Session 4 | Machine learning methods in survey statistics

Sponsored by the International Association of Survey Statisticians (IASS)

Introduction to session 4: Overview of machine learning in survey research

Natalie Shlomo
IASS President 2023-2025

University of Manchester, United Kingdom

Topics

- Overview
- Two Case studies
- Final Thoughts
- Introduction to Session 4

Areas of Survey Lifecycle with Machine Learning Potential

Survey Methodology:

- optimising data collection and adaptive/responsive survey designs
- agent based modelling/ microsimulations in the context of designing the survey experience
- predicting web / nonresponse breakoff in online web surveys for interventions

Survey Statistics:

- nonresponse classification and weighting
- imputations for unit and item nonresponse
- data integration
- statistical data editing and imputation
- satellite imagery (agriculture, building units)
- automatic coding
- small area estimation

Overview

- Buskirk, et al. (2018) provides an overview of machine learning (ML) in survey research
- Machine learning can be supervised (training with labelled data) or unsupervised (training with unlabelled data)
- Supervised learning is typically used to produce a prediction for some dependent variable while unsupervised learning might focus on pattern detection, eg. cluster analysis
- ML are algorithmic and data-driven requiring tuning parameters, eg. number of clusters, penalty parameter (amount of shrinkage) in LASSO, number of nodes in tree-based methods
- Need to make the distinction between inference and exploratory/prediction, where the latter is generally the focus of ML and here we can maximize its utility
- Essential to prepare good training data to avoid selection and algorithmic biases over time (See: Kern, et. al. (2023) discussing the impact of the annotation instrument on downstream model performance and predictions)

Overview

- Examples in Buskirk, et al. (2018) regarding mitigating negative consequences of nonresponse or item missingness:
  - For responsive survey designs, to obtain an accurate classification of which sampled units are likely to respond to a survey and which are not
  - For online survey panels, to know which respondents are likely to leave an item missing or break-off on a questionnaire and which respondents are not

- Evaluation of predictive models relies on cross-validation (to avoid overfitting):
  - Take a sub-sample of the data as the training sample, develop a predictive model, the remaining sample is the test sample and is used to evaluate accuracy (can also include a third subsample for tuning purposes)
  - Accuracy quantified through a MSE or sensitivity/specificity for classification problems

Overview

- Puts and Daas (2021) discuss ML in the context of official statistics stating that ‘applying ML learning algorithms to produce official statistics is still challenging’
- Quality standards required in official statistics and the challenges of their context with ML:
- Accessibility and Clarity:
  - Challenge on making clear how results are exactly obtained for many ML algorithms, eg. Deep learning and other neural network methods, since some of them are essentially a ’black box’
  - Use a functional approach, maybe extended with a mechanistic approach, and determine what happens when the model is ’fed’ small chunks of data.
- Coherence and Comparability:
  - How well the model is able to give a stable result over time and the correlations it has found

Overview

- Accuracy and Reliability:
  - ML algorithms can suffer from biases, eg. the annotated data set used for training (and testing), the representativeness of this data set and misclassification can bias the model developed
- Challenges still to be resolved:
  - Methodology concerning the human annotation of data
  - Sampling the population to obtain representative training sets
  - Using stratification in the context of Machine Learning
  - Data structure engineering and selection to increase the transparency of models
  - Reducing spurious correlations
  - Methodology for studying causation
  - Correcting the bias caused by the ML model
  - Dealing with concept drift (representativity over time)

Case Study 1

Natural language processing for automatic coding to predict occupation, economic activity and other classifications (Evans and Oyarzum (2021) and internal information)

- Some countries, eg. Statistics Canada, are investigating *fastText*: a neural network library for learning of word embeddings and text classification created by Facebook's AI Research lab
- *fastText* has the advantage that it works on word and n-gram embeddings and provides a score on the prediction confidence
- Traditionally, automatic coding split into two streams: manual and automatic using a variety of algorithms, eg. G-CODE (Wenzowski, 1988) and Cascot (calculates a score 0 to 100 as the probability that the code is correct) (Warwick Institute for Employment Research)
- Move to 100% automatic coding with optimal sampling methods for verification by human coders for quality assurance based on prediction score

Case Study 1

- Sample allocation (eg. stratified by prediction scores) constrained to desired level of accuracy, costs and maximum workload, to obtain output prediction error rates and update labelled data
  - To avoid risk of algorithmic biases, labelled data should only use verified codes
- In a simulation by Statistics Canada on occupation, out of 121,000 workload, 71,600 manually coded under traditional approach and 52,200 under new approach with approximately same prediction error rate
- Some caveats:
  - Use of black box *fastText* requires a good understanding of how algorithm works (i.e. should data be sorted or not)
  - Algorithm only works well if there is very high quality pre-processing and labelled training data is complete and also of high quality
  - Maintain the skills of human coders and preserve this knowledge

