Titanic Dataset
Analysis of Survivors

Introduction:

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

The goal of this project is to analyze the various factors which could have influenced the chances of survival as well as build a Random Forest model at the end which would predict whether a person would survive or not, given certain data.

Dataset:

The dataset was obtained from Kaggle. A description of the dataset is available as a separate file.

The data had 891 rows and 12 columns. A separate dataset was provided along with the first dataset which had passenger info for further 418 people but the data regarding their survival was missing from that dataset. The Random Forest model was fitted to this second dataset and the predictions for their survival were obtained.

Data Cleaning:

The following number of values were missing for the initial dataset:

Embarked	**-**	**2**
Age	**-**	**177**
Cabin	**-**	**687**

For the second dataset, the following values were missing:

Fare	**-**	**1**
Age	**-**	**86**
Cabin	**-**	**327**

Estimation for missing values:

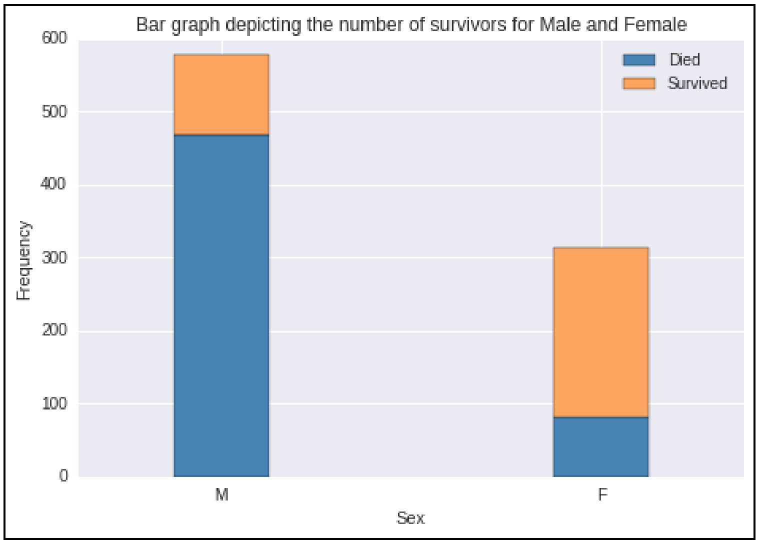
- The missing data for ages was estimated with the help of the names of the passengers. First the ages(where available) was obtained for various honorifics(Mr. , Mrs. , Master, Miss, Dr. , Rev. etc). Then the median age for each honorific was computed and this was in turn used to fill the missing ages column respective to the honorific used for the passenger.
- For missing values in Embarked column, the values were estimated using the mode of the embarkment data.
- For the missing values of Fare column, the median fares for various Passenger classes were obtained. After this, respective to the passenger class, the median fare was put as an estimation.
- The cabin data had a lot of values which were missing. Due to the lack of a proper algorithm for estimation, this column was dropped from the dataset.

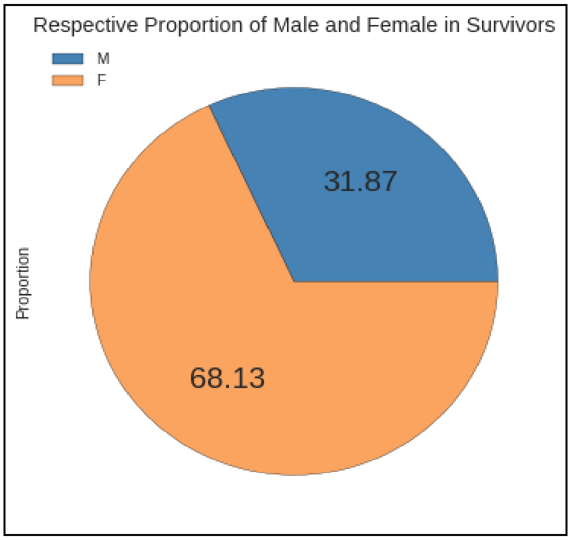
Analysis of various factors which could have influenced survival:

1. Survival by Gender:

Sex	Survived	Total	Probability of Survival	Proportion among Survivors
M	109	577	0.18891	0.318713
F	233	314	0.74204	0.681287

Clearly the Probability of Survival for women is far greater than that of men. This suggests that women were given a preference over men and were evacuated first.





