Project Final Report: Predict Diabetes with Demographics, Behaviors and Health Conditions

CPSC 5305 FQ 2019

Team members: Yu Fu, Hao Li, Nina Nguyen

Problem statement

Diabetes is a considerable burden for patients and health insurance payers in the United States. Diabetes can lead to high medical bills, impaired physical mobility, and increased risk of sudden death, ultimately causing a decline in the quality of life of patients. The factors that cause diabetes are the complex effects of genes and acquired environments. Demographics and daily human behaviors, including diet, sleep, and exercise, are important environmental factors for diabetes. Additionally, current health conditions other than diabetes, such as cancers, CVD, as well as medications and treatments, can also have correlations with diabetes.

The goal of this project is to build a predictive model for diabetes using demographics plus daily behaviors plus health conditions, then identify important variables in predicting diabetes. Finally, study the correlations between diabetes and important features. These important features could be valuable in further studies on causal relations.

Data sources

- The Behavioral Risk Factor Surveillance System (BRFSS)
    - The primary source for demographics, behavior and health condition variables
    - 2018 BRFSS data https://www.cdc.gov/brfss/annual_data/annual_2018.html

Data Preparation

- Data EDA and cleaning are finished. There are **437436 records and 176 variables (538 after one-hot encoding)** in total.
    - EDA Notebook: https://github.com/yf23/FromBehaviorToHealth/blob/master/EDA_All_Variables.ipynb
    - EDA Report (PDF): https://github.com/yf23/FromBehaviorToHealth/blob/master/EDA_Reports_All_Variables.pdf
- Mapping Don’t know / Refused / Unknown / Missing entries to NA values.
- For binary categorical variables, map values to 0 and 1.
- For multiple categorical variables, use one-hot encoding.
- For continuous variables, detect and crop outlier in [Q1 – 1.5IQR, Q3 + 1.5IQR] range.
- Unit conversion: There are data with different units within one feature. For example, 101-107 means 1-7 days per week, and 201-230 means 1-30 days per month. They are converted to the same unit.

- Some redundant and diabetes-dependent variables are dropped. Diabetes-dependent variables are those features that depend on diabetes status, such as the age of getting diabetes, the usage of insulin, etc. The model could easily classify diabetes status with these features. If the person does not have NA values for these features, this person must have diabetes.
- **Target variables (y)**: binary diabetes indicator and CVD indicator
- **Feature variables (X)** of the following categories
    - *Survey Related*: Year, month and day of the survey
    - *Demographics*: Variables including location, residence, cell phone usage, age, gender, ethnicity/race, education, marital status, employment, income.
    - *Health Status and Conditions*: Current health conditions and diseases. The target variables of diabetes and CVD are selected from this category.
    - *Cancer*: Cancer is a special module in the survey. It covers variables of conditions and treatments related to cancer.
    - *Oral Health*: Visit frequency and number of teeth removed.
    - *Behavior*: Multiple categories including *Smoking, Alcohol Consumption, Sleep, Exercise, Drive* and *Sun Exposure*
    - *Healthcare Access*: Mostly related to the frequency of doctor visits and health insurance
    - *Health Check, Test, and Treatments*: The name and frequency of tests or treatments that have been applied to a respondent person.
    - *Children*: Demographics and health conditions of the children of respondent persons.
- Missing Value Analysis
    - Number of variables with NA percentage 90% or higher: 52/176
    - Number of variables with NA percentage 80% or higher: 80/176
    - Number of variables with NA percentage 70% or higher: 85/176
    - Number of variables with NA percentage 60% or higher: 96/176
    - Number of variables with NA percentage 50% or higher: 112/176

