Audio Speech Analysis for Age Estimation through Machine Learning Regression

Andrea Lolli s346165
Data Science And Engineering
Politecnico di Torino
Turin, Italy
s346165@studenti.polito.it

Daniele Famà s345153
Data Science And Engineering
Politecnico di Torino
Turin, Italy
daniele.fama@studenti.polito.it

***Abstract*—In this report, we present a potential solution to the *Age Estimation Regression Problem* posed in the competition. The proposed approach combines provided tabular data with summary statistics extracted from the audio signals. We will aim to maximize the score in the leader-board while maintaining a good generalization ability. The approach achieved competitive results, demonstrating its effectiveness in addressing the challenges of the task.**

***Index Terms*—speech recognition, regression analysis**

I. Problem overview

We tackle the problem of target age estimation using **machine learning regression models**. Our objective is to assess the impact of different features and techniques on task performance, while balancing the trade-off between model generalization and leaderboard evaluation metric:

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \tag{1}$$

The dataset provided comprises the following components:

- **Tabular data**: This includes acoustic and linguistic features extracted from audio recordings, as well as a small set of demographic features.
- **Audio recordings**: These correspond to the tabular data and provide additional raw information.

The dataset is structured into two subsets:

- **Development set**: Contains 2,933 samples with 17 columns, including age labels, alongside a folder with the corresponding audio recordings;
- **Evaluation set**: Comprises 691 samples with 17 columns and a folder containing the associated audio recordings.

The tabular features include statistical summaries of audio attributes such as pitch, jitter, and energy, as well as features related to speaking style, such as the number of pauses, word count, character count, tempo, and others.

A. Preliminary Observations

Analysis of the development set reveals key insights into the dataset composition. Figure 1 illustrates the age distribution, showing a high concentration of individuals aged between 17 and 30 years. For individuals above the age of majority, the distribution mirrors the unbalanced age structure of the global population. Ethnicity distribution further supports this observation. The **Igbo** ethnicity is the most represented group, and its age trends align closely with the Nigerian population structure[1]. Outside the Igbo ethnicity, manual inspection reveals that only a few ethnicities are common between the development and evaluation sets. Specifically, there are 165 and 73 ethnic groups in the development and evaluation sets, respectively, with only 17 shared between them.

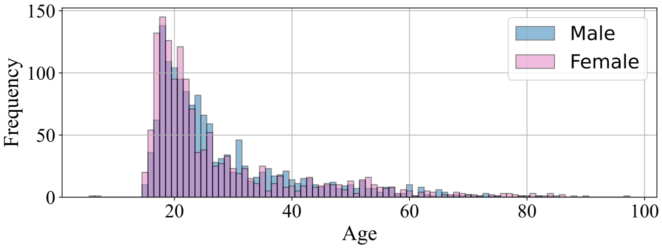


Fig. 1. Age distribution of development set broken down into genders

B. Outlier Detection & Analysis

To identify potential outliers, we computed percentiles (1st, 10th, 20th, ..., 90th, 99th) along with the minimum and maximum values of each feature distribution. Features exhibiting significant skewness and large value ranges were investigated. Outliers were defined as data points lying outside the range between the 1st and 99th percentiles, with special focus on the skewed side of the distribution.

In particular, we examined features related to pitch statistics (mean, max, min) and, separately, those related to the number of pauses and silence duration. Plotting the target age values associated with these outliers revealed that silence-related outliers were predominantly linked to adult or elderly individuals, a demographic underrepresented in the dataset, as shown in Figure 2. Retaining these data points was deemed essential to preserve this minority group's representation.

