***Choosing the dataset:***

Initially when I am choosing the dataset for the second project, I wanted to pick something that is both doable and make sense biologically. The reason for this is heavily based off my first project which did not end as well as I wish it could. This is because the dataset for my first project (dataset 1) was a survey dataset on the various aspects of health of older Americans. The questions asked were all over the place and there were no comparative questions between groups such as men or women, sick or healthy or the old and the really old. In addition, the questions in dataset 1 were also all over the place, it covers a very broad spectrum of health-related questions without going into depth on anyone of them.

The problem dataset 1 presented then is that it was very hard to perform both statistical analysis on it: how can you perform any kind of statistical analysis when (almost) each question are different, and the results don’t necessarily all align (not everyone from every state answered every question)? In addition, how can you perform machine learning algorithms that predicts the state when given a health static when for each question, each state had 1 answer?

So for my second dataset, the biggest challenge was not to find a biologically related dataset – but rather to find a biology dataset that made sense from both a biological and computational perspective: I can explain the meaning of the data to an interviewer during a presentation and I can actually write statistical and machine learning algorithms base off the dataset.

***The dataset:***

The dataset I picked was called “Birds' Bones and Living Habits, Measurements of bones and ecological groups of birds” which, as its name implies, contains the measurements of birds’ bones and their respective bird groups. Details as follows: There are many kinds of birds: pigeons, ducks, ostriches, penguins. Some are good at flying; others can't fly but run fast. Some swim under water, others wading in shallow pool, etc. According to their living environments and living habits, birds are classified into different ecological groups. There are 8 ecological groups of birds: Swimming Birds, Wading Birds, Terrestrial Birds, Raptors, Scansorial Birds, Singing Birds (Songbirds), which each bird corresponding to their own symbol: SW: Swimming Birds, W: Wading Birds, T: Terrestrial Birds, R: Raptors, P: Scansorial Birds, SO: Singing Birds. In addition, the data on their bones are such: Length and Diameter of Humerus, Length and Diameter of Ulna, Length and Diameter of Femur, Length and Diameter of Tibiotarsus, Length and Diameter of Tarsometatarsus.

Birds belong to different ecological groups have different appearances: flying birds have larger wingspan than songbirds and wading birds have long legs than raptors or swimming birds. This makes sense given that evolution has molded these birds to exhibit the bone structures best suited to survive in their respective ecological environments: it only make sense for a flying bird have long arm bones so to fly better, whereas a songbird, given its smaller statue, have small bones so it can actually fly. As such, these birds’ bones’ shapes are shaped by their living habitats – and we may be able to categorize the different birds to their respective groups based off on the length of their bones using a clustering algorithm – assuming the bones of the birds are significantly different from each other.

Right away this dataset fit what I was aiming for when looking for my second dataset: 1) it is a biological relevant dataset in that, instead of being a jumble of unrelated questions that is both impossible to analyze statically but also machine learningly, it is a clear cut dataset with a very clear focus. 2) I can actually run machine learning algorithms on the dataset, when I go through the data, I can actually envision what kind of machine learning algorithm I want to or can run. In the previous dataset it was next to impossible, I have tried to write out algorithm without much success.

***The data structure of the dataset:***

The dataset consists of 12 columns: id, huml, humw, ulnal, ulnaw, feml, femw, tibl, tibw, tarl, tarw and type. Id is numbers going from 0 – 419 and type is the bird type disclosed in the above paragraphs. The above shortnames corresponds to the names in the following order: length of humerus, diameter of humerus, length of ulna, diameter of ulna, length of femur, diameter of femur, length of tibiotarsus, diameter of tibiotarsus, length of tarsometatarsus. There are blank values spread out throughout the dataset, 15 in all which I replaced with the average value for that particular bone under that particular bird (so I took the average length value of bone X under bird Y, and replaced the empty value with that value).

In addition, the number of observations for each bird type are not equal, with the numbers being: SW, W, T, R, P, SO: 116, 65, 23, 50, 38, 128. This means that for: Swimming birds have 116 measurements, Wading birds have 65 measurements, Terrestrial birds have 23 measurements, Raptors have 50 measurements, Scansorial birds have 38 measurements and Singing birds (songbirds) have 128 measurements. This unequal distribution of the dataset will present some problems later for me, especially when I am trying to develop a machine learning algorithm to predict the bird type when given a series of bone lengths – I will explain this when I get to my machine learning algorithm. But overall, the dataset is workable, albeit with some caveats that I must keep in mind.

