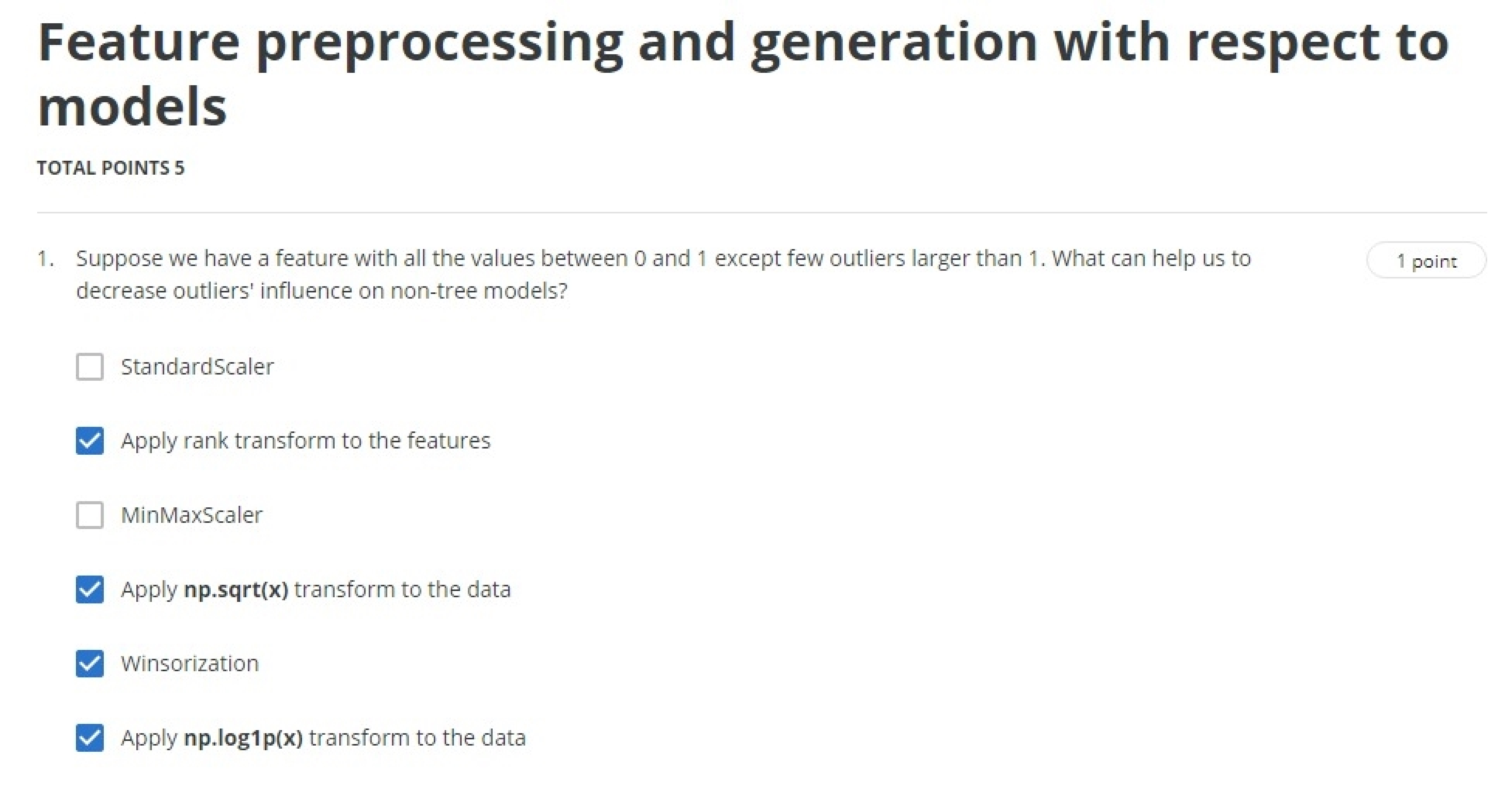


Feature preprocessing and generation with respect to models

TOTAL POINTS 5

1. Suppose we have a feature with all the values between 0 and 1 except few outliers larger than 1. What can help us to decrease outliers' influence on non-tree models?

