Machine Learning Methods, a Case Study Using an Online Web-Based Panel Survey

Yulei He*, Guangyu Zhang, Van Parsons

Division of Research and Methodology
National Center for Health Statistics
U.S. Centers for Disease Control and Prevention

**wdq7@cdc.gov*

May 17, 2021

Disclaimer: The findings and conclusions in this study are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention

Outline

- Motivation
- Background and Study Design
- Methods
- Results
- Concluding Remarks

Motivation

- Machine learning (ML) methods: an array of computer-intensive data science methods aiming at discovering patterns in data
  - Often nonparametric
  - Focus on computation and algorithms
  - Able to handle high-dimensional and complex data (e.g., with large data volume and/or many predictors, number of predictors far exceeds the sample size, as well as complicated features such as image data)
- General categories of ML
  - Supervised ML: to optimally predict or classify an outcome variable
  - Unsupervised ML: no specific outcome and to detect patterns existing in the data (e.g., cluster analysis)
- ML methods often emphasize prediction more than statistical inference
- ML methods have been widely applied to different data sources and scientific problems

Machine Learning Applied to Survey Data

- Surveys are often used to collect information from human subjects for population studies
- ML methods have begun to be applied for survey data, for example,
    - Modeling and predicting survey unit nonresponse
    - Survey item nonresponse imputation
    - Model-assisted survey estimation
    - Creation of synthetic data
    - Automatic coding from open-ended questions
    - Record linkage
- The relevant literature is increasing, for example, Survey Practice (https://www.surveypractice.org/issue/590), has a full issue devoted to introducing ML methods for survey practitioners

Machine Learning for Predicting Future Survey Data

- Many established surveys are conducted regularly and use the same (or similar) design and variables
- Can ML be used to predict survey data from future rounds based on the data and models from existing rounds?
- Can these predictions be used to generate summary statistics of variables?
- The goal of this project is to explore these ideas using a case study

Data Background: RANDS

- Research and Development Surveys (RANDS) (https://www.cdc.gov/nchs/rands/index.htm)
- A series of recruited probability-sampled commercial panel surveys started in 2015 (ongoing)
- Conducted by National Center for Heath Statistics at CDC and contracted to external vendors for data collection
- Surveys largely utilize the web mode
- Survey questions focus on a range of health-related topics including chronic conditions, healthcare access and utilization, opioid use, and COVID-19; each RANDS survey has its own focus

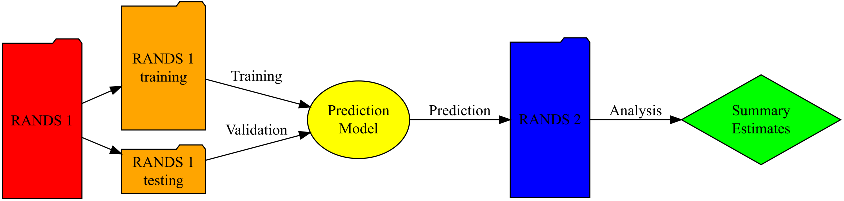
RANDS 1 and 2

- We used RANDS 1 and 2 to demonstrate the idea
- RANDS 1 (fall of 2015) and RANDS 2 (spring of 2016) data were collected by Gallup using the web mode
- RANDS 1 had 2304 completed interviews (completion rate 24%) and RANDS 2 had 2480 completed interviews (completion rate 30%)
- RANDS 1 and 2 were from the same survey panel and had almost identical questionnaires and sampling designs
- RANDS 1 and 2 data were sampled independently and were not designed to be longitudinal
- Question topics included demographics, healthcare access and utilization, chronic conditions, food security, general health, health insurance, physical activity, psychological distress, and alcohol/tobacco use
- Public-use data for RANDS 1 and 2 are available in https://www.cdc.gov/nchs/rands/data.htm

Study Design

- General idea: ML methods were developed based on RANDS 1 data and were assessed with regard to predicting data from RANDS 2
- Since the true RANDS 2 data are known, they can be used as the gold standard in the evaluation
- The performance metrics include those for individual predictions as well as those for making summary estimates using individual predictions
- Variable estimates focus on the summary statistical information (e.g., means and standard errors)