Case Study 1

- Other areas before putting into production:
  - Include how to obtain high-quality training data sets, how to monitor model decay once deployed, and how to develop user interfaces
  - Need to also assess the quality framework of automatic coding with respect to five criteria: explainability (understanding what causes a model to make particular decisions), accuracy, reproducibility, timeliness, cost effectiveness

Case Study 2

Predicting web survey breakoffs using machine learning models (Chen, Cernat and Shlomo, 2022)

- Survival models: Cox (commonly used to understand patterns of breakoffs) vs ML Lasso-Cox
- If we ignore the clustering of the questions within persons, explore other predictive ML methods (traditional and LASSO logistic regression, random forest, gradient boosting, and support vector machine)
- Compare best performing survival model with the best performing classification model to investigate whether considering the clustered data structure by the survival model improves breakoff prediction performance
- Another research question in the paper looked at types of time-varying question-level predictors:
  - 3 sets of covariates and inclusion of all of them: Demographics (Age, education,- ethnicity, student status,- marital status); concurrent (responding device, item missing, matrix question, open-ended question, question topic, and question word count), cumulative (as above but aggregated across questions, and number of times logged into survey)

Case Study 2

- Data: repeated, cross-sectional non-probability web survey administered to members of the Lightspeed Panel, an opt-in web panel in the United States
- First wave conducted between September and October 2019 while second wave collected in October 2020.
- The survey is considered appropriate to analyse:
  - Recorded breakoffs - out of the 3128 and 2370 respondents in the first and the second wave, 520 and 403 quit the survey without completing it, resulting in a breakoff rate of around 17% for both waves
  - Recorded last question respondents completed, meaning that the breakoff position is known.
  - Breakoff pattern of both waves is very similar so wave 1 is the training set applied to wave 2 in the cross-validations

Case Study 2

- Traditional Cox model performs better than LASSO Cox in predicting breakoffs, which holds true for nearly every predictor group
  - Traditional Cox has C-index between 0.68 and 0.85 across predictor groups, compared to 0.5 to 0.78 in LASSO Cox.
  - Best penalty values in LASSO Cox were close to zero except for the LASSO Cox fitted using only cumulative time-varying predictors
- Ignoring clustering, gradient boosting gives the best prediction performance across all evaluation metrics (Sensitivity, AUC, Accuracy, Specificity and Precision)
  - Gradient boosting focuses on correcting for prediction errors made by models in previous iterations and thus over time, the model makes fewer prediction errors resulting in better prediction performance

Case Study 2

Conclusions from the study:

- Comparing Gradient boosting to traditional Cox, Gradient boosting slightly better in AUC showing that considering the clustered data structure does not translate into a significant improvement in breakoff prediction
- Using values of time-varying predictors concurrent to the breakoff status is more predictive of breakoff, compared to aggregating their values from beginning of the survey, implying that respondents' breakoff behaviour is more driven by current response burden

Final Thoughts: Embedding ML into the Organization

- Develop skills training and capacity building in Data Science
- In-house expertise to evaluate emerging ML methods and take a lead in their development
- Engage with all parties (data scientists, methodologists, and subject matter experts) in discussions on applications
- Start with small projects demonstrating proof of concept and the willingness of the organization to try new methods
  - These can be followed by taking ML models into the production pipeline
  - Current applications of ML are generally well suited for prediction
  - Need to be aware of research as it evolves in allowing for statistical inference
- Get involved with international research collaborations
- Develop a quality framework for ML in official statistics

References

Buskirk, Trent D., Antje Kirchner, Adam Eck, and Curtis S. Signorino (2018) An Introduction to Machine Learning Methods for Survey Researchers. *Survey Practice* 11 (1). https://doi.org/10.29115/SP-2018-0004.

Chen, Z., Cernat, A. and Shlomo, N. (2022) Predicting web survey breakoffs using machine learning models. *Social Science Computer Review,* Vol 41 (2), 573-591.

Evans, J. & Oyarzun, J. (2021). "Need for Speed: Using *fast*Text (Machine Learning) to Code the Labour Force Survey", paper presented at *Symposium Annual Meeting - Proceedings of Statistics Canada Symposium 2021*.

Kern, C., Eckman, S., Beck, J., Chew, R., Ma, B. and Kreuter, F. (2023) Annotation Sensitivity: Training Data Collection Methods Affect Model Performance. Available at: https://arxiv.org/abs/2311.14212

Puts, M.J.H. and Daas, P.J.H. (2021) Machine Learning from the Perspective of Official Statistics, *The Survey Statistician*, Vol. 84, 12–17. http://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2021_July_N84_02.pdf

Introduction to Session 4

2:00-3:00 Machine learning procedures for the treatment of unit non-response in surveys
David Haziza, John Tsang, Khaled Larbi, Mehdi Dagdoug
and discussion

3:00-3:20 Coffee Break

3:20-3:35 State of play and perspectives on machine learning at ISTAT
Marco Di Zio

3:35-4:35 Machine learning in official statistics: towards statistical based machine learning
Marco Puts and Petrus J.H. Daas
and discussion