# Towards a Better Heart Failure Treatment: Saving One Life at a Time

***An Analytics Edge Approach***

Aziz Ayed
Maria Camila Marenco
Sara Darwish
Shaleenraj Kaur

## Table of Contents

## I. Purpose

Cardiovascular diseases (CVDs) are the number one cause of death globally, taking an estimated 17.9 million lives each year accounting for 31% of all deaths worldwide. People with cardiovascular disease or who are at high cardiovascular risk need early detection and management.

In view of that, our healthcare consulting company is tasked with helping Faisalabad Institute of Cardiology reduce the number of deaths caused by heart failure. This is through analyzing the various risk factors and their contribution to predict death due to heart failure. In addition, implementing survival analysis to observe the probability of death overtime to tackle the limitations of censored data with respect to follow-up time.

## II. Exploratory Data Analysis (EDA)

The dataset has 12 independent variables and 1 dependent variable (binary representing whether a patient dies or not during the follow-up period). The description of these variables is included in the Appendix. Before conducting the predictive analysis, data exploration is performed including the correlation plot seen in Figure 1 below which shows that there is a strong negative correlation between follow-up time and death, meaning that more people die with shorter follow-up times. However, this could potentially be due to bias in the data knowing that the follow-up period stops when the patient dies. We will further address data bias in our survival analysis section. Moreover, this plot preliminarily shows us that there is a relatively strong negative correlation with ejection fraction and a relatively weaker positive one with serum creatinine.

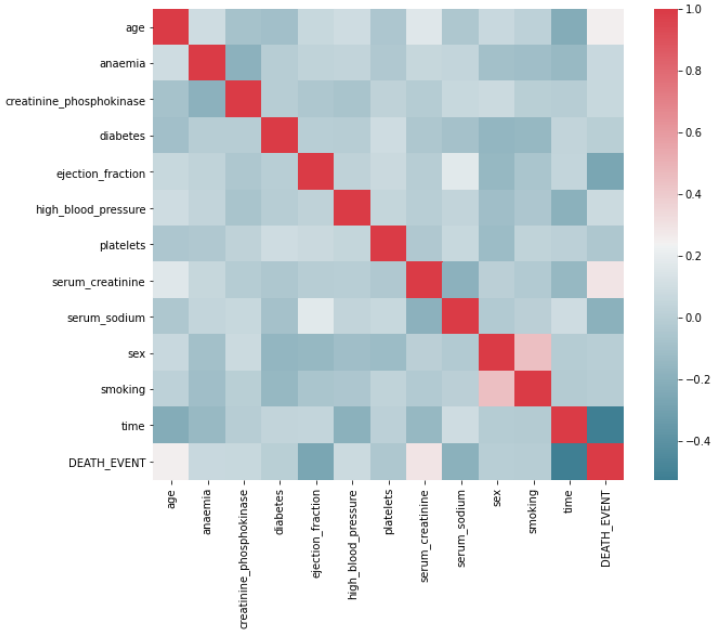


*Figure 1 - Correlation Plot*

Additional exploratory analysis done, validated withheld beliefs of factors that influence heart failures (i.e: smoking, age, diabetes and high blood pressure are correlated with higher mortality due to heart

failure). Refer to the Appendix Figures 1 - 7 for full EDA plots - conditional density plots, violin plots, etc.

# III. Predictive Modeling

## A. Traditional Predictive Models

It is important to note that the variable "time" which accounts for the duration of the follow-up period has been removed prior to fitting the models. This is mainly because our purpose is to determine patient factors that will predict if a patient will die from heart failure and help doctors make decisions with this information. Thus, the follow-up period is a biased factor completely based on the doctor's/hospital's decision and has proven to have a huge impact on death prediction. We consider excluding this variable for our prediction models will benefit our analysis.

Traditional predictive models (Logistic Regression, CART, Random Forest, XGBoost, Optimal Classification Trees and Light Gradient Boosting Machine (LGB)) were implemented and evaluated based on a 75:25 train-test split ratio. Stratified sampling was also implemented to ensure a representable sample of death observations in the test set. Furthermore, 5-fold cross validation was implemented for all models to ensure better out of sample performance. In order to obtain additional improvement from the top previous standalone models, we also tried ensemble models.The ROC curve for baseline models can be seen in Figure 8 in the Appendix.

