David DeBow
Justin Johnson
Ryan Parker
Pacific Analytics

Final Project CST 383:

The Data Science of Dog Names

Introduction

All the members of Pacific analytics enjoy the company of animals. Pacific Analytics undertook this data science project to see if we could confidently identify the gender of a dog based on its name. We noticed there are more androgynous names for animals. In the sense that a name like snow could be female or male. While a human name like John tends to skew male. With the help of machine learning, could we determine with great accuracy, above 80 percent, the gender of a dog based on its name? If so, what additional characteristics would take place?

Selection of Data

Although originating from Data.gov, it is not a federal dataset but a NY city dataset with over 500,000 rows.

The 30-megabyte file is made up of the following eight rows: AnimalName, AnimalGender, AnimalBirthYear, BreedName, ZipCode, LicenseIssueDate, LicenseExpirationDate, and extract year. Animal Name is the name of the pet. The ten most popular names are Bella, Max, Charlie, Coco, Luna, Lola, Rocky, Lucy, Daisy, and Teddy. This compares with the ten most popular baby names James, John, Robert, Michael, William, Mary, David, Joseph, Richard, and Charles. The

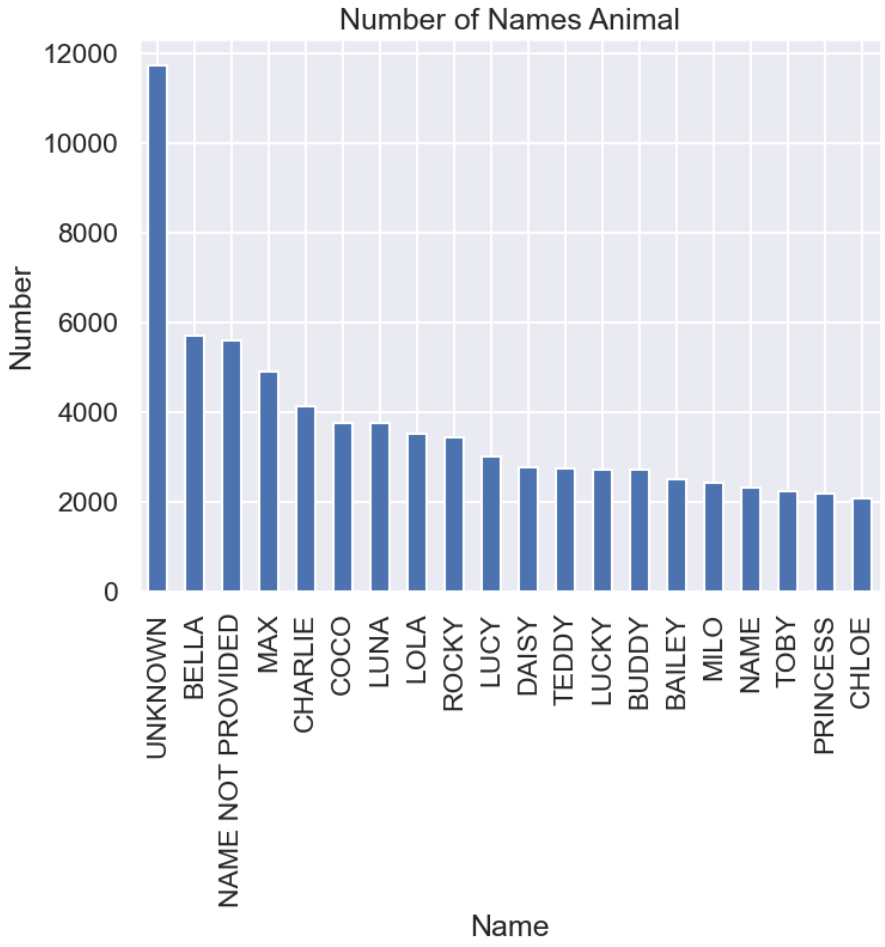
top baby names come from the Bible and are male, while the top dog names are female and whimsical. There is some crossover between Charlie and Charles. Charlie is a more informal variation of the name Charles.

Animal Gender is declared as female or male. It became transcribed to 1's and 0's during the data process. Gender means the biological sex of the dog based on its anatomy.

The zip code was the five-digit New York zip code. There are over 900 zip codes. Zip code data is echoed due to the difficulty of machine learning utilizing Zip codes.

The data is self-reported, which did cause issues in data organization. Extract year is the year the data was extracted.

Any munging of feature engineering?



We removed null values. We dropped values without a name. The above image shows the preponderance of incorrectly named dogs. Two out of the top three dog names are vaguely entered, and they are not useful for the prediction.

We removed infinite values. This was done with the command df = df[np.isfinite(df).all(1)]. The infinite value was narrowed down to Zip codes. We tried manually excluding zip codes below 0 and above 5 digits, but it wasn't adequate. We tried to keep the data true to the source. We made use of an Encoder to encode all the data into numbers. We had issues with get_dummies. The website is https://www.projectpro.io/recipes/convert-string-categorical-variables-into-numerical-variables-using-label-encoder . It is a 10 line function.

