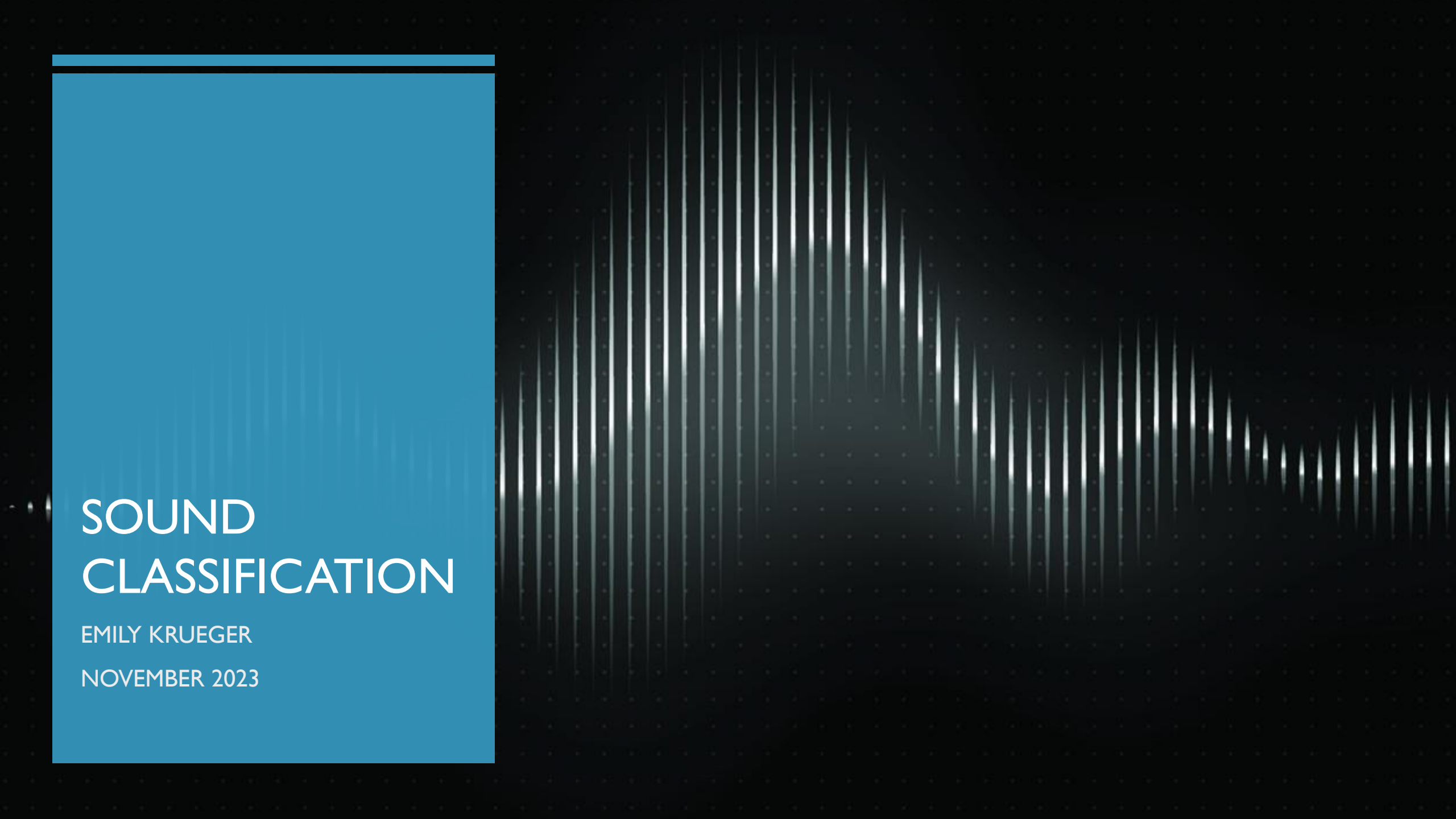
SOUND CLASSIFICATION

EMILY KRUEGER

NOVEMBER 2023



INTRODUCTION

- **Krueger Consulting** has been contracted to develop a sound classification model that can accurately distinguish between the following sound classes:

1. Music
2. Speech
3. Animal
4. Vehicle

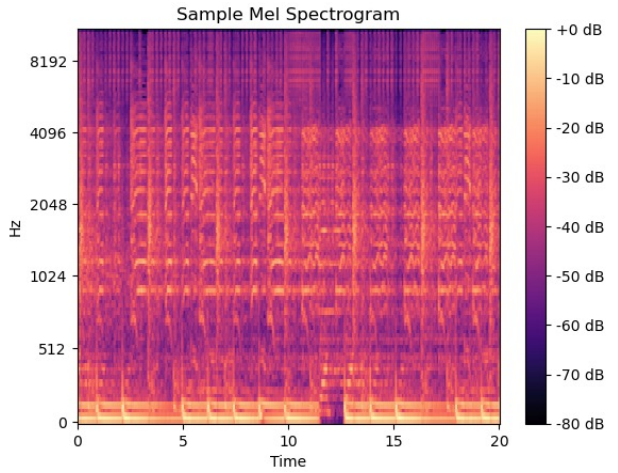
- I approached this problem in two ways:
  1. Converting audio to mel spectrograms to train a convolutional neural network for image classification
  2. Extracting numerical audio features from audio files to train various models including:
     - Random Forest
     - Logistic Regression
     - AdaBoosted Classifier (using both Decision Tree and Random Forest estimators)
     - XGBoost Classifier

DATA OVERVIEW

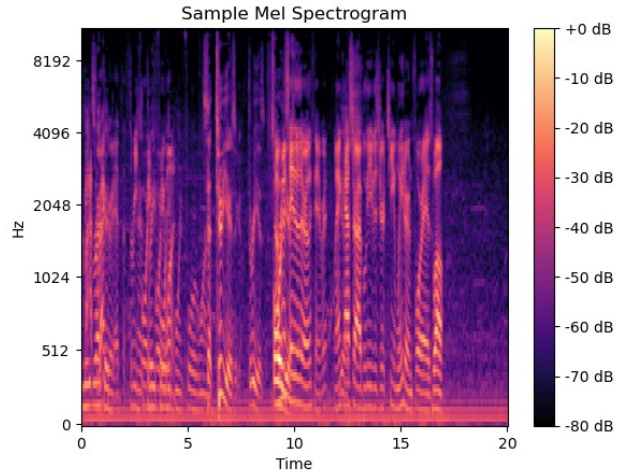
- Source: AudioSet, a publicly available dataset of approximately 2.1 million human annotated, ten second YouTube clips
  - A subset of the larger dataset was downloaded, focusing specifically on rows labelled Music, Speech, Animal, or Vehicle
- Spectrogram creation and feature extraction was conducted using Librosa, a python package for music and audio processing and analysis
- Final dataset consists of [8,941] audio files from which we extracted 64 features

MEL SPECTROGRAMS

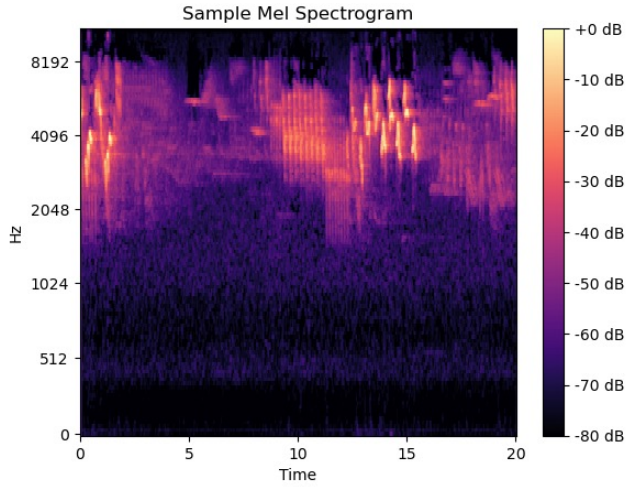
Music (EDM)



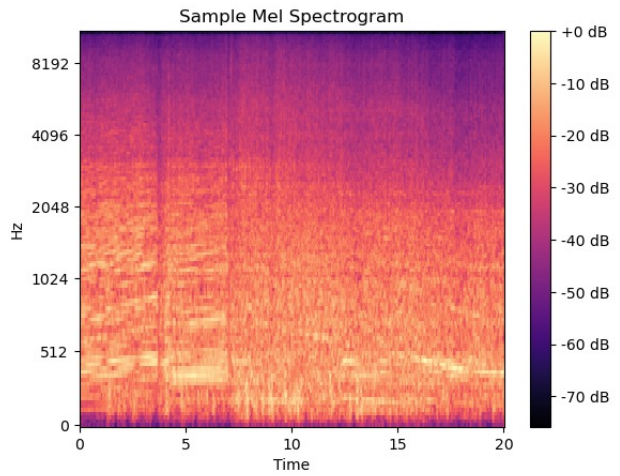
Speech (sports commentary)



Animal (Birds Chirping)

