Keyword Spotting for on-board AI: A Supervised ML Project

Han geul Lee - Brown University - https://github.com/johannes-bruhms/data1030project

Supervised ML Task: Keyword Spotting (KWS)

Goal: Train a model for 1-second audio inputs and detect which of the ‘core 10’ words was spoken.

Problem Type: Multi-class Classification

Features (X): The raw audio signal (or, more commonly, a processed representation like a Spectrogram).

Target (Y): The word label (a categorical variable, e.g., "Go").

Challenges:

- ~105k rows
- Non-IID

The Dataset

Google Speech Commands (v2) - Audio dataset of short, one-second spoken words

Dataset Statistics:

- **Total Samples:** ~105,829 audio clips
- **Total Classes:** 35 unique words (e.g., "Yes", "No", "Go", "Stop", "Cat", "Dog"...)
- **Audio Format:** 1-second .wav files (16kHz sampling rate)
- **Spoken Words:** The 35 target words.
- **Background Noise:** silence, machine hum, people talking, etc.

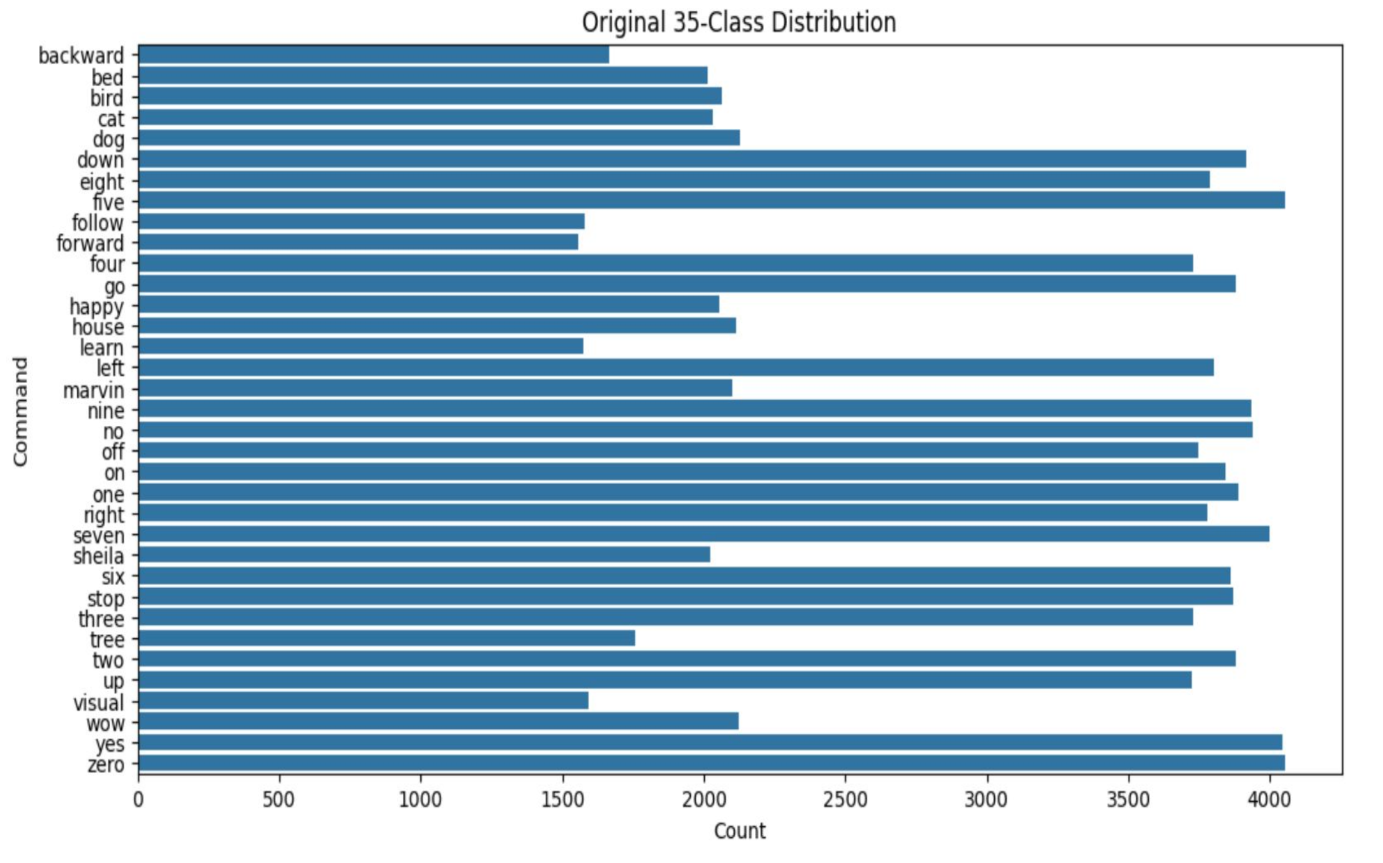
Preprocessing - Feature Engineering

- Raw 1-D 16,000 numbers
- **MFCCs** (80 features)
- **Spectral Centroid** "brightness." (2 features)
- **Zero-Crossing Rate** "noisiness." (2 features)
- **Librosa package**

extract_features( )

