Predictive Survival Analysis Proposal

2024-07-01

Aim

- Fit a model to predict in-hospital mortality among patients in the heart failure dataset.

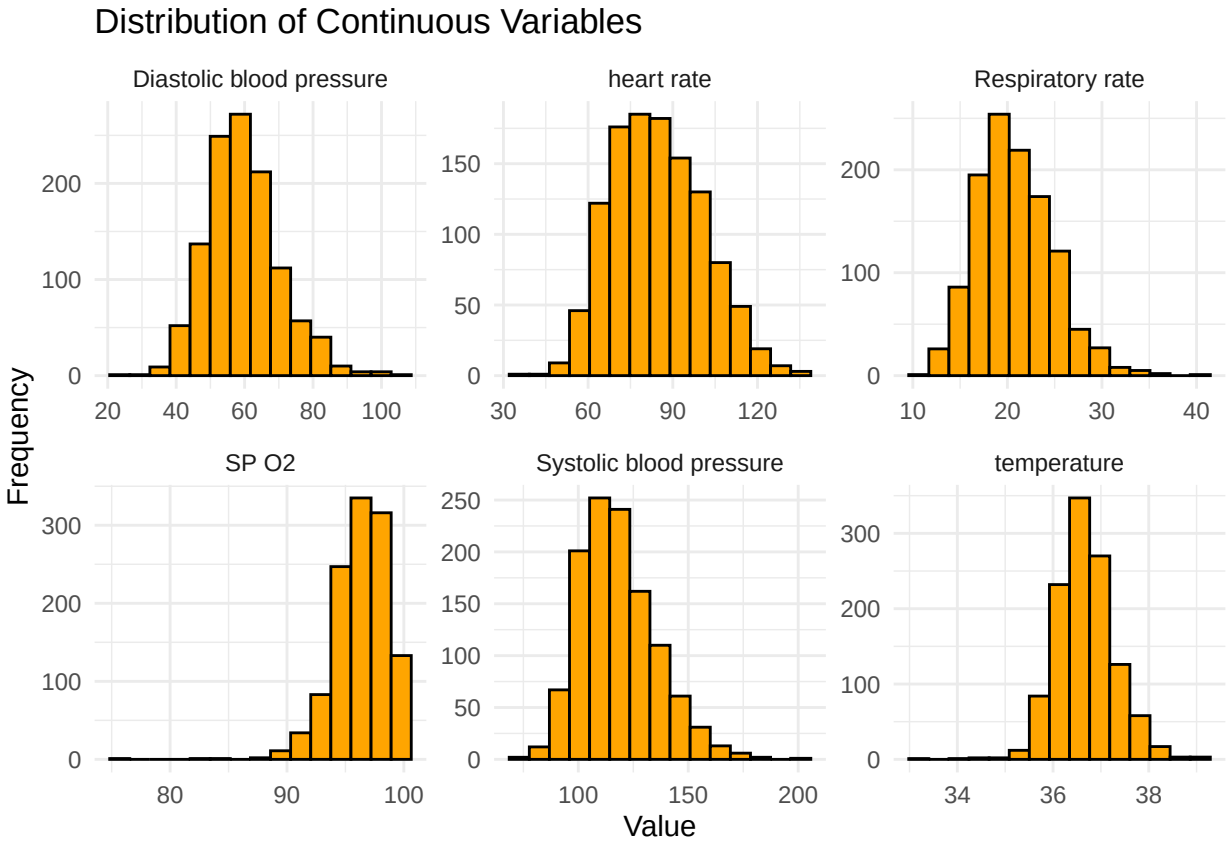
Research Questions

- What are the predictors of in-hospital mortality among ICU admitted heart failure patients?
- How do demographics, vital signs, comorbidities, and laboratory results influence the risk of hospital mortality?
- How can we ensure the interpretability of our model while also maintaining prediction accuracy?
- Which demographic factors, vital signs, comorbidities, and laboratory results are the most significant predictors of in-hospital mortality among ICU-admitted heart failure patients in the MIMIC-III database?

Table 1: Summary Statistics by Outcome

Outcome	Systolic BP (mmHg)	Diastolic BP (mmHg)	Platelets (10^9/L)	pH	INR	PT (s)	MCV (fL)
Survived	118.91	59.90	245.46	7.38	1.58	17.07	89.82
Died	112.18	57.18	216.21	7.36	1.93	20.09	90.47

EDA



t-test

```
cont <- c("age", "BMI", "heart rate", "Systolic blood pressure",
          "Diastolic blood pressure", "Respiratory rate", "temperature", "SP O2",
          "Urine output", "hematocrit", "RBC", "MCH", "MCHC", "MCV", "RDW",
          "Leucocyte", "Platelets", "Neutrophils", "Basophils", "Lymphocyte",
          "PT", "INR", "NT-proBNP", "Creatine kinase", "Creatinine",
          "Urea nitrogen", "glucose", "Blood potassium", "Blood sodium",
          "Blood calcium", "Chloride", "Anion gap", "Magnesium ion", "PH",
          "Bicarbonate", "Lactic acid", "PCO2")

for (var in cont) {
  t_test <- t.test(merged_df[[var]] ~ merged_df$outcome, data = merged_df)
  print(paste("T-test for", var))
```

```
  print(t_test)
}
```

chi-square test

```
cat <- c("gendera", "hypertensive", "atrialfibrillation", "CHD with no MI",
         "diabetes", "deficiencyanemias", "depression", "Hyperlipemia",
         "Renal failure", "COPD")

for (var in cat) {
  chi_sq <- chisq.test(table(merged_df[[var]], merged_df$outcome))
  print(paste("Chi-square test for", var))
  print(chi_sq)
}
```

