Heart Disease Prediction Model based on Prophet

Yi Lu

Department of Applied Mathematics, University of Washington, Seattle, USA

yil29@uw.edu

Abstract. Heart disease is one of the major causes of death for people of all races, genders, and nationalities. In the United States, for instance, heart disease causes more than 600,000 deaths every year and is the largest leading cause of death in 2020. A reliable heart diseases mortality prediction model could acknowledge the patients' medical professionals that the heart disease risk level of the specific group. This approach is significant in preventing further increases in heart disease mortality rates worldwide. Nowadays, multiple Machine Learning (ML) models, including hybrid models produced impressive predictions and realized that newly developed ML models might provide new perspectives on heart disease predictions. In this paper, we introduced the Facebook Prophet model (FB Prophet model), a time series prediction tool that could present seasonality in its result, since studies point out that heart disease mortality rate also shows seasonality. We produced an accuracy of approximately 94 % in predicting weekly heart disease mortality numbers in specific states. Furthermore, we explored the effects that external factors, ambient temperature, have on heart disease, and utilize this relationship in improving model accuracy.

Keywords: Heart Disease Prediction; Prophet; Pearson Correlation Coefficient.

1. Introduction

Heart disease, which includes all types of cardiovascular diseases (CVD), is a series of diseases that is the result of the change of multiple factors in the human body. Therefore, factors like landscape and distance would not limit the spread of heart disease, and it became a major threat to a considerable amount of the population in most countries. Moreover, the development level of the region and even medical resources is stronger correlated with the motility rate of heart disease. The United States as a developed country with the highest per capita GDP in medical industries is an excellent example. According to the US CDC, Heart disease causes 696,692 deaths in the year 2020, which is approximately 21% of all death and the largest leading cause of death in the US [1]. Therefore, the threat of heart disease constantly exists, and heart attack caused by heart disease kills patients in a short time, so acknowledging the patients' and local government's risk of heart disease and preparing medical resources to prevent heart attack-caused death became important. Heart disease mortality rate prediction models are forecasting tools that could promote predictions based on past data to help researchers and the public learn more about heart attack risk and prevent heart disease-caused death. As the importance of the heart disease mortality data prediction model is shown, it is necessary to understand the condition of existing forecasting tools.

As heart disease mortality numbers continue to increase, the needs for prediction models also increase over time, but the complicated mechanism of heart diseases located in human bodies created challenges in forming a reliable prediction model. The majority of existing Machine Learning forecasting models were established on specific features of human bodies. These models analyse patients' body features and their risk of death caused by heart diseases, and once the data set provides enough information for them, they could find the trend of heart disease mortality rate based on the features and produce predictions on patient heart disease mortality rate [2]. These models could provide relatively accurate predictions to specific patients based on their body features [3]. Moreover, they applied novel technics, such as a hybrid of multiple Machine Learning methods, to create more accurate models that inspire other researchers [4]. This innovative method gives helps us form improvements that could be applied in the research that stands on the perspective of improving the method's efficiency. Other perspectives such as improved classification [5] and data process [6] are also essential. However, these models have limitations caused by their data selection: they could only predict the risk of data that is collected at a specific time; they could not produce regional risk data would be time-consuming. In resolving the challenges caused by lacking time series analyzing skills, researchers presented a forecasting tool: Facebook Prophet. Facebook Prophet is an automatic forecasting tool aiming to promote reliable predictions while improving the efficiency of analysts. FB Prophet was used in predicting diseases with seasonality, such as COVID-19 and shows efficiency [7]. This paper discusses the use of FB Prophet in heart disease mortality number forecast since in the field of heart disease research, studies show that heart disease mortality also shows seasonality, and this seasonality is affected by multiple types of environmental factors (including ambient temperature) [8]. FB Prophet shows a new perspective in heart disease mortality number forecasting, and how to process this method with reliable data sets became the key.

In this paper, we utilized FB Prophet to promote predictions based on data.gov [9] dataset, and the models could forecast weekly death numbers caused by heart diseases for each state with an accuracy of approximately 93 %. Furthermore, we explored the relationship between ambient temperature and heart disease mortality rate by importing data sets of temperature for different states in the U.S. into the experiment, and then by analyzing the Pearson correction coefficient of them, we found a middle negative relationship. Since the relationship between ambient temperatures and heart disease mortality rate does exist, we performed the Facebook Prophet prediction model for temperature and plugged the result into the heart disease mortality model which has significant effects on heart disease prediction accuracy.

