Predicting Heart Failure Death Events

Dante Miller

May 10, 2021

Abstract

The purpose of this study is to determine if heart failure deaths can be detected accurately by multiple linear regression, boosted trees, and k-nearest neighbors models by examining the relationship of heart failure death events and non-death events with factors that are said to contribute to heart failures.

Introduction

Heart failure is a chronic condition in which the heart does not pump blood as well as it should. Smoking, physical inactivity, nutrition, obesity, high cholesterol, diabetes, and high blood pressure are all factors relating to heart failure. Some of these select factors are used in this study to understand their relationship to heart failure death events and non-death events. This study is important in understanding if heart failure deaths can be predicted accurately by multiple linear regression, boosted trees, and k-nearest neighbors models. It would allow us to determine if heart failure deaths can be predicted by these select model before it happens. Further research can be done on exploring other machine learning models along with improving the models used in this study by adding, deleting, or manipulating the data.

Method

The purpose of this project is to determine if heart failure deaths can be detected before they happen by multiple linear regression, boosted trees, and k-nearest neighbors models.

There are three different methods that are used to answer this question, which are the multiple linear regression model, boosted trees model, and k-nearest neighbors model. Each model splits the dataset into a training and testing set. The training set is used to create the model and the testing set is used to qualify the performance of the model.

Multiple Linear Regresssion

The multiple linear regression model uses the multiple regression formula, which predicts the outcome using independent variables.

The multiple regression formula is $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$

$\hat{Y}$ is the dependent variable of the regression. β_0 is the value at which the regression line crosses the y-axis. $\beta_1 X_1$ is the slope of the regression multiplied by the first independent variable of the regression. $\beta_2 X_2$ is the slope of the regression multiplied by the second independent variable of the regression. $\beta_n X_n$ is based on the amount of independent variables n that will be used in the formula.

The multiple regression model generates a predicted and actual value for the testing dataset along with evaluation metrics, mean error, root mean squared error, and mean absolute error, which can be used to evaluate the model and determine if heart failure deaths can be detected before it happens by a multiple linear regression model.

K-Nearest Neighbors

The k-nearest neighbors model uses the euclidean distance to classify data points based on the most similar points.

The euclidean distance formula is

$$\sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

n is the amount of euclidean vectors. i is the initial point for the formula. y_i, x_i are the euclidean vectors starting from the initial point. The formula goes through all the points until it reaches n.

The k-nearest neighbors model generates a confusing matrix, classification summary, classification report, and a knn score to evaluate the model and determine if heart failure deaths can be detected before it happens by a k-nearest neighbors model.

Boosted Trees

The Gradient Boosted Decision Trees

The gradient boosted decision trees model is an approach that fits the data using multiple simpler models, or so called base learners/weak learners. The approach optimizes the loss of the model by updating weights, which lessens the difference between the predicted value and actual value.

The gradient boosted decision trees model generates a confusing matrix, classification summary, classification report, and a gbc score to evaluate the model and determine if heart failure deaths can be detected before it happens by a gradient boosted decision trees model.

Extreme Gradient Boosting Decision Trees

Extreme gradient boosting decision trees model is an approach that builds off gradient boosted decision trees approach by improving its base framework through systems optimization and algorithmic enhancements. The extreme gradient boosted decision trees approach has better efficiency, computational speed, and model performance compared to the gradient boosted decision trees approach.

Extreme gradient boosting decision trees model generates a confusing matrix, classification summary, classification report, and a xgb score to evaluate the model and determine if heart failure deaths can be detected before it happens by a extreme gradient boosting decision trees model.

Data Analysis

Study Design and Data Collection

The data used to determine if heart failure deaths can be detected accurately by multiple linear regression, boosted trees, and k-nearest neighbors models is attained from kaggle. The variables in the dataset are age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, and death event. For this project, the variables are not manipulated. The death event variable is determined to be the outcome and the other variables are determined to be the predictors. There are thirteen variables and 299 records in the dataset. The dataset contains no null values and the variables are either numeric (data types in the dataset are float and integer) or a numeric dummy variable.

