Midterm Assignment – Due: 11/12/2014
Tom Harel (tomharel@gmail.com)
General Assembly SF – Data Science 10

Titanic: Machine Learning from Disaster

Preface: – Dataset Description:

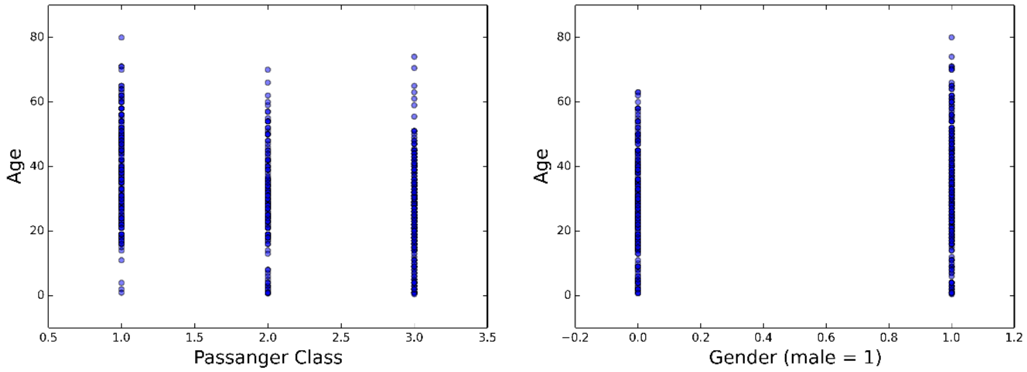
The dataset in this report is the passenger list of the historic Titanic. Every row in the dataset represents a single passenger. There are 12 columns for each passenger, which are (variable names in parentheses):

- Passenger ID (PassengerId)
- Indicator for survival (Survived) – Binary: {0 = Died, 1 = Survived}
- Passenger Class (Pclass) – Categorical: {1, 2, 3}
- Name (Name) - String
- Gender (Sex) – Categorical: {Male = 'male', Female = 'female'}
- Age (Age) – Continuous Float[1]
- Number of Siblings/Spouses Aboard – Continuous Integer
- Number of Parents/Children Aboard – Continuous Integer
- Ticket Number (Ticket) - String
- Price paid for travel (Fare) – Continuous Float
- Cabin (Cabin) - String
- Port of Embarkation (Embarked) – Categorical: {C = 'Cherbourg', Q = 'Queenstown', S = 'Southampton')

Question 1:

In the training set we have 177 entries that are missing the Age variable. I filled in these entries by taking the conditional mean of Age with respect to Pclass and Sex (both of which have no missing values): $\widetilde{Age} = E[\,Age \mid Pclass,\ Sex\,]$
I chose to use Pclass and Sex after looking at some scatter plots of Age against the other variables. Reproduced in the figure below.

Figure 1: Plots of Pclass and Sex on Age



In the plots above we see that the means and distribution seem to be different for genders and passenger classes. Table 1 depicts the predicted values used. We observe that they are different and this increases my confidence for using the conditional mean over a simple mean or mode.[2]

Table 1: Predicted Age for missing values

	Pclass 1	Pclass 2	Pclass 3
Male	41.28	30.74	26.51
Female	34.61	28.72	21.75

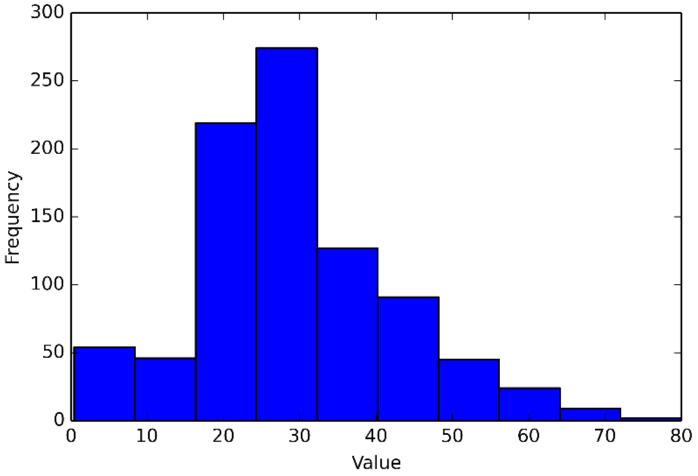
I also note that I imputed the two missing entries for Embarked by using the mode (which is 'S').