C. Audio signal review

The audio recordings are stereo, sampled at 16 bits per channel. The variable length of the recordings introduces a challenge that will be addressed later in this report. Conversion

[1] Source: **Population pyramids of Nigeria in 2024**

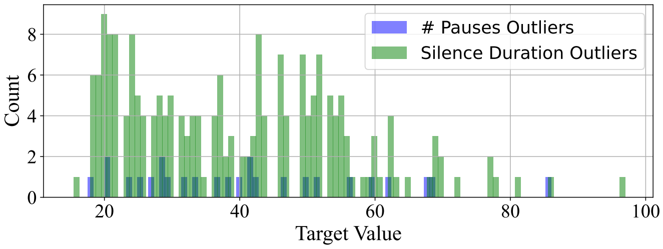
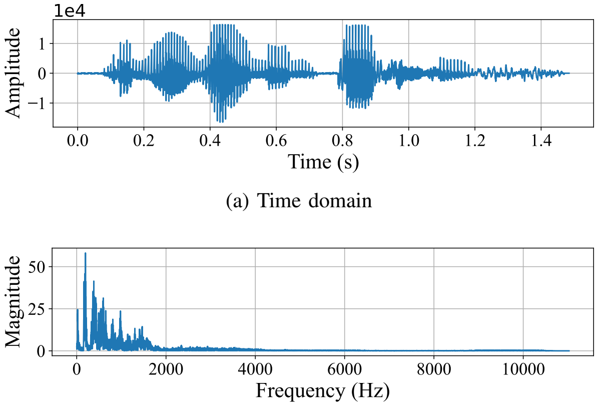


Fig. 2. Distribution of silence-related outliers and their corresponding target age values.



(a) Time domain

(b) Frequency domain

Fig. 3. Visualization of an audio signal in the time (a) and frequency (b) domain

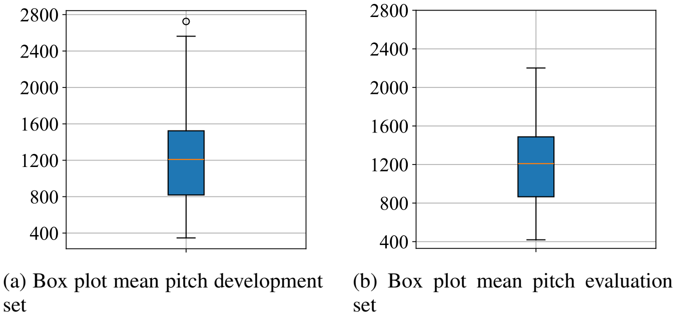
to mono is performed by Librosa, which averages the left and right channels.

The frequency domain plays a crucial role in speech analysis, as different frequency bands carry varying levels of energy [1]. Both the time (Figure 3a) and frequency (Figure 3b) domains provide complementary insights into audio signals. To leverage this, we employ spectrograms, which are widely used in speech processing. Specifically, mel-spectrograms are utilized as they represent the spectral content of audio signals while emphasizing frequency ranges relevant to human speech [2].

Furthermore, phonetic studies have demonstrated that features such as the fundamental frequency correlate with speaker age. As age increases, there is a noticeable decrease in the frequency of voiced sounds [3]. This observation underscores the importance of including both time- and frequency-domain features in modeling efforts.

D. Considerations

The unbalanced age distribution poses a risk of model overspecialization on the majority age group, potentially degrading performance for less frequent target ranges. Additionally, the high cardinality of the ethnicity feature—most of whose values are unique and non-overlapping between subsets—may lead to overfitting on training data and poor generalization to unseen data during evaluation. Moreover, dimensionality could increase substantially if this feature is encoded without careful processing.



(a) Box plot mean pitch development set

(b) Box plot mean pitch evaluation set

Fig. 4. Box plots of mean pitch in the development and evaluation sets.

Outlier removal also requires careful consideration. While outliers may distort the overall feature distributions, they can also represent valuable information, particularly for minority demographics. Thus, further investigation is needed to balance data preprocessing with retention of critical samples and models performance.

We conducted an analysis of the audio recordings. Figure 3 presents an example of a signal in the time domain, where silences are observable before and after the utterance, along with the presence of background noise.