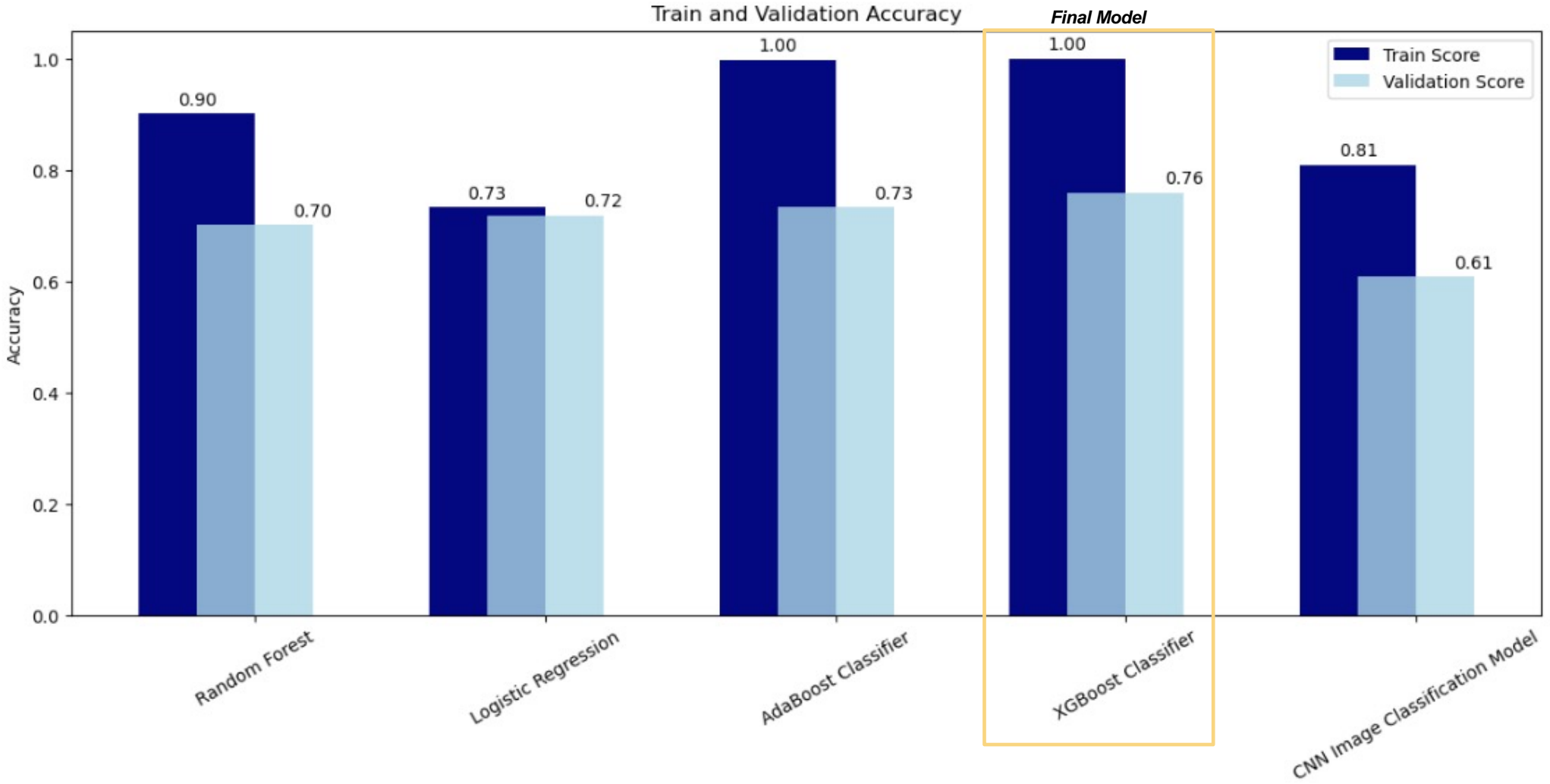


Vehicle (Engine Revving)

