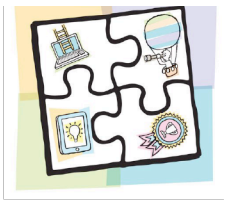


Guide to Data Quality

1. Contributors

Dr. Geraldine Gray is a lecturer with the Informatics Department at the Institute of Technology Blanchardstown. She specialises in the design and delivery of modules in the areas of data analytics, text analytics, business intelligence, and enterprise application development at levels 7, 8 and 9, and is course coordinator for ITB's online masters in Applied Data Science and Analytics. Geraldine has also supervised a number of post-graduate research students. Prior to joining ITB, Geraldine lectured in IT Tallaght, and also has a number of years of industrial experience developing software for distribution and inventory management. Research interests and peer-reviewed publications include educational data mining and learning analytics, data analytics for computer forensics and analysis of unstructured data.

2. Introduction

A successful data analytics project requires a clean sample of data that is representative of the population of interest. The data typically comes from integrating a number of data sources, and the resulting dataset is rarely clean. This document discusses data quality issues that commonly arise in such as dataset, with examples from data typical of Higher Education.

3. Using a histogram to identifying unusual values or distributions

A histogram depicts the range of values for a particular attribute. For example, Figure 1 is a histogram of CAO points. The height of each bar indicates the number of students falling within a particular range of CAO points.

A histogram can help identify unusual values in a dataset, or identify a sampling bias. CAO points should be in the range [0,625]. The presence of a number of '999' values is clearly evident in Figure 1. As these are outside the expected range, they warrant further investigation.

The histogram also highlights a number of students with low CAO points (<40). This could be indicative of a subgroup of mature student for whom prior academic performance was not available. Knowledge of the domain can help determine if this proportion of low values is correct for the student population of interest. Similarly, the overall distribution of CAO points should be representative of the expect range for the student population being studied.

4. Handling missing values

Missing values is a common occurrence across datasets. Typically a low number of missing values can be dealt with by deleting those rows with missing data. Attributes (columns) with a high proportion of missing values should also be deleted. A problem arises when the option to delete is not feasible because of the amount of information that would be lost.

There is a range of techniques for 'filling' missing values; none are ideal. The goal when filling missing values is NOT to estimate the missing value, but to select a value that PRESERVES existing patterns in the data. The more common approaches include the

following:

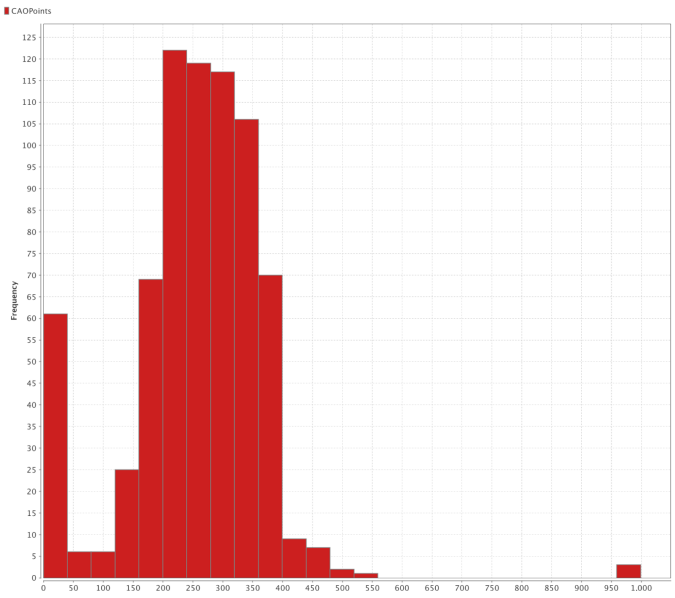
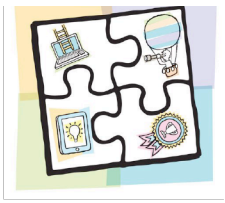


Figure 1. Histogram of CAO points

1. The simplest approach replaces missing values with the average value for that attribute, typically using the mean for numeric attributes and mode for nominal attributes. This approach ignores relationships between attributes.
2. A variation on approach 1 above is to divide the dataset into subgroups first, for example based on the class label[1]. A missing value is replaced with the average value within its subgroup. This can preserve relationships with the class label. However, it may also falsely strength such relationships.
3. A more complex approach is to use imputation, i.e. use a classification or prediction algorithm to replace missing values. The attribute with missing values becomes the class label or dependent variable; rows without missing values are used to train a model; this model is then applied to rows with missing values to predict the missing values based on existing patterns in the dataset. The advantage over the previous two techniques is that all inter-variable relationships can be considered. However, like approach 2 above, such relationships may be over-emphasised with this approach, and resulting model accuracies can be over estimated.