1 point

☐ StandardScaler

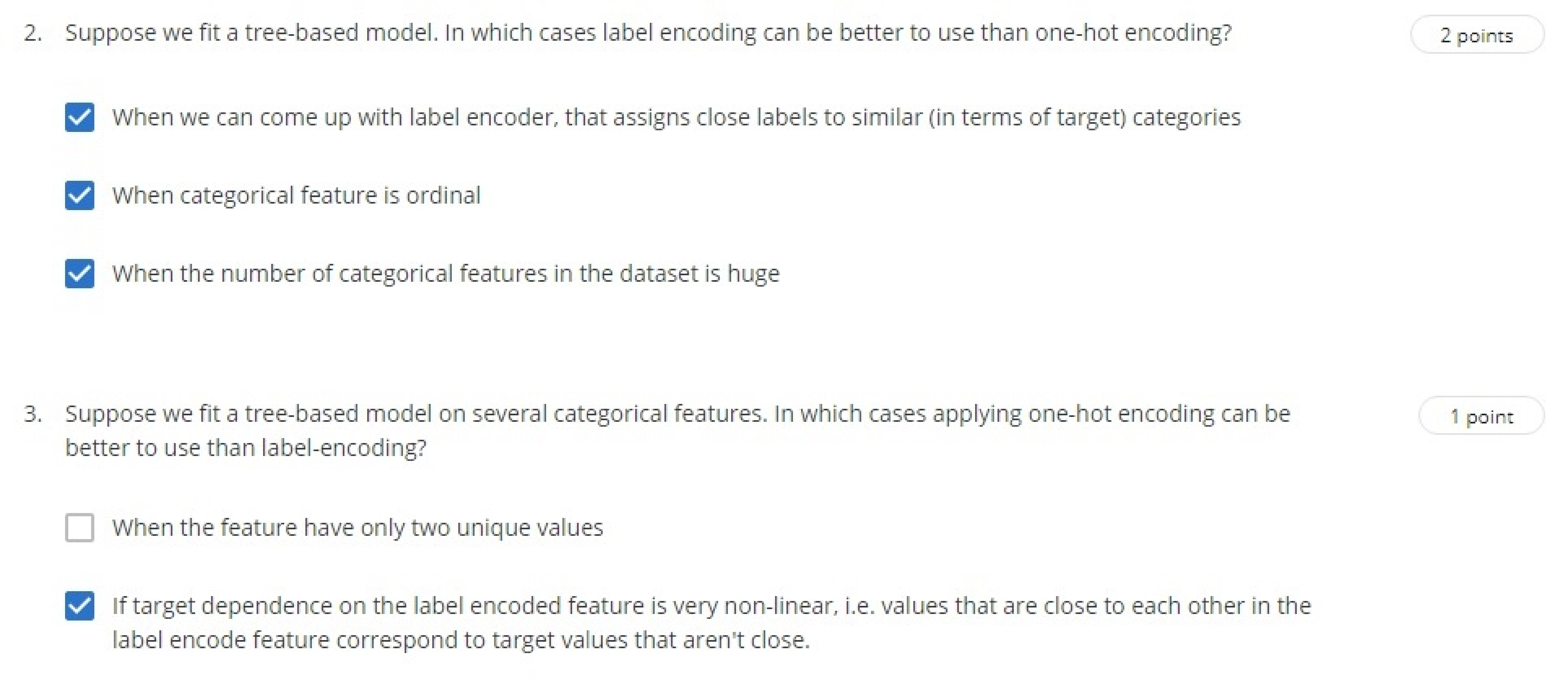
☑ Apply rank transform to the features

☐ MinMaxScaler

☑ Apply **np.sqrt(x)** transform to the data

☑ Winsorization

☑ Apply **np.log1p(x)** transform to the data



2. Suppose we fit a tree-based model. In which cases label encoding can be better to use than one-hot encoding?

2 points

☑ When we can come up with label encoder, that assigns close labels to similar (in terms of target) categories

☑ When categorical feature is ordinal

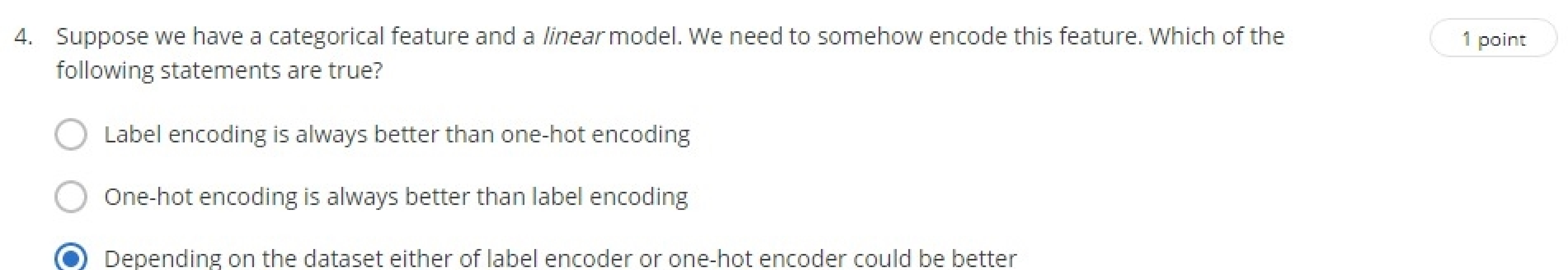
☑ When the number of categorical features in the dataset is huge

3. Suppose we fit a tree-based model on several categorical features. In which cases applying one-hot encoding can be better to use than label-encoding?

1 point

□ When the feature have only two unique values

☑ If target dependence on the label encoded feature is very non-linear, i.e. values that are close to each other in the label encode feature correspond to target values that aren't close.



4. Suppose we have a categorical feature and a *linear* model. We need to somehow encode this feature. Which of the following statements are true?

○ Label encoding is always better than one-hot encoding

○ One-hot encoding is always better than label encoding

● Depending on the dataset either of label encoder or one-hot encoder could be better