Time Series Forecasting For Mortality

1st Venkat Sairam Veeranki
Data Science
Michigan Technological University
Houghton, MI 49931 USA
veerank@mtu.edu

2nd Rahul Teja Bolloju
Data Science
Michigan Technological University
Houghton, MI 49931 USA
bolloju@mtu.edu

3rd Dhanush Biligiri N H
Data Science
Michigan Technological University
Houghton, MI 49931 USA
dnarsipu@mtu.edu

***Abstract*—This project intends to give significant insights into the causes of mortality in the United States by examining the CDC's National Vital Statistics System dataset. The dataset consists of 22 files including demographic and cause-of-death data from 2005 to 2015. We focused on the seven leading causes of death: cancer, heart disease, respiratory illness, stroke, accidents, diabetes, and old age (with Alzheimer's and Parkinson's). The project turns the dataset into a regression dataset in order to do a time series forecasting of mortality using XgBoost and a couple of other models and evaluated over various evaluation measures. The project's significance comes from its ability to improve public health outcomes, encourage innovation, and support evidence-based decision-making.**

I. Introduction

The problem under consideration is a CDC's National Vital Statistics System dataset, which releases annual detailed reports on deaths in the United States, including demographic and cause-of-death data. This dataset is a collection of CSV files, each containing a year's worth of data, and paired with JSON files containing code mappings. Analyzing this would helpus to better understand how deaths are occuring in the united state.

The project's contribution lies in providing valuable insights into the different causes of death, their demographics, and trends over time. This information is crucial for the government to design better healthcare policies and allocate resources effectively. Additionally, the fitness industry can utilize this data to develop products tailored to specific diseases and promote a healthier lifestyle. The pharmaceutical industry can leverage this information to produce much-needed medicines and vaccinations and conduct targeted research on the most commonly occurring deaths. Overall, the project's significance lies in its potential to improve public health outcomes, drive innovation, and support evidence-based decision-making.

To analyze these deaths, we have converted the problem into a time-series forecasting problem. As a general overview, our dataset has multiple features related to the deceased individuals' demographics, cause of death, date of death, race, age, and a couple of different features. In our problem, we are only trying to understand the seven major causes of death that had occurred over the years in the data. We extracted only these seven instances of death from the dataset and did feature engineering on top of those instances. These features were ultimately used in the prediction of a forecasting model.

To forecast the problem, we are converting the dataset into a regression dataset and used this data to build regression models. Specifically, we are using two different algorithms: the XgBoost and SVM regression model. We are analyzing the best model for the problem, which will accurately forecast the data. Additionally, we will find the best hyperparameters by using several different evaluation metrics, including R^2, MSE, MAE, and RMSE.

II. Related Work

The Reference[1] provides a literature review of previous studies on time series forecasting and mortality rate prediction, with a focus on Nigeria. The authors discuss various techniques used for mortality rate prediction. They highlight the strengths and weaknesses of each approach and examine the factors that influence the accuracy of the predictions, such as data quality and model selection. The results of the study showed that all three models were able to accurately predict under-five mortality rates in Nigeria. However, the ANN model performed the best, followed by the ARIMA and HWES models for time series modeling. The section concludes by emphasizing the need for accurate mortality rate predictions to inform policy decisions and improve public health outcomes in Nigeria.

III. Data

The National Vital Statistics System of the Centers for Disease Control and Prevention (CDC) is the primary source of data. The National Vital Statistics System provides complete information on births and deaths in the United States. The NVSS, established by the NCHS-National Center for Health Statistics, collects and analyzes vital statistics data for the United States. The NVSS compiles statistics from across the United States, including demographics, births, marriages, divorces, and fetal deaths. The data provided by the NVSS is frequently used by researchers, the general

public, and public health authorities.

We are considering the mortality statistics from 2005 to 2015. The data is collected and stored in two different formats, JSON and CSV. The JSON file contains mappings of key-value pairs, which are used to understand the values in the CSV file. In total, there are 22 files (11 JSON files and 11 CSV files). Each CSV file contains 77 features in a single instance of data, and in total, we have recorded 27 million instances of deaths in the data, making it a high-volume dataset. The total size of our dataset measures around 4.33 GB.

