Data Science and Engineering (DSE) Record, Volume 5, Issue 1

Customer Age Prediction in Telesales Through Voice Data Analysis Using Advanced Deep Learning Techniques

Supanut Thiengburanatam[1], Waranya Mahanan *[2], Sumalee Sangamuang[2], and Prompong Sungunnasil[2]

[1] Data Science Consortium, Chiang Mai University, Thailand
supanut_thi@cmu.ac.th
[2] College of Arts, Media and Technology, Chiang Mai University, Thailand
{waranya.m, sumalee.sa, prompong.sugunnasil }@cmu.ac.th

Abstract. Through voice data analysis, this research presents a novel deep-learning approach to predict customer age ranges in telesales. Utilizing the rich dataset from Mozilla's 'Common Voice' project, the study focuses on extracting vocal features using Librosa and building a model with TensorFlow and Keras. Based on LSTM layers, the model is trained to recognize patterns correlating vocal attributes with customer age. The research demonstrates the model's efficiency through various performance metrics, aiming to enhance customer service personalization in telesales. This research presents a novel deep-learning approach to predict customer age ranges in telesales, utilizing the rich dataset from Mozilla's 'Common Voice' project. By extracting vocal features using Librosa and building a model with TensorFlow and Keras, this study shows that LSTM layers can effectively recognize vocal attributes correlating with customer age. The results, demonstrating a validation accuracy of 54.25%, underline the potential for enhancing personalized customer service through voice data analytics. This methodological innovation represents a significant step toward practical applications in customer relationship management with advanced machine learning techniques.

Keywords: Deep Learning, LSTM, Voice Data Analysis, Age Prediction, Telesales Personalization

1 Introduction

Telesales, a critical interface in customer relations, stands to benefit significantly from advancements [10] in personalized service delivery. Traditional demographic gathering methods often lack precision and can be intrusive. In response, this study employs advanced deep learning techniques for non-intrusive and accurate age prediction from voice data, a method crucial for enhancing customer interaction in telesales. The research focuses on utilizing RNNs and LSTMs for

* corresponding author

natural language processing, tapping into the potential of these technologies to process complex vocal features [4].

This research explores the potential of deep learning to extract meaningful insights from voice data, a rich source of both explicit and implicit information. By analyzing key acoustic features like tone, pitch, and speech rate, the model aims to predict the speaker's age group accurately. The practical application of this research is poised to revolutionize the telesales industry. Equipped with accurate age-range predictions, telesales representatives can tailor their communication strategies, resulting in more effective marketing and improved customer service. This approach is expected to enhance customer satisfaction and drive sales conversions, demonstrating the transformative power of deep learning in customer-centric industries [7, 15]. The findings from this research underscore the effectiveness of our deep learning approach, particularly leveraging LSTM networks, in accurately predicting customer age ranges from voice data. We achieved a validation accuracy of 54.25%, illustrating the model's capability to capture and utilize vocal attributes for age prediction. These results advocate the potential of voice data analytics in enhancing telesales personalization, offering a promising avenue for improving customer service strategies.

2 Literature Review

The application of deep learning techniques in voice data analysis for customer age prediction within the telesales domain represents an innovative approach that builds upon existing research and leverages the strengths of advanced neural network architectures. The overview of relevant studies that inform the research objectives and methodologies outlined in the research is followed.

2.1 Deep Learning in Speech Recognition

Deep neural networks have demonstrated their effectiveness in acoustic modeling for speech recognition, as highlighted by [7]. The shared views of research groups in this field emphasize the potential of deep learning to capture the complexities of speech patterns. This foundational work underpins the application of deep learning models, including RNNs and LSTMs, in voice data analysis for age prediction.

2.2 Voice Data Analysis and Feature Extraction

Voice data analysis involves the extraction of relevant acoustic features that can provide insights into various aspects of speech. The proposed use of features such as Mel-Frequency Cepstral Coefficients (MFCC), chroma frequencies, tonnetz, and mel-scaled spectrogram align with established methods for acoustic feature extraction. Chroma feature analysis, as described [6], is particularly relevant for capturing harmonic changes in musical audio, which can be extended to voice analysis.