Methods

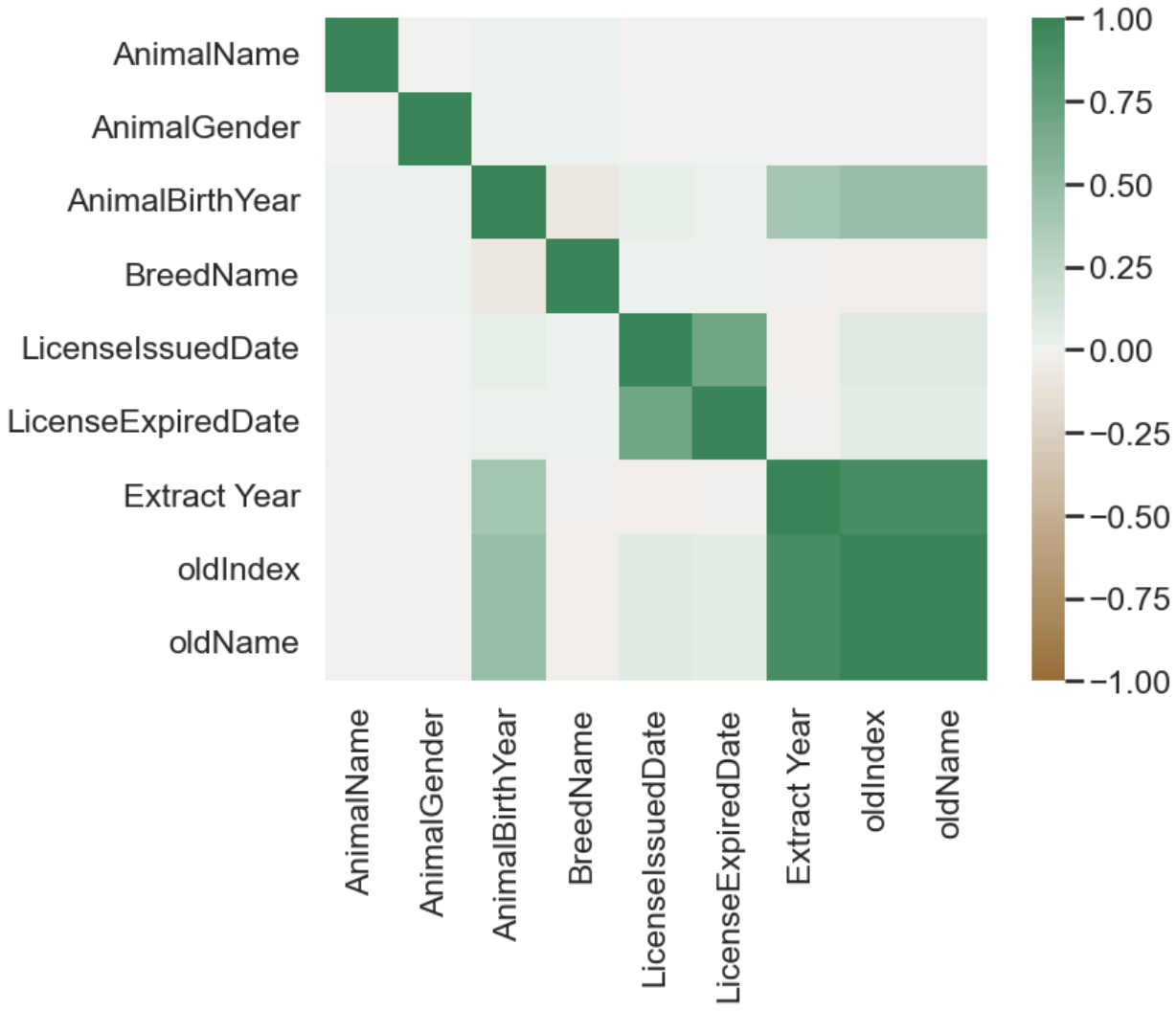
We chose to use Jupyter Notebook for the coding. We instituted a round-robin system where one person would code and then pass it to other teammates on google chat. We made use of Anaconda for the python data science library. This greatly eased the importing of modules. It was more of an issue of writing an import statement than downloading or installing it. So far, graphviz has been one of the few packages that have required an installation. In that case, it was a couple of command line commands and a few minutes of waiting.

We incorporated Encoder to encode everything into variables. We made use of KNNClassification. Initially, we tried linear regression, but we ended up with a horizontal line. There was no possible fitted slope for the data.

The source of the dataset is Data.gov. It is a large file that is approximately 30 megabytes. This is over the standard file size for files uploaded by a web browser on GitHub (25 megabytes). The file was reposted on dropbox. Once rehosted, the file was imported for download into Jupyter Notebook.