2. Methodology

2.1 FB Prophet

FB prophet tool is a time series forecasting tool. It is designed to produce predictions in a convenient, highly efficient, and user-friendly way [10].

2.1.1 Components of FB Prophet

This forecasting tool is formed by four variables: trend, seasonality, holidays and events, and errors; the equation formed by these variables is the fundamental part of the entire model.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t. \tag{1}$$

Where g is the trend, s is seasonality, h is holidays, and epsilon is an error.

The design logic of this forecasting tool is similar to constructing buildings, in which the designers construct the foundation first and then set up the following part on top of it. In this model, the trend is the foundation and the designers utilized multiple ways to make it reliable.

2.1.2 Trend Prediction

The equation of trend provides the overall forecasting for the sample. In this model, the trend equation setup the primary estimation for the sample, in another word, it has the largest effects on the forecasting. According to its importance, the researchers designed two equations for linear and non-linear samples. In the case of heart disease mortality data forecasting, we utilized the linear equation, which is not limited by the saturating growth. Furthermore, the designers utilized an adjustment variable that helps the finding of change points and reflect it in the trend.

$$g(t) = (k + \mathbf{a}(t)^{\intercal}\boldsymbol{\delta})t + (m + \mathbf{a}(t)^{\intercal}\boldsymbol{\gamma}), \tag{2}$$

Whereas k is the growth rate variable, δ has the rate adjustments variable, m is the offset parameter, and γj is set to -sj δj to make the function continuous.

As the reliable model of trend is set, the designers aim to increase the accuracy of forecasting, and the trace of seasonality and holidays (events) is their solution.

2.1.3 Seasonality and Holidays

The seasonality and holidays recognize the trends in a specific time interval or date. These two measures are utilized to find the variation of the sample that the trend could not recognize and apply these changes to the forecasting model. And seasonality and holidays work with similar logic, in which they work with 2N parameters β = [a1, b1, . . . , aN , bN], and fit the change rate into the vectors [10]. The difference is that the seasonality equation works with data in regular periods, while Holidays work with specific dates.

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right) \tag{3}$$

where P is the number of days that forms the interval (weekly, yearly), and {an, bn} matrix is the seasonality.

2.2 Pearson Correction Coefficient

Pearson correlation coefficient (PCC) is a normalized measurement of covariance, which measures the linear correlation between two sets of data. In this paper, we want to recognize the relationship between two variables, and PCC could provide numerical support to the correlation between the targeted data sets.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{4}$$

Where r is the correlation coefficient, xi is x variance, x hat i is the mean of x variance, yi is y variance, and y hat i is the mean of y variance.

3. Experiments

3.1 Data Sets

The heart disease mortality data from data.gov[9] was imported. The data set is “week counts of death by jurisdiction and select causes of death” which is a set of federal data. It includes weekly death counts for over ten leading diseases in each state in the U.S. from 2014 to 2019.

In exploring the role of temperature in the change in heart disease mortality rate, we imported a monthly temperature data set from NOAA [11]. This data set provides the monthly average ambient temperature for each state from 2014 to 2019.

3.2 Process Data

FB prophet requires the datasets to be transferred into a two columns table, including ‘ds’ and ‘y’. Therefore, the users would have to transfer the data sets that are in the form of lists into tables. We imported pandas and used the ‘DataFrame' method from pandas to combine the heart disease mortality data and the ending date of the week into the same table. When plugging the lists into the table, users should make sure that 'ds' column data are the data that define time, and 'y' column data are the values that the users aim to forecast. Besides the requirements mentioned above, the users also have to make sure the 'ds' column data is processed with to_ datetime from pandas before forecasting, otherwise, the tool will have problem reading time data. Thus, the preparation of data is finished, and the users could start forecast.

