

PM511b Data Analysis Project

George G Vega Yon

April 5, 2019

Question 1

Table 1 shows a set of descriptive statistics for the variables in the dataset. The following observations:

1. In the training dataset, only a handful of cases corresponds to patients that have died. In an ideal world we would like to have a rather balance sample size of the two cases.
2. Of all the variables, *Age*, *Glasgow Coma Scale*, and (*Ethnycity* = *Asian*) are the only ones that show to be significantly different across groups. 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 ***
<i>Ethnycity</i>			
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	

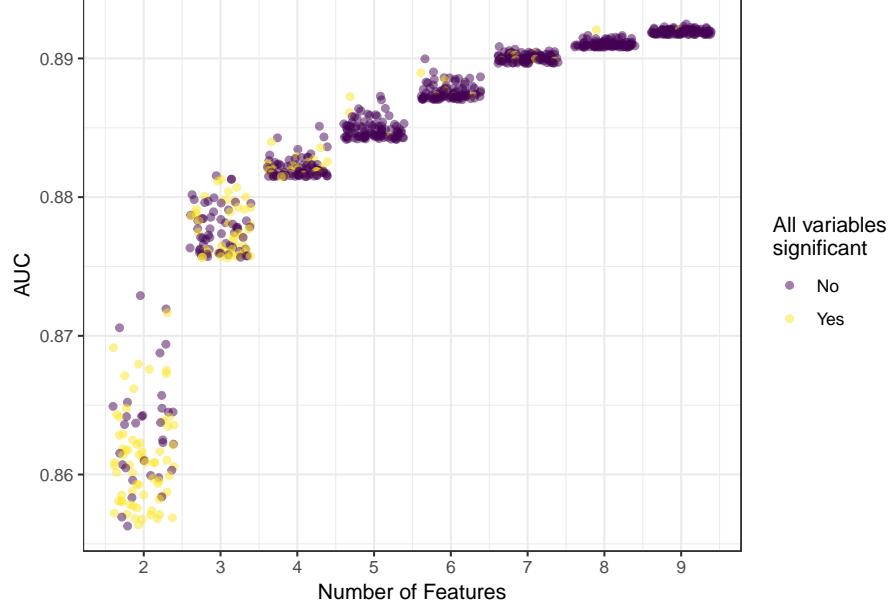


Figure 1: Distribution of the 800 different models. Overall we can see that the accuracy of the prediction model rises exponentially as a function of the number of features include in the model. At the same time, models with a large number of variables tend to be less significant in terms of having all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

Questions 2-4

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

Ultimately, I kept a set of best models instead of a single best. The algorithm that I used for such is described as follows:

1. Increase 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 (with a cardinality of 180).
2. 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 manageable, keep the best 100 in terms of AUC, and replace B with that set.
 - d. Continue 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.

Question 5

Table 2 and Table 3 show the parameter estimates of 20 different models that showed to have the highest accuracy values.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
african american \times RR (breaths per minute) ²			0.00 (0.00)							
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 \times SBP (mm Hg) ⁻¹										0.18 (0.24)
Age ⁻¹ \times Glasgow Coma Scale ⁻¹	57.27 (45.15)								111.89* (47.62)	
Age ² \times SBP (mm Hg) ⁻¹								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)	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 \times Age ⁻¹									-8.90 (6.61)	
Male \times Age ^{1/2}					-0.02 (0.06)					
Male \times Glasgow Coma Scale									0.03 (0.02)	
Male \times RR (breaths per minute) ⁻¹				-3.37 (4.07)						
Male \times SBP (mm Hg) ⁻¹	56.85* (23.17)	57.05* (24.79)	53.03* (23.06)	72.19* (32.08)	68.02 (42.65)	53.05* (23.82)	58.69* (23.68)	57.46* (23.59)		56.46* (23.56)
RR (breaths per minute) ^{1/2} \times Glasgow Coma Scale ⁻¹							0.48 (0.55)			
RR (breaths per minute) ⁻¹ \times Glasgow Coma Scale ⁻¹	49.44** (16.21)	42.21** (16.29)	46.75** (15.84)	59.15** (22.69)	47.77** (16.24)	44.33** (16.52)	53.71** (18.42)	46.81** (15.96)	58.81*** (17.29)	47.23** (15.99)
SBP (mm Hg) \times 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) \times Glasgow Coma Scale ^{1/2}									-0.03*** (0.01)	
SBP (mm Hg) ⁻¹ \times Glasgow Coma Scale ⁻¹		627.40* (280.91)								
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)	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)	-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)	-3.13*** (0.92)	-5.23*** (1.44)	-2.32*** (0.65)	-2.63*** (0.76)	-2.45*** (0.72)	-12.02** (3.75)	-3.21** (1.22)	-2.46*** (0.67)	-0.90 (0.89)	-2.50*** (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 2: Best 10 models regardless of whether all the features show as statistically significant at the 95 % level.

