CS 5304 HW1 - Tier 1 Report

By Chris M. Wang (mw866@cornell.edu)

Requirement 1

work with the Titanic Data Set from Kaggle at https://www.kaggle.com/c/titanic

Please refer to the code submission

Requirement 2

build source code in Scala or Python that runs in Spark 2.0.2 to analyze the Titanic data set.

Please refer to the code submission

Requirement 3

answer the question: “for subgroups of people boarding the Titanic, how would you maximize their individual probability of survival?”. You must define meaningful subgroups. You should submit your predictions in a file that clearly labels identity of person and the prediction.

Please refer to answers to **Requirement 5**

Requirement 4

build at least two of {Naïve Bayes, Logistic Regression, random forests, support vector machines or neural networks using the libraries of Spark.MLLib only.

Please refer to the code below.

Explain your choice;

The Random Forest classifier is chosen for the following reasons:

1. It has second best accuracy among all classifiers.
2. It provides the probability, which can be used for answer " what sorts of people were likely to survive "

investigate which features are most informative;

As the baseline, the following features are chosen based on common sense ['Pclass','Sex','Age','SibSp','Parch','Fare','Embarked'].

Then the individual features are removed to measure the (negative) impact to the prediction performance.

Accuracy	Baseline	-'Pclass'	- 'Sex'	- 'Age'	-'SibSp'	-'Parch'	-'Fare'	-'Embarked'
NaiveBayes	0.72826087	0.710382514	0.646153846	0.674157303	0.683544304	0.67816092	0.803108808	0.709677419
LogisticRegression	0.804347826	0.808743169	0.687179487	0.792134831	0.797468354	0.827586207	0.808290155	0.779569892
RandomForestClassifier	0.831521739	0.825136612	0.733333333	0.814606742	0.816455696	0.844827586	0.808290155	0.790322581
MultilayerPerceptronClassifier	0.809782609	0.759562842	0.687179487	0.837078652	0.689873418	0.643678161	0.792746114	0.784946237

Based on the results above, below is a ranking of the most informative features's:

- Sex
- Age
- Pclass

- SibSp
- Embarked
- Parch
- Fare

do at least one round of error analysis to maximize your chosen metric (F1, accuracy, weighted F1);

The prediction accuracy using features:

```
Features: 'Cabin','Sex','Age','SibSp','Parch','Fare','Embarked'
NaiveBayes                      0.710382513661
LogisticRegression              0.808743169399
RandomForestClassifier          0.825136612022
MultilayerPerceptronClassifier  0.75956284153
```

The replacement of Cablin with Pclass improves the

```
Features: 'Pclass','Sex','Age','SibSp','Parch','Fare','Embarked'
NaiveBayes                      0.728260869565
LogisticRegression              0.804347826087
RandomForestClassifier          0.83152173913
MultilayerPerceptronClassifier  0.809782608696
```

explain your choice of metric.

The accuracy is the number of correct predictions made divided by the total number of predictions made. It is chosen because it's a good performance indicator of the prediction and correlates well with the F1/weighted F1 score.

Requirement 5

complete an analysis of what sorts of people were likely to survive.

The profiles of most surviving people in the test sets are analyzed using the GaussianMixture clustering.

Cluster Means	'Pclass'	'Sex'	'Age'	'SibSp'	'Parch'	'Fare'	'Embarked'
Cluster 1	2.0882353113706267	0.8823531137782782	21.613472813187634	0.5294116726332617	0.8235293772121892	25.108456293919506	2.76470583633349
Cluster 1	1.0000001386082993	0.9999998614345726	41.204926539856274	0.5000000692945482	0.39583341708660036	109.44912135220906	1.8125001645715

Base on the clustering result above, the survivor subgroups can be generalized as follows:

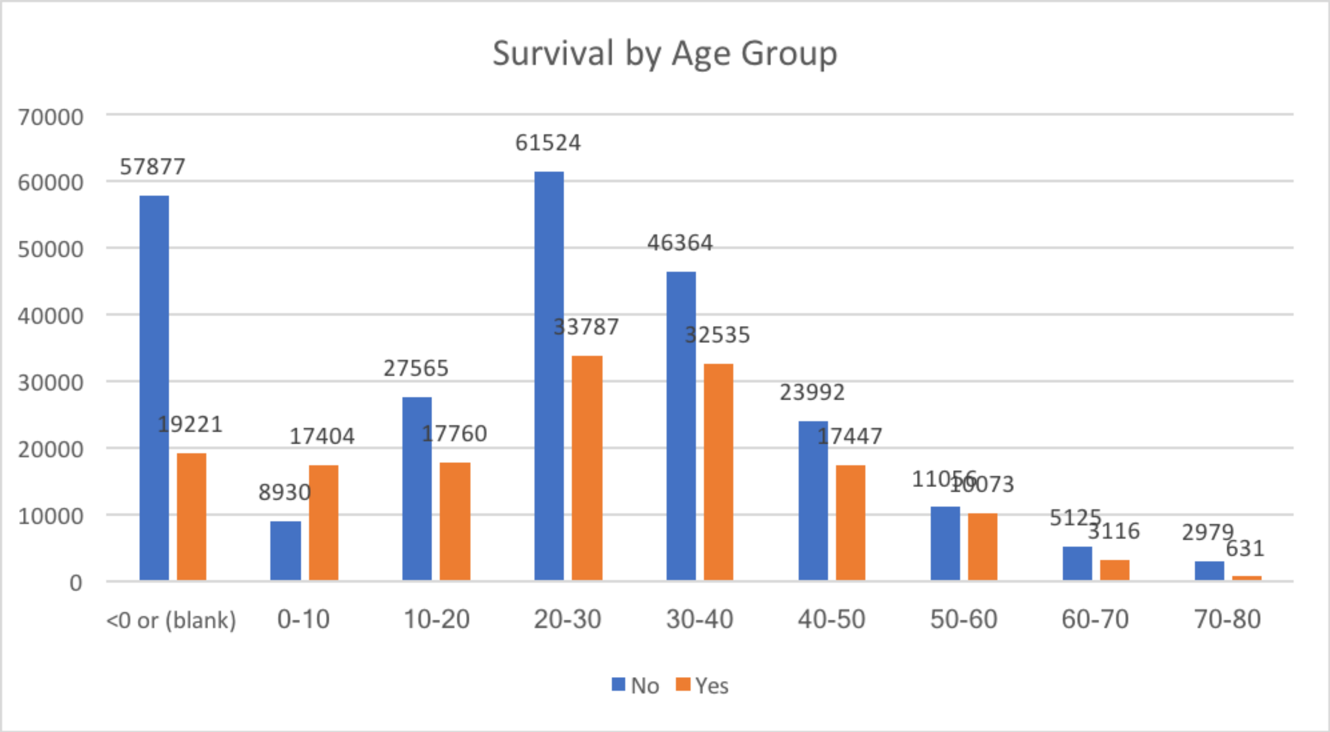
Profiles	'Pclass'	'Sex'	'Age'	'SibSp'	'Parch'	'Fare'	'Embarked'
Profile 1 - "Young Ladies"	2	Female	22-year old	with 0 or 1 sibling or spouse aboard	with 0 or 1 parents aboard	paid $25	Southampton
Profile 2 - "Rich Ladies"	1	Female	41-year old	with 0 or 1 sibling or spouse aboard	with 0 or 1 parent aboard	paid $76	Queenstown

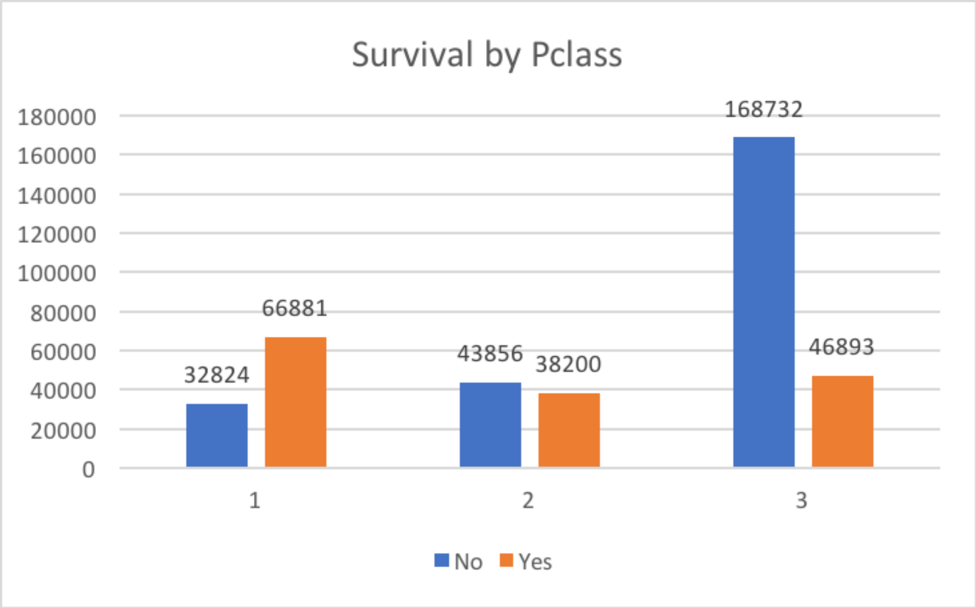
In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