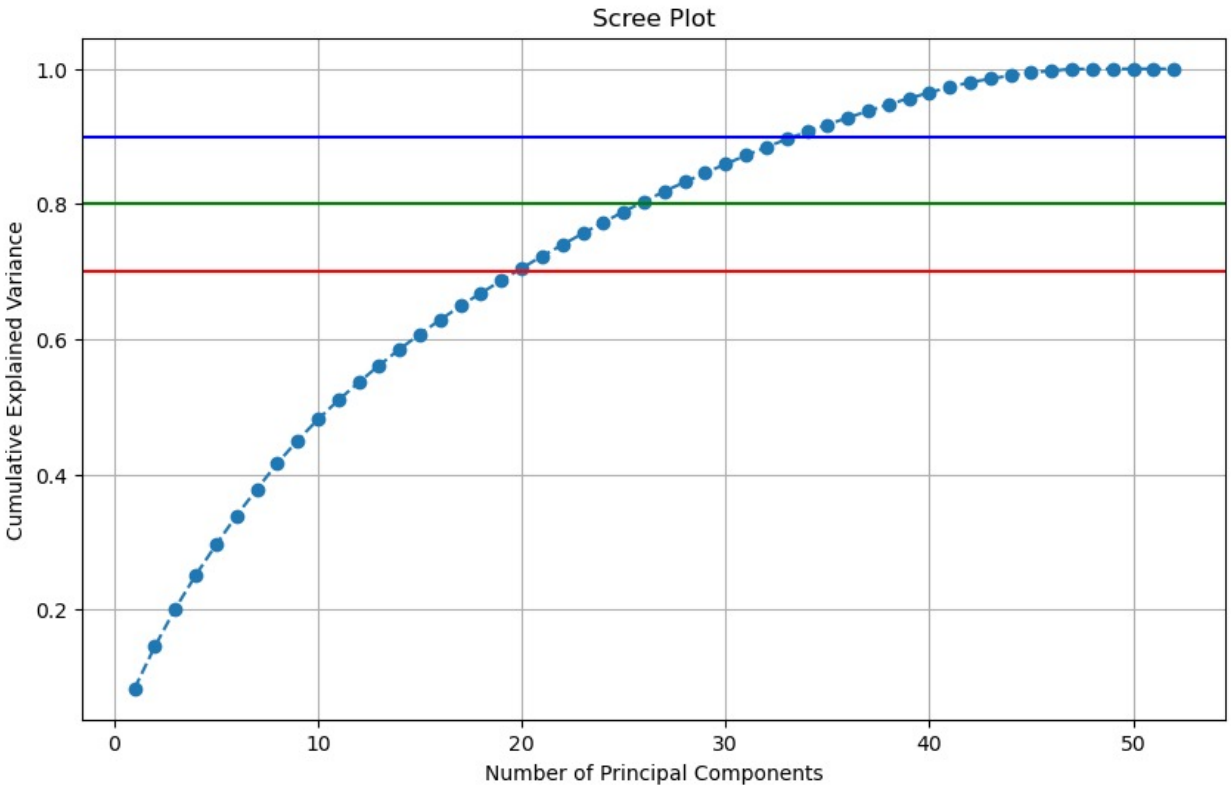
```
full_model <- glm(outcome ~ gendera + hypertensive + atrialfibrillation +
                    `CHD with no MI` + diabetes + deficiencyanemias + depression
                  + Hyperlipemia + `Renal failure` +
            COPD + age + `heart rate` + BMI + `Systolic blood pressure` +
            `Diastolic blood pressure` + `Respiratory rate` + temperature +
            `SP O2` + `Urine output` + hematocrit + RBC + MCH + MCHC + MCV +
            RDW + Leucocyte + Platelets + Neutrophils + Basophils + Lymphocyte +
            PT + INR + `NT-proBNP` + `Creatine kinase` + Creatinine +
              `Urea nitrogen` + glucose + `Blood potassium` + `Blood sodium` +
              `Blood calcium` +  Chloride + `Anion gap` + `Magnesium ion` + PH
            + Bicarbonate + `Lactic acid` + PCO2,
            data = merged_df, family = "binomial")
summary(full_model)
```

```
#current_aic <- AIC(full_model)
#while (TRUE) {
#    current_aic <- AIC(full_model)
#    reduced_model <- step(full_model, direction = "backward")
#    reduced_aic <- AIC(reduced_model)
#    if (reduced_aic > current_aic) break
#    full_model <- reduced_model
#}
#summary(full_model)
```

```
Error in step(full_model, direction = "backward") :
   number of rows in use has changed: remove missing values?
```

Variables

Top 10 features selected based on PCA loadings: gendera, Systolic blood pressure, Hyperlipemia, depression, Diastolic blood pressure, Platelets, PH, INR, PT, MCV.



Next step

- To develop a model, we plan to start with logistic regression and CARTs for a binary outcome.For logistic regression we will check for multicollinearity using VIF. We intend to include interaction terms, polynomial features or scaling variables to better capture patterns in the data.
- Advanced models like random forest, cox proportional hazards with time varying coefficients, random survival forests, hazard function, cumulative hazard function, survival function will be implemented later with the addition of the death time variable.
- For survival analysis, to evaluate the model we plan to use metrics like C-index.