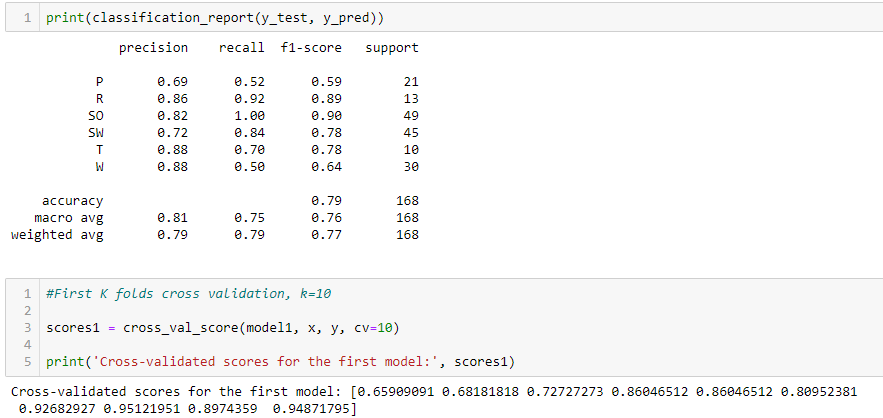
***Data analysis:***

After the data has been cleaned, I did a correlation analysis between the bones to each other. The reason for this is that I have 10 bone types, and to run any statistical or machine learning using all 10 would be a lot – and the large number of parameters could negatively impact my algorithm. So I decided to narrow down to only the bones that are the most correlated with each other. Here I define most correlated as bones with a correlation value of 0.87 or above. The reason is such: I want only the strongest correlation values, and base off my correlation analysis where the range is approximately: 0.606 -0.976. In this case I want to capture enough of bone types, but I also don’t want to include weak correlations (defined as lower than 0.8 by me, from my previous experience working on biological data). So, I picked a value where enough of the bone types can captured, but at the same time the overall correlations remain strong.

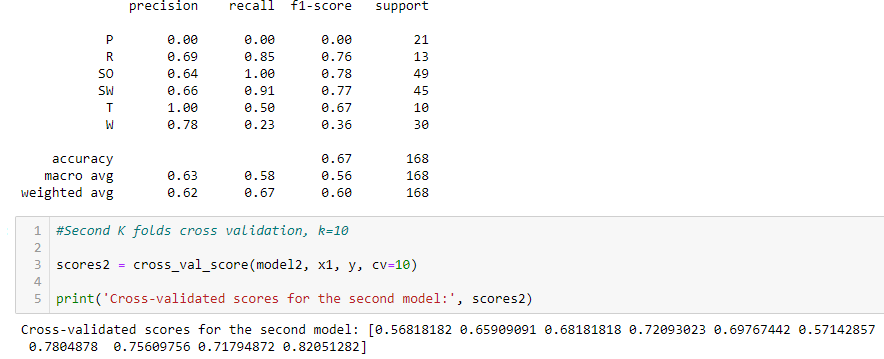
First, I composed the average and standard deviation for all of the bones for all of the birds and I constructed a boxplot from data for all of the bones for all of the birds and what I noticed is that there is a consistent trend: there are 5 long bones and 5 short bones – this is consistent across all birds, which is as expected. The length of the bones differ from bird to bird, again, to be expected: larger birds should have larger bones than smaller bones. Overall, the data looks clean, consistent, and completely workable for use to develop a machine learning algorithm.

***Machine Learning Algorithm:***

The machine learning algorithm to develop here is a clustering algorithm where given a set of bone length, it will return what bird it is. My first algorithm is a logistic regression that includes all of the bone types. Overall, it worked decently: weighted average for precision and recall are both 0.79. When I ran a 10-fold cross validation, however, the scores are all over the place: from 0.659 to 0.948. The reason is as such: the unequal distribution of the various birds is causing differences in each of my cross-validations so that each time the algorithm cross-validates, it has a different sample size, and that difference in sample size will affect the scores my cross validations produce:

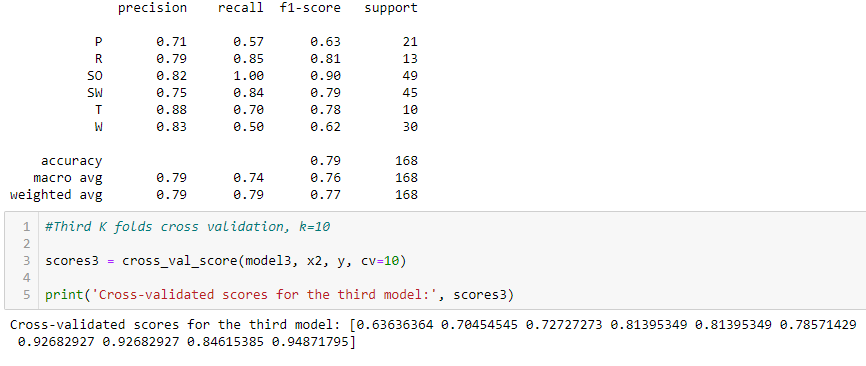


For my second logistic regression algorithm, I removed the bones with a correlation value less than 0.87 and kept only the following: humw, ulnal, ulnaw, feml, femw, tibw, tarw (refer to the key above to know what bones they are). The weighted average for precision and recall tanked to 0.62 and 0.67 respectively. The 10-fold cross validation had a range of 0.568 – 0.82. This told me that I removed too much bone types, some that had significance were removed so my algorithm actually lost power because of them. For my following algorithm I will re-evaluate the bone types and add back the ones I think will provide me with the best mode.



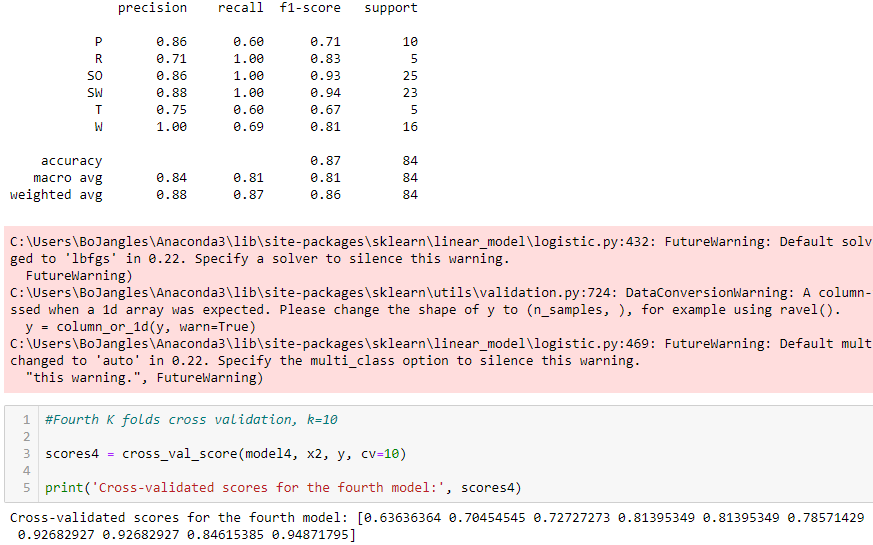
After examining the correlation analysis again, I finalized on the following bone types huml, humw, ulnal, ulnaw, feml, femw, tibl, tibw, tarw. I only took out tarl as it is the bone type that is the least correlated with all other bone types.