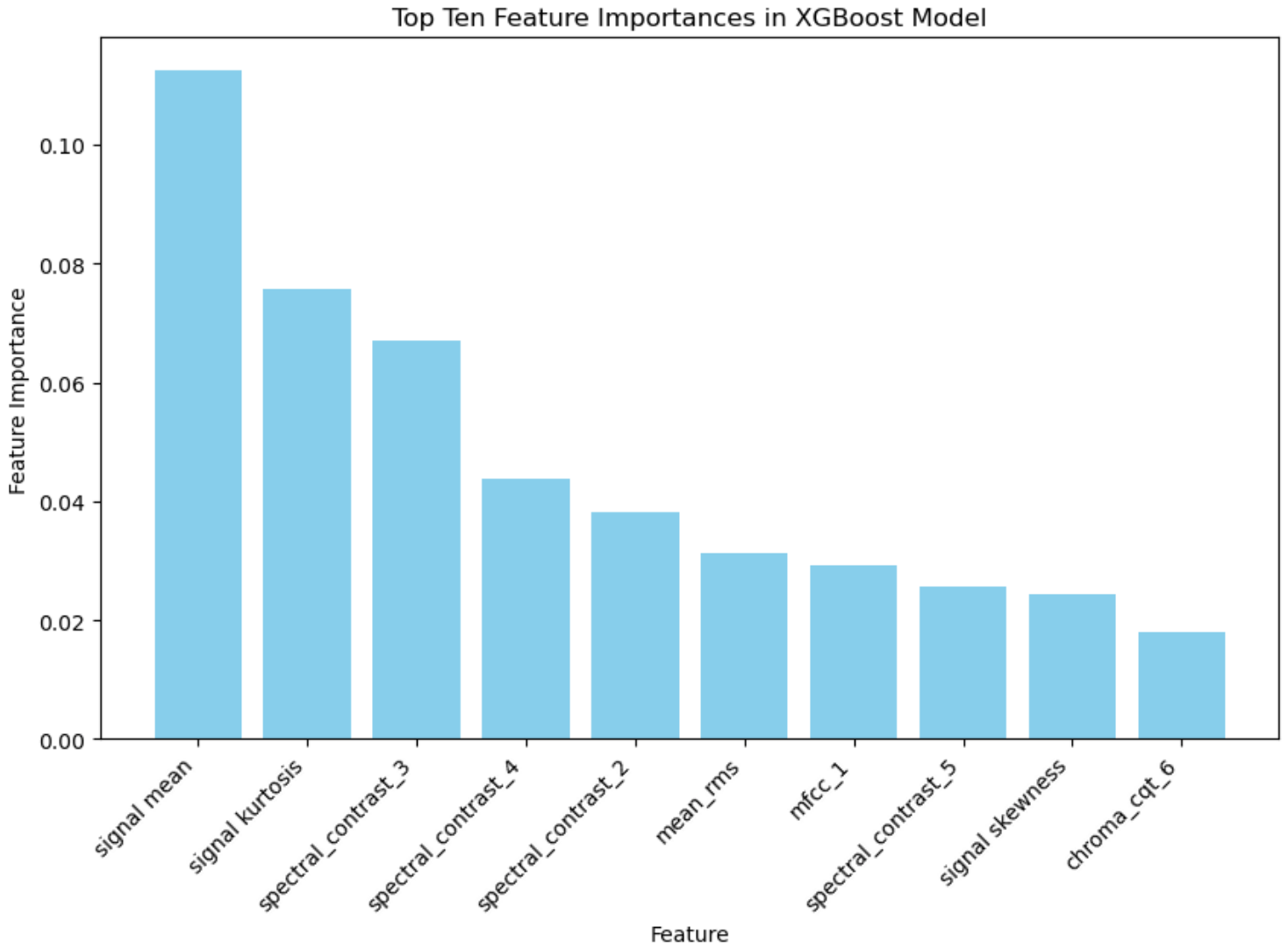


- A Mel-Spectrogram is a visual representation of audio with time on the x-axis, frequency on the y-axis, and color representing amplitude
- In creating Mel Spectrograms, the linear frequency scale of the original signal is converted to the Mel Scale, which better reflects the way in which humans perceive pitch

