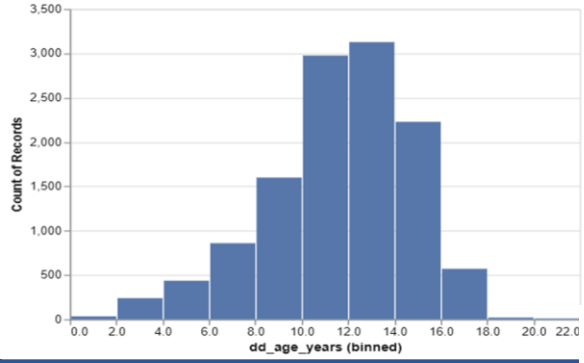
A Dog's Life

An analysis of the factors influencing dog health collected by the Dog Aging Project

A. Van Poznak, K. Bourne, A. Sachdeva

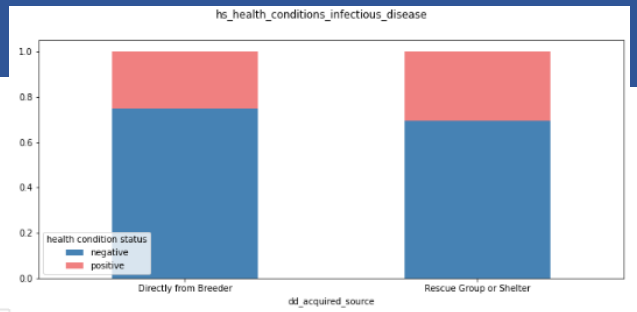
Introduction and Hypotheses

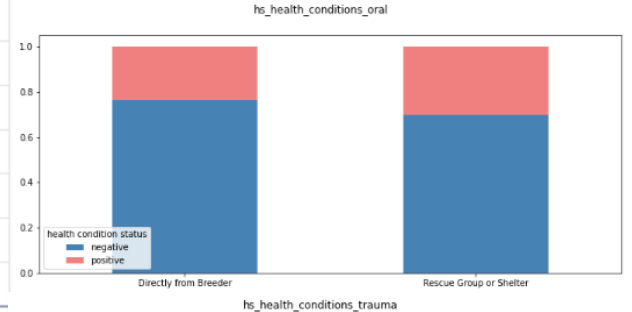
We leveraged data from the Dog Aging Project to investigate the following:
1 – Does wealth or dog sourcing influence health?
2 – Can we use advanced machine learning techniques to understand the progression of cancer in dogs?
3 – Does spaying/neutering influence dog social behaviors?

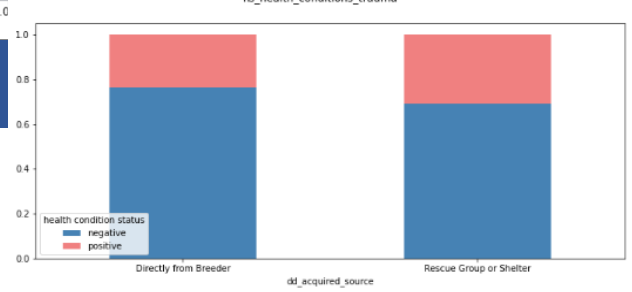


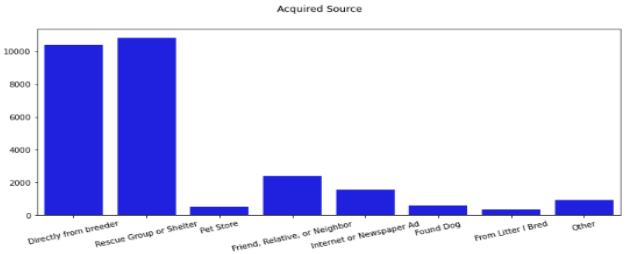
Methods

Dog healthcare data was obtained from dogagingproject.org, representing a broad set of data on 27,541 companion dogs. It contains 600+ data points on each dog. The analyses were performed in Jupyter Notebook using Scikit Learn and Tensorflow as the primary machine learning datatools. Models utilized included logistic regression, ridge regression, stochiastic Gradient Descent (SGD), Support Vector Classifier (SVC), K-Nearest Neighbors, Decision Tree, Random Forest, Gaussian Naive-Bayes, Bernoulli Naive-Bayes, Gradient Boost, XGBoost, Multi-Layer Perceptron (MLP) Neural Network, GANs, and Hidden Markov Model, and an Attentive State Space Model (ASSM).



