Predicting the Survival of Patients with Heart Failure

Presenter: Yi Cui, Yiran Li, Zipei Zhu

UNC RESEARCH

1. Motivation

2. Data Descriptions and EDA

3. Model Selection and Further Analysis

4. Future Directions

About the Heart Failure Clinical Records Dataset

- Contains the medical records of **299** patients collected from April-December 2015 in Pakistan
- 13 real-valued features characterizing the **clinical, body, and lifestyle information** for each patient (i.e., age, sex, some medical test results, etc.)
- Some binary features include anemia, high blood pressure, diabetes, sex, and smoking
- 96 dead patients and 203 survived patients

1. Motivation

2. Data Descriptions and EDA

3. Model Selection and Further Analysis

4. Future Directions

Data Descriptions

Table 1. Data Descriptions (unscaled)

Feature	Explanation	Measurement	Range
Age	Age of the patient	Years	[40, . . . , 95]
Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0,1
High blood pressure	If a patient has hypertension	Boolean	0,1
Creatinine phosphokinase	Level of the CPK enzyme in the blood	mcg/L	[23, . . . , 7861]
Diabetes	if the patient has diabetes	Boolean	0,1
Ejection fraction	Percentage of blood leaving the heart at each contraction	Percentage	[14, . . . , 80]
Sex	Woman or man	Binary	0,1
Platelets	Platelets in the blood	kiloplatelets /mL	[25.01, . . . , 850.00]
Serum creatinine	Level of creatinine in the blood	mg/dL	[0.50, . . . , 9.40]
Serum sodium	Level of sodium in the blood	mEq/L	[114, . . . , 148]
Smoking	If the patient smokes	Boolean	0,1
Time	Follow-up period	Days	[4, . . . , 285]
(target) death event	If the patient died during the follow-up period	Boolean	0,1