2. Survival by Age:

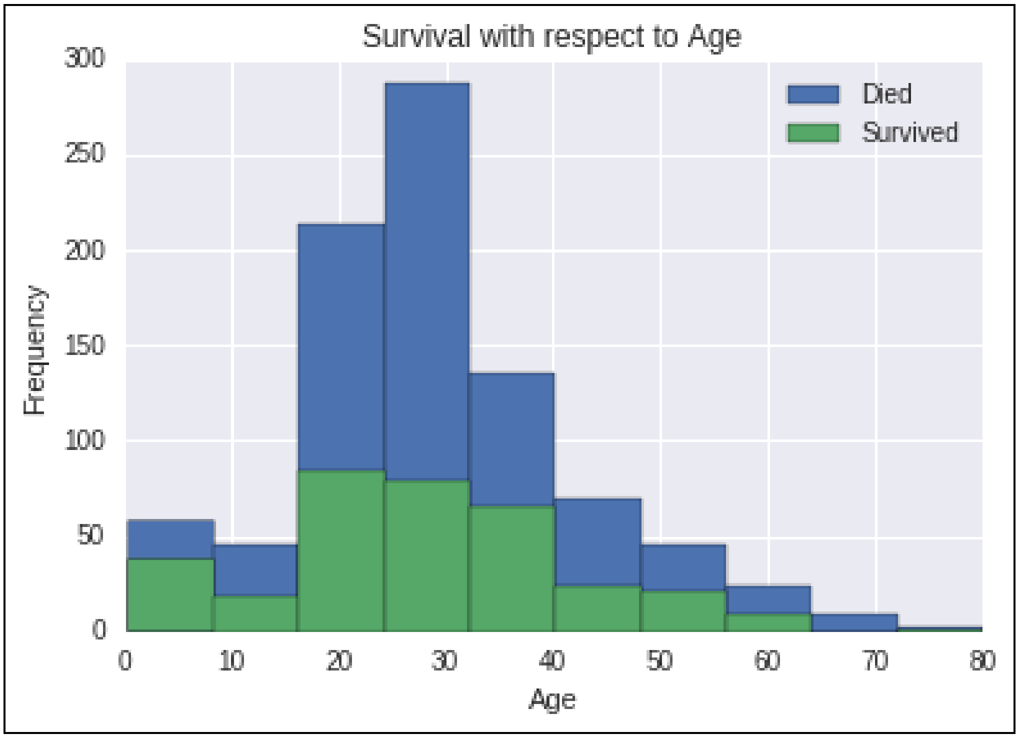
The following table illustrates the probability with respect to age.

	Survived	Died	Total	Probability of Survival
0-15	51	36	87	0.586207
16-40	236	417	653	0.361409
41-55	43	68	111	0.387387
56 & above	12	28	40	0.300000

The probability of survival is highest for ages below 16 and lowest for ages above 55. For ages between 16 and 55, it is almost similar.

This suggests that children were evacuated before adults.

Since, the probability of survival is affected a lot by the gender of the passengers, it has hard to draw any other meaningful inferences from this without isolating Gender.



The following tables illustrate the probability of survival with respect to Age for ***Women***:

	Survived	**Died**	**Total**	**Probability of Survival**
0-15	28	15	43	0.651163
16-40	168	55	223	0.753363
41-55	29	10	39	0.743590
56 & above	8	1	9	0.888889

This table is in direct contrast to the previous table. The probability of survival is the least for ages below 16 and maximum for ages above 55.

The chaos coupled with the harsh climate could have resulted in children getting separated from their parents. For the ages above 55, the number is too low to derive any meaningful inferences.

The following tables illustrate the probability of survival with respect to Age for ***Men***:

	Survived	**Died**	**Total**	**Probability of Survival**
0-15	23	21	44	0.522727
16-40	68	362	430	0.158140
41-55	14	58	72	0.194444
56 & above	4	27	31	0.129032

Here the probability of survival is the greatest for ages below 16 and lowest for ages above 55. For ages above 15, the differences in probability is not much and can be attributed to chance.