In Figure 1, the dataset death event variable is either zero (no death) or one (death). There are less non-death than there are deaths.

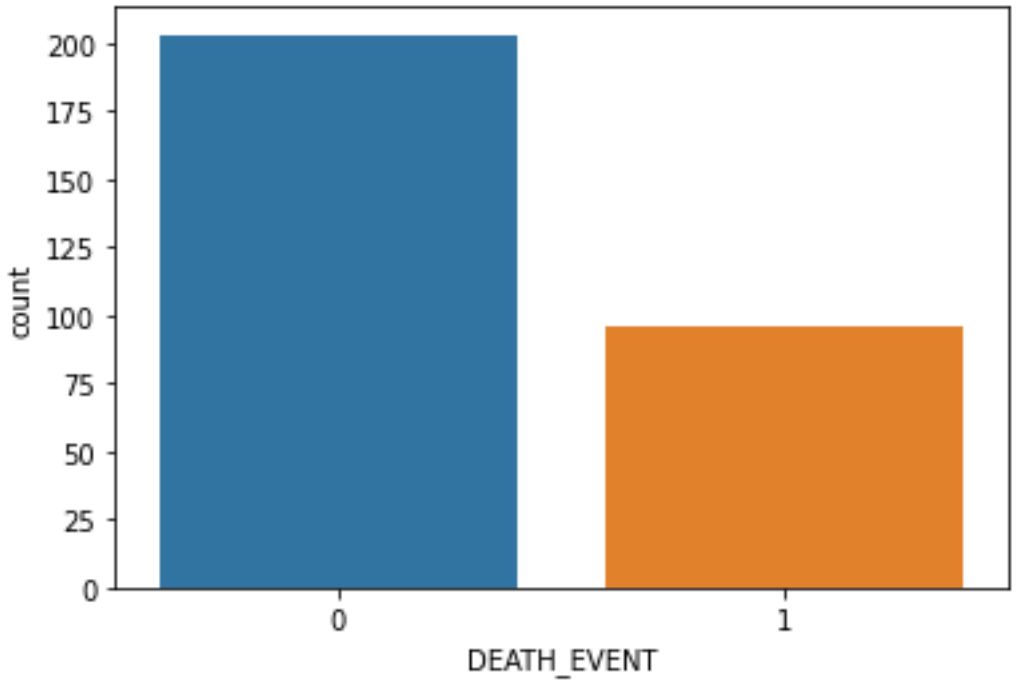


Figure 1: Death and Non-Death Counts

In Figure 2, there is a mild correlation between age and serum creatinine with death event. Death event has a negative correlation with time, serum sodium, and ejection factor. Death event has no correlation with anaemia, creatinine phosphokinase, diabetes, high blood pressure, platelets, sex, and smoking.

Figure 2: Correlation Matrix

Multiple Linear Regresssion

In Figure 3, the actual heart failure death event values are plotted alongside the values the model predicted death event to be. The actual values and predicted values do overlap.

