Supplement

Supplement1: Missing Data

The variable of interest in our analysis are position in fatality accident and age. For position in fatality accidentamong which there are 888 missing data. There are 29 categories, and 12 of them have less than 100 observations(245 in total). We only use the categories have more than 100 observations.

Supplement2&3: Model Selection and Cross-validation

We fitted three logistic regression models and used using AUC to rank models. First we randomly sampled 10 percent(7800) of the total observations for an unbiased estimate of the error rate of final model. Then we used the 90%(70266) left for model fitting and selection. We did a 10-fold cross-validation with the 70266 observations on each of our models:

Model1:

$$\text{logit}Pr(Survivor_i = 0) = \beta_0 + \boldsymbol{\beta_1} f(Position_i) + .$$

Model2:

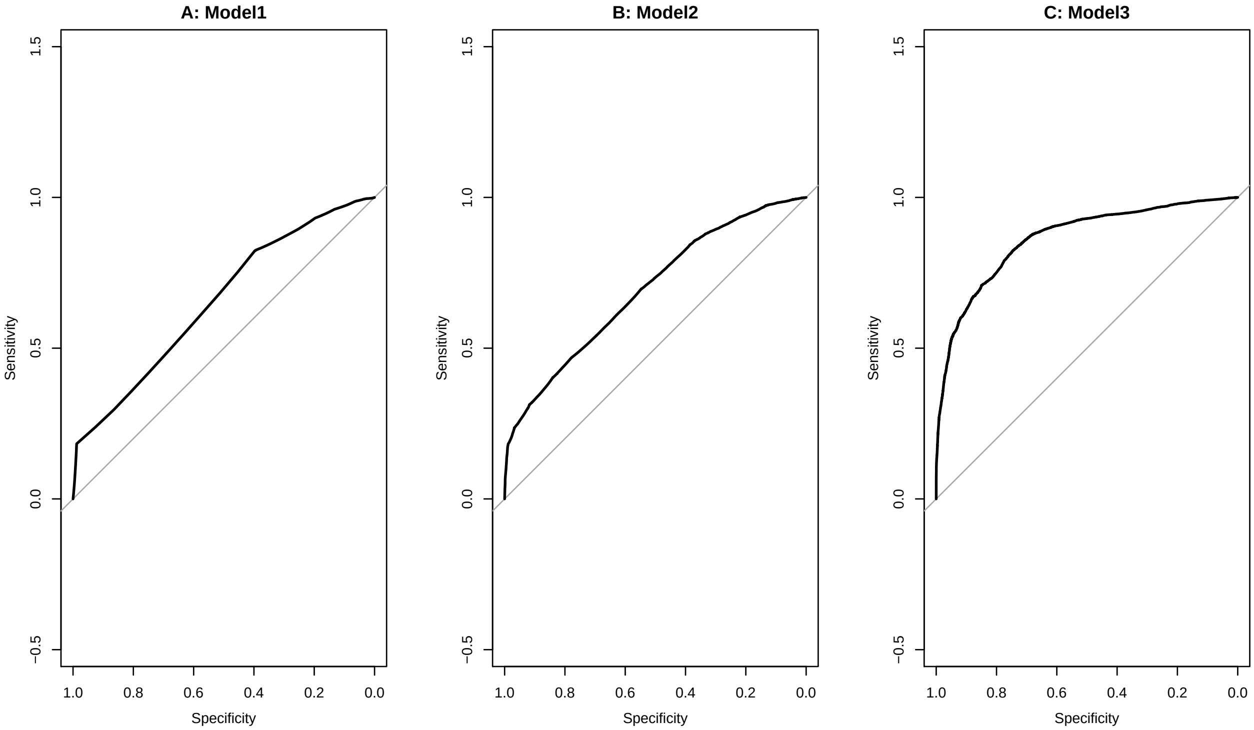
$$\text{logit}Pr(Survivor_i = 0) = \beta_0 + \boldsymbol{\beta_1} f(Position_i) + \boldsymbol{\beta_2} Age_i.$$

Model3:

$$\text{logit}Pr(Survivor_i = 0) = \beta_0 + \boldsymbol{\beta_1} Position_i + \boldsymbol{\beta_2} Age_i + \boldsymbol{\beta_3} Extricate_i + \boldsymbol{\beta_4} Restraint_i + \boldsymbol{\beta_5} Alcohol_i + \boldsymbol{\beta_6} Intersection_i$$

We used roc function(default) from Package(“pROC”)[1] to calculate ROC curve and AUC for each model.