## B. Baseline Model

The logic used to select the baseline model was to develop a simpler model that doctors could potentially use in contrast to the more complex ML models previously defined. Thus, the baseline model defined in this problem only includes variables provided by the patient to the doctor such as Sex, Age, and Smoking. This baseline input differs from the previous more "complex models", since those also include variables measured by the doctors (lab results). Finally, the same traditional predictive models stated above were fit to the baseline using only these three variables and the best-performing one based on the test AUC was chosen as a baseline model. In this case, the best performing one was the logistic regression model. The ROC curve for baseline models can be seen in Figure 9 in the Appendix.

## C. Model Comparison and Discussion

To select the best predictive model and compare the results to the baseline, the following healthcare metrics, in addition to traditional metrics, will be considered due to its importance in a real-world context:

- **Number of Surgical Interventions:** If a patient is predicted to die from heart failure, the doctor can perform a costly preventive heart bypass. This intervention reduces by 25% the mortality of heart attacks.
- **Total Deaths:** Total number of patients that actually died from heart failure.
- **Prevented Deaths:** Estimate of the total number of patients that were saved by performing a surgical intervention procedure.
- **Costs:** Total costs for each scenario including monetary values for surgical interventions and deaths. Death cost was quantified using papers published on the statistical value of life that can be seen cited in the reference section. For the cost of surgical intervention we utilized publicly

surgical rates that can also be seen in the reference section of this report. Using this information, we defined the following cost matrix in Figure 2 below:

| ***True Positive*** | ***False Positive*** |
|---|---|
| • Surgical intervention to all patients: $125 K per preventive heart bypass<br>• Death reduction of 25% of patients who had surgery ($ 8.7 M value of Life) | • Intervention to all patients: $125 K per preventive heart bypass<br>• No deaths |
| **False Negative** | **True Negative** |
| • No interventions performed<br>•. Deaths ($8.7 M value of Life) | • No interventions performed<br>• No deaths |

*Figure 2- Cost Matrix Input for the Models*

The models were compared using the previously described metrics and traditional metrics as well (AUC, False Negatives, False Positives, Accuracy and Recall) and results are shown in Table 1 and 2 in the Appendix. As expected, all machine learning models perform better than the baseline in terms of healthcare metrics (deaths and total costs) and traditional metrics. In fact, the baseline performs poorly regarding false negatives and thus death prevention, which is of extreme importance for the context of our problem.

Additionally, by analyzing the following tables, it is seen that the best predictive models in terms of healthcare metrics are Logistic Regression and XGBoost (highest number of prevented deaths and lowest total deaths and costs). This differs from the AUC ranking because in the context of our problem, false negatives rate plays a crucial role. As previously mentioned, preventing death is of extreme importance, thus having a false negative in the context of our problem is more costly than having a false positive. Therefore, the better performance of Logistic Regression and XGBoost mainly lies on their capability of having lower False Negative and not necessarily the best AUC.

With this in mind, the Logistic Regression model was selected as the best predictive model due to its performance on the healthcare metrics due to its interpretability for doctors to use given its relatively simpler nature.In addition to implementing each of the above models alone, TPOT and ensemble models were also implemented. However, they produced worse results than of our previous models.

## D. Feature Importance

The XG Boost SHAP summary plot seen below in Figure 3 is very much in line with the logistic regression feature importance analysis and Random Forest SHAP (Figure 10 and 11 in the Appendix) showing that ejection fraction, serum creatinine and age are the three most important features respectively. The figure also shows that patients with lower ejection fraction and higher serum creatinine and older ages are more likely to die.