Data Descriptions

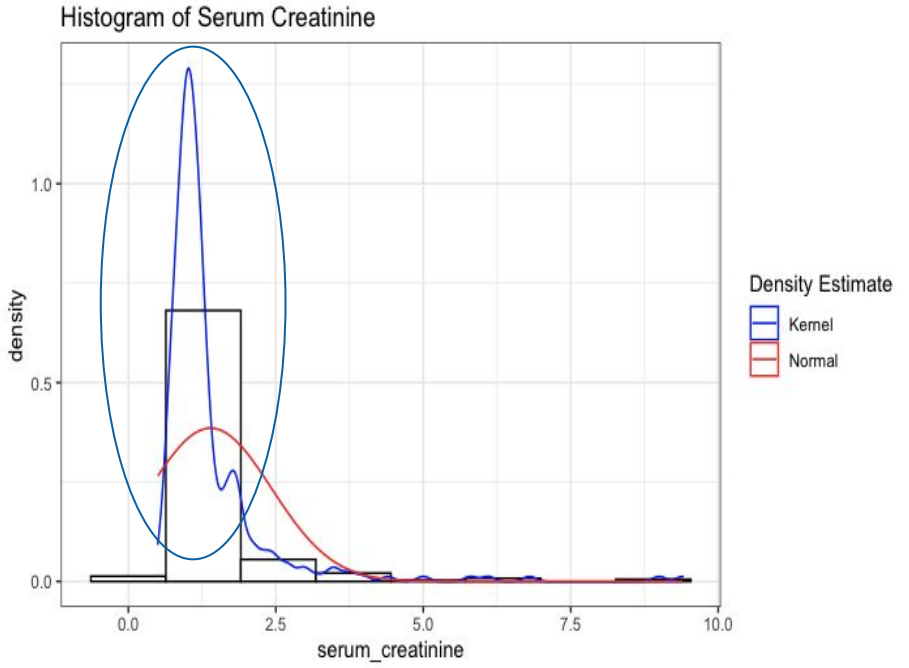
Table 2. Some Selected Binary Features

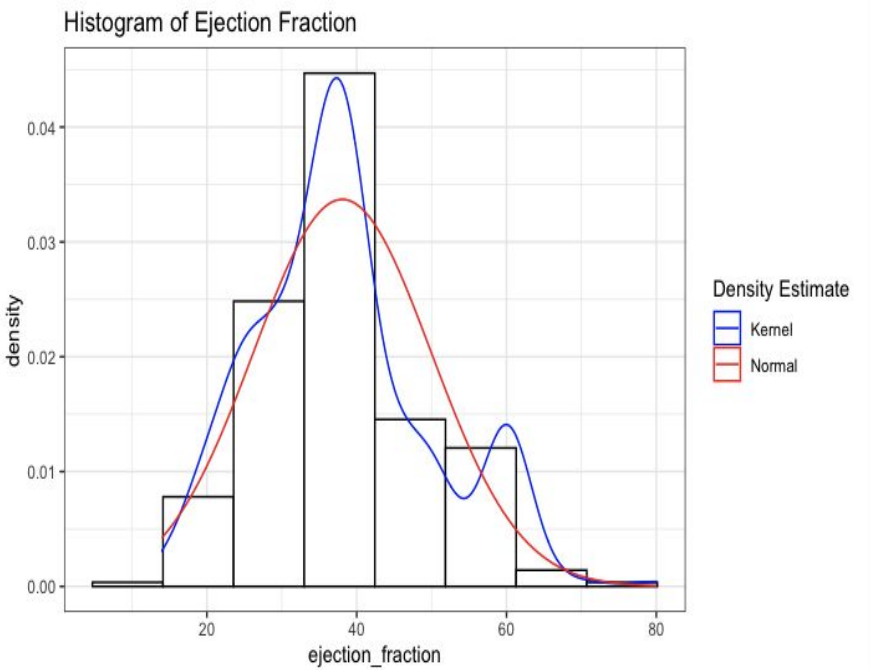
	Full sample		Dead patients		Survived patients	
Category feature	#	%	#	%	#	%
Anaemia (0: false)	170	56.86	50	52.08	120	59.11
Anaemia (1: true)	129	43.14	46	47.92	3	40.89
High blood pressure (0: false)	194	64.88	57	59.38	137	67.49
High blood pressure (1: true)	105	35.12	39	40.62	66	32.51
Diabetes (0: false)	174	58.19	56	58.33	118	58.13
Diabetes (1: true)	125	41.81	40	41.67	85	41.87
Sex (0: woman)	105	35.12	34	35.42	71	34.98
Sex (1: man)	194	64.88	62	64.58	132	65.02
Smoking (0: false)	203	67.89	66	68.75	137	67.49
Smoking (1: true)	96	32.11	30	31.25	66	32.51

Unimportant binary factor/feature

Data Descriptions

Not normal, skewness > 0





UNC RESEARCH

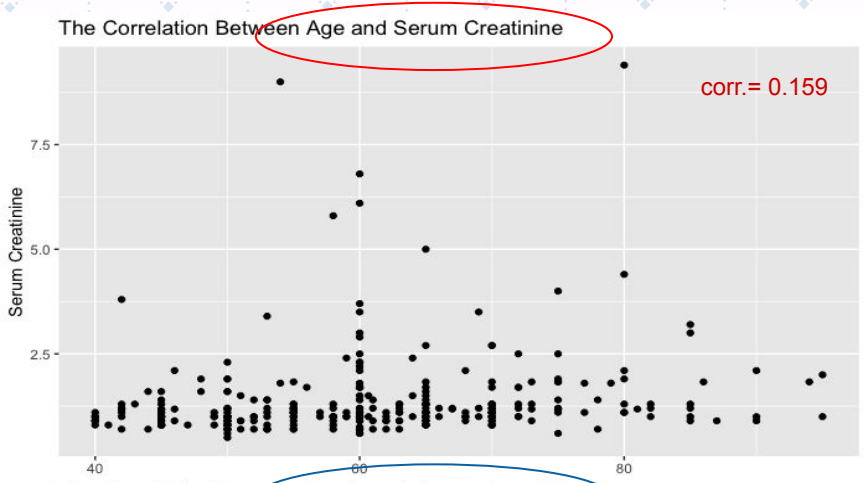
Eyeballing the data:

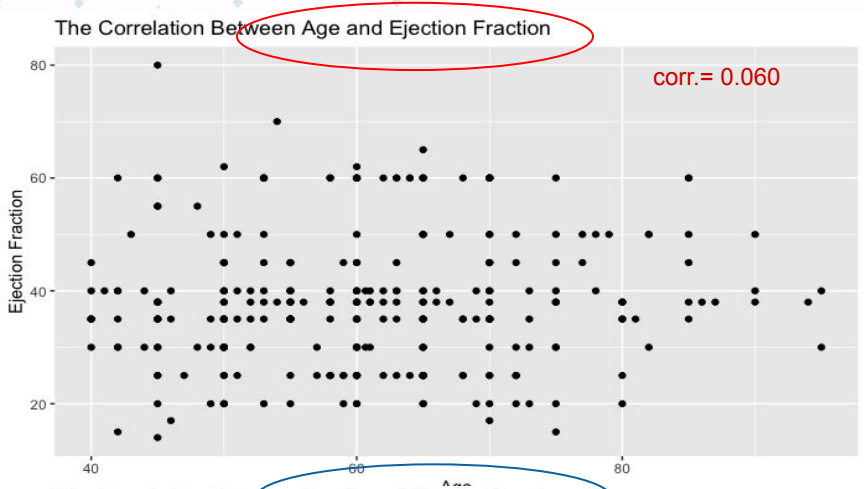
- Does higher level of serum creatinine in the blood indicates higher probability of death? Pattern

