Titanic: Machine Learning from Disaster

In this challenge, was asked to complete the analysis of what sorts of people were likely to survive. In particular, to apply the tools of machine learning to predict which passengers survived the tragedy.

Source

https://www.kaggle.com/c/titanic

Loading data

Train set:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500		S
6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583		Q

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

Descriptions:

Variable	Descriptions
survival	Survival (0 = No; 1 = Yes)
pclass	Passenger Class(1 = 1st; 2 = 2nd; 3 = 3rd)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Let's start with the name field. And build a new predictive variables.

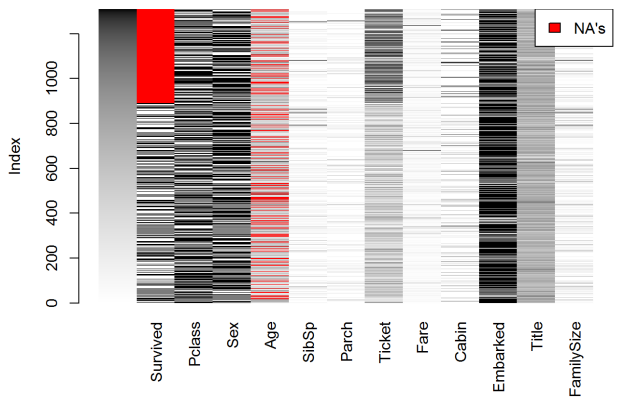
This step will be very useful when we well be using Random forest algoritm. WE create variable "Familysize"="Parch"+"Sibsp"+1. But this variable has lots of cathegories, in some cases it's problematic. So, we add new variable FamilyId2 that contains less cathegories than variable FamilyId. We substitute families that has less than three members as "small". Random Forest algoritm is imposible for variables that have more than 53 cathegories.

We create "combine set", that is a union of test and train sets. He has 1309 observations, it means that abroad were 1309 passengers.

Missing values

Let's see on the situation with NA's in "combine set".

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0         418           0           0           0         263
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           1           0           0
##       Title  FamilySize     Surname    FamilyID   FamilyID2
##           0           0           0           0           0
```



There are lots of missing values, if we later want to make a statistical analysis of this data, it will be impossible to realize, that's why, we must to substitute this missing values by new values, one way is to predict new values by using random forest or decision tree.

Complete data set without Na's.

Here we recieve a new "test" and "training" sets with new variables, without "NA's".

Train set:

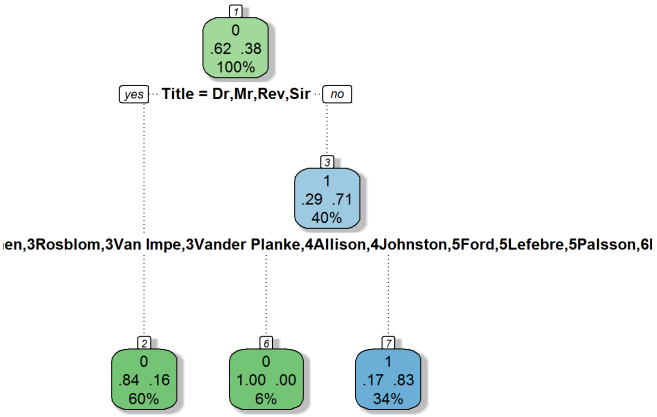
PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Title	FamilySize	FamilyID	FamilyID2
1	0	3	male	22	1	0	7.2500	S	Mr	2	Small	Small
2	1	1	female	38	1	0	71.2833	C	Mrs	2	Small	Small
3	1	3	female	26	0	0	7.9250	S	Miss	1	Small	Small
4	1	1	female	35	1	0	53.1000	S	Mrs	2	Small	Small
5	0	3	male	35	0	0	8.0500	S	Mr	1	Small	Small
6	0	3	male	31	0	0	8.4583	Q	Mr	1	Small	Small

Missing values prediction.

So, move on to the next step.

One way to measure the prediction ability of a model is to test it on a set of data not used in estimation. Data miners call this a “testing set” and the data used for estimation is the “training set”.

Split data into training and testing sets. And apply **decision tree** algorithm for prediction.