II. Proposed Approach

A. Data preprocessing

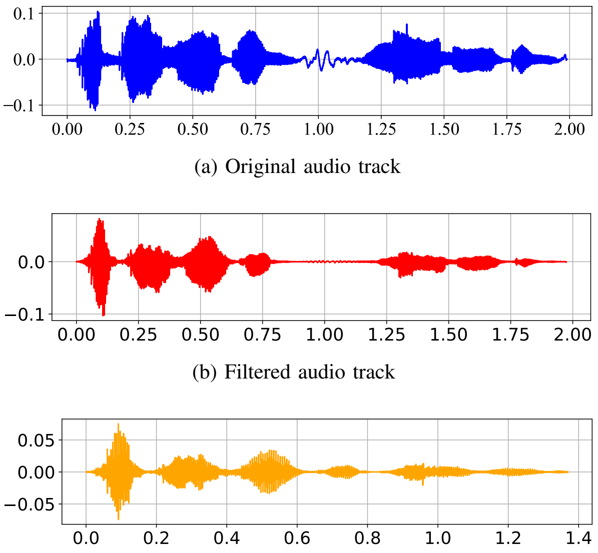
1) Tabular Data: Univariate outlier detection focused on pitch-related features (mean, max, min).

Boxplots (Figure 4) were used to compare the distributions of these features in the evaluation and development sets. Initially, outliers were defined as points outside the 1st to 99th percentile range. These thresholds were refined by aligning them with the maximum range observed in the evaluation set for each feature. This adjustment retained potentially linked points between the two sets while effectively eliminating genuine outliers.

For multivariate outlier detection, we employed DBSCAN, a density-based clustering algorithm that identifies clusters of densely packed points and labels points in sparse regions as outliers [4]. To determine the optimal density threshold (ϵ), we analyzed the distances to each point's 12th nearest neighbor using a k-distance plot. The ”elbow” of the plot guided the selection of ϵ, balancing the inclusion of dense clusters and the exclusion of noise. Points not meeting DBSCAN's density criteria were labeled as outliers. These points were associated with instances that occurred very frequently, meaning their removal would not disrupt the overall data distribution.

Finally, indices of outliers identified through both univariate and multivariate analyses were combined and flagged for removal as part of the preprocessing pipeline.

We examined the correlation between features and observed a strong linear relationship between *num_words* and *num_characters*. Both features exhibited equal correlations



(a) Original audio track

(b) Filtered audio track

(c) Processed audio track

Fig. 5. Preprocessing raw audio data

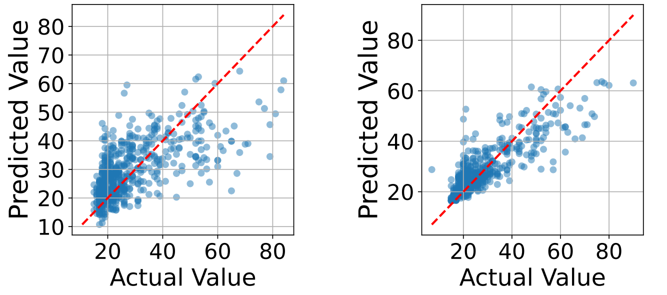
with other variables, indicating redundancy. Consequently, one of these features was removed to avoid unnecessary duplication of information.

Categorical preprocessing involved binarizing the *gender* column and converting the ethnicity column into a binary *isIgbo* feature, indicating whether an instance belonged to the Igbo ethnicity.

Scaling techniques, such as *Min-Max Scaling* and *Standard Scaling*, were applied to non-binary numerical features. To analyze feature relevance, we employed a Random Forest model for feature importance analysis. Results indicated that the *isIgbo* column had minimal contribution to the model, making it a candidate for removal. To evaluate the impact of feature removal, we tested the Random Forest regressor after excluding the *isIgbo* and *num_words* columns. The results showed negligible changes in model performance, confirming these features could be safely removed as part of the preprocessing pipeline.

