PM511b Data Analysis Project

George G Vega Yon

April 19, 2019

Question 1

Table 1 shows a set of descriptive statistics for the variables in the dataset, in particular, comparing means and proportions between surviving and deceased patients. From it we can see the following regarding the training dataset (`val = 0`):

1. Only a handful of cases (174 out of 2789) corresponds to patients that have died. This means that, considering the hetorgeneity of the features that we will be using in our predictive modeling, we don't have anough cases to train our data with.
2. Of all the variables, *Age*, *Glasgow Coma Scale*, and *(Ethnycity = Asian)* are the only ones that show to be significantly different across groups (pvalues less than 0.001, 0.001, and 0.01 respectively). This makes them more likely to be meaningful (significant) in the prediction model that will be described in the next section.
3. Finally, the training dataset has a significant high proportion of males in both groups (above 76%). This creates a future problem in terms of external validity of the prediction model on females.

The next section discusses the process which I used to build the predictive models.

Table 1: Descriptive statistics by status (In-hospital mortality). In the case of continuous variables, standard errors are showed in parenthesis, otherwise it shows the proportion within the group. For the significance of the differences $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

	Survived	Died	Dif. in Means (statistic)
Age	41.08 (17.40)	50.45 (20.79)	-5.81 ***
Male	1998 (76.41%)	137 (78.74%)	0.37
SBP (mm Hg)	133.97 (25.29)	127.74 (41.72)	1.95
RR (breaths per minute)	20.18 (6.34)	21.30 (9.46)	-1.55
Glasgow Coma Scale	14.13 (2.52)	8.94 (5.14)	13.22 ***
Ethnicity			
Asian	165 (6.31%)	20 (11.49%)	7.42 **
African American	482 (18.43%)	31 (17.82%)	0.02
Non-Hispanic white	918 (35.11%)	65 (37.36%)	0.29
Hispanic white	1050 (40.15%)	58 (33.33%)	3.23
Total	2615	174	

Questions 2-4

For the model selection part, I use an automatic feature selection process that maximizes the prediction accuracy. In particular, the following algorithm builds the model incremnetally by adding one variable at a time from a pool of variables that includes: the baseline features, monotonic transformations of those, and interaction terms.

Ultimately, I kept a set of best models instead of a single best, and what's more, throughout the process the algorithm never pics a single model, on the contrary, it carries around a set of 100 bests models. The algorithm used is described in what follows:

1. Start by increasing the features space by: (a) transforming continuous variables with the functions $()^{-1}$, square-root, and $()^2$ and (b) generating interaction terms across all features. Denote the set of features S. In our data, this set has a cardinality of 186.[1]
2. In order to maintain the feature space relatively compact, I then estimated logistic regression models using one of S at a time, including an intercept, and only kept those variables that showed to be significant at the 95% level. Recall that S also includes interaction terms and transformed verions of the baseline data. At this point the pool of features reduced to 110.
3. Once the pool of features to be used in the predicting model has been generated, initiate the set of best parameters $B = S \times S$ ($\binom{|S|}{2}$ possible models to be estimated), then:
    a. Generate all possible combinations of (B, S), this is $B' = B \times S$
    b. For all $b \in B'$, estimate a logistic regression model, and calculate Area Under the Curve (AUC) using the observed and predicted values.
    c. In order to keep the number of models manageble, keep the best 100 in terms of AUC, and replace B with that set. This is a crucial step since if we don't reduce the number of models that will be used for the next iteration of the algorithm, the number of models to estimates becomes unmanageable[2]
    d. Contiue with the next step

I repeated steps a through d 8 times so that we have 800 possible models to be selected for our predictive problem. Figure 1 shows the distribution of the aforementioned models.

The process as a whole consisted on estimating about 75,000 models, as shown in the Table 2.

Table 2: Number of unique fitted models per number of features included. In total we estimated about 75,000 different specifications to try to find the one which maximized the prediction accuracy measured as AUC.

