DATA-583 : EDA

Baldeep Dhada, Han Chen, Somya Nagar

2024-03-09

Dataset Description

Summary

This dataset includes the medical histories of 299 patients diagnosed with heart failure, with the target variable being ‘death_event.’ Each patient’s profile consists of 13 clinical attributes.

variables	Type	description	values
age	Integer	age of the patient (years)	[40,95]
anaemia	Binary	decrease of red blood cells or hemoglobin	0,1
creatinine_phosphokinase	Integer	level of the CPK enzyme in the blood(mcg/L)	[23,7861]
diabetes	Binary	if the patient has diabetes	0,1
ejection_fraction	Integer	percentage of blood leaving the heart at each contraction (%)	[14,80]
high_blood_pressure	Binary	if the patient has hypertension	0,1
platelets	Continuous	platelets in the blood(kiloplatelets/mL)	[25100,850000]
serum_creatinine	Continuous	level of serum creatinine in the blood (mg/dL)	[0.5,9.4]
serum_sodium	Integer	level of serum sodium in the blood (mEq/L)	[113,148]
sex	Binary	woman or man	0,1
smoking	Binary	if the patient smokes or not	0,1
time	Integer	follow-up period (days)	[4,285]
death_event	Binary	if the patient died during the follow-up period	0,1

Statistical Descriptive Analysis

Target: As Figure 1 shows below, our targeted variables `death_event` is unbalanced with 203 survival observations vs 96 dead observations.

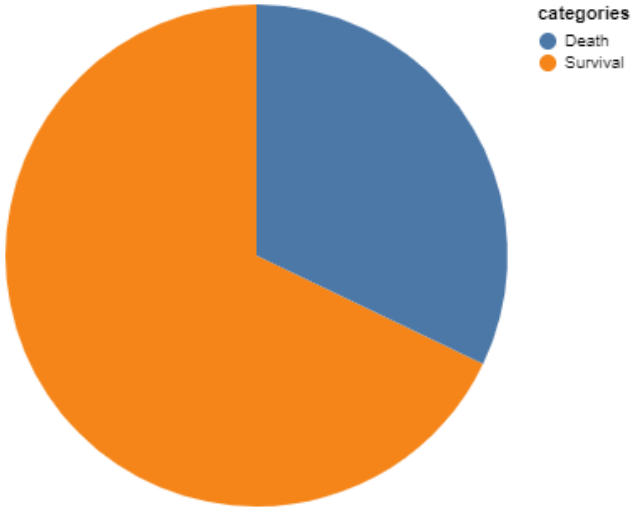

Figure 1: Distribution of target variable

Quantitative Variables

We visualize 7 quantitative variables with histograms and fit each distribution of them, except for ‘time,’ which has two peaks. The upper four plots comply with the gamma distribution, and the lower ones follow the normal distribution

