[Bird Song Identification]

[Gold Team]

Data Science Capstone Project
Exploratory Data Analytics Report

Date:

[05/23/2024]

Team Members:

Name: Jonathan Watkins

Name: Joseph Trybala

Name: Max Song

[The purpose of this report is to describe the exploratory data analytics. It includes five major sections:

1. Analyzing the basic metrics of variables: data types, size, descriptive statistics
2. Non-graphical and graphical univariate analysis: identifying unique value and counts, histogram, box plots, etc.
3. Missing value analysis and outlier analysis
4. Feature engineering and analysis: correlation analysis, dimensionality reduction, deriving new variables
5. Appendix]

Analysis of the basic metrics of variables

[In this section, we identify all the variables in the dataset and conduct the basic metrics of the variables. What are the data types (numerical/categorical, discrete or continuous, ordinal or nominal) and size? Provide the descriptive statistics of the variables such as mean, standard deviation, min, max, percentiles, etc.]

-

Numerical Features present in the metadata are alphabetically ordered:

- **Bitrate_of_mp3**: Numerical, Continuous // The bitrate at which the MP3 file is encoded, provided in kbps (kilobits per second)
    - Mode (more important than mean):
    - Min:
    - Max:
- **Duration:** Numerical, Continuous // The recording length as measured in seconds (per minute). Within the metadata file, the number of seconds is always reported as a whole number and not rounded up at all, e.g. a recording with a duration of 1.9s is still counted as having a length of 1 in the .csv file. Some recordings (40 in total) values are zeroes but a manual search on xeno-canto points confirms these are typically between 0.5s and .9s in length.
    - Mean: 54
    - Min: (arbitrarily as the files are only whole numbers) > 0 < 1
    - Max: 3552
- **Elevation**: Numerical, Continuous // The elevation at which the recording was made. The data points in the elevation column contained several anomalies in terms of types of entry which needed to be cleaned for proper use. Refer to the notebook for visuals and more specifics. Outliers need to be taken under the microscope here as the min. e.g. does not appear correct as the recoding was done in Texas (not -3670 m. below sea level) and the max 7760(m) in Utah, which is not the case.
    - Mean: 604.13 (rounded)
    - Min: -3670.0
    - Max: 7760.0
- **Latitude**: Numerical, Continuous // Combined with longitude, this geolocates our point of recording.

- **Longitude**: Numerical, Continuous // Combined with latitude, this geolocates our point of recording.
- **Ratings:** The ratings in this dataset are for the recordings. Ratings per recording are an averaged value from a community rating process.
    - Average: 3.26
    - Min: 1
    - Max: 5
- **Sampling_rate**: Numerical, Continuous // The number of samples of audio recorded per second measured in Hertz (Hz). It determines the frequency range and the quality of the audio recording, where higher sampling rates translate to higher resolution fidelity and capturing more details of the sound (by capturing more data).
    - Mode (more important than mean): 44100 (Hz)
    - Min: 8000 (Hz)
    - Max: 48000 (Hz)

Categorical Features present in the metadata are alphabetically ordered:

- **Author**: Categorical, Nominal // The name of the author of the audio clip, the recordist in Xeno-Canto terminology
- **Background**: Categorical, Nominal // A list of species found in addition to the target species in the background of the audio clip.
- **Bird_seen**: Categorical, Nominal // A true or false value, truth indicates that the recordist made visual identification of the individual making the call
- **Channels:** Categorical, Nominal // Either 1(mono) or 2 (stereo) used to record the birds for playback purposes. No more in-depth exploration done.
- **Country**: Categorical, Nominal // The country where the recording is located
- **File_type**: Categorical, Nominal // A singular value represents the whole dataset, .mp3.
- **Filename**: Categorical, Nominal // A combination of the XC_ID and the file_type. Inconsequential for us.
- **License**: Categorical, Nominal // Type of license with which the data was provided
- **Location**: Categorical, Nominal // The local town, park, or proper place at which the data is recorded. Varying degrees of geographical information given.
- **Playback_used:** Categorical, Nominal // If yes, recordist/author used a synthetic “pre-recorded” bird call of the same species to elicit calls/song/chirps etc. from the bird being pursued.
- **Primary_label**: Categorical, Nominal // The primary species represented in the recording
- recordist: Categorical, Nominal // Same value as Author
- **Sci_name**: Categorical, Nominal // Scientific name of the primary species of the recording
- **Secondary_labels**: Categorical, Nominal // Same value as Background
- **Species**: Categorical, Nominal // Same value as Background
- **Time**: Date/Time (or Categorical, Nominal if descriptive), both are found in our dataset. The vast majority are found in Date/Time. While not as easy, we feel as though date/time provides too much to convert to categorical. We are setting all data to Date/Time numerical by finding average times per am/pm, and then setting all AM or PM values to the average.