Study Design Diagram



Method: Dependent and Independent Variables

- We use public-use data from RANDS 1 and RANDS 2
- Dependent variable (label): for demonstration, we use body mass index (BMI) as a continuous dependent variable in this talk
- Independent variables (features):
  - Demographic variables (e.g., age, sex, race, region, education, income, marital status, employment status, housing status)
  - Health conditions (e.g., chronic conditions such as diabetes, hypertension, asthma, emphysema/chronic bronchitis or COPD, whether taking medications for some conditions, and self-rated health status)
  - Access to healthcare (e.g., health insurance coverage, delayed or can't afford healthcare, using internet for health information and appointments)
  - Health behaviors (e.g, tobacco use, alcohol use, physical activity)
  - Psychological distress variables
  - Sampling design variables (final survey weight)
  - Total 68 variables (survey questions), which can be dummy coded to 120 factors

Method: ML Methods

- We use several established ML methods (Boehmake and Greenwell 2020; James et al. 2015; Hastie et al. 2008)
- Linear regression
- Subset regression
- Regularized regression (Ridge regression, LASSO, Elastic net)
- Principal component regression (PCR)
- Partial least squares (PLS)
- Multivariate adaptive regression splines (MARS)
- K-Nearest Neighbours (KNN)
- Regression tree-based methods
    - Bagging, random forests, gradient boosting (GBM and XGBoost)
- Support vector machines (SVM)
- Deep-learning neural networks

Method: Setup of ML Application

- Divide RANDS 1 into the training sample (80%) and testing sample (20%)
- ML methods developed and tuned from the RANDS 1 training sample are applied both to the RANDS 1 testing sample and to the full RANDS 2 data to assess their performance
  - Evaluation using the RANDS 1 testing sample are for within-study prediction
  - Evaluation using the RANDS 2 data are for out-of-study prediction
- Performance measures include
  - Mean squared error (MSE): the average of squared deviations between predicted and observed values
  - Survey weighted estimates and standard errors for the dependent variable

Method: Complicated Issues in Pre-Processing

- Four outlier/extreme values in the dependent variable (e.g., BMI $<$ 10 or BMI $>$ 60)
- Missing values in both the dependent and independent variables
- Some continuous independent variables are nonnormal (e.g., highly skewed)
- How to deal with survey weights in ML?
- We follow some standard statistical practices as follows
  - Set the outliers or extreme values as missing
  - Impute the missing data using model-based multiple imputation and use one imputed dataset
  - Apply necessary transformation and standardization to continuous independent variables
  - Treat survey weights as an independent variable in ML methods and obtain weighted estimates in the evaluation

Method: A Few Practical Remarks in the Implementation of ML Methods

- All methods are implemented in R (sample code available upon request)
- For many ML methods, there exist alternative R packages in implementing the same method; we use the packages introduced in our references and our programs are not exhaustive
- Parameter tuning by cross-validation (e.g., 10-fold) is important in ML: some R packages can do the tuning automatically; otherwise we use the R caret package to do the tuning
- Parameter tuning for complex algorithms (e.g., gradient boosting) can take a long time
- If sometimes the program (e.g., for deep-learning) is not converging (e.g., MSE is not decreasing after certain iterations), a simple option is to change to a new random seed
- Testing different random seeds is necessary

Summary of the Results: MSE

Table 1: Prediction MSEs from RANDS 1 Testing Data and RANDS 2

Method	RANDS 1 testing	RANDS 2
Linear	30.52	31.76
Subset by stepwise	30.36	31.41
Subset by leap	30.79	31.42
Ridge	30.63	31.19
LASSO	30.55	30.93
Elastic	30.50	30.97
PCR	30.69	31.55
PLS	30.67	31.52
MARS	31.33	35.34
KNN	36.78	35.56
Tree	33.41	33.96
Bagging	32.58	32.21
Random Forest	30.88	30.25
GBM	30.32	30.19
XGBoost	30.83	30.44
SVM	31.37	31.33
Deep-learning	32.46	34.48

Summary of the Results: Estimates

Table 2: Relative Biases (%) of the Mean BMI Using Predictions for RANDS 2

Method	Unweighted (29.06)	Weighted (28.78)
Linear	-.65	-.69
Subset by stepwise	-.79	-.94
Subset by leap	-.93	-.94
Ridge	-.79	-.73
LASSO	-.76	-.63
Elastic	-.76	-.66
PCP	-.69	-.83
Partial	-.69	-.90
MARS	1.69	1.91
KNN	-4.51	-3.75
Tree	-1.07	-.38
Bagging	-.89	-.10
Random Forest	-.45	.00
GBM	-.55	-.24
XGBoost	-.41	-.14
SVM	-1.68	-1.25
Deep-learning	-.83	-.76