To illustrate the potential impact on model accuracy of each of the three approaches above, Table 1 shows the result of applying each technique to a dataset of 500 students. *CAOPoints*, *age* and *gender* were used to train two types of models: a regression model predicting first year GPA; and classification model predicting a *pass* (GPA>=2) or a *fail* (GPA<2.0). The original dataset had no missing values. 23% of values for *CAOPoints* were randomly selected for deletion to simulate missing values. Models were trained on the



original dataset, the dataset with missing values, and the dataset generated from each of the three approaches above.

[1]When training a classification model, the class label is the attribute the model is trained to

With the exception of replacing values with the subgroup mean for regression modelling, all approaches overestimated model accuracy. Replace by imputation gave the most realistic results for regression; replace by mean give the most realistic results for classification with *k*-Nearest Neighbour (*k*-NN).

Table 1. A comparison of model accuracies using different techniques to replace missing values

Predicting GPA using Linear regression; 500 rows, 114 with missing values:

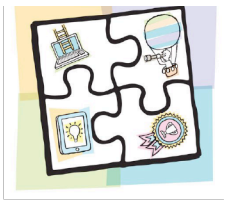
	Original dataset	**Dataset with missing values (not replaced)**	**Replace with mean**	**Replace with imputation (regression)**	**Replace with subgroup mean**
RMSE	0.981	0.988	0.986	0.982	0.29
R^2	0.131	0.103	0.118	0.118	0.924

Predicting *Pass* / *Fail* using *k*-NN with k=3; 500 rows, 114 with missing values:

	Original dataset	**Dataset with missing values (not replaced)**	**Replace with mean**	**Replace with imputation (*k*-NN)**	**Replace with subgroup mean**
Accuracy	63.40%	66.80%	64.20%	66.60%	65.80%
Recall on fail	48.09%	51.91%	49.18%	51.37%	51.37%
Geometric Mean	59.00%	63.00%	60.00%	62.00%	62.00%

5. Working with unbalanced datasets

An unbalanced dataset is one where a particular subgroup is underrepresented. For example, a class with a small failure rate will have few examples of the behaviour



patterns of failing students. This may cause a problem when building a classification model as the minority class can be ignored without significant impact on model accuracy. For example, a simple prediction that all students will pass, applied to a class group with a 10% failure rate, will achieve 90% accuracy.

There are two things to consider which will be discussed below:

1. Use of a pre-processing techniques to rebalance the dataset.
2. Appropriate measures of model accuracy for an unbalanced dataset

Rebalancing the dataset can be done in two ways:

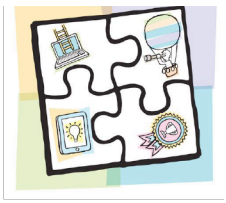
1. If the dataset is large enough, the majority class can be under-sampled so that it is a similar size to the minority class. For example, suppose a dataset of 10,000 students has 500 examples of 'fail'. The 9,500 sample of students that '*passed*' can be sampled to a size close to 500, resulting in a balanced dataset of about 500 in class '*pass*' and 500 in class '*fail*'. This option is viable if the resulting dataset has sufficient data to represent the patterns of both classes.
2. The alternative option is to oversample the minority class. For example, this could be achieved by replicating rows in the minority class using bootstrap sampling, or by creating synthetic instances of the minority class using a techniques such as SMOTE (Synthetic Minority Over-sampling Technique). There are drawbacks to both approaches. Replicating the minority class can result in over fitting a model, as examples used to test the model may also have been used during training. Although it is possible to over-sample the training dataset only. Creating synthetic samples does not necessarily produce a sample representative of the population of interest.

