documentation

jazzmoe

November 9, 2018

Things to check (Cross when done)

- Run regression on age and impute ages of missing values
- ~~Make a dummy for cabin no cabin~~
- Classify different cabin categories
- Classify "children" using age and parent/siblings variable
- Search for special titles in the name category and test impact on sirvival rate
- a "Countess" might be more likely to survive than Mr. Huber
- Test simple logit model
- Test random forest
- Test other machine learning models
- Investigate different variable selection | classify importance of variables for classification
- Compare different models

Data Documentation

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes pclass: A proxy for socio-economic status (SES) 1st = Upper 2nd = Middle 3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way... Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way... Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children travelled only with a nanny, therefore parch=0 for them.

Exploratory Data Analysis

Analyze here relationships of variables. z.B Plot of fare and pclass

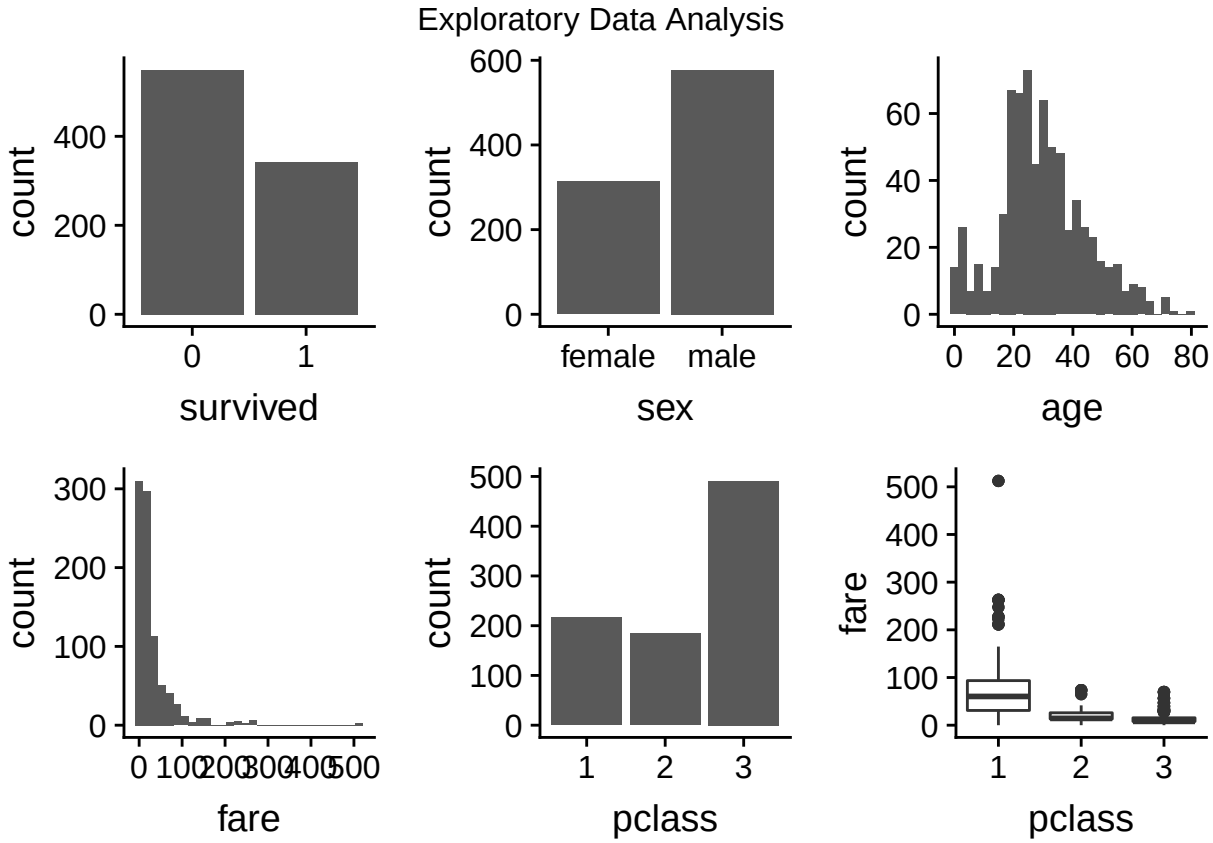
```
glimpse(Train)
```

```
## Observations: 891
## Variables: 17
## $ passengerid <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,...
```