2) *Audio Features:* Raw recordings (figure 5a) are denoised through *noiserecude* method [5, 6]. Then we filtered the frequncies below 40Hz and above 11025Hz corresponding to the Nyquist frequency. Figure 5c shows the processed recording where we trimmed out the pauses before and after the utterance as well as pauses within the recordings. While silences might provide useful information- for instance, an older speaker may exhibit a slower speaking rate- this information is already available in the tabular data. After processing the raw data, we addressed the varying lengths of the recordings by extracting statistical measures from the mel-spectrogram.

We divided the Mel spectrogram into six bands based on human perception, ranging from low to high frequencies with intermediate categories. In the frequency domain, we collapsed these bands by extracting the weighted mean frequency. The weights for this calculation are derived from the energy values



(a) Random Forest Regressor regressor before tuning

(b) Random Forest Regressor after tuning

Fig. 6. Actual values vs Predicted ones

of each frequency component. The weighted mean frequency for each band $\boldsymbol{B}$ at each time frame t is defined as:

$$\mu_B(t) = \begin{cases} \dfrac{\sum_{i\in B} f_i S_{i,t}}{\sum_{i\in B} S_{i,t}} & \text{if } \sum_{i\in B} S_{i,t} \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

$S_{i,t}$ is the energy value from the Mel spectrogram at the i-th Mel band and t-th time frame. This formula calculates the centroid frequency for each frequency band and time frame by taking the weighted average of the center frequencies, where the weights are the corresponding energy values. If the total energy within a band for a specific time frame is zero, the weighted mean frequency is defined as 0 Hz to avoid division by zero.

Statistical measures such as mean, standard deviation, kurtosis and skew have been extracted for each band of the spectrogram and other common audio features such as Mel-Frequency Cepstral Coefficients (MFCCs) and Spectral features. We also extracted fundamental frequency computing, in addition, its minimum and maximum.

B. *Model selection*

To select the models for evaluation, we conducted a general assessment of various candidates. The models with the best scores in a simple test setting were grouped into two categories: advanced models, utilizing boosting techniques such as GradientBoost and CatBoost, and classical models, including SVM and Random Forest.

For each category, we built a model pipeline comprising the following components: a data standardization step using `ColumnTransformer`, a sampling strategy such as `RandomOverSampling`, a dimensionality reduction technique like PCA, and the respective model. We optimized these pipelines by exploring a range of hyperparameters and selecting the best-performing combinations of sampling and dimensionality reduction techniques through a `RandomSearch` approach. This process identified Random Forest as the top performer in the classical models category and CatBoost in the advanced models category, both with the support of the *Linear Discriminant Analysis* dimensionality reduction technique.

TABLE I
HYPERPARAMETER CONSIDERED

Model		
Random Forest	n_estimators	randint(80, 400)
	max_depth	None, $2 \rightarrow 6$ step 1
	min_impurity_decrease	uniform(0.0, 0.02)
	min_weight_fraction_leaf	uniform(0.0, 0.02)
CatBoost	iterations	randint(500, 2500)
	depth	$2 \rightarrow 6$ step 1
	learning_rate	uniform(0.01, 0.1)
	l2_leaf_reg	randint(1,11)
	bagging_temperature	uniform(0, 2)

Figure 6 presents the plots comparing predicted values against actual values. As evident from the data, the errors are more pronounced for predictions involving adult and elderly subjects. We hypothesize that this behavior arises due to the scarcity of such data points. During testing on the development set (using train-test splits), the model encounters a lack of representation for these age groups, forcing it to generalize predictions based on the most frequent values in the dataset. Consequently, when the model is trained on the entire development set for evaluation on the test set, after the final tuning, we expect improved performance due to the inclusion of more diverse data points in the training process.

