Mean Arterial Pressure and Heart Rate Prediction

Lexie Sun, Todd Zhang

1. Abstract

Arterial Pressure and Heart Rate are accurate reflections of the cardiovascular condition of a person. Therefore, being able to predict those two metrics can be crucial in the pre-intervention of cardiovascular disease.

In this paper, our team used machine learning algorithms to predict arterial pressure and heart rate of patients based on their body measures and various other factors. We were able to achieve a combined R-Squared score of 0.90 when comparing our prediction to the actual results.

2. Team

Our team name on Kaggle is Toxic and team members are Qingyi(Lexie) Sun (Kaggle ID: 1910766), and Xintao(Todd) Zhang (Kaggle ID: 4020474).

3. Dataset Description and EDA

The dataset records personal information such as gender and age and various unspecified information of 2321 patients. A large portion of those patients have multiple records on file. The sizes of the training and testing datasets are 129 MB and 22.6 MB respectively.

The average age of patients was 64.8 years old (See appendix fig. 1) and the ratio between male and female patients is 3:2 (fig. 2). The distribution of mean arterial pressure is right-skewed (fig. 3) whereas the distribution of heart rate is roughly normally distributed with a mean of 83.4 and standard deviation of 14.3 (fig. 4)

4. Feature engineering and machine learning methods:

After our exploratory data analysis and initial model fitting, we discovered that the target variable 'Mean Heart Rate' is highly correlated with the unspecified feature 'xx5' whereas no strong correlation was found between 'xx5' and 'Mean Arterial Pressure'. Therefore we applied different preprocessing methods to the dataset when analyzing the two different target variables, essentially creating two copies of the original datasets for each variable.

We also discovered that the number of records differs from patients to patients. Some patients have hundreds of records whereas others have less than 10. To prevent the models from fitting too heavily on the patients that have more records, we used downsampling techniques in the feature engineering process when analyzing both target variables.

Y_mean_hr (Mean Heart Rate)

Feature engineering:

For the target 'y_mean_hr', we grouped the dataset by the feature 'key' which indicates the number of visits for each patient and then took the mean of all the remaining features. This action reduced the size of the training set from 1348470 records to 44949 records.

Model Selection:

Linear Regression

Since the target 'y_mean_hr' is highly related to some features, the linear regression model performs well on 'y_mean_hr'. We randomly selected 80% of the dataset as the training dataset and reserved 20% of the dataset as the test dataset. The R Squared score was 0.939 on the test dataset.

*Random Forest**

We applied XGBoost algorithm on the target variable and randomly selected 80% of the dataset as the training dataset and reserved 20% of the dataset as the test dataset. The R Squared score was 0.945 on the test dataset, which makes random forest the best performing model in predicting 'Y_mean_hr'.

Y_mean_map (Mean Arterial Pressure)

Feature Engineering:

For the target 'y_mean_map', we again grouped the dataset by the feature 'key', and then took the value of the most recent record of all the features. We also found out that some patients have multiple 'key's, which means that even after data aggregation, the records of some patients still make up a large proportion of the training dataset. We decided to take the most recent 10 records of each patient to arrive at a more balanced dataset (if the patient has less than 10 records, we kept them all).

We also took the log value of 'Y-mean-map' as the target variable when we trained our models since it displays a right-skewed distribution.

Model Selection:

Neural Networks

We applied different configurations of neural networks on the target variable and randomly selected 80% of the dataset as the training dataset and reserved 20% of the dataset as the validation dataset. The highest R Squared score was 0.915 on the validation dataset.

XGBoost

We applied XGBoost algorithm on the target variable and randomly selected 80% of the dataset as the training dataset and reserved 20% of the dataset as the validation dataset. The R Squared score was 0.895 on the validation dataset.

*Random Forest**

We also applied the Random Forest model on the target variable and used a random search CV to find the best model parameters.We randomly selected 80% of the dataset as the training dataset and reserved 20% of the dataset as the test dataset. The R Squared score was 0.906 on the validation dataset, which makes random forest the best performing model in predicting 'Y_mean_MAP'.

5. Experimental results

Since there are two distinct target variables in our analysis, the experiment results are split into two parts.

Y_mean_hr (Mean Heart Rate)

The best performing model in predicting Mean Heart Rate is random forest. The R Squared score was 0.945.

Y_mean_map (Mean Arterial Pressure)

The best performing model in predicting Mean Arterial Pressure is also random forest. The R Squared score was 0.906.

