The University of Texas at Austin
Chandra Department of Electrical and Computer Engineering
Cockrell School of Engineering

Predicting Suicide Rates within the United States

Jesus Hernandez, Arnav Kithania, Ayan Chaudhry, Nicholas St. Martin

Abstract

Our study examined diverse datasets encompassing age, race, sex, and other demographics to train predictive models for estimating suicide risk. XGBoost and 5 Nearest Neighbors classifiers performed the best. A text-based classifier was trained on a relevant dataset and achieved high area under ROC. We hope to improve suicide prevention efforts by enabling the identification of high-risk individuals and providing them with better support.

Background

Suicide rates have grown signifcantly within the United States with over 50,000 people losing their lives to mental stressors in 2024.

The goal of this project is to classify demographics as either high or low risk in regards to suicide using the Suicide Rates Datsets. Most publicly data available is comprised of death rate per 100,000 individuals by various demographics, and most labeled-text datasets are scraped from social media platforms and volunteer labeled.

Our goal is to create a starting point for studying suicide rates in relation to demographics using machine learning classification techniques.

Data preprocessing

Two datasets will be utilized in our findings:

1. Suicide Rate Dataset: Non-user friendly data. Each input varied in the types of features it displayed.
2. Text Classification Dataset: Text extracted from suicide related columns on Reddit, classified by volunteers as “suicidal” or “non-suicidal".

Demographics Dataset

Preprocessing involved duplicate removal and text analysis for feature extraction. Extracted features were then one-hot encoded for use in classifiers. The risk level classification feature was determined using 2 standard deviations above the mean as our threshold.

Text Classification Dataset

Preprocessing involved transforming text into vector-space representations using Word2Vec, chosen because of the small nature of individual data points and performance.