3.3 Process Method

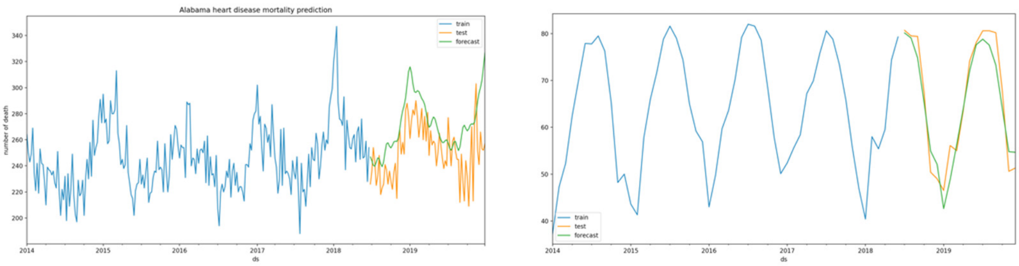
In producing the forecast, we set up the imported the Prophet method from the prophet in the first hand. After the method is available, the users could start the process of model setup, and the users could adjust the model setup based on their data sets and forecasting goal. The three most important adjustments that the users have to decide are growth type, seasonality type and seasonality mode. The growth type is the linear and nonlinear trend types, which would affect the equation utilized by the model; seasonality types decide the regular period of seasonality, for instance, daily, weekly, and yearly, and they could exist at the same time; seasonality mode decides how the model process the data set and recognize its seasonality. Based on the data set utilized in this paper, this paper utilized linear trends with both weekly and yearly seasonality, and for processing seasonality, the multiplicative mode is the choice. The model is stored in a variable Model. As the model is set, we set the first 75% of heart disease mortality data to be the training data, and the rest 25% to be the test data to check the accuracy. The training set was fitted into the model by using the fit method (Model.fit(train)). The forecasting tool is set, and the user could use predict tool in the Prophet file to process the data set. In this case, the data set is the test set, and after we process it with predict (Model.predict()), the prediction data is collected and could be compared with the real data to get the accuracy of this model.

On the other hand, this paper explores the relationship between heart disease mortality rate and environmental factors, ambient temperature for instance. We took the heart disease mortality data of one state and combine the number of deaths of the same month into one entry, so the monthly number of deaths could be collected. And since the monthly average temperature of states is imported in this paper, the data sets required for calculating the Pearson correlation coefficient are collected. This paper imported “personr” from the scipy.stats file, and imported the two data sets to produce a coefficient.

Beyond the method of finding the relationship between temperature and heart disease mortality data, this paper also discusses how to apply temperature effects to the prediction model. The first step is creating a forecasting model for temperature. We set up a similar model in the signs of progress discussed above, but since this data set is monthly collected, only yearly seasonality is applied. And this paper also collected accuracy data for the temperature forecast. After having the temperature data in the future, the second step is plugging them into the forecasting model of heart disease mortality numbers. We utilized to "holiday" section in the prophet model to accomplish this goal. The holiday section requires the inputs to be a table of dates, and past dates could help the model predict future dates, for example, using past ten years' thanksgiving dates and next year's thanksgiving date as a list, the model would learn from the data change in past ten thanksgiving and apply this change to the next thanksgiving. In this paper, we used the coldest month in the past years and the coldest month predicted in the next year as a list (weeks of those months) and added them as holidays (events) into the model to improve accuracy.

4. Results and Analysis

The FB Prophet prediction model for heart disease mortality data produced an accuracy of approximately 93 % for each state (Table. 1). The variation in accuracy is caused by multiple factors, and we recognized there are peaks in the number of deaths that occurs every year that affect the forecast. Take the forecast of Alabama (Fig. 1a) as an example, there is a significant peak in the winter of 2018, and a relatively low number of deaths in the winter of 2019. It is almost impossible for the model to predict such variations unless the researchers could find some factors that cause these types of changes and apply the mechanism to the model. According to the graph and studies on heart disease mortality rate, we recognized ambient temperature as an affecting factor [8].



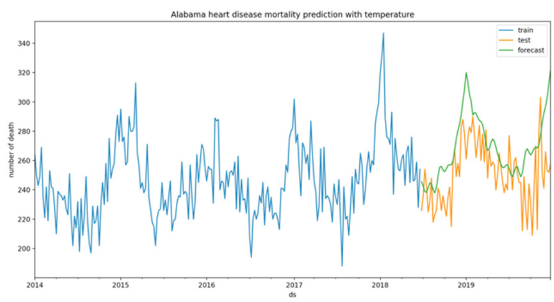
(a) **(b)**

Fig 1. Alabama: (a) weekly death forecasting model
(b) monthly temperature forecasting model

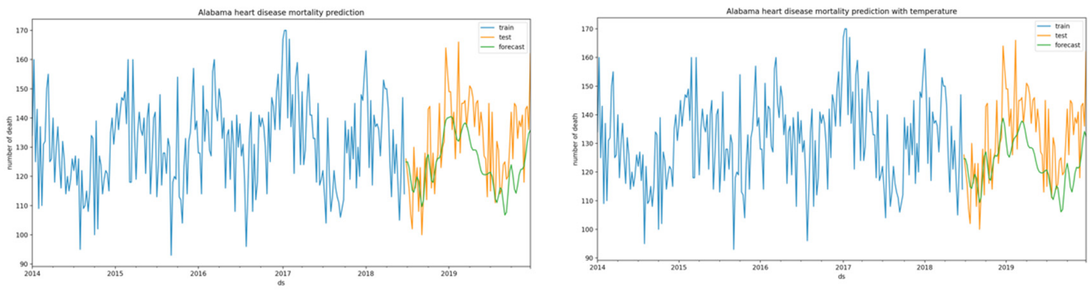
Table 1. States: Pearson correlation and comparison of model accuracy

