Speakers' age prediction with eXtreme Gradient Boosting

Massimiliano Carli
Data Science and Engineering
Politecnico di Torino
Torino, Italy
s337728@studenti.polito.it

***Abstract*—In this report, we address the *Speakers' Age Prediction Problem*, concerning the prediction of a person's age using an audio sample of their voice. The proposed solution consists of extracting various features directly from audio samples, such as Mel-log spectrograms, Mel-Frequency Cepstral Coefficients (MFCCs), and silence contour information, combining them with other descriptors of audio samples, such as the number of words spoken in the audio, pitch information and others. We then fine-tuned an eXtreme Gradient Boosting regressor (XGBoost), demonstrating how XGBoost outperforms our baseline, represented by a simple Ordinary Linear Regressor (OLR) trained on the same features.**

***Index Terms*—Age prediction, XGBoost, Mel-log spectrogram, MFCCs, Silence contour**

I. Problem Overview

The goal of this task is to predict speakers' age from their audio recordings, with evaluation based on the Root Mean Squared Error (RMSE). The dataset provided consists of two parts:

- **Development set:** containing labeled audio samples with some pre-extracted features and corresponding age labels.
- **Evaluation set:** containing unlabeled audio samples, with the same pre-extracted features as in development, for which predictions are submitted for scoring in the competition leaderboard.

The pre-extracted features provided include measures such as ethnicity of the speaker, gender of the speaker, pitch, jitter, shimmer, energy, zero-crossing rate (ZCR), spectral centroid mean, tempo, harmonic-to-noise ratio (HNR), number of words, number of characters, number of pauses, and silence duration.

II. Proposed approach

A. Preprocessing

The development dataset provided did not present any missing value. Nevertheless, some features required some preprocessing in order to be properly used for our regression task. More specifically the `'tempo'` feature was provided as a string presenting other characters at the start and end of the value, this required to preprocess the entire feature by trimming on both sides and converting in a float number.

We encountered that the ground truth label representing the person's age was mainly frequent in 15 years old, 25 years old range, with substantially lower frequencies as the age increased. As can be seen in Figure 1. Such skewness could be addressed with a box-cox transformation. Nevertheless, it resulted in poorer capacity of the model to predict final data, we therefore opted to avoid scaling and train our model on original labels.

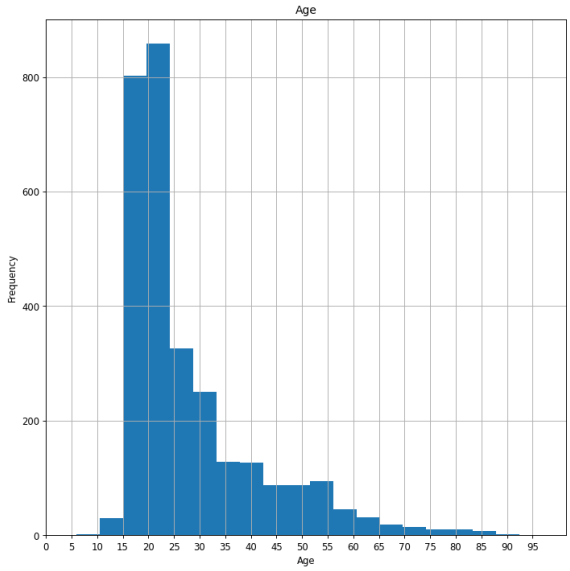

Fig. 1. Histogram of the age ground-truth label. Most speakers are ranged between 15 years old and 25 years old. Indicating imbalance in the dataset, with young speakers being overrepresented compared to older ones.