Num. Features	Num. Fitted Models
2	5995

[1] I the case of the variable GCS, we could have transformed it into a categorical variable following the description in the dataset. I chose not to do so under the rationale that (1) if associated with the dependent variable, the association should be monotonic and (2) if not completely linear, the applied monotonic transformations should be able to capture part of its effect.

[2] In our case, wanted us to compute all the possible combinations (excluding the empty model), we would need to estimate 1.2980742×10^{33} models.

Num. Features	Num. Fitted Models
3	10008
4	10425
5	10395
6	10063
7	9257
8	9598
9	9852
Total	75593

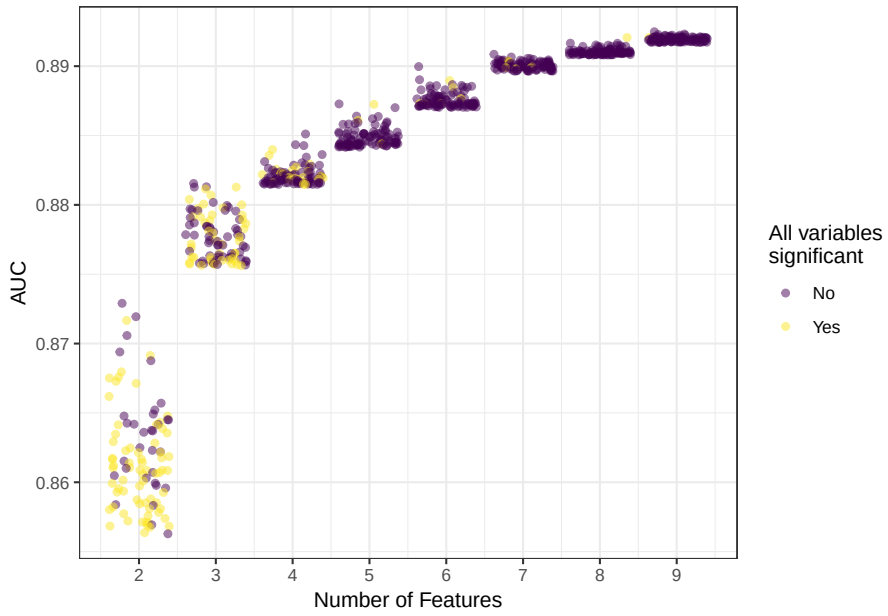

Figure 1: Distribution of the 800 different models. Overall we can see that the accuracy of the prediction model rises exponentially as a function of the number of features incude in the model. At the same time, models with a large number of variables tend to be less signficant in terms of having all coefficients simultaneously significant, i.e. we say that a model is significant if all variables are statistically significant in the model.

Question 5

Table 5 and Table 6 show the parameter estimates of 20 different models that showed to have the highest accuracy values, and Table 3 and Table 4 show their corresponding prediction accuracy levels in (`val = 0`) and out of sample (`val = 1`).

Table 3: In and out-of-sample AUC levels with the corresponding 95% CI for the top 10 models regardless of whether all features are significant. The corresponding coefficients of these models are reported on Table 5

Model	All feat. signif.	AUC	AUC out of sample
1	No	0.892 [0.871; 0.914]	0.822 [0.745; 0.899]
2	No	0.892 [0.871; 0.914]	0.819 [0.740; 0.898]
3	No	0.892 [0.871; 0.913]	0.821 [0.744; 0.899]
4	No	0.892 [0.871; 0.913]	0.825 [0.749; 0.901]
5	No	0.892 [0.871; 0.913]	0.825 [0.748; 0.901]
6	No	0.892 [0.871; 0.913]	0.817 [0.737; 0.897]
7	No	0.892 [0.871; 0.913]	0.823 [0.746; 0.900]
8	No	0.892 [0.871; 0.913]	0.825 [0.749; 0.901]
9	No	0.892 [0.871; 0.914]	0.815 [0.732; 0.897]
10	No	0.892 [0.871; 0.913]	0.825 [0.749; 0.901]

Table 4: In and out-of-sample AUC levels with the corresponding 95% CI for the top 10 models constrained to the set of those where all the features should be significant. The corresponding coefficients of these models are reported on Table 6.