C. Hyperparameters tuning

Separetely we detected two category of hyperparameters:

- `top_db`: as this parameter decreases, a more aggressive cut is applied to recordings;
- Random Forest and CatBoost hyperparameters

We manually tried several configuration of the parameter regarding the aggressiveness of the cut of the silences. It has been applied a more gentle (+10 db) cut the silences throughout the utterance than to trim the audio. We conducted a *random search* with a narrower range of hyperparameters, centered around the best values identified in the initial search. However, a grid search approach was not feasible due to computational constraints. Table I shows the hyperparameter considered for the random search.

III. RESULTS

The tuning shows that both the Random Forest and CatBoost achieve better results using when top_db $=35$. We also found out that results tend to be higher when the removal of outliers is not performed. Through the Random Search, the best configuration for random forest was found for {min_impurity_decrease: 0,02, min_weight_fraction_leaf: 0,02, n_estimators: 166, oob_score: True}(RMSE ≈ 8,77) whereas the best configuration for catboost was found for {bagging_temperature: 1,2, iterations: 2113, l2_leaf_reg: 2, learning_rate: 0,01} (RMSE ≈ 8,58). Having trained the best performing models on the whole development set, the public scores obtained are 9,091 for the random forest and 9,01 for the catboost (other configuration reached a public score 8,6 but we didn't evaluate them).

To provide a baseline for comparison, a naive solution was implemented. The steps for the naive approach were as follows:

- Applying one-hot encoding to the ethnicity feature;
- Using no feature extraction from audio recordings;
- Removing highly correlated features, such as *num_world* and *num_character*;
- Training a default Random Forest model on the processed features.

This naive approach resulted in a public leaderboard score of 10.1, demonstrating the advantages of our proposed methods over the baseline.

IV. DISCUSSION

The proposed approach achieved very promising results (1^{st} position in public score using `catboost`), significantly improving upon those obtained with a naive implementation. This improvement is attributed to the use of both time-domain and frequency-domain statistical measures. The application of dimensionality reduction techniques combined with a n exhaustive research of the best model's parameters lead us to achieve very good score in public leaderboard.

Empirically, boosting regression algorithms, such as `CatBoost`, demonstrated superior performance in terms of *RMSE*. This can be attributed to their robustness in handling complex relationships, ability to avoid overfitting, and flexibility in hyperparameter optimization [7]. These qualities collectively contributed to the effectiveness of our approach and the significant improvement over simpler implementations.

Several aspects could be explored to further enhance the approach:

- Better data extraction from spectrogram. Spectrogram could be chunked working in both frequencies bands and time windows. This approach allows for the capture of localized frequency patterns and temporal features, offering finer granularity in feature extraction.
- Specific data mining pipeline constructed to improve regression model based on gradient boost. By focusing on improving the pipeline around the gradient boosting framework, we could potentially reduce overfitting, boost model accuracy.
- Given the imbalance in the dataset, a potential improvement could involve generating synthetic data for adult age groups to better balance the distribution.

The results demonstrate accurate estimations for data points corresponding to the most frequently represented target group, while showing larger errors for older individuals.

REFERENCES

[1] V. S. Kone, A. Anagal, S. Anegundi, P. Jadhav, U. Kulkarni, and S. Meena, "Voice-based gender and age recognition system," in *2023 International conference on advancement in computation computer technologies (InCACCT)*. IEEE, 2023, pp. 74–80.

[2] J. Gowdy and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, vol. 3, 2000, pp. 1351–1354 vol.3.

[3] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013, special issue on Paralinguistics in Naturalistic Speech and Language. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230812000101

[4] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[5] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, p. e1008228, 2020.

[6] T. Sainburg, "timsainb/noisereduce: v1.0," jun 2019.

[7] A. V. Dorogush, V. Ershov, and A. Gulin, "Catboost: gradient boosting with categorical features support," 2018. [Online]. Available: https://arxiv.org/abs/1810.11363