Capstone Project for Springboard Data Science Intensive Course
Alona Varshal
July 2016

Project Idea 1:

Wine Quality Data Set (UCI Machine Learning Repository)
(http://archive.ics.uci.edu/ml/datasets/Wine+Quality)

The wine quality data set available at UCI Machine Learning Repository provides information on eleven physicochemical parameters as well as a sensory parameter for Portuguese "Vinho Verde" wines. Grapes are indigenous to the Vinho Verde area in Portugal. Physicochemical parameters include fixed acidity, citric acid content, residual sugar, total sulfur dioxide, among others. The sensory parameter was determined from evaluations by a minimum of three sensory assessors who graded the wine according to whether the wine was bad or good. The following questions can be explored using the dataset:

- Can the wines' physicochemical characteristics predict the human sensory panel?
- Are there any differences in the parameters that can better predict red wines from the white wines? What differentiates the red wines from the white wine based on the physicochemical properties?
- Are all 11 parameters measured necessary or are there only a few of them needed to predict the human sensory panel? Is there a single parameter that can predict wine quality or is there a set?
- Which physicochemical parameter best indicates poor wine quality? Best wine quality?
- Are the 11 parameters sufficient to predict the human sensory panel?
- Is there a better machine learning technique that can better show patterns in the data?

Project Idea 2:

Twin Gas Sensor Arrays Data Set (UCI Machine Learning Repository)
(http://archive.ics.uci.edu/ml/datasets/Twin+gas+sensor+arrays)

A gas sensor made up of an array of eight metal oxide (MOX) sensors was tested whether it can be calibrated to measure concentrations of various gases of environmental significance (ethanol, methane, ethylene and carbon monoxide). This project attempts to see whether the findings of the paper published using the data are reproducible. It was found that the sensor can be calibrated and that one unit's result can be mapped to the other unit.

Project Idea 3:

Austin (Texas, USA) Animal Center Intakes and Outcomes (separate data)

Austin is known to be an animal-friendly city. Does the open data on the city's animal center intakes and outcomes reflect this? There are numerous aspects that can be examined by looking at the data. Does the rise in intake of animals correlate with any town events? What breed, sex, age among dogs is the most popular to be lost? Which area in Austin has the most lost animal? Which animal gets adopted most? Which breed of dogs get adopted most? What is the average time spent by animals in the center before adoption? Does the state of the animal affect probability of being adopted?