The survival probability for children is the greatest here but it is still lower than that of women. This seems to reinforce the previous theory that the resultant chaos and climate put the children to a disadvantage despite of them being given a preference.

Clearly, the men made way for women and children, and hence their probability of survival is the lowest.

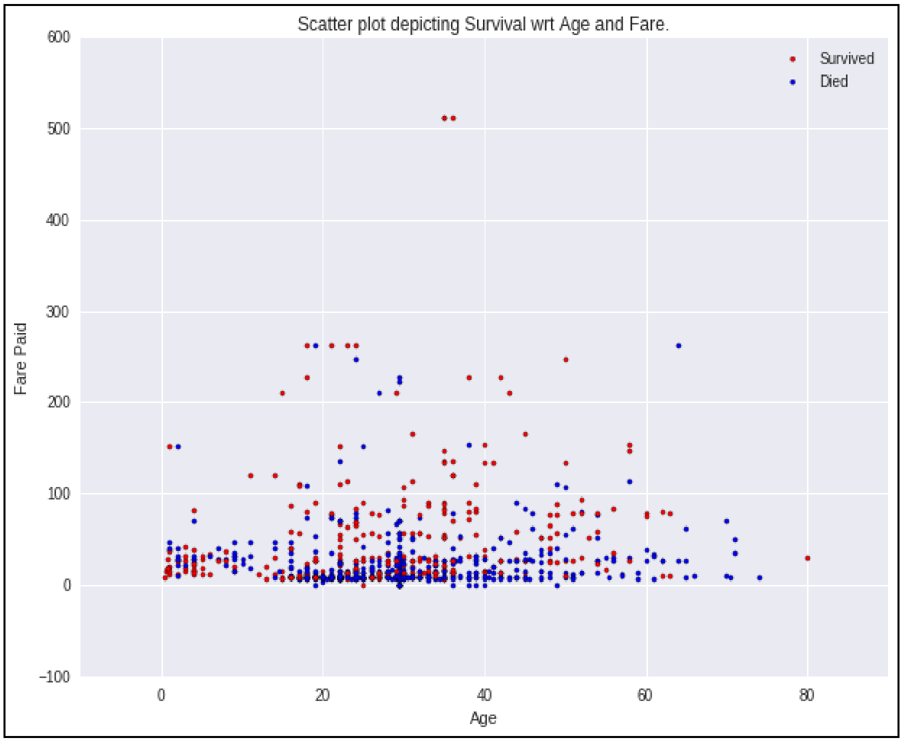
3. Survival by Fare Paid:

The following table illustrates the probability with respect to Fare.

	Survived	Died	Total	Probability of Survival
0-10	67	269	336	0.199405
11-30	139	182	321	0.433022
31-100	97	84	181	0.535912
100-200	25	8	33	0.757576
201-300	11	6	17	0.647059
301 and above	3	0	3	1.000000

From the above it is clear that the probability of survival is directly correlated to the amount of fare paid by an individual. As fare increases, so does the probability of survival.

