Capstone Project – The Machine Learning Model of Heart Failure Prediction

Alec Xu
29 Aug 2020

1. Introduction of Background

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Heart failure is a common event caused by CVDs and it means the heart is unable to pump sufficiently to maintain blood flow to meet the body's needs.

Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management wherein a machine learning model can be of great help. That is the starting point that I will use different modelling methods to estimate the risk that a person may get heart failure as an effective reference to support the media examination of hospital.

The target audience of this modelling studying is not just the medical researchers or doctors, but also the patients who have strong concerns on their health threatened by cardiovascular diseases to raise their awareness of some negative signals of heart failure.

2. Data Cleaning and Splitting

This dataset contains 12 features that can be used to predict mortality by heart failure is obtained from Kaggle with the link below.
https://www.kaggle.com/andrewmvd/heart-failure-clinical-data?select=heart_failure_clinical_records_dataset.csv

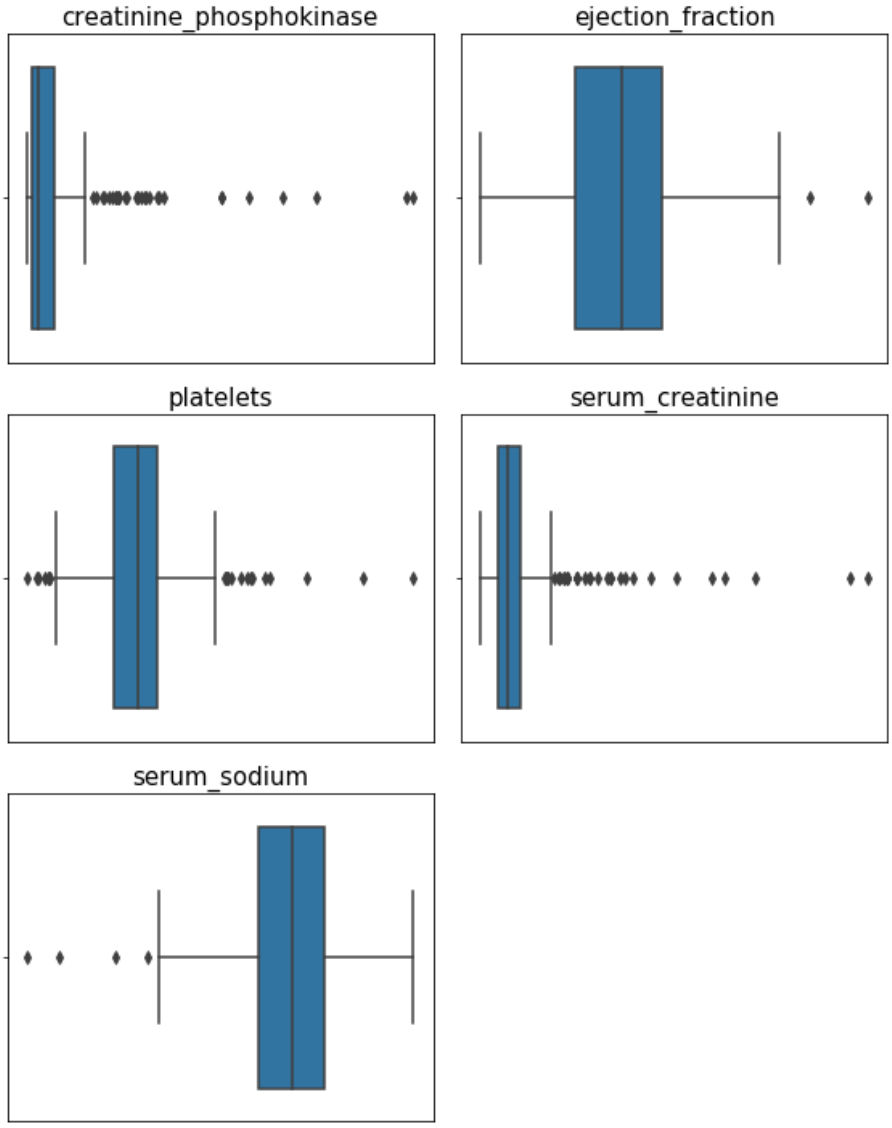
The 12 attributes are age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, death event. And the total sample scope is 299.

