THE UNIVERSITY OF HONG KONG

SCHOOL OF COMPUTING AND DATA SCIENCE
DEPARTMENT OF COMPUTER SCIENCE

FITE7410 Financial Fraud Analytics

Date: Dec 7, 2024 **Time: 6:30 – 8:30pm**

Answer ALL questions in SEPARATE ANSWER BOOK provided.

Question No.	**Mark**
1 (20 marks)	
2 (20 marks)	
3 (30 marks)	
4 (30 marks)	
Total (100 marks)	

Only approved calculators as announced by the Examinations Secretary can be used in this examination. It is candidates' responsibility to ensure that their calculator operates satisfactorily, and candidates must record the name and type of the calculator used on the front page of the examination script.

This is an open book examination. Candidates may bring to their examination any printed/written materials.

University Numbers: ____________________

Brand and Type of Calculator: ____________________

QUESTION 1.

QUESTION 1.1 (5 marks)
Please name 3 key legislations in relation to Anti-money Laundering / Counter Terrorist Financing in Hong Kong.

QUESTION 1.2 (5 marks)
Please name 3 situations when a financial institution has to conduct customer due diligence.

QUESTION 1.3 (10 marks)
You are a data scientist working for a credit card company. The company wants to develop a model to predict whether a new customer is likely to default on their credit card payments. You decide to build a decision tree model to predict default risk.

a) You can use the Gini Importance to select the root node. Show the calculation steps for decrease in Gini impurity (decimal up to 3 digits) for the following two features with splitting criteria:
   - Feature 1: Credit Score < 700
   - Feature 2: Debt-to-income ratio < 0.3

Age	Income	Credit Score	Debt-to-income ratio	Number of credit cards	History of late payments	Default
25	50,000	720	0.3	2	0	No
30	75,000	780	0.2	1	0	No
35	100,000	850	0.1	3	0	No
40	125,000	700	0.4	4	1	Yes
45	150,000	800	0.3	2	0	No
50	175,000	650	0.5	5	2	Yes
55	200,000	750	0.2	3	0	No
60	225,000	820	0.1	1	0	No
65	250,000	700	0.3	4	1	Yes
70	275,000	780	0.2	2	0	No
75	300,000	650	0.4	5	2	Yes
80	325,000	750	0.3	3	0	No
85	350,000	820	0.2	1	0	No
90	375,000	700	0.4	4	1	Yes

b) Based on the result of (a), which feature should be used at the root node. Explain your choice of the root node.

QUESTION 2. (20 marks)

Case Background

A whistle blower complaint was received by the Country Chief Executive Officer (CEO) in China of a USA manufacturing company targeting at their Sales Director in Shanghai.

There were allegations of diversion of the company's business to private business held by the Sales Director's family members as well as fraudulent expense claims made by the Sales Director during the last 2 years.

Please provide an Action Plan setting out how you would investigate the potential irregularities/ issues highlighted above based on what you have learned in fraud investigation.

Please also explain what technology you would use to assist with the investigation, e.g. data analytics and/or any other tools.

QUESTION 3. (30 marks)

A financial institution is developing a machine learning model to predict fraudulent transactions in real-time. They have collected a dataset of 1 million credit card transactions, labeled as either fraudulent (1) or legitimate (0). The dataset is imbalanced, with only 10,000 transactions being fraudulent. The team has trained and evaluated four different machine learning algorithms. The following table summarizes the confusion matrix results on the test set:

Logistic Regression

	Predicted Fraudulent	Predicted Legitimate
Actual Fraudulent	700	300
Actual Legitimate	700	29,300

Random Forest

	Predicted Fraudulent	Predicted Legitimate
Actual Fraudulent	800	200
Actual Legitimate	800	29,200

SVM

	Predicted Fraudulent	Predicted Legitimate
Actual Fraudulent	750	250
Actual Legitimate	750	29,250

Neural Network

	Predicted Fraudulent	Predicted Legitimate
Actual Fraudulent	900	100
Actual Legitimate	850	29,150

a) Calculate the Recall, Precision and F1-Score for the above 4 algorithms. Show the calculation steps ((decimal up to 3 digits) .

b) Explain the implications of high and low Precision, and the implications of high and low Recall. Discuss the trade-off between Precision and Recall in the context of this case study.

c) What are the implications of high false positives in this scenario? List at least 3 implications. Suggest how to reduce these implications.

d) Based on the results, which model would you recommend for deployment? Explain your reasoning.

e) Describe how you would explain the predictions of the chosen model in (d) to stakeholders who are not familiar with machine learning.

QUESTION 4. (30 marks)

Multiple choice questions, choose **ONE** answer for each question. **2 marks for each question** with correct answer, **2 marks penalty** for incorrect answer or for selecting more than one answers for the same question. Maximum mark in Question 4 is 30 marks. The minimum mark in Question 4 is 0 marks. Select your answer with care.

Answer ALL questions in ANSWER BOOK (i.e. NOT on this question paper).

(Questions continue on the next page.)

=== END OF PAPER ===