	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* (0.66)				-1.69** (0.66)		
Male \times Age ^{1/2}								0.06* (0.03)		
Male \times RR (breaths per minute) ⁻¹			7.58* (3.05)	6.93* (3.07)			7.07* (3.12)			
Male \times SBP (mm Hg) ⁻¹	54.87* (23.10)	54.87* (23.10)			52.43* (22.86)	52.59* (22.97)			62.38** (24.06)	58.95* (23.64)
RR (breaths per minute) ⁻¹ \times Glasgow Coma Scale ⁻¹	46.60** (15.80)	46.60** (15.80)			49.76** (16.12)					
SBP (mm Hg) \times Glasgow Coma Scale	-0.00*** (0.00)	-0.00*** (0.00)	-0.00*** (0.00)	0.01* (0.00)	-0.00*** (0.00)	-0.00*** (0.00)	0.01* (0.00)	0.01* (0.00)	-0.00* (0.00)	0.01* (0.01)
SBP (mm Hg) \times Glasgow Coma Scale ^{1/2}				-0.04*** (0.01)			-0.02* (0.01)	-0.04*** (0.01)		
SBP (mm Hg) ^{1/2} \times Glasgow Coma Scale ^{1/2}							-0.50* (0.22)			-0.66** (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*** (0.65)	-2.34*** (0.65)	-0.93* (0.45)	4.88** (1.82)	-2.30*** (0.66)	-1.14* (0.49)	5.73* (2.32)	5.21** (1.79)	-2.73** (0.87)	5.76* (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 3: Best 10 models in which all the features show as statistically significant at the 95 % level.

Question 6

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

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

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:

$$\Pr(\text{die} = 1 | X = x) = \frac{\exp \hat{\theta}^t x}{1 + \exp \hat{\theta}^t x} \quad (2)$$

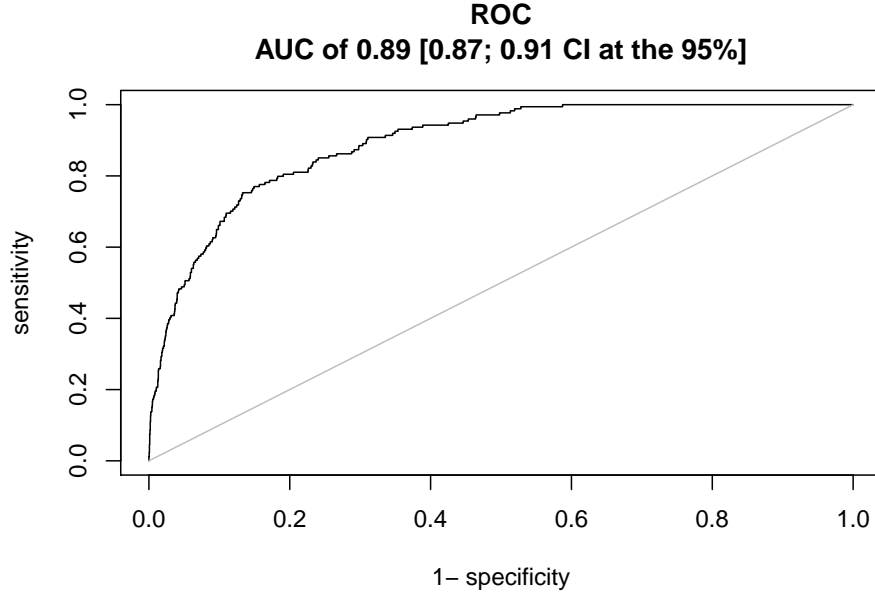


Figure 2: ROC AUC curve for the selected model. The model shows a good performance within sample with an AUC value that lives within 0.89 and 0.91. A very close range, but rather high accuracy level.

Questions 7 and 8

Figure 2 and Table 4 show the AUC curve and classification table for the selected model on the training data, 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.

Table 4: Classification table. Rows show the number of individuals as by the predictions, while the columns classify individuals by their observed feature.

