Analysis of survival data aboard the RMS Titanic

Introduction:

The accidental sinking of the RMS Titanic when she collided with an iceberg is one of the largest documented civilian maritime disasters of the last century. A large number of people when the ship crashed into an iceberg on the night of April 15th 1912.
This paper explores a dataset consisting of several data points for passengers of the Titanic, including whether they survived or not. The paper seeks to explore whether there were specific data points that contributed to the survival of individuals, using several supervised-learning based machine-learning tools for classification.

1.1 Summary of the data set and handling of missing values

The data set was pulled from www.kaggle.com , which is running a competition to see whether disasters like the Titanic can be predicted based on passenger information.

The data set consists of the following data points. Columns in red had missing values:

1. Passenger id # - this was disregarded as this was just an up-counting number
2. Survived (1/0) - Used in determining accuracy of any algorithm.
3. Socio-economic class indicator (1,2,3 indicating high, normal or low social standing)

4. Passenger Age - missing values were substituted with a value of 29.5, as this was the average age of people (with this parameter specified) on board the ship. Further, ages were converted to classes 0 and 1 denoted by age brackets < 18 and >=18 respectively. This was done to test the hypothesis that more children survived than adults.
5. Passenger Sex
6. Passenger Name - this attribute was discarded.
7. Presence of Spouse/Siblings on the ship
8. Presence of Parents / Elders
9. Ticket number - This attribute was discarded as it was merely a code.
10. Ticket price - Missing values were substituted based on parameter 3, the socio-economic status indicator of the passenger. Given that there were 3 classes of tickets (1st, 2nd, and 3rd), the average ticket prices (averaged across all customers in a particular class) of GBP 84, GBP 20 and GBP 13 were used to fill in values for the missing ticket prices.
11. Cabin - The Titanic was divided into Decks A-G. The passenger quarters were located on decks C-G, with C housing primarily first class passengers, D holding second class passengers, and F and G housing second and third class passengers. G deck was located just above the waterline. For missing cabin numbers, a mapping between Socio-Economic status was once again used, with classes 1,2 and 3 assumed to be on decks C, D and G respectively.
12. Port of embarkation - Passengers boarded the TItanic at 3 ports: Southampton in

England, Cherbourg in France and Queenstown in Ireland. Since the bulk of passengers boarded at Southampton, this was used to fill in missing values. Note, that a better approach would be to do parsing on passenger name, and infer nationality, and appropriately set the departure port.

2.0 - Hypothesis

The paper aims to capture whether some criteria in the data set had more of an impact on passenger survival than others, and if so, which variables were these?

2.1 - Techniques employed

The output variable is a binary number, 0 and 1, implying that this is a fairly obvious classification problem - did the passenger in question survive or not.

3 Different machine learning techniques were used - Naive Bayes Classifier using a Gaussian distribution, Random Forests with 10 estimators, and an SVM with a linear kernel.

The simulation was run in using scikit learn and numpy modules available in the open source distribution of python.

In all cases, classifiers were built using 5-fold cross validation on the data set using scikit-learn's cross validation framework.

2.1.1 - Types of tests run

- Individual attributes - All classifiers were run using each attribute individually.
- All attributes - The training set was run against all classifiers with the valid data points included as mentioned above.
- All attributes without the sex of the passenger.

2.1.2 - Results of tests

- **Individual attributes** - The performance was varied for most attributes individually. The top 3 attributes that best predicted results were passenger sex, fare and social class with accuracies 0.79, 0.68 and 0.679 respectively as depicted in figure 1. Passenger sex was easily the best classifier observed here.
  Surprisingly, age bracket wasn't a particularly good classifier here. The reason is most likely the training set. Of the 113 rows of training data that represented children (<18 years), only 61 survived (53.98%). Further, 19.75% of the data set consists of passengers with no age specified. Any assumptions we would make about age would be speculative at best.