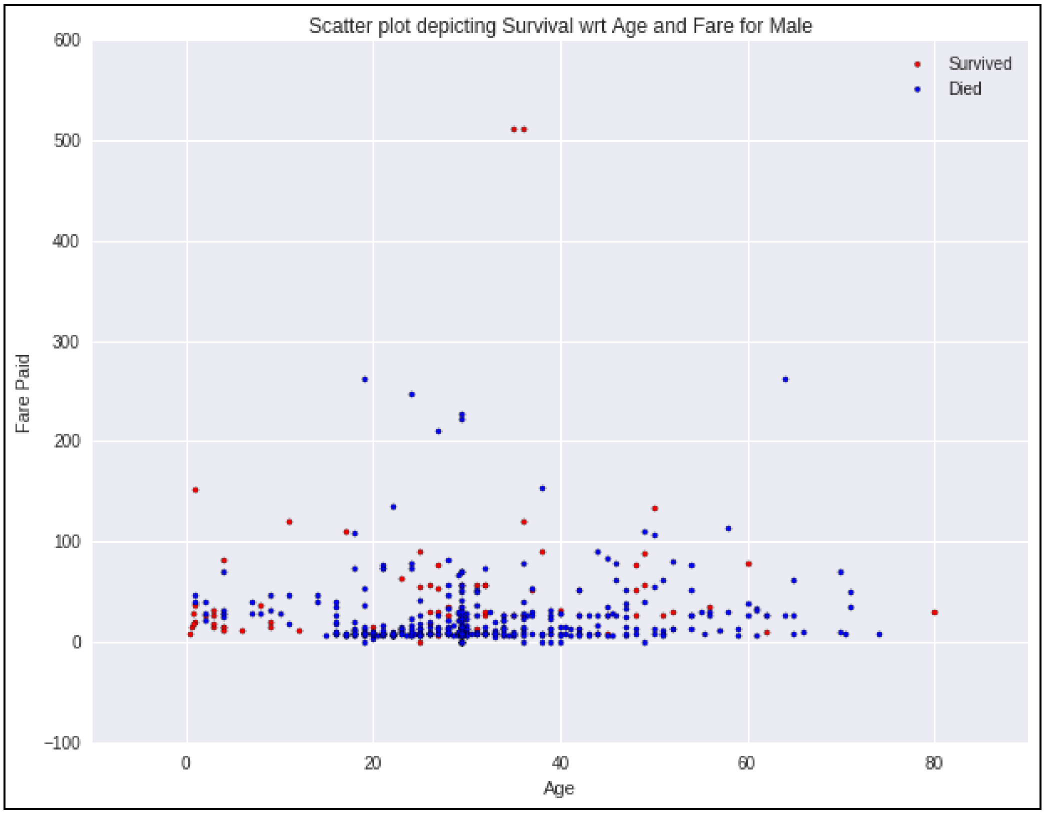
Only in the range 201-300 we see a drop in the probability. To analyze the cause of this we isolate the probabilities for different genders.



The following table illustrates the probability with respect to Fare for ***Men***:

	Survived	**Died**	**Total**	**Probability of Survival**
0-10	29	243	272	0.106618
11-30	41	143	184	0.222826
31-100	32	70	102	0.313725
100-200	5	6	11	0.454545
201-300	0	6	6	0.000000
301 and above	2	0	2	1.000000

Here also, the probability of survival increases with an increase in fare. Only in the range 201-300 there is a drastic fall. On careful examination of the data it was found that 2 out of the 6 passengers were from the same family. Also, 4 of the 6 passengers had cabins in the same section. There are very few passengers in this class and as a result it is affected by isolated incidents.

