Age Prediction Using Regression Model

- Submitted to: Sir Ehsaan Ali
- Submitted on: 10/5/2024

Table of Contents:

Project Overview

» Objectives

The project aims to predict the age of speakers from audio recordings, leveraging the Common Voice dataset. This predictive task holds significance in various applications, including voice-based personalization, targeted advertising, and forensic voice analysis. The dataset comprises audio recordings with associated labels for speaker age, encompassing a diverse range of speakers across different ages and genders. To achieve the goal of age prediction, the project entails several key steps. Firstly, feature extraction techniques are employed to extract relevant acoustic features from the audio recordings, such as pitch, formant frequencies, intensity, duration, and spectral features. These features provide valuable insights into the characteristics of the voice signal that may correlate with the speaker's age. Subsequently, data preprocessing steps are performed to handle missing values and normalize or scale the extracted features. The dataset is already divided into training and testing sets for model evaluation. Next, regression algorithms suitable for predicting continuous-valued attributes, such as age, are trained using the extracted acoustic features as input and the corresponding age labels as target variables. Regression algorithms such as linear regression, support vector regression, decision tree regression, or neural network regression can be utilized for this purpose. Finally, the trained regression model is evaluated on the testing set using appropriate evaluation metrics such as mean squared error (MSE), mean absolute error (MAE), or R-squared (R^2) coefficient of determination. The performance of the model in accurately predicting speaker age across different age groups is assessed, providing insights into the effectiveness of the predictive model.

Problem Statement

The objective of this project is to predict the age of speakers based on acoustic features extracted from audio recordings. This prediction can have various practical applications such as voice-based personalization, targeted advertising, and forensic voice analysis. The dataset provided consists of audio recordings accompanied by labeled speaker ages, encompassing a diverse range of speakers across different ages and genders.

» Requirements

Data Analysis and Signal Processing Skills: Proficiency in data analysis techniques, particularly in extracting relevant acoustic features from audio signals using Librosa.

Machine Learning Knowledge: Understanding of regression algorithms and model evaluation metrics (e.g., MSE, MAE, R-squared) for training and evaluating predictive models.

Programming Proficiency in Python: Strong programming skills in Python for implementing data preprocessing, model training, and evaluation scripts using the specified libraries.

» Input and Output Specifications

The input for this project consists of audio recordings of speakers.

Each audio recording is expected to be in a digital format compatible with the selected audio processing libraries (e.g., Librosa).

The dataset includes a diverse range of speakers across different ages and genders, ensuring variability in voice characteristics.

The primary output of the project is the predicted age of the speakers based on the extracted acoustic features from the audio recordings.

The predicted ages are continuous-valued attributes, indicating the estimated age of each speaker.

Methodology

Feature Extraction:

Acoustic features are extracted from the audio recordings using signal processing techniques.

Key features for speaker age prediction include:

» Pitch (Fundamental Frequency - F0): Reflects the perceived tone of the voice.
» Formant Frequencies: Provide information about vocal tract size and shape, correlated with vowel sounds.
» Intensity: Represents the energy or loudness of the voice.
» Duration: Indicates the length of speech segments.
» Spectral Features: Frequency content of the voice signal, derived from Fourier transform or spectrogram.

These extracted features are organized into data frames for subsequent analysis and visualization to identify trends.

Data Preprocessing:

Missing value handling (imputation) is performed to address any null values in the dataset, using techniques such as mean/median/mode imputation, forward/backward fill, or dropping null values.

Feature scaling or normalization is applied to ensure uniformity in feature magnitudes and scales, which is crucial for effective model training.

Model Selection and Training:

A regression algorithm is chosen for predicting continuous-valued attributes like age.

Commonly considered regression algorithms include:

» Linear Regression
» Support Vector Regression
» Decision Tree Regression
» Neural Network Regression

The selected regression model is trained using the extracted acoustic features as input and the corresponding age labels as target variables.

Model Evaluation:

The trained regression model is evaluated on a separate testing set using appropriate evaluation metrics such as:

» Mean Squared Error (MSE)
» Mean Absolute Error (MAE)
» R-squared (R^2) Coefficient of Determination

Understanding the Dataset

Understanding Audio Data Representation:

Waveform Representation:
The waveform represents the audio signal in the time-domain, depicting changes in air pressure over time. By plotting the waveform, we visualize the temporal oscillations of the signal. It provides information about amplitude, frequency, and duration of sound events.

Temporal Resolution:
Examining smaller segments of the waveform allows us to observe fine-grained details of the audio signal. A small window of the waveform showcases rapid oscillations, indicating the frequency and intensity variations within a short duration.