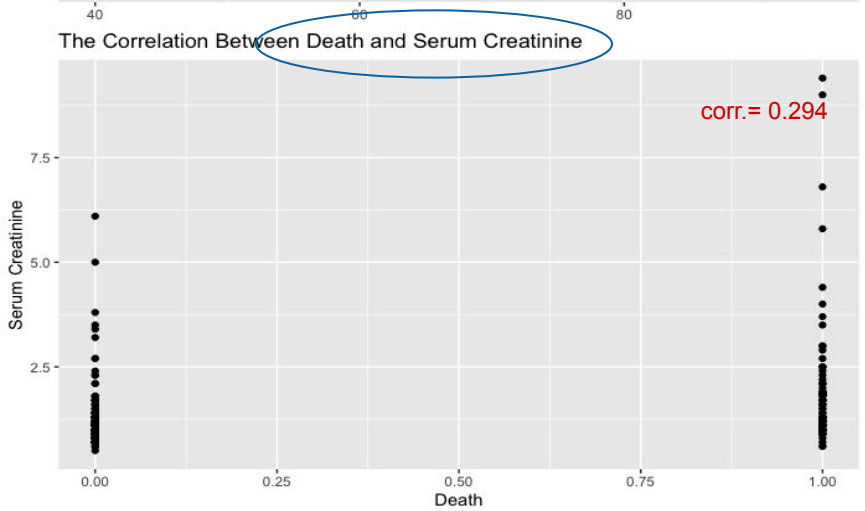
serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
9.40	133	1	1	10	1
9.00	137	0	0	196	1
6.80	146	0	0	43	1
6.10	131	1	0	107	0
5.80	134	1	0	26	1
5.00	130	0	0	207	0
4.40	133	1	0	41	1
4.00	131	1	1	10	1
3.80	128	1	1	250	0
3.70	134	1	0	96	1
3.50	134	1	0	30	1
3.50	136	1	1	187	0
3.40	145	1	0	105	0
3.20	138	0	0	94	0
3.00	132	1	0	28	1
3.00	142	0	0	30	1
2.90	127	1	1	64	1
2.70	116	0	0	8	1

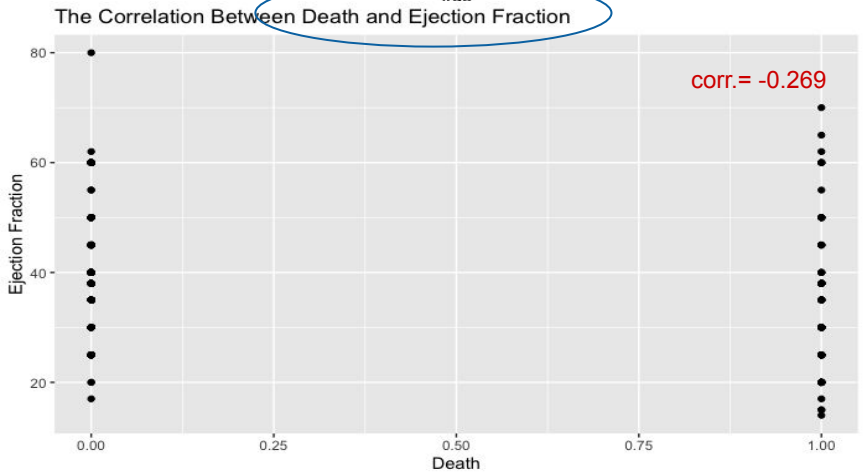
serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
1.00	140	1	0	206	0
1.00	136	0	0	210	0
1.00	142	1	1	214	0
1.00	133	1	0	215	0
1.00	139	1	0	215	0
1.00	142	1	1	216	0
1.00	139	1	0	230	0
1.00	138	1	0	233	0
1.00	140	0	0	237	0
1.00	132	1	0	244	0
1.00	137	1	1	245	0
1.00	137	1	0	247	0
1.00	132	0	0	250	0
1.00	140	1	1	258	0
0.90	140	1	1	10	1
0.90	140	1	0	14	1
0.90	140	1	1	20	1
0.90	130	1	0	38	1
0.90	139	1	1	71	0
0.90	140	0	0	74	0
0.90	138	1	0	88	0

What is the correlation between some features and death?