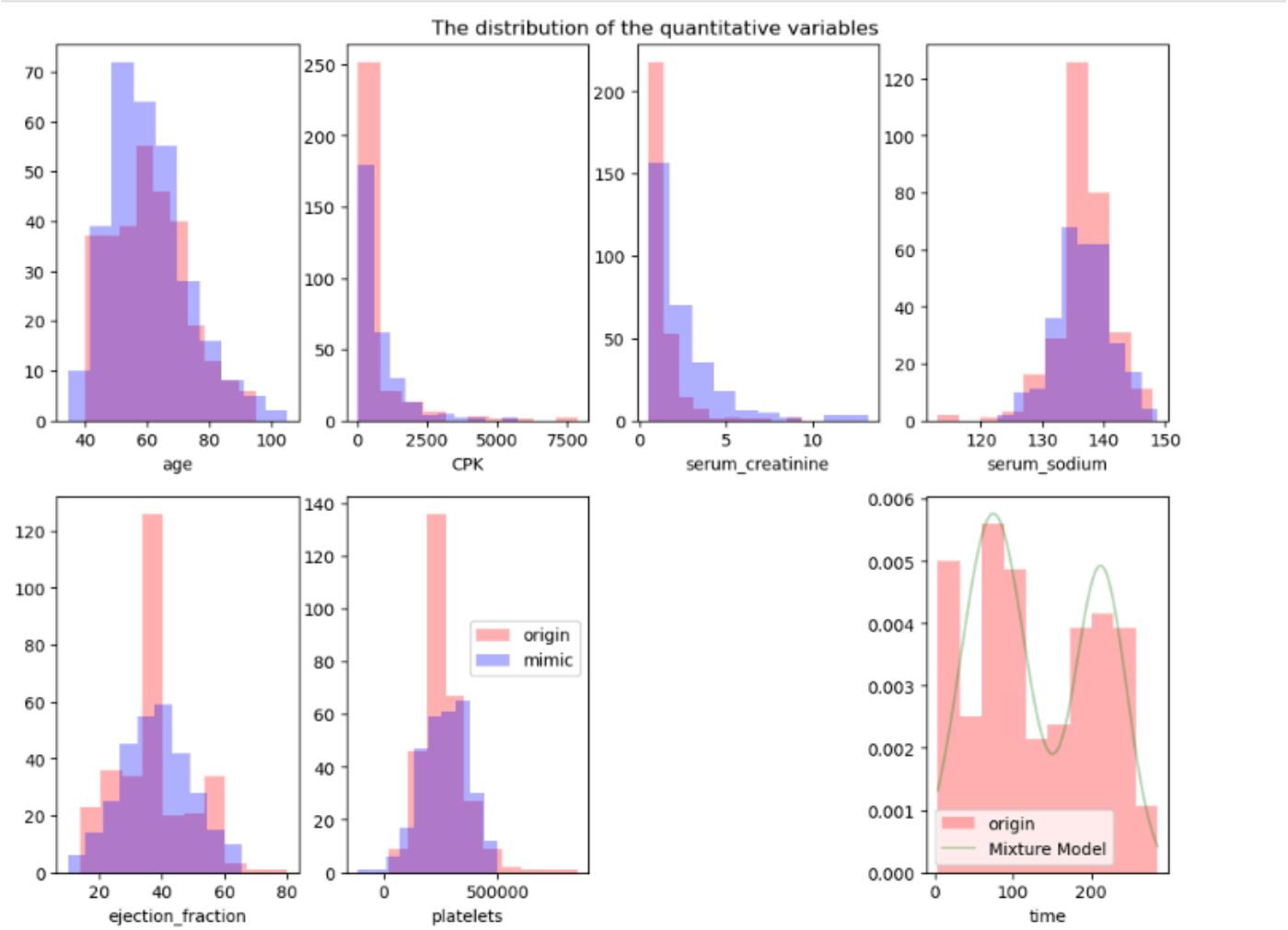

Figure 2: Distribution of Quantitative Variables

As shown in the table below, these variables exhibit different scales, indicating the need for standardization. Based on descriptive statistics, the variables do not appear to have significant differences between the deceased and surviving groups, except for **`time`** .

	Full sample			Dead patients			Survived patients		
Numeric feature	Median	Mean	σ	Median	Mean	σ	Median	Mean	σ
Age	60.00	60.83	11.89	65.00	65.22	13.21	60.00	58.76	10.64
Creatinine phosphokinase	250.00	581.80	970.29	259.00	670.20	1316.58	245.00	540.10	753.80
Ejection fraction	38.00	38.08	11.83	30.00	33.47	12.53	38.00	40.27	10.86
Platelets	262.00	263.36	97.80	258.50	256.38	98.53	263.00	266.66	97.53
Serum creatinine	1.10	1.39	1.03	1.30	1.84	1.47	1.00	1.19	0.65
Serum sodium	137.00	136.60	4.41	135.50	135.40	5.00	137.00	137.20	3.98
Time	115.00	130.30	77.61	44.50	70.89	62.38	172.00	158.30	67.74

Figure 3 : Statistical quantitative description of the numeric features

Combined with the correlation matrix, we can also observe a stronger relationship between the follow-up time and the target variable.

