The Titanic:
Would you have survived?

by George Manuelpillai

RMS *Titanic* was a British passenger liner that sank in the North Atlantic Ocean in the early morning of 15 April 1912 after colliding with an iceberg during her maiden voyage from Southampton, UK to New York City, US. The sinking of *Titanic* caused the deaths of more than 1,500 people in one of the deadliest peacetime maritime disasters in modern history. The RMS *Titanic,* the largest ship afloat at the time it entered service, was the second of three *Olympic* class ocean liners operated by the White Star Line, and was built by the Harland and Wolff shipyard in Belfast with Thomas Andrews as her naval architect. Andrews was among those lost in the sinking. On her maiden voyage, she carried 2,224 passengers and crew. (wikipedia)

The Data

891 entries, 0 to 890
Data columns (total 12 columns):

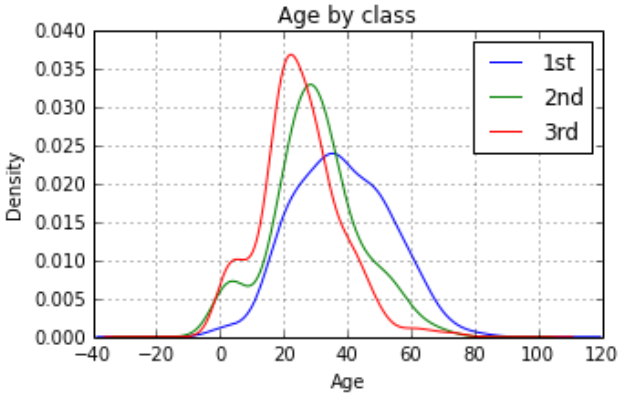
PassengerId	891 non-null int64
Survived	891 non-null int64
Pclass	891 non-null int64
Name	891 non-null object
Sex	891 non-null object
Age	714 non-null float64
SibSp	891 non-null int64
Parch	891 non-null int64
Ticket	891 non-null object
Fare	891 non-null float64
Cabin	204 non-null object
Embarked	889 non-null object

Feature Engineering

- Add family size = Sibsp + Parch +1
- Fare -15 zero values. These are across classes and obviously not free tickets. Replace with median fare per class.
- Sex - create new column “Gender” 0=Female, 1=Male

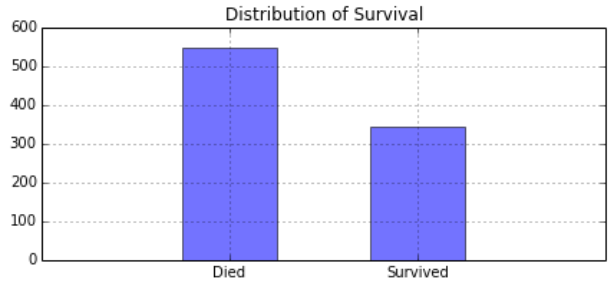
Missing values

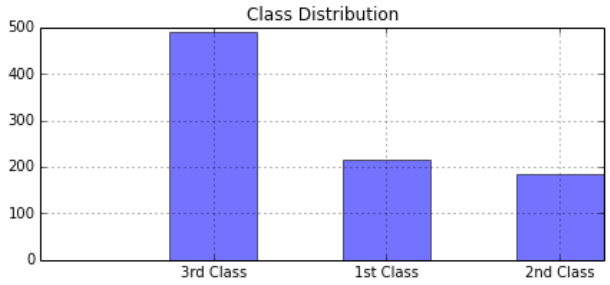
- Age - We can see that the age of passengers has different distributions based on class. So we will fill in missing values with median age per class

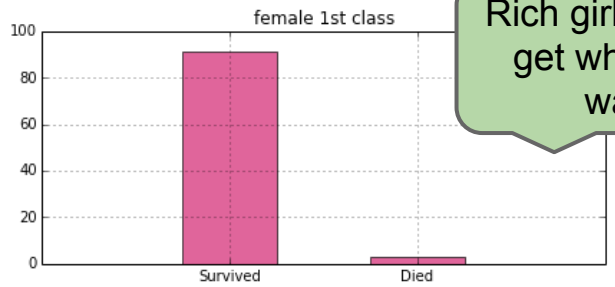


- Cabin - Did passengers in cabins closer to the lifeboats have a better chance at survival? Possibly. But hard to do with Logistic Regression. Skip it.
- Embarked - It is doubtful that port of embarkation had an effect of survival and in addition, it is only 2 values. So we will replace missing values with “S” as 664 passengers embarked from that port, using the mode function