1. Motivation

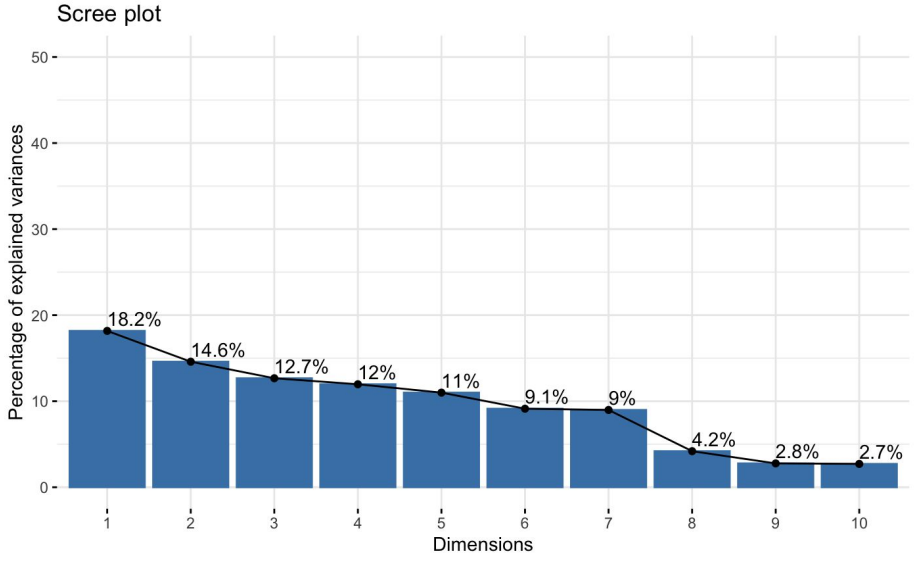
2. Data Descriptions and EDA

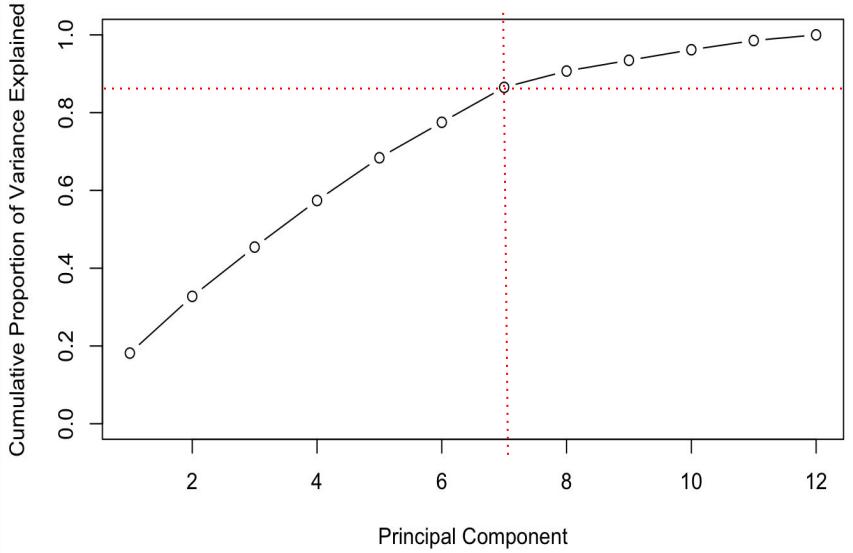
3. Model Selection and Further Analysis

4. Further Directions

Model Selection and Further Analysis:
Principal Component Analysis (PCA)

Model Selection and Further Analysis (PCA)



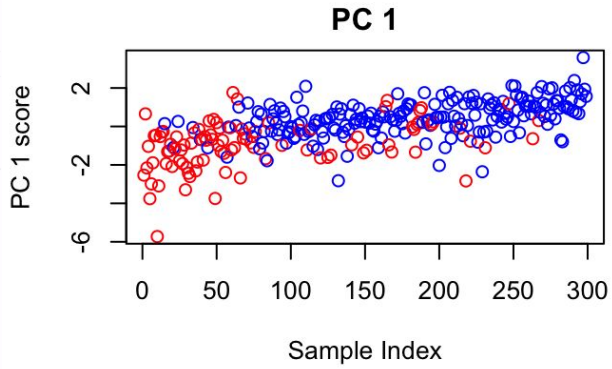


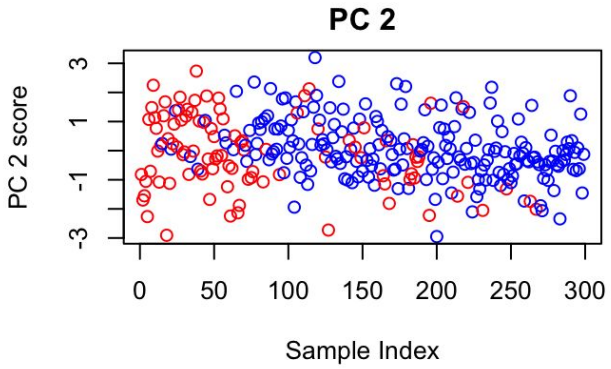
- PC1's variance percentage is only 18.2%
- 7 principal components are required to explain at least 80% of the variation in the data

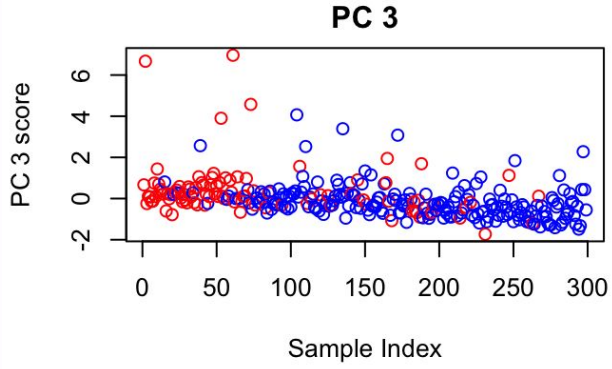
It doesn't seem like PCA has done a good job in reducing the dimensionality of the data

