The RMS Titanic - A Machine Learning Exercise

Mike Dean

Abstract:

April 15, 1912 was the date of one of the most famous tragedies in Human History. Most people know the story of the RMS Titanic; the ship was considered "unsinkable" by the standards of the time, but after an unexpected encounter with an iceberg, this claim was quickly disproven. Of the approximately 2224 passengers and crew aboard for this maiden voyage from Southampton, England to New York City, approximately 1502 died. As was, and still is, common practice, the evacuation of the Titanic followed a "women and children first" protocol. Sex and age were not the only factors to determine whether a passenger survived. Through multiple machine learning techniques, I hope to be able to illustrate why some survived while others perished.

Introduction

Machine learning is a method of statistical analysis in which a computer program is used to determine the best algorithm to fit a dataset with known outcomes in order to predict the outcomes of a similar dataset. The former dataset mentioned above is known at "training" data, while the latter is considered "test" data.

About The Data

This dataset comes from a current open competition at Kaggle.com. Kaggle refers to itself as "The Home of Data Science", hosting competitions in Data Analysis that anyone can enter with prizes of up to $500,000. The Titanic competition, however, is solely for learning how to conduct machine learning experiments and offers no monetary reward.

As mentioned earlier, in order to conduct a machine learning exercise, you need to have a "training" dataset and a "test" dataset. In this case, Kaggle has provided these sets as separate files. Both the training data and the test data contain the following headings. First is "PassengerId", which is used only to index the passengers for the data. Next is "pclass", which corresponds to the passenger's class on the ship, this can have a value of "1", "2", or "3". This can be considered an estimate of socioeconomic status. Next is "name", which just contains the name of the passenger. Sex is either "male" or "female". Age, if known, is an integer if at least one year old, any age under one is a decimal value. The number of siblings or spouses of the passenger on board; siblings including only brothers, sisters, stepbrothers, or stepsisters and spouses only counting husbands and wives. The number of parents and children on board, where a parent is a mother or a father while a child includes stepchildren. The ticket number for the passenger. The fare paid for the ticket. The cabin number, and finally, the port of embarkation which can be either "C" for Cherbourg, "Q" for Queenstown, or "S" for the ship's port of origin, Southampton. Additionally, the training data contains a column labeled "Survived" which can either be a "0" or a "1" with a "1" indicating that the passenger survived.

It is important to note a few things about this data. First of all, it does not include the crew at all. While the reasoning is not clear, it is most likely due to the fact that a much larger proportion of the crew did not survive. Additionally, the sample would be skewed to include more men as only 23 of the nearly 900 crew members were female. **(Encyclopedia Titanica)**. Finally, the crew would most likely also contain a number of statistical anomalies as the

more honorable upper-level crew members would have decided to "go down with the ship".

The data includes a total of 1309 passengers, 891 of which are included in the training data with the remaining 418 in the test set. In the training set, the median passenger age is 28 years, 314 passengers were female while 577 were male. The majority of these passengers, 644, embarked from Southampton while 168 boarded in Cherbourg and the remaining 77 in Queenstown. The median fare paid for a ticket was $14.45. There were 216 first class passengers, 184 second class passengers, and 491 in third class. Finally, of the 891 passengers in the training data, 342 survived while the remaining 549 were among the dead.

Methods

The main purpose of this experiment was to learn more about machine learning and statistical methods. Therefore, I utilized a few different analytical methods through the use of Microsoft Excel, Python, and R. Each of these tools succeeded with differing rates on finding a useful algorithm from the training data. While they are all powerful tools, R came out ahead by being the most useful tool for analyzing the data with Python in a close second place. While Excel has the most intuitive user interface, it is much more useful for looking at relationships in data rather than machine learning.

Excel

First, using Excel, I used pivot tables to examine the relationship between different variables and the survival rate. Specifically, I explored the “women and children” by comparing survival rates with sex and age, as well as passenger class. While this method does not allow for a sophisticated algorithm to be developed, it helped me to visualize the data and see initial patterns that may be useful to follow as my methods became more sophisticated. I was also able to use the patterns witnessed in the pivot tables of the training data to apply simple algorithms to the test data.

Python

After the initial patterns were found in Excel, I was able to apply some of this information in a more sophisticated manner using Python. First, I used simple data analysis tools, namely, csv and numpy. The csv package simply allows for the reading, writing, and manipulation of text files saved in the common comma separated value (csv) format. Numpy, on the other hand is slightly more powerful, it is a package used for scientific computation. The most important feature for my purposes is its ability to create and manipulate n-dimensional arrays of data. The main advantage to using Python, or any scripting language, is the fact that it can handle data much more elegantly. A python program is run from a python script file, rather than running the analysis from the actual data file. This helps to preserve the original dataset and it preserves the operations you run on the data so doing analyses on different files that include the same type of data is somewhat trivial.

Python and Pandas

While my initial experiments with Python were essentially the same operations as the pivot tables in Excel, I also implemented a random forest using the Python package, "pandas". The "pandas" package for Python is a

Data Analysis Library that allows for the use of decision trees and other statistical models. The first challenge I faced while using "pandas" was the fact that it can only use floating point numbers for inputs and outputs. For the purposes of the Random Forest in Python, in the “Sex” category converted “female” and “male” to “0” and “1” respectively.
Additionally, in the category of “Embarked”, I set the value for “Southampton” to “1”, “Cherbourg” to “2”, and “Queenstown” to “3”.

These were the only string values replaced with numerical values in this Random

Forest. The columns labeled “Name”, “Ticket”, and “Cabin” were all dropped from this analysis as they contained non-numeric values.