Ethnicity feature was also highly biased towards specific categories, namely `'igbo'` ethnicity and `'english'` ethnicity, which were the two most common one among all possible 165 options, as can be seen in the histogram in Figure 2. To address this problem we encoded each possible value of ethnicity in 3 main categories: 'igbo', 'english' and 'others', with the latter representing a default placeholder for all infrequent remaining ethnicities. We justify this solution since numerical or boolean values are excepted by our models (XGBoost and OLR, see II-B) to be properly trained, and a naive dummy encoding on the whole values of ethnicity would

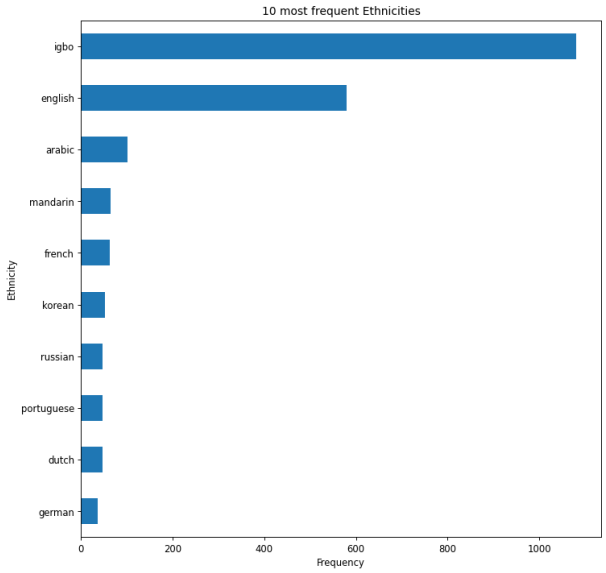

Fig. 2. Histogram of ethnicity distribution. The majority of speakers are 'Igbo', followed by 'English'. An elbow-like shape is observed, with other ethnicities being significantly underrepresented. By clustering all non-Igbo and non-English ethnicities as 'others', we obtain a set of 1,273 points, which is comparable to the 1,081 Igbo and 579 English speakers, despite the latter potentially being slightly underrepresented.

have result in a sparse matrix.

In addition to the pre-extracted features, we computed other measures directly from audio samples, more specifically

- Log-Mel Spectrogram
- MFCCs
- Silence contour

Log-Mel Spectrogram is a time-frequency representation of audio that uses the Mel scale for the frequency axis and applies a logarithmic transformation to the amplitude. This spectrogram mimics the way the human ear perceives sound by emphasizing frequencies that humans are more sensitive to [1]. The computation involves segmenting the audio into overlapping windows, for then extracting statistical components for each. In this case, a window size of 2048 samples is used, with successive windows defined every 512 samples. The resulting spectrogram is then mapped to 20 Mel-scaled frequency bands. Essentially, a spectrogram is a two-dimensional matrix, with one dimension representing the Mel-frequency bands and the other representing time intervals. For each Mel-frequency band, the mean and standard deviation are calculated as final features for our model, i.e. 40 new features in our dataset.

Mel-Frequency Cepstral Coefficients (MFCCs) [2] [3], similarly to the log-mel spectrogram, are representations of the audio spectrum derived from the Mel scale. They are defined as coefficients that describe the shape of the power spectrum of an audio signal. MFCCs are particularly popular in voice-recognition related tasks, including speech recognition, speaker identification, and emotion detection. MFCCs can be computed with a partial-overlapping windowing technique, in our case we adopted windows 2048 samples long, shifted every 512 samples, for each window we extracted 20 MFCCs, for then calculating the mean and standard deviation of each coefficient, virtually obtaining a comprehensive representation of 20 MFCCs for each audio. i.e. 40 new features in our dataset.

Silence contour quantifies the distribution of silence within an audio recording. By dividing the signal into 20 non-overlapping windows, we measured the proportion of silence in each window using a predefined threshold for amplitude of 20 db. The resulting contour provides valuable information for identifying pauses or silence intervals, we computed the overall mean silence duration over the 20 windows of the given audio sample, together with its standard deviation and median, obtaining 3 new features in our dataset.

Furthermore, we combined pre-computed features to capture non-linear insights. For instance, words per second, pitch range, mean-to-max pitch ratio, energy-to-duration ratio, energy-to-silence ratio, number of words per silence, and silence ratio (proportion of silence in the audio).