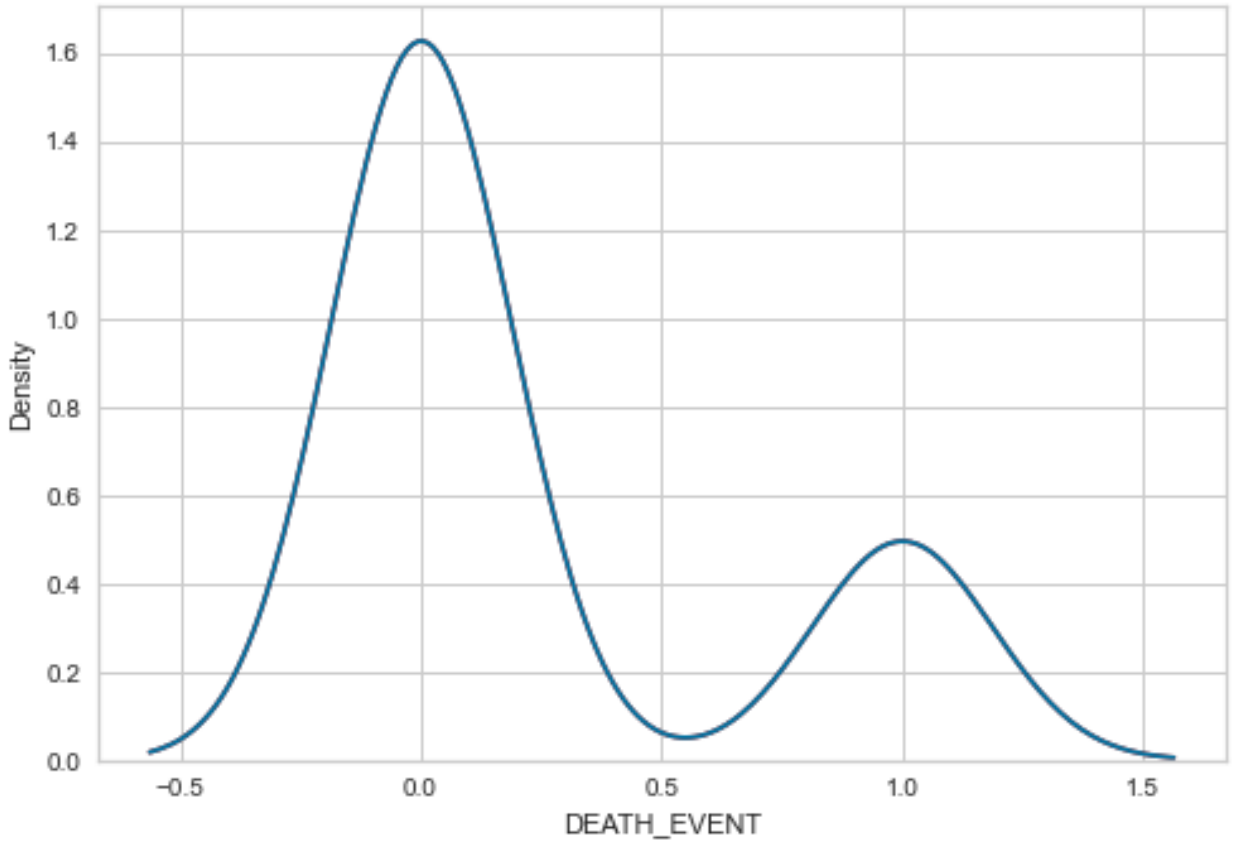
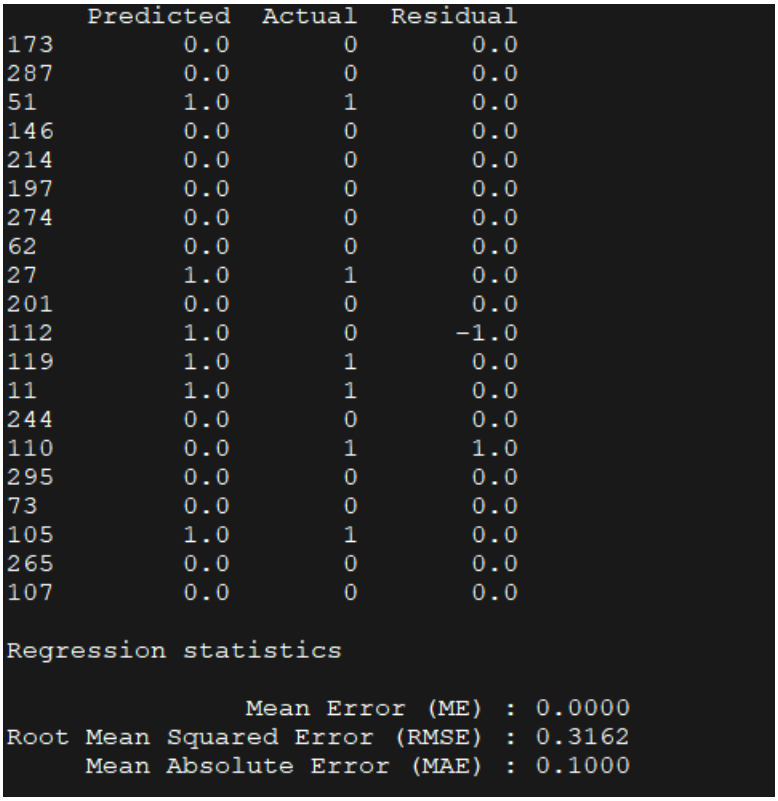