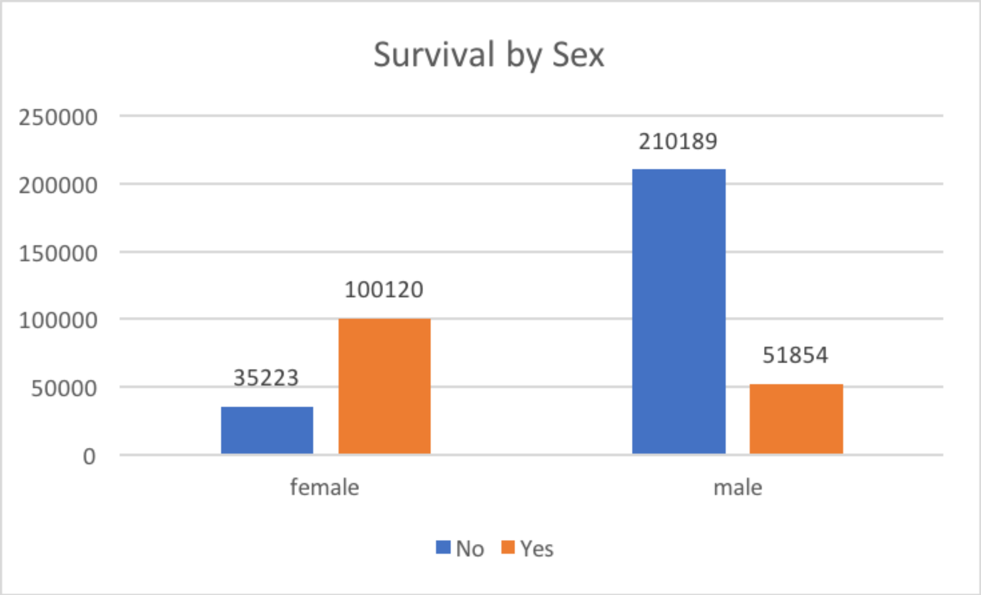
Please refer to the prediction file in the ./output folder.