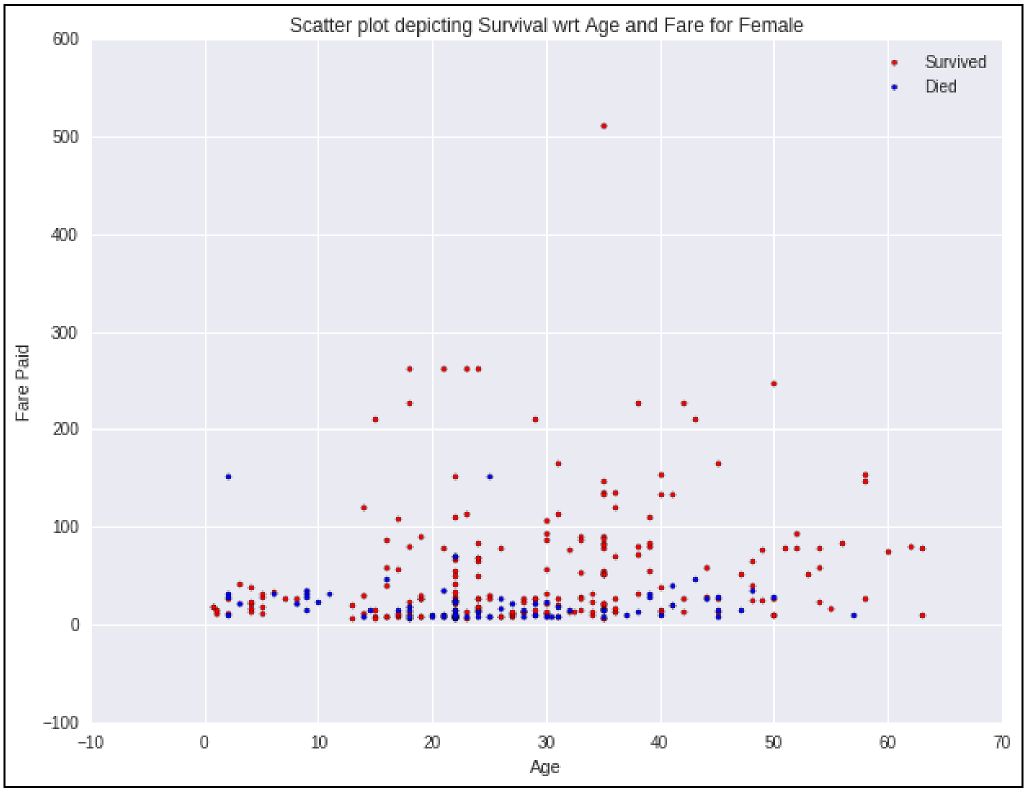


There seem to be no patterns here. The points representing survival seem to be all over the graph. There is a bit of a cluster for ages below 10 which further proves that the chances of survival reduces with an increase in age for men.

The following table illustrates the probability with respect to Fare for ***Women***:

	Survived	Died	Total	Probability of Survival
0-10	38	26	64	0.593750
11-30	98	39	137	0.715328
31-100	65	14	79	0.822785
100-200	20	2	22	0.909091
201-300	11	0	11	1.000000
301 and above	1	0	1	1.000000

From this table it is clear that there was definitely a class bias. The probability of survival increases with an increase in fare paid.



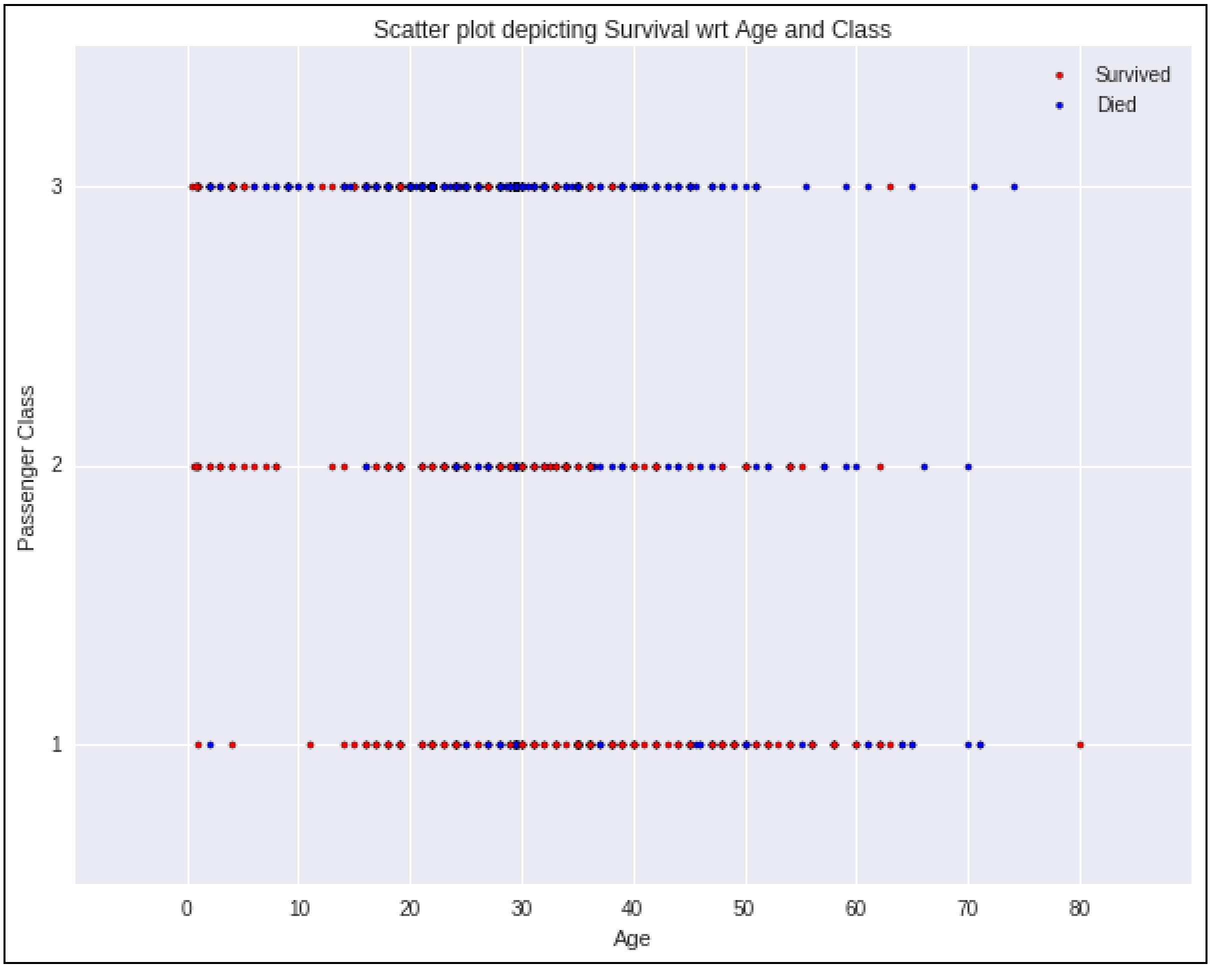
From the graph it seems that the points representing female passengers who died seem to be clustered around fares less than 30. There seems to be no obvious correlation with ages as such.