There are a number of measures of model accuracy, some of which are more appropriate for estimating accuracy of a model applied to a test dataset that has class imbalance.

1. The simplest approach is overall model accuracy, calculated as the number of correct predictions divided by the total number of instances. However, this does not give an indication of model accuracy for individual classes.
2. Accuracy of individual classes can be recorded in a number of ways, most commonly as *precision* or *recall*. *Precision* is the number of predictions that are correct for a particular class; *recall* is the actual number of instances in a class that were predicted correctly.
3. Geometric mean is an overall model accuracy calculated from the root of the product of each class recall. Each class, regardless of size, has equal weight in the overall accuracy calculation.

Calculations for each of these are given below, based on the confusion matrix in Table 2. The confusion matrix shows the results of a classification algorithm applied to 680 students. 180 students failed, of which the algorithm predicted 120 correctly. 500 students passed, of which the algorithm predicted 450 correctly.

Table 2. Confusion matrix



	Predicted *pass*	**Predicted *fail***
Actual *pass*	450	50
Actual *fail*	60	120

Overall model accuracy: $\frac{450+120}{450+120+50+60}$ = 0.84 (84%)

Recall on *fail*: $\frac{120}{120+60}$ = 0.67 (67%) Precision on *fail*: $\frac{120}{120+50}$ = 0.71 (71%)

Recall on p*ass*: $\frac{450}{450+50}$ = 0.9 (90%) Precision on *pass*: $\frac{450}{450+60}$ = 0.88 (88%)

Geometric mean: $\sqrt{0.67 * 0.9} = 0.77\ (77\%)$

6. Recognising attributes with poor information content

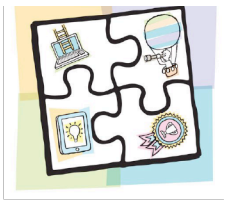
In data analytics, a role of nominal attributes (non numeric) is identification of subgroups in a dataset for which specific patterns can be identified. For example, defining sub groups based on gender, course of study, or age band. If a nominal attribute has a different value for every row of data (i.e. the most frequent value occurs once or twice) then it is not useful in identifying subgroups. Conversely, if a nominal attribute has a mode close to the sample size, it has similarly poor information content as virtually every row has the same value. A numeric attribute with an unusually low standard deviation may similarly indicate a lack of variability resulting in poor information content. Examples of each scenario are given in Table 3.

7. A note on data types

Some algorithms, such as linear regression, SVM or neural networks, require all attributes to be numeric. Avoid the temptation to convert a nominal attribute to a numeric one. This infers relationships between values that may not be valid. For example, mapping *course name* to a numeric value as in column 2 of Table 4 infers that *law* is more similar to *computing* than *engineering*, which is invalid. However, representing some numeric characteristic about the course, such as *numbers enrolled* or *CAO points*, is a valid, numeric description of the course.

Table 3. Attributes with low information content

Attribute	**Least frequent value**	**Most frequent value (mode)**	**Comment**
Gender	Female (2)	Male (500)	Poor information content: Most rows have the same value of 'male'. This may also indicate a



			sampling bias.
Student ID	B00001234 (1)	B00004321 (2)	Poor information content: Each row has a different value.
Course	Law (80)	Business (200)	Identifiable subgroups.

Attribute	Mean	Standard deviation	Comment
GPA	2.1	0.001	Poor information content: All examples have similar GPA.
GPA	2.1	1.3	Mean and standared deviation indiacte a resaonble distribution for GPA given a valid range of [0,4] .