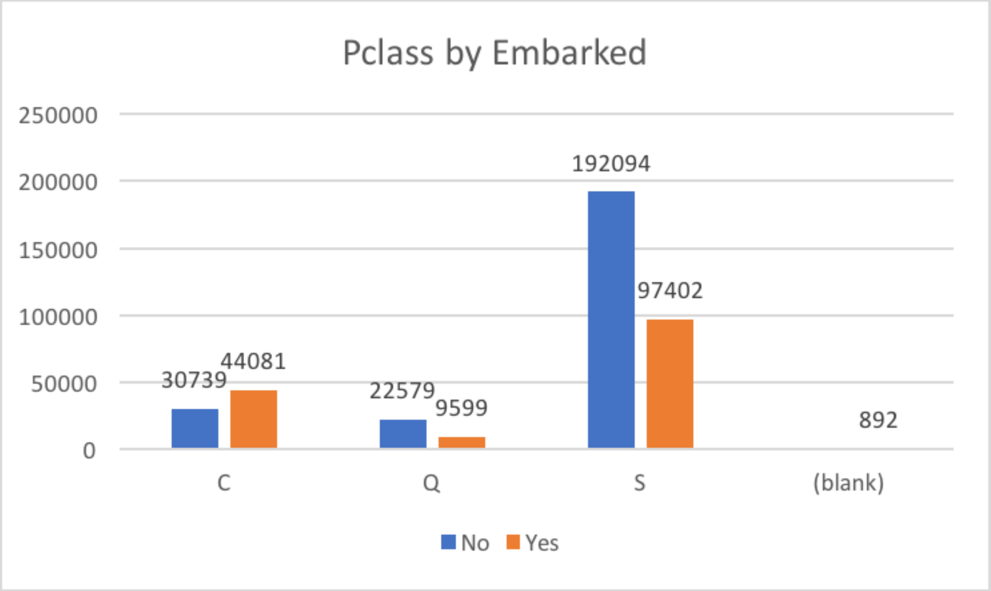
Addendum

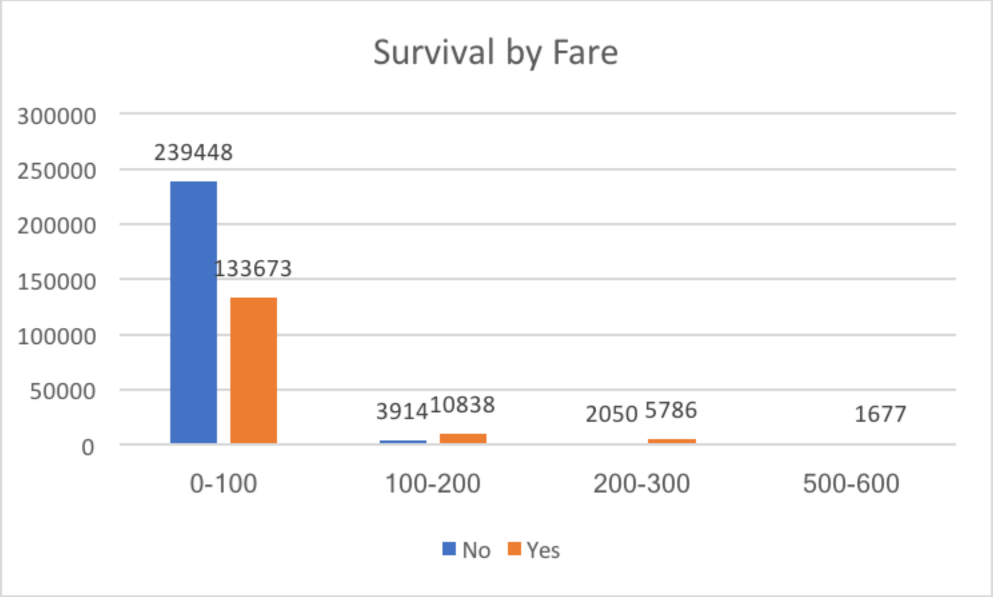
a statistical summary of the original data