Summary of Results: Additional Remarks

- For linear regression, the R-square is around 30%
- For the prediction MSE from RANDS 2, the random forest and gradient boosting methods seem to reduce the prediction error the most compared with the linear regression (around 4% reduction)
- Some ML methods (e.g., KNN, MARS, and deep-learning) can increase the MSE compared with the linear regression; possibly due to overfitting/overparametrization since the sample size is only around 2000
- The range and scale of MSEs are similar between RANDS 1 testing sample and RANDS 2, suggesting that the prediction performance is consistent for within-study and out-of study
- The patterns are not sensitive to the random seeds in ML

Summary of Results: Additional Remarks

- Most of the ML methods preserve the mean estimates well with relative biases less than 1%
- However, the standard errors based on predictions (not shown) are considerably smaller (around 50% less) than those from the observed values
- The reduction of the variance from predictions is expected because the mean predictions are less noisy than the observed data conditional on the model (Little and Rubin 2020)
  - Need to be cautious if predictions are mechanically used to create summary estimates

Summary of Results: Important Predictors

- Unlike linear regression, most ML methods cannot provide coefficient estimates and standard errors for the features used
- For most ML methods, variable importance for the features can be determined using vip() in R
- The top important features have some overlap but can also be somewhat different across different ML methods
    - For linear regression, the top important features include self-rated health status, alcohol use, tobacco use, feeling worthless, and physical activity
    - For random forest, the top important features include self-rated health status, diagnosed diabetes, diagnosed hypertension, taking hypertension medication, age
    - For gradient boosting, the top important features include self-rated health status, diagnosed diabetes, family income, alcohol use, and the survey weight
- Subject-matter input may be necessary in choosing among ML methods

Histogram of Predictions Based on Random Forests

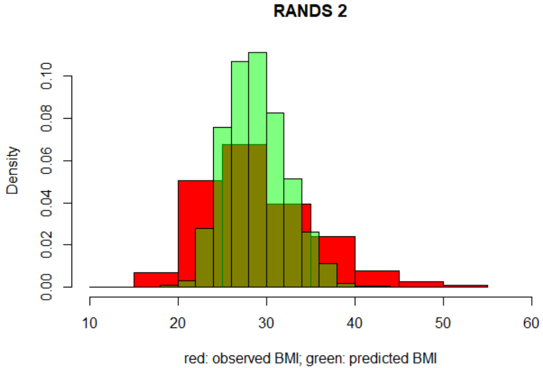


Figure 1: Histograms of Predicted and Observed BMI from RANDS 2. BMI: body mass index. RANDS: Research and Development Survey.

Scatter Plot of Predictions Based on Random Forests

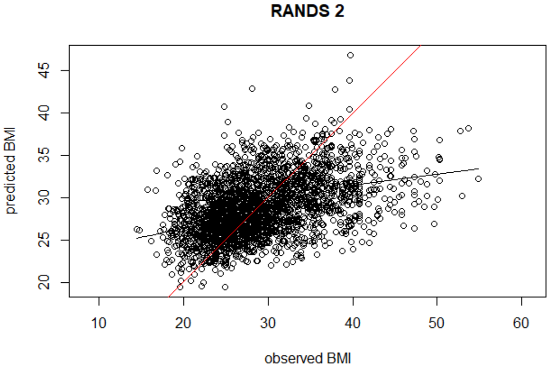


Figure 2: Scatter plot of observed vs predicted BMI from RANDS 2. BMI: body mass index. RANDS: Research and Development Survey. Red: 45 degree line; dark: the lowess curve.

Conclusion and Discussion

- Demonstrated the use of ML methods for predicting major health outcomes and for making estimates for surveys using a case study
- Survey estimates for the overall mean based on predictions are close to the original ones, but the standard errors are considerably reduced
- The performance of the predictions is impacted by sample size, quality of survey data (e.g., coverage/response/measurement error), and number of variables included
- Extend the research to other outcome variables (e.g., a binary outcome variable)
- Apply the methods to different surveys
- Methodological research areas in applying ML methods to survey data
    - The optimal method to handle survey nonresponse
    - How to incorporate survey weights and other design information
    - How to preserve the variation if predictions are to be used for making estimates

Major References

- Hastie, T., Tibshirani, R., and Friedman, J. (2009) The Elements of Statistical Learning. 2nd Edition. Springer.
- Boehmke, B. and Greenwell, B. (2020) Hands-On Machine Learning with R. Chapman and Hall/CRC.
- Gareth, J., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning. Springer.