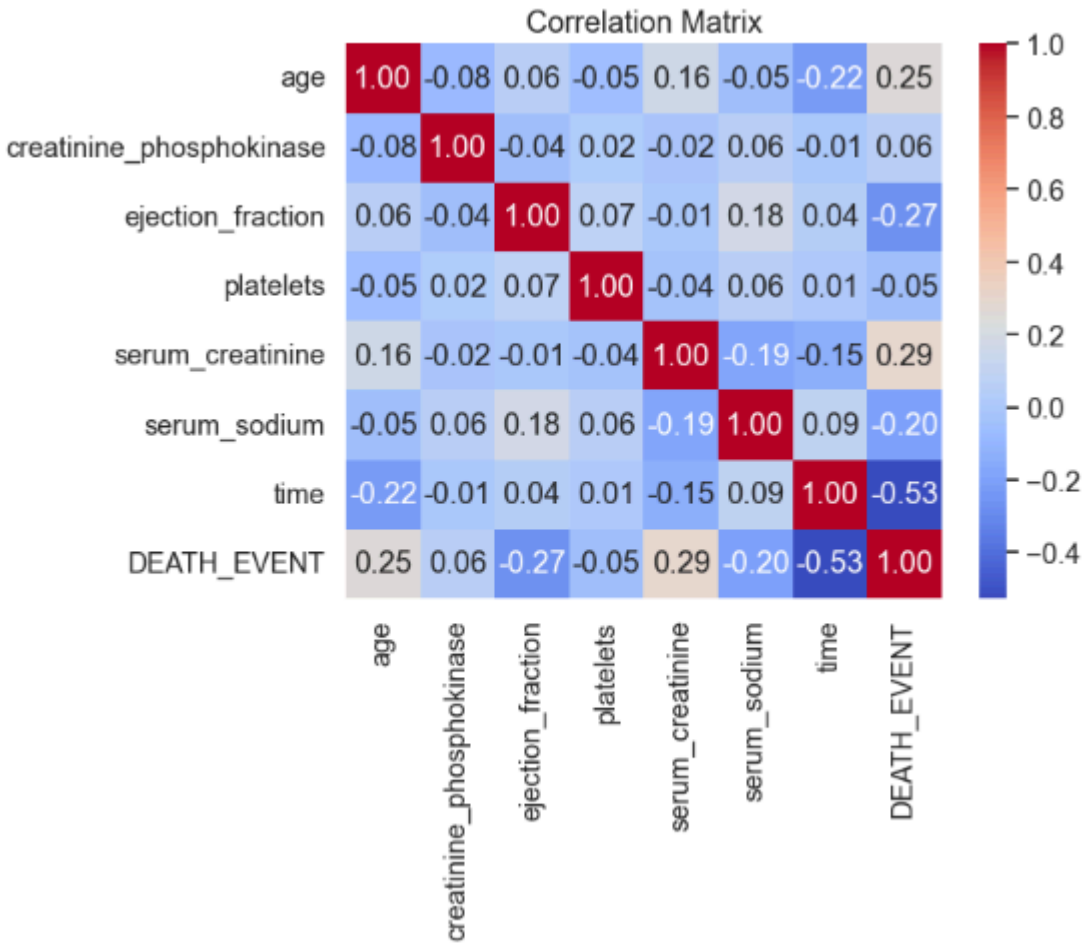

Figure 4: Correlation matrix

Categorical Variables

We visualize 5 categorical variables against the target variable “death_event.” As the plots indicate, in our dataset, the number of males is larger than females, and individuals with a history of ‘anaemia,’ ‘high blood pressure,’ ‘diabetes,’ and ‘smoking’ are more prevalent than those without such histories. These variables demonstrate bias. Therefore, we need to perform oversampling to preprocess our data.

