Ben Rapp, Jared Huzar, Craig Fox
1/26/2022
Projects in Data Science

Best Practices

Technical

1. **Check Distributions**
    a. Considering that we have numeric data describing certain features of our audio clips, it will be worthwhile to examine the distributions of our numeric data and see if they inform the audio clips themselves. It will also be important to extract other features from the audio data that aren't given to us in the Kaggle dataset. We should use histograms, CDF's and other statistical plots to get a visual understanding of the data, and we can also use descriptive statistics to understand the distributions as well (mean, media, quantiles, etc.). These things will allow us to detect outliers in the data that might not be immediately obvious from listening to the recordings or viewing the waveforms / spectrograms.
2. **Consider Noise**
    a. There will quite literally be random noise in our audio data, and it will be important to take this into account when we are examining the data. We can use descriptive statistics such as Signal-to-Noise Ratio (SNR) to quantify the random noise in the dataset, and we will likely revisit this statistic when preprocessing the dataset to prepare it for ML algorithms. It's also important to note that the measured noise in our dataset should be contributing to our displayed level of confidence in various statistics about the data. If we are going to create metrics that describe the audio data for given subsets of the population, we will need to make sure that there is an associated confidence with that measurement.
3. **Look at Examples**
    a. In order to understand the audio data, we are going to have to find a way to meaningfully analyze each audio file. We will have to show the data visually using waveforms and spectrograms. In order to make sure that this process is working correctly, it will be important to test and debug the code we use that transforms these files into more interpretable formats. It's also going to be important to consider examples from a number of different groups. We will want to check out audio clips from each of the different classes, and for larger classes of species, we will want to look at values from different parts of various feature distributions (e.g. for species X, examine the top and bottom 5% of examples with a given feature value)
4. **Slice your Data**
    a. There are a few different immediately obvious slices in our dataset (species, song type, frequency cutoffs, etc.) that will be worth comparing, especially if the

size of each subset is similar. It will not be safe to compare slices that are very dissimilar in size, such as the largest and smallest group of species in the dataset. However, looking at metrics across similarly-sized groups will be worthwhile to do. We might want to analyze certain features across species to learn what is similar and different about them, and also to expose problems with the dataset that might be solved by augmenting the data or preprocessing it in a certain way further on in the data science process.

Process

1. **Check For What Shouldn't Change**
    a. When determining what variables are most important in our machine learning model, it will be important to identify which features are highly divergent across time, across the population, and/or across the species. This can be achieved by splicing the data based on various metrics, i.e, species, song type, frequency, etc.). Based on what's consistent and what is not, we can appropriately weight variables and determine which are most important for distinguishing species' in the model. Also, examining the relationship between features will help us identify confounding variables and edit our model sufficiently to address the redundancy in the data.
2. **Make Hypotheses and Look for Evidence**
    a. Even before performing exploratory data analysis, we will make hypotheses on what variables we think are important, what distribution the data may follow, etc. For example, if we believe certain frequencies are associated with a certain species, we can look for correlation between the variables. In general after hypothesis building, we will then look for evidence and determine the accuracy of our hypotheses using statistical measures and reproducible tests. In this process of exploration, we may come across other important trends which will contribute to our understanding.
3. **Check For Reproducibility**
    a. To determine the robustness of certain conclusions we derive from the data, i.e, distribution of data, importance of certain variables, etc, we will test these conclusions across various time scales and subsamples of the dataset as a whole. Properly slicing the data will best enable testing of reproducibility. Also, we are given a training dataset of true positives and one of false positives we can check for reproducibility and consistent results across both in our model. Robustness to minor perturbations is essential in a dataset like ours, rainforest noises, which is subject to extensive random noise.
4. **Exploratory Analysis from End to End Iteration**
    a. We will employ a more vertical splicing model in which we perform complete analysis of the data from processing to testing. Based on our results we can then tweak different steps in the pipeline, and even experiment with different combinations of steps. This will allow us to properly and more frequently

incorporate the feedback we get when we test our model, resulting in more opportunities for improvement.

Mindset

1. **Data analysis starts with questions, not data or a technique**
    a. We will ask questions whenever you do not understand a decision or an analysis. We will all review each other's parts and ask questions about methodology and results to try to find holes in it. Before we write any code or view any data, we will ask each other questions about what we hope to achieve from the next step. We will try to ask good questions to ensure we gain useful insights. We will not criticize each other for asking basic questions and instead use it as an opportunity to verify that we are all on the same page and are clear about what we are doing. There are no stupid questions!
2. **Be both skeptic and champion**
    a. We will not rely on one piece of data to draw a conclusion. For any insight we have, we will show other pieces of evidence that support it or a lack of evidence opposing it. We will try to prove ourselves wrong in the process until we create a solid case supporting our insight. We will show someone else the data without telling them our analysis and see if they draw the same conclusions and have the same reasoning.
3. **Correlation != Causation**
    a. We will have the whole team try to provide a field/business explanation for why two factors are related for any correlations we find. If possible, we will try to create new tests or use different data to validate our hypothesis. We will consult outside sources to see if others also have a justification for the data. We will also look for potential ways we could receive a false positive. We will use statistical knowledge to see the chance the correlation is just a statistical anomaly.
4. **Expect and accept ignorance and mistakes**
    a. We will discuss what gaps in knowledge we have and how certain we are about any assumptions or conclusions we have. We will try to find ways to fill in the gaps or increase our certainty if possible. We will also take into account the uncertainty of the result whenever making a decision using that information. We will acknowledge that everyone in the group will make mistakes. We will accept these mistakes and try to provide constructive feedback to each group member so they can learn from it. We will not criticize each other or blame each other for any mistakes that are made and only use them as a learning opportunity.

Inspiring Quote We Believe In

“Therefore, your initial focus should not be on perfection but on getting something reasonable all the way through. Leave notes for yourself and acknowledge things like filtering steps and unparseable or unusual requests, but don't waste time trying to get rid of them all at the beginning of exploratory analysis.”

Practices We Were Confused About

- **Practical Significance**
    - Will practical significance be important for our purpose on this project? It seems like this would be a useful thing to consider in an A/B testing scenario or for a business use case, but should we be considering the practical significance of differences between subsets of our data?
- **Standard vs. Custom Statistics**
    - For audio data what kind of standard metrics would we validate? This seems more applicable to a situation in which we would be judging the effect of the introduction of a new aspect. Given the task we have is this an essential concern?