B. *Model selection*

We implemented eXtreme Gradient Boosting (XGBoost) regressor [4] as our model for prediction. Gradient boosting models build an ensemble of decision trees sequentially, with each tree correcting the errors of the previous one. XGBoost is an enhanced version of traditional gradient boosting, offering improvements in speed, scalability, and most importantly, it incorporates regularization techniques (L1 and L2) to prevent overfitting. Furthermore, since the nature of regressors to return decimal values as prediction, we rounded the result to the most close integer value, this post-processing procedure did not determine a worse RMSE, while producing a coherent output wth the nature of the datasets provided.

We furthermore compared XGBoost with an Ordinary Linear Regressor (OLR) trained on the same features as a benchmark for our results.

C. *Fine-tuning*

Tuning our model consisted of an 80/20 holdout on our development set, reserving 20% of the data as a test after training and validation. For the remaining data, we implemented Grid Search Cross Validation with 3 folds. While XGBoost is highly effective for machine learning tasks, its performances depend on the selection of hyperparameters which span among many options, making the fine-tuning computation heavier[1]. The hyperparameters we decide to tune can be broadly categorized into three groups: (i) tree structure, (ii) regularization, and (iii) sampling. However, the distinction between these groups is

[1]We could reduce time for training thanks to multi-threading, and the implementation of XGBoost, such implementation also support GPU-based training and prediction, future work could rely on this option for further speeding up fine-tuning phase.

TABLE I
HYPERPARAMETERS FINE-TUNING. WITH $k = 3$ FOLDS RESULTED IN 8748 DIFFERENT FITS.

Parameter	*Values*
max_depth	3, 5
n_estimators	500, 1000
gamma	0.3, 0.5, 0.8
reg_alpha	0, 10, 100,
reg_lambda	0, 10, 100
subsample	0.4, 0.6, 0.8
colsample_bytree	0.4, 0.6, 0.8
learning_rate	0.01, 0.05, 0.1

not always clear, for instance certain tree-structure parameters, such as maximum depth, can also induce regularization. More specifically, we selected:

- **Tree Structure**: Maximum depth of trees (max_depth), number of trees (n_estimators), and minimum loss reduction required to split a node (gamma).
- **Regularization**: L1 regularization (reg_alpha) and L2 regularization (reg_lambda).
- **Sampling**: Fraction of samples (subsample) and features (colsample_bytree) used for fitting each tree.
- **Learning Rate**: Controls the contribution of each tree to the final model.

The values of these hyperparameters that passed under validation are reported in Table I. These values came from multiple rounds of Grid Search, where we tested different ranges for each parameter and narrowed them down to the ones that could more probably report the best combination.

III. RESULTS

The best configuration found for hyperparameters of XGBoost are

- `max_depth`: 5
- `n_estimators`: 1000
- `gamma`: 0.8
- `reg_alpha`: 0
- `reg_lambda`: 100
- `subsample`: 0.4
- `colsample_bytree`: 0.4
- `learning_rate`: 0.05

These resulted in an RMSE score on the local test set (i.e. the 20% holdout slice of the dataset mentioned before, unseen at fitting-time) of 9.199. XGBoost outperformed an OLR model trained on the same features, which produced a local RMSE score of 9.954, underlining the better capability of XGBoost to model non-linear relationships between data. The model was then re-fitted on the whole data available in development to optimize the final prediction, which produced a 10.283 RMSE on the public leaderboard.

IV. DISCUSSION

In this work, we explored the problem of predicting speakers' age from audio samples using an eXtreme Gradient Boosting (XGBoost) regressor. Challenges such as dataset imbalance and feature encoding biases were addressed with preprocessing strategies, nevertheless a more balanced dataset could have improved generalization of the model. By leveraging both pre-extracted features and additional audio descriptors - including Log-Mel spectrograms, MFCCs, silence contours, and derived non-linear relationships - we demonstrated that XGBoost outperforms a baseline model trained with ordinary linear regression. Hyperparameter fine-tuning further optimized the performance of XGBoost, achieving sufficient results in the development test set.

REFERENCES

[1] E. B. Goldstein, *Sensation and perception*. Wadsworth/Thomson Learning, 1989.

[2] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122 136–122 158, 2022.

[3] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various mfcc implementations on the speaker verification task," in *Proceedings of the SPECOM*, vol. 1, no. 2005. Citeseer, 2005, pp. 191–194.

[4] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.