4. Survival by Passenger Class:

The following table represents the probability of survival with respect to the class the passengers were travelling in.

	Survived	Died	Total	Probability of Survival
1	136	80	216	0.629630
2	87	97	184	0.472826
3	119	372	491	0.242363

It further becomes clear that there was definitely a class bias while the people were being evacuated. The probability of survival falls as we go from 1st class to 3rd class.

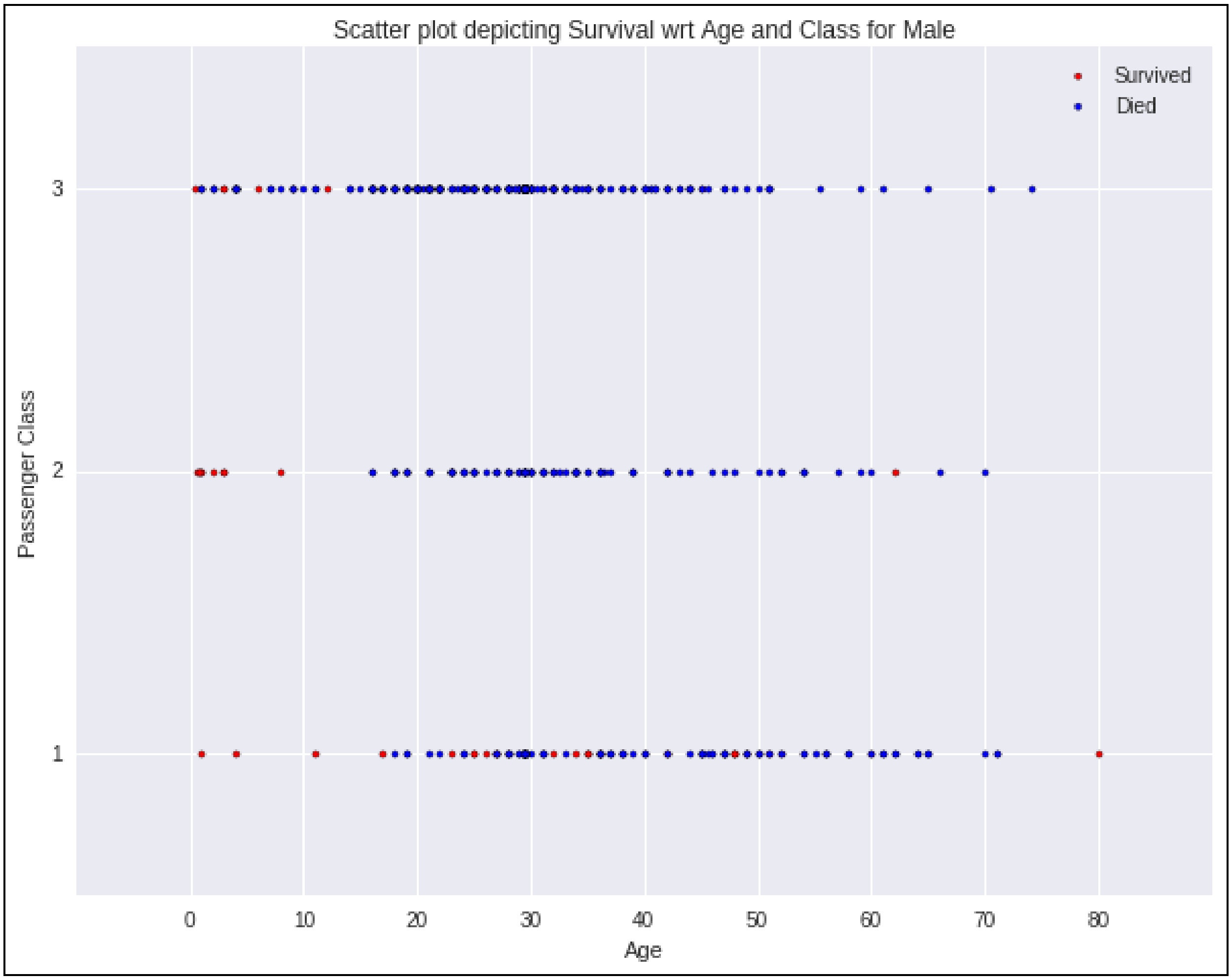