IV. DataPreprocessing

Given the large number of possible causes of death, it was necessary to narrow the scope and focus on a smaller subset of causes that were deemed to be the most significant. In our analysis, we considered the seven major causes of death in the US: Cancer, Heart Disease, Respiratory Issues, Stroke, Accidents, Diabetes, and Old Age (including both Alzheimer's and Parkinson's). Figure 1 shows the counts of deaths attributed to each disease in the US. Heart disease is the leading cause of death in the US.

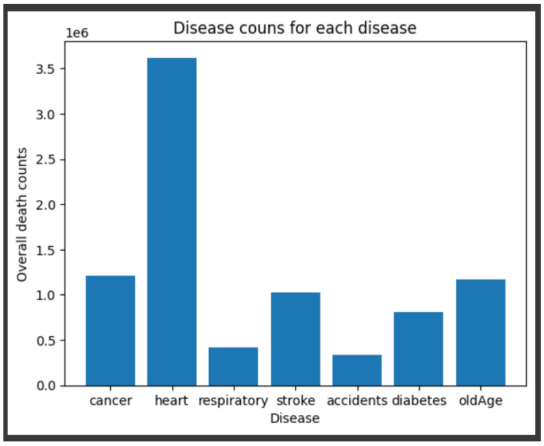


Fig. 1.

Out of the 77 features that we are considering, we have selected 6 features from the CSV data file for our analysis. These features are: Curren_data_year, Month_of_death, Resident_status, Sex, Race_recode_5, and 358_cause_recode. The table below (Table 1) provides a detailed explanation of the variable types and data types for each of these features.

Each of the JSON files contains a mapping of individual features in the CSV file. The mapping for each variable can be understood from Table 2 below.

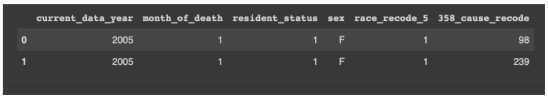
After combining both datasets from the CSV and JSON files for each year, we can consider a sample instance of our final dataset shown in Figure[2]. The "current_data_year" attribute indicates the year in which the death occurred, such as 2005. The "month_of_death" attribute specifies the month of the year in which the death occurred, with 1 indicating January. The "resident_status" attribute describes the residency status of the individual at the time of their death, where 1 indicates residency in the US. The "Sex" attribute indicates the gender of the person, in our case, denoting M and F. The "race_recode_5" attribute indicates the race of the individual, with 1 indicating "White". Finally, the CSV file includes a key that describes the causes of death. In our sample, it indicates "Malignant melanoma of the skin".

TABLE I

Feature	Variable Type	Data Type
Current_data_year	Nominal	Integer
Month_of_death	Nominal	Integer
Resident_status	Nominal	Integer
Sex	Nominal	String
Race_recode_5	Nominal	Integer
358_cause_recode	Nominal	Integer

TABLE II

Feature	JSON Key- value pair
Current_data_year	{'2005':'2005'}
Month_of_death	{'01': 'January', '02': 'February', '03': 'March', '04': 'April', '05': 'May', '06': 'June', '07': 'July', '08': 'August', '09': 'September', '10': 'October', '11': 'November', '12': 'December'}
Resident_status	{'1': 'RESIDENTS', '2': 'INTRASTATE NONRESIDENTS', '3': 'INTERSTATE NONRESIDENTS', '4': 'FOREIGN RESIDENTS'}
Sex	{'F': Female,'M':Male}
Race_recode_5	{'0': 'Other (Puerto Rico only)', '1': 'White', '2': 'Black', '3': 'American Indian', '4': 'Asian or Pacific Islander'}
358_cause_recode	{"001": "I. Certain infectious and parasitic diseases (A00-B99)", "002": "Intestinal infectious diseases (A00-A09)", "003": "Cholera (A00)", "004": "Other intestinal infectious diseases (A01-A08)", "005": "Typhoid fever (A01.0)",.....} Here each key is a code given to a certain disease present in the Value of that key, we have almost 358 causes of death.



	current_data_year	month_of_death	resident_status	sex	race_recode_5	358_cause_recode
0	2005	1	1	F	1	98
1	2005	1	1	F	1	239

Fig. 2.

