# Regression Models for Speaker Age Estimation

Pietro Rossi and Matteo Destino
*Politecnico di Torino*
Student id: s330750, s329337
s330750@studenti.polito.it, s329337@studenti.polito.it

***Abstract*—The goal of this project is to develop a data science pipeline for predicting the speaker's age for each audio recording. This is achieved by extracting a diverse set of speech-derived features and computing various statistical descriptors as model inputs. Our approach first evaluates a Random Forest Regressor as the primary baseline. Performance is then further improved using Gradient Boosting and XGBoost, yielding satisfactory results.**

## I. Problem Overview

This project aims to develop a regression pipeline that predicts age for each spoken sentence using both extracted and existing dataset features.

The dataset consists of 3,624 speech samples, with 2,933 samples in the development set and 691 in the evaluation set. Each sample includes various acoustic and linguistic characteristics, along with metadata such as gender and ethnicity. The target variable is the speaker's age. We will need to use the development set to build a regression model to correctly label the points in the evaluation set.

All audio signals in the dataset are sampled at the same fixed rate, but their lengths differ due to variations in speech patterns. Some speakers are more verbose or repeat sentences, while others speak only briefly. Additionally, while voices remain clearly audible, certain recordings contain background noise.

To better understand the nature of data, we analyze signals in the time and frequency domains, as shown in Figures 1 and 2, providing complementary insights.

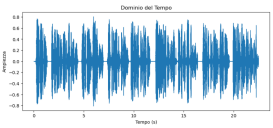


Fig. 1. Time-domain representation of an audio sample showing amplitude variations

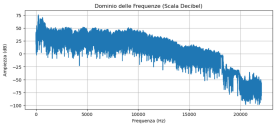


Fig. 2. Frequency-domain representation (db scale) of an audio sample, showing spectral content

A detailed exploratory data analysis reveals key insights:

- **Age distribution**: The dataset is **imbalanced**, with a higher concentration of younger speakers and fewer older speakers, as shown in Figure 3. This imbalance could lead to biased model predictions, favoring the more represented age groups.
- **Numerical Features**: Several features, such as energy and jitter, exhibit skewed distributions with long tails, which can negatively impact the model. Others, like *HNR* and *mean_pitch*, display a bimodal distribution, suggesting the presence of distinct speaker groups. Features such as *num_pauses* and *silence_duration* are concentrated near zero, potentially introducing bias into the regression model. Additionally, *num_words* and *num_characters* fall within a limited range of discrete integer values. Most remaining features follow an approximately Gaussian distribution.
- **Categorical features**: Categorical features require careful handling. The *gender* feature, with 'male' and 'female' categories (correcting 'famale' to 'female'), is well-balanced. In contrast, *ethnicity* has 221 distinct values, some well-represented and others with few instances. Including all ethnicities could increase dataset size and model complexity.

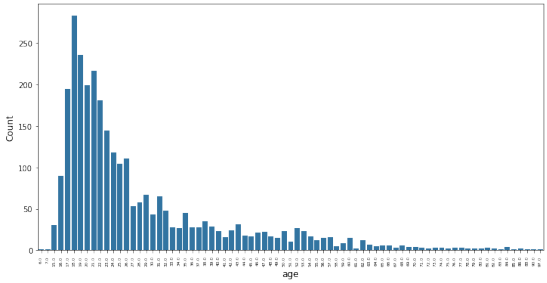


Fig. 3. Target variable 'age'

## II. Proposed Approach

### A. Preprocessing

We tested various preprocessing steps and implemented only those that significantly improved data quality and performance.

- **Age Distribution Adjustment:** To reduce the impact of infrequent extreme values, we removed samples with ages above 86 or below 10 from the training set, as they could negatively impact the model.
- **Normalization:** Normalization is unnecessary for our tree-based models, as they split data based on feature thresholds rather than distance metrics. Consequently, we did not apply normalization.
- **Feature Selection and Transformation:**
  - The *sampling rate* (22050 Hz) was removed as it was constant across all samples and non-informative.
  - The *tempo* feature, originally in a non-numerical format, was converted into a float for consistency.