Model	All feat. signif.	AUC	AUC out of sample
1	Yes	0.892 [0.871; 0.913]	0.824 [0.748; 0.901]
2	Yes	0.892 [0.871; 0.913]	0.824 [0.748; 0.901]
3	Yes	0.890 [0.869; 0.912]	0.823 [0.744; 0.901]
4	Yes	0.890 [0.869; 0.912]	0.820 [0.740; 0.900]
5	Yes	0.890 [0.869; 0.912]	0.830 [0.756; 0.905]
6	Yes	0.890 [0.869; 0.911]	0.822 [0.745; 0.899]
7	Yes	0.890 [0.868; 0.912]	0.821 [0.741; 0.901]
8	Yes	0.890 [0.869; 0.911]	0.817 [0.736; 0.897]
9	Yes	0.890 [0.869; 0.911]	0.827 [0.752; 0.902]
10	Yes	0.889 [0.867; 0.911]	0.823 [0.745; 0.901]

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
african american × RR (breaths per minute)^2			0.00							
			(0.00)							
Age × RR (breaths per minute)	0.00***	0.00***	0.00***	0.00***	0.00***	0.00***	0.00***	0.00***	0.00***	0.00***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Age × SBP (mm Hg)$^{-1}$										0.18
										(0.24)
Age^{-1} × Glasgow Coma Scale^{-1}	57.27								111.89*	
	(45.15)								(47.62)	
Age^2 × SBP (mm Hg)$^{-1}$								0.00		
								(0.00)		
Glasgow Coma Scale × SBP (mm Hg)^2	0.00**	0.00**	0.00**	0.00**	0.00**	0.00***	0.00**	0.00**	0.00***	0.00**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Glasgow Coma Scale^{-1}						18.49**				
						(6.74)				
Glasgow Coma Scale$^{1/2}$						2.30*				
						(1.00)				
hispanic white × Age^{-1}									−8.90	
									(6.61)	
Male × Age$^{1/2}$					−0.02					
					(0.06)					
Male × Glasgow Coma Scale									0.03	
									(0.02)	
Male × RR (breaths per minute)$^{-1}$				−3.37						
				(4.07)						
Male × SBP (mm Hg)$^{-1}$	56.85*	57.05*	53.03*	72.19*	68.02	53.05*	58.69*	57.46*		56.46*
	(23.17)	(24.79)	(23.06)	(32.08)	(42.65)	(23.82)	(23.68)	(23.59)		(23.56)
RR (breaths per minute)$^{1/2}$ × Glasgow Coma Scale^{-1}							0.48			
							(0.55)			
RR (breaths per minute)$^{-1}$ × Glasgow Coma Scale^{-1}	49.44**	42.21**	46.75**	59.15**	47.77**	44.33**	53.71**	46.81**	58.81***	47.23**
	(16.21)	(16.29)	(15.84)	(22.69)	(16.24)	(16.52)	(18.42)	(15.96)	(17.29)	(15.99)
SBP (mm Hg) × Glasgow Coma Scale	−0.00*	0.00	−0.00**	−0.00**	−0.00**	−0.00	−0.00	−0.00**	0.00*	−0.00**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
SBP (mm Hg) × Glasgow Coma Scale$^{1/2}$									−0.03***	
									(0.01)	
SBP (mm Hg)$^{-1}$ × Glasgow Coma Scale^{-1}		627.40*								
		(280.91)								
SBP (mm Hg)^2 × Glasgow Coma Scale^{-1}	0.00	0.00*	0.00*	0.00	0.00*		0.00	0.00*		0.00*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		(0.00)	(0.00)		(0.00)
SBP (mm Hg)^2 × Glasgow Coma Scale$^{1/2}$	−0.00*	−0.00*	−0.00*	−0.00*	−0.00*	−0.00***	−0.00*	−0.00*		−0.00*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		(0.00)
SBP (mm Hg)^2 × Glasgow Coma Scale^2	−0.00***	−0.00***	−0.00***	−0.00**	−0.00***	−0.00***	−0.00***	−0.00***	−0.00***	−0.00***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
(Intercept)	−3.13***	−5.23***	−2.32***	−2.63***	−2.45***	−12.02**	−3.21**	−2.46***	−0.90	−2.50***
	(0.92)	(1.44)	(0.65)	(0.76)	(0.72)	(3.75)	(1.22)	(0.67)	(0.89)	(0.69)
AIC	936.92	932.06	937.05	937.92	938.38	934.62	937.75	937.65	933.35	937.81
BIC	996.26	991.39	996.38	997.26	997.71	993.96	997.08	996.99	992.69	997.14
Log Likelihood	-458.46	-456.03	-458.52	-458.96	-459.19	-457.31	-458.87	-458.83	-456.68	-458.90
Deviance	916.92	912.06	917.05	917.92	918.38	914.62	917.75	917.65	913.35	917.81
Num. obs.	2789	2789	2789	2789	2789	2789	2789	2789	2789	2789