Fourier Transform and Frequency Domain:
The Fourier transform converts the audio signal from the time-domain to the frequency domain, revealing the spectral composition of the signal. It decomposes the signal into its constituent frequencies, providing insights into the frequency distribution.

Frequency Representation:

Plotting the Fourier transform reveals the distribution of frequencies present in the audio signal. By analyzing the frequency spectrum, we can identify dominant frequency components and their respective magnitudes.

Spectrogram Analysis:
Spectrograms offer a comprehensive visualization of audio data, combining time and frequency information. Using techniques like Short-time Fourier Transform (STFT), Mel spectrogram, and Mel-frequency cepstral coefficients (MFCCs), we can analyze audio signals in both the time and frequency domains.

STFT Spectrogram:
STFT spectrograms provide a detailed representation of the audio signal, with time on the x-axis and frequency on the y-axis. The color intensity indicates the magnitude of each frequency component at different time intervals.

Mel Spectrogram:
Mel spectrograms use the Mel scale to represent frequencies, mimicking human perception of pitch. This scale enhances the discriminative power of the spectrogram, allowing for better understanding of audio features.

MFCCs:
MFCCs provide a compact representation of audio signals by extracting a small number of features. These coefficients capture essential characteristics of the audio, making them suitable for various machine learning applications.

Exploratory Data Analysis (EDA) on Audio Dataset:

Investigation of Features Distribution:

» *Target Features:* Initial exploration involves examining the class distributions of potential target variables, including age and gender. Bar plots are utilized to visualize the distribution of age groups and gender categories within the dataset.

» *Extracted Features:* Further analysis focuses on the distribution of extracted acoustic features such as fundamental frequency (f0), tempo, and duration. Histograms are employed to visualize the value distributions of these features, highlighting potential skewness in the data.

Feature Correlation:
Correlation analysis is performed to assess relationships between features within the dataset. A heatmap is generated to visualize the correlation matrix, revealing potential associations between acoustic features and demographic variables such as age and gender.

Spectrogram Features:

» *Spectrogram Extraction:* Mel spectrograms are computed for audio recordings to capture spectral characteristics. Spectrograms are resized to a uniform length of 3 seconds to facilitate comparative analysis.

» *EDA on Spectrogram Features:* Average mel spectrograms are visualized separately for male and female speakers, highlighting differences in spectral content. Gender-based

differences in voice characteristics, such as lower frequencies in male speakers, are identified through comparative analysis of spectrograms.

Feature Extraction

Onset Detection:
Onset detection is a fundamental technique for identifying the beginning of sound events within an audio signal. Librosa's onset_detect() function allows us to detect onsets, which are indicative of the start of spoken words or sound events. By plotting onsets alongside the waveform, we can visualize the temporal distribution of these events within the audio signal.

Length of an Audio Recording:
The duration of an audio recording is a crucial metric that influences the number of spoken words or sound events contained within it. By computing the length of the recording and the average speed of speech (words per second), we gain insights into the pacing and density of spoken content.

Tempo:
Tempo refers to the speed or pace of speech within an audio signal. By analyzing the tempo, we can quantify the rhythmic pattern of speech and identify variations in speaking speed. Librosa's beat.tempo() function allows us to estimate the tempo in beats per minute (bpm), providing valuable information about the cadence of speech.

Fundamental Frequency:
The fundamental frequency (f0) is the lowest frequency at which a periodic sound appears, also known as pitch. By extracting the fundamental frequency from an audio signal, we can analyze the pitch contour of speech and identify variations in vocal tone. Librosa's pyin() function provides a probabilistic approach to estimate the fundamental frequency.

Algorithm Implementation

Linear Regression Model:

Linear regression is a simple yet effective method for predicting a continuous target variable (in this case, age) based on one or more independent variables (features extracted from audio signals). The choice of linear regression is motivated by its interpretability, ease of implementation, and ability to capture linear relationships between features and the target variable. Linear regression provides coefficients that indicate the strength and direction of the relationship between each feature and the target variable, allowing for meaningful interpretation of the model.

The Ordinary Least Squares (OLS) method for estimating the coefficients in linear regression.

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} = (X'X)^{-1}X'Y$$

X represents the design matrix, which contains the predictor variables (features) of dataset. Each row corresponds to a data point, and each column corresponds to a feature. It has dimensions $n \times (p+1)$, where n is the number of observations and p is the number of features (excluding the intercept).

Y is the response vector, which contains the target variable you want to predict. It has dimensions $n \times 1$.

The formula a = $(X'\ X)^{-1}\ X'\ Y$ calculates the coefficient vector b for the linear regression model. This vector contains the estimated coefficients for each feature, including the intercept.

