

THEME A: CONCEPTS · A4.2 · HL ONLY

Data preprocessing — Mark scheme

33 marks · covers A4.2.1–A4.2.3

How to mark. Each creditworthy point is on its own line and ends with a semicolon (;). A line beginning with ... is an *expansion* of the line above it, so the pair is worth the marks shown. A slash (/) separates acceptable alternative wordings. Words in (brackets) are not needed for the mark. Mark positively: credit what the student has got right rather than penalising what is missing, and accept any valid alternative not listed here. Do not award more than the maximum for a part. Where a later part uses a wrong answer from an earlier part correctly, award follow-through marks (FT). HL marks parts that assess higher level only content.

Q	ANSWER	MARKS
HL 1(a) A4.2.1	Award [2] max Award [1] per problem correctly identified, [2] max. Any two. The record for A104 is duplicated; The annual income for A103 is missing; The age recorded for A105 (203) is impossible, so it is incorrect data or an outlier; The City values are inconsistently formatted, with both Leeds and leeds appearing; AnnualIncome is inconsistently formatted, with 38,500 written with a comma while other values are plain numbers, so the column is text rather than numeric;	2
HL 1(b) A4.2.1	Award [4] max Award up to [2] per technique (point plus expansion), [4] max, for two different techniques. Deletion Remove the whole record, or the attribute, from the data set; ...(which) is simple, but discards the other valid information about A103 and shrinks the data set; Imputation Replace the missing value with a statistic calculated from the other records, such as the mean or median income; ...(which) keeps the record, though it reduces the natural variation in the data and may bias the model; Predictive modelling Use the other attributes, such as age and employment years, to predict what the missing income is likely to be; ...(which) is more accurate than a simple average, but is more complex and introduces the assumptions of the second model; Do not award [4] for two variations of the same technique, for example mean imputation and median imputation.	4
HL 1(c) A4.2.1	Award [1] max Any one, [1] max. The model would learn from an impossible value, distorting the relationship it finds between age and default; As an extreme outlier it would pull any scaling or averaging badly out of range, compressing all the genuine ages together;	1

Q	ANSWER	MARKS
	Predictions for real applicants would become less accurate;	
HL 2(a) A4.2.1	Award [3] max Award [1] per point up to [3] max. A duplicated record gives that one applicant twice the influence of every other applicant on what the model learns; The model is therefore biased towards the pattern shown by A104, who did not default; If the duplicate appears in both the training and the test set, the model is tested on a record it has already seen; ... (so) the measured accuracy would look better than the model's real performance; Any statistic calculated from the data, such as the mean income, would also be distorted;	3
HL 2(b) A4.2.1	Award [3] max Award [1] per point up to [3] max. Both columns should be addressed for full marks. Leeds and leeds are the same city, but the algorithm compares values literally and would treat them as two different categories; ... (so) the records for Leeds would be split between two groups and any pattern in them weakened; AnnualIncome containing 38,500 with a comma cannot be read as a number; ... (so) the whole column would be treated as text, and no arithmetic, scaling or comparison could be performed on it; Consistent formatting means each real-world value is represented in exactly one way throughout the data set;	3
HL 3(a) A4.2.1	Award [4] max Award [1] per point up to [4] max. The answer must use the ranges given. Age spans a range of about 62 while AnnualIncome spans a range of about 391,000; Many algorithms measure the distance or similarity between records, and that calculation is dominated by whichever attribute has the largest numeric range; ... (so) a difference of 40 years in age would count for almost nothing beside a difference of a few thousand in income; The model would effectively ignore age, even if age is genuinely predictive of default; Scaling puts both attributes onto a comparable range, so each contributes according to its actual importance; It also helps algorithms that use gradient descent to converge faster and more reliably;	4
HL 3(b) A4.2.1	Award [2] max Award [1] for the correct choice and [1] for the justification. No mark for the choice alone. Standardization; Normalization rescales using the minimum and maximum, so a single very high income sets the top of the range and squashes every ordinary income into a narrow band near zero; ... (whereas) standardization rescales to a mean of 0 and a standard deviation of 1, so extreme values become large scores but do not compress the rest of the data; Accept a well-argued case for normalization if the candidate states that the outliers would first be removed or capped during cleaning. Do not award the justification mark for simply restating the definitions without linking them to the effect of the outliers.	2
HL 4(a) A4.2.2	Award [2] max Award [1] per point up to [2] max. Must refer to keeping some attributes and discarding others. The bank is deciding which of the 180 attributes to keep and which to discard before training;	2

Q	ANSWER	MARKS
	It retains the most informative ones, those that actually help predict whether an applicant defaults; It discards attributes that are irrelevant, or redundant because another attribute already carries the same information; The rows themselves are unchanged; it is the columns that are reduced;	
HL 4(b) A4.2.2	Award [4] max Award [1] per point up to [4] max. Both methods must be addressed and the contrast made clear. Filter methods judge each attribute using a statistical measure of its relationship with the target, before and independently of any model; They are fast and can be applied once to all 180 attributes, but they ignore how attributes behave in combination; Wrapper methods train and evaluate an actual model on different subsets of attributes and keep the subset that performs best; They account for interactions between attributes and are usually more accurate, but they are far more computationally expensive because a model is trained for every subset tried; ...(so) filter methods suit a first pass over a large number of attributes, while wrapper methods suit refining a smaller shortlist; Also accept a comparison mentioning embedded methods, which perform the selection as part of training the model itself.	4
HL 5(a) A4.2.2	Award [2] max Award [1] for the point and [1] for the expansion. ApplicantID is a unique identifier, so it carries no information about whether that person will default; ...(and) it is different for every record, so it cannot describe a pattern that generalises to new applicants; The model could latch onto the identifier and appear to fit the training data; ...(but) it would fail completely on new applicants, whose identifiers it has never seen;	2
HL 5(b) A4.2.2	Award [2] max Award [1] for the point and [1] for the expansion. Any one benefit. Training uses less computation and less memory; ...(so) the model trains faster and can be retrained more often; Removing irrelevant attributes reduces the noise the model can fit to; ...(so) it is less likely to overfit and should generalise better to new applicants; A model built on fewer attributes is easier to interpret; ...(so) the bank can explain to a regulator or a customer why an application was refused;	2
HL 6 A4.2.3	Award [4] max Award [1] per point up to [4] max. The curse of dimensionality describes the problems that arise as the number of attributes in a data set grows large; The volume of the space grows exponentially with the number of dimensions, so the available records become sparse within it; ...(so) far more data would be needed to cover that space adequately, and the sample size required grows quickly; Distance measures become less meaningful, because in high dimensions all points end up roughly the same distance apart;	4

Q	ANSWER	MARKS
	...(which) undermines algorithms such as K-NN and clustering that rely on distance; With many attributes and limited records, the model can fit noise rather than signal, so it overfits; Computation time and memory usage rise, making training slower and more expensive; The data also becomes impossible to visualise directly, making it harder to inspect;	

Total for this paper: 33 marks