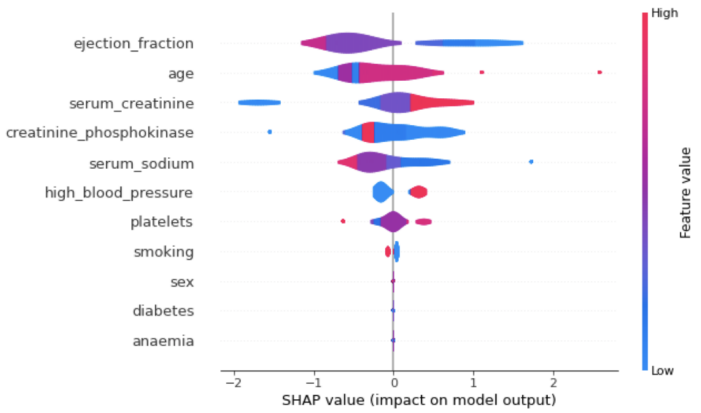


*Figure 3 - XG Boost SHAP Summary Plot*

# IV. Survival Analysis

## A. Dealing with Censored Data

In our dataset, we are dealing with censored data. Indeed, we are looking at the death_event during the follow-up period of a patient. A value of 0 simply means that the patient did not die during that follow-up period, but could as well have died one day later without it being registered. As seen in Figure 4 below, we draw 25 random uncensored events (death = 1), and we fix a follow-up period of 150 days for each one of them. Among these 25, 5 of them (20%) will be registered as not dead, even though they died a few days later. Thus, we decided to fit a survival model to our dataset, namely Cox Proportional Hazard Models in order to be able to detect this type of patients.

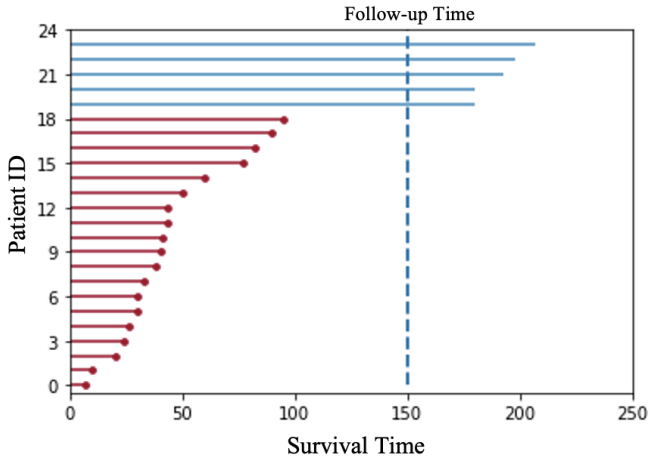


*Figure 4 - Observation of Censored Data*

## B. Cox Proportional Hazard Models

With this family of models, we are interested in modeling the hazard function $\lambda(t, x) = \lambda_0(t)e^{\theta^T X_i}$ This function can be interpreted as the probability of an individual dying at time t given that he hasn't died yet. In other terms, this allows us to look at the probability of a patient being alive over a window of time (biggest follow-up period in the dataset) rather than a point estimation after a given follow-up period. Different varieties of survival models exist, but this one seems particularly suited for our problem as it models a multiplicative effect of the covariates on the hazard. Intuitively, this seems to capture the idea of comorbidities in our scenario (for example, smoking should be even worse for people with diabetes). We use the C-Index to evaluate the performance of our model.

## C. Results

After fitting our Cox Model (using a second-order algorithm such as Newton-Raphson), we start by making sure its predictions make sense. Thus, by fixing an arbitrary threshold at 0.55 and only looking at the uncensored data (death events), we observe an accuracy (TPR) of 74% (this is simply a sanity check and does not provide further information about the quality of the model). Survival curves for different patients can be seen in Figure 12 in the Appendix.

After looking at the censored data, we observe an interesting phenomenon illustrated in Figure 5 below which shows that some individuals registered as not dead have low probabilities of surviving after the end of their follow-up periods. For example, out of the over 203 individuals registered as not dead, 12 reach a probability of survival lower than 30%, which is potentially 6% of false negatives (over the total negatives) in our dataset.

