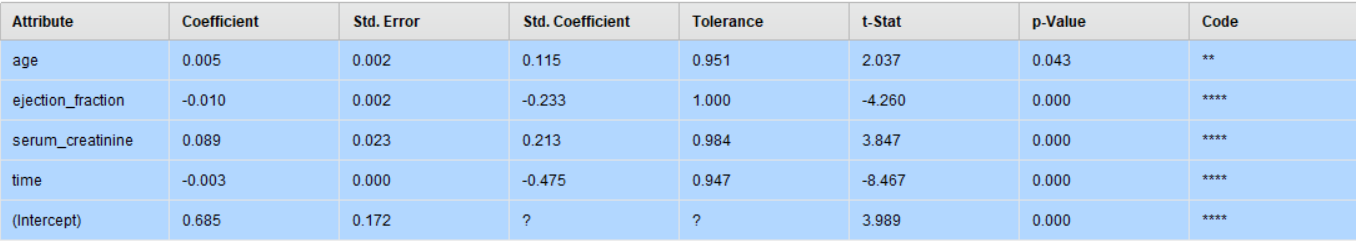
1. A multiple linear regression model depicts the relationship between one quantitative response or dependent variable and several predictor variables, whereas logistic regression models the relationship between a discrete dependent variable and one or more predictor variables. All else being equal, if the dependent variable is binary or categorical in nature with only a set number of outcomes, a logistic regression model is a better fit; if the dependent variable is a continuous quantitative variable with infinite outcomes, a multiple regression model would yield a better outcome. While multiple regression is used to predict the numerical value of a response variable, logistic regression is used for classifying a variable into one of a set number of categories.
2. I used the following data set for answering this question:

<https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records>

This dataset contains the medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features: age,anaemia, high blood pressure, level CPK enzyme in blood, diabetes, ejection fraction, platelets, sex, serum creatinine, serum sodium, smoking, time, death (target variable).

I used RapidMiner to run the data set using linear regression first to see if it could accurately predict a death event, and this was the output of the process (I partitioned the data into training and test sets using a 70-30 split, and then optimized the model by using a set of predictor variables that had statistically significant p-values):

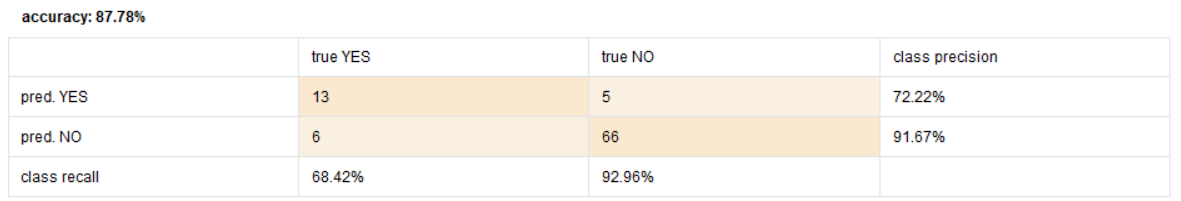


RMSE = 0.333 and R squared = 0.362.

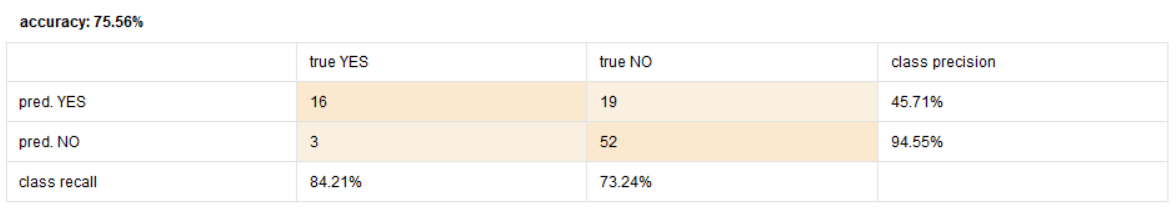
A low R squared value indicates that the predictor variables in this model do not do a particularly good job explaining the variances in the response variable, death event. Plus, the model yielded predicted values of death event as greater than 1 and less than 0 which is outside the possible range of values.

