INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 4 - V7I4-1655)
Available online at: https://www.ijariit.com

Survival analysis of heart failure patients using Machine Learning on an imbalanced dataset

Mohammed Mafaz
mafazazhaar@gmail.com
Loyola College, Chennai, Tamil Nadu

ABSTRACT

In this paper, we have focused on survival analysis of heart failure patients. The number of individuals diagnosed with coronary failure is increasing and projected to rise by 46 percent by 2030, leading to quite 8 million people with coronary failure. The reasons for increase in heart failure is due to increase in the number of cases involving high blood pressure, valve disease, thyroid disease, kidney disease and diabetes [1]. With the growth of machine learning, data mining, statistical analysis, data-modelling predicting whether the person will survive [2] or not after the heart failure is possible and it becomes very crucial.

Keywords— *SMOTE, Random Forest, Confusion Matrix*

1. INTRODUCTION

The dataset chosen for analysis is secondary in nature which contain 299 records of patients, in 13 columns (Including the dependent variable). The dataset has no missing values but then it is highly imbalanced. The data has been taken from UCI Machine learning repository.

Variable information:

- age: age of the patient (years)
- anemia: decrease of red blood cells or hemoglobin (Boolean)
- high blood pressure: if the patient has hypertension (Boolean)
- creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
- diabetes: if the patient has diabetes (Boolean)
- ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
- platelets: platelets in the blood (kilo platelets/mL)
- sex: woman or man (binary)
- serum creatinine: level of serum creatinine in the blood (mg/dL)
- serum sodium: level of serum sodium in the blood (mEq/L)
- smoking: if the patient smokes or not (Boolean)
- time: follow-up period (days)
- [target] death event: if the patient deceased during the follow-up period (Boolean)

2. OBJECTIVES AND METHODOLOGY

The objective of this paper is to do survival analysis of heart failure patients using machine learning algorithm. The dataset used in the survival analysis of heart failure patients for this paper is slightly imbalanced. Only 32% of the values from the total dataset fall under survived category. Since the dataset is highly imbalanced, we make use of a method which is called ***SMOTE***. It s one of the most commonly used oversampling methods to solve the imbalance problem. In this paper, we use a classification algorithm where the entire dataset is made use of, excluding the outliers. With this classification algorithm, we build a model that gives us the confusion matrix. This classification algorithm is called ***Random Forest***.

A. SMOTE

SMOTE [3] stands for synthetic minority oversampling technique. is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesizes new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class. After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data.

B. Random Forest

Random forest, like its name implies, consists of an outsized number of individual decision trees that operate as an ensemble. Each individual tree within the random forest spits out a category prediction and therefore the class with the foremost votes becomes our model's prediction. an outsized number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

Random forest [4] in data analytics returns a confusion matrix. This confusion matrix will give the accuracy of the classification model. In Random Forest, studies with small to moderate sample size overestimate the effect measure. Thus, large sample sizes are required for Random Forest to supply enough numbers in both categories of the variable.

C. Confusion Matrix

In classification analysis, a table of confusion (sometimes also called a confusion matrix) [5], may be a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives. this enables more detailed analysis than mere proportion of correct classifications(accuracy).

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 3.1 Confusion Matrix

3. ANALYTICS

The dataset used for analysis has 299 observations and 13 variables with reference to Figure 4.1.

```
In [43]: heart.shape
Out[43]: (299, 13)
```

Figure 4.1 The number of observations in the dataset

In Figure 4.2, we get the first five observations of our dataset.

```
In [45]: print(heart.head())
    age  anaemia  creatinine_phosphokinase  diabetes  ejection_fraction  \
0  75.0        0                  6.366470         0                 20
1  55.0        0                  8.969669         0                 38
2  65.0        0                  4.983607         0                 20
3  50.0        1                  4.709530         0                 20
4  65.0        1                  5.075174         1                 20

   high_blood_pressure  platelets  serum_creatinine  serum_sodium  sex  \
0                    1  12.487485          0.641854      4.867534    1
1                    0  12.481270          0.095310      4.912655    1
2                    0  11.995352          0.262364      4.859812    1
3                    0  12.254863          0.641854      4.919981    1
4                    0  12.697715          0.993252      4.753590    0

   smoking  time  DEATH_EVENT
0        0     4            1
1        0     6            1
2        1     7            1
3        0     7            1
4        0     8            1
```

Figure 4.2 The First five observations of the dataset

Now, we check the correlation between the variables using the heatmap. (Figure.4.3)

```
In [15]: sns.heatmap(heart.corr())
Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x2a1a1c6dd00>
```