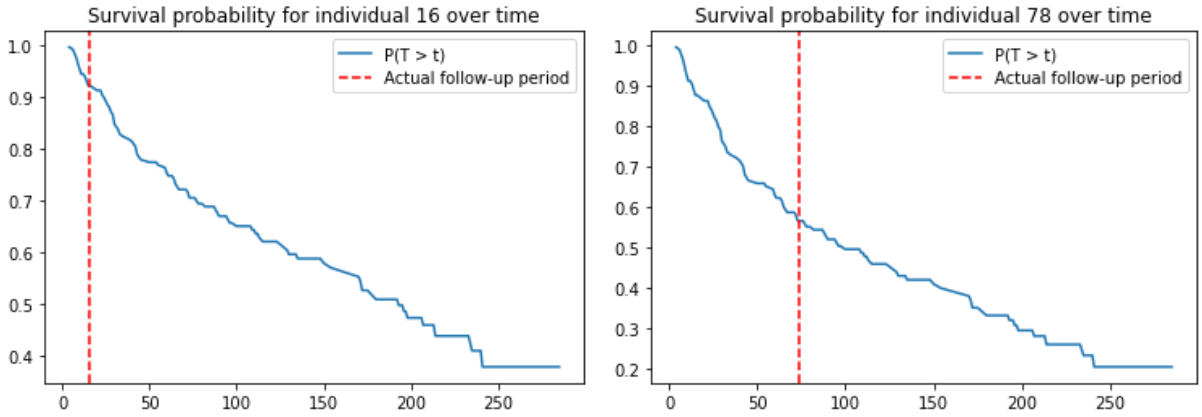


*Figure 5 - Survival Probability of Patients Labeled 0*

## V. Conclusions and Impact

Our survival model allows us to characterize the risks of death over time for different patients instead of giving a prediction of death at a given point of time. It is hard to compute an accuracy because here, we are assuming that the 0 of our dataset might be falsely labelled. However, given that this model has a slightly worse TPR than the traditional models, one can imagine them being used conjointly. For example, predicting the patients that are likely to die using a classical ML method, and give a more nuanced prediction for the patients predicted as 0. Another possible usage could be urgence ranking for surgeries, giving higher priority to patients that are more likely to die in the next few days. Finally, this gives way more interpretability to the feature importance ranking that we did earlier as we are now able to plot curves of partial effects of a given predictor on the outcome of a patient (see Figure 13 in the Appendix). Moreover, using SHAP explanation force plots, doctors can identify for each individual patient what are his particular risk factors. Thus, they can lead non critical patients towards improving their quality of life with concrete day-to-day actions (improving ejection fraction with sports, quit smoking, etc.). By using our analysis, patients with high risk can be detected early and doctors can identify factors that affect heart failure the most and therefore improve treatments accordingly.. In addition, the survival analysis can guide doctors to choose the optimal follow up time.

## VI. Appendix

**Variable Descriptions:**

- **Age:** Age of the patient
- **Anaemia:** 1 if the person has anaemia , 0 otherwise
- **Creatinine_phosphokinase:** Level of the CPK enzyme in the blood (mcg/L)
- **Diabetes:** 1 if the person has diabetes , 0 otherwise
- **Ejection_fraction:**Percentage of blood leaving the heart at each contraction (percentage)
- **High_blood_pressure :** 1 if the person has high blood pressure, 0 otherwise
- **Platelets:** Platelets in the blood (kiloplatelets/mL)
- **Serum_creatinine:** Level of serum creatinine in the blood (mg/dL)
- **Serum_sodium:** Level of serum sodium in the blood (mEq/L)
- **Sex:** 1 if men, 0 otherwise
- **Smoking:** 1 if the patient smokes, 0 otherwise
- **Time:** Duration of the follow-up period
- **Death Event:** 1 if the person dies during the follow-up period, 0 otherwise

**EDA Plots:**

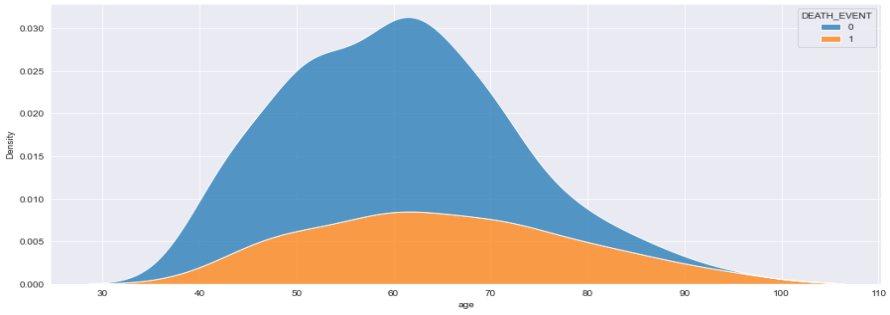


*Figure 1- Conditional density lot of age and probability of death*

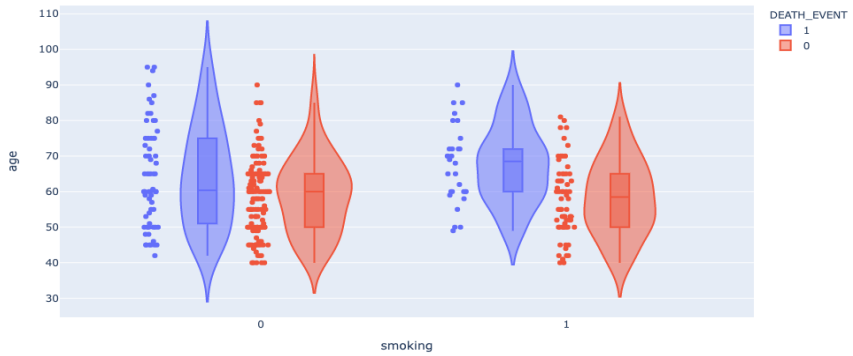


*Figure 2- Relations between variables showing people that smoke are more likely to suffer heart failure compared to people who don't*

