University of London

Assessment Coversheet

Complete this coversheet and read the instructions below carefully.

Candidate Number: EX0829
Refer to your Admission Notice

Degree Title: BSc
e.g. LLB

Course/Module Title: Data Science
As it appears on the question paper

Course/Module Code: CM3005
This is in the top right corner of the question paper. If there is more than one code, use the first code.

Enter the numbers, and sub-sections, of the questions in the order in which you have attempted them:
Q2 (a) [i][ii] (b) (c) (d)

Q3 (a) (b) (c) (d)

Date: 14 MARCH 2023

Instructions to Candidates

1. Complete this coversheet and begin typing your answers on the page below, or, submit the coversheet with your handwritten answers (where handwritten answers are permitted or required as part of your online timed assessment).
2. Clearly state the question number, and any sub-sections, at the beginning of each answer and also note them in the space provided above.
3. For typed answers, use a plain font such as Arial or Calibri and font size 11 or larger.
4. Where permission has been given in advance, handwritten answers (including diagrams or mathematical formulae) must be done on light coloured paper using blue or black ink.
5. Reference your diagrams in your typed answers. Label diagrams clearly.

The Examiners will attach great importance to legibility, accuracy and clarity of expression.

Begin your answers on this page

Candidate Number:EX0829

Question 2

(2a)

(i) For House Prices, we may commonly use regression algorithms. For example:

1. Linear Regression: This algorithm is used to predict the continuous target variable, such as house prices. The input data should contain features such as the number of bedrooms, bathrooms, square footage, and location.

2. Decision Trees: This algorithm is used to predict the target variable by recursively partitioning the input data based on the features. For house price prediction, the input data should contain categorical and numerical features such as the age of the house, proximity to schools, shopping malls, and transportation.

The type of dataset used for predicting house prices through machine learning algorithms usually contains both **categorical** and **numerical features**.
Categorical features, such as location, type of house, and other character type input value about the house condition, offer critical information for determining the house price.
Numerical features, including the number of bedrooms, bathrooms, and those numerical input value about the property of the house, also play a significant role.

(ii) For Measure Customer Habits, we may commonly use clustering algorithms. For example:

1. K-Means Clustering: This algorithm is utilized to group customers with similar behaviour based on features such as purchase history, demographics, and other relevant customer behaviour data.

2. Random Forest: This algorithm is also be useful for predicting customer habits, especially when dealing with datasets containing numerous features.

The type of dataset required to measure customer habits through machine learning algorithms should encompass relevant information on customer purchases, demographics, and other behaviour-related data. Additionally, transactional data, such as the date of purchase and the products purchased, can prove to be valuable in assessing and analyzing customer habits.

(2b)

Let A be the event that the red pill was drawn from the first box, and B be the event that a red pill was drawn (from either box).

	RED	BLUE	IN TOTAL
BOX 1	7	3	10
BOX 2	12	8	20
IN TOTAL	19	11	30

We want to find P(A|B), the probability that the red pill was drawn from the first box given that a red pill was drawn.

According to the problem statement, we know:

P(A) = 1/2, since both boxes are identical and the pill was drawn from one of them at random.
P(B|A) = 7/10, since the first box has 7 red pills out of 10 total pills, and we are given that a red pill was drawn from this box.
P(B|not A) = 12/20, since the second box has 12 red pills out of 20 total pills, and we are given that a red pill was drawn from one of the boxes. not A means the event that the pill was drawn from the second box.

By using Bayes' theorem, we have:

P(A|B) = P(B|A) * P(A) / P(B)
where
P(B) = P(B|A) * P(A) + P(B|not A) * P(not A)

Substitute the values we know:

P(B) = (7/10 * 1/2) + (12/20 * 1/2) = 13/20

P(A|B) = (7/10 * 1/2) / (13/20) = 7/13 = 0.53846(5.d.p)
Therefore, the probability that the red pill was drawn from the first box given that a red pill was drawn is 7/13, approximately is 53.8% (3 significant figure).

(2c)

The preferred method for dealing with missing data depends on the nature and extent of the missing values. However, the most common approach is imputation, which involves estimating the missing values using other available data.

‘ffill’ and ‘bfill’ are two types of imputation techniques used to fill missing values in a dataset. ffill (forward fill) replaces missing values with the most recently observed value in the same column, while bfill (backward fill) replaces missing values with the next observed value in the same column.

ffill is preferred over bfill when the missing values occur at the beginning of the dataset, as bfill will propagate the missing values forward. However, both techniques have limitations in that they do not take into account the relationships between variables, and may introduce bias or distort the distribution of the data.

In Scikit-learn (SKlearn) library, there are several imputation methods available, including:

1. Mean imputation: This involves replacing missing values with the mean of the non-missing values in the same column. It is a simple and effective method that works well when the missing values are missing at random and the dataset is not too large.

2. K-nearest neighbors imputation: This involves estimating the missing values by using the values of the k-nearest neighbors in the dataset. It works well when there is a correlation between missing values and the values of other variables in the dataset. However, it may be computationally expensive for large datasets and requires the choice of an appropriate k value.

(2d)

1. Ambiguity in Names: The name "Martin O'Neill" has multiple possible interpretations, such as whether "O'Neill" is a middle name or part of a hyphenated last name. This can create ambiguity in tokenization and named entity identification.