2.3 Audio Analysis

The advancements in audio analysis, especially in speech and voice recognition, have progressed notably over recent years. The shift from basic waveform analysis to more sophisticated methods, such as Fourier Transformations in the 1990s marked the early integration of machine learning in audio signal processing. The advent of deep learning in the 21st century, particularly through Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have significantly revolutionized audio analysis. These developments have led to major breakthroughs in areas like speech recognition and voice biometrics, enhancing the capability to detect complex patterns in audio data. This progress in deep learning models has substantially impacted various audio processing tasks [13]. In the development market for audio feature extraction, several notable libraries include:

Essentia Developed by the Music Technology Group at Universitat Pompeu Fabra, Essentia is a comprehensive library for audio analysis and music information retrieval. It provides an extensive set of algorithms for feature extraction, making it suitable for both research and industry applications. The Essentia is known for its efficiency in processing and extracting a variety of audio features, including those relevant to music and speech analysis.

pyAudioAnalysis This library offers a wide range of audio analysis capabilities, including feature extraction, classification, and segmentation. It's versatile and well-suited for tasks that involve machine learning in audio data processing. The library provides a comprehensive set of tools for extracting features from audio signals, which can be used for various applications like speech recognition and music analysis.

Librosa a Python library for audio and music analysis, plays a crucial role in the proposed methodology for preprocessing voice data. [12] introduced librosa and its application in audio and music signal analysis in Python, making it a valuable tool for extracting and processing acoustic features from voice data.

For this research, the Python library Librosa is chosen for audio analysis due to several key reasons. Firstly, Librosa is specially tailored for music and audio analysis, making it highly effective for complex audio data processing. It boasts a comprehensive set of features crucial for speech analysis, including MFCCs, chroma features, and spectral contrast. Additionally, its user-friendly interface and compatibility with other Python libraries like TensorFlow and Scikit- learn ensure ease of use and integration. Librosa also benefits from an active community and extensive documentation, which are essential for research and development. Finally, its suitability for nuanced speech analysis aligns perfectly with the project's focus on customer age prediction in telesales. This combination of specialized features and user accessibility makes Librosa the preferred choice for this project.

In the domain of telesales, the predictive analysis of customer demographics through voice data has been significantly propelled by the incorporation of advanced deep learning methodologies. The study "Age group classification and gender recognition from speech" by [14] exemplifies the potent capabilities of neural networks in categorizing age and discerning gender, indicating substantial progress in tailored customer interactions. Furthermore, the research "Voice-based Gender and Age Recognition System" by [9] reveals the potential of temporal convolutional neural networks in the identification of gender and age, which augments the analytical depth of voice data interpretation. Concluding these notable advancements [11] contribute critical insights into the precision of neural networks for demographic predictions, cementing the value of these technologies in the enhancement of demographic prediction through voice analysis.

3 Metodology

This research has used voice data analysis to predict customer age ranges in telesales. The methodology is grounded in advanced deep learning techniques, employing a series of steps from data collection, preprocessing, and feature extraction to model development, training, and evaluation. The workflow is designed to ensure the research's integrity, reproducibility, and ethical compliance, particularly in handling voice data as shown in Fig. 1.