Analysis

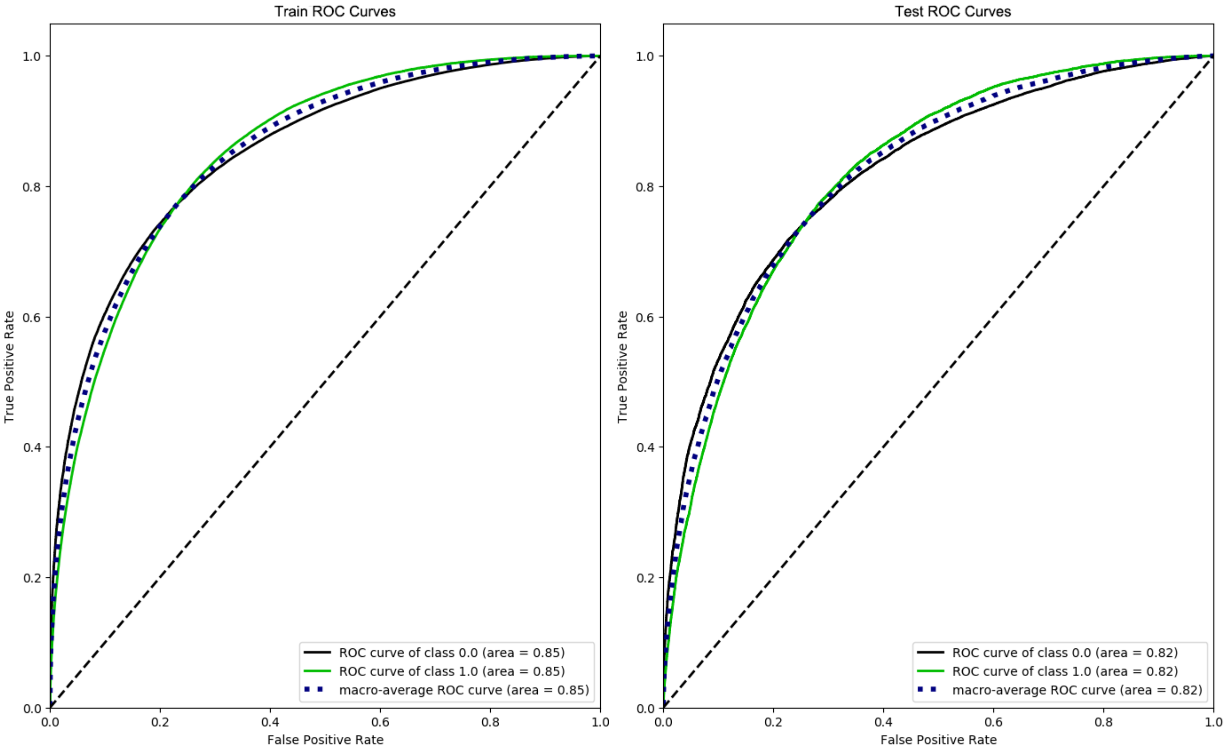
- **Goal**: Build a model to predict y using X. Tune hyperparameters of the model for better prediction. Interpret the prediction result and import features based on the model.
- **Train/Test split** in 8:2 ratio.
- **Metric of Choice**: AUC ROC score
    - ROC is the measurement of the sensitivity (the percentage of diabetic people who are correctly identified as diabetic) and specificity (the percentage of healthy people who are correctly identified as being healthy) the model.
    - ROC has 1-specificity (False Positive Rate) as the x-axis and sensitivity (True Positive Rate) as the y-axis. Each point on the ROC curve is the value corresponding to a different threshold of probability that the classifier predicts the record as positive.
    - AUC (Area Under Curve) of ROC measures the probability that a classifier will rank a randomly chosen positive record higher than a randomly chosen negative one. Higher AUC ROC scores generally mean a better model.
    - Why not accuracy? Accuracy is not a good metric here because the diabetes status indicator is skewed with a lower percentage of 1’s.