```
##           Reference
## Prediction   0   1
##          0 153  27
##          1  11  76
```

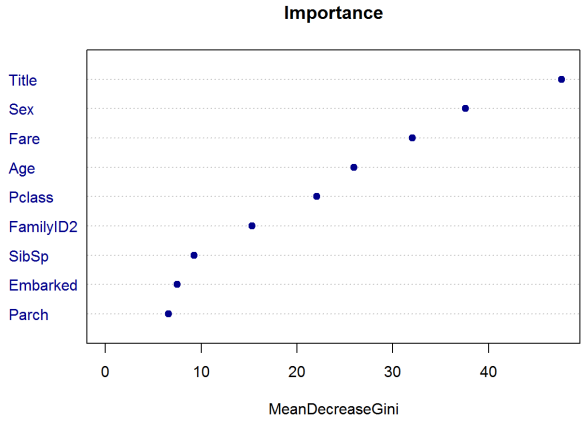
```
##  Accuracy
## 0.8576779
```

Let's apply Random Forest algorithm.

Add to prediction those varibles that we created later (“Title”,“FamilyID2”)

```
##           Reference
## Prediction   0   1
##          0 149  29
##          1  15  74
```

```
## Accuracy
## 0.835206
```



Forest of conditional inference trees

Let's try a forest of conditional inference trees. They make their decisions in slightly different ways, using a statistical test rather than a purity measure, but the basic construction of each tree is fairly similar. It will give us the most precise result. Also we used Logistic Regression for prediction, but have choosen the most accurate result.

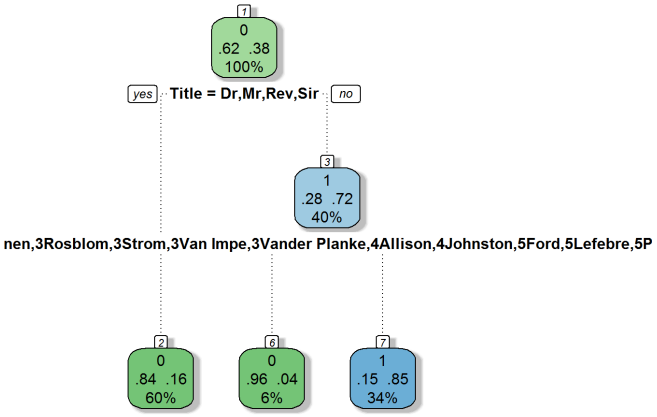
```
##           Reference
## Prediction   0   1
##          0 157  32
##          1   7  71
```

```
##  Accuracy
## 0.8539326
```

Test set prediction

Testing model. So, here we move to next step, from training model to testing. Now let's apply random forest and decision tree to predict “Survived” variable in test set.

Decision Tree



Forest of conditional inference trees

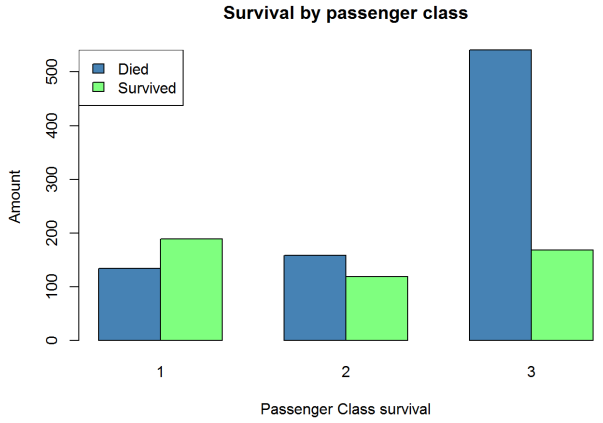
First 15 rows of survive prediction:

```
##     PassengerId Survived
## 1           892        0
## 2           893        0
## 3           894        0
## 4           895        0
## 5           896        0
## 6           897        0
## 7           898        1
## 8           899        0
## 9           900        1
## 10          901        0
## 11          902        0
## 12          903        0
## 13          904        1
## 14          905        0
## 15          906        1
```

Split training set into two sets “Male” and “Female”

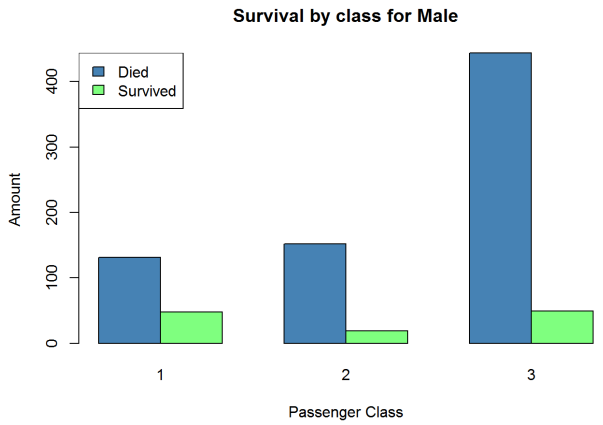
Survival that depends on passenger class.

Class	Amount	Died	Survived	Probability
1	323	134	189	0.5851393
2	277	158	119	0.4296029
3	709	541	168	0.2369535



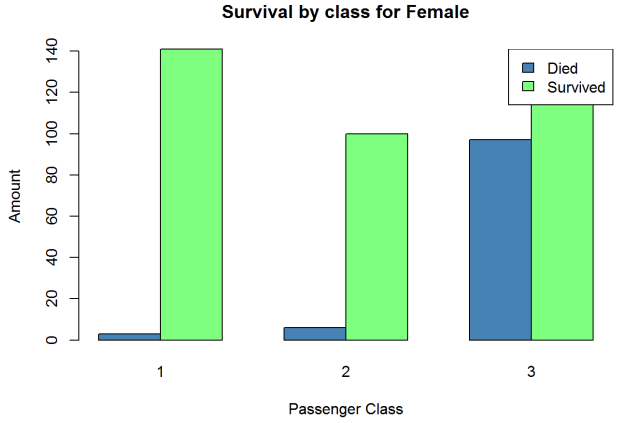
For Male

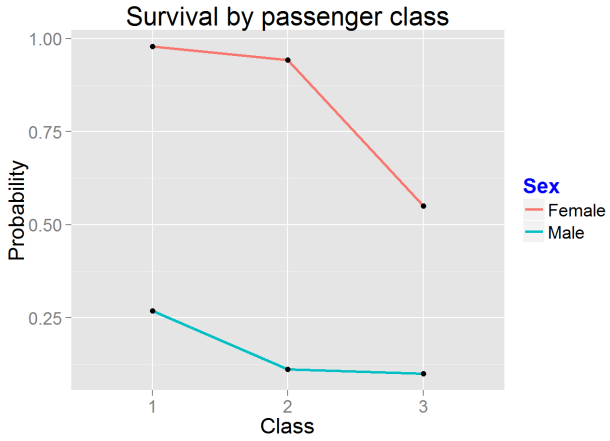
Class	Died	Survived	Probability
1	131	48	0.2681564
2	152	19	0.1111111
3	444	49	0.0993915



For Female

Class	Died	Survived	Probability
1	3	141	0.9791667
2	6	100	0.9433962
3	97	119	0.5509259





Survival rate by Age

General case

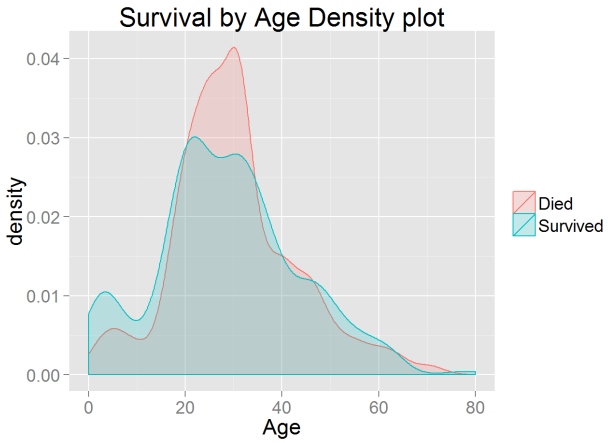
Age	Amount	Died	Survived	Probability
Less28	639	399	240	0.3755869
Greater28	670	434	236	0.3522388
Less12	111	53	58	0.5225225

For Female

Age	Amount	Died	Survived	Probability
Less28	255	72	183	0.7176471
Greater28	211	34	177	0.8388626
Less12	54	24	30	0.5555556

For Male

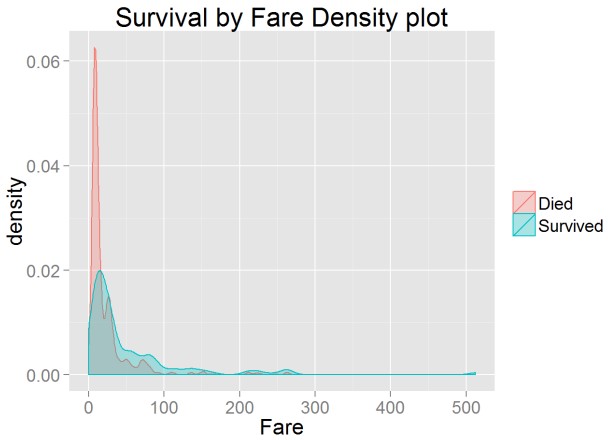
Age	Amount	Died	Survived	Probability
Less28	384	327	57	0.1484375
Greater28	459	400	59	0.1285403
Less12	57	29	28	0.4912281



Survival that depends on Fare

Let's see on deviation of ticket price from the mean price.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.896  14.450  33.300  31.280 512.300
```