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

Table 5: Best 10 models regardless of whether all the features show as statistically significant at the 95 % level. The corresponding AUCs are reported in Table 3.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Age $\times$ RR (breaths per minute)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Age$^{1/2}$ $\times$ Glasgow Coma Scale^{-1}									0.85^{*}	
									(0.35)	
Glasgow Coma Scale $\times$ SBP (mm Hg)^2	0.00^{**}	0.00^{**}	0.00^{**}	0.00^{***}	0.00^{***}	0.00^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Glasgow Coma Scale$^{1/2}$				-1.67^{*}				-1.69^{**}		
				(0.66)				(0.66)		
Male $\times$ Age$^{1/2}$								0.06^{*}		
								(0.03)		
Male $\times$ RR (breaths per minute)$^{-1}$			7.58^{*}	6.93^{*}			7.07^{*}			
			(3.05)	(3.07)			(3.12)			
Male $\times$ SBP (mm Hg)$^{-1}$	54.87^{*}	54.87^{*}			52.43^{*}	52.59^{*}			62.38^{**}	58.95^{*}
	(23.10)	(23.10)			(22.86)	(22.97)			(24.06)	(23.64)
RR (breaths per minute)$^{-1}$ $\times$ Glasgow Coma Scale^{-1}	46.60^{**}	46.60^{**}			49.76^{**}					
	(15.80)	(15.80)			(16.12)					
SBP (mm Hg) $\times$ Glasgow Coma Scale	-0.00^{***}	-0.00^{***}	-0.00^{***}	0.01^{*}	-0.00^{***}	-0.00^{***}	0.01^{*}	0.01^{*}	-0.00^{*}	0.01^{*}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
SBP (mm Hg) $\times$ Glasgow Coma Scale$^{1/2}$				-0.04^{***}			-0.02^{*}	-0.04^{***}		
				(0.01)			(0.01)	(0.01)		
SBP (mm Hg)$^{1/2}$ $\times$ Glasgow Coma Scale$^{1/2}$							-0.50^{*}			-0.66^{**}
							(0.22)			(0.22)
SBP (mm Hg)^2 $\times$ Glasgow Coma Scale^{-1}	0.00^{*}	0.00^{*}	0.00^{*}			0.00^{*}				
	(0.00)	(0.00)	(0.00)			(0.00)				
SBP (mm Hg)^2 $\times$ Glasgow Coma Scale$^{1/2}$	-0.00^{*}	-0.00^{*}	-0.00^{**}		-0.00^{*}	-0.00^{*}			-0.00^{**}	
	(0.00)	(0.00)	(0.00)		(0.00)	(0.00)			(0.00)	
SBP (mm Hg)^2 $\times$ Glasgow Coma Scale^2	-0.00^{***}	-0.00^{***}	-0.00^{***}	-0.00^{***}	-0.00^{***}	-0.00^{**}	-0.00^{***}	-0.00^{***}	-0.00^{***}	-0.00^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
(Intercept)	-2.34^{***}	-2.34^{***}	-0.93^{*}	4.88^{**}	-2.30^{***}	-1.14^{*}	5.73^{*}	5.21^{**}	-2.73^{**}	5.76^{*}
	(0.65)	(0.65)	(0.45)	(1.82)	(0.66)	(0.49)	(2.32)	(1.79)	(0.87)	(2.40)
AIC	936.52	936.52	943.34	936.42	938.76	949.76	936.97	942.64	947.74	943.58
BIC	989.93	989.93	990.81	983.89	986.22	997.23	984.44	990.10	995.21	985.12
Log Likelihood	-459.26	-459.26	-463.67	-460.21	-461.38	-466.88	-460.48	-463.32	-465.87	-464.79
Deviance	918.52	918.52	927.34	920.42	922.76	933.76	920.97	926.64	931.74	929.58
Num. obs.	2789	2789	2789	2789	2789	2789	2789	2789	2789	2789