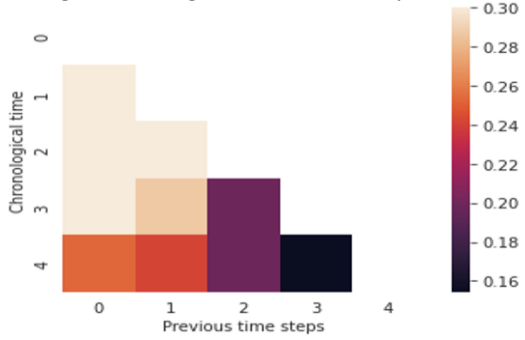


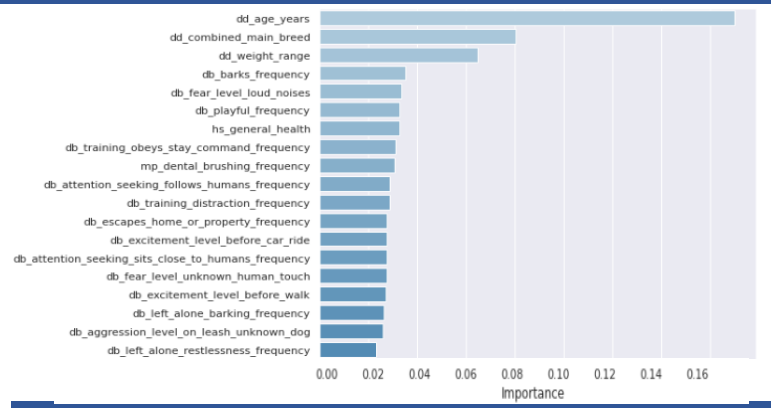




Results

- No relationship was found between owner wealth or median income on the health of dogs in the study. While certain health conditions were found to be present for adopted dogs more frequently than dogs from a breeder, a Random Forest classifier built to predict the presence health conditions was inaccurate.
- The cancer trajectory analysis component of this project resulted in the generation of a model with significant predictive power up to 3 time steps (doctor visits) in advance in predicting an occurrence of some type of cancer. In addition, an analysis with Random Forest indicated that age, breed, and weight range stand apart as the most predictive factors in determining if a dog has cancer.
- After concluding the neutering/spaying analysis component of this project, we were left with a firm affirmation that the relationship between spaying and neutering and the prevalence of certain undesirable behaviors in male and female dogs warrants further study. Although the models we constructed weren't perfect predictors, the performance scores most produced, despite the underlying data quantity problem, were firmly average and too high to discount entirely.





Conclusion

- For the analysis of wealth and sourcing on dog health, no demonstrable relationship was identified. Although a potential relationship was identified between dog source and reported infectious disease, oral health conditions, and trauma, an attempt at building a Random Forest model to predict the presence of these conditions was largely inaccurate and likely suffered from a small data set size.
- For the cancer trajectory and prediction section, less complex (more traditional) machine learning models had a difficult time handling the highly complex nature of healthcare data and producing useful and consistent disease predictions. However, the Random Forest model did offer up insights into what features in the dataset had the most predictive power. More advanced models, like ASSM, seem capable of predicting dog disease trajectories effectively, but the data available in this dataset fell short in providing the granularity needed to approach these diseases from a broader staging perspective (i.e. stage I cancer vs stage IV cancer), and is limited to simply predicting the presence or lack of presence of cancer.
- While model performance for the neutering/spaying component was somewhat fair, with less complex classifiers like Gaussian Naïve Bayes and Logistic Regression fairing slightly better, I believe that future researchers who are able to obtain a comprehensive dataset will be able to produce far more robust and productionizable results. All in all, lack of size was undoubtedly one of the biggest impediments to model performance for this analysis.