[1] Non-integer ages exist in the data. There is someone with Age = 0.83 for example.

[2] In my iPython Notebook I also calculated the Standard Deviations for these estimates. They were all in the range of roughly 12-15.

Question 2:
First, I note that I decided to treat all variables used in the Logistic Regression model below as categorical, with the exception of Age. The rationale behind this approach is that imposing a linear relationship (where one might not truly exist) is not desirable. Also, since most of these variables have only a few discrete values it will not hurt the model's integrity to treat them as dummies. However, Age has more values and its distribution, as depicted in the histogram in Figure 2 is almost normal. Hence I will normalize[3] the values and treat it as a continuous variable.

Figure 2: Age Histogram



Second, I note that all dummy variables I created are of the form OriginalVarName_Value. For example, if Pclass takes the values 1, 2 and 3. The dummies will be Pclass_1, Pclass_2 and Pclass_3.

Third, I note that to avoid multicollinearity I will exclude one of the dummies in each regression model.

Model:
The model fits Survival on Age, Pclass, Gender, Embarked, Parch and SibSp (as defined in the preface, all dummied out and excluded one). After the initial deployment my analysis showed that using 0.61 as a threshold gets the best results.

The model estimated coefficients are:[4]

AgeN	**Pclass_1**	**Pclass_2**	**ismale**	**Embarked_S**	**Embarked_C**	**Parch_1**	**Parch_2**	**Parch_5**
-0.455186	2.040292	1.036882	-2.359933	-0.260524	0.168329	0.27918	-0.170706	-0.316847

Parch_4	**Parch_6**	**SibSp_1**	**SibSp_3**	**SibSp_4**	**SibSp_2**	**SibSp_5**	**SibSp_8**	**intercept**	**Parch_4**
-0.4634	-0.1983	0.1864	-0.9951	-0.7005	-0.03608	-0.4399	-0.6572	0.3248	-0.4634

Looking at the size of the coefficient (either in the table directly above, or in Figure 4 in the appendix), it seems that the passenger classes are highly predictive of survival rates. It is better to have been in the 1st and 2nd classes over the third (dummy excluded). This makes intuitive sense because upper classes were given priority access to the lifeboats (as seen in the movie). Also, gender is very important, as chances of survival for woman were much higher, which was also explained in the movie as the men allowed the women to go first. The rest of the variables are less predictive and but are included for completeness. I note that Age does not seem to be a big predictor (since the coefficients are normalized and are mean zero this coefficient would have been lower if I had not normalized). This might be due to the fact that it will make sense for the elderly and children will be given priority. So in a linear sense Age should not perform well.

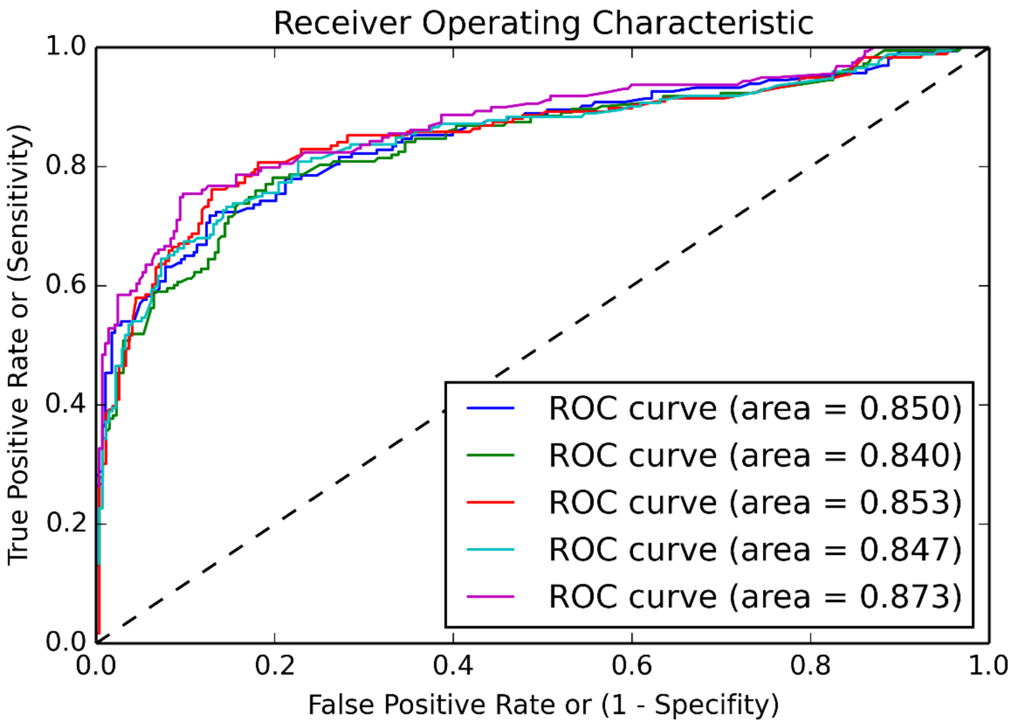
Question 3: Please see iPython Notebook.
With respect to the number of folds, it did not seem to matter so I chose 8 folds because the data has 891 rows and I wanted the smaller set to be of at least 100 rows. My cross-validation scores were roughly around 0.8.