    - Why not F1 score? The most suitable threshold here might not be 0.5 as classifiers normally do, and the threshold could always be adjusted in real scenarios, which might require higher sensitivity or higher specificity. AUC ROC will consider and integrate over all the thresholds from 0 to 1 and score the model by generalizing performance on different thresholds. In our experiments, if the grid search uses the F1 score, it finds the most overfitting model that gives a very high training F1 score but a very low testing F1 score. Using the AUC ROC score gives us a better and not overfitting model.
- **Model of Choice**: XGBoost
    - XGBoost is an ensemble method of decision trees using a gradient boosting method. Trees are constructed in a sequence that learns to predict the residual errors of the previous tree.
    - Handle NA values by learning the direction of NA values at each split.
    - Hyperparameters to tune using grid search with 5-fold cross-validation (optimal value in bold red)
        - Number of trees in the ensemble: [150, **250**, 350, 450]
        - Maximum depth of trees: [5, **6**, 7, 8, 9]
        - Number of features (proportion of number of X columns) used in each tree: [**0.65**, 0.75, 0.85, 0.95]
        - Learning rate: [**0.1**, 0.25, 0.5]
    - Due to the size of the dataset, the grid search was limited both to the number of hyperparameters and the range of search. The grid search took 19 hours on a 64-core Google Cloud VM using a parallel search.
- **Feature Importance**:
    - SHAP (SHapley Additive exPlanations) Values
        - A consistent and accuracy feature importance score
        - Can be used in model interpretation (More on results)
        - SHAP value is the measure of impact to y variable. The positive or negative value indicates a positive or negative impact. For each feature, each record has a SHAP value. The mean(|SHAP|) across all records is the feature importance score.
    - Weight: The number of times a feature is used to split the data across all trees
    - Gain: The average training loss reduction gained when using a feature for splitting

Results / Summary / Conclusion

- **Model Evaluation**
    - Train AUC ROC: 0.85
    - Test AUC ROC: 0.82
    - The test and train are close, indicating no significant sign of overfitting.
- **Model Interpretation**
    - The ROC Curve gives the ability to choose the prediction threshold on different requirements of sensitivity and specificity.
    - The upper-right part of the curve is high sensitivity but very low specificity, meaning the model could predict most people as diabetic using a low threshold.

  - The lower-left part of the curve is low specificity but high specificity, meaning the model is using a high threshold to reject most people as being diabetic.
  - The balance range seems to be the middle of the curve, where sensitivity is around 0.8, and the corresponding specificity is around 0.7. That means the model could correctly identify 80% of the diabetic patients and 70% of the healthy patients.



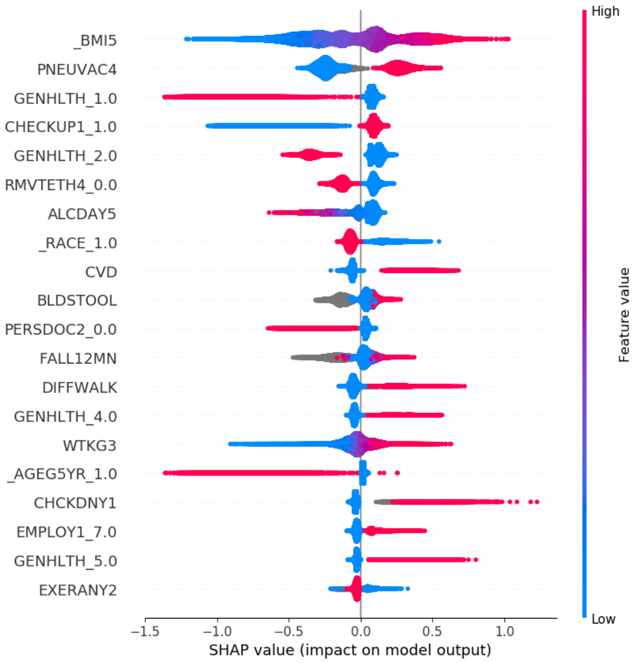
  - Overall, although not ideal, the model is useful in predicting diabetes with given features of demographics, behaviors, and health conditions. That meets our goal of building a predictive model using the features from the BRFSS dataset.
- **Feature Importance**
  - Out of 538 features (after one-hot encoding), there are 95 features with no feature importance scores. That means these features are not used to construct any tree in the ensemble model.
  - The top 20 features of each feature importance score take large amounts of total scores. That means the model relies heavily on these top features.
    - Mean(|SHAP|): 60.03%
    - Gain: 40.38%
    - Weight: 39.80%
  - The table below shows the top 20 import features ranked by mean absolute SHAP.

