Assignment 1

Seongming Lee (01247436), Yuxuan Luo (01376247) and Nina Hauser (01418616)

Question 1

(1) A linear separation could use $\phi(x) = ||x||_2$, since the inner class (black rectangle) has a strictly smaller absolute value than the outer class (blue rectangle).

(2) The two dimension graph can be transformed to a one dimension by using $\phi(x) = x_1 * x_2$. As the blue class consists of x_1 and x_2 values with different signs, it will have strictly negative values, while the black class with equal signs for x_1 and x_2 has strictly positive values.

(3) Using $\phi_1(x_1) = x_1$,$\phi_2(x_1) = x_1^2$, a parabola can be used to separate the two classes. Consequently, blue class values will now be strictly bigger than black class values and the classes can therefore be split successfully.

Question 2

1. Cost of Readmissions

23 percent of the patients were readmitted in the timespan of one year. As the loss in Medicare reimbursements are estimated to be 8,000 USD per readmitted patient, the values of 7,984,000 USD is derived by the formula:

$$Cost\ without\ CareTracker = Number\ of\ Patients\ * Share\ of\ Readmissions\ *\ Cost\ of\ Readmission$$

$$= 382\ * 0.2277 * 8000 = 7,984,000$$

2. Cost of Caretaker

Caretaker reduces the likelihood of readmissions by 40 percent, while costing 1,200 USD per patient. Subsequently, 4,790,400 USD lost by medical reimbursements remain and program costs add up to 5,258,400 USD. Based on the cost analysis in the prior questions, CareTracker would actually increase costs for Tahoe. Consequently, Tahoe should not implement CareTracker for all AMI patients.

$$Cost\ with\ Caretracker = Cost\ of\ Readmissions\ + Program\ Costs$$

$$= 4,790,400 + 5,258,400 = 10,048,800$$

$$Cost\ Difference = Cost\ without\ Caretracker\ - Cost\ with\ CareTracker$$

$$= 7,984,000 - 10,048,800 = -2,064,800$$

3. Cost for patient-specific use of CareTracker

With a perfect classifier, Tahoe could reduce costs as it would on give CareTracker do those patients who would be readmitted in the near future. The 1,200 USD implementations costs are therefore only applicable to the correctly identified `100*round(readmissions,2)` share of patients. The upper savings bond is `cost_diff2` USD.

$$New\ Cost\ with\ Caretracker = Cost\ of\ Readmissions\ + Program\ Costs$$

$$= 4,790,400 + 1,197,600 = 5,988,000$$

$$Cost\ Difference = Cost\ without\ Caretracker\ - New\ Cost\ with\ CareTracker$$

$$= 7,984,000 - 5,988,000 = 1,996,000$$

4. A Simple Classification Algorithm

The best value for S* is 41 with savings of 136,800 USD.

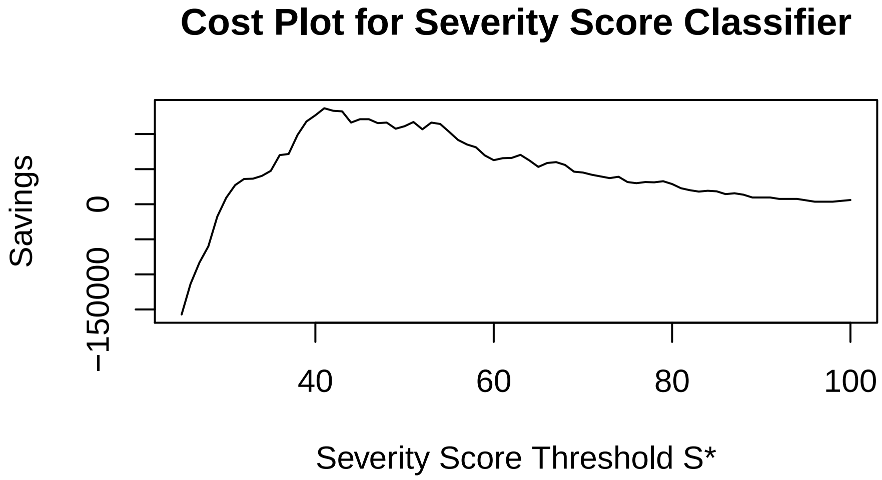


Table 1: Best Simple Classifier

S*	Savings
41	136800

5. A Sophisticated Classification Algorithm

The GLM classifier predicts readmission for any probability equal or bigger than 0.5. 80.35 percent of the observations are accurately classified, with only 3.93 percent of cases where a high risk of readmission is missed and 15.72 percent observations that are given CareTracker without cause.

Table 2: Confusion Matrix

	0	1
0	3212	172
1	689	309

6. Cost Savings with Sophisticated Algorithm

The best value for p* is 0.4 with savings of 495,200 USD.