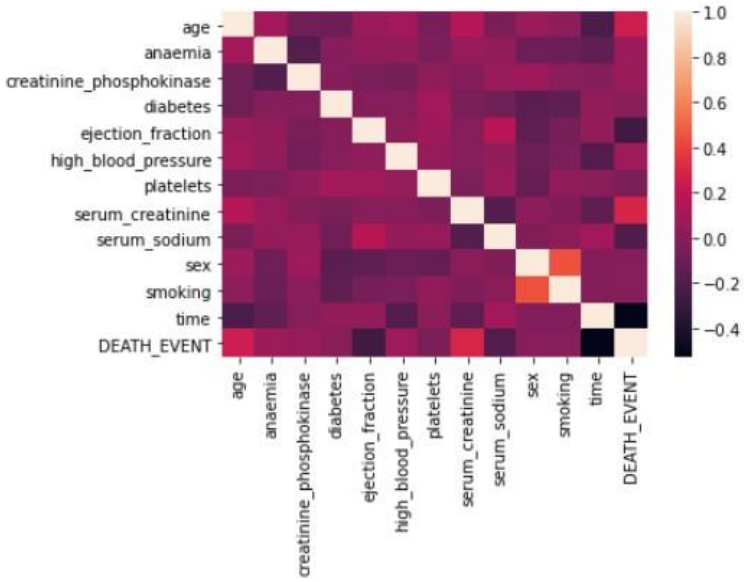

Figure 4.3 Correlation Heatmap

Then, we check if the variables have outliers using boxplot. If the variables have outliers, we rectify the outliers using inter quartile range concept. We also check if the variables are skewed, if they are skewed, we normalize the variables by log or reciprocal transformation. From the Figure 4.4, we observe that only 32% cases are of survived patients and rest all the cases are of people who died.. Hence, our dataset is slightly imbalanced.

```
In [21]: sns.countplot(y)
Out[21]: <matplotlib.axes._subplots.AxesSubplot at 0x2a1a06eb700>
```

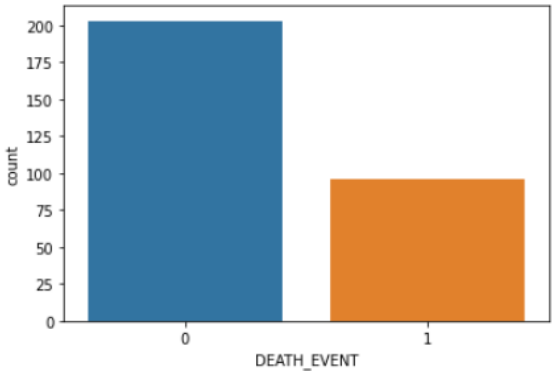

Figure 4.4 Death Vs Survived (0 Vs 1)

To make the dataset balanced, we apply the SMOTE method in Figure 4.5.

```
In [22]: from imblearn.over_sampling import SMOTE
         samp = SMOTE()
         x_res1,y_res1=samp.fit_sample(X,y)
         sns.countplot(y_res1)
Out[22]: <matplotlib.axes._subplots.AxesSubplot at 0x2a1a0798850>
```

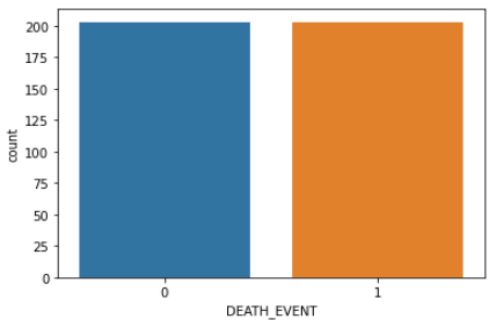

Figure 4.5 Count plot after SMOTE

Now, we fit the classification algorithm, Random Forest on our dataset and we use 100 decision trees and have given the maximum depth of each decision tree as 40.

Random Forest

```
from sklearn.ensemble import RandomForestClassifier
RFC=RandomForestClassifier(n_estimators=100,max_depth=40,criterion='entropy',random_state=0)
RFC.fit(x_train,y_train)
ypred=RFC.predict(x_test)
```

Figure 4.6 Random Forest

The overall accuracy of the model is 86% and since the accuracy for death class is 81% and the accuracy of survived class is 89.9% as per Figure 4.7, we have achieved a good prediction model for survival analysis of heart failure patients.

```
In [62]: cm=confusion_matrix(y_test,ypred)
         print(cm)
         print('Overall accuracy score is ',accuracy_score(y_test,ypred))
         print('Accurcay for death class (0) is ' , cm[0,0]/(cm[0,0]+cm[0,1]))
         print('Accuracy for survived class (1) is ' , cm[1,1]/(cm[1,0]+cm[1,1]))

         [[43 10]
          [ 7 62]]
         Overall accuracy score is  0.860655737704918
         Accurcay for death class (0) is  0.8113207547169812
         Accuracy for survived class (1) is  0.8985507246376812
```

Figure 4.7 Confusion Matrix and Accuracy score of the dataset

4. CONCLUSION

Using SMOTE and Random Forest algorithm, we have built an efficient model for survival analysis of heart failure patients with the overall accuracy of 86%. Since, the accuracy for death class and the accuracy of survived class is also high, the model is efficient and not biased.

5. REFERENCES

[1] " Analysis of heart failure patients, Tanvir Ahmad, Journal of Machine Learning Research, 2014.

[2] "Collett D. Modelling Survival Data in Medical Research", 2nd ed. Taylor & Francis; 2011.

[3] "SMOTE: Synthetic Minority Over-sampling Technique", Nitesh V. Chawla, Kevin W. Bowyer Journal of Artificial Intelligence, 2012.

[4] "Analysis of a Random Forests Model", Gerard Biau, Journal of Machine Learning Research, 2012.

[5] "A Novel Approach to Compute Confusion Matrix for Classification of n-Class Attributes with Feature Selection", V Mohan Patro, Transactions on Machine Learning and Artificial Intelligence, Volume 2 No 4, Aug (2014)