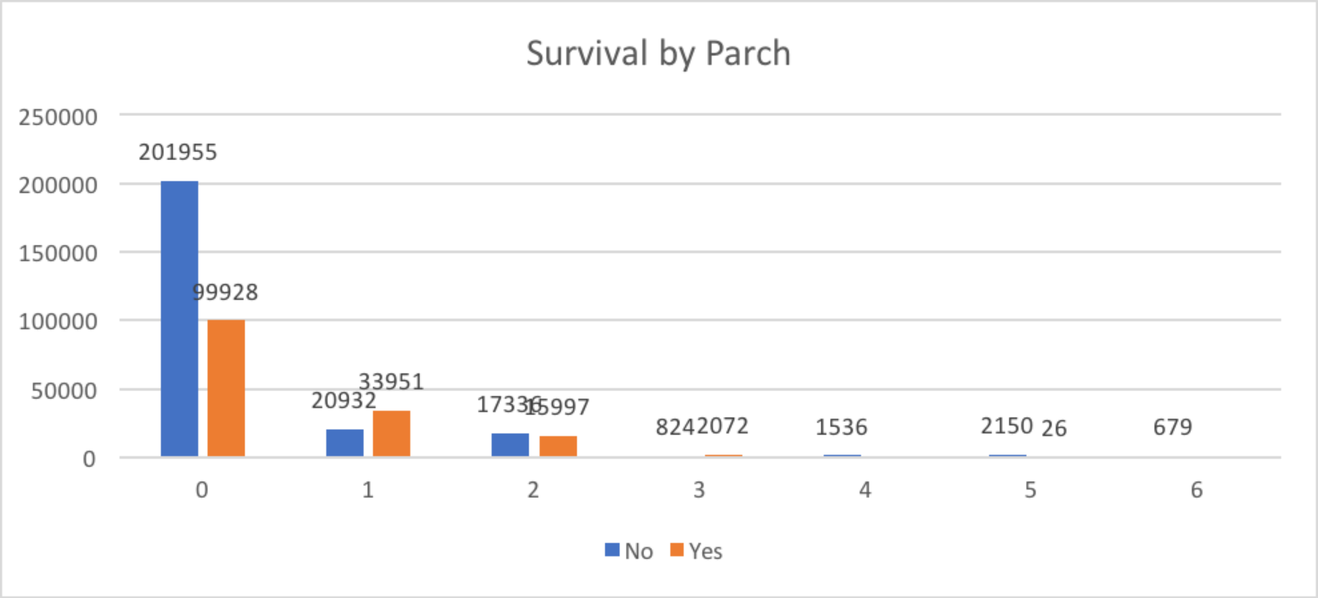

Survival by Age Group
70000
60000
50000
40000
30000
20000
10000
0
57877
19221
8930
17404
27565
17760
61524
33787
46364
32535
23992
17447
11056
10073
5125
3116
2979
631
<0 or (blank)
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
No
Yes

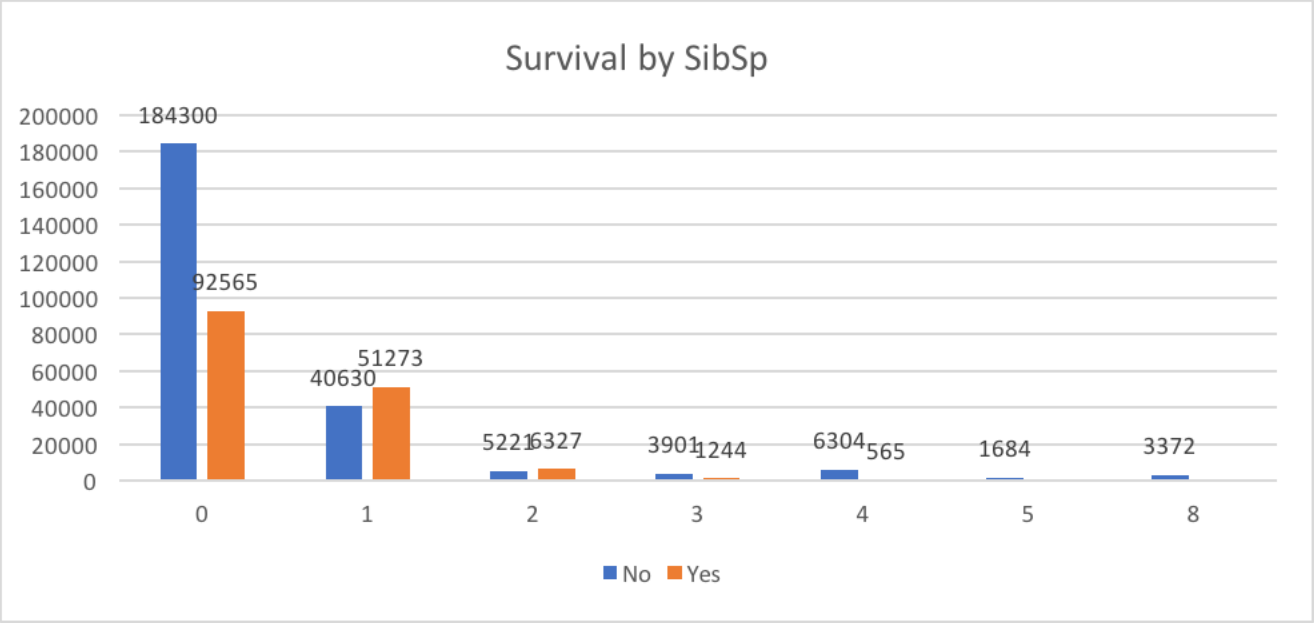

Survival by Pclass
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
32824
66881
43856
38200
168732
46893
1
2
3
No
Yes


Survival by Sex
250000
200000
150000
100000
50000
0
210189
100120
35223
51854
female
male
No
Yes


Pclass by Embarked
250000
200000
150000
100000
50000
0
192094
97402
44081
30739
22579
9599
892
C
Q
S
(blank)
No
Yes


Survival by Fare
300000
250000
200000
150000
100000
50000
0
239448
133673
3914
10838
2050
5786
1677
0-100
100-200
200-300
500-600
No
Yes





a discussion of model convergence

All the classification models benchmarked in the code converges well during runtime. In contrast, the clustering model (GaussianMixure) did happen to converge to local maximum given certain seed.

a summary of the compute requirements for processing the data.

Based on measurement on my MacBook (Early-2016), the resident memory consumed by Spark-related processes is from 385MB to 787MB. This is in line with the relatively small size of the datasets: test.csv (28K) and train.csv (60K)