[3] Normalization of the form: xnorm = ((x - mean(x)) / sd(x))
[4] See Appendix for Figure # which shows the size of the coefficients on a bar chart.

Question 4:

I chose to split the (train) data into a train set and a test set. Because this split will affect our ROC curve's shape I decided to plot 5 of them to show that the overall shape and area under the curve is preserved, regardless of split (I note that for this analysis I used a 50/50 split). The average Area Under the Curve my model depicts is roughly 0.85.

Figure 3: ROC Curves with AUC



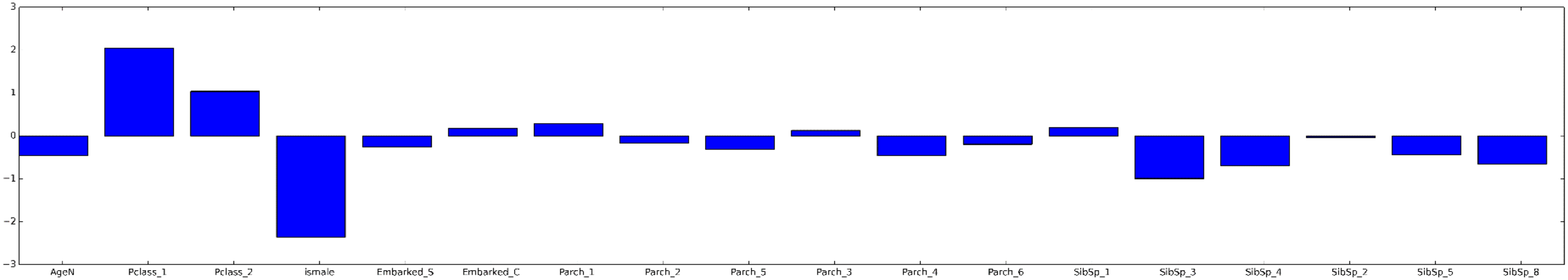
I think this model reached this level of accuracy due to the use of non-parametric regressors and by including most of the available indicators. I also implemented regularization at 0.5, which my analysis showed gives the best cross validation results (see iPython Notebook)

When uploading this model to Kaggle it scored **0.78947**

For next steps, I think I should consider evaluating introducing a new variable, Family Size, which will engulf both Parch and Sibsp, still treating it as categorical. I might also look into using Age and Age^2, because there seems to be a non-linear (if any) relationship between Survival and Age. Last, I might also consider experimenting with extracting data from the passenger name and salutation, as well as consider analyzing the Cabin variable.

Appendix

Figure 4: Model Coefficients Bar Chart


AgeN
Pclass_1
Pclass_2
ismale
Embarked_S
Embarked_C
Parch_1
Parch_2
Parch_5
Parch_3
Parch_4
Parch_6
SibSp_1
SibSp_3
SibSp_4
SibSp_2
SibSp_5
SibSp_8