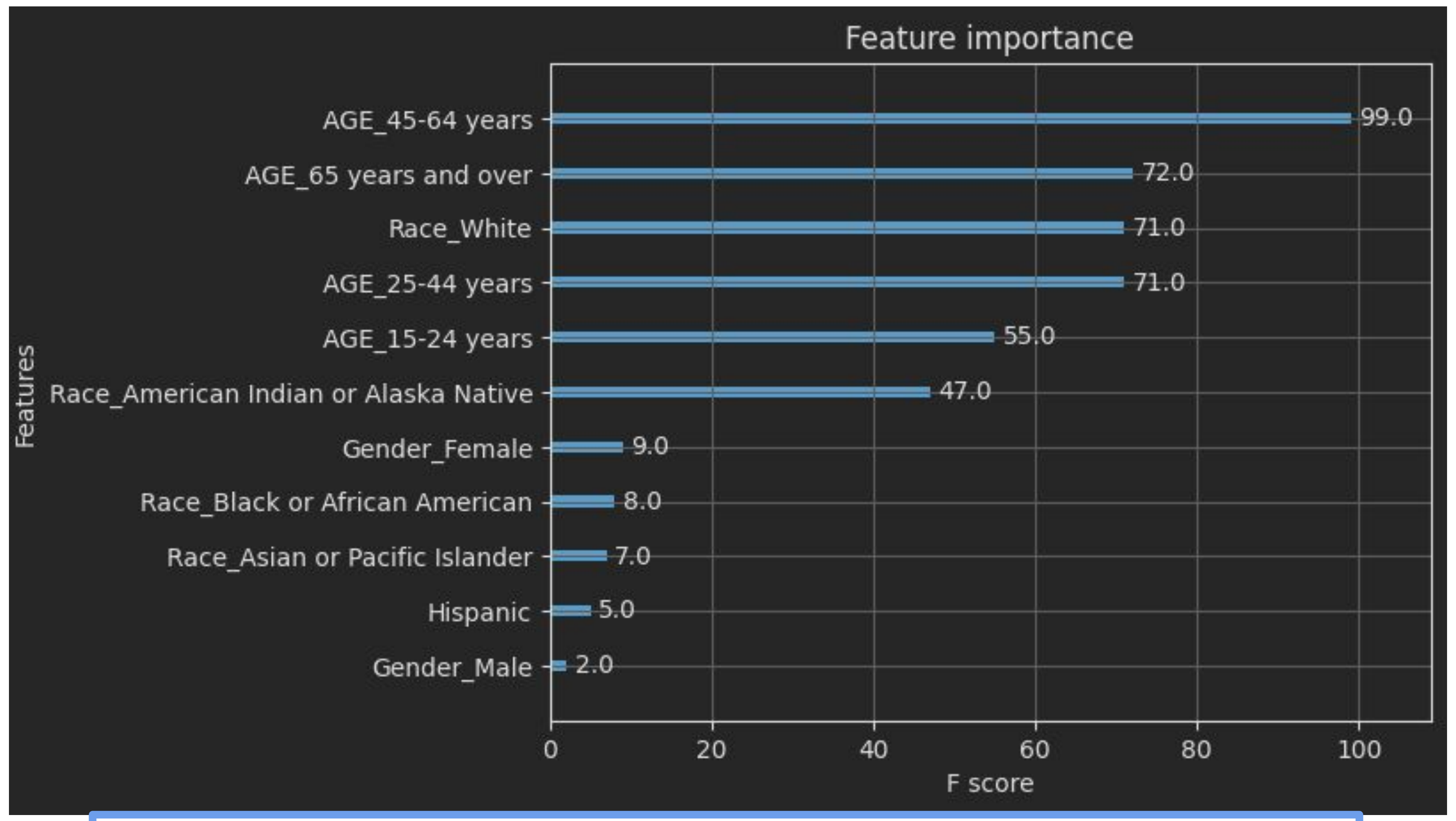

Fig. (a) XGBoost Model Feature Importance

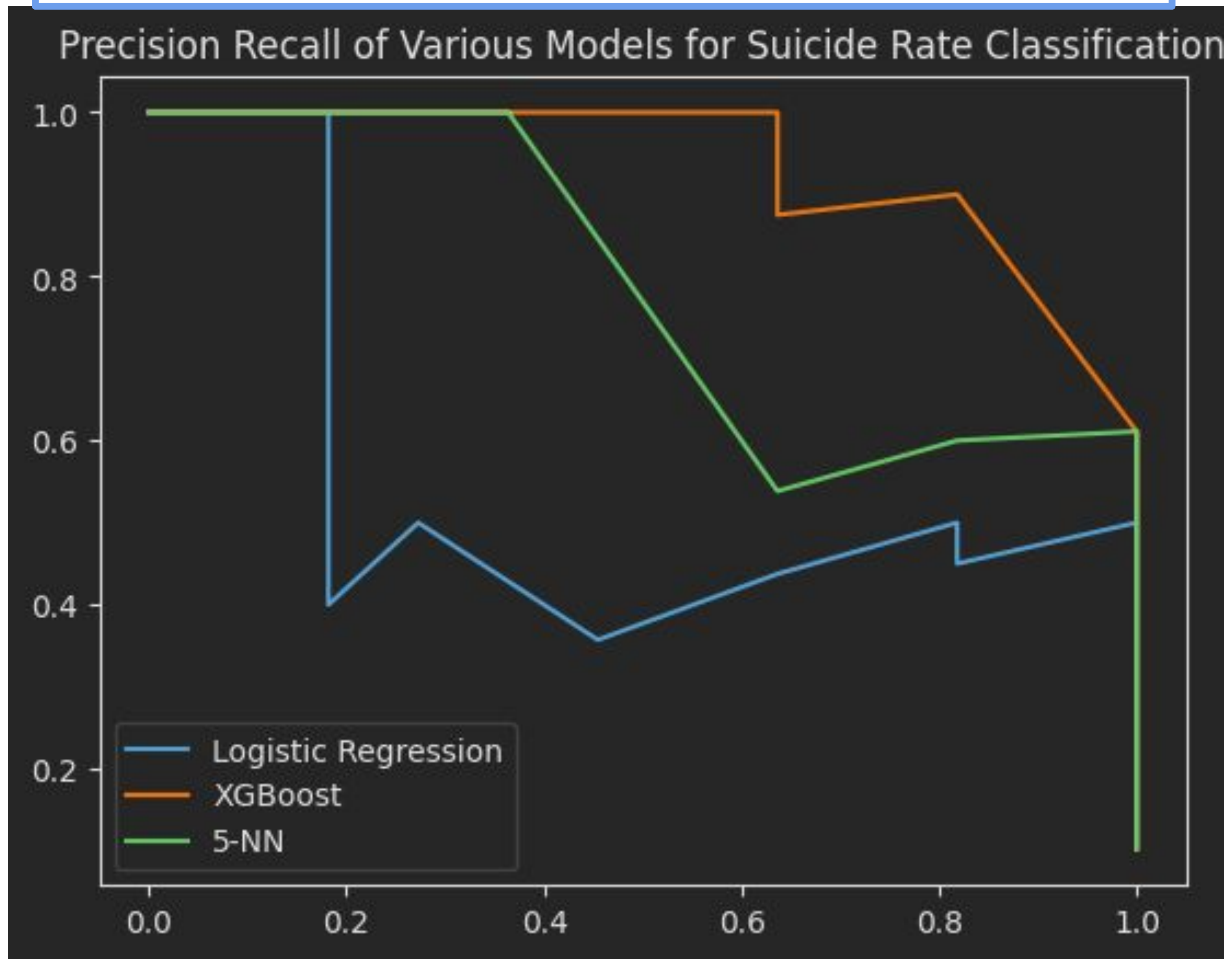

Fig. (b) Precision-Recall for Different Classifiers

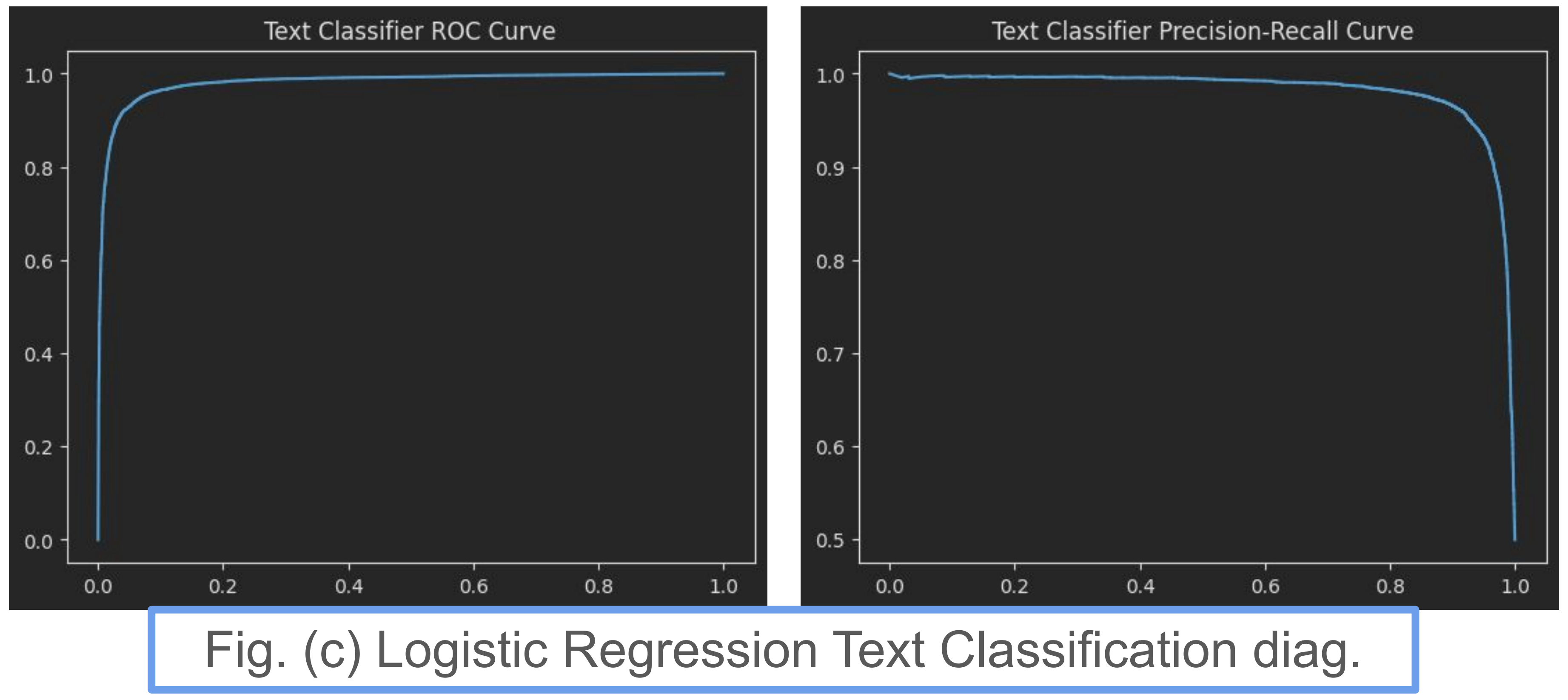

Fig. (c) Logistic Regression Text Classification diag.

Results & Important Features

Demographics Dataset

During evaluation, a 0.1 test-train split was during testing and evaluation. Due to the nature of the problem, the dataset demonstrates a sparse count of severe cases compared to non-severe cases. This implied the use of the True Positive Rate and Precision-Recall curves as model evaluation metrics.

We found that the 5-nearest neighbors model had the best TPR, with a rate of 0.978, followed by tuned XGBoost (0.960), SVM (0.950), and Logistic Regression (0.935). XGBoost gave a much more well balanced precision-recall curve than the 5-NN model as well.

Text Classification Dataset

The classification dataset was balanced, so ROC-AUC was used for model evaluation. A low-L2-regularization logistic regression model was chosen based on computing constraints and performance, yielding a score of 0.9811 with a similar train-test split.

Implications & Conclusions

From the demographic dataset analysis, XGBoost revealed age as the most significant factor in identifying individuals at risk. Specifically, ages 45-64 exhibited the strongest correlation, followed by those 65 and older, as depicted in Figure (a), illustrating the feature importance from XGBoost.

In the text classification dataset, after encoding the words appearing in specific phrases, notable hot words such as "help," "die," and "me" emerged, indicating potentially negative contexts. Leveraging this insight, monitoring various social media platforms for these key terms can aid in identifying individuals who might be at greater risk and could benefit from outreach efforts.

References:

[1] Centers for Disease Control and Prevention. (2021, April 21). Death rates for suicide, by sex, race, Hispanic origin, and age: United States. https://catalog.data.gov/dataset/death-rates-for-suicide-by-sex-race-hispanic-origin-and-age-united-states-020c1

[2] "Suicide and Depression Detection" by Nikhileswar Komati is licensed under CC BY-SA 4.0