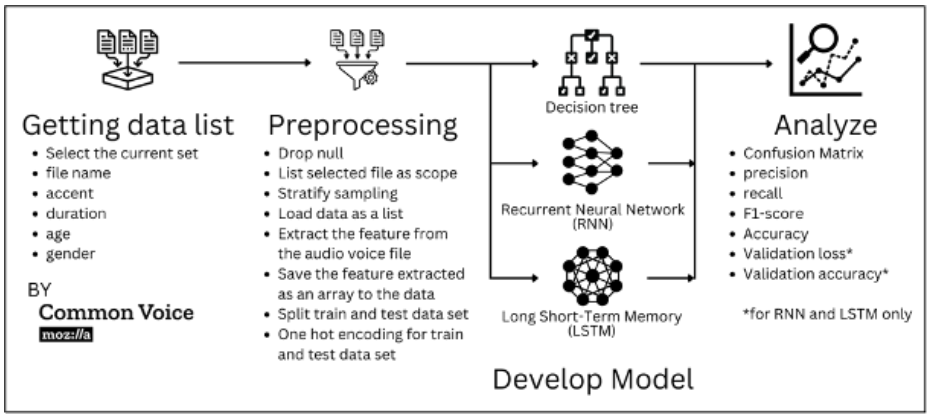


Fig. 1: Voice Data Model Development Workflow

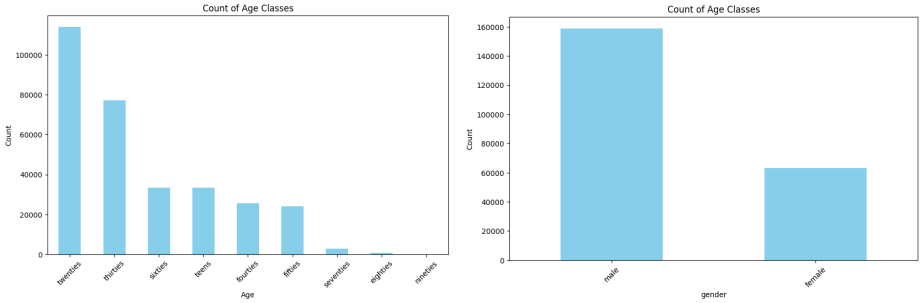
3.1 Data Collection and Privacy Considerations

The foundation of any project pivots crucially on the dataset it employs. This research is predicated on the objective of training a model to discern multiple age brackets, necessitating a substantial corpus of vocal samples. To this end,

the dataset was curated from an official open-source initiative that amasses voice recordings from registered volunteers, subject to preliminary public validation. This approach is congruent with the project's aims and is particularly suited to the constraints imposed by the research timeline. Nonetheless, to safeguard personal information, the identity of each voice sample will be anonymized, with only the national accent and age utilized for research purposes. This methodology ensures both the research's integrity and the participants' privacy.

3.2 Data Preprocessing and Feature Extraction

The initial dataset, comprising over two million vocal samples, underwent extensive preprocessing to refine and tailor it for the objectives of this study. Initially, the data exploration phase revealed nine distinct age categories. To align with the target demographic of the telesales sector and minimize potential bias stemming from underrepresentation from Fig. 2. This decision ensures that the model focuses on the most relevant age ranges.



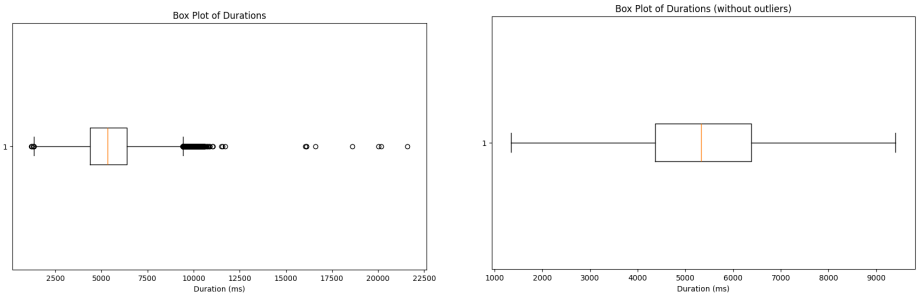
(a) Age class in raw data (b) Gender class in raw date

Fig. 2: The dataset summary

The initial exploratory analysis delineated nine age catergories within the dataset. Aligning with the telesales demographic, categories 'seventies' , 'eighties', and 'nineties' were omitted due to sparse data representation.