2. Special Characters: The use of an apostrophe in "O'Neill" can pose a challenge for tokenization, as it may be interpreted as a separate token or as part of the named entity. Similarly, the use of hyphens in "January-June" and "2018-19" can complicate tokenization and named entity recognition.

3. Contextual Disambiguation: In the sentence "Result: it didn't work out", it is unclear what the "it" refers to without further context. This lack of clarity can make it difficult to accurately identify the named entity being referred to.

4. Seasonal References: The reference to the "2018-19 season" may be interpreted as a single named entity or two separate entities. For example, it can be separated into two entities called “ 2018” and “19 season” . Moreover, depending on the domain, there may be multiple possible interpretations of what constitutes a "season". This can lead to ambiguity in named entity recognition.

Question 3

(3a)

Bayes' theorem is a mathematical formula used to calculate the probability of an event based on prior knowledge of conditions that might be related to the event. The theorem is written as:

P(A|B) = P(B|A) * P(A) / P(B)

where P(A|B) is the conditional probability of event A given event B has occurred, P(B|A) is the conditional probability of event B given event A has occurred, P(A) is the prior probability of event A, and P(B) is the prior probability of event B.It allows us to update our prior beliefs or assumptions about the probability of an event in light of new evidence or data.

By Bayes Theorem,

the prior probability is the initial probability of a hypothesis before considering any new evidence or data. It represents the probability of the hypothesis based on prior knowledge or experience.

The likelihood is the probability of observing the evidence given a hypothesis. It represents the degree to which the observed evidence supports or contradicts the hypothesis.

(3b)

Formula:

s.f. = Significant Figure

Precision = (true positive)/(predicted positives)
= (true positive)/[(true positives)+(false positives)]

Recall = (true positive)/ (actual positive)
= (true positive)/[(true positive)+(false negative)]

Average Precision = (Sum of precision)/ (total number of precision)

Average Recall = (Sum of Recall)/ (total number of recall)

F-measure = 2 * [(precision * recall) / (precision + recall)]

For Dog Class:

True positive for Dog (real dog): 20

False positive for Dog (cat or rabbit but predicted as dog): 3+1 =4

False negative for Dog (Dog but predicted as cat or rabbit): 6+4 =10

Precision = 20 / (20+4) = 20/24 = 5/6 = 0.833 (3 s.f.)

Recall = 20 / (20+10) = 20/30 = 2/3 = 0.667 (3 s.f.)

For Cat Class:
True positive for Cat (real cat): 16
False positive for Cat (dog or rabbit but predicted as cat): 6+2 =8
False negative for Cat (Cat but predicted as dog or rabbit): 3+1 =4

Precision = 16 / (16+8) = 16/24 = 2/3 = 0.667 (3 s.f.)
Recall = 16 / (16+4) = 16/20 = 4/5 = 0.800 (3 s.f.)

For Rabbit Class:
True positive for Rabbit (real rabbit): 7
False positive for Rabbit (dog or cat but predicted as rabbit): 4+1 =5
False negative for Rabbit (Rabbit but predicted as dog or cat): 1+2 =3

Precision = 7 / (7+5) = 7/12 = 0.583 (3 s.f.)
Recall = 7 / (7+3) = 7/10 = 0.700 (3 s.f.)

Average Precision = ((5/6) + (2/3) +(7/12)) / 3 = (25/12) / 3 = 25/36 = 0.694 (3 s.f.)
Average Recall = ((2/3) + (4/5) + (7/10)) / 3 = (130 / 60) /3 = 13/18 = 0.722 (3 s.f.)

F measure of Dog class:
2 * [((5/6) * (2/3)) / ((5/6) + (2/3))] = 20/27 = 0.741 (3 s.f.)

F measure of Cat class:
2 * [((2/3) * (4/5)) / ((2/3) + (4/5))] = 8/11 = 0.727 (3 s.f)

F measure of Rabbit class:
2 * [((7/12) * (7/10)) / ((7/12) + (7/10))] = 7/11 = 0.636(3 s.f)

	PRECISION	RECALL	F MEASURE
DOG	0.833	0.667	0.741
CAT	0.667	0.800	0.727
RABBIT	0.583	0.700	0.636
AVERAGE	0.694	0.722	

(3c)
Here we use an example of security checking, a false positive can be more detrimental than a false negative.
For example, consider a security system that scans individuals entering a restricted area for potential threats like weapons or explosives. A false positive in this case would occur if the system incorrectly identifies a harmless item as a weapon or explosive, causing unnecessary panic and disruption to operations, which could result in harm to innocent individuals.
Conversely, a false negative in this case would occur if the system fails to detect an actual weapon or explosive, which could lead to a serious security breach and

endanger people's lives. Therefore, in this scenario, a false positive is more costly than a false negative because it could result in harm to innocent individuals and disrupt operations.

(3d)

Here we use a example of medical testing. In the medical testing about serious disease, such as COVID-19, a false negative result could result in significant negative consequences.

For example, if a medical test fails to detect a serious disease, it may prevent the timely implementation of necessary medical interventions, which may exacerbate the patient's condition, leading to increased morbidity and even mortality. Let’s say if a COVID-19 patient gets false negative test, then he may still carry the virus and infect others.

On the other hand, a false positive result, while potentially causing distress and leading to unnecessary treatment, does not pose an immediate risk to the patient's health. Therefore, in medical testing, the cost of a false negative may be greater than that of a false positive.