Data Mining, Spring 2019 Projects List

Analytical Project Assignment
Project A
BRFSS: Factors predictive of high pneumonia vaccination rates

Associated data set

Download data from: https://www.cdc.gov/brfss/annual_data/annual_2017.html

Data description

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world. For more information go to: https://www.cdc.gov/brfss/about/index.htm

References:

Documentation about the data: https://www.cdc.gov/brfss/annual_data/annual_2017.html

Tool to check validity of trends you generated from data: https://www.cdc.gov/brfss/brfssprevalence/

Your tasks

1. Download the 2017 data, summarize and review the data to identify and visually present potentially interesting aspects of it. ~4-5 visuals / Tableau dashboards

2. How much information is available around pneumonia vaccination? Number of respondents, prevalence by region, distribution by gender & age

3. CDC believes that apart from age and region there are other factors related to/predictive of pneumonia vaccination rates. You're asked to analyze the data and answer the following questions:

   - Is the data fit to explore and identify such factors?

   - If yes, what factors are predictive of high pneumonia vaccination rates?

   - CDC wishes to launch a marketing campaign to increase the vaccination rates. Please make recommendations to them based on your analysis from above.

Analytical Project Assignment

Project B

BRFSS: Sleep quality across the country

Associated data set

Download data from: https://www.cdc.gov/brfss/annual_data/annual_2017.html

Data description

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world. For more information go to: https://www.cdc.gov/brfss/about/index.htm

References:

Documentation about the data: https://www.cdc.gov/brfss/annual_data/annual_2017.html

Tool to check validity of trends you generated from data: https://www.cdc.gov/brfss/brfssprevalence/

Your tasks

1. Download the 2017 data, summarize and review the data to identify and visually present potentially interesting aspects of it. ~4-5 visuals / Tableau dashboards

2. Analyze various questions on sleep quality and sleep hours present in the survey and present their response rates across demographics and region.

    - Are the various questions on sleep correlated? Which question(s) would be the ideal indicator to represent sleep quality of a respondent?

3. Based on the sleep quality indicator you've identified, you're asked to analyze the data and answer the following questions:

    - Can we predict if the respondent is likely to get "enough sleep" based on other information from the survey?

    - If yes, what factors are predictive of low sleep quality?

Analytical Project Assignment

Project C

Human Activity Recognition (credits: Prof Artur Dubrawski @ Auton Lab, CMU)

Associated data sets

Dataset-har-PUC-Rio-ugulino.csv

Data description

Human activity recognition is a growing field with numerous applications. Companies like BodyMedia and Fitbit sell personal activity monitors worn on the left arm and waist, respectively. One can imagine numerous applications for this technology. For the purposes of this project, suppose your company would like to construct a similar device. Toward that end, you have been provided a dataset containing approximately 8 hours of accelerometer data for each of 4 individuals. Records represent 3-axis acceleration measurements taken from 4 accelerometers worn on the waist, left thigh, right arm, and right ankle. Each measurement is taken over a time window of 150ms and presented in temporal order, without a time stamp. The activity of each participant is categorized into 5 classes; sitting, sitting-down, standing, standing-up, and walking. For more information you can visit http://groupware.les.inf.puc-rio.br/har#ixzz2aUaBROdz.

Your tasks

1. Determine whether or not the activity of an individual can be correctly detected from the given accelerometer data.

2. Derive features that exploit the temporal nature of these data. Determine the informativeness of these features and compare with an instantaneous approach (i.e. using only the acceleration measures for the current instant of time).

3. If you are able to reliably detect the current activity, investigate how long it takes to detect a change from one type of activity to another. Try to calculate new features that improve your change-point detection.

4. If only a single accelerometer can be worn, determine which location waist, left thigh, right arm, or right ankle results in the best activity classifier in terms of both classification accuracy and change point detection.

5. Summarize your findings and provide recommendations to your company.

Analytical Project Assignment

Project D

Understanding patterns of road safety

Associated data sets

DfTRoadSafety_Accidents_2012.csv
Road-Accident-Safety-Data-Guide.xls

Data description

The associated data is a subset of the public collection of data of the circumstances of personal injury road accidents in Great Britain in 2012. The statistics relate only to personal injury accidents on public roads that are reported to the police, and subsequently recorded, using the STATS19 accident reporting form. Information on damage-only accidents, with no human casualties or accidents on private roads or car parks is not included in this data. Please note that some of the integer values are codes rather than actual values, please refer to the supplied code book (Road-Accident-Safety-Data-Guide.xls) to understand the actual meanings. This file may contain references to additional datasets that can be ignored.

Very few, if any, fatal accidents do not become known to the police although it is known that a considerable proportion of non-fatal injury accidents are not reported to the police. Figures for deaths refer to persons killed immediately or who died within 30 days of the accident. This is the usual international definition, adopted by the Vienna Convention in 1968.

As well as giving details of date, time and location, the accident file gives a summary of all reported vehicles and pedestrians involved in road accidents and the total number of casualties, by severity.

Key tasks

1. Identify a group of features that you think will be most useful. You may use a combination of intuition and variable selection techniques for this task. If you wish, you may derive new variables which you think could be useful in your analysis and discard variables which you think may not be. Search for interesting patterns in data by summarizing and visually inspecting it.

2. Based on the available data, what are the factors that best discriminate between different severities of accidents? Can the knowledge of these factors be helpful in the practice of reducing frequency of higher-severity accidents?

Can the given data be used to infer areas for improvement within different police jurisdictions? If so, provide the reasoning for your analysis with examples.

Analytical Project Assignment
Project E
<Bring Your Own Project>

Associated data sets: It should be a publicly available data set

- <provide link to data set you'll be using>

Data description

Provide a brief description of what is present in the data set and a high-level data dictionary (key fields that are interesting to you

Your tasks

Come up with a list of questions you will be exploring – should include an exploratory component and a predictive component.