- $y = a_0 + a_1x_1 + a_2x_2$

Now, the regression equation consists of the elements of a calculated from the previous equation. Here, x1 and x2 represent the independent variables and y represents the dependent variable. So, if you know the values of the x1 and x2 you can predict the value of y.

Result and Evaluation

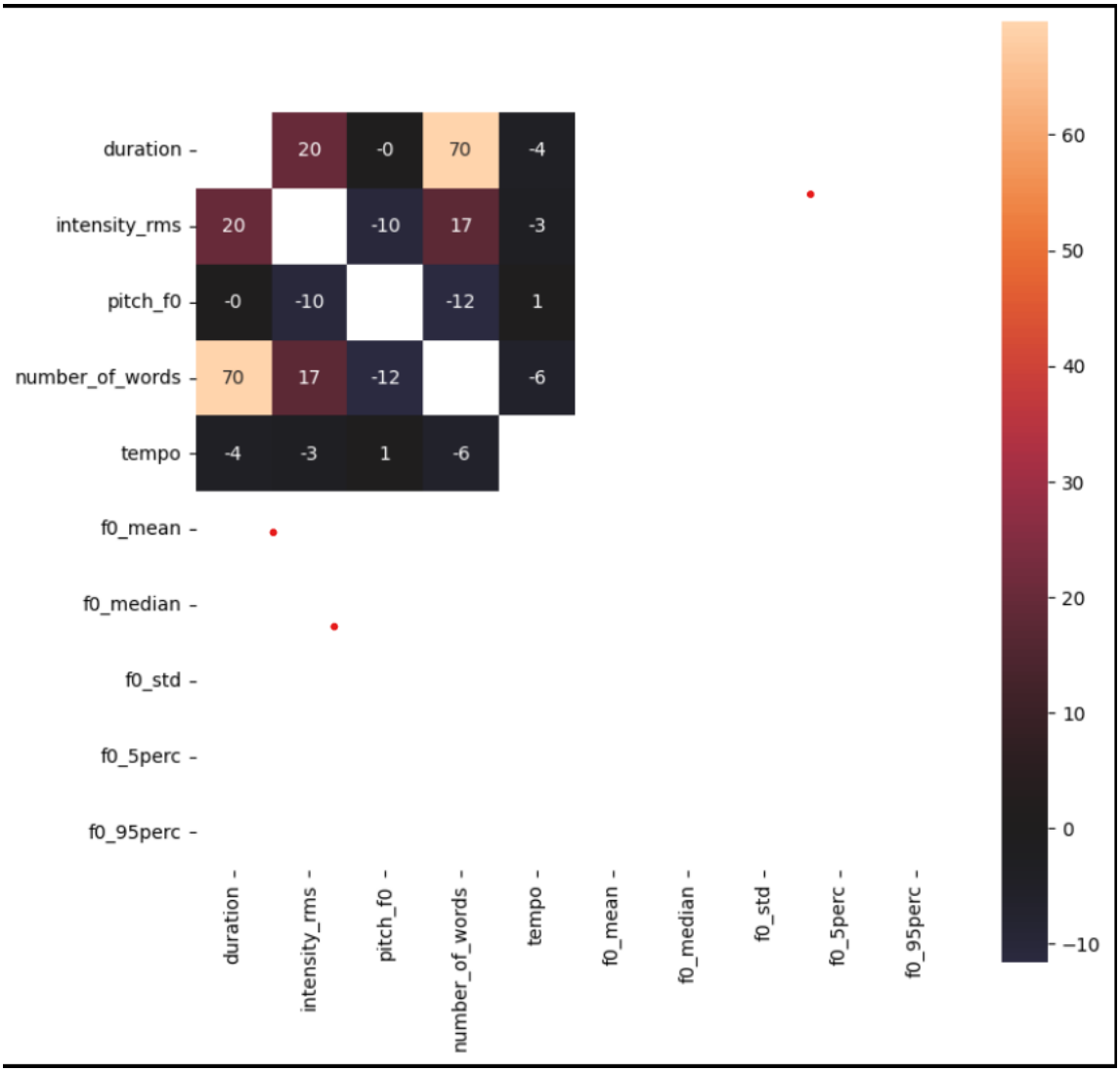
After obtaining the coefficient vector a, you can use it to make predictions on new data points by multiplying the predictor variables with the corresponding coefficients and summing them up. Model performance can be evaluated using various metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), or $R2$ (coefficient of determination).