From the graph we can see that most people who survived, belonged to 1st class.

The following table represents probability of survival with respect to class for ***Men:***

	Survived	Died	Total	Probability of Survival
1	45	77	122	0.368852
2	17	91	108	0.157407
3	47	300	347	0.135447

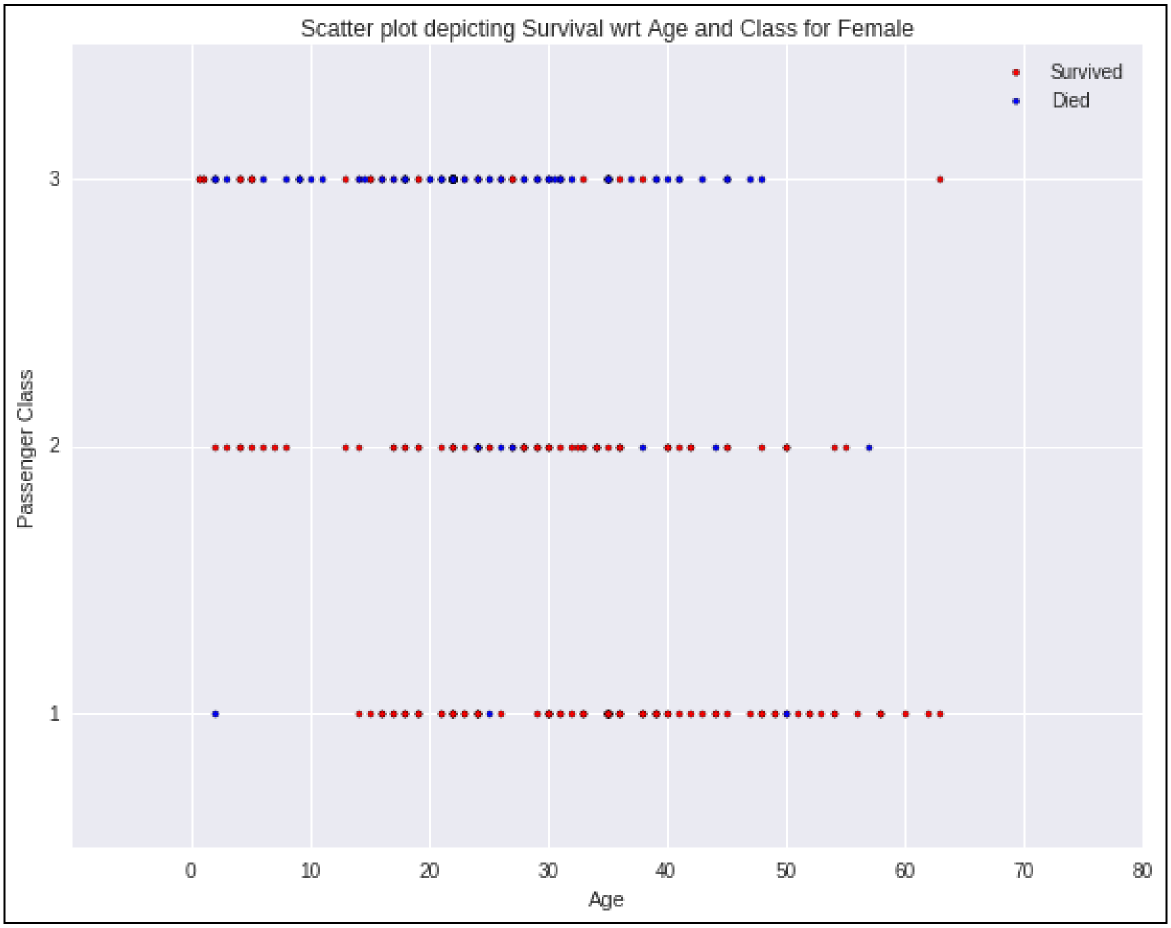
The trend continues here also. Passengers travelling in 1st class had much higher chances of survival than those travelling in 2nd or 3rd class.