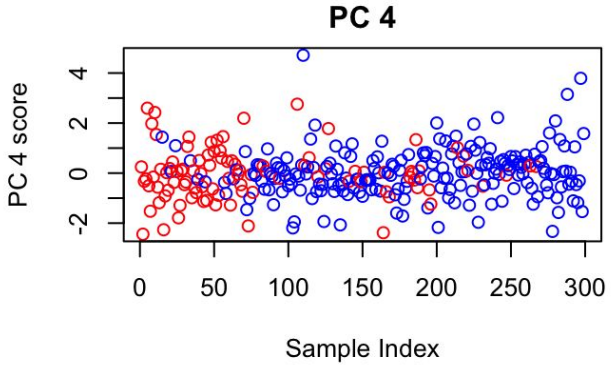
UNC RESEARCH

Model Selection and Further Analysis (PCA)







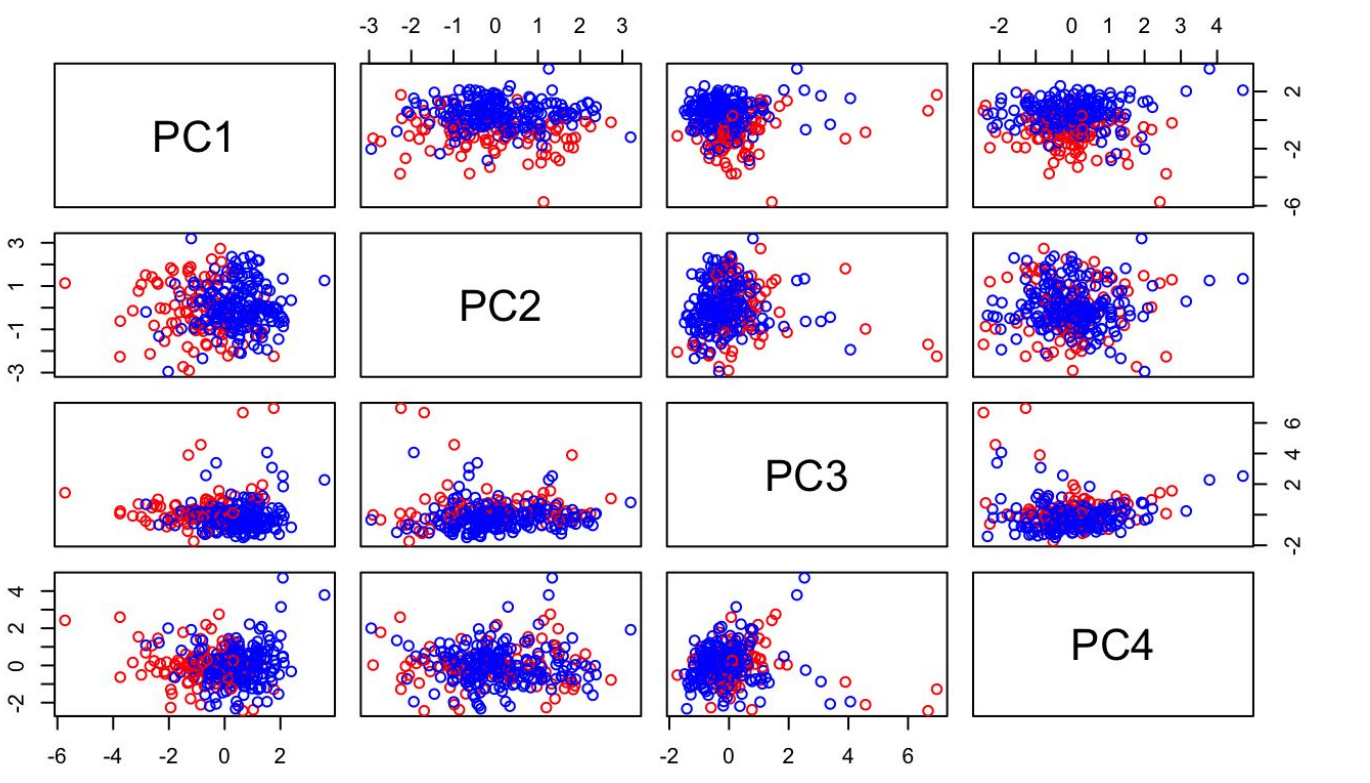


dead
survived

- Hardly no PCs appear to be helpful for separating the data

UNC RESEARCH

Model Selection and Further Analysis (PCA)