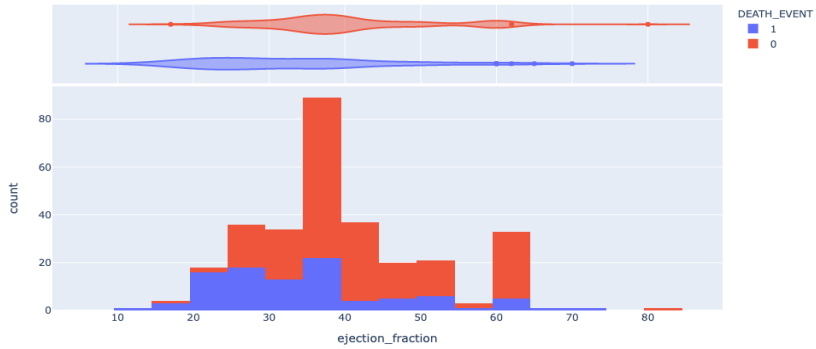


*Figure 3 - Patients with less ejection fraction are more likely to die of heart failure*

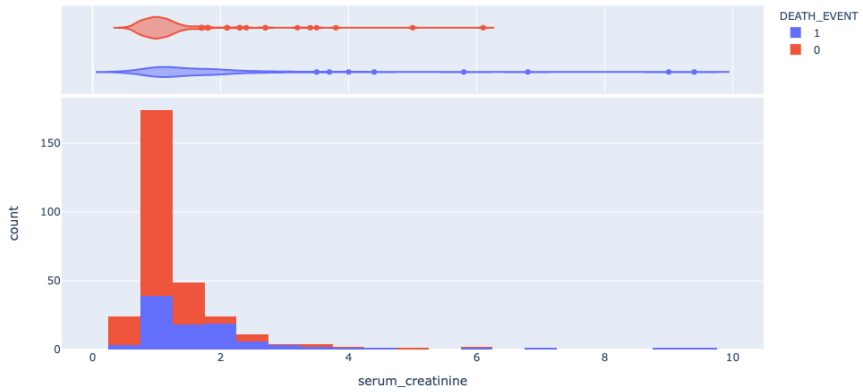


*Figure 4 - Higher levels of serum creatinine indicates a higher probability of heart failure*

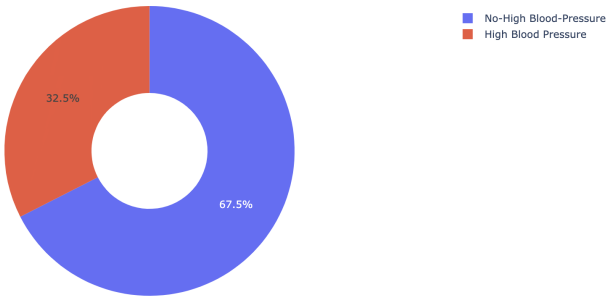


*Figure 5 - People that has no high blood pressure has a higher chance of surviving compared to people with higher blood pressure*

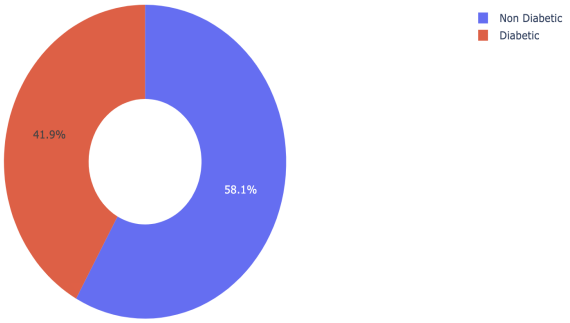


*Figure 6 - People that have no diabetes has a higher chance of surviving compared to people with diabetes*