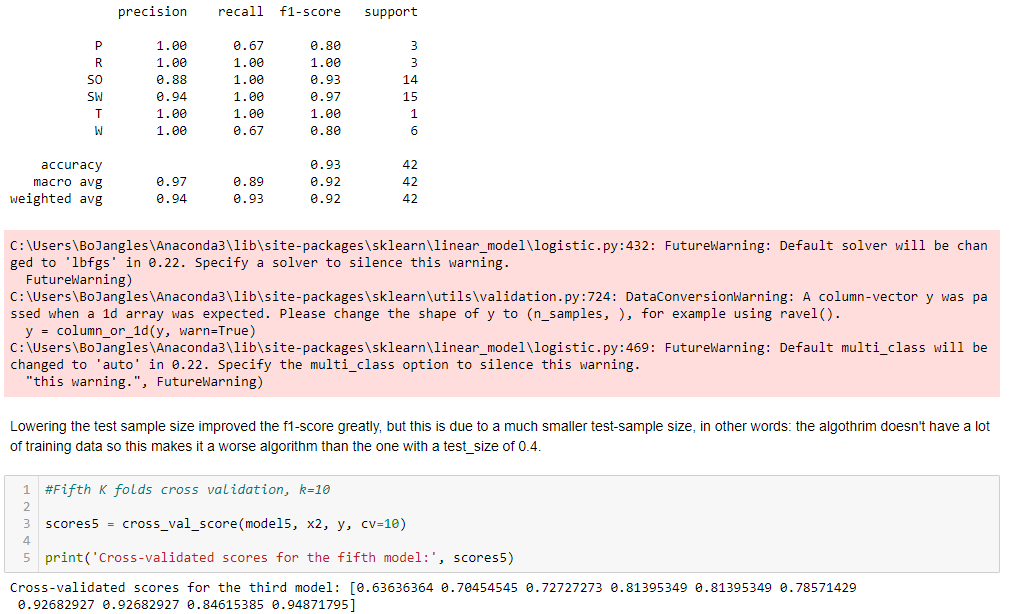
For my third logistic regression algorithm, I used the set described above, the weighted average for precision and recall returned to 0.79. The 10-fold cross validation had a range of 0.636 – 0.948. So basically I am back at where I started – this makes sense since I only removed 1 bone type, my results shouldn’t be too different from when I included all bone types (see next page).



For my next several algorithms I stuck to the one mentioned previously, except I tweaked the test\_size parameter to see whether or not I can improve my algorithm.

**Test\_size = 0.2**



**Test\_size = 0.1**

My conclusions from tweaking the test\_size parameter is that, don’t. The reason being is that while the scores for my precision and recall went up pretty dramatically (0.88 and 0.87 for the test\_size = 0.2 and 0.94 and 0.93 for test\_size = 0.1) – this is due to the algorithm overfitting my data because the overall sample size shrunk. Given that my dataset doesn’t have millions of data points, a small test\_size value proves to be disastrous because the algorithm will overfit. Looking at the 10-fold cross validation for each of the algorithms, the range remained high.

At this point, I decided to pick another, more powerful algorithm: the random forest algorithm. The reason being is that random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction. In this sense it is much more powerful than a regular logistic regression – instead of 1 model, there are many, and the most common prediction becomes the model’s prediction. Two heads (in this case, many heads) are better than one. In fact, generally only academics use logistic regression; most industry professionals use random forest instead.

Given my small sample size, I decided to set my test\_size at 40 but change around the n\_estimators parameter to see whether having more models will improve the predictions. With my first random forest having a n\_estimators of 50 (50 decision trees) and including all bone types, I got a weighted average precision and recall of 0.8 and a 10-fold cross-validation range scattered between 0.432-0.930. Once again, the uneven bird type sampling is harming my machine learning algorithm, and there is no real way to absolve this problem without collecting more data (I considered about deduce the amount of bird type data to the bird type with the least amount of data, but this would present its own problems as I am essentially throwing data away and having an even smaller dataset will definitely make my machine learning algorithm prone to overfitting. I consider the uneven bird type data the lesser evil here).

For my second random forest I used a n\_estimators of 100 and included all bone types, my weighted average precision and recall improved to 0.82 and a 10-fold cross-validation range of 0.477-0.897 with the bulk of the results in the 0.7 range. This is an improvement over my first forest, which is to be expected as I doubled the decision trees. In this vein, I ran a third random forest algorithm with all bone types except tarl, with a n\_estimators of 100. My weighted average precision and recall was 0.80 and a 10-fold cross-validation range of 0.454-0.872 with the bulk of the results in the 0.7 range. Overall, this isn’t a significant improvement over my second random forest algorithm, and for my final random forest algorithm I picked a n\_estimator of 500 using all bone types except tarl. My weighted average precision and recall was 0.78 and 0.79 and a 10-fold cross-validation range of 0.477-0.872 with the bulk of the results in the 0.7 range. Not much improvement and the 10-fold cross-validation remains scattered due to the uneven bird type sampling.

***Conclusion:***

My conclusion from this machine learning project is that I will choose a random forest model with n\_estimator of 100 and including all bone types (so the second random forest algorithm I tried). As all other alternatives did not yet significantly better results but these alternatives involve dropping data. This random forest algorithm I picked, on average, is able to correctly classify 80% of the results. But on any given prediction there could be a wide range of predictions, so this is definitely the user needs to keep in mind moving forward with this algorithm. See figure below:

