Speaker's Age estimation

Simone Fassio
Politecnico di Torino
Student id: s342462
s342462@studenti.polito.it

***Abstract*—This paper presents a robust solution to the task of voice-based age estimation using the Voice Age Estimation Dataset, which comprises acoustic, linguistic, and speaker metadata features. The dataset exhibits significant age and ethnicity imbalances, posing challenges in model training and evaluation. To address these, I implemented a comprehensive preprocessing pipeline, including selective one-hot encoding and advanced audio feature extraction. The regression pipeline leverages an ensemble model combining Support Vector Regression (SVR) and Ridge Regression, achieving good results. SVR with an RBF kernel effectively models non-linear relationships, while Ridge Regression captures linear dependencies, with their ensemble showcasing complementary strengths. The inclusion of audio-derived features, such as pitch, energy levels, and silence duration, significantly improved model performance. The analysis highlights the importance of feature engineering and ensemble strategies for addressing dataset imbalances and enhancing predictive accuracy. Residual analysis indicates strong performance in the age range with dense data representation, while underscoring the need for rebalancing techniques to better capture underrepresented age groups. This work demonstrates the effectiveness of combining diverse features and advanced regression techniques for voice-based age prediction.**

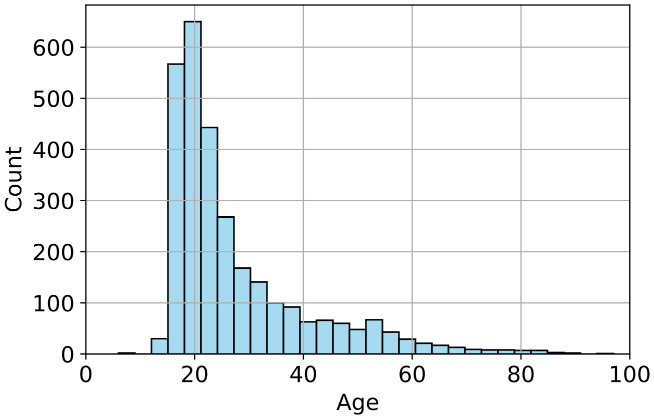

Fig. 1. Distribution of Speakers' Ages in the dataset.

I. Problem overview

The proposed competition addresses a regression problem using *Voice Age Estimation Dataset*, which consists of two distinct sets:

- a development set, including 2,933 samples.
- an evaluation set, including 691 samples.

Each sample corresponds to a spoken sentence, with the speaker's age as the continuous target variable. The dataset encompasses three categories of features:

- Acoustic properties: Pitch, energy levels, jitter, shimmer, and spectral descriptors.
- Linguistic features: Number of words/characters, and speaking rate (tempo) and prosodic elements (pauses, silence duration).
- Speaker metadata: Gender, ethnicity.

Additional audio-derived features were extracted to enhance model precision. The dataset contains no missing values (NaN), eliminating the need for imputation or data removal. The project's objective is to build a regression pipeline leveraging these features for age prediction.

The dataset exhibits a pronounced imbalance in the distribution of speaker ages, as illustrated in Figure 1. The majority of samples are concentrated in the 15–30-year age range, forming a distinct peak that reflects a strong overrepresentation of younger speakers. Beyond this range, the frequency of samples declines sharply, with progressively fewer instances in older age groups. This skewed distribution introduces a bias toward younger speakers. The model is likely to prioritize learning patterns associated with younger speakers due to their dominance in the dataset. As a result, predictions for speakers over 40 years old may suffer from higher errors, as the model encounters fewer training examples for these age groups. The dataset exhibits distinct patterns in the distribution of gender and ethnicity labels. The dataset is well-balanced by gender, with nearly equal representation of male and female speakers. This balance mitigates gender-related biases in age prediction, ensuring that the model does not disproportionately favor one gender over the other. On the other hand, the dataset includes a total of 165 distinct ethnicities, reflecting broad demographic diversity. However, the distribution is highly irregular. The most prevalent ethnicity, Igbo, accounts for approximately one third of all samples, while the second most common, English, represents approximately one sixth. Other notable ethnicities, such as Arabic, Mandarin and French, form smaller but significant subgroups. The remaining ethnicities are sparsely represented, creating a long-tail distribution. This imbalance in ethnicity raises concerns about potential ethnicity-specific bias, as the model may prioritise patterns from over-represented groups (e.g., Igbo) while underperforming for under-represented ones.

II. Proposed approach

A. Preprocessing

To ensure the machine learning model performs optimally for age estimation from speaker voice, a detailed data preprocessing pipeline was implemented. This section outlines the categorical encoding methods, feature extraction techniques, and considerations for feature selection.

a) Categorical Feature Encoding: The dataset contains two categorical features: gender and ethnicity. Proper encoding was applied to transform these features into a format suitable for machine learning algorithms: For the gender feature, one-hot encoding was initially applied. However, to prevent collinearity among the encoded features, only the `gender_male` feature was retained as a binary indicator. Encoding the ethnicity feature presented challenges due to the high cardinality of categories. A direct one-hot encoding approach would have significantly increased the feature space dimensionality, potentially biasing the model towards these features. To address this, only the three most represented ethnicities (Igbo, English, and Arabic) were retained, and one-hot encoding was applied to these categories. This selective encoding effectively limits the increase in dataset cardinality while retaining relevant diversity.

b) Audio Feature Extraction: To analyze speaker voice characteristics, several features were extracted from the audio tracks. These features were computed using the `librosa` Python library and included pitch, energy levels, silence detection, and speech rate.

The pitch was computed using `librosa.pyin`, which estimates the fundamental frequency (F0) of voiced frames. To exclude silent frames, only voiced regions were considered. The following statistical measures were derived:

- Mean, maximum, minimum, and standard deviation of the pitch.
- Jitter, calculated as the relative mean absolute difference between consecutive pitch periods, expressed as a percentage.

Energy features were extracted using the root mean square (RMS) of the audio signal:

$$\text{RMS}_{dB} = 20 \cdot \log_{10}\left(\frac{\text{RMS}}{\max(\text{RMS})}\right) \tag{1}$$

The mean, minimum, maximum (99th percentile), and standard deviation of the RMS in decibel scale were calculated. These features capture variations in loudness and dynamics of the audio.

Two methods were used to detect pauses and silences:

1) Speech frames computed by `librosa` were used to determine the total duration of silence. Silent runs (minimum length of 5 frames to avoid spurious pauses) were analyzed to compute their count, mean duration, and standard deviation.
2) A relative energy threshold (10% of the 99th percentile of energy levels) was applied to identify silent regions. The total duration of these regions was computed.

Both methods yielded complementary features, which were retained for model training to capture different aspects of silence and pauses.

The speech rate was calculated based on the number of detected onsets:

$$\text{Speech Rate} = \frac{\text{Number of Onsets}}{\text{Speech Duration}} \cdot 60 \tag{2}$$

where the speech duration excludes silent regions. This metric provides an estimate of articulation speed.

To ensure robustness all provided and computed features were retained for the number of words that was excluded due to its perfect correlation with the number of characters. The total number of input features was kept at 33, striking a balance between richness and complexity. This approach accounts for the potential noise in the audio signals and ensures a comprehensive representation of speaker characteristics.

B. Model selection

Age estimation from speaker voice is formulated as a regression task. Several models were considered, including random forest, linear regression with ridge and lasso regularization, and support vector regression (SVR). After experimentation, ridge regression and SVR emerged as the most effective approaches. While lasso regression excluded too many features, ridge regression performed well but was slightly outperformed by random forest and SVR.

a) Support Vector Regression (SVR): SVR is a supervised learning algorithm based on support vector machines (SVMs) and is particularly effective for regression tasks [1]. The SVR model attempts to find a function $f(x)$ that has at most ϵ deviation from the actual target values y, while minimizing model complexity. The optimization problem can be formulated as:

$$\min_{w,\xi,\xi^*} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{3}$$

subject to:

$$y_i - \langle w, x_i\rangle - b \leq \epsilon + \xi_i, \tag{4}$$

$$\langle w, x_i\rangle + b - y_i \leq \epsilon + \xi_i^*, \tag{5}$$

$$\xi_i, \xi_i^* \geq 0, \tag{6}$$

where w is the weight vector, ϵ is the margin of tolerance, and C is the regularization parameter. By introducing kernel functions, SVR can model complex, non-linear relationships between the input features and the target variable. In this work, the radial basis function (RBF) kernel was utilized, allowing the model to capture non-linear dependencies effectively [2].

b) Ridge Regression: Ridge regression, is a linear regression method that includes an L_2 penalty term to prevent overfitting by discouraging large coefficients [3]. The optimization problem for ridge regression can be expressed as:

$$\min_{w} ||y - Xw||^2 + \alpha||w||^2 \tag{7}$$

where w is the coefficient vector, α is the regularization parameter that controls the trade-off between model complexity and data fit. Ridge regression is particularly effective for datasets with multicollinearity, as it shrinks the coefficients of less relevant features towards zero without excluding them entirely. This ensures that all features contribute to the prediction, albeit to varying degrees.

c) Model Ensemble: To improve performance, an ensemble approach combining ridge regression and SVR was implemented [4]. The model ensemble leverages the strengths of both methods: ridge regression captures linear relationships effectively, while SVR with the RBF kernel models non-linear dependencies [5]. Instead of averaging the outputs of the two models, a ridge regressor (trained without an intercept) was used to merge their outputs. This approach resulted in a significant reduction of the root mean squared error (RMSE) on the public evaluation set.

d) Data Standardization: Before training, all features were standardized to have zero mean and unit variance. Standardization ensures that all features contribute equally to the model and prevents features with larger scales from dominating the optimization process. This step is particularly crucial for models like SVR, which are sensitive to feature scaling.

The final model architecture is depicted in Figure 2, highlighting the preprocessing pipeline and the ensemble approach.

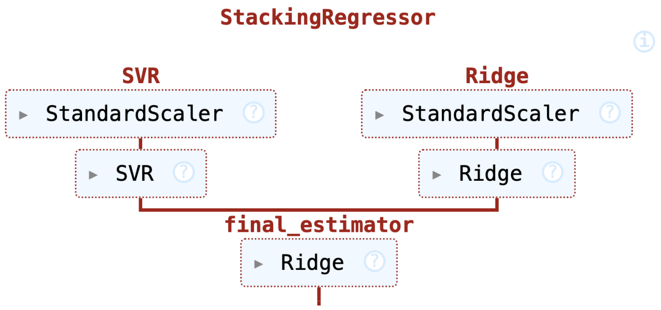

Fig. 2. Model ensemble architecture.

C. Hyperparameters tuning

Hyperparameter tuning was conducted for both SVR and ridge regression to optimize their performance. SVR is particularly sensitive to its hyperparameters: the regularization parameter C and the ϵ parameter that defines the margin of tolerance. Ridge regression, on the other hand, has a single hyperparameter α, which controls the strength of the L_2 regularization.

A grid search was performed over all possible combinations of these hyperparameters to identify the optimal settings [6]. The best results were obtained with $C = 50$, $\epsilon = 1$, and $\alpha = 0.01$. For the final estimator, the default $\alpha = 1$ performed well.

In SVR, a high regularization parameter C allows the model to better fit the training data by increasing the tolerance for points outside the ϵ, but it also increases the risk of overfitting. In ridge regression a low α value reduces the strength of regularization, enabling the model to place greater emphasis on all features. The combination of these settings in the ensemble allows SVR to capture non-linear relationships with high precision, while ridge regression models the linear aspects effectively.

III. Results

The model demonstrated robust performance, achieving an RMSE of 9.35 on the public evaluation dataset. A detailed comparison of the RMSE results is provided in Table I, which includes: Cross-validation RMSE on the development set, and RMSE on the public evaluation set. Moreover, for comparison, the results obtained by the Random Forest Regressor using only the features included in the original dataset are also provided. This baseline serves as a reference to evaluate the effectiveness of the additional audio-derived features and the ensemble approach.

TABLE I
Comparison of RMSE Across Models

Model	CV RMSE	Test RMSE
Random Forest (Baseline)	10.70	10.21
Ridge Regression	10.68	9.79
SVR	10.08	9.51
Ensemble (SVR + Ridge)	**9.91**	**9.35**

The results clearly demonstrate the synergy between SVR and ridge regression. By combining the strengths of both methods, the ensemble approach significantly reduced RMSE without introducing excessive computational complexity. Notably, the ridge regression model is efficient to fit and predict, while the SVR model effectively captures non-linear patterns.

Additionally, the coefficients of the final estimator (0.75 for SVR and 0.28 for ridge regression) highlight that SVR, which individually performed better, contributes more to the final prediction. However, the contribution of ridge regression remains essential and non-negligible.

IV. Discussion

The ensemble model successfully leveraged the complementary strengths of SVR and ridge regression, achieving an RMSE of 9.35 on the public dataset. The inclusion of audio-extracted features, such as duration, silence duration (via both methods), and energy in the decibel scale, significantly enhanced the model's predictive power compared to using only the original features (RMSE of 9.8).

A brief analysis of the ridge regressor coefficients highlights the importance of specific features in age estimation. As shown in Table II, the top 10 coefficients reveal that speech duration, number of characters, and silence detection play a critical role in capturing vocal characteristics associated with age. Notably, features prefixed with an underscore (e.g., _duration, _std_pause_length) were extracted from the audio tracks, emphasizing the value of audio-derived features in improving model performance. These features, combined with

TABLE II
LARGEST 10 FEATURE COEFFICIENTS FROM RIDGE REGRESSION

Feature	**Coefficient**
_duration	16.62
zcr_mean	2.46
ethnicity_english	2.30
_std_pause_length	1.75
_energy_db_std	-1.62
_energy_db	-2.21
silence_duration	-2.79
_silence_duration	-2.93
spectral_centroid_mean	-3.08
num_characters	-7.27

linguistic and metadata attributes, contribute to the robustness and accuracy of the age estimation model.

The residual plot of the ensemble model in Figure 3 shows the difference between actual and predicted age values (residuals) against the predicted age [7]. However, a funnel-shaped pattern is observed, suggesting increasing variance in residuals for higher predicted ages. This indicates that the model performs best in the age range where it has the most data to learn from. As a consequence, the model struggles to generalise effectively in less-represented age ranges, due to the limited diversity of training samples in those regions. Overall, the residual plot confirms good model performance, as most residuals remain within a reasonable range without systematic bias. The wider range of residuals for older predicted ages highlights a potential need for data augmentation or rebalancing strategies to improve model performance on under-represented age groups.

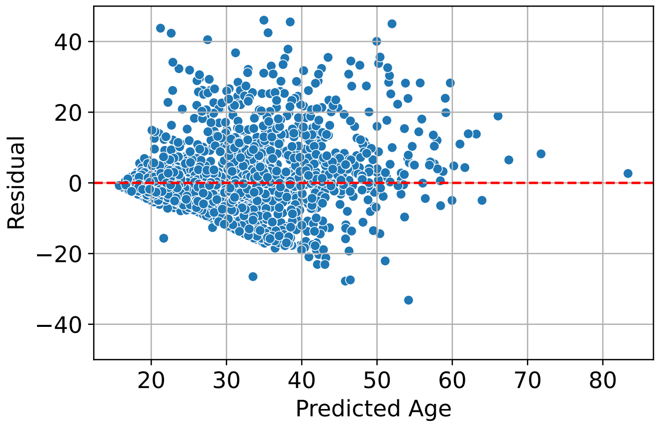

Fig. 3. Residual plot of the model ensemble.

Additional feature extraction techniques can be implemented in addition with alternative ensemble strategies or different models to further enhance performance. However, the result obtained is more than satisfactory for the purposes of the analysis.

REFERENCES

[1] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Statistics and computing*, vol. 14, pp. 199–222, 2004.

[2] M. Huckvale and A. Webb, “A comparison of human and machine estimation of speaker age,” in *Statistical Language and Speech Processing: Third International Conference, SLSP 2015, Budapest, Hungary, November 24-26, 2015, Proceedings 3*, pp. 111–122, Springer, 2015.

[3] D. W. Marquardt and R. D. Snee, “Ridge regression in practice,” *The American Statistician*, vol. 29, no. 1, pp. 3–20, 1975.

[4] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, “Ensemble approaches for regression: A survey,” *Acm computing surveys (csur)*, vol. 45, no. 1, pp. 1–40, 2012.

[5] Y. Frayman, B. F. Rolfe, and G. I. Webb, “Solving regression problems using competitive ensemble models,” in *Australian Joint Conference on Artificial Intelligence*, pp. 511–522, Springer, 2002.

[6] K. Smets, B. Verdonk, and E. M. Jordaan, “Evaluation of performance measures for svr hyperparameter selection,” in *2007 International Joint Conference on Neural Networks*, pp. 637–642, IEEE, 2007.

[7] J. Belloto and T. Sokolovski, “Residual analysis in regression,” *American Journal of Pharmaceutical Education*, vol. 49, no. 3, pp. 295–303, 1985.