Analysis of 911 Emergency Call Recordings Using Machine Learning Techniques

Nathaniel Maxey

December 19, 2024

Abstract

This study analyzes a dataset of 911 emergency call recordings collected by the late Gary Allen, a former editor and publisher of Dispatch Monthly magazine and the 911 Dispatch website. The dataset consists of critical and unusual 911 calls, accompanied by metadata detailing call type, date, state, citizen initiation status, and potential deaths. The audio files were initially in MP3 format and converted to WAV, then processed to extract Mel Frequency Cepstral Coefficients (MFCCs) for further analysis. Despite filtering out incomplete or corrupted data, a subset of the dataset was used to classify calls into critical and non-critical events based on metadata. Several machine learning techniques, including Random Forest, Support Vector Machines (SVM), Logistic Regression, k-Nearest Neighbors (KNN), Decision Trees, and Naive Bayes, were applied. The Random Forest model demonstrated the highest accuracy, achieving perfect classification results, followed by KNN and SVM. These results suggest that machine learning models can effectively predict critical events from 911 call data.

1 Introduction

The dataset used in this analysis contains 911 emergency call recordings, which were collected from public sources by the late Gary Allen, a former editor and publisher of Dispatch Monthly magazine and the 911 Dispatch website. These recordings, mostly in MP3 format, represent critical or unusual 911 calls, providing valuable second-hand experience for public safety dispatchers. The collection includes various types of emergencies, such as airline crashes, fires, and other significant events. Along with the recordings, metadata was manually coded, which includes detailed information about each call, such as the date, state, citizen initiation status, and whether deaths or potential deaths occurred. Additionally, the metadata includes the type of call (e.g., fire, medical emergency, accident), the website from which the call was sourced (911 Dispatch website), and an indicator of whether the event was a false alarm. Some recordings contain partial audio or censored material, while others include intros indicating their legitimacy for court use.

2 Methods

I followed several steps to process the dataset, including data cleaning, feature extraction, and model development. Initially, I converted the MP3 files into WAV format using the `tuneR` package, ensuring that all audio files could be processed uniformly. Subsequently, the WAV files were converted to mono, as many of the files were originally in stereo. This was achieved using the `mono()` function in the `tuneR` package. The next step involved extracting Mel Frequency Cepstral Coefficients (MFCCs) from each audio file using the `seewave` package. MFCCs are commonly used in speech and audio signal processing because they capture the spectral characteristics of audio signals. However, during this process, some difficulties arose due to the diversity and size of the dataset. Many of the 911 call recordings had missing data or incomplete audio, and as a result, many recordings were filtered out due to the lack of sufficient data for meaningful analysis. This posed a challenge in ensuring a robust dataset for feature extraction, as incomplete or corrupted files had to be excluded.

After processing, the features were aggregated by calculating the column means of the MFCC matrices, providing a condensed set of 13 features per recording. Additionally, a binary label for "critical events" was created based on metadata, where a critical event was defined as one where there were deaths or potential deaths, excluding false alarms. This label was merged with the extracted MFCC features based on the matching filenames between the metadata and the audio files.

For supervised learning, I employed a range of machine learning techniques, including Random Forest, Support Vector Machines (SVM), and Logistic Regression, to classify the 911 calls into critical and non-critical events. To address class imbalance, I applied the Synthetic Minority Over-sampling Technique (SMOTE) to oversample the minority class in the dataset. I then split the data into training and test sets, training the models on the training set and evaluating them using confusion matrices and ROC curves. Finally, the models were compared to determine the best performer in terms of accuracy, sensitivity, and specificity for predicting critical events in the 911 call data.

3 Results

In the results, multiple classification models were tested on the SMOTE-balanced dataset, and performance metrics were calculated to evaluate each model's effectiveness.

The logistic regression model exhibited low significance for the predictor variables, as indicated by p-values near 1, with a warning about non-convergence and extreme fitted probabilities (0 or 1). The model's accuracy was 89.29%, with a sensitivity of 1.0000 and specificity of 0.8500. However, it showed a relatively low positive predictive value (0.7273) and a balanced accuracy of 0.9250.

The Support Vector Machine (SVM) model achieved an accuracy of 89.29%, with a sensitivity of 0.8750 and specificity of 0.9000. It had a balanced accuracy of 0.8875, demonstrating reliable performance in distinguishing between the two classes.

The Random Forest model outperformed others, achieving perfect accuracy (100%) with sensitivity and specificity both at 1.0000. The model demonstrated a kappa value of 1, indicating ideal agreement between predicted and actual class labels.

The k-Nearest Neighbors (KNN) model also showed high performance, with an accuracy of 92.86%, sensitivity of 1.0000, and specificity of 0.9000. The balanced accuracy was 0.9500, highlighting the model's strong classification ability.

The Decision Tree model performed moderately, with an accuracy of 85.71%, sensitivity of 0.7500, and specificity of 0.9000. Its balanced accuracy was 0.8250, reflecting its ability to classify the data but with less precision than other models.

The Naive Bayes model had an accuracy of 82.14%, with a sensitivity of 0.7500 and specificity of 0.8500. It showed a balanced accuracy of 0.8000, suggesting that it was less effective than the other models in terms of classifying both positive and negative outcomes.

These results suggest that Random Forest is the most effective model for this dataset, followed closely by KNN and SVM, which also performed well. Although the logistic regression model yielded a high accuracy, it showed issues with convergence and significance of predictors, and the Decision Tree and Naive Bayes models provided more moderate results.

4 Conclusion

This analysis demonstrates the potential of machine learning models, particularly Random Forest, to effectively predict critical events from 911 call data. Although there were challenges in processing a large and diverse set of 911 calls, including the filtering of recordings due to missing data and the conversion of audio formats, the methods employed allowed for successful classification of critical and noncritical events. Future work may include refining the feature extraction process and addressing the limitations posed by missing or incomplete data.