Age Estimation Regression Project

Mustafa Kerem KOSE
Politecnico di Torino
Student id: s339018
s339018@studenti.polito.it

***Abstract*—In this report, a regression-based approach is presented to tackle the age estimation problem using features of speech signals. The proposed solution involves extracting of audio-related information from the provided .wav files and the computation of meaningful statistical features, such as standard deviation and other audio-derived metrics. Additional features are analyzed to identify attributes that can enhance the predictive power of the regression model. The approach is focused on feature engineering and model optimization, with performance evaluated using the root mean square error (RMSE) metric.**

I. Problem Overview

The objective of this project is to develop a regression model capable of estimating the age of speakers based on features extracted from their speech signals. The dataset provided for this task contains a total of 3,624 samples, divided into two subsets:

- *Development set*: Contains 2,933 samples, including both the input features and the target age labels, intended for training and validating the models.
- *Evaluation set*: Contains 691 samples, with only the input features provided, used for the final prediction of the model.

By conducting an analysis of the development set, several observations can be made. Firstly, the sampling rate is consistent across all rows with a frequency of 22.5 kHz, indicating uniformity in the audio signal processing. Additionally, the dataset contains three categorical attributes that require careful handling during model training. Notably, while the *tempo* attribute is initially classified as categorical, it is, in fact, represented as a float value within a list, necessitating specific preprocessing steps to ensure an accurate analysis.

An examination of the dataset reveals that there are no missing values, ensuring the completeness of the provided data. Lastly, the final column provides the file path to the corresponding audio recordings for each participant, facilitating access to the raw audio data.

Upon analyzing the correlation matrix, it is evident that the *silence duration* feature exhibits a strong positive correlation with age, with a coefficient of 0.51 as can be seen in Figure 1. This suggests that longer durations of silence are indicative of older speakers, likely due to slower speech patterns or increased pauses. Additionally, the *zcr_mean* (mean zero-crossing rate) emerges as the second most correlated feature with age, further highlighting its importance. These insights emphasize the need to prioritize these features in the modeling process, as they hold significant predictive power for estimating the age of the speakers.

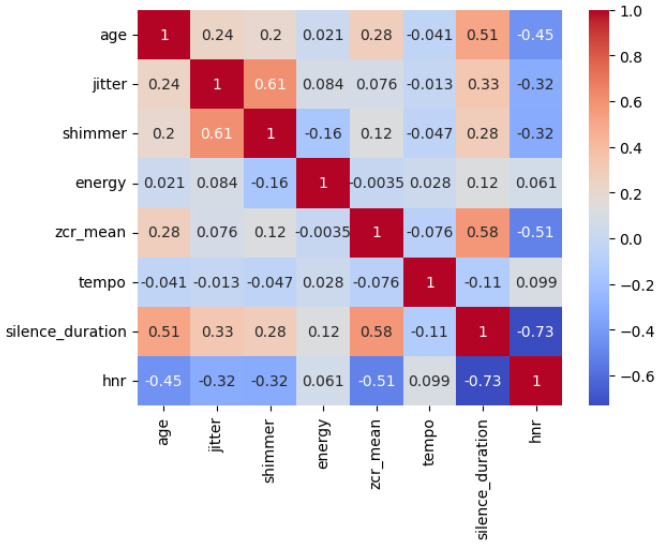

Fig 1: Heatmap of feature correlations

II. Proposed Approach

A. Preprocessing

In this phase, several steps were undertaken to prepare the data for modeling. Categorical features, such as gender and ethnicity, were converted into numerical representations using *one-hot encoding*. Additionally, the *tempo* attribute was addressed, and audio files were processed using the *Librosa* library to extract critical insights, including spectrogram features, MFCCs (Mel-frequency cepstral coefficients), and RMS (Root Mean Square energy) since they are highly used for predicting age [1]. These newly derived features were added to the dataset to enrich the feature set and enhance model performance.

Next, an analysis of the feature distributions revealed potential outliers in the *max pitch* variable. Specifically, the minimum value for maximum pitch was observed to be 935 Hz, which is unusually close to the minimum pitch values, suggesting the presence of anomalies. To address this, outlier removal was performed by excluding the 5% of the distribution, as identified in Figure 2. This step ensured a more robust dataset by mitigating the impact of extreme values.