As part of our mortality forecasting project, we created a new data frame that includes key features for predicting mortality outcomes. To build models on the data, we converted the "sex" feature from a string data type to a boolean type, with "0" representing male and "1" representing female. Additionally, we combined the "Current_data_year" and "month_of_death" columns into a single column named "month." This column ranges from 1 to 132, with "1" representing January 2005 and "132" representing December

2015.

To better understand the impact of different causes on mortality rates, we replaced the ”358_cause_recde” feature with a numerical equivalent. Specifically, we assigned numeric codes to each cause, with ”cancer” assigned a code of 1, ”heart” assigned a code of 2, ”respiratory” assigned a code of 3, ”stroke” assigned a code of 4, ”accidents” assigned a code of 5, ”diabetes” assigned a code of 6, and ”oldAge” assigned a code of 7. We calculated the total death counts by grouping over each variable described above.

In Figure 3, we can observe the plot of death counts over the months. We can see a periodic trend for every 12 months in the death counts.

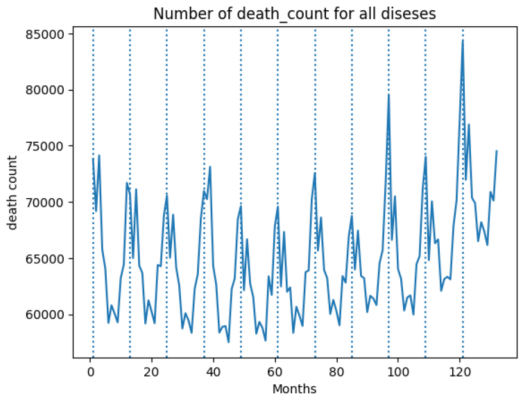


Fig. 3.

To analyze the trend in death count over the month, we grouped the values by month and calculated the cumulative sum of these counts. We found a linear growth in deaths over the months, as shown in Figure 4.

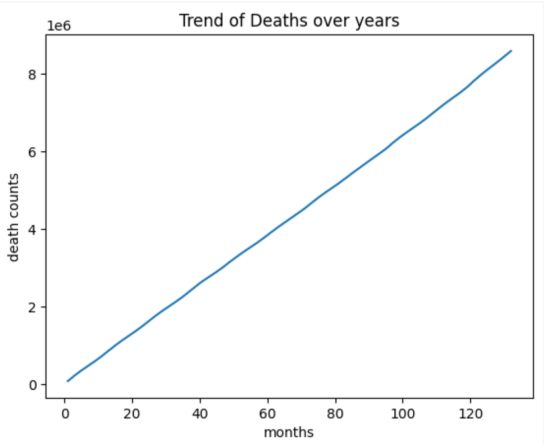


Fig. 4.

Since we observed that the death counts were periodic throughout the year, we sought to examine the monthly variations during the 11-year period. Upon analyzing the data presented in Figure 5, we identified a pattern similar to that observed in the yearly cycles depicted in Figure 3.

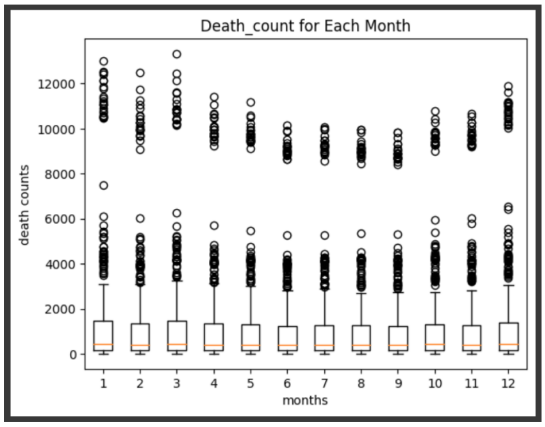


Fig. 5.

In our mortality analysis project, we utilized a variety of statistical techniques to uncover patterns and trends in the data. One approach that we employed was to calculate the average death counts for different subgroups within the data set. To achieve this, we applied a custom function named ”avg_cal” to our data frame using various sets of grouping variables. Initially, we computed the average death count by month, resident status, and cause of death, and stored the results in a new column named ”resstatus_avg_death.” This enabled us to compare mortality rates between various categories of residents (such as resident vs. non-resident) and different causes of death, while controlling for the effect of time. Subsequently, we calculated the average death count by month, sex, and cause of death, and stored the results in a new column called ”sex_avg_death.” This analysis allowed us to examine gender-based differences in mortality rates while controlling for time and cause of death. Lastly, we computed the average death count by month, race, and cause of death, and stored the results in a new column named ”race_avg_death.” This analysis allowed us to explore racial disparities in mortality rates, while controlling for time and cause of death.