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time
0	75.0	0	582	0	20	1	265000.00	1.9	130	1	0	4
1	55.0	0	7861	0	38	0	263358.03	1.1	136	1	0	6
2	65.0	0	146	0	20	0	162000.00	1.3	129	1	1	7
3	50.0	1	111	0	20	0	210000.00	1.9	137	1	0	7
4	65.0	1	160	1	20	0	327000.00	2.7	116	0	0	8

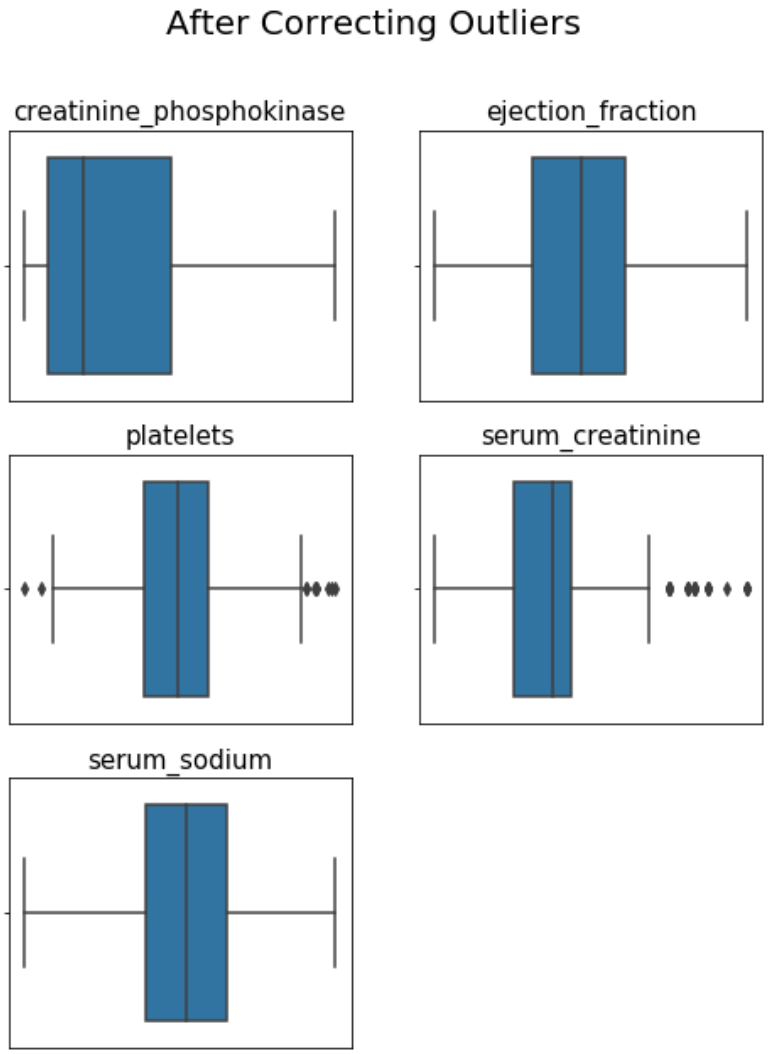
After loading data, firstly I will check the data types of each column and also whether there

are null values in the dataset to ensure there is no blank entry.

Then the sample distribution should be plotted to see any factors that having outliers which will affect the accuracy of the machine learning models to be built. Here I only choose five categories, which are 'creatinine_phosphokinase', 'ejection_fraction', 'platelets', 'serum_creatinine', 'serum_sodium' because the values of the other categories are all reasonable considering their natures.



Obviously, the plots above show outliers existing in those selected columns of categories. So I identify the data point which is not within the range of [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR] as outliers and used mean value of the corresponding features to replace their values. Now the data set has been processed to be ready for modelling.



<Figure size 432x288 with 0 Axes>

After that, I will drop the death event column, which means the person had heart failure (0) or not (1), from the dataset and treat it as y, the dependent variable, while the 12 factors will be independent variables X. I also normalize the X values and split all the 299 data referring to the ratio of 7:3 as training data and test data which can be applied to the machine learning in the next step.

3. Methodology

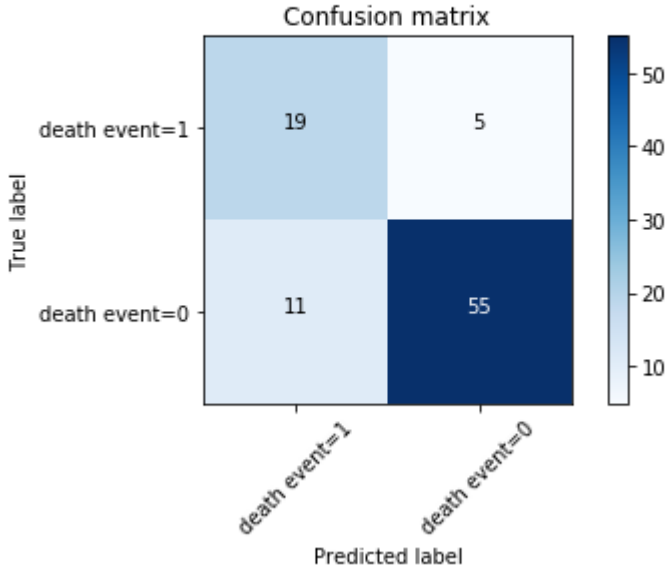
To build a proper prediction model for heart failure event, I will use three modelling methods one by one which are logistic regression, SVM and decision tree in order to compare their model accuracy to obtain the most appropriate one.