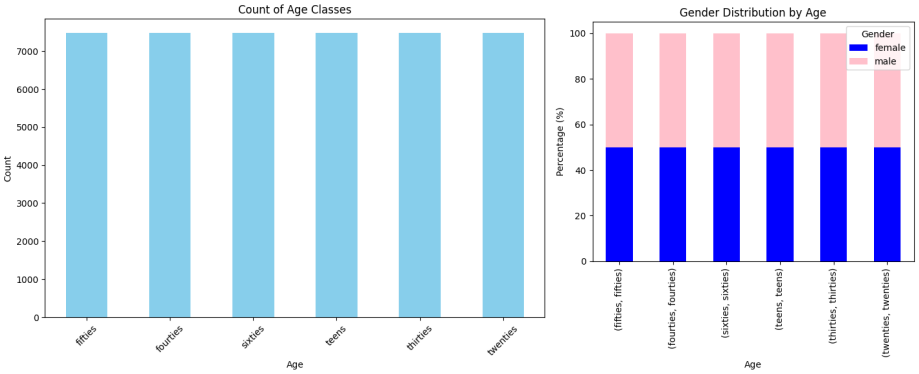
Durational analysis identified outliers, prompting their exclusion based on the Interquartile Range (IQR), retaining samples surpassing the Lower Quartile threshold is 4.4 second. This procedure, underpinned by the feature extraction requirements, culminated in a refined dataset conducive to robust model training, visualized in Fig. 3. Moreover, gender and age stratification in sampling was applied to avert overfitting, thus optimizing the age prediction's robustness, depicted in Fig. 4.

Preprocessing of the audio files was conducted using Librosa, a Python library for audio and music analysis. The audio data was segmented into fixed lengths to standardize the input, and several features were extracted to capture the characteristics of the voice data related to age prediction:



(a) Box plot Voice duration (b) Box plot without Outlier

Fig. 3: Voice Duration data distribution



(a) Age Balance (b) Gender Balance

Fig. 4: Pre-possessed for Data Balance

MFCC (Mel-Frequency Cepstral Coefficients): These coefficients provide a representation of the power spectrum of a sound.

Chroma Frequencies: This feature represents the energy distribution of pitches [1] within an audio signal.

Tonnetz (Tonal Centroid Features): These capture the tonal aspects of sound, related to the harmonic content.

Mel-Scaled Spectrogram: This spectrogram is used to represent the spectral density of sound over time.

The features were then compiled into a single array to form the final dataset for model training.

3.3 Model Development and Training

The model was developed using TensorFlow [3] and Keras. A Sequential model was chosen, incorporating LSTM layers to capture the temporal structure of

the audio features effectively. The model architecture also included Dense layers with ReLU activation functions for nonlinear processing and a final layer with a softmax activation function to predict the probability of each age category. The RNN layer is typically same as the LSTM layer and the final layer with a SoftMax activation function to predict the probability of each age category.

The training involved multiple epochs with batch processing, minimizing binary cross-entropy loss using the Adam [8] optimizer. The model's performance was monitored through accuracy metrics and validated against a hold-out dataset. The training process demonstrated convergence, as visualized by decreasing loss and increasing accuracy over epochs.

Performance metrics such as classification metrics like accuracy, precision, recall and F1-score were calculated to evaluate the model's prediction capabilities. Additionally, a confusion matrix was generated to analyze the model's classification performance across different age groups.

The LSTM and RNN model's learning dynamics are visualized through training and validation loss and accuracy plots, demonstrating the model's ability to learn from the data over time without overfitting, as indicated by the close alignment of training and validation metrics.

4 Results

Implementing the deep learning model for predicting customer age ranges in the telesales industry yielded promising results. The LSTM and RNN network, complemented with features derived from Librosa, was trained on a diverse dataset sourced from Mozilla's 'Common Voice' project. The model architecture proved effective in learning the nuances of vocal attributes, and its performance was evaluated on several key metrics.

4.1 Dataset

The dataset employed in this research was sourced from Mozilla's Common Voice project, a globally inclusive initiative to create an expansive database for speech recognition software[2]. This project collects voice data from volunteers worldwide, ensuring a rich diversity of accents, languages, and age groups. The dataset includes voice samples of varying durations and is freely available, aligning with open-source principles and ethical considerations, which is made available under a Creative Commons Zero License (CC0). This licensing ensures that all voice data is public domain, offering an ethical framework that sidesteps privacy concerns typically associated with personal data. The CC0 license allows for the free use of the data without any legal or technical restrictions, making it an ideal resource for academic and commercial research that aims to respect and protect individual privacy. This framework aligns with the research's commitment to uphold ethical standards in data usage while enabling the exploration of advanced deep learning techniques for age prediction.