Next, I had to deal with missing values. There are many passengers in both datasets whose Ages are missing. In order to account for this, I replaced missing values with one of six median age values based on the Sex and the Class of the passenger. These values are as follows: First class females; 35, second class females; 28, third class females; 21.5, first class males; 40, second class males; 30, and finally third class males; 25.

The final manipulations I had to make to this data is dropping rows with missing information. Unfortunately, other than Age, not much else can be generalized and inferred from looking at the rest of the data so and rows that contained null data were dropped from the algorithm.

Once the data was completely usable by Python and Pandas, I was able to implement a Random forest. A Random forest is essentially a group or “forest” of decision trees. Decision trees are used to classify data. In a decision tree, a piece of data is submitted to the “root” of the tree which uses a classifier to determine which “branch” it should proceed down in order to determine its value when it reaches the outermost branch. One of the disadvantages of classic decision trees is overfitting to the training data. Random forests help to generalize the prediction algorithm by creating numerous decision trees and randomly submitting a subset of the training data to the decision trees while reserving the remainder of the data for error calculation and to determine the importance of each variable. (**Liaw)**

R

I started off in R by essentially performing the same analysis I did in Excel as well as the analysis in Python described above. My first R-specific experiment was a simple decision tree. Just as with Python, I ignored the values for the “Cabin”, and “Ticket” entries.

Prior to moving on to more powerful machine learning techniques, I manipulated the data so more of it could be used for analysis. Feature engineering is another important aspect of machine learning, one in which seemingly unusable data is altered in some algorithmic way in order to allow for statistical analysis to be performed on it. For this dataset, I engineered a few features. The first features I engineered were based on the passenger name.

Using R’s built-in functions, I was able to split the names of all of the passengers into three parts, saving the title and the surname. For example, the “Name” variable of the first passenger in the training set, “Braund, Mr. Owen Harris”, was divided into the category “Surname” which is “Braund” and the category “Title” which is “Mr”. After combining some of the more rare titles they become much more useful. Using the Surname, combined with the “FamilySize” (a combination of the “SibSp” and the “Parch”) I was able to assign identification numbers to each of the families in the dataset.

After feature engineering was complete, I ran a much more powerful random forest than in I was able to in Python due to a richer dataset.

Results

It is commonly known that there were fewer survivors than fatalities on the Titanic, so the first submission I made to the Kaggle competition was simply to assume that everyone in the test set died. This results in 62.679% accuracy.

Excel

Using Excel, I was able to determine statistically that females were much more likely to survive than males. Using pivot tables and conditional statements, by simply assuming that all males died and all females survived, accuracy rises to 76.555%.

Python

Python allowed for some slightly more powerful analysis. By factoring in fare paid as well as Sex, accuracy increases to 77.99%. I was also able to implement a simple random forest which actually decreased the accuracy to 77.512%.

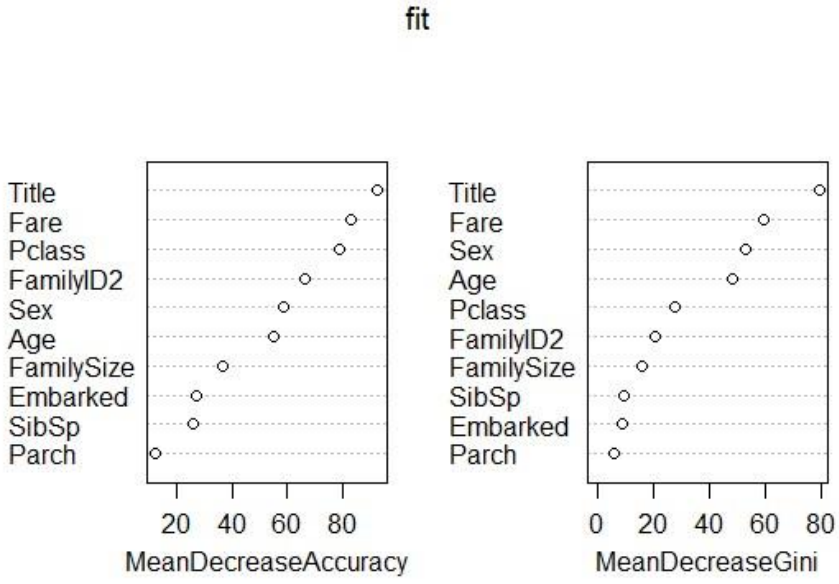
R

As described in the "Methods" section, I was able to use feature engineering to best utilize the data. Using a random forest combined with this improved dataset, accuracy rises to its highest value, at 81.34%. What is most interesting about this approach is the importance of certain variables that were not even considered in my previous statistical analyses. This plot graphs the variables based on their importance in determining survival. As you can see in the plot, the most important variable in determining survival was the passenger's title. Additionally, it appears that the number of siblings and spouses and the number of parents and children on their own were not important in determining survival, but Family Size, and FamilyID, which were both functions of those less important variables.

Conclusion

Machine learning is a powerful method of statistical analysis. It allows patterns in data to be seen that may not have been previously. Although the Titanic's "women and children first" policy of evacuation was reinforced by all of these experiments, some new information was discovered as well. The fact that passenger's titles played an important role in their survival is particularly interesting as initially this information was ignored.

Feature engineering has proven to be an important step in creating a successful algorithm. Hopefully this success can continue to be improved upon as more advanced techniques of machine learning can be applied.



Encyclopedia Titanica - Female Titanic Crew. (2013, November 13). Retrieved November 22, 2014, from http://www.encyclopedia-titanica.org/titanic-female-crew/

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R news*, *2*(3), 18-22.