State name	Forecasting accuracy (%)	Forecasting accuracy with temperature (%)	Pearson correlation coefficient
Alabama	90.8	91.4	-0.457
California	95.7	96.1	-0.534
Georgria	95.8	95.8	-0.472
Oregon	91.8	91.2	-0.433
Texas	95.1	95.0	-0.480
Utah	90.7	90.5	-0.413

In this paper, we applied the Pearson correlation coefficient, and with the data sets plugged in, there is a middle negative correlation (Table 1.) between temperature and heart disease mortality number. This relationship shows that when the temperature gets low, the number of deaths caused by heart disease tends to increase. According to the existing relationship, we applied the FB Prophet model to the average temperature and collected data for several states. The result shows an accuracy of about 95%. The error might be caused by the small sample or extreme climate, but this model could provide a reliable forecast in the coldest month of the year for most states (Fig. 1b). Therefore, we utilized the temperature model as a tool to identify the coldest weeks of the year and apply the forecast to the heart disease mortality number predicting model as a holiday. And by observing the comparison in Table 1., we confirm that this attempt shows improvements in some cases, take Alabama as an example (Fig. 2a), the temperature factor improved its forecasting accuracy at a rate of 0.4%, but in other cases like Oregon (Fig. 2b & 2c), it would not help. When matching the Pearson correlation coefficient with the effects that the temperature factor has on the model, we found out that a stronger relationship does not guarantee a higher rate of accuracy improvement after adding temperature as a factor. There are still factors that are not involved and more data need to be collected to increase this model's reliability.



(a)



(b) **(c)**

Fig 2. Forecasting models: (a) Alabama weekly death with temperature (b) Oregon weekly death (c) Oregon weekly death with temperature

5. Conclusion

Forecasting heart disease mortality number for a specific region in time series is significant in managing medical resources and will save lives. FB prophet could provide more accurate models in multiple diseases to the local government when they are planning to manage medical resources. Furthermore, since this model is working from a new perspective, it would not affect the development of human health condition features-based machine learning models, and they can collaborate to save more lives. And more research-based the relationship between heart disease and external factors is needed. In the future, we plan to collect weekly temperature data to help increase the accuracy of temperature accuracy and its importance in heart disease predictions. Furthermore, the application of technics such as a hybrid of multiple Machine Learning method is also a way that we want to explore.

References

[1] Murphy S L, Kochanek K D, Xu J, et al. Mortality in the United States, 2020[J]. 2021.

[2] Repaka A N, Ravikanti S D, Franklin R G. Design and implementing heart disease prediction using naives Bayesian[C]//2019 3rd International conference on trends in electronics and informatics (ICOEI). IEEE, 2019: 292-297.

[3] Ramalingam V V, Dandapath A, Raja M K. Heart disease prediction using machine learning techniques: a survey[J]. International Journal of Engineering & Technology, 2018, 7(2.8): 684-687.

[4] Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques[J]. IEEE access, 2019, 7: 81542-81554.

[5] Khan M A. An IoT framework for heart disease prediction based on MDCNN classifier[J]. IEEE Access, 2020, 8: 34717-34727.

[6] Fitriyani N L, Syafrudin M, Alfian G, et al. HDPM: an effective heart disease prediction model for a clinical decision support system[J]. IEEE Access, 2020, 8: 133034-133050.

[7] Battineni G, Chintalapudi N, Amenta F. Forecasting of COVID-19 epidemic size in four high hitting nations (USA, Brazil, India and Russia) by Fb-Prophet machine learning model[J]. Applied Computing and Informatics, 2020.

[8] Stewart S, Keates A K, Redfern A, et al. Seasonal variations in cardiovascular disease[J]. Nature Reviews Cardiology, 2017, 14(11): 654-664.

[9] National Center for Health Statistics. Weekly counts of deaths by state and select causes, 2014-2019[J]. 2022.

[10] Taylor S J, Letham B. Forecasting at scale[J]. The American Statistician, 2018, 72(1): 37-45.

[11] NOAA National Centers for Environmental Information. Climate at a glance: Statewide time series[J]. 2022.