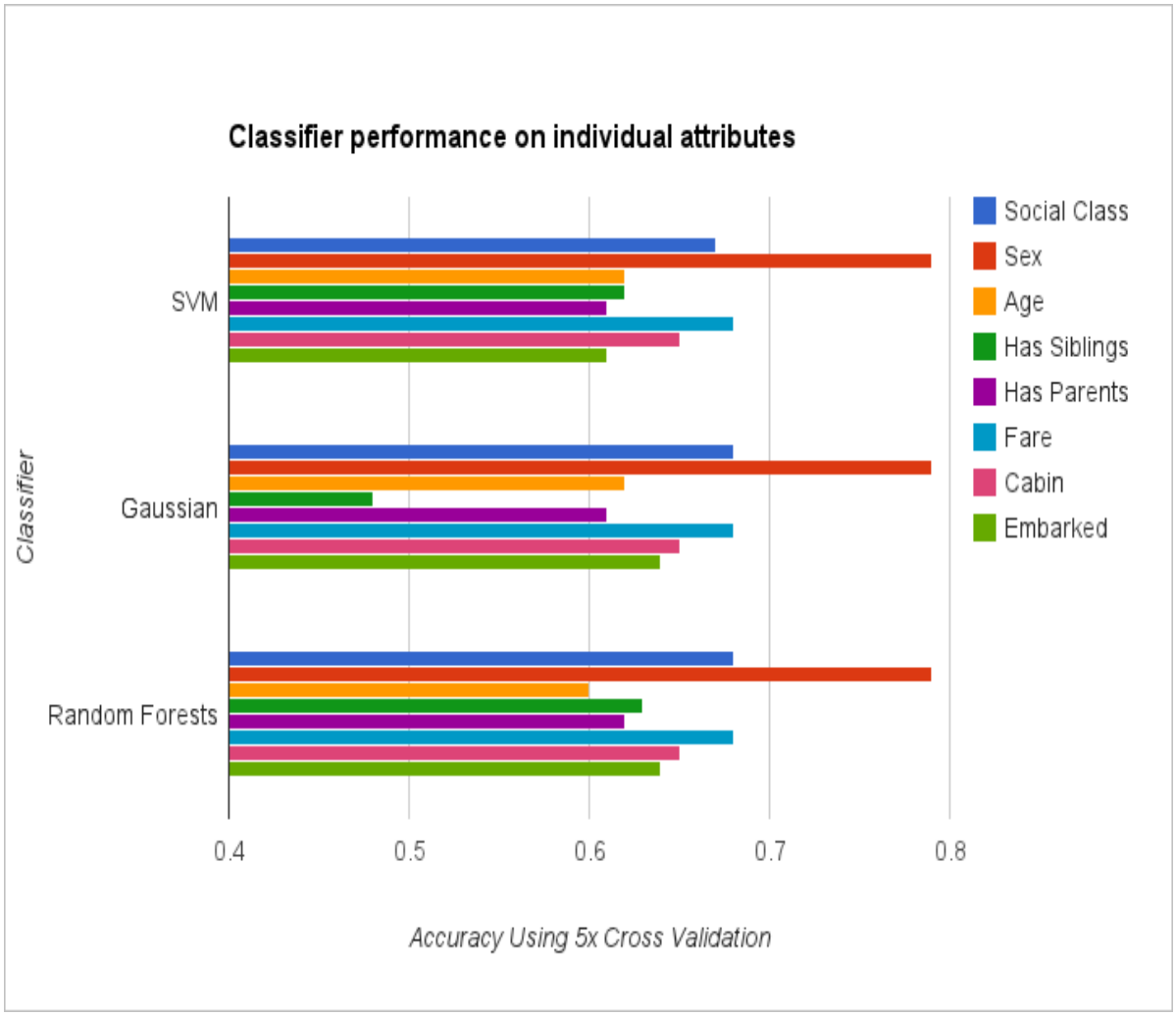

Figure 1 - each attribute individually scored using each classifier

- All attributes - SVMs, Naive Bayes and the random forest produced results with accuracies 0.79, 0.76 and 0.80. The random forest with 10 estimators won out as the best estimator using all variables. It is likely that the forest used a tree that split on passenger sex and came up with a result closest to if sex alone had been used to construct the classifier. See figure 2.
- All Attributes w/o sex - The performance of all the classifiers deteriorated

significantly as can be seen from figure 2.