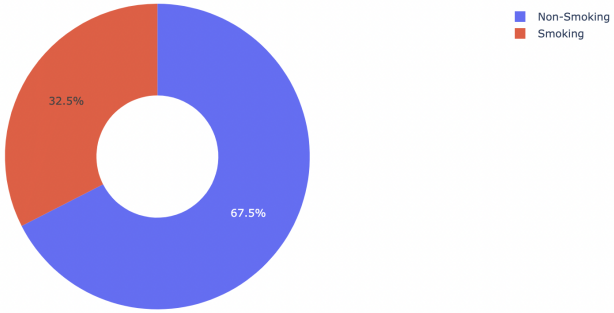


*Figure 7 - People that dont smoke has a higher chance of surviving compared to people that smokes*

| **Models** | **Number of interventions** | **Total Deaths** | **Prevented Deaths** | **Total Costs** |
|---|---|---|---|---|
| **Random Forests** | 10 | 23 | 1 | 201,350,000.00 |
| **XGBoost** | 19 | 21 | 3 | 185,075,000.00 |

| | | | | |
|---|---|---|---|---|
| **Logistic Regression** | 18 | 21 | 3 | 184,950,000.00 |
| **CART** | 20 | 21 | 3 | 185,200,000.00 |
| **LGB** | 16 | 22 | 2 | 193,400,000.00 |
| **OCT** | 20 | 21 | 3 | 185,200,000.00 |
| **Baseline Model** | 4 | 24 | 0 | 209,300,000.00 |

Table 1*: Model comparison regarding Healthcare Metrics*

| ***Model*** | ***Best Parameters*** | ***Accuracy*** | ***Recall*** | ***Precision*** | ***Test AUC*** | ***TN*** | ***TP*** | ***FN*** | ***FP*** |
|---|---|---|---|---|---|---|---|---|---|
| ***RF*** | *Entropy, Max depth: 4* | *0.73* | *0.29* | *0.70* | *0.82* | *48* | *7* | *17* | *3* |
| ***XG Boost*** | *Depth 1, # of Trees: 100* | *0.77* | *0.54* | *0.68* | *0.77* | *45* | *13* | *11* | *6* |
| ***Logistic Regression*** | *Penalty: L1 Norm with parameter 0.751* | *0.79* | *0.54* | *0.72* | *0.72* | *46* | *13* | *11* | *5* |
| ***CART*** | *Gini Max Depth: 4* | *0.73* | *0.50* | *0.60* | *0.69* | *43* | *12* | *12* | *8* |
| ***LGB*** | *Depth 4, n_estimators: 100* | *0.71* | *0.38* | *0.56* | *0.77* | *44* | *9* | *15* | *7* |
| ***OCT*** | *Depth 2, Min Bucket 15* | *0.73* | *0.50* | *0.60* | *0.67* | *43* | *12* | *12* | *8* |
| ***Baseline*** | *Penalty: L1 Norm with parameter 0.721* | *0.68* | *0.08* | *0.50* | *0.60* | *50* | *2* | *22* | *2* |