Figure 3: Heart Failure Death Event Actual Values vs Predicted Values

In Figure 4, the mean error, root mean squared error, and mean absolute error are acceptable because they are not large values. When looking at the predicted values vs actual values, the actual values are either one or zero and the predicted values seem to be either one or zero. The actual values and predicted values seem to match for the most part.



	Predicted	Actual	Residual
173	0.0	0	0.0
287	0.0	0	0.0
51	1.0	1	0.0
146	0.0	0	0.0
214	0.0	0	0.0
197	0.0	0	0.0
274	0.0	0	0.0
62	0.0	0	0.0
27	1.0	1	0.0
201	0.0	0	0.0
112	1.0	0	-1.0
119	1.0	1	0.0
11	1.0	1	0.0
244	0.0	0	0.0
110	0.0	1	1.0
295	0.0	0	0.0
73	0.0	0	0.0
105	1.0	1	0.0
265	0.0	0	0.0
107	0.0	0	0.0

Regression statistics

Mean Error (ME)	0.0000
Root Mean Squared Error (RMSE)	0.3162
Mean Absolute Error (MAE)	0.1000

Figure 4: Multiple Linear Regression Statistics

K-Nearest Neighbors

In Figure 5, the confusion matrix has an accuracy of 0.71 percent. 127 values are predicted to be no death events and are actually no death events. 21 values are predicted to be death events and are actually death events. 15 values are predicted to be death events, but are actually no death events. 37 values are predicted to be no death events, but are actually death events. The precision in the classification report is 0.74 percent for no death and 0.58 percent for death. The accuracy is 0.69.

```
Confusion Matrix:
[[107  15]
 [ 37  21]]
Classification Summary
Confusion Matrix (Accuracy 0.7111)

       Prediction
Actual   0   1
     0 107  15
     1  37  21
None
Classification Report
              precision    recall  f1-score   support

           0       0.74      0.88      0.80       122
           1       0.58      0.36      0.45        58

    accuracy                           0.71       180
   macro avg       0.66      0.62      0.63       180
weighted avg       0.69      0.71      0.69       180

Score
0.7111111111111111
```

Figure 5: K-Nearest Neighbors Summary

Boosted Trees

Gradient Boosted Trees

In Figure 6, the confusion matrix has an accuracy of 0.8333 percent. 74 values are predicted to be no death events and are actually no death events. 26 values are predicted to be death events and are actually death events. 11 values are predicted to be death events, but are actually no death events. 9 values are predicted to be no death events, but are actually death events. The precision in the classification report is 0.89 percent for no death events and 0.70 percent for death events. The accuracy is 0.84.

```
Confusion Matrix:
[[74 11]
 [ 9 26]]
Classification Summary
Confusion Matrix (Accuracy 0.8333)

       Prediction
Actual  0  1
     0 74 11
     1  9 26
None
Classification Report
              precision    recall  f1-score   support

           0       0.89      0.87      0.88        85
           1       0.70      0.74      0.72        35

    accuracy                           0.83       120
   macro avg       0.80      0.81      0.80       120
weighted avg       0.84      0.83      0.83       120

Score
0.8333333333333334
```

Figure 6: Gradient Boosted Trees Summary

Extreme Boosted Trees

In Figure 7, the confusion matrix has an accuracy of 0.8667 percent. 74 values are predicted to be no death events and are no actually no death events. 30 values are predicted to be death events and are actually death events. 11 values are predicted to be death events, but are actually no death events. 5 values are predicted to be no death events, but are actually death events. The precision in the classification report is 0.94 percent for no death events and 0.73 percent for death events. The accuracy is 0.88.

```
Confusion Matrix:
[[74 11]
 [ 5 30]]
Classification Summary
Confusion Matrix (Accuracy 0.8667)

       Prediction
Actual  0  1
     0 74 11
     1  5 30
None
Classification Report
              precision    recall  f1-score   support

           0       0.94      0.87      0.90        85
           1       0.73      0.86      0.79        35

    accuracy                           0.87       120
   macro avg       0.83      0.86      0.85       120
weighted avg       0.88      0.87      0.87       120

Score
0.8666666666666667
```