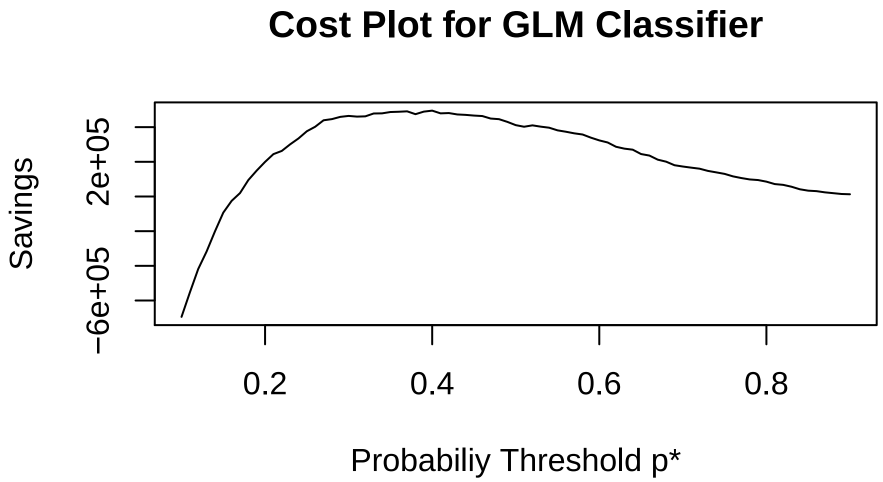
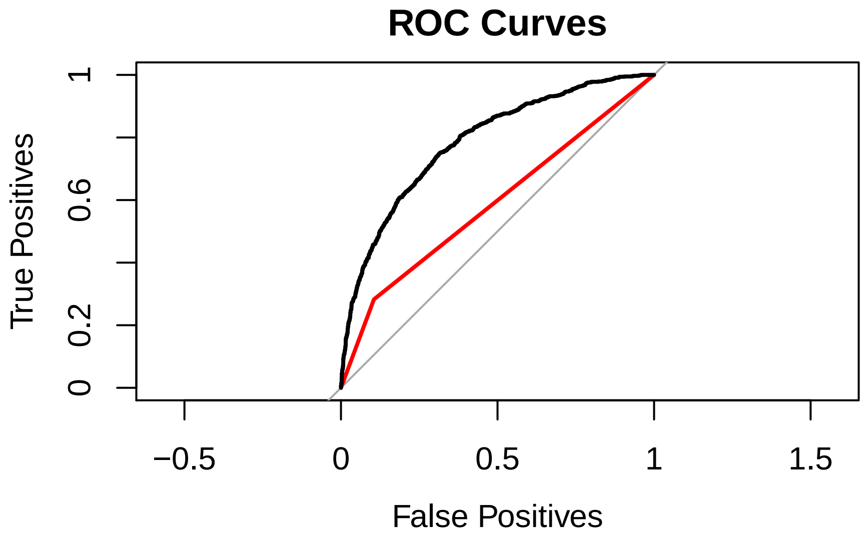


Table 3: Best Advanced Classifier

S*	Savings
0.4	495200

Comparing the simple (red) to the GLM classifier (black), it becomes visible that the GLM clearly exceeds the performance of the simple classifier.



	Model 1
(Intercept)	-0.12^{*}
	(0.06)
age	-0.00
	(0.00)
female	0.02
	(0.01)
flu_season	0.11^{***}
	(0.01)
ed_admit	-0.02
	(0.02)
severity.score	0.00^{***}
	(0.00)
comorbidity.score	0.00^{***}
	(0.00)
AIC	3883.30
BIC	3934.38
Log Likelihood	-1933.65
Deviance	620.13
Num. obs.	4382

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: GLM model

Question 3

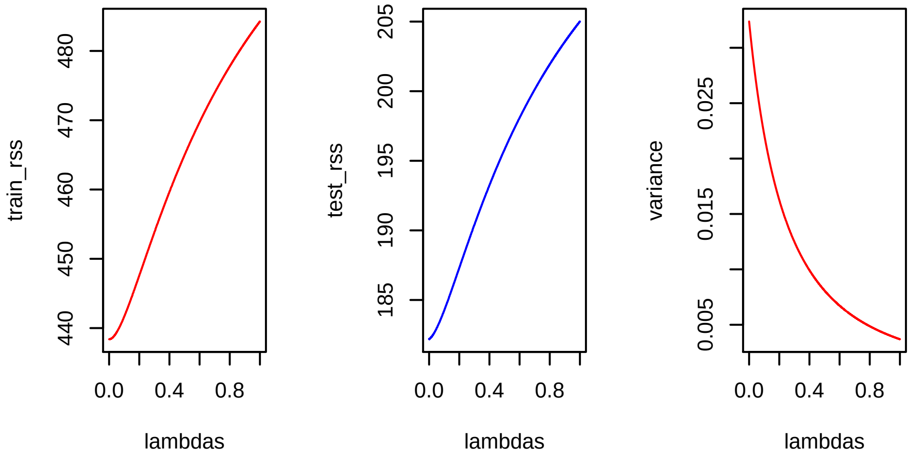
(a) With an increasing penalty value λ, the training RSS will steadily increase. By increasing λ, the parameter estimates will all be shrunk to zero, deviating from non-regularized model estimates and thus bringing a steady increase in training RSS.

(b) For test RSS, it will decrease initially, and then start increasing in a U shape. Since coefficients estimates are forced to decrease in the beginnging, the test RSS will decrease slightly as the model is less overfitting. However, if λ keeps increasing, some necessary coefficients will be shrunk to 0 and removed from the model, leading to an increase in the test RSS.

(c) Variance will decrease steadily. Since large coefficients generally have more variability than smaller coefficients, adding the penalization term reduces variance.

(d) Bias will steadily increase. Higher values of the penalty parameter λ constrains parameter estimates, and thus a regularized model with higher λ will be more biased compared to non-regularized models.

(e) The irreducible error remains constant, since it is independant of the model, and consequently independant of the value of λ.


train_rss
440
450
460
470
480
0.0
0.4
0.8
lambdas
test_rss
185
190
195
200
205
0.0
0.4
0.8
lambdas
variance
0.005
0.015
0.025
0.0
0.4
0.8
lambdas