3.1 Logistic Regression Model

I run the logistic regression model with training set X_train & y_train and then enter X_test in test set to get the predicted yhat. To evaluate the model, confusion matrix is built to compare

the y_test with yhat. From the matrix it is easy to find that 74 out of 90 prediction can match the test. And the prediction of death event 0 has a very high rate of precision which is 92%.

```
Confusion matrix, without normalization
[[19  5]
 [11 55]]
```




```
# Calculate f1 score
print (classification_report(y_test, yhat))
```

	precision	recall	f1-score	support
0	0.92	0.83	0.87	66
1	0.63	0.79	0.70	24
micro avg	0.82	0.82	0.82	90
macro avg	0.77	0.81	0.79	90
weighted avg	0.84	0.82	0.83	90

The f1 score and Jaccard Index are also calculated below for evaluation reference. Both scores are around 0.82, meaning that this model has high reliance.

```
# Calculate f1 score and jaccard index for logistic regression
from sklearn.metrics import f1_score
f1_score(y_test, yhat, average='weighted')
```

```
0.8278659611992946
```

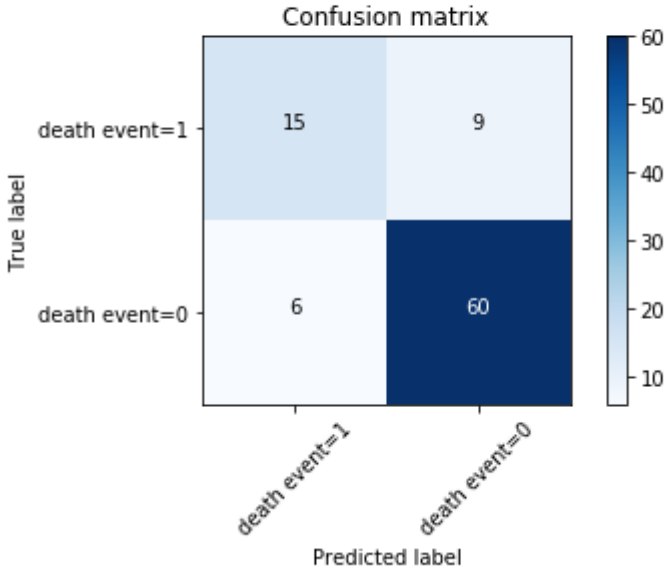
```
from sklearn.metrics import jaccard_similarity_score
jaccard_similarity_score(y_test, yhat)
```

```
0.8222222222222222
```

3.2 SVM Model

Next I follow the same flow to build the SVM model. In this model's confusion matrix, 75 out of 90 prediction can match the test and the precision accuracy of predicting death event 1 has increased to 71%.

```
Confusion matrix, without normalization
[[15  9]
 [ 6 60]]
```



```
# Calculate f1 score
print (classification_report(y_test, ypred))
```

```
              precision    recall  f1-score   support

           0       0.87      0.91      0.89        66
           1       0.71      0.62      0.67        24

   micro avg       0.83      0.83      0.83        90
   macro avg       0.79      0.77      0.78        90
weighted avg       0.83      0.83      0.83        90
```

The f1 score and Jaccard Index are also calculated below for evaluation reference. Both scores are around 0.83, meaning that this model is also reliable.

```
# Calculate f1 score and jaccard index for SVM
from sklearn.metrics import f1_score
f1_score(y_test, ypred, average='weighted')
```

```
0.8296296296296297
```

```
from sklearn.metrics import jaccard_similarity_score
jaccard_similarity_score(y_test, ypred)
```

```
0.8333333333333334
```

3.3 Decision Tree Model

The last model is decision tree which I choose depth of 15 and fit the model with training set.

```
# Decision Tree Model
from sklearn.tree import DecisionTreeClassifier
HFTree = DecisionTreeClassifier(criterion="entropy", max_depth = 15)
HFTree
```

```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=15,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
```

After getting the predicted values of y, I directly calculate the accuracy score which is 0.81.

```
# Evaluation of decision tree model
from sklearn import metrics
import matplotlib.pyplot as plt
print("DecisionTrees's Accuracy: ", metrics.accuracy_score(y_test, predTree))
```