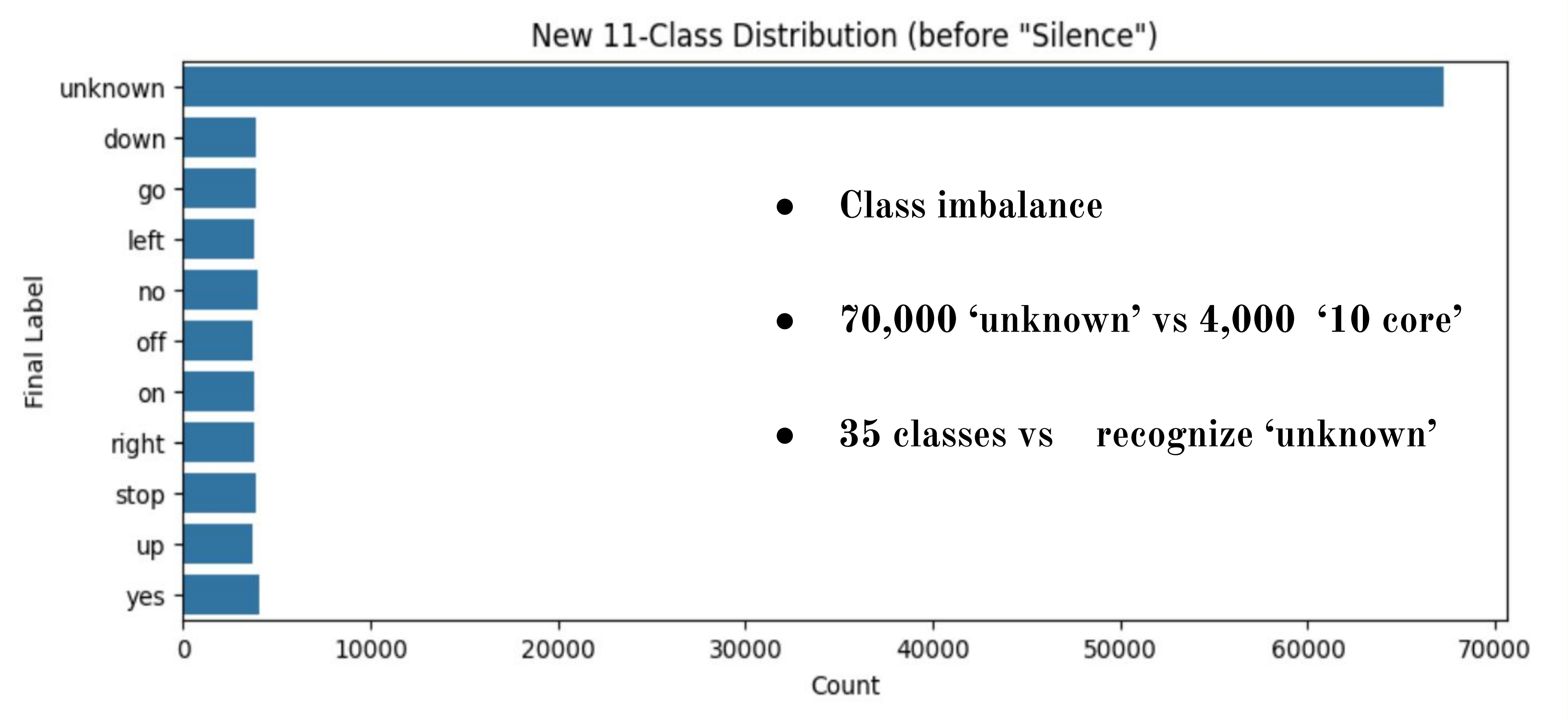
n_mfcc=40	calculates 40 coefficients.
Mfccs_mean	(40 features)
mfccs_std	(40 features)
spec_centroid_mean	(1 feature)
spec_centroid_std	(1 feature)
zcr_mean	(1 feature)
zcr_std	(1 feature)

