Snizhana Kurylyuk

Abstract

An estimated 34.2 million people have diabetes (10.5 percent of the U.S. population). An estimated 7.3 million adults ages 18 years or older have diabetes but are undiagnosed (21.4 percent of adults with diabetes). I partnered with a healthcare organization to create a classification model to predict patient diabetes status. I used the Behavioral Risk Factor Surveillance System (BRFSS) diabetes data which contained over 250 thousand data points and I built six various classification algorithms. Then I tuned the hyperparameters and selected the best model to present to my client. Overall I was focusing on the best accuracy of my model.

Design

To create a classification model to be used for prediction requires identifying which features would be most useful. I performed some exploratory data analysis.

Then once preliminary models were built, functions were used to find the best scoring hyperparameters. Cross-validation of the model was then used to check the validity of the model.

Once a model was selected the model was run against validation data and once again scored. With an acceptable score, the model was then used on test data and ultimately used to predict whether or not a patient had diabetes.

Data

The dataset contains 253,680 survey responses with 22 features for each. The data set comes from Kaggle but was originally produced by the Behavioral Risk Factor Surveillance System (BRFSS) , a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. This specific data set is from 2015.

Features

- High BP (0 = no high BP, 1 = high BP)
- High Cholesterol (0 = no high Cholesterol, 1 = high Cholesterol)
- Cholesterol Check (0 = no cholesterol check in 5 years, 1 = yes cholesterol check in 5 years)

- Body Mass Index
- Smoker
- Stroke (0 = no, 1 = yes)
- Heart Disease or Attack (Coronary heart disease (CHD) or myocardial infarction (MI) 0 = no, 1 = yes)
- Physical Activity (physical activity in past 30 days - not including job 0 = no, 1 = yes)
- Fruits (Consume Fruit 1 or more times per day 0 = no, 1 = yes)
- Veggies (Consume Vegetables 1 or more times per day 0 = no, 1 = yes)
- Heavy Alcohol Consumption (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week, 0 = no)
- Any Healthcare (Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no, 1 = yes)
- No Doctor because of cost (Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no, 1 = yes)
- General Health (Would you say that in general your health is: scale 1-5 - 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor)
- Mental Health
- Physical Health
- Difficulty Walking (Do you have serious difficulty walking or climbing stairs? 0 = no, 1 = yes)
- Sex (0 = female, 1 = male)
- Age (13-level age category)
    - 1 = Age 18 - 24
    - 2 = Age 25 to 29
    - 3 = Age 30 to 34
    - 4 = Age 35 to 39
    - 5 = Age 40 to 44
    - 6 = Age 45 to 49
    - 7 = Age 50 to 54
    - 8 = Age 55 to 59
    - 9 = Age 60 to 64
    - 10 = Age 65 to 69
    - 11 = Age 70 to 74
    - 12 = Age 75 to 79
    - 13 = Age 80 or older
- Education
    - 1 = Never attended school or only kindergarten

    - 2 = Grades 1 - 8 (Elementary)
    - 3 = Grades 9 - 11 (Some high school)
    - 4 = Grade 12 or GED (High school graduate)
    - 5 = College 1 year to 3 years (Some college or technical school)
    - 6 = College 4 years or more (College graduate)
- Income
    - 1 = Less than $10,000
    - 2 = $10,000 to less than $15,000
    - 3 = $15,000 to less than $20,000
    - 4 = $20,000 to less than $25,000
    - 5 = $25,000 to less than $35,000
    - 6 = $35,000 to less than $50,000
    - 7 = $50,000 to less than $75,000
    - 8 = $75,000 or more

Algorithms

Feature Engineering

1. Mapping latitude and longitude to 3-dimensional coordinates so nearby continuous values would also be close in reality
2. Converting categorical features to binary dummy variables
3. Combining particular dummies and ranges of numeric features to highlight strong signals and illogical values for waterpoint status identified during EDA
4. Selecting subsets of the total unique values for categorical features that were converted to dummies, according to the number of samples they were associated with and their contribution to certain statuses

Models

Logistic regression, k-nearest neighbors, and random forest, decision tree, stacking resembling, and voting ensembling classifiers were used before settling on random forest as the model with strongest cross-validation performance. Random forest feature importance ranking was used directly to guide the choice and order of variables to be included as the model underwent refinement.

Snizhana Kurylyuk

Model Evaluation and Selection

The entire training dataset of 250,000 records was split into 80/20 train vs. holdout, and all scores reported below were calculated with 10-fold cross validation on the training portion only. Predictions on the 20% holdout were limited to the very end, so this split was only used and scores seen just once.

The official metric for DrivenData was classification rate (accuracy); however, class weights were included to improve performance against F1 score and provide a more useful real-world application where classification of the minority class (functional needs repair) would be essential.

Algorithms

Logistic Regression:

- Accuracy: 0.7896
- Precision: 0.3639
- Recall: 0.6476
- F1: 0.4659
- Cross Validation Accuracy Score : 0.74455

10 nearest neighbors validation metrics:

- Accuracy: 0.7673
- Precision: 0.3128
- Recall: 0.5360
- F1: 0.3951
- Cross Validation Accuracy Score : 0.77854

Random Forest validation metrics:

- Accuracy: 0.7698
- Precision: 0.3488
- Recall: 0.7195
- F1: 0.4698

- Cross Validation Accuracy Score: 0.80256

Decision Tree validation metrics:

- Accuracy: 0.7699
- Precision: 0.3330
- Recall: 0.6218
- F1: 0.4337
- Cross Validation Accuracy Score: 0.71748

Stacking Ensembling validation metrics:

- Accuracy: 0.8623
- Precision: 0.5388
- Recall: 0.1988
- F1: 0.2905
- Cross Validation Accuracy Score: 0.92498

Voting Ensembling validation metrics:

- Accuracy: 0.7974
- Precision: 0.3560
- Recall: 0.5296
- F1: 0.4258
- Cross Validation Accuracy Score: 0.8623

Final random forest 5-fold CV scores:

- Accuracy: 0.8295
- Precision: 0.3963
- Recall: 0.3869
- F1: 0.3915
- Cross Validation Accuracy Score: 0.91611

Snizhana Kurylyuk

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting

Communication

The presentation can be viewed here.