Using the same data set, I then built a logistic regression classification model, and then adjusted the cut off probabilities in order to improve the precision of death predictions.

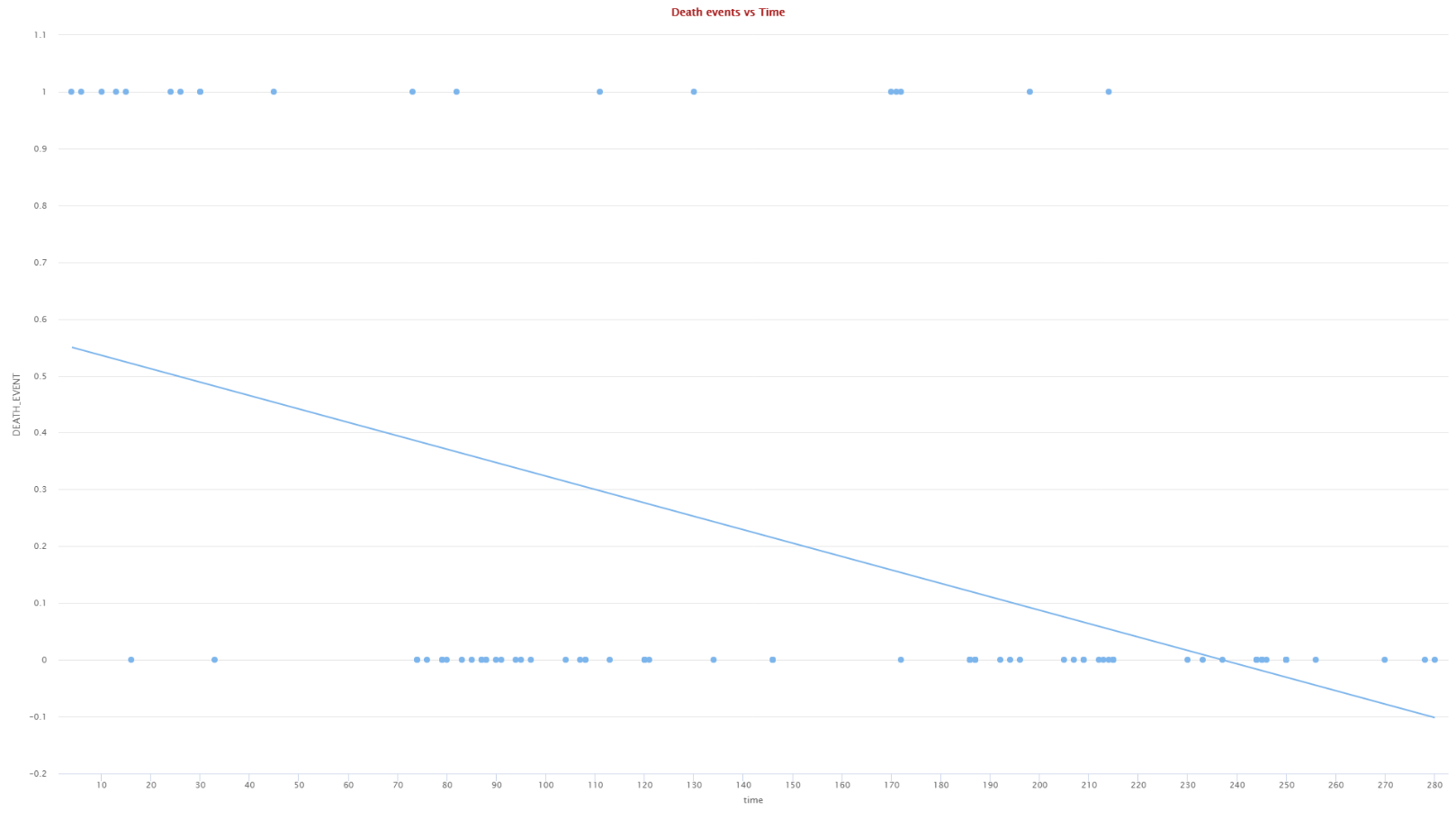
This was my classification metric/performance vector when my probability was set to 0.5:



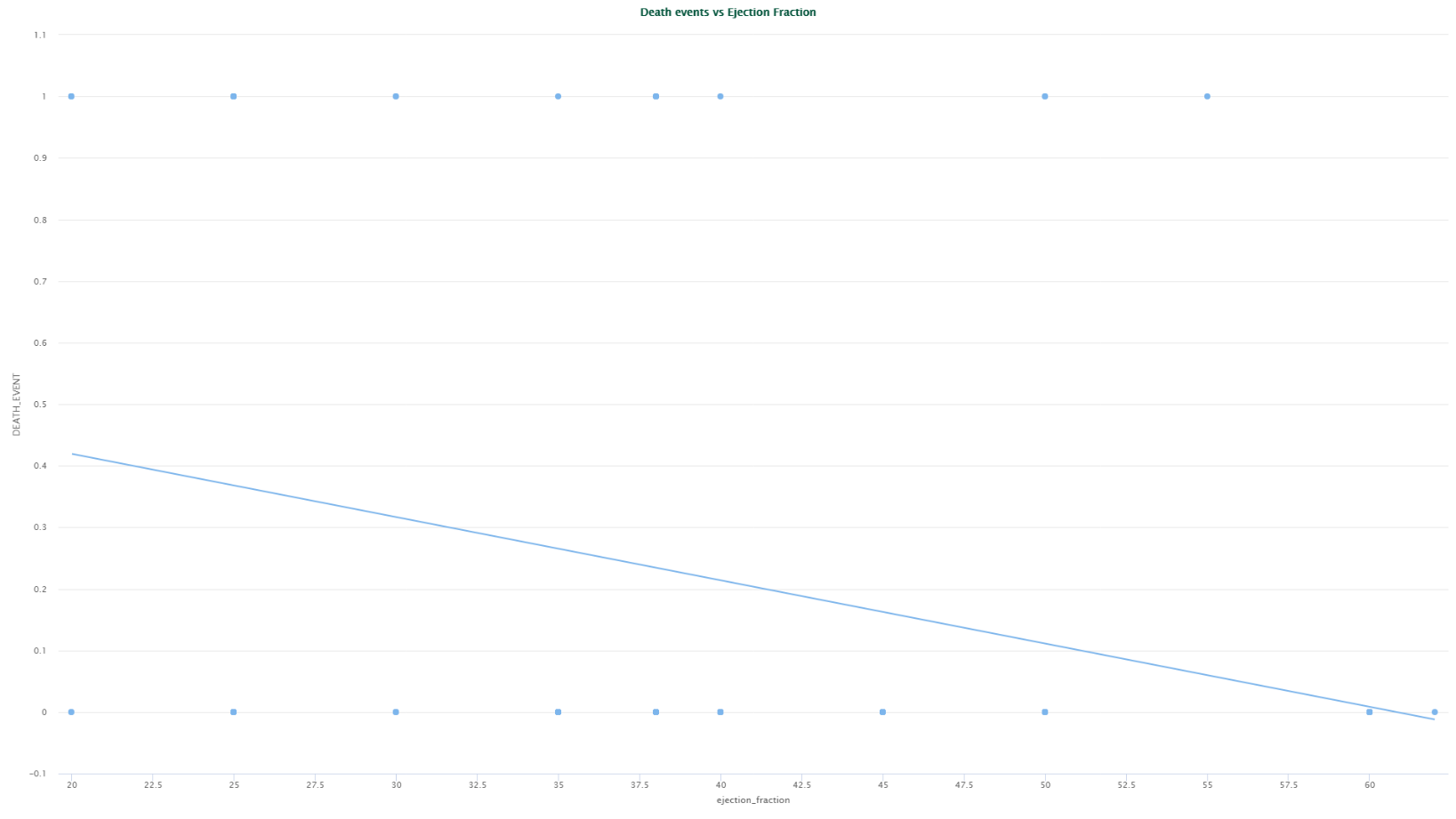
By increasing the threshold to 0.7, I was able to increase the precision of a true death event, although it was at the cost of the overall accuracy of the model:



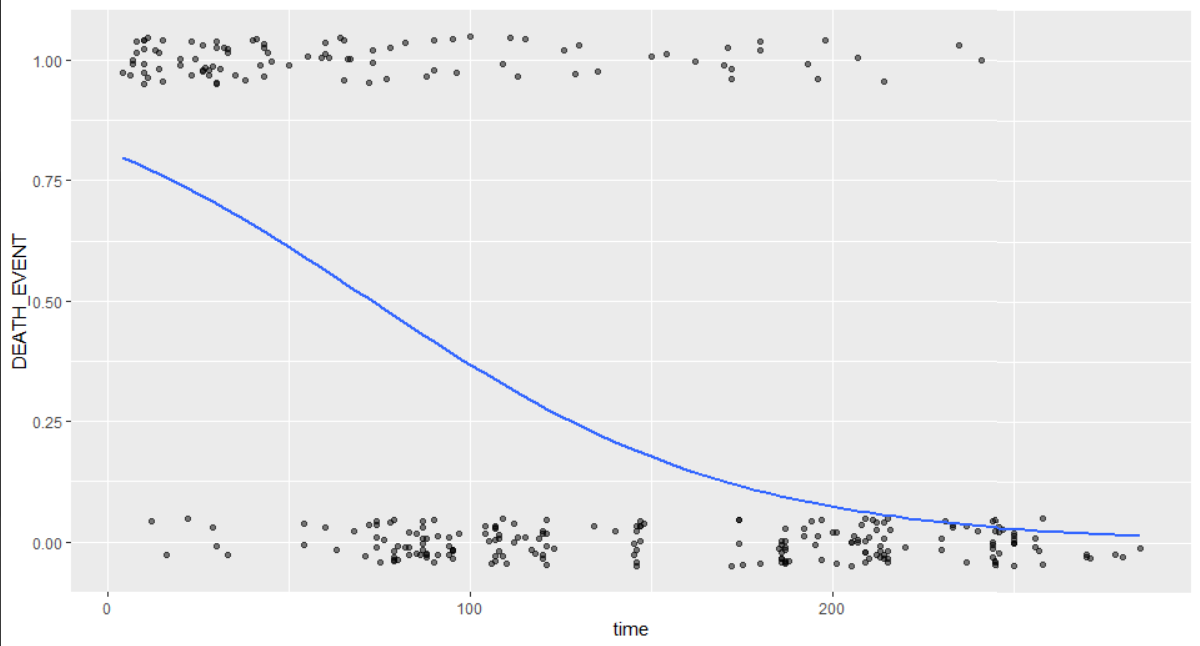
The following is a scatterplot with “deaths” on the Y-axis and “time” (measuring the follow-up period in days) on the X-axis:



Below is another scatterplot with “deaths” on the Y-axis and “ejection fraction” (percentage of blood leaving the heart at each contraction) on the X-axis:



The scatterplot below shows a logistic regression curve for our data set, using “deaths” on the Y-axis and “time” on the X-axis:



From the above illustrations, we can conclude that a linear trend line is not a good fit for these data points, and that a multiple regression model is not the most effective way to predict a response variable with two discrete binary outcomes. In this example, a logistic regression model is a better fit.