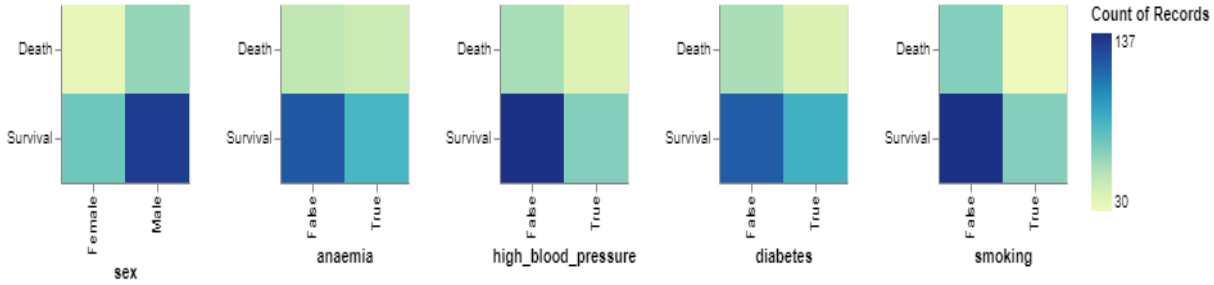

Figure 5: Heatmap of Categorical Variables

Application of different statistical analysis techniques

Preprocessing

Figure 3 shows that the predictors have vastly different scales or units. Creatinine phosphokinase has the largest mean and standard deviation in the data set. Serum Creatinine has the lowest mean and standard deviationIts important to standardize the data with mean of 0 and standard deviation of 1 using z-score normalization. This process will be repeated for each predictor. This will prevent features with larger scales from dominating the learning process.

As highlighted above, there is a huge imbalance in the response variable `death_event` with 203 survival observations vs 96 dead observations. Dealing with imbalanced datasets poses a challenge because the majority of machine learning methods tend to overlook the minority class, resulting in poor performance. In most cases, it is often the performance on the minority class that holds the utmost significance. Hence, we will attempt to balance using the SMOTE (Synthetic Minority Oversampling Technique) algorithm.

Feature Selection Techniques

We will be using several feature selection techniques to choose a subset of relevant and important features from our dataset. Feature selection will help prevent overfitting by reducing the complexity of the model. Unnecessary features might introduce noise or irrelevant information, which can degrade model performance. Irrelevant or noisy features can lead to overfitting, where the model captures patterns specific to the training data but fails to generalize to new data.

In our case we will be deploying a classification tree instead of a regression tree because our reponse variable is reported as a 0 or a 1. Feature importance in classification trees is typically calculated based on how much each feature contributes to the reduction in impurity using the Gini Index by summing up the impurity decreases for each node where a particular feature is used to split the data.

A variance importance plot will be generated after deploying a Random Forest model. This plot will rank the features and show us the drop in the models accuracy as features are excluded. By examining the variable importance scores, you can decide to eliminate or retain certain features. Features with low importance scores may be candidates for removal if the goal is to simplify the model without significantly sacrificing predictive performance. The random forest will be fine tuned with different number of variables tried at each split.

We will plot the reduction in coefficients at different levels of lambda by a LASSO. You can access the coefficients to identify important features. Features with non-zero coefficients are considered selected by the Lasso. LASSO is preferred over ridge regression because LASSO forces some coefficients to 0 while ridge regression will only force them ‘close’ to 0. The removal of some variables will take care of any multicollinearity problems.

A z-test will be used to check the significance of the difference between the distributions. We will split the data into 2 groups one for 1 (death) and the other for 0 (survival) and calculate the marginal distribution of each feature in each group. Features with more negative Z-scores are less likely to fail and features with more positive z scores are more likely to fail.

Different methods may show that different variables are important. If I'm running a LASSO I'm not allowing interactions between the predictors but Random Forests and classification trees model interactions. Therefore, its natural for different methods to select different subsets of variables since they are allowing different behavior among the predictors. If these methods display conflicting results in terms of feature selection we will use the suggested features from each model and select the set of features that result in the lowest misclassification rate.

Predictive Techniques

We are using supervised machine learning approaches to to predict the death_event response variable. We will be utilizing Parametric models like, Logistic regression and GAMS as well as Non-Parametric models such as a Decision Tree, Random Forest, Gradient boosting, k-nearest neighbors, negative binomial regression.

Logistic regression will be deployed because our response variable is binary. This model uses the logistic (sigmoid) function to model the relationship between the predictor variables and the probability of the outcome being in the positive class. An S shaped curve will be plotted that compresses any real valued number to the range [0,1]. We can predict death if $\hat{y} > 0.5$ and survival otherwise.

Generalized Additive Models (GAMs) are a flexible extension of traditional linear regression models that allow for non-linear relationships between predictors and the response variable. First we would use B-splines to model non-linear relationships in numerical variables and combining them with dummy variables for categorical variables. And then we will print out the summary to see the detailed result, giving interpretation.

In situations where the decision boundary exhibits a high degree of non-linearity, we have opted for the implementation of a non-parametric algorithm known as **k-nearest neighbors (KNN)**. KNN will dominate logistic regression if the decision boundary is highly non-linear.