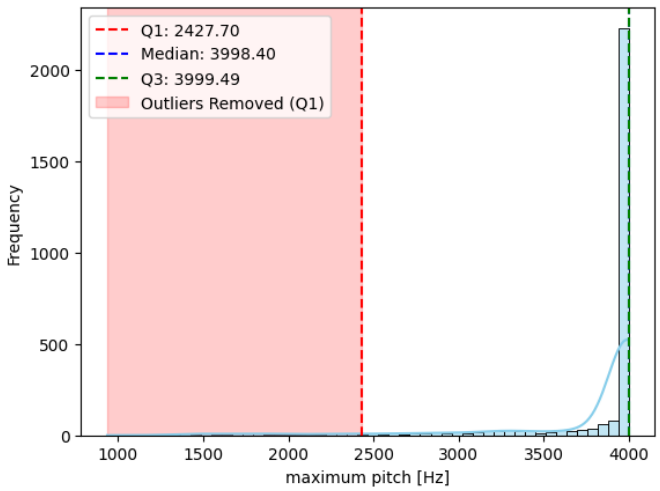

Fig 1: Histogram of the maximum pitch with percentiles

Subsequently, the correlations between features were analyzed, and it was identified that silence duration and jitter might hold significant importance for predicting age. Longer silence durations combined with higher jitter values were observed to potentially indicate slower speech patterns or less stable vocal control, traits often associated with older speakers. To capture this potential relationship, a new composite variable was created by multiplying silence duration and jitter, combining their effects into a single attribute for better predictive power.

In a similar way, another composite variable was created by multiplying silence duration and the number of pauses, resulting in the feature *sil_num*. This new attribute captures both the frequency and duration of pauses, providing a holistic measure of speech fluency and hesitation patterns, which are often correlated with age. Furthermore, the interaction between the Harmonics-to-Noise Ratio (HNR) and the shimmer was examined, and their product, *hnr_shimmer*, was generated as a feature. This attribute combines vocal stability and amplitude irregularity, both of which are indicative of age-related changes in speech. These engineered features aim to capture nonlinear relationships and enhance the model's ability to distinguish age-related vocal characteristics, leveraging insights from speech patterns and voice quality.

To enhance the representation of the shimmer feature, a logarithmic transformation was applied. This step was crucial to address the positively skewed distribution of shimmer, where most values are near zero, with a few outliers significantly larger than the rest. By taking the logarithm, the feature's range was compressed, reducing the impact of extreme values while spreading out smaller ones. This transformation not only stabilized the variance but also brought the feature closer to a normal distribution.

B. Model selection

In this phase, two machine learning models are analyzed: Random Forest Regressor and Gradient Boosting Algorithm (XGBoost). Both models are selected based on their ability to handle structured data, flexibility in capturing non-linear relationships, and robustness for regression tasks like predicting age from speech features.

- Random Forest Regressor: it was chosen for its simplicity, robustness, and interpretability. This ensemble method builds multiple decision trees independently and averages their predictions, making them less sensitive to noise and outliers compared to single decision trees. Random Forest is particularly effective in handling large feature spaces and datasets with irrelevant or redundant features, as it selects random subsets of features at each split. Moreover, its ability to compute feature importance allows for insights into which features contribute most to the model's predictions. Given the mix of acoustic and linguistic features in this dataset, Random Forest was a logical choice [2] for an initial benchmark due to its low risk of overfitting and its strong performance out of the box.

- Gradient Boosting Algorithm (XGBoost): this algorithm was selected for its powerful gradient boosting framework, which builds trees sequentially, focusing on reducing the residual errors of previous trees as shown in Figure 3. XGBoost is known for its efficiency and flexibility, with features like regularization to prevent overfitting, built-in support for missing values, and advanced handling of nonlinear relationships. For this project, where the goal is to predict age - a complex target influenced by multiple interdependent features - XGBoost's ability to optimize through boosting iterations makes it highly suitable [3]. Additionally, XGBoost supports various hyperparameter tuning options, which enable the model to adapt better to the specific patterns in the dataset.

Since both models can benefit from hyperparameter optimization, Grid Search will be utilized to evaluate and fine-tune their performance. This approach ensures that each model is tested across a range of parameter combinations, allowing for the identification of the optimal configuration for this specific task.

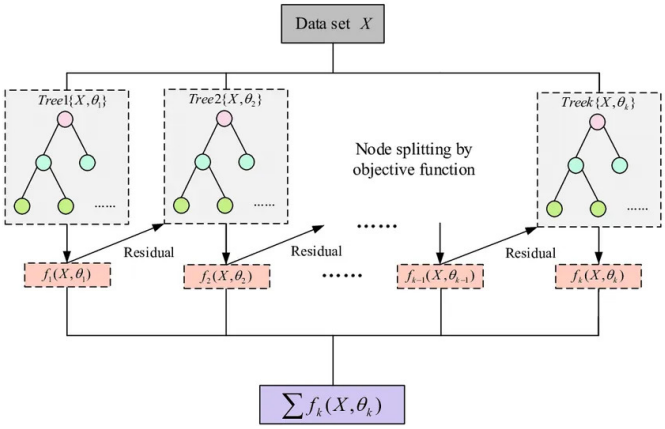

Fig 3: The diagram of the Gradient Boosting process in XGBoost, where sequential trees correct residual errors, and

the final prediction is the sum of all tree outputs optimized by the objective function ($\sum f_k(X, \theta_k)$).

C. Hyperparameters tuning

In this section, hyperparameter tuning was performed using grid search to optimize the Gradient Boosting model. The parameters were selected and shown in Table 1, as they have the most significant impact on the model's ability to generalize and balance bias and variance. In detail, the *learning rate*, also referred to as *eta*, determines how much the model adjusts its predictions at each boosting iteration. Smaller values of the learning rate ensure that each tree contributes only incrementally to the overall prediction, providing greater stability and reducing the likelihood of overfitting. However, smaller learning rates typically require more boosting rounds (i.e., more estimators) to achieve convergence, as each tree's impact is minimized, as shown in Figure 4. In contrast, larger learning rates can speed up the training process but may lead to instability or overfitting if the steps are too aggressive.

On the other hand *number of estimators* determines the number of boosting rounds to balance underfitting and overfitting.

The *max depth* parameter limits tree complexity to avoid overfitting, and the *subsample* introduces randomness by using a fraction of the data for each tree to improve generalization. The chosen ranges allow for a comprehensive exploration of these trade-offs while remaining computationally efficient.

Model	Parameter	Values
Gradient Boosting (XGBoost)	*max_depth*	{3, 4, 5}
	n_estimators	{50, 100, 200}
	subsample	{0.8, 1.0}
	learning rate	{0.01, 0.05, 0.1}
Random Forest	*n_estimators*	{100, 250, 400}
	max_depth	{None, 5, 10, 20}
	max_features	{sqrt, log2}
	criterion	{poisson, squared_error}

TABLE I
HYPERPARAMETER GRID FOR GRADIENT BOOSTING AND RANDOM FOREST

III. RESULTS

The best hyperparameters for the Random Forest Regressor were identified as follows: criterion set to squared_error, maximum depth set to None, maximum features set to log2, and n_estimators set to *400*. With this configuration, the model achieved a Root Mean Squared Error (RMSE) of 10.072 on the test set, demonstrating its ability to capture the underlying patterns in the data moderately.

Allowing unrestricted tree depth (None) enabled the model to fully explore the complexity of the relationships within the dataset, ensuring that no important patterns were left unmodeled. Additionally, the choice of *log2* for max_features introduced randomness compared to *sqrt* by limiting the number of features considered at each split. This not only helped to prevent overfitting, but also ensured that the model maintained generalizability when applied to unseen data.

Increasing the number of estimators to 400 created a strong ensemble of trees. By averaging the outputs of multiple trees, the model reduced variance and minimized the impact of noisy data points. While increasing the number of estimators generally improves performance by reducing variance, the gains tend to diminish beyond a certain point due to computational costs and diminishing returns. Overall, this combination of hyperparameters provided a well-balanced approach, optimizing both accuracy and generalization for the given dataset.

Finally, the use of both squared_error and poisson allows the model to explore trade-offs between sensitivity to large deviations and robustness to outliers. In this task, Poisson's ability to handle non-uniform distributions and provide robust predictions likely led to its superior performance over squared error.

The optimal hyperparameters for the Gradient Boosting model were determined to be learning_rate at 0.05, max_depth at 4, subsample at 0.8 and n_estimators at 200. This configuration was determined through grid search, aiming to balance the model's accuracy, generalization capabilities, and computational efficiency, ensuring reliable performance for the task. The model achieved a Root Mean Squared Error (RMSE) of 9.892, which is the best among the tested models and well-suited for this type of task.

Setting the learning_rate to 0.05 enables the model to improve predictions incrementally, striking a balance between stability and convergence speed. It allows the model to achieve high performance without requiring an excessively large number of trees. Limiting the max_depth to 4 ensures that the individual trees remain shallow, which helps to prevent overfitting while still capturing important relationships in the data.

With n_estimators set to 200, the model benefits from a robust ensemble of trees, correcting errors iteratively without introducing unnecessary computational overhead. Lastly, Together with the learning rate, this ensures that the model avoids both underfitting and overfitting. This combination of parameters achieves a strong balance, allowing the model to effectively learn from the data while maintaining generalization to unseen samples.

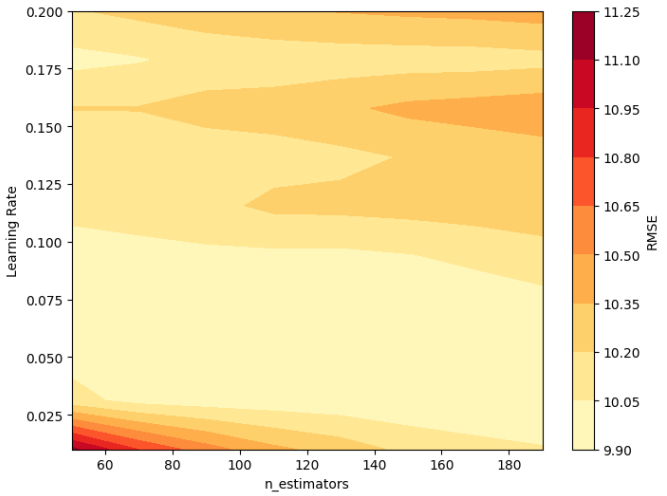

Fig 4: Effect of Learning Rate and n_estimators on RMSE

IV. Discussion

When comparing the results obtained from the Gradient Boosting model to the baseline public value of 11.179 RMSE, it is evident that the model achieved a moderate improvement with an RMSE of 9.892. This indicates that the selected hyperparameters and features were effective in capturing the patterns within the dataset. However, there is still significant room for improvement, as the dataset contains diverse and valuable insights about the audio files that remain underutilized.

One potential avenue for improvement is the extraction of more advanced features from the spectrograms of the audio signals. Spectrograms are visual representations of the frequency spectrum of the signal over time and provide rich information about the temporal and spectral structure of speech. By analyzing the spectrogram, features like harmonic content, phonemes duration, and transient patterns can be extracted, which may further enhance the model's ability to predict the speaker's age.

Another promising approach would be to incorporate Convolutional Neural Networks (CNNs) into the pipeline. CNNs are particularly well-suited for processing spectrograms [4], as they are designed to automatically learn spatial hierarchies of features from images. By treating spectrograms as image-like inputs, CNNs can capture both local and global patterns, such as variations in pitch, energy, and rhythm, which are closely related to age prediction. This method would allow the model to directly leverage the raw audio's spectral information, potentially leading to significant performance improvements.

In conclusion, while the current results represent a step forward, exploring advanced feature extraction techniques and integrating CNNs could unlock the full potential of the dataset, offering substantial improvements in predictive accuracy. These considerations highlight the opportunities for further research and experimentation in this domain.

References

[1] L. k. Durgam and R. k. Jatoth, "Age estimation based on mfcc speech features and machine learning algorithms," in *2022 IEEE International Symposium on Smart Electronic Systems (iSES)*, pp. 398–401, 2022.

[2] M. Berardi, E. Hunter, and S. Ferguson, "Talker age estimation using machine learning," *Proceedings of Meetings on Acoustics*, vol. 30, p. 040014, June 2017. Epub 2018 Oct 25.

[3] K. Narayana and R. Surekha, *Age Prediction by Speech: A Machine Learning Approach Using a Common Speech Dataset*. IEEE, 2024.

[4] F. Wolf-Monheim, "Spectral and rhythm features for audio classification with deep convolutional neural networks," *arXiv preprint arXiv:2410.06927*, 2024.