6. Github repo:

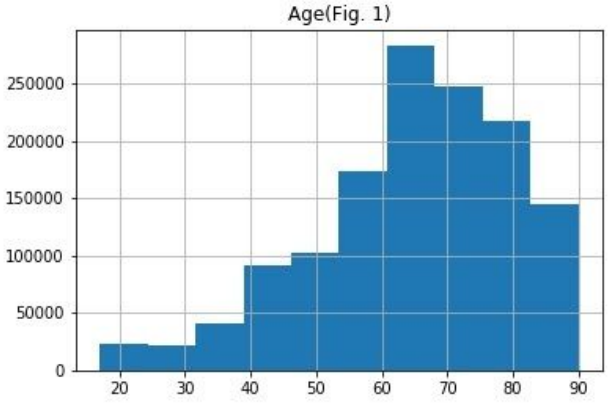
https://github.com/USF-ML2/final-project-toxic

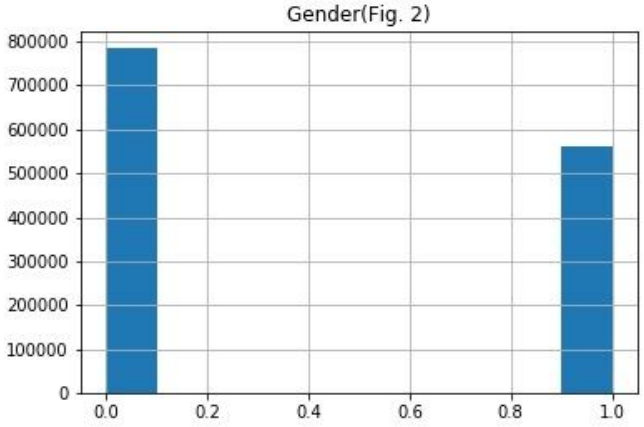
7. Responsibility:

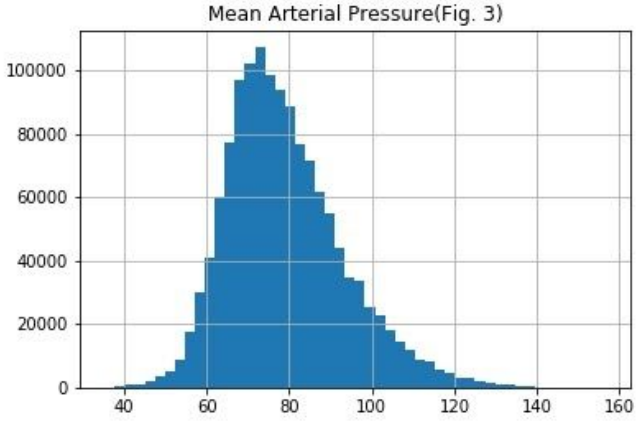
Qingyi Sun: Feature engineering and model selection

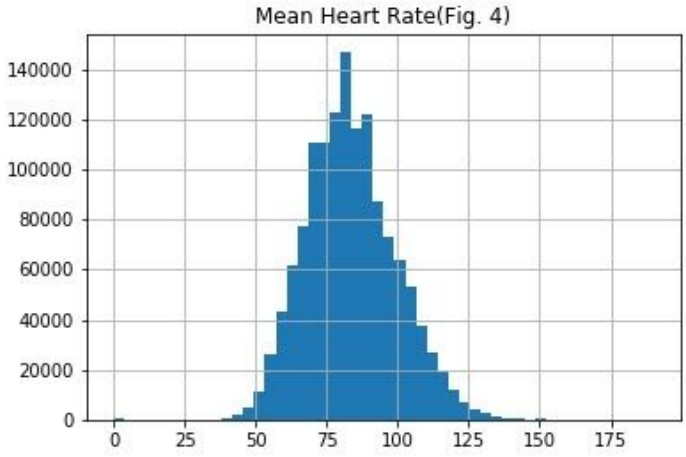
Todd Zhang: EDA, model selection and report.

8. Appendix


Age(Fig. 1)
250000
200000
150000
100000
50000
0
20
30
40
50
60
70
80
90


Gender(Fig. 2)
800000
700000
600000
500000
400000
300000
200000
100000
0
0.0
0.2
0.4
0.6
0.8
1.0


Mean Arterial Pressure(Fig. 3)
100000
80000
60000
40000
20000
0
40
60
80
100
120
140
160


Mean Heart Rate(Fig. 4)
140000
120000
100000
80000
60000
40000
20000
0
0
25
50
75
100
125
150
175