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

Table 6: Best 10 models in which all the features show as statistically significant at the 95 % level. The corresponding AUCs are reported in Table 4.

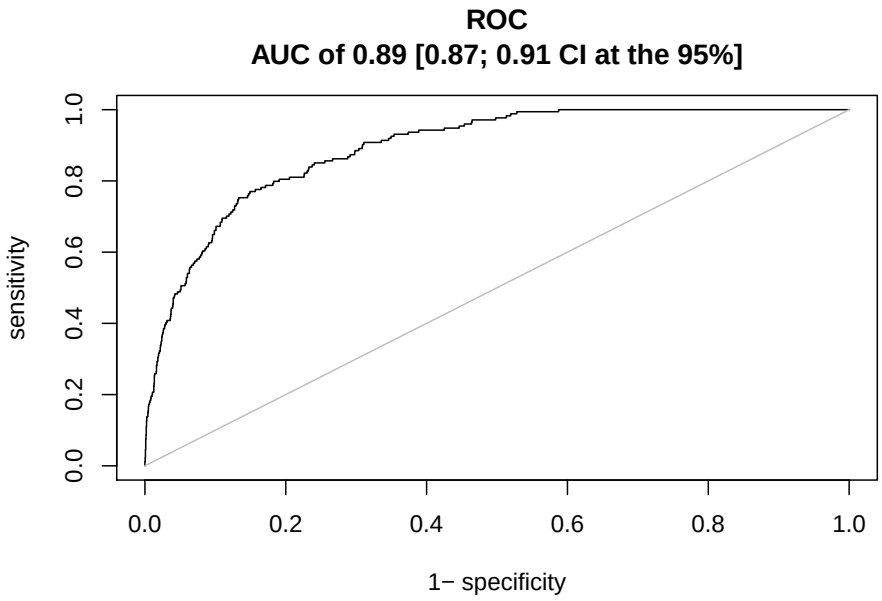

Figure 2: ROC AUC curve for the selected model. The model shows a good performance within sample with an AUC value that lives within 0.89 and 0.91. A very close range, but rather high accuracy level.

Question 6

If we chose the model with the highst AUC considering only those that have all of its features' as statistically signifiant predictors, we obtain the following linear prediction model:

$$\hat{\theta}^{\mathsf{t}}x = -2.34 \times \text{(Intercept)} + 0.00 \times \text{Age} \times \text{RR (breaths per minute)} + 0.00 \times \text{Glasgow Coma Scale} \times \text{SBP (mm Hg)}^2 + 54.87 \times \text{Male} \times \text{SBP (mm Hg)}^{-1} + 46.60 \times \text{RR (breaths per minute)}^{-1} \times \text{Glasgow Coma Scale}^{-1} - 0.00 \times \text{SBP (mm Hg)}^2 \times \text{Glasgow Coma Scale}^2 + 0.00 \times \text{SBP (mm Hg)}^2 \times \text{Glasgow Coma Scale}^{-1} - 0.00 \times \text{SBP (mm Hg)}^2 \times \text{Glasgow Coma Scale}^{1/2} - 0.00 \times \text{SBP (mm Hg)} \times \text{Glasgow Coma Scale} \tag{1}$$

From the linear prediction, we can then compute the probability of dying by exponentiating the previous equation and dividing the number by 1 + itself, this is:

$$\Pr\left(die = 1 | X = x\right) = \frac{\exp \hat{\theta}^{\mathsf{t}}x}{1 + \exp \hat{\theta}^{\mathsf{t}}x} \tag{2}$$

Questions 7 and 8

Figure 2 and Table 7 show the AUC curve and classification table for the selected model on the training data (`val = 0`), while Figure 3 shows the AUC ROC curve in the case of using the model to predict the dependent variable with out-of-the-sample data (`val = 1`).