The following table represents probability of survival with respect to class for ***Women:***

	Survived	**Died**	**Total**	**Probability of Survival**
1	91	3	94	0.968085
2	70	6	76	0.921053
3	72	72	144	0.500000

The probability of survival is the greatest for women travelling in first class whereas the probability of survival for women travelling in 3rd class is almost half of that of 1st class or 2nd class. This proves to be further evidence of class bias.



From the graph we can see that the women passengers who died seem to be evenly distributed in 3rd class. So, the low probability of survival is not because of the women being of a certain age but because of class bias

Random Forest Model:

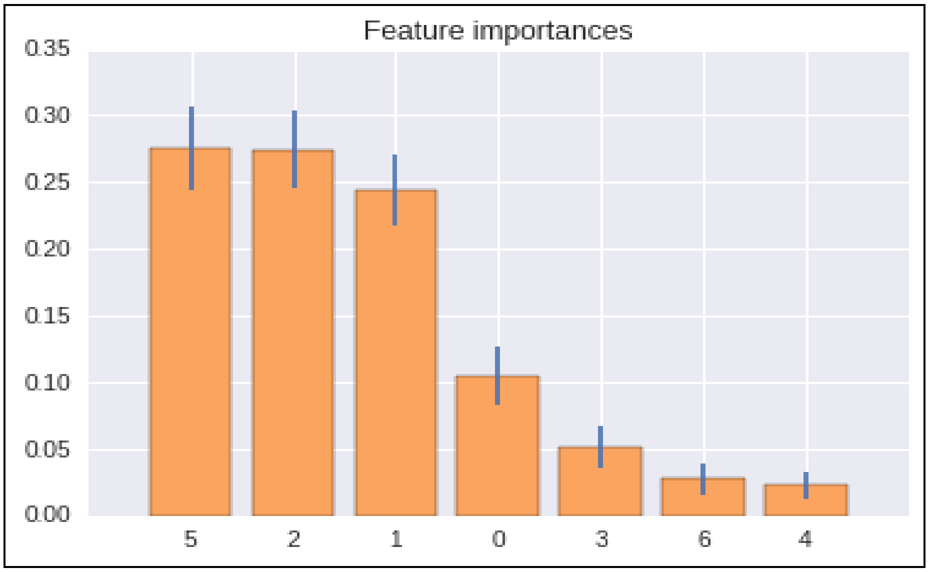
Based on the data for 891 passengers, a random forest model was constructed. Following were the parameters passed in it:

n_estimators = 100 max_features = None

criterion = 'gini' / min_samples_split = 5

oob_score = True

Following graph represents the feature importance got by the Random Forest Model:



0 - Passenger Class

1 - Sex

2 - Age

3 - Siblings / Spouse

4 - Parents / Children

5 - Fare

6 - Embarked

The random forest model so obtained was used to predict the Survival of the passengers in the second dataset. The dataset was then submitted to Kaggle and an accuracy score of 78.845 was obtained.