VARIABLE	MEAN(|SHAP|)	GAIN	WEIGHT
_BMI5	0.29008311	56.2712	955
PNEUVAC4	0.246306852	151.764	186
GENHLTH_1.0	0.215565026	187.4385	97
CHECKUP1_1.0	0.192156404	141.0423	134

GENHLTH_2.0	0.191416338	185.7621	92
RMVTETH4_0.0	0.11579673	267.0226	46
ALCDAY5	0.110556245	62.10222	192
_RACE_1.0	0.102761239	45.04941	117
CVD	0.093895532	122.2237	118
BLDSTOOL	0.088536993	124.0847	93
PERSDOC2_0.0	0.087854497	55.53914	76
FALL12MN	0.086158559	46.05168	243
DIFFWALK	0.08476235	520.1821	79
GENHLTH_4.0	0.074216522	101.6607	101
WTKG3	0.0736752	16.34008	593
_AGEG5YR_1.0	0.059232209	63.29371	55
CHCKDNY1	0.052875798	57.8411	140
EMPLOY1_7.0	0.049120281	46.31621	63
GENHLTH_5.0	0.040119227	56.38427	84
EXERANY2	0.038051486	17.9177	80

- Most of the top 20 features are health condition related variables (BMI, general health condition, vaccines, CVD, etc.). But there are also some demographic variables (Race, Age, Employment) and drink as behavior variables. Other behavior variables rank lower in importance. Sleeping time ranks at 34, and smoking ranks at 76.

- **Feature Interpretation**
  - One nice thing about SHAP values is that it is possible to see how the change of feature values influences its impact on the y variable using a distribution plot on SHAP values.



  - High BMI (_BMI5) has a strong positive impact on diabetes, while low BMI has a strong negative impact.

    - Pneumonia (lung infection) vaccination (PNEUVAC4) is recommended for diabetes patients. Our model catches the relationship.
    - People identifying as being good health (GENHLTH_1.0) has a strong negative impact on diabetes. However, identifying as being poor health who do not have a strong positive impact on diabetes. There is a similar pattern on people does not do routinely checkup with the doctor. (CEHCKUP1_1.0 = 0)
    - High alcohol consumption (ALCDAY5) seems to have a negative impact on diabetes. But which direction is the causal relationship? It could be heavy drinking cause not having diabetes, or not having diabetes cause heavy drinking, or there are other undiscovered common causes for both heavy drinking and diabetes. The model is not sufficient to confirm these hypotheses.
    - For demographics, being not white (_RACE_1.0) has a positive impact on having diabetes; being young (_AGEG5TR_1.0) has a strong negative impact on having diabetes; being retired (EMPLOY1_7.0) has a positive impact on having diabetes.
    - Other important health conditions that have positive impacts on having diabetes include CVD, difficulty at walking, fall in the past 12 months, having kidney checked, and generally felling poor health.
- Overall, we build a model that is able to predict diabetes using demographic, behavior, and health condition features from BRFSS data. We identify the most important features are health condition variables, following by demographic variables. Behavior variables are generally less important than the other two categories. We study the correlations between important features and diabetes, yet the causal relationship is not given by the model. Therefore, our goal for the project is reached.

Challenges

- Large data is much harder (and more expensive!) to process than expected. The EDA took a long time for 188 features, and the training and grid search took longer.
- Choose correct performance measuring metric matters. When we chose the F1 score as the metric, the grid search always searched for the very high F1 but overfitting parameters.
- Software engineer skills really help in data cleaning and training. A little additional coding time could make the experiments much easier.

Future Work

- Model fine-tuning with more hyperparameters and larger search space
- The connection between multiple years of BRFSS survey data
- Population segmentation studies by demographics, by behaviors, or by unsupervised clustering
- Some minority groups might not be well represented in the survey and worth studying. Some data augmentation methods could be tested. The minority groups are the groups with low prevalence rates in the combination of various demographic variables. For example, we might only have one record represents the group “native American + 70-79 years old + <15000 annual income + high school education”. As a result, we will have a very high bias in these groups.
- Experiments to test the causal hypothesis that can be derived from the current predictive model

Citation

This project is inspired by Behavior Predictor from PricewaterhouseCoopers DoubleJump Health.

https://www.pwc.com/us/en/industries/health-industries/library/doublejump/behavior-predictor.html

https://www.pwc.com/us/en/industries/assets/pwc-2019-us-hi-behavior-predictors-placemat-rev6.pdf

Thanks this article that introduces SHAP values.

https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27