

THEME A: CONCEPTS · A4.2 · HL ONLY

Data preprocessing

Original practice questions · 33 marks · about 50 minutes · covers A4.2.1–A4.2.3

Instructions. Answer all questions. The maximum mark for this paper is [33 marks]. The marks available are shown in brackets. Parts badged **HL** assess higher level only content; standard level students may leave these out. **A4.2 is a higher level only topic**, so every question in this paper is HL.

1

Total: 7 marks

A bank is preparing a data set to train a model that predicts whether a loan applicant will default. An extract from the raw data is shown in Table 1.

Table 1: extract from the LOAN_APPLICATIONS data set

ApplicantID	Age	AnnualIncome	EmploymentYears	City	Defaulted
A101	34	41200	6	Leeds	No
A102	29	38,500	3	leeds	No
A103	52	(blank)	21	Bristol	Yes
A104	41	55000	12	Cardiff	No
A104	41	55000	12	Cardiff	No
A105	203	47000	9	Bristol	Yes

- (a) **HL** Identify **two** data quality problems visible in Table 1. [2]

- (b) **HL** Describe **two** different techniques the bank could use to deal with the missing annual income for applicant A103. [4]

- (c) **HL** State **one** consequence of training the model without correcting the age recorded for applicant A105. [1]

The bank continues to clean the data set.

- (a) **HL** Explain why the duplicated record for applicant A104 must be removed before training. **[3]**

- (b) **HL** Explain why the values in the City and AnnualIncome columns must be transformed into a consistent format. **[3]**

In the cleaned data set, Age ranges from 18 to 80 while AnnualIncome ranges from 9,000 to 400,000.

- (a) **HL** Explain why the bank scales these two attributes before training the model. **[4]**

- (b) **HL** A small number of applicants earn far more than the rest. Identify whether normalization or standardization would cope better with these values, and justify your answer. **[2]**

4

Total: 6 marks

The full data set holds 180 attributes for each applicant. The bank decides to use only some of them.

- (a) **HL** Outline what the bank is doing to its 180 attributes when it carries out feature selection. [2]

- (b) **HL** Describe the difference between filter methods and wrapper methods of feature selection. [4]

5

Total: 4 marks

The bank reviews which of the 180 attributes to keep.

- (a) **HL** Explain why ApplicantID should not be used as a feature. [2]

- (b) **HL** Outline **one** benefit to the bank of training on fewer features. [2]

6

Total: 4 marks

Reducing the number of attributes is described as addressing the curse of dimensionality.

- HL** Explain what is meant by the curse of dimensionality. [4]

END OF QUESTIONS