*Table 2: Model comparison regarding Classical Metrics*

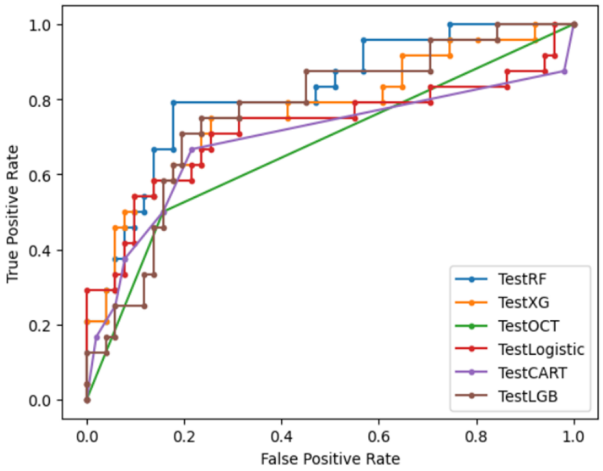


*Figure 8 - ROC of Different Models*

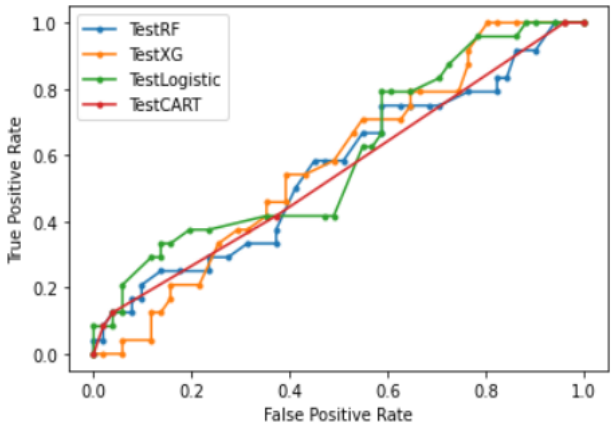


*Figure 9 - Test AUC and ROC Curve Comparison for Baseline*

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               9.670e+00  6.357e+00   1.521   0.1282
age                       5.904e-02  1.945e-02   3.036   0.0024 **
anaemia                  -1.487e-01  4.127e-01  -0.360   0.7186
creatinine_phosphokinase  1.085e-04  2.018e-04   0.538   0.5909
diabetes                  1.590e-01  3.991e-01   0.398   0.6903
ejection_fraction        -8.501e-02  2.006e-02  -4.238 2.26e-05 ***
high_blood_pressure      -2.608e-01  4.215e-01  -0.619   0.5362
platelets                -2.542e-06  2.301e-06  -1.104   0.2694
serum_creatinine          2.676e-01  2.412e-01   1.110   0.2672
serum_sodium             -5.880e-02  4.479e-02  -1.313   0.1892
sex                      -1.967e-01  4.727e-01  -0.416   0.6773
smoking                  -4.289e-01  4.855e-01  -0.883   0.3770
time                     -2.057e-02  3.446e-03  -5.971 2.36e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 10 - Logistic Regression Feature Importance*

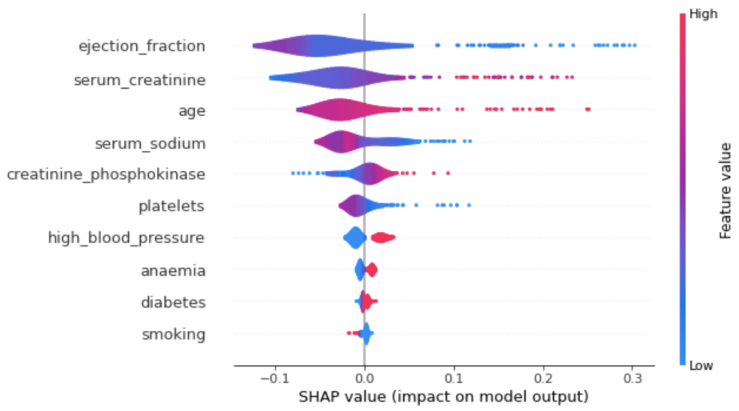


*Figure 11 - Random Forest SHAP Summary Plot*

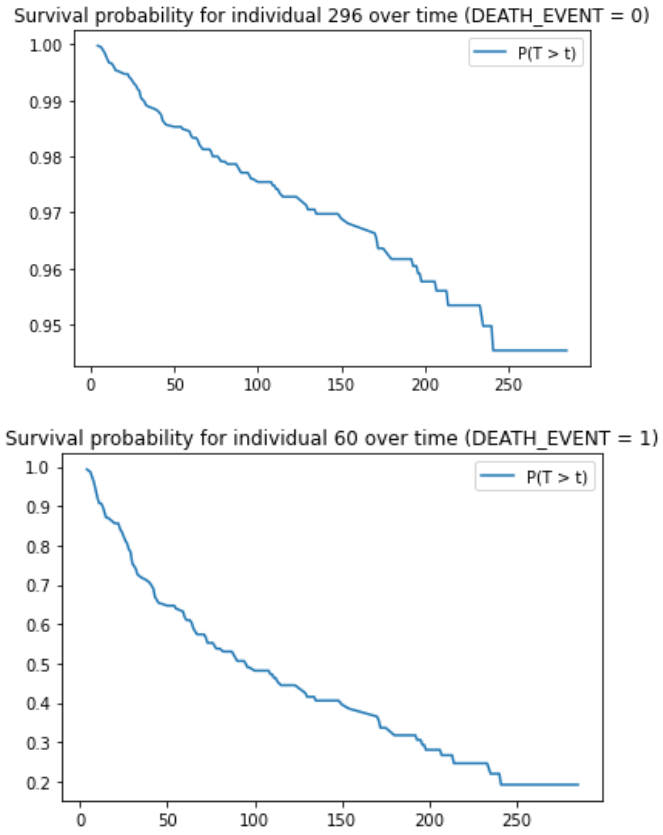


*Figure 12 - Survival Curves for Two Patients*

These curves show survival rates for two types of patients: (i) A patient did not die (Left) with survival curve very high, near 95%, at the end of the 250 day window, (ii) A patient who did die (Right) with survival curve very low, near 2%, at the end of the 250 day window. Showing our survival curves make sense with the data.

