documentation

jazzmoe

November 9, 2018

Things to check

- Many age variables are missing which might cause bias | run regression on age and estimate ages of missing values
- make a dummy for cabin no cabin | classify different cabin categories
-

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

Exploratory Data Analysis

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

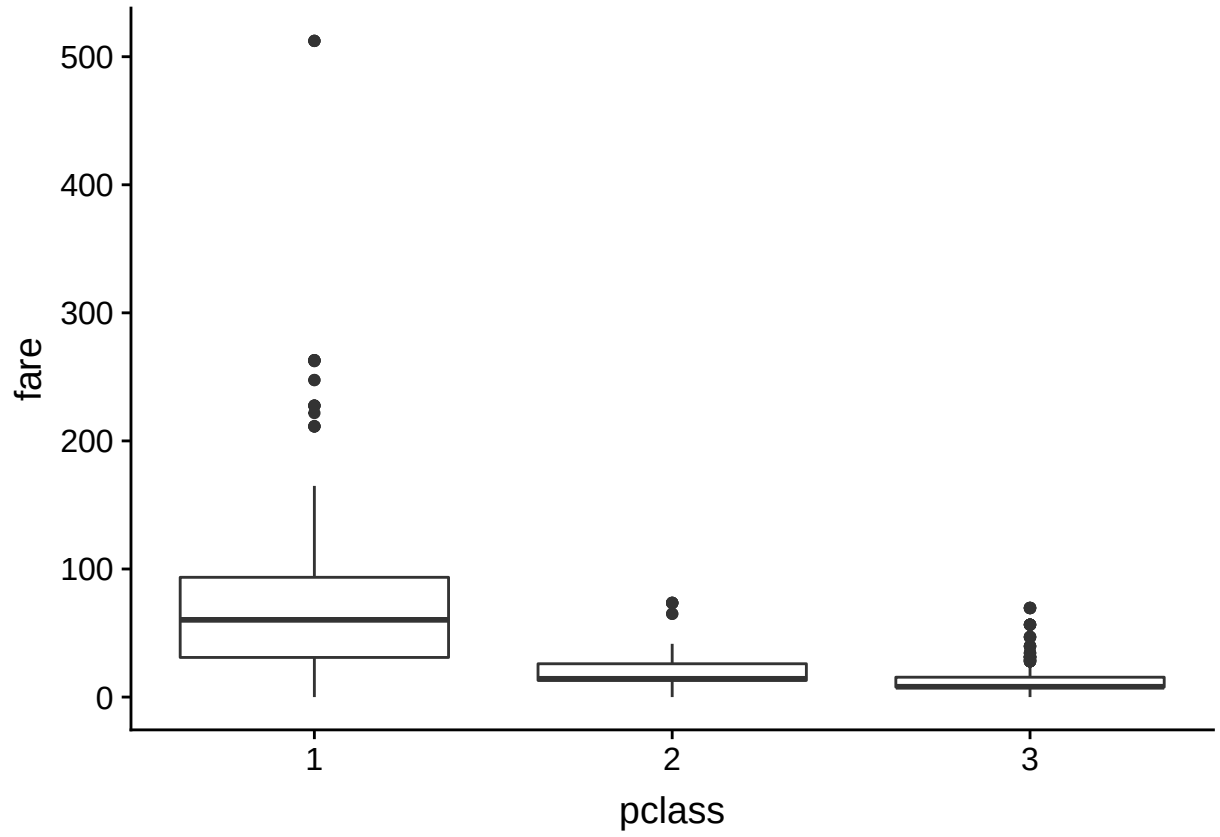
```
glimpse(Train)
```

```
## Observations: 891
## Variables: 15
## $ 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,...
```

```
p5
```

```
## NULL
```

```
p6
```



Missing Data

```
naFrame %>% kable
```

passengerid	survived	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	pclass1	pclass2
0	0	0	0	0	177	0	0	0	0	687	2	0	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

Training Models

Random forrest with 10-fold cv

```
summary(Mod3)
```

```
##                 Length Class      Mode
## call               5   -none-     call
## type               1   -none-     character
## predicted        572   factor     numeric
## err.rate        1500   -none-     numeric
## confusion          6   -none-     numeric
## votes           1144   matrix     numeric
## oob.times        572   -none-     numeric
## classes            2   -none-     character
## importance         6   -none-     numeric
## importanceSD       0   -none-     NULL
## localImportance    0   -none-     NULL
## proximity          0   -none-     NULL
## ntree              1   -none-     numeric
## mtry               1   -none-     numeric
## forest            14   -none-     list
## y                572   factor     numeric
## test               0   -none-     NULL
## inbag              0   -none-     NULL
## xNames             6   -none-     character
## problemType        1   -none-     character
## tuneValue          1   data.frame list
## obsLevels          2   -none-     character
## param              1   -none-     list
```

```
confusionMatrix(Mod3)
```

```
## Cross-Validated (10 fold, repeated 5 times) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction    0    1
##          0 52.4 12.4
##          1  6.7 28.5
##
##  Accuracy (average) : 0.8087
```

Best model so far Mod3 with accuracy of 0.83.