Identifying Rhode Island characteristics and predicting whether a response came from a resident of Rhode Island or another New England state using 2017 BRFSS data

Rizki Syarif

***Abstract*— This is a short report analyzing a subset of the 2017 Behavioral Risk Factor Surveillance System (BRFSS) data from the Center for Disease Control and Prevention (CDC), comparing Rhode Island to other New England states and using machine learning methods to associate a set of responses to which state the participant is a resident of.**

I. INTRODUCTION

The 2017 BRFSS data [1] is a table compilation of responses from telephone interviews done with participants from across the United States, containing about 358 features or columns. Using this data, I present a brief analysis that identify features that distinguish Rhode Island from at least three other New England states, within a select subset of features. I also report evaluations of at least three classifiers that predicts a state given these subset of BRFSS features.

II. METHODS

The general strategy to identify features of Rhode Island compared to other New England states are as follows. The BRFSS data is re-processed and fit using a statistical model that associates the features to the state where the participant belongs to. More specifically, the problem is constructed as as a binary classification problem where Rhode Island is compared against non-Rhode Island states. To extract the distinguishing characteristics the feature importance of the model after the fit is quantified, depending on the model.

In evaluating the binary classifiers I use the receiver operating characteristics (ROC) and the area under the curve (AUC) to compare performances of the models.

The training and test sets are constructed such that they contain equal amount of Rhode island participants and non-Rhode Island participants.

A. Preprocessing

A total of 22 features were selected that represents participant's demographic information and health status. Most of them were selected from the 2017 calculated variables [2] of BRFSS data. The selected features are either ordinal or categorical data types. The categorical features are transformed into one-hot-vectors, and the ordinal types are preprocessed such that they are consistent continuous variables where the "not sure/missing/refused" responses are treated as separate features of binary data type. Missing values of the categorical data are replaced with the appropriate number category and missing values of ordinal types were replaced by the mean value of the responses.

The three other New England states (arbitrarily) selected for the purpose of this task are Vermont, New Hampshire, and Massachusetts.

B. Machine Learning Models.

In this task, four models are considered: Logistic Regression, k-Nearest Neighbours, LightGBM [3], and Neural Network. These models were chosen because of their different capacities to fit linear and non-linear patterns.

C. Validating Models

To validate the models, cross validation method with 70:30 train:test set split is used. ROC and AUC are used to compare performances across models.

III. RESULTS

The final training and test ROC of the logistic regression is shown in Figure 1, which has a AUC of 58%. The most distinguishing feature of Rhode Island can be read from the feature importance of the

model, shown in Figure 2 where odds of whether a participant associated with a set of features or responses is a Rhode Island resident is shown. Participants who identifies themselves as Hispanic (*_RACE_8*) and who have had their cholesterol checked in the past five years (*_CHOLCH1*) are more likely to be associated with Rhode Island than the other states. Consistently, LightGBM's feature importance, shown in Figure 3 also indicates the same two features as being the most deciding features that associates a participant to Rhode Island. Moreover, employment status (*EMPLOY_3*) is also shown to be an importance feature in both models.

Figure 4 shows the performance comparisons of the four models considered in this report. Although LightGBM is shown to be the superior performance[1], all models are consistently showing AUC of just under 60%.

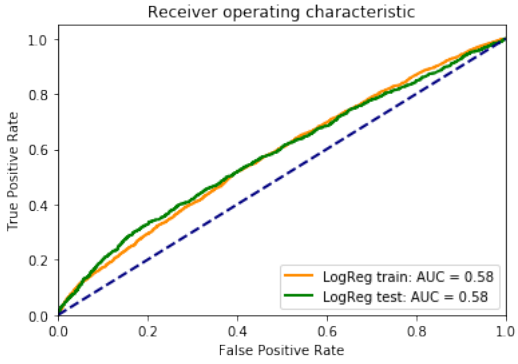

Fig. 1. ROC and AUC of logistic regression model training and test set.

	Coefficients	odds
_RACE_8	0.606230	1.833507
_CHOLCH1_1	0.248155	1.281659
EMPLOY1_3	0.225174	1.252541
_RACE_2	0.191504	1.211070
MARITAL_5	0.114917	1.121781

	Coefficients	odds
_CHOLCH1_2	-0.249219	0.779409
_PHYS14D_NA	-0.187931	0.828672
_RACE_1	-0.181395	0.834106
EMPLOY1_2	-0.115271	0.891125

Fig. 2. Fit parameters (coefficients) of the logistic regression model and odds of whether a participant is a resident of Rhode Island or not, given a set of responses.

[1]Taking into account statistical fluctuations, this superiority may not be significant compared to the second best model.

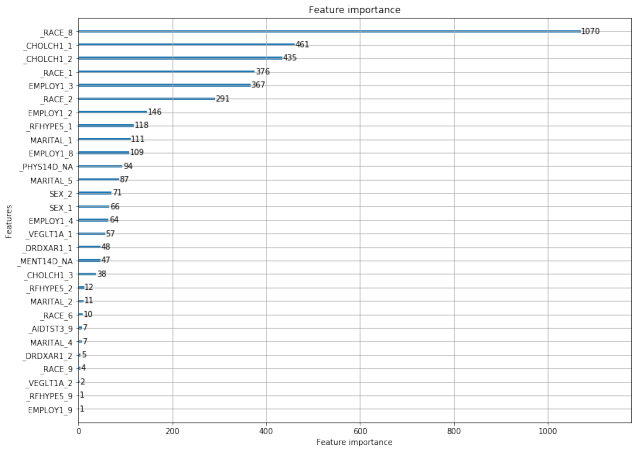

Fig. 3. Feature importance of the LightGBM model using the test set.

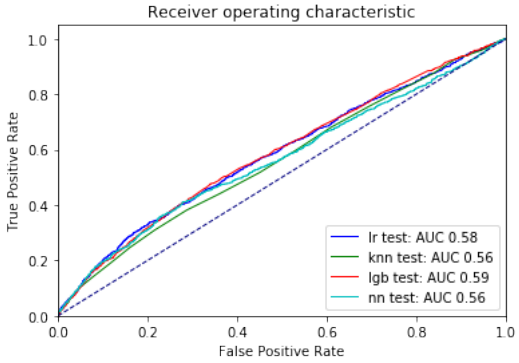

Fig. 4. ROC curves and AUC of four models test set: Logistic regression, kNN, LightGBM, Neural Network

IV. DISCUSSION

A brief analysis of a subset of columns of the 2017 BRFSS data is presented. Given the time constraint, some processes in the analysis are not the most optimal. For example, more time could be given to selecting the most interesting subset BRFSS features. Certainly, domain knowledge would be helpful in determining more interesting set of features.

Also, more experimentation could also be done on the way the training sample is constructed.

REFERENCES

[1] https://www.cdc.gov/brfss/annual_data/annual_2017.html
[2] https://www.cdc.gov/brfss/annual_data/2017/pdf/2017-calculated-variables-version4-508.pdf
[3] LightGBM: A Highly Efficient Gradient Boosting Decision Tree