1. Problem Description, Overview and Interpretations

Predicting Heart Disease Mortality with Machine Learning Techniques
Lorenzo Negri, July 2018

Problem Description

The goal is to predict the rate of heart disease (per 100,000 individuals) across the United States at the county-level from other socio-economic indicators.

The target variable *heart_disease_mortality_per_100k* for each row of the test data set must be in a positive integer with no decimals points.

The job is to:

- Train a model using the inputs from two different files with several missing values from: train_values.csv and the labels train_labels.csv
- Predict integers for each row in test_values.csv for which, there is no information about heart disease mortality rate.
- Output the predictions in a format as requested by the client to be scored.

The predicting variable is a numeric quantity, so this is a regression problem. To measure regression performance, the client will use a metric called root-mean-squared error. It is an error metric, so lower value is better (as opposed to an accuracy metric, where a higher value is better).

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(\widehat{y_n} - y_n)^2}{N}}$$

Where $\hat{y}$ is the predicted value and y is the actual value. The best possible score is zero, but the worst possible score can be infinite.

Overview

The target variable: *heart_disease_mortality_per_100k* (a positive integer available in: *train_labels.csv*) will be joined later to the dataset in order to train a model and predict the target variable. For the analysis in this document, the label (target variable) was joined has the first step, but this will not be the first step of the train machine-learning model workflow because the label can create overfitting results when managing and predicting the missing data of the train dataset.

Exploring data by summary statistics, and by creating visualizations, made possible to identify several potential relationships between features and heart disease mortality, leading to create (after analysing, cleaning and filtered the data) a machine learning regression model to predict an integer number of heart disease mortality per 100k giving some related variables.

Interpretations of Analysis and Experiments with the dataset

The biggest factor that increase imprecise predictions, are the managing of missing data for some key features like:

- **health__homicides_per_100k** – represents the number of deaths by homicide per 100,000 population (National Center for Health Statistics). With more than half of the data not available (1967 missing values over 3198 total), has +0.48 rate of positive linear correlation measure with the label (*heart_disease_mortality_per_100k)*. This creates imprecise predictions and cleaning of other missing data.
- **health__pct_excessive_drinking** – represents the percent of adult population that engages in excessive consumption of alcohol (Behavioural Risk Factor Surveillance System). With 978 missing values.
- **health__pct_adult_smoking** – percent of adults who smoke (Behavioural Risk Factor Surveillance System). With 464 missing values.

A good problem solving of the missing data in this project is one key factor to obtain better predicting regression model results.

Many features found to be important when regenerating and managing key feature missing data, some are:

- **area__rucc** – Rural-Urban Continuum Codes "form a classification scheme that distinguishes metropolitan counties by the population size of their metro area, and nonmetropolitan counties by degree of urbanization and adjacency to a metro area. Can affect significantly managing missing data for *health__homicides_per_100k* and *health__pct_excessive_drinking*.
- **area__urban_influence** – Urban Influence Codes "form a classification scheme that distinguishes metropolitan counties by population size of their metro area, and nonmetropolitan counties by size of the largest city or town and proximity to metro and micropolitan areas." Resulting helpful for managing missing data when predicting *health__pct_adult_smoking*.

Converting these categorical features including *econ__economic_typology* to indicator values (binary features for each unique values) can help have better results when managing prediction of the missing values.