4.2 Model Performance

The LSTM model, utilizing Librosa features, achieved a training accuracy of approximately 62.40% and a validation accuracy of 54.25%. These figures suggest that the model has effectively captured the underlying patterns in the data without succumbing to overfitting, a common challenge in deep learning applications. The confusion matrix for the LSTM model, presented in Fig. 7a, provides further insight into the model's classification accuracy across different age categories, with notable precision demonstrated in the 'sixties' age group. This evidences the model's ability to generalize across the dataset, validating the robustness of its predictive power.

Fig. 5: The result of LSTM's validation loss and validation accuracy

Fig. 6: The result of RNN's validation loss and validation accuracy

Evaluation Metrics The evaluation of model performance in age group classification, as illustrated in Table 1, reveals distinctive characteristics for each model. Precision, recall, and F1-score provide valuable insights into their effectiveness across various age groups, complemented by the model's training and

validation loss. As demonstrated in the training and validation loss graph, the model shows a consistent decrease in loss over epochs, which suggests an effective learning rate and model fit. The closely tracking validation loss indicates good generalizability and minimal overfitting. Concurrently, as per Fig. 5 for LSTM model and Fig. 6 for RNN model, the training and validation accuracy graphs exhibit an upward trend, affirming the model's capability to accurately predict the age groups with a high degree of certainty. These combined metrics from the graphs and Table 1 paint a comprehensive picture of the model's robust performance.

Upon scrutinizing the comparative results of the predictive models, several patterns emerge. The Long Short-Term Memory (LSTM) model exhibits superior precision across the majority of the age categories, with a particularly noteworthy precision score of 0.76 in the 'sixties' category. This suggests that the LSTM model has a robust capacity for correctly identifying true positive cases within this demographic.

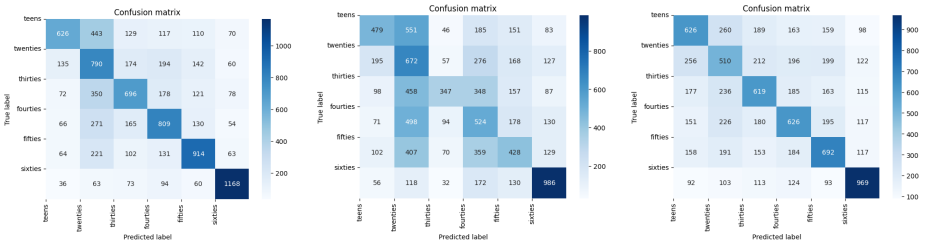
The Recurrent Neural Network (RNN) model presents a diverse set of results. While its precision appears modest, the model's recall rates are commendable, especially in the 'twenties' category, signifying a high efficiency in capturing all relevant instances within this age bracket. This capability is indicative of the RNN model's potential utility in applications where maximizing the identification of positive instances is critical.

In contrast, the Decision Tree (DT) model exhibits a more uniform performance with respect to the F1-score, signifying a harmonious balance between precision and recall. Despite this, the model shows a discernible variance in precision and recall across different age categories, potentially hinting at a susceptibility to overfitting or underfitting in specific groups.

The 'twenties' age group presents an area of challenge for all models, as evidenced by lower precision and recall scores. This phenomenon could be attributed to the intrinsic variability or non-distinctive features present within this age group's dataset, necessitating further investigative research.

Table 1: The outcomes for each predictive model

	Precision			*Recall*			*F1-score*			*Support*
Age	LSTM	RNN	DT	LSTM	RNN	DT	LSTM	RNN	DT	
teens	0.56	0.49	0.44	0.50	0.38	0.43	0.52	0.43	0.43	1495
twenties	0.38	0.28	0.31	0.42	0.28	0.31	0.40	0.28	0.31	1494
thirties	0.47	0.51	0.44	0.48	0.29	0.44	0.47	0.37	0.44	1495
fourties	0.55	0.31	0.43	0.51	0.39	0.43	0.53	0.35	0.43	1495
fifties	0.57	0.34	0.46	0.60	0.47	0.45	0.58	0.40	0.45	1495
sixties	0.76	0.62	0.63	0.75	0.67	0.64	0.76	0.64	0.64	1495
Total	3.29	2.55	2.72	3.26	2.48	2.70	3.26	2.47	2.70	



