Data Analysis on Depression

Aug 2019

Siyu Zhang

Motivation – Depression Prediction

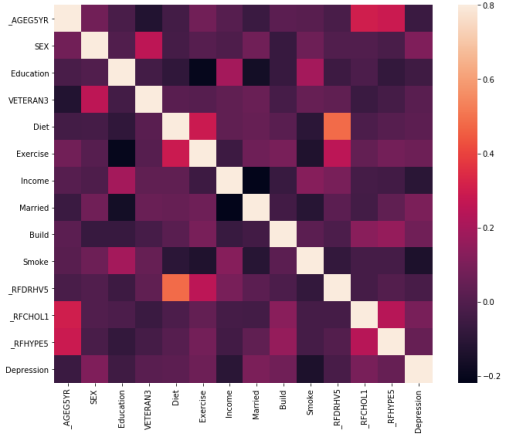
- The project is designed for a health management tool to help clients to prevent depression
- Potential Variables:
  - Age, gender, education, veteran, income, marital status
  - Diet, exercise, body build(BMI), smoking, drinking alcohol
  - Cholesterol level, blood pressure
- Target:
  - Diagnosed depression
- Fun Result:
  - A married female with an unhealthy diet, smoking, drinking with high cholesterol is more likely to have depression
  - Higher income with no smoking or alcohol would less likely to have depression

Data Collection - BRFSS

- Data Source: The Behavioral Risk Factor Surveillance System(BRFSS)
  - Ongoing health-related telephone surveys
  - Collect data on health-related risk behaviors, chronic health conditions, and use of preventive services from the noninstitutionalized adult population
  - Contains 358 variables for a total of 450,016 observations in 2017
  - It’s based on a large stratified random sample

Dataset Analysis

- Data Processing:
  - Fill NA
  - Data binning
- Check correlations



Symbol	Description	Data Binning
_AGEG5YR	Reported age in five-year age categories calculated variable	NA
SEX	Respondents sex	nA
EDUCA	Education level	Narrow to 4 categories, value 1 for not graduate High School or Unknown, value 2 for graduated High School, value 3 for attended college or technical school, value 4 for graduated from college or technical school.
VETERAN3	Are you a veteran	NA
FRUIT2	How many times did you eat fruit	Narrow to 2 categories, value 1 for having fruit or vegetables every week, value 2 for once more than a week or never or unknown
FRUITJU2	How many times did you drink 100 percent pure fruit juices	
FVGREEN1	How many times did you eat dark green vegetables	
VEGETAB2	How many times did you eat other vegetables	
EXEROFT1	How many times walking, running, jogging, or swimming	Narrow to 3 categories, value 1 for exercise every week, 2 for exercise every month, 3 for never or unknown
INCOME2	Income level	Narrow to 4 categories, value 1 for less than $50,000, value 2 for $50,000 - $75,000, value 3 for greater than $75,000, value 4 for unknown
MARITAL	Marital status	Narrow to 2 categories, value 1 for married, value 2 for all other situations
_BMI5	Body Mass Index	Narrow to 4 categories, value 1 for underweight, value 2 for normal weight, value 3 for overweight, value 4 for obese
SMOKE100	Smoked at least 100 cigarettes	Narrow to 4 categories, value 1 for Every day smoker, value 2 for Someday smoker, value 3 for Former smoker or unknown, value 4 for Non-smoker
SMOKDAY2	Frequency of days now smoking	
_RFDRHV5	Heavy alcohol consumption	NA
_RFCHOL1	High blood cholesterol	NA
_RFHYPE5	High blood pressure	NA
ADDEPEV2	Ever told you had a depressive disorder	Narrow to 2 categories, value 1 for yes, value 2 for all other situation

Modeling Result Interpreting

- Model Logistic Regression
- The confusion matrix
  - Overall accuracy: 0.80
  - The model has better result predict non-depression samples, the precision is 0.81 and the recall is 0.99
  - As for the depression samples, the precision is 0.56, but the recall is only 0.05.
- Interpreting
  - The cause of depression could be very complicated and not easy to define.
  - The data analyzed is self-reported health survey data, which might be biased due to the respondents' lack of awareness of their risk status.