By applying these calculations to our data frame, we have been able to identify important trends and patterns in mortality rates for various subgroups of the population. These insights can assist in guiding public health interventions and policy decisions aimed at reducing mortality and improving overall health outcomes in the United States. This conversion has enabled us to more effectively analyze the relative impact of different causes of death on overall mortality rates in the United States and has provided a more precise framework for our statistical modeling and forecasting efforts. By carefully selecting and transforming key features in our data set, we have been able to conduct a more robust and accurate analysis

of mortality trends.

We have created additional variables in addition to those mentioned earlier, specifically for the variables 'death_count', 'resstatus_avg_death', 'sex_avg_death', and 'race_avg_death'. We created these columns by generating lag variables and shifting the data by adding 12 months to all the variables. This was done because we discovered a yearly periodic mortality pattern. We then converted the nominal variables in the dataset into one-hot encoded variables. In the end, our dataset contains of 24,188 rows and 28 columns, totaling 2.4 MB.

These conversions enabled us to analyze the impact of different causes of death on overall mortality rates in the United States more effectively. They also provided a more precise framework for our statistical modeling and forecasting efforts. By meticulously selecting and transforming key features in our dataset, we were able to conduct a more robust and accurate analysis of mortality trends.

V. Model

To gain a better understanding of mortality trends in the United States and improve the accuracy of future mortality rate forecasts, we undertook a comprehensive data analysis using a variety of statistical and machine learning techniques. The analysis was based on a large dataset of mortality data, which was thoroughly processed and cleaned to ensure accuracy and consistency.

Once the data had been cleaned and processed, we carried out a series of descriptive analyses to identify key trends and patterns in mortality rates over time. We utilized a range of techniques such as time series analysis and regression analysis to investigate the relationship between mortality rates and various demographic and environmental factors, including age, sex, geographic location, and climate.

Our project involved the creation of two distinct models for mortality forecasting: SVM Regression and XGBoost. SVM Regression is a statistical approach that analyzes the relationship between one or more predictor variables and a response variable. It is frequently used in fields such as economics, social sciences, and health research to quantify the relationship between variables and make predictions about future outcomes based on past data.Because it can handle both linear and nonlinear relationships between the input and output variables, SVM regression is a viable option for predicting models. Additionally, it can handle input spaces with many dimensions and is resilient to outliers in the data. Both univariate and multivariate forecasting may be done using SVM regression, and it is simple to adapt to handle time series data. For feature selection and model justification, the model offers a clear understanding of the relative weights of each input variable. Additionally, SVM regression has a solid theoretical background, which promotes good generalization performance and can provide information about the data's structure.

In contrast, XGBoost is a well-known machine learning algorithm used for regression, classification, and ranking problems. Because it can handle a wide range of data types, including numeric, categorical, and ordinal data, XGBoost is a well-liked option for predicting models. It is also very scalable and capable of handling big datasets with a lot of dimensions. Because of its quick execution speed and ability to train models on parallel computing systems, XGBoost is a good option for complex forecasting projects. Strong regularization approaches in the model help to prevent overfitting, which is important in forecasting since the model must be able to successfully handle brand-new, unforeseen data. Additionally, XGBoost offers interpretable feature importance ratings that can be used to find the factors that are most crucial for predicting.

To construct our machine learning algorithms on the data, we separated the data into two components: the training data and the test data. As the data is time series data, we employed the first 122 months as the training data and the remaining data (months 122 to 132) as the test data. We utilized GridSearchCV to train both models over a range of hyperparameters to determine the best ones. Finally, we compared the two models using their optimal hyperparameters.