EDA - I - Initial Classes



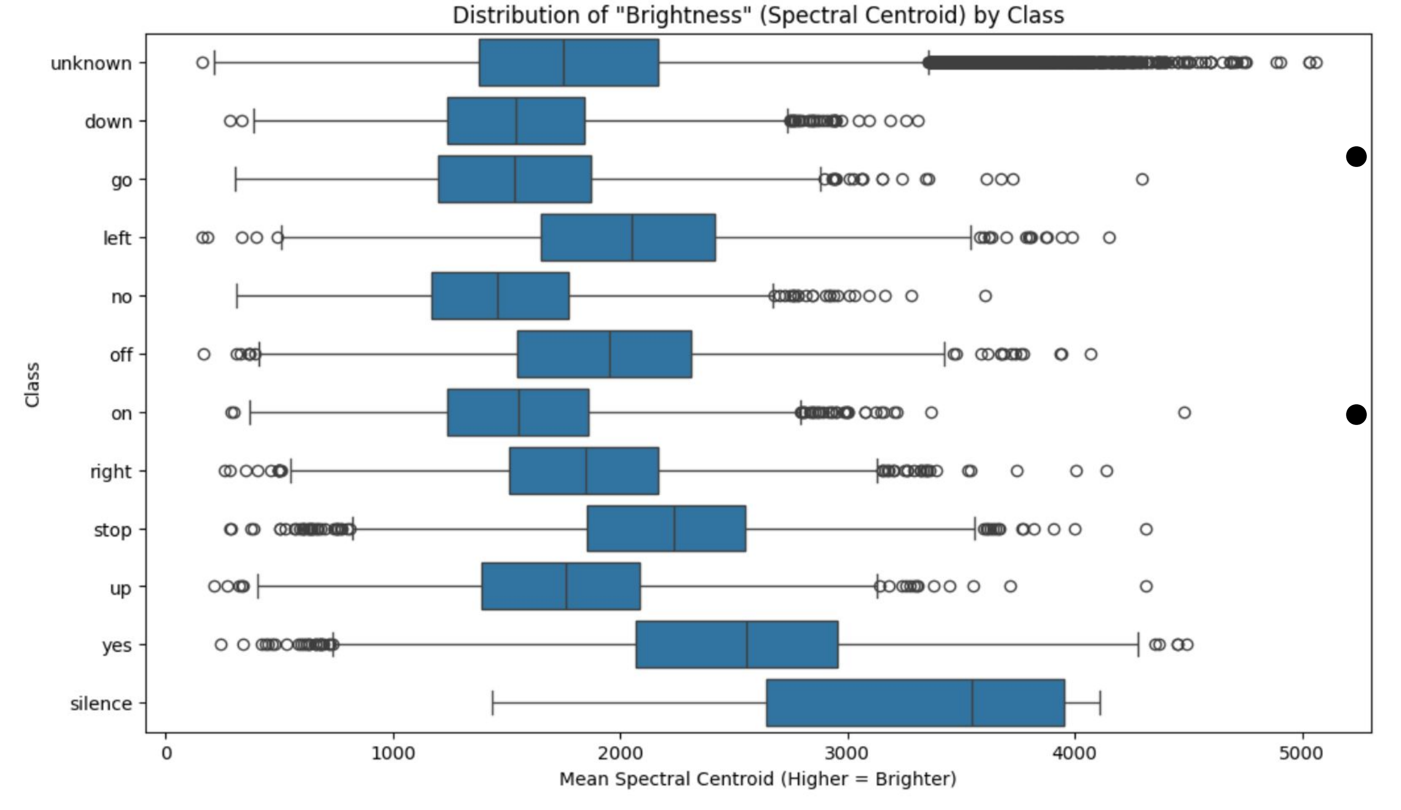
- **[10 ‘core’ classes + ‘unknown’]**
- **High-value vs Low-value**
- **35 classes vs recognize ‘unknown’**

EDA - II - New Distribution & Imbalance



- **Class imbalance**
- **70,000 ‘unknown’ vs 4,000 ‘10 core’**
- **35 classes vs recognize ‘unknown’**

EDA - III - Feature Separability



- 'Silence' vs spoken words.
- Core classes show “sound profiles”

Data Splitting Strategy: Speaker-Based (GroupShuffleSplit)

- **Prevent data leakage**
- **Split the list of 2,618 speakers 80/10/10**
- **Preprocessing Pipeline:** `ColumnTransformer => StandardScaler`
- **Final Data Shape:**
  - **Before: (105,829 samples, 1 audio file)**
  - **After: (105,829 samples, 84 scaled features)**