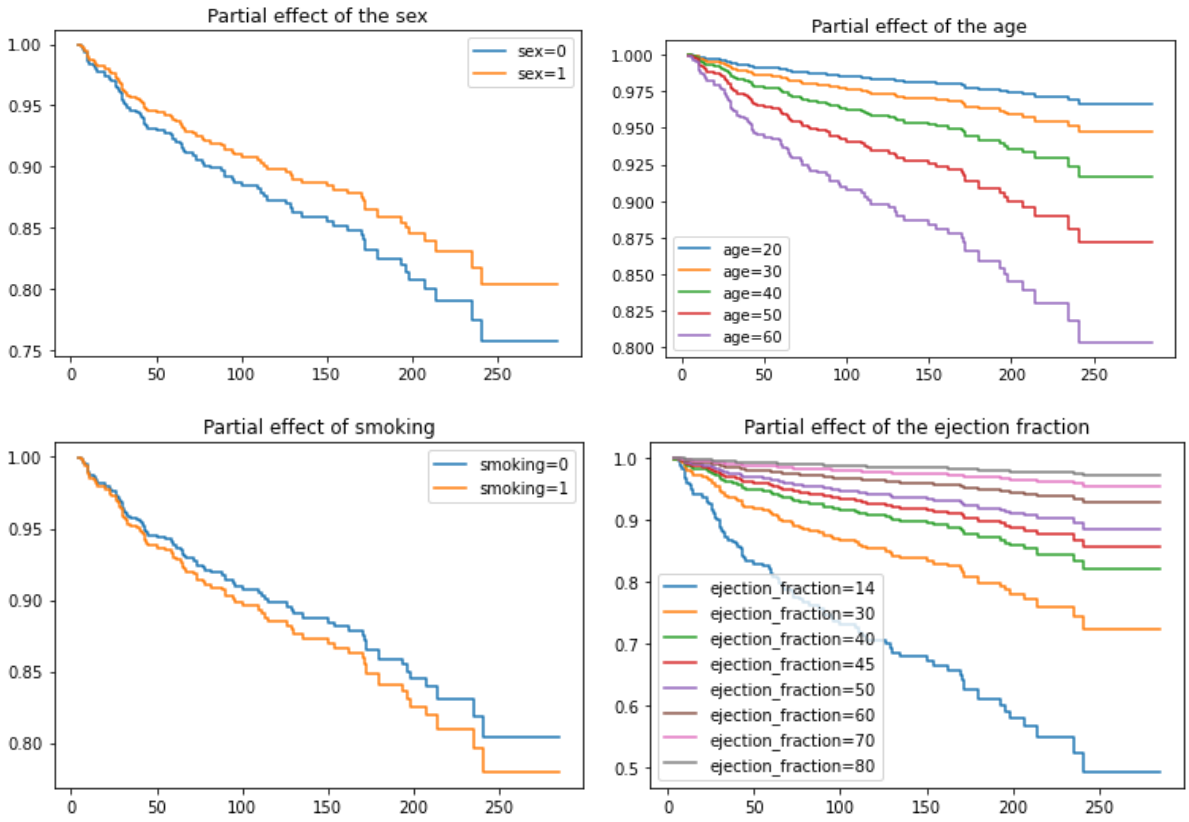


*Figure 13 - Effects of Features on the Survival Models*

Cox models are similar to logistic regressions in their very design. We can now concretely interpret the effect of a covariate on the survival rate. For example, we can see that women have 10% less chance of survival by the end of the 300 days period post-appointment.

## VII. References


Kniesner, T. J., & Viscusi, W. K. (2019, May 16). *The value of a statistical life*. SSRN. Retrieved December 4, 2021, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3379967.

Frakt, A. (2020, May 11). *Putting a dollar value on life? governments already do*. The New York Times. Retrieved December 4, 2021, from https://www.nytimes.com/2020/05/11/upshot/virus-price-human-life.html.

Statista. (2021, June 21). *Cost of an angioplasty in various countries 2019*. https://www.statista.com/statistics/189973/cost-of-an-angioplasty-in-various-countries/

*Survival models - princeton university*. (n.d.). Retrieved December 4, 2021, from https://data.princeton.edu/wws509/notes/c7.pdf.