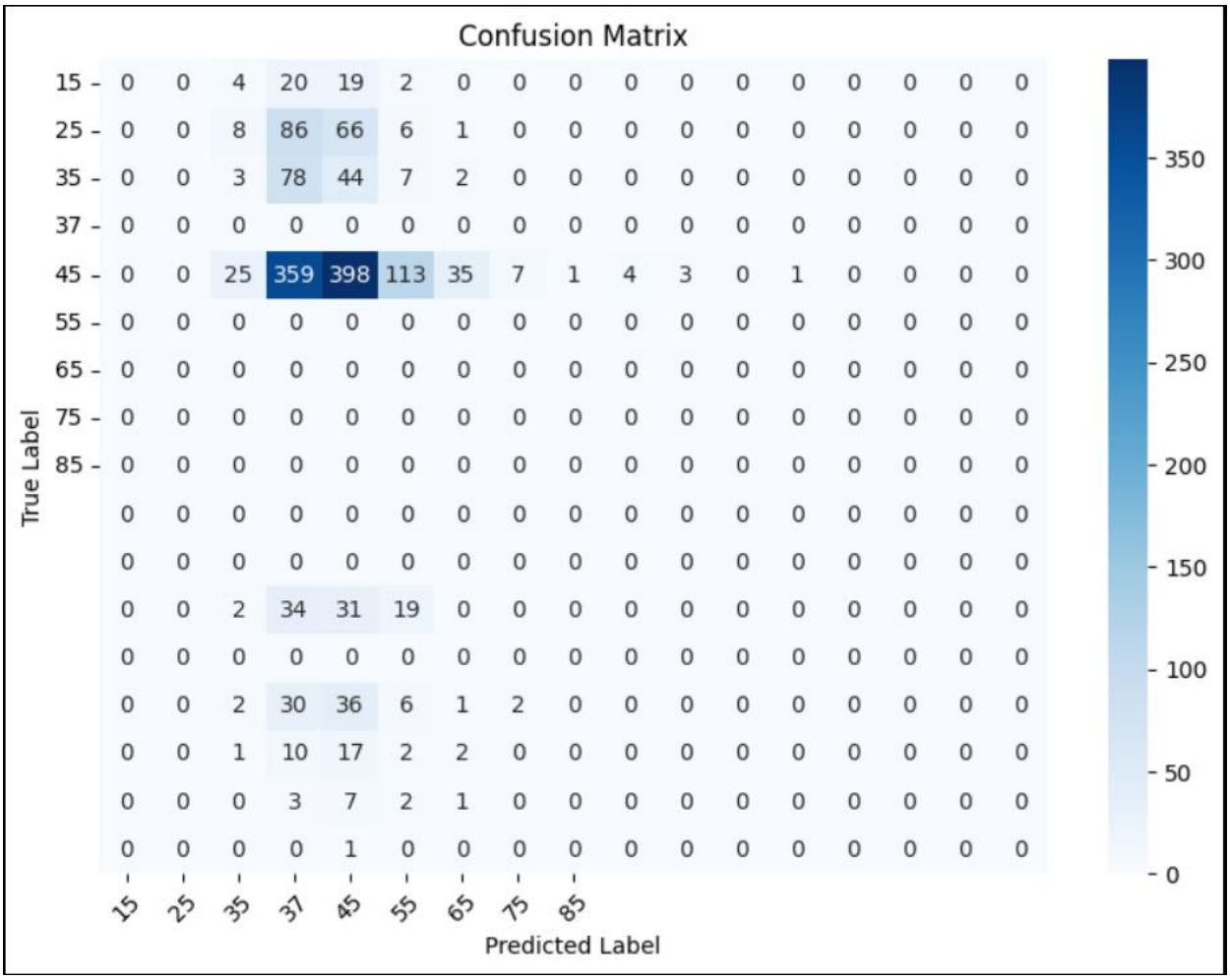
Assumptions:
The OLS method assumes that the errors (residuals) are normally distributed with constant variance (homoscedasticity) and are independent of each other. It also assumes that the relationship between the predictors and the response variable is linear.

Loss Function:
To generate a loss function for the provided manual linear regression model, mean squared error (MSE) loss evaluation metric was used. The goal of the loss function is to quantify the error between the predicted values and the actual target values.

Report: Age Prediction Using Regression Model

Visualization:





The mean squared error is a common loss function used in linear regression and is defined as the average squared difference between the predicted values and the true values.

References

- Notter, M. (2024, January 5). Age prediction of a speaker's voice - EPFL Extension School - Medium. *Medium*. https://medium.com/epfl-extension-school/age-prediction-of-a-speakers-voice-ae9173ceb322

- *ResearchGate*. (n.d.). ResearchGate. https://www.researchgate.net/publication/371445307_Voice-based_Gender_and_Age_Recognition_System/link/64bfc395b9ed6874a54496c4/download?

_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19