Hong Kong Horse Race

Sahba Salarian

Feb 2019

Dataset information

The Hong Kong Horse racing dataset from http://fisher.stats.uwo.ca/faculty/aim/2019/9850/data/:

The data structure shows 925 rows with 2 variables.

```
str(df)
```

```
## 'data.frame':    925 obs. of  2 variables:
##  $ PROB: num  0.044 0.105 0.178 0.137 0.021 0.027 0.191 0.033 0.124 0.089 ...
##  $ WIN : int  0 0 0 0 0 0 0 0 0 1 ...
```

Data Engineering

We check if any unknown value (NA) exists in the data set. Since there is no unknown values the structure of dataset remains unchanged. In case of exsiting NA values we can remove the observation or replace it by the coloumn mean in some cases.

Train & Test

The data is split between two randomly selected datasets of train and test. The splitting procedure is done randomly with specific seed value to make our model reproducable. Train set has 70% of all the data.

Logistic Model

Considering the binary class of the WIN variable which is our target response, logistic regression is used to model the fit. Fit summary and coefficients can be found in Table 1.

Table 1: summary of logistic regression fit with all inputs

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1822	0.2453	-12.97	0.0000
PROB	7.9517	1.4514	5.48	0.0000

Based on this fit we can consider our hpothesis to be a line as: winning probability = Pr(WIN=1|PROB) logit(winning probability)= -3.1822 + 7.9517(PROB)

ROC curve

ROC curve shows how well the fit developed over the train dataset can predict well, over the test data set. The area under the ROC curve, AUC value, which is a measure for the fit prediction capacity is 0.712.

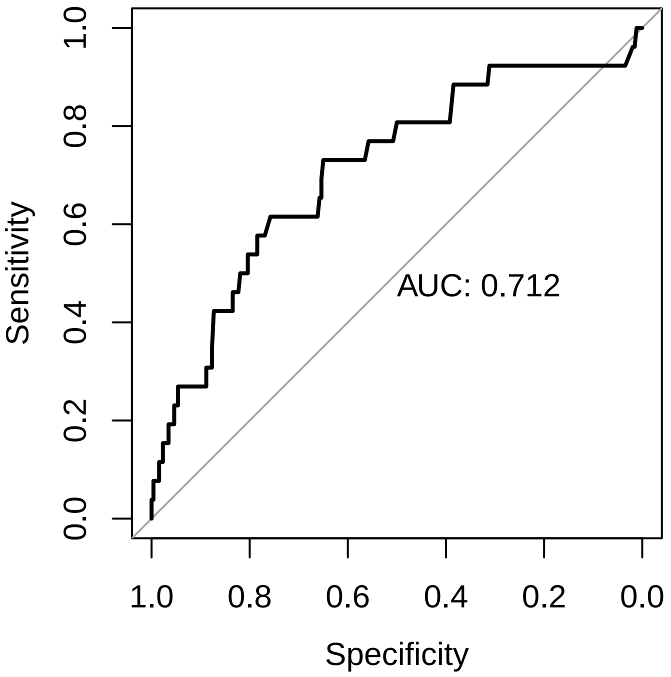

Figure 1: ROC for simple Logistic Fit