(a) LSTM (b) RNN (c) Decision Tree

Fig. 7: Confusion matrix predictions for different models

Overall, the choice of the best model depends on specific priorities. LSTM demonstrates a balance between precision and recall, making it suitable for general age group classification tasks. However, for tasks where precision or recall is paramount, such as targeting a specific age group, Decision Tree or RNN may be more appropriate, respectively. The suitability of each model should align with the project's unique requirements and objectives.

Comparative Analysis The LSTM model outperformed the SimpleRNN and Decision Tree models. LSTM's superior handling of sequential data and its ability to remember long-term dependencies make it well-suited for voice data analysis, unlike the SimpleRNN, which is more limited in this capacity. The Decision Tree, while interpretable, lacks the nuanced understanding of temporal features present in voice data.

Practical Implications The LSTM model's ability to predict customer age brackets with high accuracy presents significant opportunities for personalization in telesales. This predictive power can lead to enhanced customer service, fostering better customer relationships and potentially higher sales conversion rates. Future improvements could include exploring more complex models and feature sets to enhance performance further.

5 Conclusion

The completion of this study showcases the efficacy of deep learning techniques in discerning customer age groups through voice data within the telesales domain. The LSTM model, enhanced by RNNs, achieved a notable validation accuracy of 54.25%, supporting the theory that vocal features can accurately predict age ranges. These results confirm the model's applicability in real-world scenarios.

For telesales, this model's integration can revolutionize customer interaction by offering age-specific personalization, which may significantly enhance customer satisfaction and increase sales conversions.

Future research may involve real-time deployment to test the model's live interaction performance, data diversification to include various age groups and languages for broader applicability, and experimentation with other language models like BERT model, in particular, employed for its excellence in understanding the nuances of spoken language, enhancing the interpretation of speech context [5].

This research's innovative approach could lead to groundbreaking enhancements in customer service across various sectors, suggesting a transformative potential for customer relationship management leveraging voice data analytics.

6 Acknowledgement

We thank the Data Science Consortium, Chiang Mai University, for supporting us in this research.

References

1. Chroma feature analysis and synthesis. https://www.ee.columbia.edu/ dpwe/resources/matlab/chroma-ansyn/, accessed: date-of-access
2. Common voice. https://commonvoice.mozilla.org/ (2024)
3. Abadi, M., Agarwal, A., Barham, P., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems (2016)
4. Chollet, F.: Deep learning with python (2021)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
6. Harte, C., Sandler, M., Gasser, M.: Detecting harmonic change in musical audio (2006). https://doi.org/10.1145/1178723.1178727
7. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups (2012). https://doi.org/10.1109/MSP.2012.2205597
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014)
9. Kone, V.S., Anagal, A., Anegundi, S., Jadhav, P., Kulkarni, U., M, M.S.: Voice-based gender and age recognition system (2023). https://doi.org/10.1109/InCACCT57535.2023.10141801
10. Kuncheva, L.I., Hadjitodorov, S.T.: Using diversity in cluster ensembles (2004). https://doi.org/10.1109/ICSMC.2004.1399790
11. Kwasny, D., Hemmerling, D.: Gender and age estimation methods based on speech using deep neural networks (2021). https://doi.org/10.3390/s21144785, https://doi.org/10.3390/s21144785
12. McFee, B., Raffel, C., Liang, D., Ellis, D.P.W., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python (2015)
13. Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S., Sainath, T.: Deep learning for audio signal processing (2019), https://arxiv.org/abs/1905.00078
14. Sánchez-Hevia, H.A., Gil-Pita, R., Utrilla-Manso, M., et al.: Age group classification and gender recognition from speech with temporal convolutional neural networks (2022). https://doi.org/10.1007/s11042-021-11614-4

15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017), retrieved from https://arxiv.org/abs/1706.03762