Table 4. Converting a nominal attribute to numeric

	Invalid numeric description	**Valid numeric descriptions**	
Course	**Map to numeric**	**Numbers enrolled**	**CAO points in year *X***
Engineering	1	100	450
Computing	2	150	460
Business	3	300	350
Law	4	80	500

8. How much data is enough?

A learning analytics project based on a small number of classes may not have sufficient examples of all possible patterns to accurately model that dataset. The gives rise to the question: how much data is enough? There are some simple rules of thumb around the number of rows being a multiple of the number of attributes (e.g. 20*x*). This may work for a small number of attributes, but is more difficult to estimate for larger number of attributes that may have many possible distinct values. One approach that can indicate an appropriate sample size is 'progressive sampling', which works as follows:

Starting with a small sample, estimate the accuracy of models produced by that sample size using an approach such as cross validation. Gradually increase the sample size, estimating model accuracy at each step. Initially, model accuracy will **change** as the sample size increases. At some point. model accuracy should converge. Lack of convergence in model accuracy indicates insufficient data. Figure 2 depicts model

accuracy of sample ratios from 5% up to 100%. Model accuracy started to converge once the sample size was 80% of the full dataset.

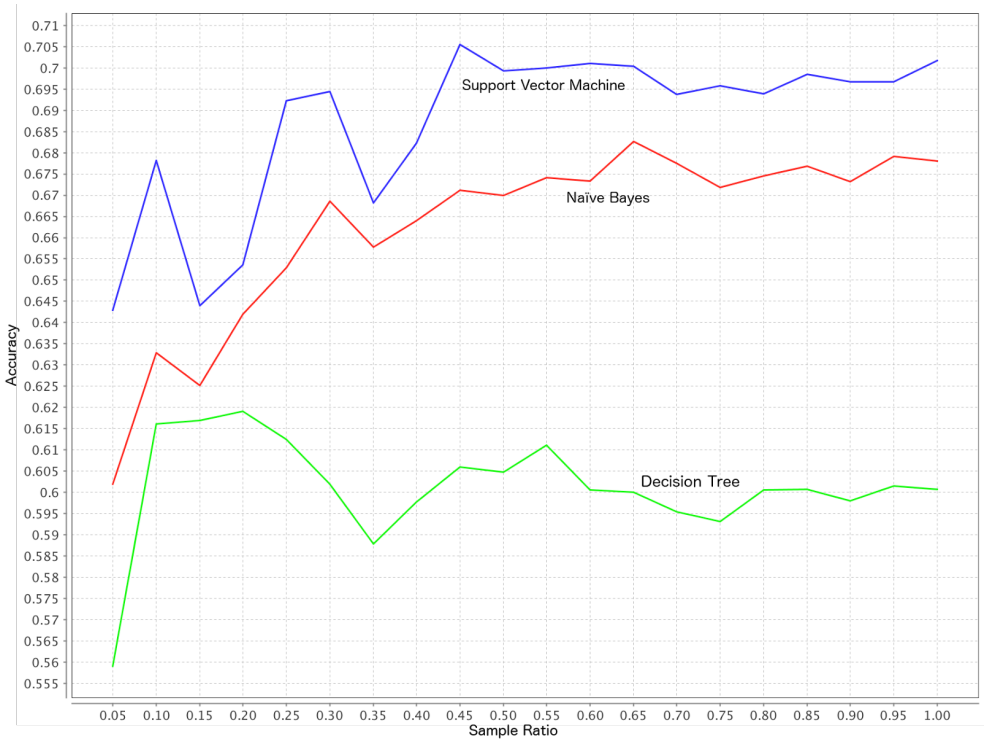
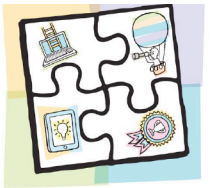


Figure 2. Progressive sampling

9. Conclusion

Real data in rarely clean. Factors discussed in this report of bias, invalid values, imbalanced data, insufficient data, and attributes with low information content, will all impact on resulting model accuracy. However, an awareness of quality issues, and an understanding of the impact of approaches to 'fix' quality issues, will ensure interpretation of resulting models is more realistic. It is important to understand that outputs from learning analytics are not infallible; they are only as good as the quality of data and methods used to generate them. To quote the statistician George Box, "*all models are wrong, but some are useful*"[2]. Our role in learning analytics is to ensure our interpretation of data is indeed useful.



[2]Referenced in: Bergner, Y. 2017. Measurement and its Uses in Learning Analytics. In C. Lang, G. Siemens, A. Wise, and D. Gavšević, Eds. *Handbook of Learning Analytics.* 1st ed. SoLAR, p. 34-48. Available online at: https://solaresearch.org/hla-17/