TABLE III

	SVM Regression	XGB Regression
Best Hyper Parameters	{'C'= 100, 'epsilon'= 2, 'gamma'= 0.001, 'kernel'= 'rbf'}	{'colsample_bynode'= 0.6, 'colsample_bytree'= 0.6, 'gamma' = 0.01, 'learning_rate'= 0.01, 'max_depth'= 9, 'n_estimators'= 500}
R^2	-0.02302	0.9986
RSME	1255.1523	45.0914
MAE	345.1669	16.8137
MAPE	994.5659	45.5060

We have observed that XGBoost regression outperformed other models, as indicated by its superior evaluation metrics such as R^2, RMSE, MAE, and MAPE. To visualize the model's performance, we plotted the predicted death count against the test data death counts over the respective months. The plot reveals that the predicted values generated by the model are quite similar to the actual test data, indicating that the model is accurately predicting the death counts over time.

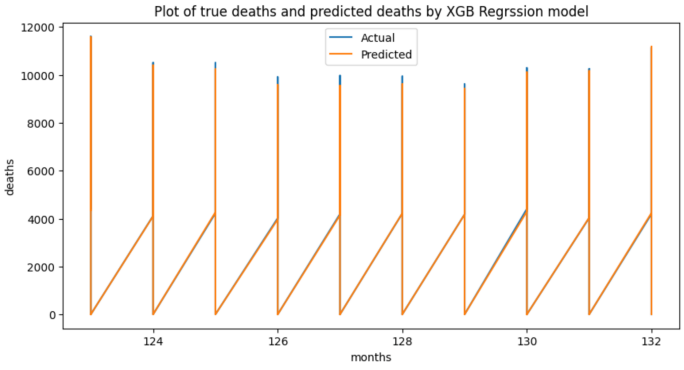


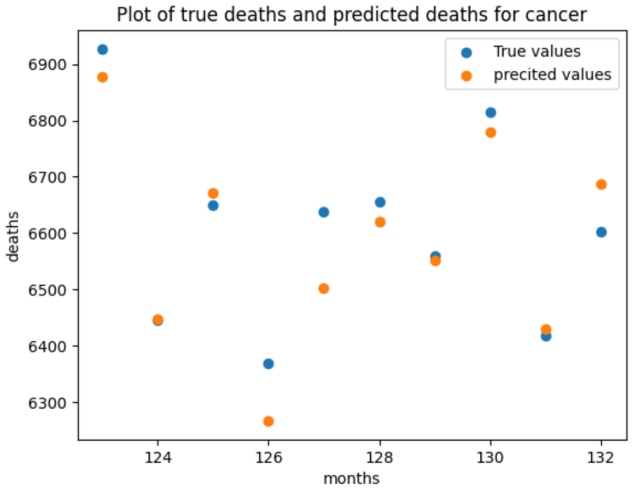
Fig. 6.

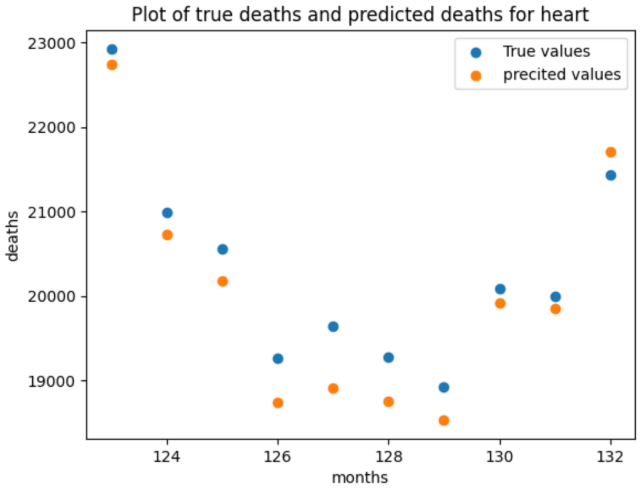
On the other hand, the plot of predicted deaths for SVM regression failed to capture a significant amount of information from the dataset and was not as effective in forecasting, as shown in the graph below.

