Let the data speak: Machine Learning methods for data editing and imputation

Paper by: Felibel Zabala

Presented by: Amanda Hughes

September 2015

Objective

- Machine Learning (ML) methods can be used to help us analyse and understand erroneous data and non-response in various data collections.

- This presentation aims to communicate machine learning methods we have explored to assist in developing sound editing and imputation methodology using Statistics New Zealand's Household Economic Survey as a case study.

Context: Statistics New Zealand

- The Household Economic Survey is currently migrating to the Household Processing Platform.

- Edit rules are currently being developed.
  - Editing system will have a contextual editor that provides users with a relational view of data requiring manual editing.
  - To assist in the development of the contextual editor, we are exploring the use of association data mining in relation to the creation of editing rules.

Association rule mining: Introduction

- Introduced by Agrawal et al in 1993.

- Originated from analysing a market basket of transactions to generate association rules that describe which items from transactions tend to occur together.
  - Item associations are generated based on:
    - The strength of the association,
    - The frequency of the occurrence and,
    - The predictive utility of the relationship.

Association rule mining: Definition

- Association rule:
  - An implication expression of the form X → Y, where X and Y are disjoint itemsets.
  - Example:

    {Milk, Diapers} → {Beer}
- Rule evaluation measure:
  - Support (s):
    - Fraction of transactions that contain both X and Y.
  - Confidence (c):
    - Measures how often items in Y appear in transactions that contain X.

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coca Cola
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coca Cola

- Example:

  {Milk, Diapers} → {Beer}

  s = σ (Milk, Diapers, Beer) / |T| = 2/5 = 0.4

  c = σ (Milk, Diapers, Beer) / σ (Milk, Diapers) = 2/3 = 0.67

Association rule mining: Goal

- The goal of association rule mining is to find all rules with support and confidence above defined thresholds.
  - First generate all combinations of items whose support ≥ *minsup* (called frequent itemsets)
  - Then extract all the *high-confidence rules* from the frequent itemsets.

- The most popular algorithm used in association rule mining is the *Apriori* algorithm (*arules).*

Association rule mining: Limitation

- Association rules: applied to categorical data.
- HES: mostly quantitative data, so had to use ordinal association work around.

- Ordinal association rule mining is done using the following steps:
  - Find ordinal rules with a minimum confidence (using a version of the *Apriori* algorithm).
  - Identify data items that break the rules and can be considered outliers.

Association rule mining: HES data

- Investigated using unedited HES data consisting of 4,292 records described by 33 attributes.
  - Used ordinal rules to illustrate identification of outliers.
  - Age and income are converted to ordinal attributes
    - Age into five-year age groups for 15-64 and 65+
    - Income into percentiles of the income distribution of the data set.

Association rule mining: HES data, results

- Association rules with minimum confidence equal to 0.15 were extracted.
  - Association rule: Highest qualification → Total income

Highest educational qualification	Income group					
	<P10	P10-P25	P25-P50	P50-P75	P75-P90	>P90
No qualification	0.136	0.174	0.349	0.234	<0.15	<0.15
Some secondary school certification	0.456	0.222	0.000	0.000	<0.15	<0.15
Level 4 certificate	<0.15	<0.15	0.221	0.332	0.187	<0.15
Levels 5&6 certificate	<0.15	0.15	0.237	0.260	0.177	<0.15
Bachelor degree & Level 7 certificate	<0.15	<0.15	0.206	0.216	0.219	0.217
Postgraduate and honours degree	<0.15	<0.15	<0.15	0.268	0.298	0.169
Masters degree	<0.15	<0.15	<0.15	0.268	0.183	0.235
Doctorate degree	<0.15	<0.15	<0.15	<0.15	0.351	0.378

- Association rule: Age → Total income

Age group	Income group					
	<P10	P10-P25	P25-P50	P50-P75	P75-P90	>P90
Age15_19	0.549	0.265	0.190	<0.15	<0.15	<0.15
Age20_24	0.471	0.258	0.364	0.185	<0.15	<0.15
Age25_29	<0.15	<0.15	0.234	0.369	<0.15	<0.15
Age30_34	<0.15	<0.15	0.258	0.288	0.171	<0.15
Age35_39	<0.15	<0.15	0.237	0.234	0.176	<0.15
Age40_44	<0.15	<0.15	0.182	0.252	0.231	<0.15
Age45_49	<0.15	<0.15	0.256	0.268	0.163	0.171
Age50_54	<0.15	<0.15	0.221	0.278	0.161	0.150
Age55_59	<0.15	<0.15	0.249	0.263	0.190	<0.15
Age60_64	<0.15	<0.15	0.269	0.319	0.163	<0.15
Age65+	0.182	0.227	0.256	0.187	<0.15	<0.15

Household Economic Survey (HES)

- HES income + expenditure (+ wide range of demographic information)
- For a personal income questionnaire to be a response in the current HES, all key questions must have a valid answer.
- Current method: nearest neighbour donor imputation.

Household Economic Survey (HES)

- Proposed methodology: same imputation methodology but income questionnaire divided into three modules:
  - Jobs module, government transfers module and the investment module.

- A previous SNZ project investigated and recommended the use of a ML method as a standard tool to create homogeneous imputation classes.

Classification methods for imputation: Classification and Regression Trees

- Imputation is done within homogeneous classes to minimise the potential non-response bias.

- Sometimes a large number of variables are available to form imputation classes
  - A Statistics New Zealand project proposed the use of decision tree (or classification) learning methods (CART) to determine the useful variables to create imputation classes.