- **Correlation Analysis and Feature Refinement:**
  - *num_words* and *num_characters* had a perfect correlation (1.0), making one redundant. We removed *num_words* to avoid duplication.
  - Some features showed significant correlation with age, particularly *HNR* (-0.45), *num_pauses* (0.48) and *silence_duration* (0.51), suggesting their possible relevance for prediction.
  - *num_characters* and *silence_duration* exhibited high correlation (0.92). Although initially considered redundant, experiments confirmed that both features provided valuable information, so we retained them.
- **Outlier Analysis**
  To improve model robustness, we explored outlier handling by analyzing feature distributions and visualizing them with box plots. For some features (*mean_pitch*, *shimmer*, *zcr_mean*, *spectral_centroid_mean*) we tested removing values beyond the first and third quantiles, as well as a more conservative approach focusing only on the most extreme outliers. However, both approaches increased RMSE, indicating that extreme values in speech data are important and reflect natural variability. Moreover, further reducing the already limited dataset hindered the model's ability to generalize.
- **Data Transformations**
  To address skewed distributions and large value ranges in features like *energy* and *min_pitch*, we tried to apply standardization and logarithmic transformations. These techniques are commonly effective for long-tailed distributions, but in our case, they did not result in significant improvements.
- **Discretization** Another approach we implemented was the discretization of features like *num_characters*, *num_pauses*, and *silence_duration*, where values were grouped into a few distinct ranges. However, this method did not result in a substantial improvement.
- **Categorical Feature Encoding:** Since Random Forest Regressor requires numerical input, categorical variables were transformed accordingly.
  - **Gender:** Encoded as a binary variable (male = 1, female = 0) using dummy encoding.
  - **Ethnicity:** Given its high cardinality, different encoding strategies were evaluated using a Random Forest Regressor:
    - Removing ethnicity led to a higher RMSE, indicating its relevance.
    - One-hot encoding improved performance but significantly increased dimensionality.
    - PCA applied after standard scaling did not enhance results compared to one-hot encoding.
    - To balance information retention and dimensionality, ethnicities with fewer than 10 speakers were grouped into an *other* category. This threshold was selected empirically, as it minimized RMSE.

    The best approach retained for subsequent analysis was one-hot encoding with grouped ethnicities.
- **Feature Extraction**
  We extracted speech-related features using `librosa` [1] and computed statistical descriptors to capture essential information, drawing inspiration from [2].
  - Before eliminating *num_words*, we introduced a new feature, **words_per_second**, which measures the rate of spoken words within a given time frame.
  - **Spectrogram:** The spectrogram represents how energy is distributed across frequencies over time, with age-related changes influencing these patterns. We used 40 mel frequency bins to capture the frequency range and extracted the mean and standard deviation for each bin.
  - **Mel-Frequency Cepstral Coefficients (MFCCs):** MFCCs model speech's spectral properties. We extracted 13 MFCCs, along with their first and second derivatives (*delta* and *delta-delta*), computing their mean and standard deviation.
  - **Spectral Features:** We extracted *spectral rolloff, flatness, contrast, bandwidth, chroma STFT, and polyfeatures* to capture additional frequency characteristics, retaining only the means.
  - **Rhythm Features:** We analyzed speech rhythm using tempogram features, calculating both the mean and standard deviation of the tempogram and the mean of the tempogram ratio for the first 13 bins, as these reflect articulation speed and fluency.

  To address the issue of noisy audio, we applied the *reduce_noise* function from the *noisereduce* package before extracting the same features. This approach is recommended by [3]. However, the results were worse compared to those obtained from the original, non-denoised audio.

### *B. Model selection*

The following regression algorithms have been tested:

*a) Random Forest Regressor:* This model combines predictions from multiple decision trees, each trained on random subsets of data and features, to improve predictive accuracy and reduce overfitting. Random forests are robust to outliers and handle non-linear relationships effectively.

*b) Gradient Boosting Regressor:* This algorithm builds an ensemble of decision trees sequentially, where each tree minimizes the errors of the previous ones. It often outperforms random forests but is more prone to overfitting, requiring precise tuning of parameters.

*c) Extreme Gradient Boosting:* is an advanced machine learning algorithm based on the gradient boosting framework. XGBoost optimizes performance by using regularization techniques to control overfitting and parallelized tree construction for faster training.

For each model, several variations were implemented to enhance performance:

- **Baseline model**: A basic version with default parameters was used as a starting point.

- **Weighting schemes**: Different weighting strategies were tested to assign varying importance to specific age groups.
- **Variance regularization**: Transformations were applied to the target variable to reduce prediction variance.
- **Data balancing**: Under-sampling and over-sampling techniques were employed to mitigate dataset imbalance.
- **Model optimization**: More complex models were tested to better capture problem complexity and identify suitable hyperparameter ranges.

These strategies guided the selection of hyperparameters for the final grid search, as detailed in the following section.

To ensure a robust estimation of the RMSE, both standard K-fold and stratified K-fold cross-validation were used. The 5-fold standard K-fold reduced RMSE variability caused by different random seeds and training-test splits, while stratified K-fold maintained age group distribution, reducing variance in error estimates across different folds.