	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 fit, which which means that it could be the case that we are overfitting the data. That said, part of it is not surprising since (1) the number of events in which an individual dies in the data set is rather small and (2)

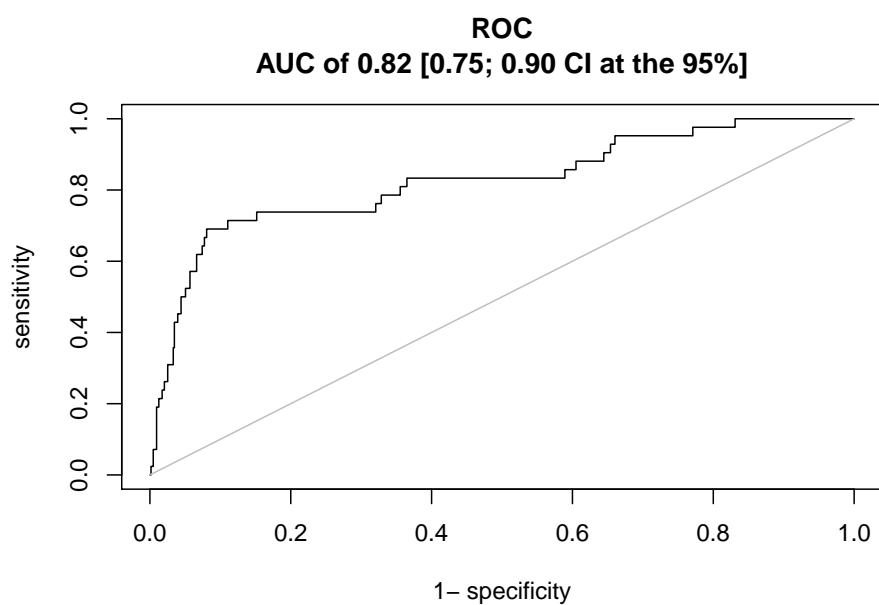


Figure 3: AUC ROC curve for out-of-sample predictions. Compared to the within sample predictions, this is doing worse. With a lower observed AUC of 0.82 and a CI of 0.87-0.90, the model seems to be experiencing overfitting.

Question 9

In this paper we have build a classification model for prediction of the likelihood of dying for patients in an hospital using age, gender, and some other measurements regarding the patients' health status. As a difference from a statistical inference scenario, here we are applying a heuristic that tries to identify the best prediction model using brute force to calculate thousands of plausible model candidates. Table 5 shows a summary of the number of models estimated per number of features selected.

Table 5: 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
3	10008
4	10425
5	10395
6	10063
7	9257
8	9598
9	9852
Total	75593

The large sample space of potential models to use includes, besides of the baseline features, interaction terms and transformations of the independent variables

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Age			0.02** (0.01)		0.02** (0.01)	0.02*** (0.01)
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)
Glasgow Coma Scale						-0.20*** (0.02)
Glasgow Coma Scale \times SBP (mm Hg) ²	0.00** (0.00)	0.00 (0.00)	0.00*** (0.00)			
Male \times SBP (mm Hg) ⁻¹	54.87* (23.10)	38.23 (23.65)	49.92* (23.91)			
RR (breaths per minute) ⁻¹ \times Glasgow Coma Scale ⁻¹	46.60** (15.80)	38.38* (17.16)	43.67** (14.72)	60.97*** (14.09)	46.50** (14.75)	
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)	-0.00*** (0.00)	
(Intercept)	-2.34*** (0.65)	-4.74*** (0.91)	-5.95*** (0.51)	-4.92*** (0.40)	-5.17*** (0.41)	-3.35*** (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*** (0.61)	1.92*** (0.26)	2.13*** (0.26)	2.05*** (0.26)	2.01*** (0.26)
Not expected to survive without operation		3.43*** (0.62)	3.33*** (0.28)	3.42*** (0.27)	3.48*** (0.27)	3.51*** (0.27)
AIC	936.52	794.36	785.61	807.44	800.92	792.15
BIC	989.93	871.49	839.01	843.04	842.46	827.75
Log Likelihood	-459.26	-384.18	-383.80	-397.72	-393.46	-390.07
Deviance	918.52	768.36	767.61	795.44	786.92	780.15
Num. obs.	2789	2789	2789	2789	2789	2789

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 6: Estimated coefficients of the model selected in part 1 versus models generated after including the ASAPS variable in the feature set.

Question 10

Table 7: 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	Yes	0.930 [0.913; 0.947]	0.801 [0.727; 0.875]
4	Yes	0.924 [0.906; 0.942]	0.784 [0.706; 0.863]
5	Yes	0.924 [0.906; 0.942]	0.791 [0.714; 0.868]
6	Yes	0.926 [0.907; 0.945]	0.806 [0.730; 0.883]