Figure 7: Extreme Boosted Trees Summary

Conclusions

Multiple Linear Regression Model Conclusion

The mean error, root mean squared error, and mean absolute error are acceptable because they are not large value. The predicted values and actual values overlap for the most part. These results suggest that the model can predict death events and non-death events accurately. The issue with this model might be the lack of data. The model was trained and tested using a dataset that might not be an appropriate sample size for the population this study is looking at. It does answer the question on whether heart failure deaths can be detected accurately by a multiple linear regression model. Heart failure death events can be detected accurately by a multiple linear regression model. Can the model be improve by manipulating the data, adding variables, or deleting variables? Would a model that is trained and tested on a larger dataset detect heart failure deaths accurately? These are two new questions brought up after creating the multiple linear regression model.

K-Nearest Neighbors Model Conclusion

The accuracy of 0.7111 percent and the precision of 0.74 percent for non death events and 0.58 percent for death events are not acceptable for this model. These results suggest that the model can not predict death events and non-death events accurately. The issue with this model might be due to keeping all the variables. It does answer the question on whether heart failure deaths can be detected accurately by a k-nearest neighbors model. Heart failure deaths can not be detected accurately by a k-nearest neighbors model. Would the deletion or addition of variables improve the model? This is the new question brought up after creating the k-nearest neighbors model. Another problem is the model being trained and tested using a dataset that might not be an appropriate sample size for the population this study is looking at.

Gradient Boosted Trees Model Conclusion

The accuracy of 0.8333 percent and the precision of 0.89 percent for non death events and 0.70 percent for death events are acceptable for this model. These results suggest that the model can not predict death events accurately, but can predict non-death events accurately. The issue with this model might be due to keeping all the variables. It does answer the question on whether heart failure deaths can be detected accurately by a gradient boosted trees model. Heart failure deaths can not be detected accurately by a gradient boosted trees model. Would the deletion or addition of variables improve the model? This is the new question brought up after creating the gradient boosted trees model. Another problem is the model being trained and tested using a dataset that might not be an appropriate sample size for the population this study is looking at.

Extreme Boosted Trees Model Conclusion

The accuracy of 0.8667 percent and the precision of 0.94 percent for non death events and 0.73 percent for death events are acceptable for this model. These results suggest that the model can not predict death events accurately, but can predict non-death events accurately. The issue with this model might be due to keeping all the variables. It does answer the question on whether heart failure deaths can be detected accurately by a extreme boosted trees model. Heart failure deaths can not be detected accurately by a extreme boosted trees model. Would the deletion or addition of variables improve the model? This is the new question brought up after creating the extreme boosted trees model. Another problem is the model being trained and tested using a dataset that might not be an appropriate sample size for the population this study is looking at.

References

Heart Failure Prediction. https://kaggle.com/andrewmvd/heart-failure-clinical-data. Accessed 11 May 2021.

Kiernan, Diane. "Chapter 8: Multiple Linear Regression." Natural Resources Biometrics, Open SUNY Textbooks, 2014. milnepublishing.geneseo.edu, https://milnepublishing.geneseo.edu/natural-resources-biometrics/chapter/chapter-8-multiple-linear-regression/.

Seif, George. "A Beginner's Guide to XGBoost." Medium, 14 Feb. 2021, https://towardsdatascience.com/a-beginners-guide-to-xgboost-87f5d4c30ed7.

Subramanian, Dhilip. "A Simple Introduction to K-Nearest Neighbors Algorithm." Medium, 3 Jan. 2020, https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e.

Yıldırım, Soner. "Gradient Boosted Decision Trees-Explained." Medium, 17 Feb. 2020, https://towardsdatascience.com/gradient-boosted-decision-trees-explained-9259bd8205af.