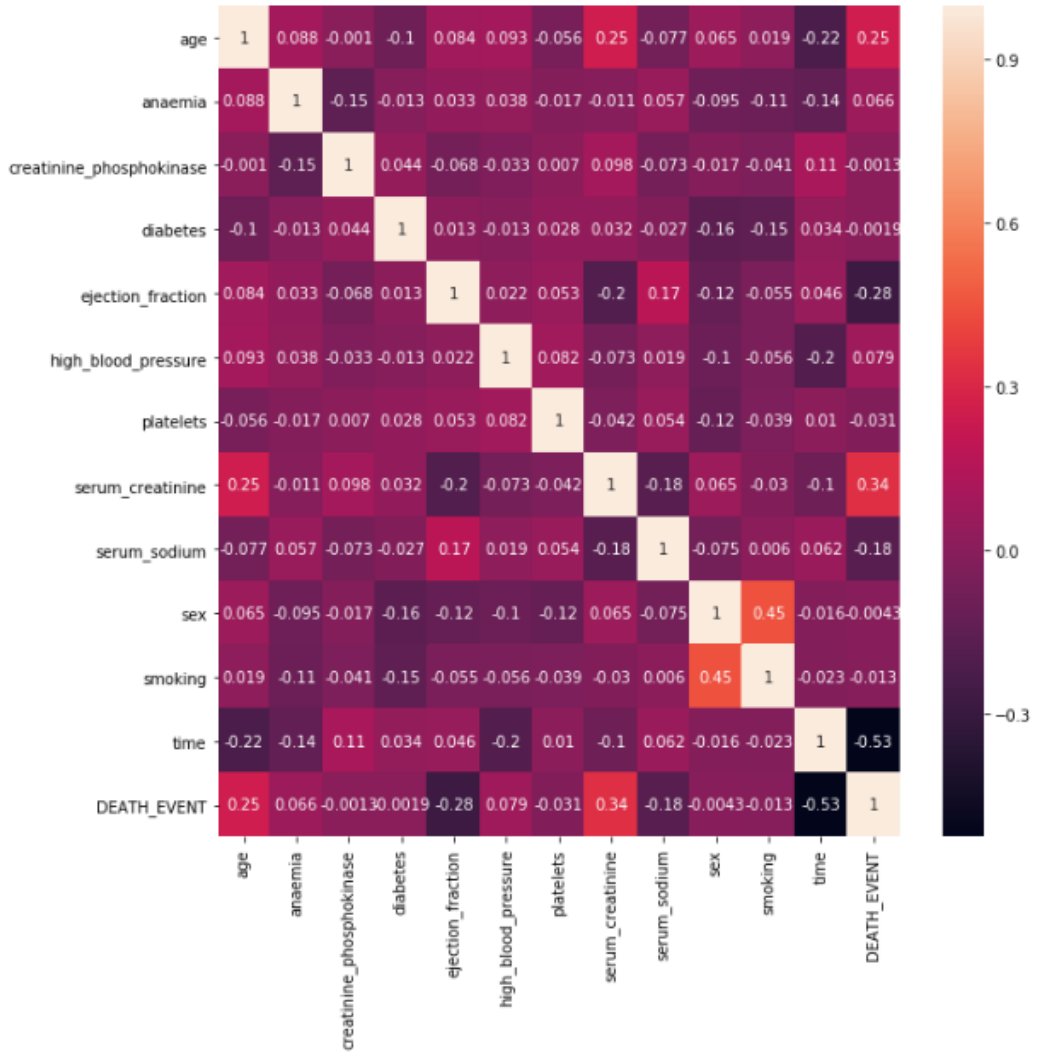
```
DecisionTrees's Accuracy:  0.8111111111111111
```

4. Result

According to the accuracy of three model, the SVM model attains the highest score 0.83 which is slightly higher than the scores of logistic regression model 0.82 and decision tree model 0.81. Therefore, in this project SVM model shows the best prediction of heart failure event.

```
# Check correlation between factors
plt.figure(figsize = (10, 10))
sns.heatmap(df.corr(), annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2932623a9e8>
```



And I also checked the correlations among all factors and from the chart above four attributes, age, anaemia, high blood pressure and serum creatinine have more significant positive influence on the decrease of patient with heart failure.

5. Discussion

This project still has some limitation that can be further studied:

- It is unknown that whether other models like random forest can have a better prediction performance than three models I used.
- The 12 attributes included in this project may not be fully cover all the important factors that will impact on the heart failure.
- The model accuracy may be higher if some of the insignificant factors can be removed from the data sets.

6. Conclusion

In this capstone project, I tested three machine learning models to predict whether a patient would suffer the risk of heart failure. And the SVM model stands out to be the most reliable model that worth further study by medical researcher and if doctors collected those 12 factors of patients, they may enter those data into the SVM model to forecast their heart failure risks and prepare more careful nursing to those patients predicted to be burdened with the cardiovascular disease.