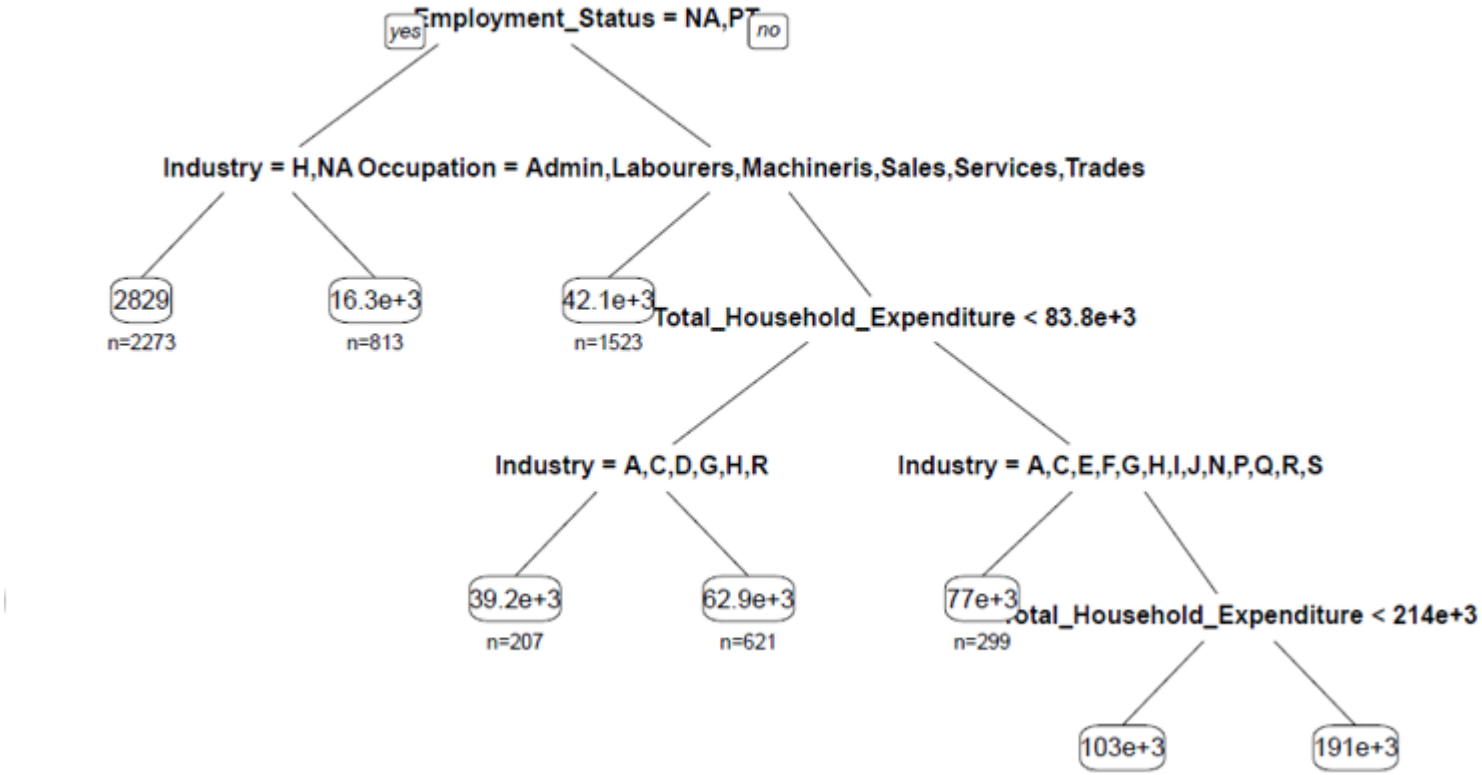
Classification and Regression Trees

- The basic structure of a classification or a regression tree consists of a root node which grows through a series of splits to create terminal nodes.
- Different criteria are used for splitting

Classification and Regression Trees: Case study

- CART was used to explore matching variables needed to impute for missing income from the three income modules using the nearest neighbour imputation methods.
  - Data from HES 2010/11 was used as training samples to identify the regression tree that will be used for future data.
  - Used the R-package, *Rpart (Recursive Partitioning)*.

Regression Tree for Jobs



Classification and Regression Trees: Case study

- The complexity parameter is used to control the size of the classification tree and to select the optimal tree size.
- Tree construction does not continue unless it would decrease the overall lack of fit by a factor of cp.
- If a smaller cp was used, say, 0.001, more branches would have been produced which would result in the use of more imputation cells.

Classification and Regression Trees: Case study, weights.

- To find the corresponding weights, we used random forests (R-package *randomForest).*
- Random forests can be used to rank the importance of variables in a classification problem.
- A random forest combines the predictions made by multiple decision trees, where each tree is generated using a random vector generated from some fixed probability distribution

Classification and Regression Trees: Case study, results

Variable	Weight
Employment status	10
Industry classification (for main job)	7
Occupation (for main job)	6
Total household expenditure	5
Age	5
Total income from government transfers	5
Labour force status	4
Highest educational qualification	3
Total investment income	1

Classification and Regression Trees: Case study, evaluation

- Results evaluated, using HES 2012/13 data.
- 5% and 10% of missing income records were simulated from the true income records.
- Canceis was used impute the missing records using the matching variables recommended.
- The presence of bias was measured comparing the imputed data to the true observed data.
- RESULT: no evidence of bias for the imputed values.

Conclusion

- We have found useful applications of machine learning methods for data editing and imputation.
- Association rule mining allows us to extract some rules from datasets.
- Classification methods allow us to create homogeneous imputation classes, and enable us to determine efficient matching variables for use in nearest neighbour imputation.

Future work

- In the future we plan to explore the use of cluster analysis for data editing and imputation.
- Future training on data editing and imputation will include relevant topics on machine learning methods.
- We also plan to develop a user interface that will make it easier to investigate machine learning methods for data editing and imputation.