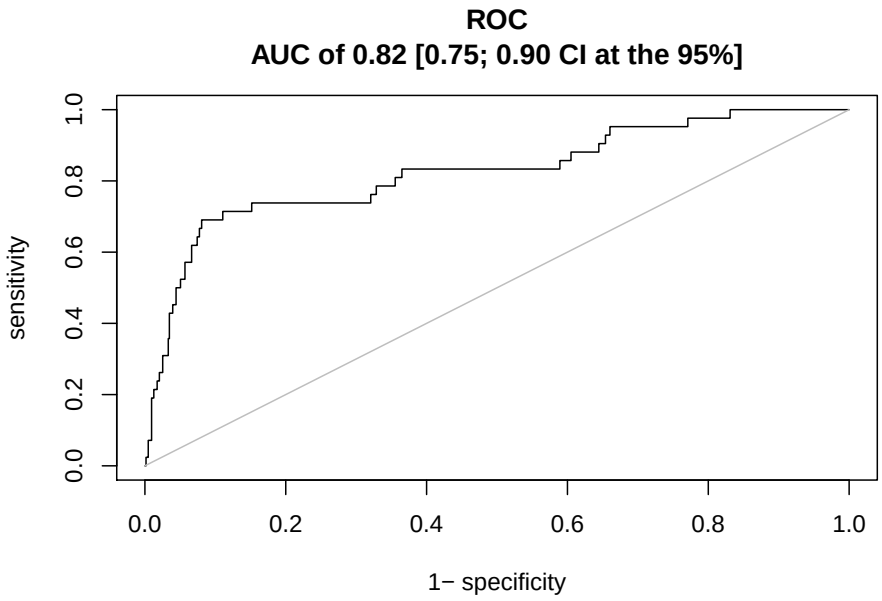

Figure 3: AUC ROC curve for out-of-sample predictions. Compared to the within sample predictions, this is doing worse. With a lower observed AUC of 0.82 and a CI of 0.87 0.90, the model seems to be experiencing overfitting.

Table 7: Classification table. Rows show the number of individuals as by the predictions, while the columns classify individuals by their observed feature. Classification was done on the basis of closest integer, in other words, rounding the predicted values so that those reach either 0 or 1.

	Obs. 0s	Obs. 1s
Predicted 0s	0.93	0.05
Predicted 1s	0.01	0.01

From what we observe on Figure 2 and Figure 3, the model is clearly doing better within sample prediction (89% AUC) vs out-of-sample prediction (84% AUC), which means that it could be the case that we are overfitting the data. In order to reduce overfitting, we could hae tried incorporating cross-validation to the feature selection process. Using cross-validation we could have picked features (and in essence, models) by looking at those that maximized the average AUC across the folds (partitions of the data as a produc of cross-validation).

Question 9

In this paper we have build a classification model to predict the death of patients who suffered of a head injury which caused them attending a hospital's emergency room. Using age, gender, ethnicity, and some other measurements regarding the paitiens' health status such as symbolic blood preassure, respiratory rate, and the Glasgow Coma Scale, we trained various logistic

regression models–which corresponds to the branch of supervised learnig in machine learning literature. The final model itself was build incrementally using an automatic feature selection process which tried to increment prediction accuracy by adding one feature at a time, while at the same time only preserving models in which all the features were statistically significant. The whole process required estimating dozens of thousands of models with various different sets of features.