General case

Fare	Amount	Died	Survived	Probability
Greater30	344	152	192	0.5581395
Less30	965	681	284	0.2943005

For Male

Fare	Amount	Died	Survived	Probability
Greater30	176	133	43	0.2443182
Less30	667	594	73	0.1094453

For Female

Fare	Amount	Died	Survived	Probability
Greater30	168	19	149	0.8869048
Less30	298	87	211	0.7080537

Survival that depends on gender

Sex	Died	Survived	Probability
female	106	360	0.7725322
male	727	116	0.1376038

Survival depends on Embarked variable

General case

Embarked	Amount	Died	Survived	Probability
Cherbourg	270	132	138	0.5111111
Queenstown	123	70	53	0.4308943
Southampton	914	631	283	0.3096280

For Male

Embarked	Amount	Died	Survived	Probability
Cherbourg	157	123	34	0.2165605
Queenstown	63	60	3	0.0476190
Southampton	623	544	79	0.1268058

For Female

Embarked	Amount	Died	Survived	Probability
Cherbourg	113	9	104	0.9203540

Queenstown	60	10	50	0.8333333
Southampton	291	87	204	0.7010309

Survival that depends on Family size

General case

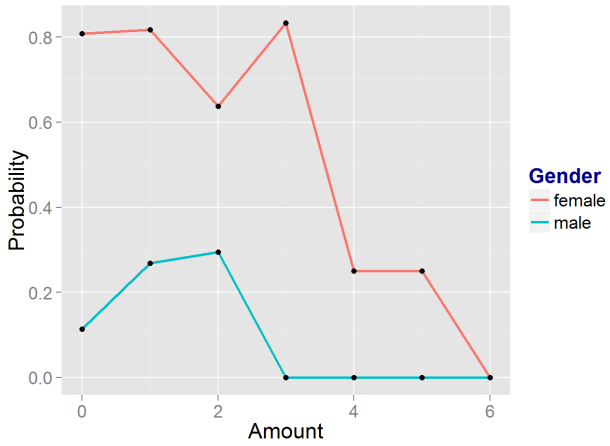
	Amount	Died	Survived	Probability
0	0	684	318	0.317
1	1	76	94	0.553
2	2	56	57	0.504
3	3	3	5	0.625
4	4	5	1	0.167
5	5	5	1	0.167
6	6	2	0	0.000