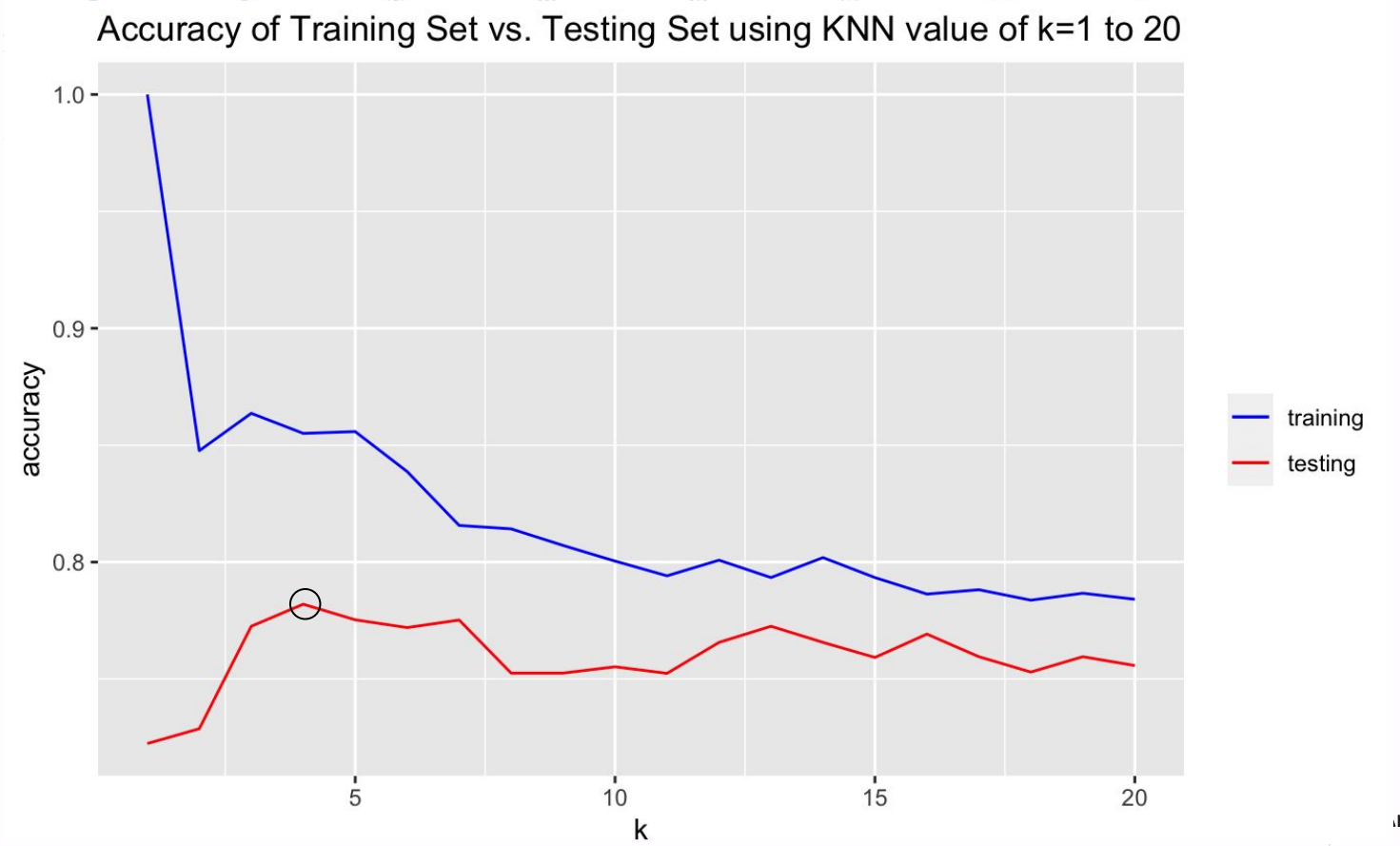


- After controlling for PC2, PC3 and PC4, **dead patients** in red seem to have **lower PC1** than the patients survived.
- There does not seem to be any clustering depending the levels of PC2, PC3 and PC4

UNC RESEARCH

Model Selection and Further Analysis:
K Nearest Neighbor (kNN)

Model Selection and Further Analysis (kNN)



Model Selection and Further Analysis:
Logistic Regression (baseline)

Model Selection and Further Analysis (Logistic Regression)

- Remove the variable "time"
- Focus on the clinical features and discover something meaningful about them

Feature	Explanation	Measurement	Range
Age	Age of the patient	Years	[40, . . . , 95]
Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0,1
High blood pressure	If a patient has hypertension	Boolean	0,1
Creatinine phosphokinase	Level of the CPK enzyme in the blood	mcg/L	[23, . . . , 7861]
Diabetes	if the patient has diabetes	Boolean	0,1
Ejection fraction	Percentage of blood leaving the heart at each contraction	Percentage	[14, . . . , 80]
Sex	Woman or man	Binary	0,1
Platelets	Platelets in the blood	kiloplatelets /mL	[25.01, . . . , 850.00]
Serum creatinine	Level of creatinine in the blood	mg/dL	[0.50, . . . , 9.40]
Serum sodium	Level of sodium in the blood	mEq/L	[114, . . . , 148]
Smoking	If the patient smokes	Boolean	0,1
~~Time~~	~~Follow-up period~~	~~Days~~	~~[4, . . . , 285]~~
(target) death event	If the patient died during the follow-up period	Boolean	0,1

Model Selection and Further Analysis (Logistic Regression)

- Build the model and take a look at their coefficients:

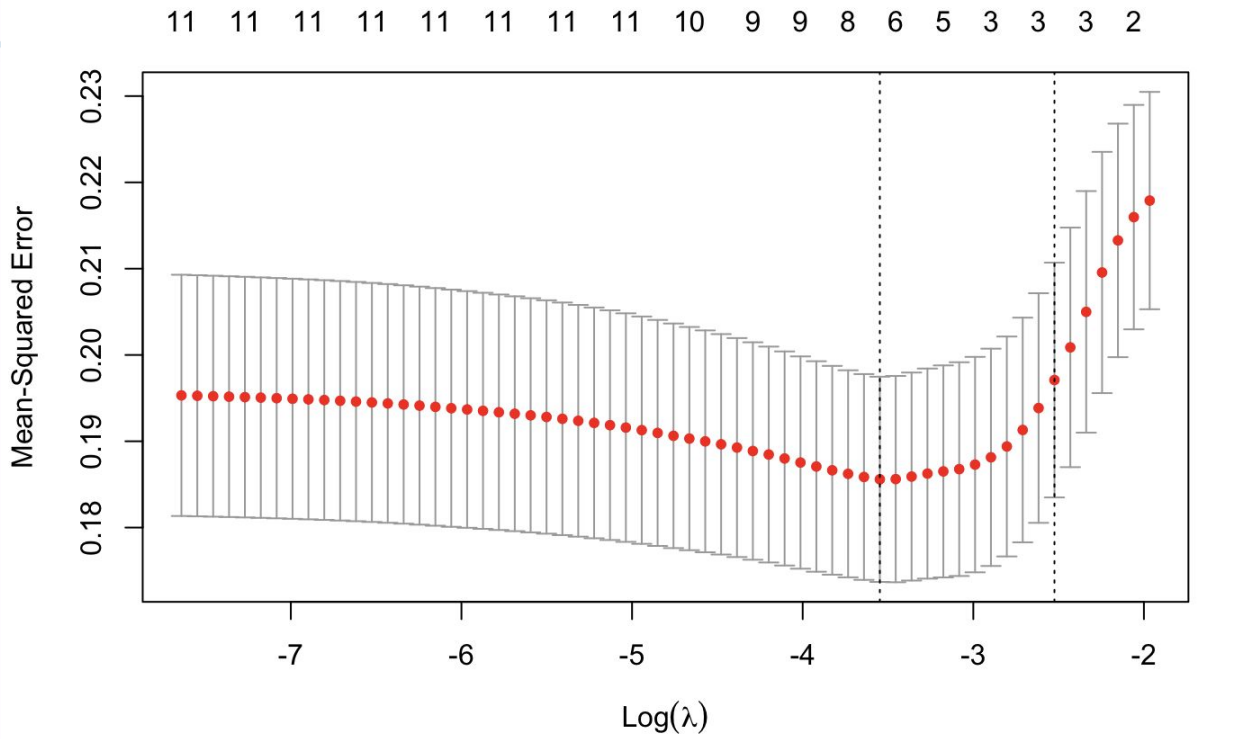
```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.3138     0.4156  -3.161 0.001571 **
age                        0.7793     0.1738   4.483 7.35e-06 ***
creatinine_phosphokinase   0.2742     0.1457   1.883 0.059737 .
ejection_fraction         -0.7917     0.2017  -3.925 8.66e-05 ***
platelets                 -0.1777     0.1800  -0.987 0.323733
serum_creatinine           0.6334     0.1816   3.487 0.000488 ***
serum_sodium              -0.1703     0.1677  -1.015 0.309962
anaemia                    0.1921     0.3451   0.557 0.577737
diabetes                   0.4117     0.3377   1.219 0.222784
high_blood_pressure        0.6124     0.3449   1.776 0.075806 .
sex                       -0.2085     0.3933  -0.530 0.596071
smoking                   -0.2581     0.4002  -0.645 0.519031
---
```

- Accuracy: 0.6833333

Model Selection and Further Analysis:
Sparse Linear Regression - the LASSO

Model Selection and Further Analysis (LASSO)

- Remove the variable "time"
- Build the model, select the best lambda using cross-validation (best lambda = 0.0288)



UNC RESEARCH

Model Selection and Further Analysis (LASSO)

- Use the best lambda to estimate the coefficients

```
12 x 1 sparse Matrix of class "dgCMatrix"
                                      s0
(Intercept)                 0.31742208
age                         0.08009002
creatinine_phosphokinase    0.00614508
ejection_fraction          -0.09596062
platelets                   .
serum_creatinine            0.09028105
serum_sodium               -0.02537043
anaemia                     .
diabetes                    .
high_blood_pressure         0.01038856
sex                         .
smoking                     .
```

Results from logistic regression:

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.3138     0.4156  -3.161 0.001571 **
age                        0.7793     0.1738   4.483 7.35e-06 ***
creatinine_phosphokinase   0.2742     0.1457   1.883 0.059737 .
ejection_fraction         -0.7917     0.2017  -3.925 8.66e-05 ***
platelets                 -0.1777     0.1800  -0.987 0.323733
serum_creatinine           0.6334     0.1816   3.487 0.000488 ***
serum_sodium              -0.1703     0.1677  -1.015 0.309962
anaemia                    0.1921     0.3451   0.557 0.577737
diabetes                   0.4117     0.3377   1.219 0.222784
high_blood_pressure        0.6124     0.3449   1.776 0.075806 .
sex                       -0.2085     0.3933  -0.530 0.596071
smoking                   -0.2581     0.4002  -0.645 0.519031
---
```

- The associated test error: 0.1905741

Model Selection and Further Analysis:
Decision Trees

Decision Trees

- The usefulness of decision trees
  - Recall pros: **a white box model**, simple to understand and interpret, able to validate the results from other methods; Cons: larger variance in results (e.g., the tree structure may change a lot when there is just a small change in data)

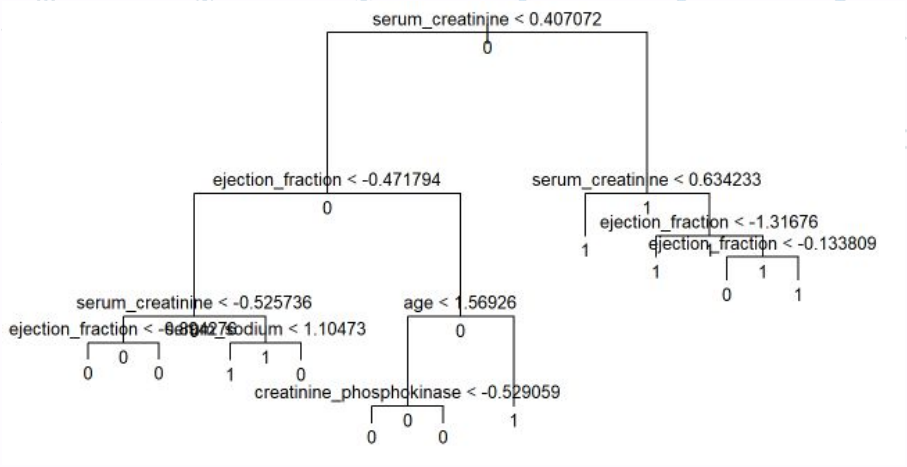
- Results#1: fit a classification tree on **the whole data set**

```
## Classification tree:
## tree(formula = death ~ . - DEATH_EVENT - time - sex, data = scale_heart_temp)
## Variables actually used in tree construction:
## [1] "serum_creatinine"         "ejection_fraction"
## [3] "serum_sodium"             "age"
## [5] "creatinine_phosphokinase"
## Number of terminal nodes:  11
## Residual mean deviance:  0.8297 = 238.9 / 288
## Misclassification error rate: 0.1773 = 53 / 299
```

- There are **six variables** actually used as **internal nodes** in the tree: serum_creatinine, ejection_fraction, age, serum_sodium, creatinine_phosphokinase, and sex.
- Overall, the **training error rate** is **17.73%**. This relatively small deviance indicates a tree that provides a good fit to the training data.

Decision Trees

- The structure of the tree:



- We can see the most important feature is **serum_creatinine (+)** since the first branch differentiates the level of serum creatine in the patient's blood using a threshold of 0.407. Plus, this feature is used once again in the following internal nodes.
- The other most important feature seems to be **ejection_fraction (-)**, which is used three times in the following internal nodes.

Decision Trees

- Results#2: Cross validation
  - To properly evaluate the tree, we must estimate the test error besides the training error
  - Split the data into a training set (80% of all samples) and a test set (20%)
  - This approach leads to correct predictions for around 75% of the data points in the test data test.
  - (11+34)/(11+34+7+8) = 0.75

```
##              death.test
## tree.predict  1  0
##            1 11  8
##            0  7 34
```

1. Introduction

2. Data Descriptions

3. Model Selection and Further Analysis

4. Further Directions

Some Takeaways and Future Directions

- Results
  a. Since our data is not very highly dimensional, unsupervised methods such as PCA and clustering do not yield good results, as expected.
  b. Both EDA and our baseline model (i.e., the logistic regression) show that the level of serum creatinine in the patient's blood (**serum_creatinine**) and the the speed of bloodstream through the heart (**ejection_fraction**) are two most important predictors along with the age.
  c. Other methods such as LASSO and decision trees yield **similar results** to the logistic regression model. LASSO successfully reduces the variables to six, while the decision tree also shows good performance and indicates the importance of the two features above.

- Future directions?
  a. Try bagging, random forests, and boosting on the decision tree
     - To see if we can boost the performance of our model (at the expense of less interpretability)
  d. Try SVM to see if this method could result in better predictive accuracy
     - SVM is relatively newer than the other learning methods (e.g., decision trees)

Thank you!

Yi Cui, Yiran Li, Zipei Zhu
yicui@unc.edu
yiran1@live.unc.edu
zipei_zhu@kenan-flagler.unc.edu

UNC RESEARCH