Question 10

For this last question, as a difference from the previous part of this work, I only analyzed a small group of models to be compared to the current best model which has an AUC of 0.89 (95% CI [0.871; 0.913]) for the training data and a 0.82 (95% CI [0.748; 0.901]) for the out-of-sample data, versus an extension of that adding the ASAPS variable.

As shown in Table 8, adding the ASAPS variables causes changes in the coefficients of the current best-model. In particular, we see that in order to go back to having a model in which all the features are significantly associated with the outcome, we need to remove various terms of the original model. The columns “Model 2” through “Model 4” show the steps I followed from adding the new variables to remove some of the predictors in order to get to a model where all predictors are significant.

Regarding the prediction accuracy itself, Table 9 shows the corresponding AUC values for in-sample and out-sample data predictions. While the new feature significantly increases accuracy compared to our baseline model (rises to almost 93%), the accuracy of out-of-sample data worseners, and the same happens in the last two especifications. Ultimately, what we observe here is an increasing overfitting of the data while at the same time we are doing worse outside of the training data set; hence, I would keep the model in which the ASAPS variables are not included.

Table 9: Comparing the model selection in part 1 with variants after including the ASAPS variable. While in general the new variable is not improving the out of sample AUC, it does improve significantly the within sample AUC up to about 0.03 units.

Model	All feat. signif.	AUC	AUC out of sample
1	Yes	0.892 [0.871; 0.913]	0.824 [0.748; 0.901]
2	No	0.929 [0.912; 0.946]	0.800 [0.725; 0.875]
3	No	0.930 [0.914; 0.947]	0.795 [0.720; 0.870]
4	Yes	0.929 [0.912; 0.946]	0.790 [0.712; 0.869]

	Model 1	Model 2	Model 3	Model 4
Age $\times$ RR (breaths per minute)	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	(0.00)	(0.00)	(0.00)	(0.00)
Glasgow Coma Scale $\times$ SBP (mm Hg)^2	0.00^{**}	0.00	0.00^{***}	0.00^{***}
	(0.00)	(0.00)	(0.00)	(0.00)
Male $\times$ SBP (mm Hg)$^{-1}$	54.87^{*}	38.23	42.68	
	(23.10)	(23.65)	(22.74)	
RR (breaths per minute)$^{-1}$ $\times$ Glasgow Coma Scale^{-1}	46.60^{**}	38.38^{*}	55.94^{***}	52.03^{***}
	(15.80)	(17.16)	(14.13)	(13.75)
SBP (mm Hg) $\times$ Glasgow Coma Scale	-0.00^{***}	-0.00		
	(0.00)	(0.00)		
SBP (mm Hg)^2 $\times$ Glasgow Coma Scale^{-1}	0.00^{*}	0.00		
	(0.00)	(0.00)		
SBP (mm Hg)^2 $\times$ Glasgow Coma Scale$^{1/2}$	-0.00^{*}	-0.00		
	(0.00)	(0.00)		
SBP (mm Hg)^2 $\times$ Glasgow Coma Scale^2	-0.00^{***}	-0.00^{*}	-0.00^{***}	-0.00^{***}
	(0.00)	(0.00)	(0.00)	(0.00)
(Intercept)	-2.34^{***}	-4.74^{***}	-5.63^{***}	-5.12^{***}
	(0.65)	(0.91)	(0.48)	(0.40)
Mild sys. dis.		0.03		
		(0.68)		
Severe sys. dis.		0.49		
		(0.66)		
Severe sys. dis. that is a const. threat to life		2.15^{***}	1.99^{***}	2.00^{***}
		(0.61)	(0.26)	(0.26)
Not expected to survive without operation		3.43^{***}	3.26^{***}	3.31^{***}
		(0.62)	(0.28)	(0.27)
AIC	936.52	794.36	791.06	794.24
BIC	989.93	871.49	838.52	835.77
Log Likelihood	-459.26	-384.18	-387.53	-390.12
Deviance	918.52	768.36	775.06	780.24
Num. obs.	2789	2789	2789	2789

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

Table 8: Estimated coefficients of the model selected in part 1 versus models generated after including the ASAPS variable in the feature set.