### C. Hyperparameters tuning

Hyperparameter tuning was performed for the three models to optimize performances. Grid search explored various hyperparameter combinations, but high computational costs limited the range. Therefore, we concentrated on a select subset of key hyperparameters, chosen for their influence on the model. For the *RandomForestRegressor*, *n_estimators* was varied to control the number of decision trees, while *max_depth* and *min_samples_leaf* were adjusted to balance model flexibility and variance reduction. For the *GradientBoostingRegressor*, additional parameters such as *learning_rate* and *subsample* were included to manage the trade-off between convergence speed and overfitting, while *max_features* was explored to capture relevant feature interactions. The *XGBRegressor* utilized advanced regularization terms, such as *reg_lambda* and *reg_alpha*, to penalize over-complex models and improve generalization.

In Tables I II III are the proposed hyperparameters for each model.

TABLE I
HYPERPARAMETER VALUES TESTED FOR RANDOMFORESTREGRESSOR

| Hyperparameter | Values Tested |
|---|---|
| n_estimators | [200, 500, 1000] |
| max_depth | [10, 20, None] |
| min_samples_leaf | [2, 5] |

TABLE II
HYPERPARAMETER VALUES TESTED FOR GRADIENTBOOSTINGREGRESSOR

| Hyperparameter | Values Tested |
|---|---|
| n_estimators | [1000, 1500, 2000, 2500] |
| learning_rate | [0.005, 0.01, 0.02] |
| max_depth | [4, 6] |
| min_samples_leaf | [5, 10] |
| subsample | [0.6, 0.7, 0.8] |
| max_features | [`sqrt`, `log2`, 0.5] |

TABLE III
HYPERPARAMETER VALUES TESTED FOR XGBREGRESSOR

| Hyperparameter | Values Tested |
|---|---|
| n_estimators | [500, 1000, 1500, 2000] |
| learning_rate | [0.005, 0.01, 0.05] |
| max_depth | [3, 6] |
| subsample | [0.8, 1.0] |
| reg_lambda | [0.1, 1, 10] |
| reg_alpha | [0, 0.1, 1] |
| min_child_weight | [3, 5] |

## III. RESULTS

The models were compared using the average RMSE across 5 stratified K-folds, grouping age ranges into 10-year intervals and preserving their distribution between training and validation sets.

- The **baseline model**, a *RandomForestRegressor* with default parameters, was used to evaluate preprocessing strategies, achieving an initial RMSE of 10.6654 without considering ethnicity and 10.2167 with the previously chosen encoding.
- Various **target transformations**, including square root, Box-Cox, and logarithmic, were tested to address increasing variance with larger target values. The logarithmic transformation proved most effective, reducing RMSE to 10.1581.
- A **more complex** *RandomForestRegressor* (*n_estimators = 200*, *max_depth = 10*, *min_samples_leaf = 5*) was trained, but it did not significantly improve performance.
- A detailed analysis revealed that dataset imbalance led to better performance for well-represented age groups and underperformance for older individuals. To mitigate this, **weighting strategies** and sampling techniques were tested, with threshold weights achieving the best results, reducing RMSE to 10.1132.
- Retaining **outliers** proved beneficial, as their removal increased RMSE to 10.2131.
- **Feature extraction** improved RMSE to 9.8555, but the high dimensionality of the dataset increased computational costs and training time.
- We trained on the full dataset of 271 features, then applied **feature selection** to retain the most informative ones, excluding noise and redundant features. Selecting the top 40 features by importance reduced the RMSE to 9.6558. The optimal number of features was determined empirically, balancing dimensionality reduction and prediction performance.
- A final **grid search** on this reduced dataset yielded an RMSE of 9.481 on the test set, evaluated by a submission, with parameters *RandomForestRegressor* (*n_estimators = 1000*, *max_depth = 20*, *min_samples_leaf = 2*).

The same pipeline was applied to *GradientBoostingRegressor* and *XGBRegressor*. Strategies implemented in the Random Forest model, such as logarithmic transformation and weighted training, improved performance for these models as well. We also found that selecting features based on the importance from

each model, instead of relying on the Random Forest's feature importance, resulted in better outcomes.

Preliminary tests with a wide range of hyperparameter configurations helped identify effective settings within the time constraints.

- For *GradientBoostingRegressor* (*n_estimators = 1000*, *learning_rate = 0.01*, *max_depth = 6*, *min_samples_leaf = 5*, *subsample = 0.8*, *random_state = 0*), selecting the top 60 features reduced RMSE to 9.3647,
- For *XGBRegressor* (*n_estimators = 1000*, *learning_rate = 0.01*, *max_depth = 6*, *min_samples_leaf = 5*, *subsample = 0.8*, *random_state = 0*, *reg_lambda = 1*, *reg_alpha = 0.5*), the top 70 features achieved an RMSE of 9.2113.

Final grid searches further optimized both models, as shown in the Table IV. A graphical summary of the various steps and their corresponding RMSE values is shown in Figure 4.