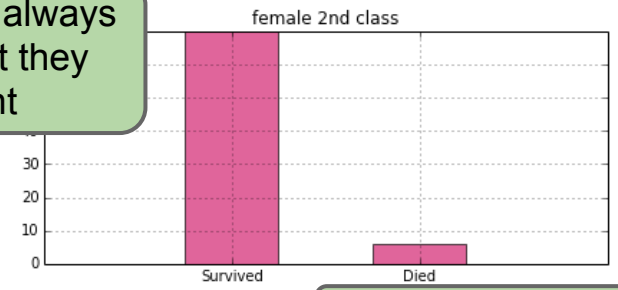
The Data - Visuals

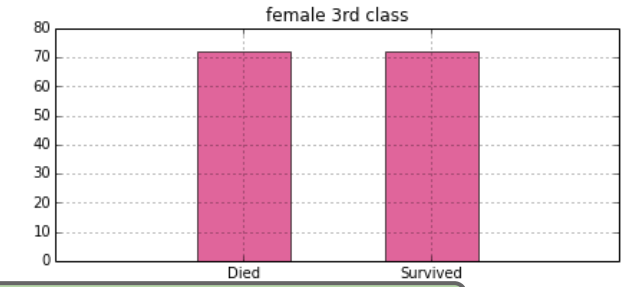




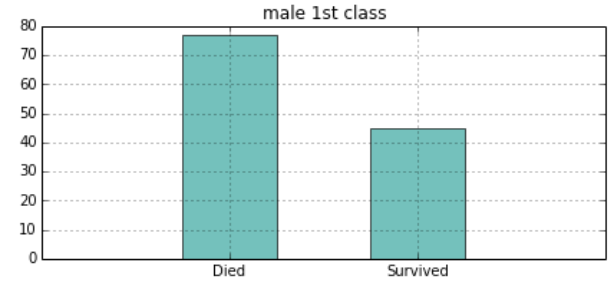


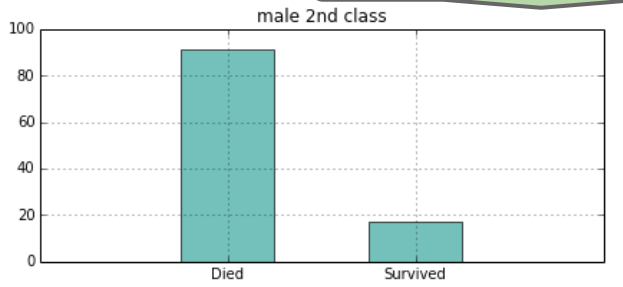
Rich girls always get what they want

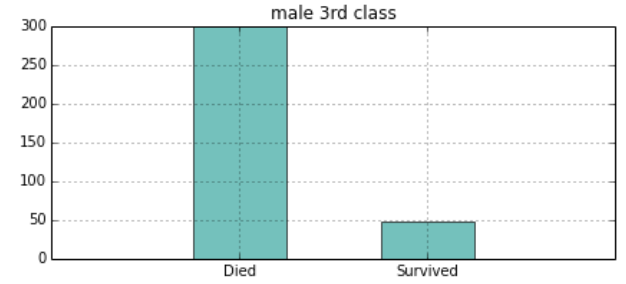




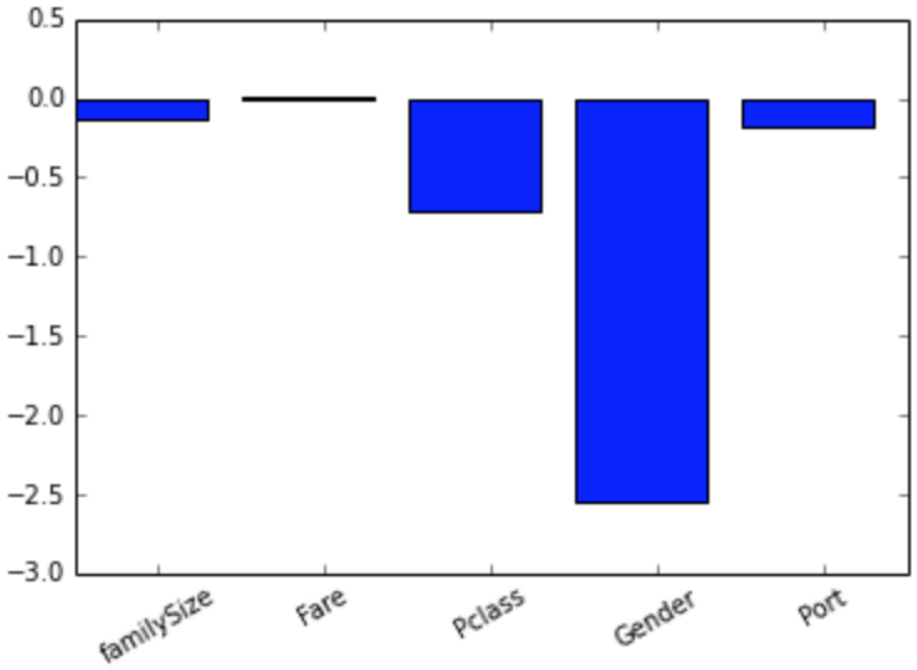
"BE A MAN". Save others, not yourself







Logistic Regression



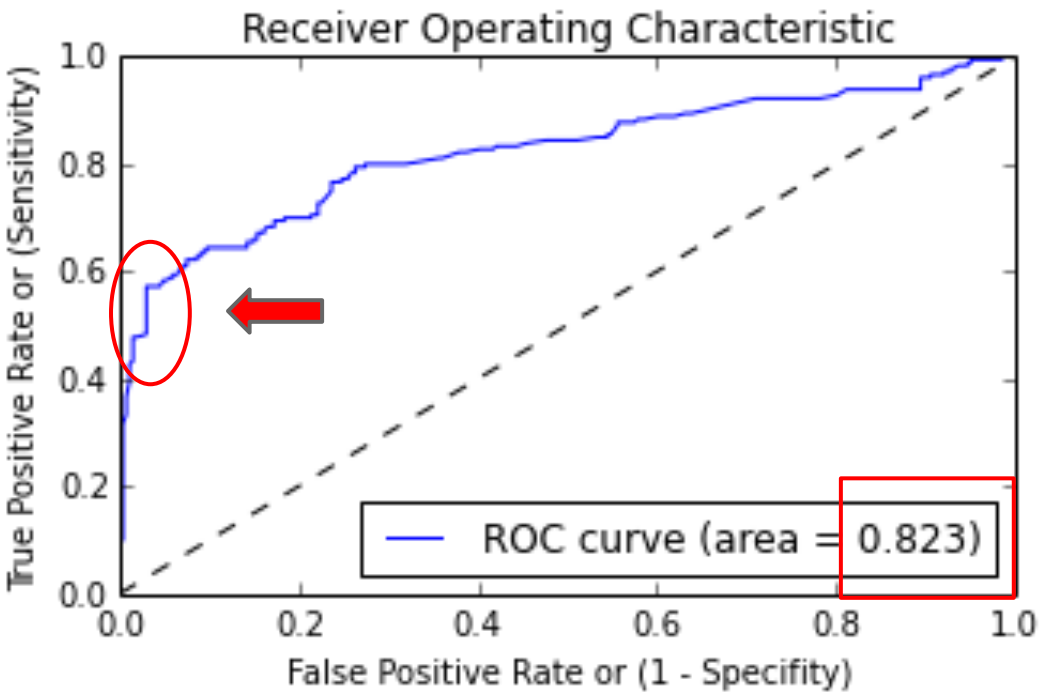
KFolds: Using cross validation on the model and a kfold of 12, I was able to get a prediction score of 80%. Given that my kaggle score was 78.47%, my prediction score could not increase much further. A kfold of 5-12 is more than sufficient

The coefficients on the left show that gender and class were the most influential factors in determining survival. This is evident in:

- High percentage of women in first class surviving, probably due to class structures that existed at that time
- High percentage of males perishing regardless of class. The men were probably tasked with aiding women and children into lifeboats.

If you were part of a family, and not the alpha male, you probably had a better chance of beating the odds. Did not investigate.

ROC / AUC



As you can see from the graph, I can increase the true positive rate while not sacrificing on my false positive rate.

I varied my threshold value to take advantage of this. By adjusting my threshold value from .50 to .58, I was able to increase my score.

882	new	**George Manuelpillai**	0.78469	3	Wed, 12 Nov 2014 05:25:46

Your Best Entry
You improved on your best score by 0.00957.

You just moved up 413 positions on the leaderboard. Tweet this!

	Gender, Price and Class Based Model	0.77990