Classification trees are prone to over fitting therefore we will prune the tree using the misclassification rate with a cost complexity tuning parameter. The tree will be pruned to the respective number of terminal nodes using the lowest misclassification error generated from cross validation.

Cross validation won't be necessary for the **Random Forest** Model because out of bag error from bagged models can be compared to cross validated error from other models. Out of bag estimation is an efficient way for providing realistic long term error. The goal is to reduce the error rate by different the number of variables tried at each split.

Gradient boosting models are incredibly powerful and popular which is why we decided to include this technique. The distribution for this model will be binomial. The model will be built iteratively with 5000 trees but the model obtained at the iteration where cross-validation performance is optimal will be used as the final model.

Grid search will be utilized for finding the best parameters. Each technique (except Random Forests) will be k-fold cross validated to provide a more robust and reliable estimate of a model's performance compared to a single train-test split. It helps assess how well a model generalizes to new, unseen data. Without cross-validation, the model evaluation might be highly dependent on the specific random choice of data points in the training and testing sets. Our data set is limited to only 299 rows, cross-validation allows for more efficient use of the available data. Instead of allocating a fixed portion for training and testing, cross-validation uses the entire data set for both training and testing in different folds, maximizing information extraction from the limited data. Cross validation

also ensures we are not over fitting to the training data and helps generalize our model to new data. The misclassification rate for both the training and testing split will be reported. If they are very similar or identical it would mean that our model generalizes well to new data therefore we are not overfitting to the training data.

Every model will be compared using a confusion matrix, misclassification error, logloss, F1 score. F1 score will be heavily emphasized because there is an imbalance in the response variable in the data set.

Scientific Questions

The medical dataset at hand contains valuable information and can help in answering some crucial questions, and understanding trends that provide useful insights about patients health. It has great applications in a predictive landscape and we will try to use the predictors and create models to make predictions. The scientific questions that we will try to answer are discussed below:

1. **Out of all the predictors available, which of them are significant in determining whether the patient dies within the follow up period?**

- Identifying the key clinical variables associated with the risk of death is crucial for understanding disease progression and tailoring treatment plans. It can guide clinicians in prioritizing interventions and monitoring strategies for at-risk patients.
- Heart failure is characterized by complex mechanisms and is affected differently by different clinical factors. Hence, having a better understanding of what predictors are significant would prove to be vital information to any physician.
- Having the answer to this question would have a great significance in the problem that we are trying to solve with this project.

  Statistical Techniques:

- This is a variable selection problem. We will be using different statistical techniques to try to answer this question using this dataset. They are discussed below:
    - LASSO Regression (L1 Regularization): Lasso regression can be used to perform variable selection by penalizing coefficients of less important variables to zero. This helps in identifying the subset of clinical variables with non-zero coefficients, indicating their importance in predicting death events.
    - Feature Importance Analysis: Analyzing feature importance from decision trees or random forests helps identify which clinical variables are most predictive of death events. This insight can guide clinical decision-making and risk stratification strategies.
    - Z-score: We can use the Z-score to standardize variables. This normalization technique ensures that variables are on the same scale, facilitating comparisons and interpretations across different features.

2. **Can we predict the likelihood of death during the follow-up period based on the patient's clinical attributes?**

- Heart failure is a condition that presents with various complications and and clinical presentations. Predicting the likelihood of death due to heart failure during the follow-up period can prove to be crucial.
- This question is of great importance in clinical practice as it helps healthcare providers identify patients at higher risk of mortality, allowing for timely interventions and personalized treatment strategies.

- By identifying patients at higher risk of mortality, healthcare providers can implement targeted interventions and monitor the disease progression more closely. It will also have applications in adjusting the follow-up period of the patients accordingly.

  Statistical Techniques:

  - Random Forests: Decision tree-based algorithms like random forests can be employed for classification tasks. They can capture nonlinear relationships between predictors and the likelihood of death.
  - Logistic Regression: Logistic regression can be used to model the probability of death (binary outcome) based on the patient's clinical attributes. It provides the probability of death given the values of the predictor variables.
  - Gradient Boosting Machines (GBM): GBM algorithms like XGBoost or LightGBM can effectively handle complex relationships and interactions between variables, thus aiding in predicting death events.