For Female

	Amount	Died	Survived	Probability
0	0	56	237	0.809
1	1	16	72	0.818
2	2	25	44	0.638
3	3	1	5	0.833
4	4	3	1	0.250
5	5	3	1	0.250
6	6	1	0	0.000

For Male

	Amount	Died	Survived	Probability
0	0	628	81	0.114
1	1	60	22	0.268
2	2	31	13	0.295
3	3	2	0	0.000
4	4	2	0	0.000
5	5	2	0	0.000
6	6	1	0	0.000



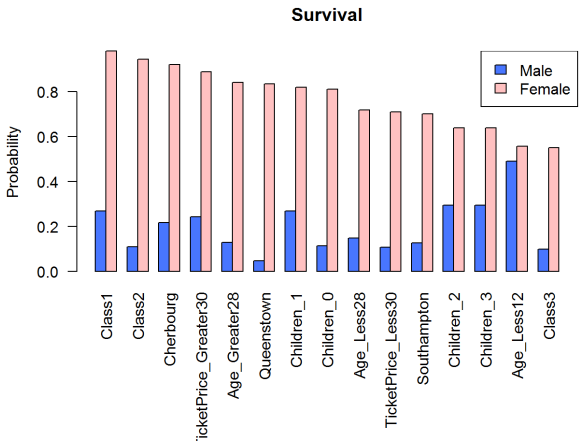
Survival “Title” variable

	Amount	Died	Survived	Probability
Col	4	2	2	0.500
Dr	8	5	3	0.375
Lady	4	1	3	0.750
Master	61	32	29	0.475
Miss	260	71	189	0.727
Mlle	3	0	3	1.000
Mr	757	676	81	0.107
Mrs	197	34	163	0.827
Ms	2	1	1	0.500
Rev	8	8	0	0.000
Sir	5	3	2	0.400

General table

Uner each column name we have a probability to survive for male and female respectively.

Sex	Class1	Class2	Class3	Age_Less28	Age_Greater28	Age_Less12	TicketPrice_Greater30	TicketPrice_Less30	Cherbourg	Queenstown	Southampton	Children_0	Children_1	Children_2	Children_3
Male	0.268	0.111	0.099	0.148	0.129	0.491	0.244	0.109	0.217	0.048	0.127	0.114	0.268	0.295	0.295
Female	0.979	0.943	0.551	0.718	0.839	0.556	0.887	0.708	0.920	0.833	0.701	0.809	0.818	0.638	0.638



Probability table

For Male

	Class1	Class2	Class3	Age_Less28	Age_Greater28	TicketPrice_Greater30	TicketPrice_Less30
Cherbourg	0.2425	0.1640	0.1580	0.1825	0.1730	0.2305	0.1630
Queenstown	0.1580	0.0795	0.0735	0.0980	0.0885	0.1460	0.0785
Southampton	0.1975	0.1190	0.1130	0.1375	0.1280	0.1855	0.1180
Children_0	0.1910	0.1125	0.1065	0.1310	0.1215	0.1790	0.1115
Children_1	0.2680	0.1895	0.1835	0.2080	0.1985	0.2560	0.1885
Children_2	0.2815	0.2030	0.1970	0.2215	0.2120	0.2695	0.2020
Children_3	0.2815	0.2030	0.1970	0.2215	0.2120	0.2695	0.2020

For Female

	Class1	Class2	Class3	Age_Less28	Age_Greater28	TicketPrice_Greater30	TicketPrice_Less30
Cherbourg	0.9495	0.9315	0.7355	0.8190	0.8795	0.9035	0.8140
Queenstown	0.9060	0.8880	0.6920	0.7755	0.8360	0.8600	0.7705
Southampton	0.8400	0.8220	0.6260	0.7095	0.7700	0.7940	0.7045
Children_0	0.8940	0.8760	0.6800	0.7635	0.8240	0.8480	0.7585
Children_1	0.8985	0.8805	0.6845	0.7680	0.8285	0.8525	0.7630
Children_2	0.8085	0.7905	0.5945	0.6780	0.7385	0.7625	0.6730
Children_3	0.8085	0.7905	0.5945	0.6780	0.7385	0.7625	0.6730