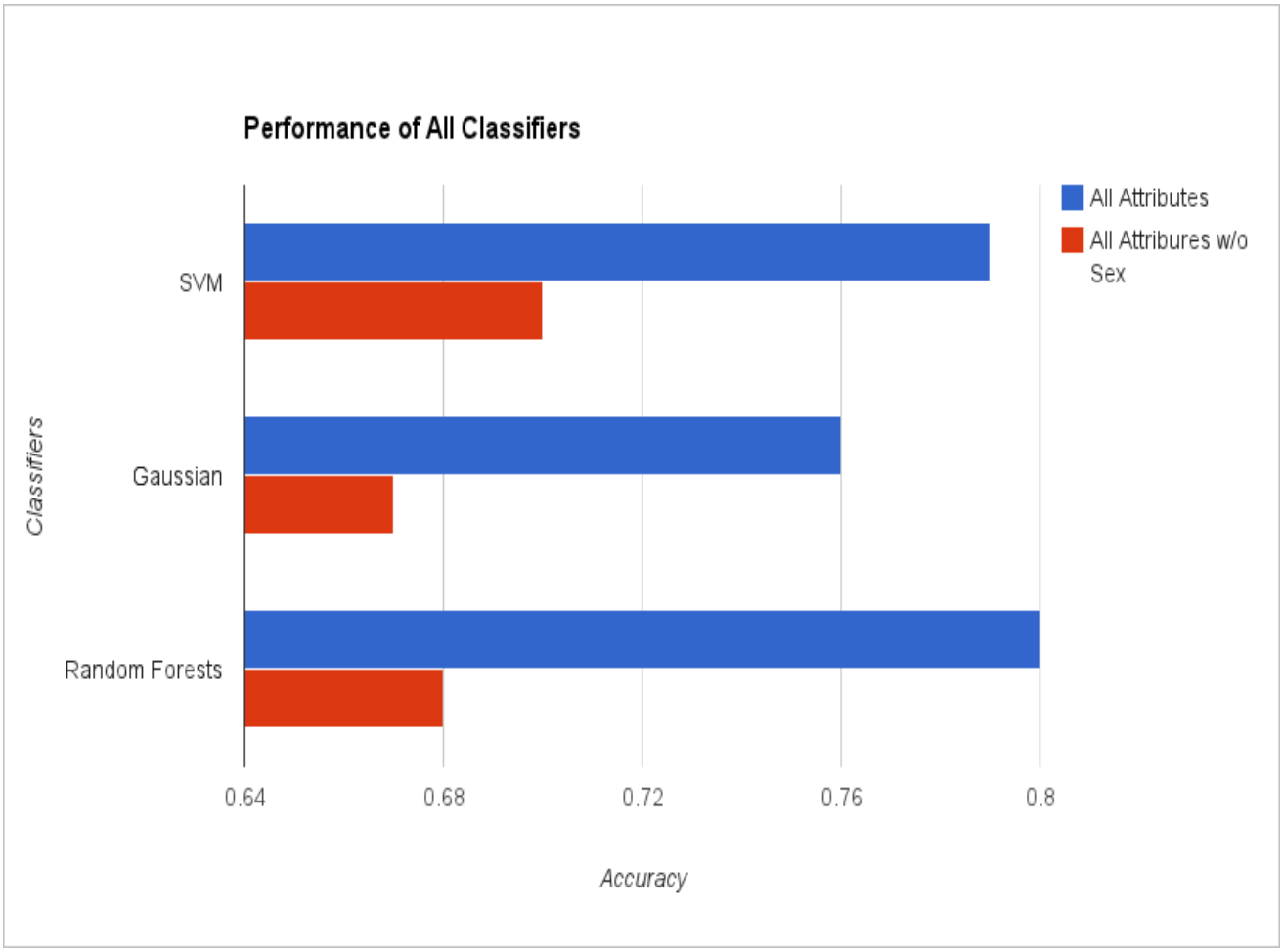

Figure 2

2.1.3 Conclusion:

From the training data set, this is corroborated by 233/314 or 74.2% of female passengers having survived, and 109/577 or 18.89% of men surviving, or more specifically, 81.11% not surviving. All models built produce results that are very supportive of this data, that sex itself was the single biggest determining factor in the training data that affected the outcome of the passenger.