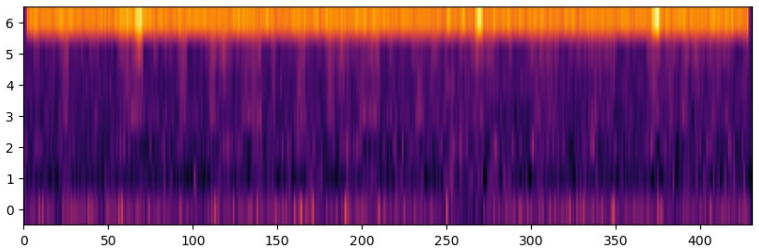
MODEL COMPARISON



FEATURE IMPORTANCE - XGBOOST



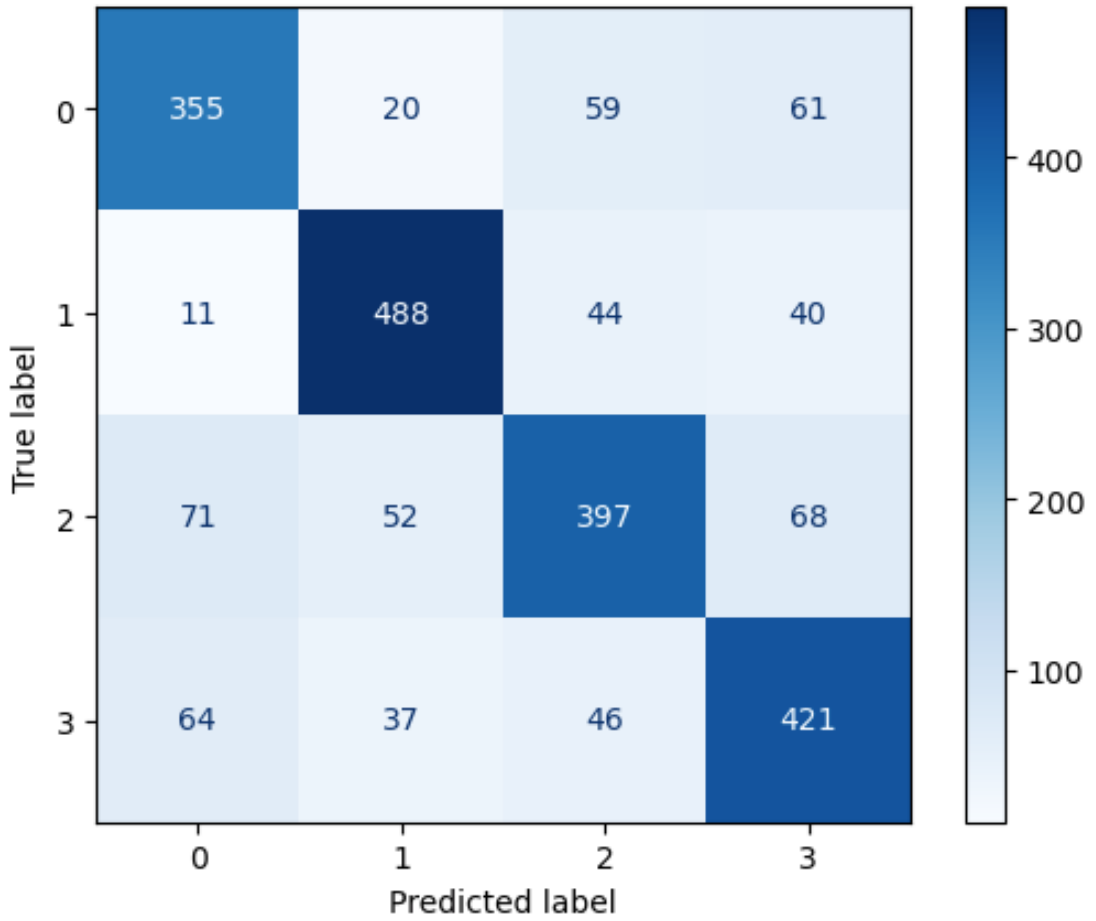
- Signal mean is the most meaningful feature in determining sound classification
- Spectral contrast across four different frequency bands is also impactful in determining sound classification
- Example of visualized spectral contrast with six bands:



FINAL MODEL EVALUATION ON UNSEEN DATA

- ***Business Case:*** **McStreamy**, a top streaming platform, has thousands of users uploading audio files on a daily basis. The platform hosts both music and podcasts. The platform is in need of a model that can accurately distinguish between music, speech, vehicle, and animal in order to i) label and segment out music and podcasts and ii) reject any attempts to upload audio that is not music or podcasts, such as animal and vehicle sound

Our model was tested on data obtained from McStreamy and yielded the following results:



Accuracy: 74.4%

Precision: *74.3%*

Recall: *74.4%*

F1: *74.3%*

CONCLUSION

- **Recommendations:**
  - I recommend using this model to classify various sounds. For example, one could feed this model a set of audio files from a streaming platform in order to segment out songs and podcasts and discard audio that is not a song or podcast, such as animal and vehicle sounds
  - One could also use this model as a starting point to develop a virtual assistant that can both recognize and interpret speech and music, provide song information, or respond to requests

Additional Considerations/Next Steps:

- Given the likelihood of overfitting in most models, this is just a starting point. Some next steps include.
  1. Train models on a larger dataset
  2. Add additional sound classes to increase number of use cases

BIOGRAPHY

Emily Krueger

Email: ekrueger1217@gmail.com

M: 732-403-4566

Github: https://github.com/ekrueger1217

LinkedIn: https://www.linkedin.com/in/emily-krueger-058513103/