Initial Capstone Project Ideas - V2

Updated: November 27th, 2017

1. **Purpose:** In efforts to understand trends in pet adoption outcomes, the Austin Animal Center has provided data relating to the pets in their adoption center. The data includes details of the pets such as animal type, age, gender, breed, color, etc., as well as the outcome of the pets at the end of their stay at the Austin Animal Center. I am proposing to explore the pet adoption center data to predict the likelihood of adoption for pets, as well as possibly use this information to propose actions that could improve the performance of the center.

   Dataset: https://www.kaggle.com/c/shelter-animal-outcomes/data

   Questions to be explored:
   (1) What are the most important factors that influence whether or not a pet finds a home in this area? (explore factors related to all outcome types, including: Adoption, Return to owner, Euthanasia, Death, Transfer)
   (2) How accurately can a predictive model identify pets that are likely to have difficulties being adopted? (i.e. probability of adoption above/below a specific threshold value)
   (3) If so, can the insights we gain from successful adoptions be used to generate suggestions for actions that could potentially improve the chances of adoption for some pets? (e.g. early transfer for pets that typically don't get exposure at the center we are investigating, or better selection of pets to prioritize for fostering, etc.)

2. **Purpose:** Ronny Kohavi and Barry Becker extracted various records from the 1994 census bureau database that combine traits of individuals - such as age, education, marital status, occupation, and gender - with information that indicated whether or not their annual income exceeded $50k USD. Using the 1994 census data, I will explore the most important factors relating to whether or not a person makes over $50k per year and build a classification model to generate predictions on a person's income relative to this threshold. Although this data was taken over 20 years ago and may not be accurately representative of the financial climate of today, I believe it is still a valuable exercise to extract the trends from this time period.

   Dataset: https://www.kaggle.com/uciml/adult-census-income/data

   Questions to be explored:
   (1) What were the most important factors in 1994 that determined the income group of those surveyed?

(2) Using basic feature engineering techniques, are there additional features that I can produce from this information that will be more valuable indicators of a person's income group?

3. **Purpose:** I believe that using machine learning techniques to aid in medical diagnosis can be one of the most powerful and impactful applications of data science techniques. In a dataset published by the University of Wisconsin relating to breast tumor measurements, this dataset contains "Features...computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image." I would like to study this dataset and explore the possibility of predicting the probability that a tumor is either benign or malignant given the measurement data.

   Dataset: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/data

   Questions to be explored:
   (1) How accurately can one classify a tumor as either benign or malignant based solely on physical measurement characteristics such as size, density, concavity and symmetry of the tumor sample cell nuclei?
   (2) How does this approach to diagnosis compare with modern malignant tumor detection methods? Or alternatively, is there additional measurement information which is likely to improve the model performance?
   *(2nd question will be answered by combining model results with background literature research on the FNA biopsy technique using additional resources, such as this article: https://emedicine.medscape.com/article/1819862-overview)