Results

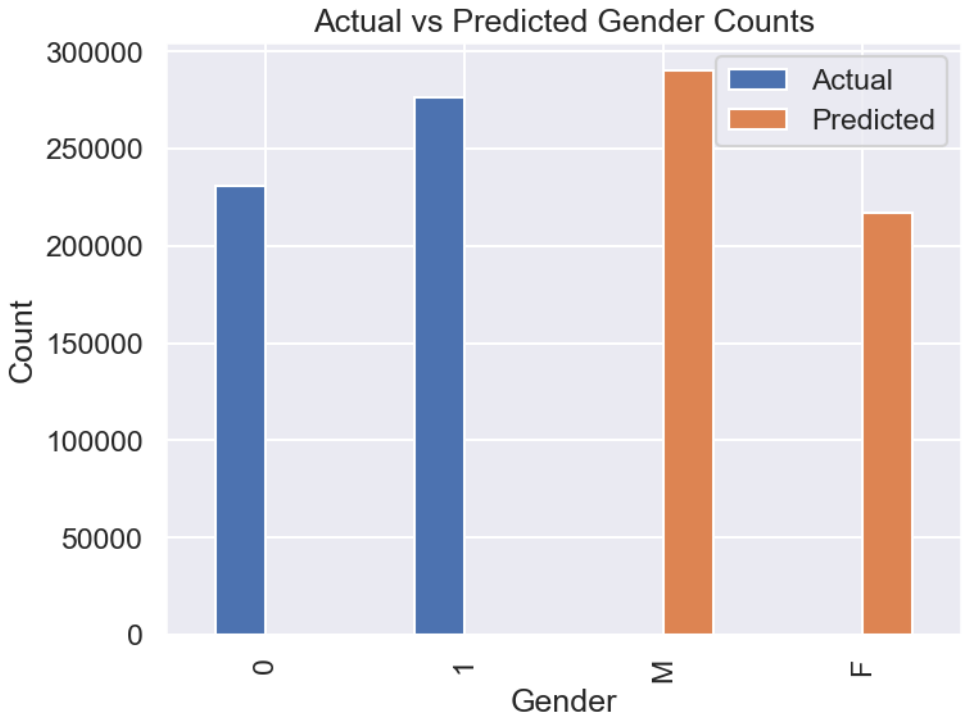
The question we began with was, With the help of machine learning, could we determine with great accuracy, above 80 percent, the gender of a dog based on its name? The answer to this question was very positive. Yes, we were able to reach 80 percent. In fact, we got to approximately 86 percent accuracy. It could go higher with additional data munging. We expelled major data discrepancies such as null and placeholder names, but we should have verified all the names.



Above, a heat map to show the relationship between various values.

The results are not of random luck. Various data points were tested. The nearest neighbors' k values were optimized to reach the best result. More than half the predictor markers produced poor results, results that were below 80. Some results were around the chance of luck or, more specifically, a coin flip.

Initially, the results were even worse. Linear regression produced nonsensical results. This led to the use of KNN Classification. There don't seem to be coefficients and intercepts associated with naming a dog.



We have a visualization for our conclusion. The "Actual vs. Predicted Gender Counts" graph demonstrates that Males were overpredicted and females were underpredicted. I believe that this is due to gender-neutral names leaning towards boys. It is more socially acceptable for a girl to have a masculine name than for a guy to have a feminine name.

There is an 86 percent accuracy rate. For a 500 thousand data set, that is an error rate of 70 thousand, for an average difference of 35 thousand in each prediction.

Discussion

The answer implies that there is no wrong way to name your dog. Other research has found that names tend to be more common for dogs or humans (citation). While dogs and human names can be interchangeable, cough, Charlie. You are more likely to meet a Chewbacca-named dog than a person.

Future research entails seeing the specific differences in the accuracy of name predictions for dogs, finding the ratio of dog names to dog populations, and comparing it to human names to human populations.

Why do the predictions matter for the gender of the dog? For one, it is helpful for animal control to determine the number of spays and neuters needed. Not all dogs will be fixed, but utilizing a

prediction on top of a prediction could help. A second reason is that it can help determine the ratio of male and female dogs and can hint at zipcodes underreporting dogs.

There is a superficial curiosity that is satisfied by knowing the gender of an animal. Is it an outdated construct in today's society? In a gender-equitable society, it shouldn't necessarily matter for a pet. It is possible an 86 percent accuracy record with gender names is representative of a society that is liberal. Maybe in a more conservative part of the United States, the KNN classification test would provide more accuracy.

Summary

People can name their dogs however they want. There is nothing stopping someone from naming a female dog a traditional name. Therefore it is expected to reach a different level of accuracy. There are a few names that go either way. With machine learning, you have an 85 percent chance of getting the sex right. Oddly enough, the name is almost the ideal situation. Adding other variables and factors can tank the ratings. This is due to a smaller sample size.