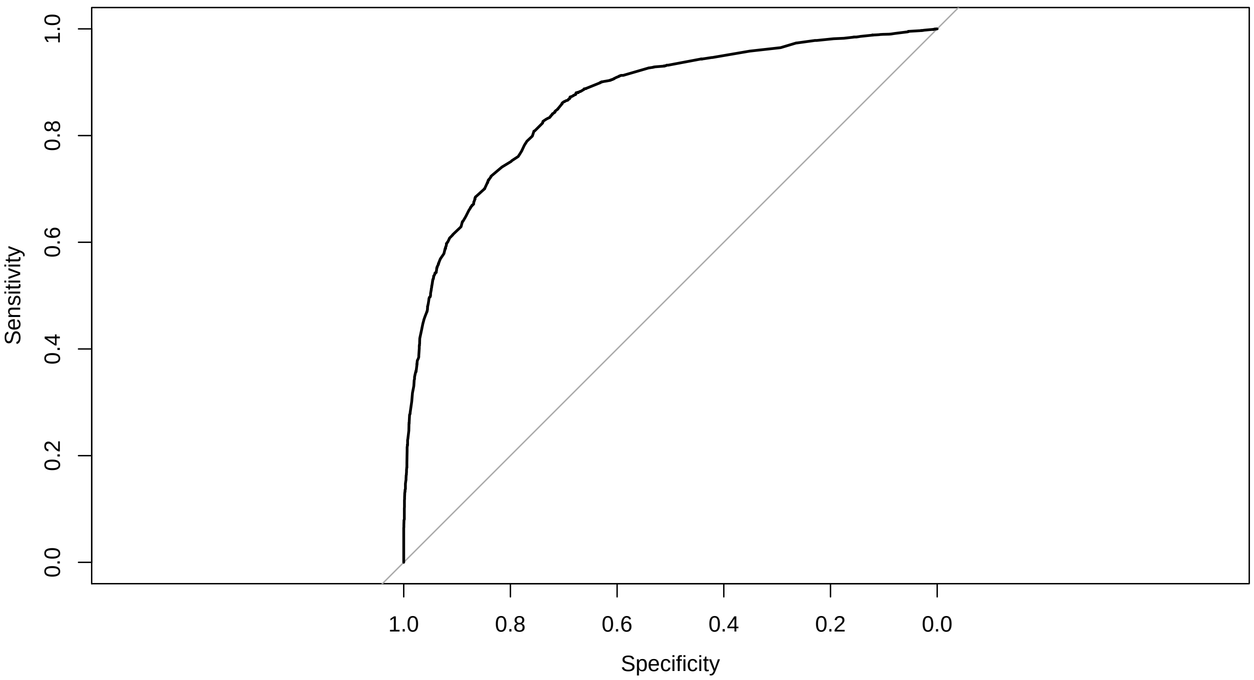
Figure1:ROC curve For Model selection



We ranked models with AUC(Model1 0.65,Model2 0.69,Model3 0.86).

For estimate of the error rate of final model, we used the 10 percent(7800) of the total observations we preserved at the begining, the ROC curve showed below, with AUC 0.86

Figure2 :ROC curve For Final model



Reference

[1]Xavier Robin, Natacha Turck, Alexandre Hainard, et al. (2011) "pROC: an open-source package for R and S+ to analyze and compare ROC curves". BMC Bioinformatics, 7, 77. DOI: 10.1186/1471-2105-12-77.