Assignment 4

Analysis Report

Dropping unwanted columns : "Name","PassengerId","Cabin","Ticket"

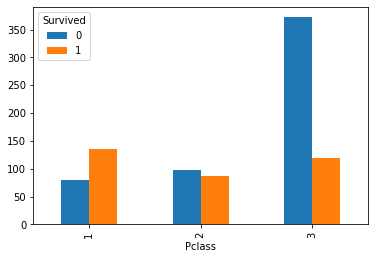
Reasons are:

“Name” : too many unique values not helpful feature

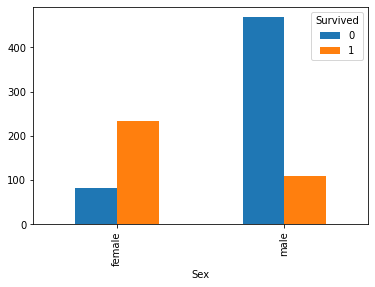
“PassengerId” : too many unique values not helpful feature

“Ticket” : too many unique values not helpful feature

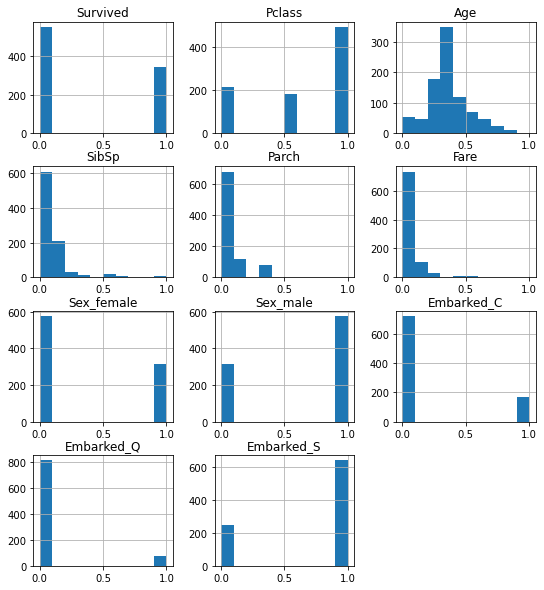
“Cabin” : too many Null values also too many unique values not helpful feature



In the above graph we can see the that highest number of survivors are from Pclass 1 on the other side most people in pclass 3 were unable to survive. People in Pclass 2 has roughly the same number of survivors and non-survivors. Also we can conclude that Pclass 3 has most numbers of people 450+, Pclass 2 has a little less than 200 people and pclass 1 has slightly more than 200 people.



From the above graph we can infer that there are little less than 350 females and around 250 of them survived on the other hand there are around 600 males and only around 100 males were able to survive.



From the above histogram keeping focus on “Port of Embarkation”, most people with C and Q were not able to survive and most people from S survived.

Train Test split ideal scenarios:

test for 15.000000000000002% of data

[0.77631579 0.82894737 0.86842105 0.78947368 0.82894737 0.78947368

0.82894737 0.77333333 0.74666667 0.81333333]

Mean score: 0.8043859649122809

Max score: 0.868421052631579

Min score: 0.7466666666666667

test for 20.000000000000004% of data

[0.79166667 0.88888889 0.77464789 0.8028169 0.78873239 0.81690141

0.8028169 0.76056338 0.77464789 0.81690141]

Mean score: 0.8018583724569641

Max score: 0.8888888888888888

Min score: 0.7605633802816901

test for 25.000000000000007% of data

[0.79104478 0.88059701 0.74626866 0.8358209 0.74626866 0.80597015

0.76119403 0.7761194 0.78787879 0.8030303 ]

Mean score: 0.7934192672998643

Max score: 0.8805970149253731

Min score: 0.746268656716418

test for 30.000000000000004% of data

[0.88888889 0.79365079 0.80952381 0.79032258 0.79032258 0.79032258

0.82258065 0.72580645 0.79032258 0.79032258]

Mean score: 0.7992063492063491

Max score: 0.8888888888888888

Min score: 0.7258064516129032

test for 35.00000000000001% of data

[0.81034483 0.79310345 0.81034483 0.77586207 0.79310345 0.84482759

0.75862069 0.74137931 0.79310345 0.80701754]

Mean score: 0.7927707199032062

Max score: 0.8448275862068966

Min score: 0.7413793103448276

test for 40.000000000000014% of data

[0.74074074 0.83333333 0.77777778 0.7962963 0.81132075 0.81132075

0.79245283 0.75471698 0.75471698 0.81132075]

Mean score: 0.7883997204751921

Max score: 0.8333333333333334

Min score: 0.7407407407407407

test for 45.00000000000001% of data

[0.81632653 0.83673469 0.73469388 0.81632653 0.81632653 0.79591837

0.7755102 0.75510204 0.75510204 0.81632653]

Mean score: 0.7918367346938775

Max score: 0.8367346938775511

Min score: 0.7346938775510204

From the above output for different slips, we can see: a 30% test split gives the most accurate model. As the mean, max, and min accuracy of 30% test split is the highest amongst all the splits.

Classification report for 0.5 threshold

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| False | 0.79 | 0.86 | 0.83 | 160 |
| True | 0.77 | 0.67 | 0.71 | 108 |
| accuracy |  |  | 0.78 | 268 |
| macro avg | 0.78 | 0.76 | 0.77 | 268 |
| weighted avg | 0.78 | 0.78 | 0.78 | 268 |

Classification report for 0.75 threshold

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| False | 0.70 | 0.99 | 0.82 | 160 |
| True | 0.95 | 0.36 | 0.52 | 108 |
| accuracy |  |  | 0.74 | 268 |
| macro avg | 0.82 | 0.67 | 0.67 | 268 |
| weighted avg | 0.80 | 0.74 | 0.70 | 268 |

Here we can see 0.5 threshold has higher accuracy then 0.75 threshold.

Reason being with 0.75 threshold is good in predicting True values as precision for True is 0.95. Also recall for False value is 0.99. this makes this model bias in predicting True values. But 0.5 threshold is unbiased and based on the precision and accuracy we can say it can make more correct decisions in determining both True and False values.