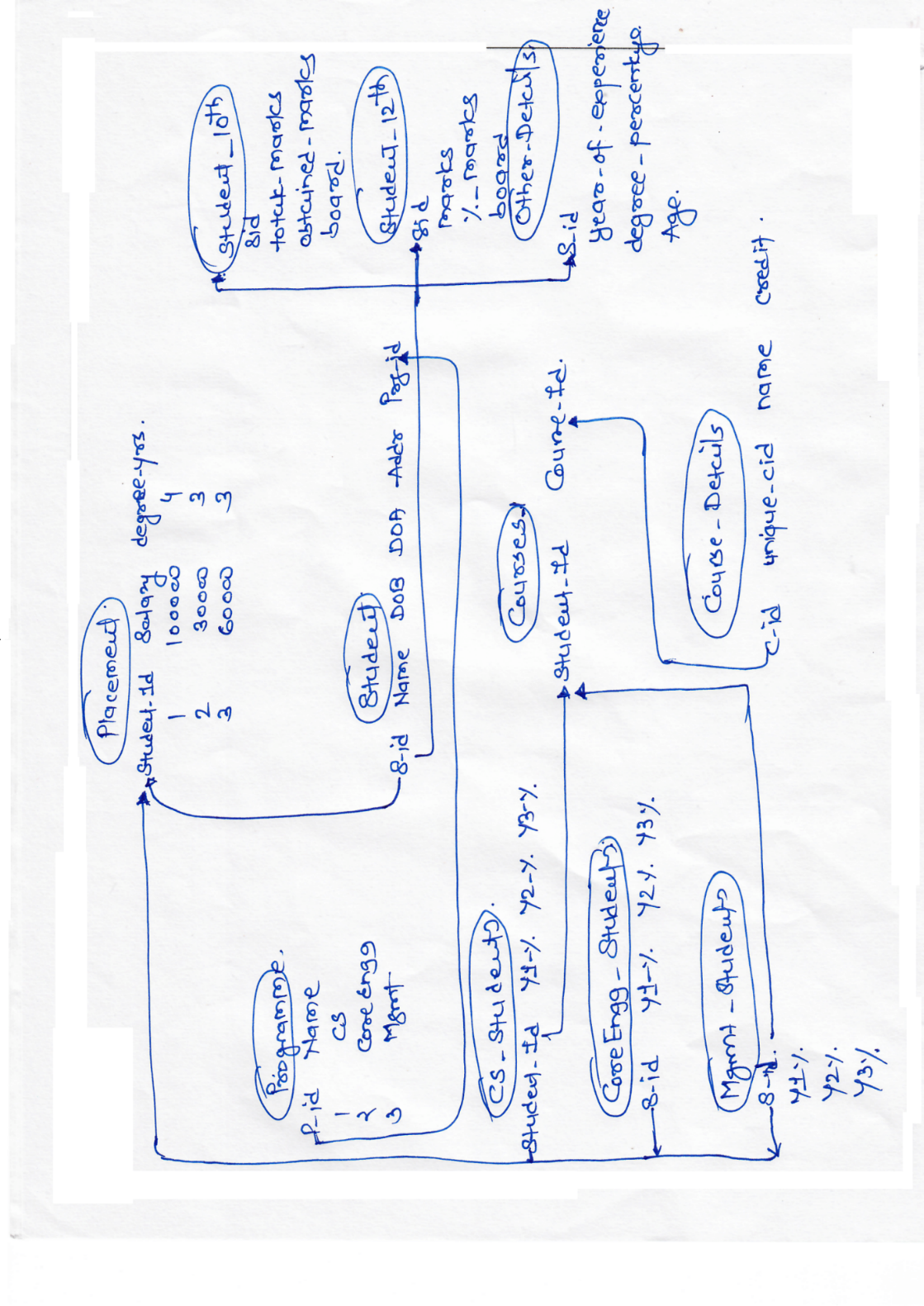
Each level -1 mark, 1 \* 3 = 3 marks

Q4. The dean of placement division of Premiere Institute has collected data on their recent placement. To attract good students, it is important for the school to ensure that the students are placed with good salary package. The dean believed that the salary earned by a student at placement dependent on several variables.

The data collected by Dean has the following structure associated with it (look at next page). But he is not able to identify the significant attributes that are helping the students to earn good salaries. You are consulted by him to help in this exercise.

1. List down two major ambiguities which are present in the given ER diagram.
2. Flatten out the data set by applying the data preprocessing techniques learnt in the classroom.
3. Using the flattened out dataset, derive a model (just based on the intuition) that will predict what range of salaries can be earned by the student?

**(To be answered between page 15 to 17)** **[2 + 3 + 1 = 6]**



Answer:

1. Which student id needs to be considered? It’s referred differently all over the places.

What is the duration of the programmers? The year wise marks, degree marks are scattered.

Each ambiguity -1 mark, 1 \* 2 = 2 marks

1. Flattened data set

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Student\_id | Name | Age | DOA | Addr | Programme | SSC marks | SSC Board | HSC marks | HSC board |
| Year\_Of\_Exp | Degree\_perc | Y1% | Y2% | Y3% | Y4% | Degree\_years | Salary |  |  |

Removal of duplicate attributes – 1 mark

Removal of irrelevant attributes – 1 mark

Converting the attributes into appropriate forms – 1 mark

1. Student can derive logistic regression model just taking two salary ranges into consideration like HIGH or LOW

Or they can think of building a decision tree where the outcome can be discrete salary ranges like LOW, AVERAGE, and HIGH etc.

Any one of the above model – 1 mark