- **Title**: Categorical, Nominal // Same value as XC_ID
- **Type**: Categorical, Nominal // What type of sound was recorded, descriptive, could contain labelling of sex
- **Url**: Categorical, Nominal // Uses XC_ID to direct to the URL origin of the audio data
- **Xc_id**: Categorical, Nominal // The unique ID that Xeno-Canto provides for the audio clip

Irrelevant:

- **Ebird_code:** Categorical, Nominal // A shorted form of species name, serving as a form of unique ID. Irrelevant to our purposes.
- **Recordist:** same as Author. Duplicate data.

Non-graphical and graphical univariate analysis

[In this section, we identify the list and number of unique values for each variables and provide the histogram and box plots to understand the distribution of the data.]

Univariate analysis can be found in full in attached .ipynb.

Missing value analysis and outlier analysis

[In this section, we identify the missing values and outliers and determine how we handle these values before analysis.]

Reference .ipynb for full discussion and code.

MISSING VALUES:

Included are the feature names, approach taken, and rows lost if any.

- latitude/longitude (removed): 429 rows lost, 1.8% of our dataset

- background (imputed): none lost, 53.7% of our dataset

- type (removed):: 29 rows lost, 0.1% of our dataset

- bird_seen (imputed): none lost, 8.2% of our dataset

- playback_used (imputed): none lost, 7.7% of our dataset

- date (imputed): none lost, <0.1% of our dataset

- time (remove) (for now), 1240 rows lost, >3% of our dataset

Discussion: We are not missing any audio data. Some metadata is missing. Time missing values will be imputed after better error correction. Handled during processing and EDA via imputation and amputation. Discussion in full in .ipynb.

OUTLIERS:

- The major outlier identified was elevation. There may be outliers intrinsic to the audio data that need

some form of correction, but as of right now that is a hard compute task and should not be undertaken for EDA.

Feature engineering and analysis

[In this section, we identify the variables that are useful for predictive modeling and machine learning through correlation analysis. You may also reduce the dimension or derive new variables so that the predictive modeling can be more efficient and effective.]

- Place feature engineering features here so I can work it in.

Discussion: Our model will focus on the learning that can be imparted from the spectrogram representation of the audio files. This means that the creation of the spectrogram is feature engineering. We are turning this problem from an audio classification to a visual classification problem. There is no need to identify specific standout features of this, as the entirety of the audio is important for us, noise and signal. To completely eliminate noise will introduce instability into the model. During training and testing we will be supplying our attempts with multiple versions of data, including filtered and unfiltered data. Some techniques which we intend to test for filtering include pre-emphasis (López-Espejo et al., 2024), spectral subtraction (instead of spectral gating) (E. Verteletskaya, 2011), and median filtering. These are all relatively simple and computationally inexpensive techniques that should provide some value. Based on the attempts of others, we should be able to get competitive results without these techniques, but we will nonetheless train a model with and without this data for sake of comparison.

Work and research is currently ongoing into signal isolation, we do not have an automated process for this yet, nor are we sure we will be able to create one for this project. The intent is to further refine the training data, and assist future theoretical production scenarios, by isolating the areas of an audio (spectrogram) file by identifying areas of the sound which have the most spectro-energy and cutting away all but that area. To enable consistent training, if we are able to build an automated isolation process, we will bookend synthetic white noise to the audio snippet..

Appendix

[Provide the code or pseudo code, and any other information in the appendix here.]

The code can be found in the attached Jupyter notebook.

Table of Contributions

The table below identifies contributors to various sections of this document.

	Section	Writing	Editing
1	**Analysis the basic metrics of variables**	**Joe**	**All**
2	**Non-graphical and graphical univariate analysis**	**Jonathan, Max**	**All**
3	**Feature engineering and analysis**	**All**	**All**
4	**Appendix**	**All**	**All**

Grading

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.

References:

Ekaterina, Verteletskaya & Simak, B.. (2011). Noise Reduction Based on Modified Spectral Subtraction Method. IAENG International Journal of Computer Science. 38.

Iván López-Espejo and Aditya Joglekar and Antonio M. Peinado and Jesper Jensen (2024). On Speech Pre-emphasis as a Simple and Inexpensive Method to Boost Speech Enhancement. Accessed May 23 2024, https://doi.org/10.48550/arXiv.2401.09315