THE UNIVERSITY OF HONG KONG

FACULTY OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE

FITE7410 Financial Fraud Analytics

Date: 8 December 2022 (THU) **Time: 6:30-8:30pm**

Answer ALL questions in SEPARATE ANSWER BOOK provided.

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

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.

QUESTION 1.

QUESTION 1.1.

Below is the result of a Logistic Regression model to predict a fraudulent company based on the financial ratios.

	Model 1		**Model 2**		**Model 3**	
	Coefficients	**p-value**	**Coefficients**	**p-value**	**Coefficients**	**p-value**
Net profit/Equity (X1)	0.519	0.107	0.703	0.173	0.624	0.003
Cash/Current liabilities (X2)	1.109	0.205	1.936	0.011	1.228	0.004
Total liabilities/Total assets (X3)	5.242	0.000	4.766	0.000	5.619	0.000
Sales/Fixed assets (X4)	3.048	0.053	0.029	0.026	2.722	0.010
Inventories/Total assets (X5)	5.104	0.067	4.263	0.051	5.832	0.002
Constant (C)	-7.279	0.000	-5.768	0.000	-7.578	0.000
R-square	0.6842		0.7406		0.7408	
Adjusted R-square	0.6765		0.7338		0.7262	

(a) Based on the above result table, which model you would choose as the final fraud prediction model? Explain your answer based on the R-square and Adjusted R-square values.

(b) Based on your choice in (a), write down the formula for this selected Logistic Regression model.

(c) Using this Logistic Regression model, would the following new observation be classified as fraudulent financial statement (assume we use a threshold of 0.5)?

ID	**X1**	**X2**	**X3**	**X4**	**X5**
1003	0.036	0.80	2.56	2.27	0.103

Show the steps of your calculation. NO marks will be given if only final values are written as the answers.

QUESTION 1.2.

What is Money Laundering? Please use an example to illustrate how it works.

QUESTION 1.3.

Please name 4 types of entities covered by the Anti-money Laundering Ordinance in Hong Kong.

QUESTION 2.

Case Background

A whistle blower complaint was received by the CEO of an Information Technology company in the USA headquarters, targeting at their operation in Shanghai.

There were allegations on inflation of revenue, falsification of expenses and bribery during the past 2 years.

As an external consultant to the CEO of the USA headquarters, 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.

Your task is to select the most suitable fraud detection model to predict listed companies that are fraudulent or not based on the financial variables derived from their published financial statements. 4 machine learning models are trained.

QUESTION 3.1 Performance of ROC AUC in two different data sets with 4 models.

Model	**AUC (imbalanced data set)**	**AUC (balanced data set)**
Logistic regression (LR)	0.719	0.757
Random forest (RF)	0.562	0.635
Support Vector Machine (SVM)	0.509	0.574
Decision Tree (DT)	0.505	0.513

Answer **ALL** of the questions below:

(a) List **ONE** possible technique that you could use to transform the imbalanced dataset to balanced dataset. Explain the advantages and disadvantages of this selected technique.

(b) Based on results of AUC, which dataset (imbalanced vs balanced) you would select to build your fraud detection model? Explain your answer with reference to AUC results and list all the reasons to support your choice.

QUESTION 3.2 Confusion matrix of 4 models.

Model	True negative (TN)	False positive (FP)	False negative (FN)	True positive (TP)
LR	5313	50	22	33
RF	5330	33	15	40
SVM	5309	54	14	41
DT	5323	40	32	23

Answer **ALL** of the questions below:

(a) Calculate the Accuracy, Recall and Precision for the 4 models and fill in the table below.

(NOTE: Please write your answer in the ANSWER BOOK)

Model	Accuracy	Recall	Precision
LR			
RF			
SVM			
DT			

Show the steps of your calculation. NO marks will be given if only final values are written as the answers.

(b) Based on all the performance metrics above, which model you would select as the fraud detection model? Explain your choice.

QUESTION 3.3 Assume you conducted additional work to fine tune each model with additional linguistic variables. Below is the new performance results.

Model	Recall	Precision	AUC
LR + linguistic variables	99.010%	99.497%	0.801
RF + linguistic variables	99.136%	99.171%	0.791
SVM + linguistic variables	98.974%	98.337%	0.780
DT + linguistic variables	98.995%	99.542%	0.781

Answer **ALL** of the questions below:

(a) Assuming the business objective is to prefer a fraud model which can detect fraud as far as possible, which model you would recommend. Explain your choice.

(b) Assuming there are business costs for false alarms and at same time also prefer a model that can detect fraud as far as possible, which model you would recommend. Explain your choice.

QUESTION 4. (Not for circulation)

Multiple choice questions, choose **ONE** correct answer for each question. Enter your answers in the **SEPARATE ANSWER BOOK** provided. **2 marks for each question** with correct answer, **2 marks penalty** for incorrect answer or 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.

=== END OF PAPER ===