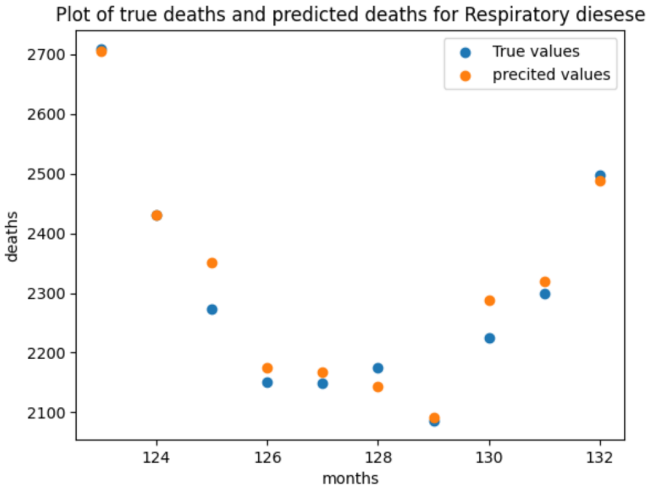
Fig. 7.

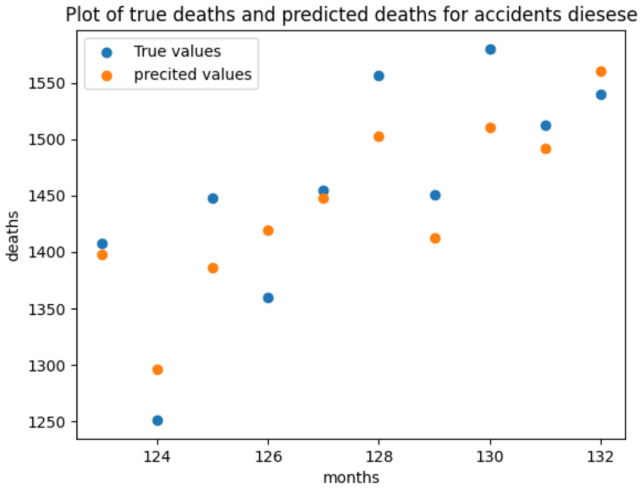
VI. Conclusion

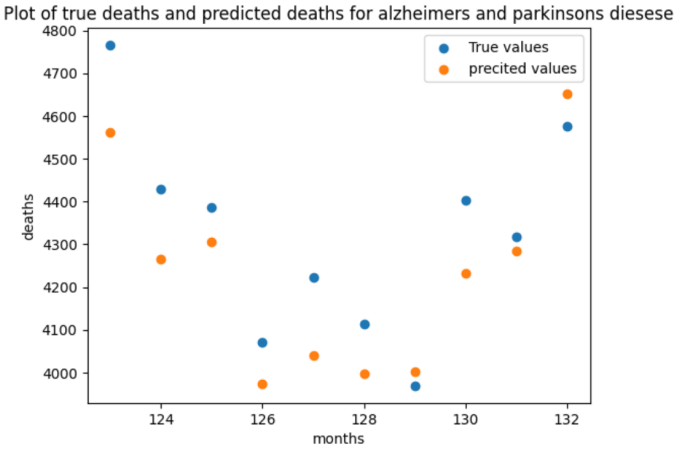
Overall, the models exhibited a very low mean absolute percentage (MAP) error across the six different types of deaths. This suggests that the models were relatively accurate in predicting the death counts for each category.











The below table shows MAE for each individual plots shown above.

Disease	MAE
Cancer	0.007309
Heart	0.01805
Respiratory	0.0114
Stroke	0.0316
Accidents	0.0265
Alzheimer's and Parkinson's disease	0.0264

REFERENCES

[1] Adeyinka, D. A., Muhajarine, N. (2020). Time series prediction of under-five mortality rates for Nigeria: comparative analysis of artificial neural networks, Holt-Winters exponential smoothing and autoregressive integrated moving average models. BMC Medical Research Methodology, 20(1). https://doi.org/10.1186/s12874-020-01159-9

[2] "Machine Learning for Mortality Prediction of Patients With Acute Myocardial Infarction" by B. E. Sadik and colleagues (2020)

[3] "Machine learning algorithms for predicting outcomes in trauma: A systematic review" by S. I. Alexander and colleagues (2020) "Predicting Heart Disease Mortality Rates using Machine Learning Algorithms" by N. T. Nguyen and colleagues (2019)

[4] "Predicting Death from Census Data: A Machine Learning Approach" by A. M. Albrecht and colleagues (2019)

[5] "Predicting Mortality in the ICU using Machine Learning and Physiological Data" by H. Shetty and colleagues (2017)