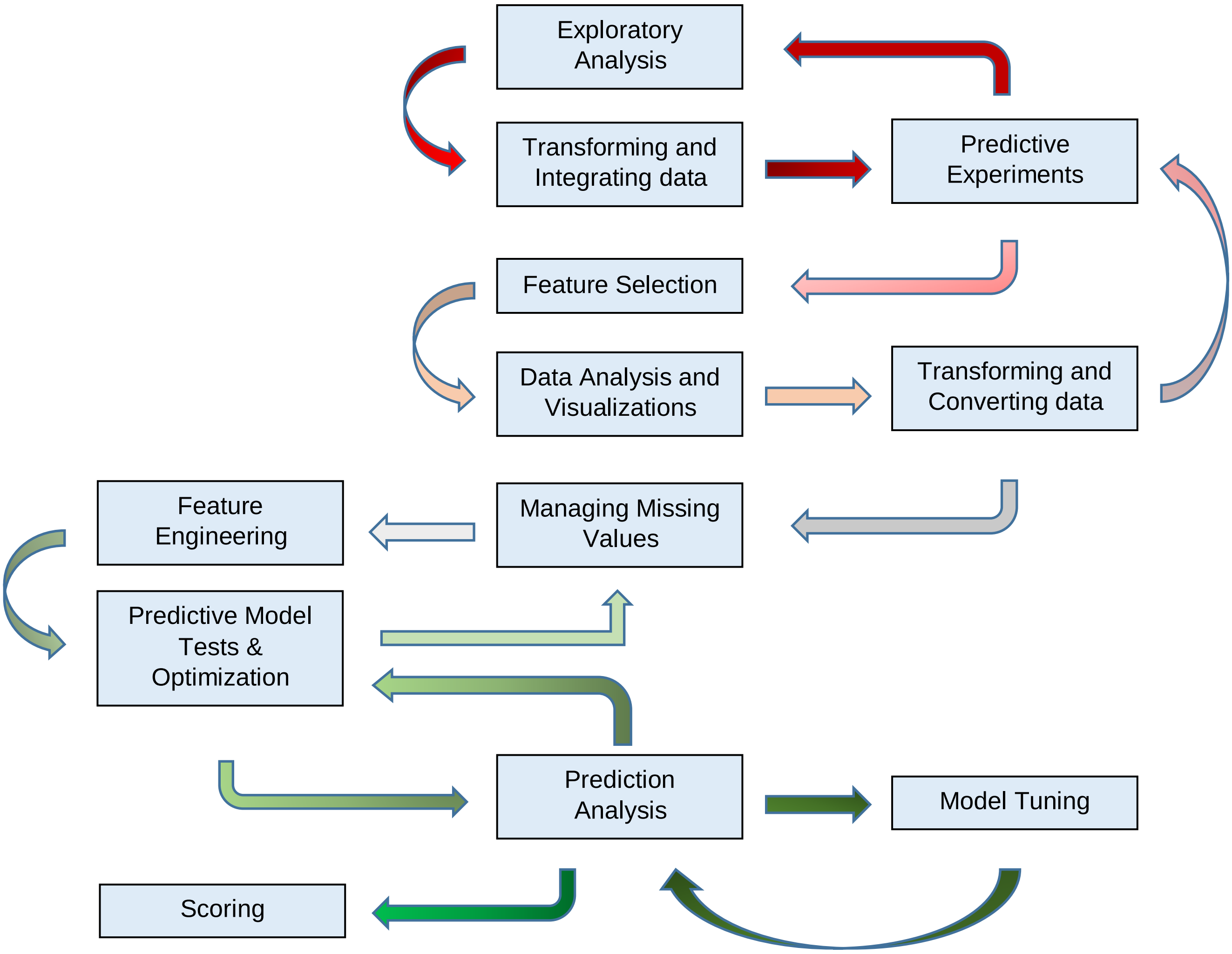
Many factors can help machine-learning algorithm to increase precise predictions of the number of deaths per 100k from heart disease in this problem, some significant features found in the analysis and experiments were:

- **health__pct_physical_inacticity** – the percent of adult population that is physically inactive (National Center for Chronic Disease Prevention and Health Promotion). The number of deaths by heart disease increases when the percent of adult inactivity is greater for a particular county.

The feature has the highest coefficient of determination and the highest score when evaluating feature importance of the best-evaluated model performance.

- **demo__pct_adults_bachelors_or_higher** – the percent of adult population which has a bachelor's degree or higher as highest level of education achieved (US Census, American Community Survey). Counties with higher percentage of bachelor’s degree tend to have a lower mean number of deaths by heart disease than others, which in turn tend to chain in similar way to the other education demographics features.
- **health__pct_adult_smoking** – the percent of adults who smoke (Behavioral Risk Factor Surveillance System). Counties with a higher percent tend to increase heart disease death rate.
- **health__pct_diabetes** – percent of population with diabetes (National Center for Chronic Disease Prevention and Health Promotion, Division of Diabetes Translation). There appears to be a positive correlation between deaths by heart disease and diabetes percent, in which high diabetes population counties, tend to have greater number of deaths by heart disease.

Scheme approach of workflow applied for solving the problem


Exploratory Analysis
Transforming and Integrating data
Predictive Experiments
Feature Selection
Data Analysis and Visualizations
Transforming and Converting data
Feature Engineering
Managing Missing Values
Predictive Model Tests & Optimization
Prediction Analysis
Model Tuning
Scoring