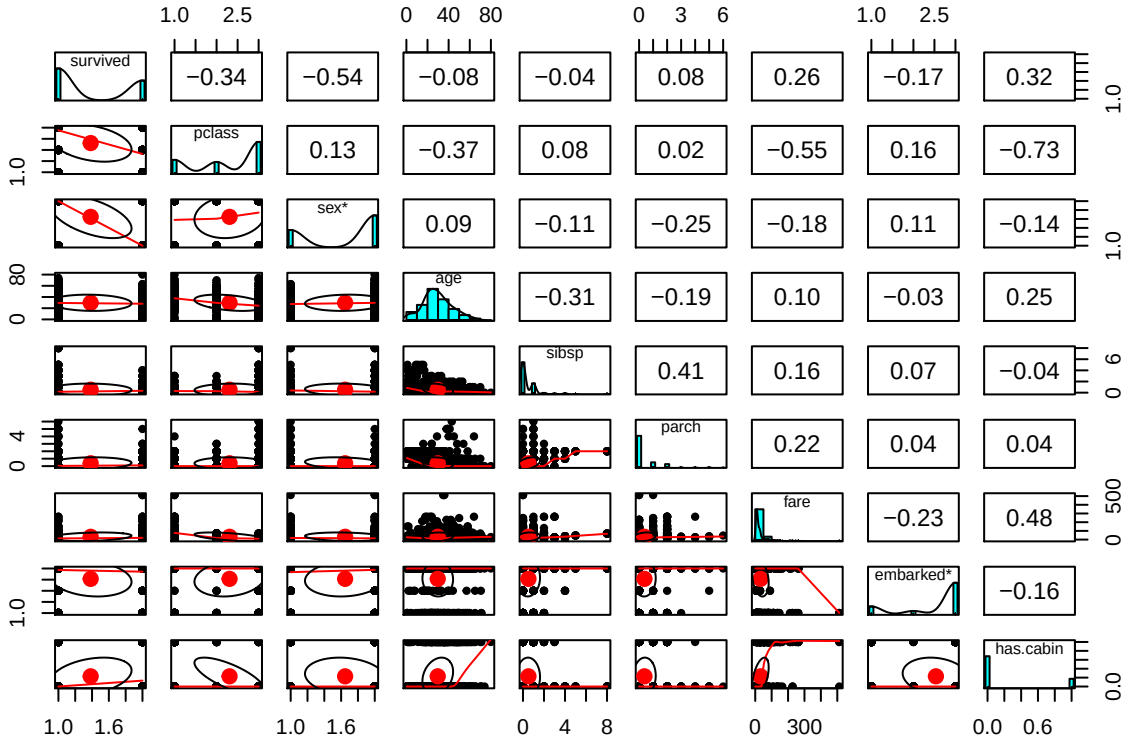
```
## $ survived    <fct> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0,...
## $ pclass      <int> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3,...
## $ name        <chr> "Braund, Mr. Owen Harris", "Cumings, Mrs. John Bra...
## $ sex         <chr> "male", "female", "female", "female", "male", "mal...
## $ age         <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14, 4, 58, 20, ...
## $ sibsp       <int> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4,...
## $ parch       <int> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1,...
## $ ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 3101282", "1138...
## $ fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, ...
## $ cabin       <chr> NA, "C85", NA, "C123", NA, NA, "E46", NA, NA, NA, ...
## $ embarked    <chr> "S", "C", "S", "S", "S", "Q", "S", "S", "S", "C", ...
## $ pclass1     <dbl> 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,...
## $ pclass2     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0,...
## $ pclass3     <dbl> 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1,...
## $ has.cabin   <dbl> 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0,...
## $ big.family  <fct> 1, 1, 0, 1, 0, 0, 0, 4, 2, 1, 2, 0, 0, 6, 0, 0, 5,...
```

```
grid.arrange(p.survived, p.sex, p.age1, p.fare, p.pclass, p.fare.pclass,
             top = "Exploratory Data Analysis",
             nrow = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 177 rows containing non-finite values (stat_bin).
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
pairs.panels(TrainCorr) # corr matrix
```



Missing Data

Numbers of NAs per variable.

```
naTrainFrame[,c(2:12)] %>% kable
```

survived	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
0	0	0	0	177	0	0	0	0	687	2

```
naTestFrame[,c(2:11)] %>% kable
```

pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
0	0	0	86	0	0	0	1	327	0

Embarked

```
emb.surv
```

```
##
##        0   1
##    C  75  93
##    Q  47  30
##    S 427 217
```

```
emb.surv.chisq
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  Train$embarked and Train$survived
## X-squared = 26.489, df = 2, p-value = 1.77e-06
```

Peoplo who embarked in Southhampton have significantly higher likelihood of dying.

Age

Impute age with a regression or some ML algo -> add imputed age to classification -> more information

Training Models

Random forrest with 10-fold cv

```
print(Mod3)
```

```
## Random Forest
##
## 712 samples
##   7 predictor
##   2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 509, 510, 509, 509, 510, 509, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.8011352  0.5747635
##    5    0.8120752  0.6007411
##    8    0.8071066  0.5939173
##   11    0.7947561  0.5698546
##   15    0.7880140  0.5566681
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 5.
```

```
confusionMatrix(Mod3)
```

```
## Cross-Validated (10 fold, repeated 5 times) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction    0    1
##          0 52.9 12.5
##          1  6.3 28.3
##
##  Accuracy (average) : 0.812
```

Best model so far Mod3 with accuracy of 0.83.