TABLE IV
BEST HYPERPARAMETERS AND RMSE FOR EACH MODEL

| Model | Best Parameters | RMSE |
|---|---|---|
| RandomForest | {'max_depth': 20,'min_samples_leaf': 2, 'n_estimators': 1000} | 9.481 |
| GradientBoosting | {'learning_rate': 0.01, 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 5, 'n_estimators': 2500, 'subsample': 0.8} | 9.132 |
| XGBRegressor | {'learning_rate': 0.01, 'max_depth': 6, 'min_child_weight': 5, 'n_estimators': 2000, 'reg_alpha': 0.1, 'reg_lambda': 0.1, 'subsample': 0.8} | 9.113 |

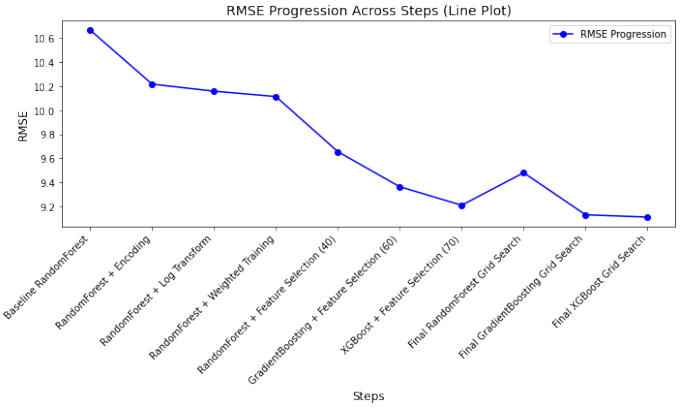


Fig. 4. RMSE progression across steps (Line Plot).

## IV. DISCUSSION

The proposed solution demonstrates a significant improvement compared to the simpler models initially analyzed, confirming the effectiveness of the methodologies employed. However, the complexity of the problem remains evident. The dataset provided includes a limited number of samples, many of which are concentrated within specific age ranges. Expanding the dataset by collecting additional samples, particularly for less-represented age ranges, could substantially enhance model performance.

An analysis of the feature importance in the final model, Figure 5, revealed that *silence_duration* is the most influential variable, likely due to age-related changes in speech rhythm, cognitive processing, and articulation. Other notable features include *num_characters* and the *ethnicity_english* feature, followed by several others with roughly equal importance. This highlights the effectiveness of combining diverse features to enhance the model's predictive capabilities.

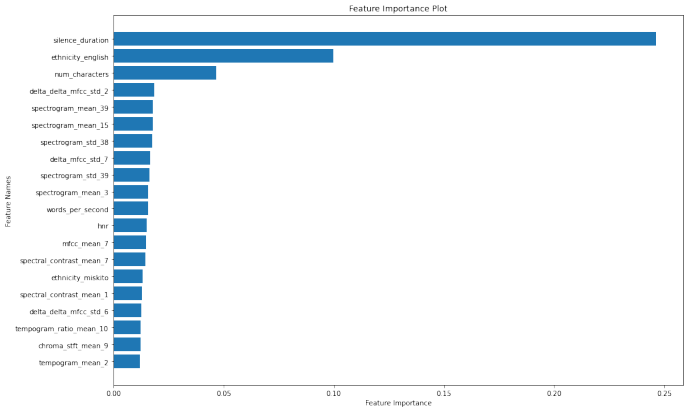


Fig. 5. Features importance of top 20 features

The feature extraction process relied on the `librosa` library to compute statistical descriptors from audio signals. Future work could explore neural networks to automatically derive high-level representations. Furthermore, the current feature set, while informative, represents only a fraction of the potential information contained in the audio files. For example, adding domain-specific measures like prosodic, spectral, or phonetic patterns could enhance the feature space, but would increase dataset dimensionality.

Resource constraints also limited the exploration of a broader range of hyperparameter combinations, transformations, weighting strategies, and sampling techniques. Testing these alternatives could potentially yield better results.

Additionally, while this project focused on tree-based models and gradient boosting, future work could explore neural networks, which are often well-suited for tasks involving complex, high-dimensional data such as audio. Although this approach was not pursued due to the focus of the current study, it represents a promising direction for further investigation.

In conclusion, this work underscores the importance of careful feature selection, dataset balancing, and hyperparameter tuning in addressing the challenge of predicting speaker age from audio data. While the results achieved are promising, they also highlight the need for continued experimentation, feature enrichment, and dataset expansion to further refine the model's performance.

## REFERENCES

[1] "Feature extraction - librosa."
[2] M. Fabien, "Sound feature extraction."
[3] M. Notter, "Age prediction of a speaker's voice," 2022.