Discovery of Survival behind the Titanic Sunk: A Data Mining Approach based on the Survival's Information

Pengshengnan Cheng(V00838497) and Mingyu Du(V00815833)

Department of Computer Science, University of Victoria

Contents

1 Introduction

1. What is Kaggle?

Kaggle was founded as a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models.[1]

2. Background information of the Titanic

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.

3. What we do?

In this project, we complete the analysis of what kinds of people were likely to survive. Specifically speaking, we make advantage of using Weka, which has integrated with many machine-learning oriented algorithms, to predict which passengers survived the tragedy. Besides, we find that Microsoft Excel can be effective tool to do the data cleaning when the Weka is not able to clean the data and we would like better desired results.[2]

2 Related Study Field

Travel tragedy could happen at any time. If it is possible to find out the potential relationships between some certain patterns such as age, we believe that we can help insurance company to target audience with more accuracy so that people can be more prepared to encounter unexpected events.

We start with doing research data retrieved from Kaggle website, the Titanic shipwreck is being well known to all over the world. Many studies have been done for researching for reasons that why this tragedy happened and hope that such a tragedy could be avoided. That is to say, the design and construction of ship or airplane can be more effectively designed in case of emergency on the means of traveling. The resources can be more wisely allocated.

Moreover, our study can help research and rescue teams or agencies to be professionally prepared as we may develop model to anticipate the best rescue plan, in a statistically way. In other words, the rescue team has more information so the rescue team members can put more efforts into saving those have higher chance of surviving.

All in all, we believe our study can be beneficial to the three aspects mentioned above.

3 Data Processing

The preprocessing is aiming at clean and data attributes separation. So we have the following list to do:

1. Data Cleaning

2. Useless information deletion and Attributes separation

3. Transform csv file into arff file.

1. Data Cleaning

The data csv file is download from Kaggle Titanic: Machine Learning from Disaster project. Every column is an attribute. In the Name attribute we need to delete symbols such as “,”, “’” and () to make sure Weka can load data csv file correctly. We use replace from Excel to do this.

2. Useless information deletion

   In Name attribute, the whole name and the title such as Mr., Miss., and Mrs. are in there. We only keep first name (We think the name will not make a difference for probability of survive). And we add a new attribute for title. Also we think Ticket attribute does not influence the prediction, because this attribute basically provide the tickets number. So we decide to delete Ticket attribute.

3. Transform csv file into arff file

   The Weka can load csv file directly. We can use Weka Explorer open data csv file. Because we want to predict Survived attribute, we need to use Edit to make Survived as class and save as arff file. The Weka read Survived attribute as an Numeric attribute, so we can change this into Nominal by edit arff file. In the same way we change Pclass to Nominal from Numeric. In order to make data mining part more quickly, we also choose Normalize filter and apply on attributes.

   Also we need to do everything above to test data csv file to make sure we can get a match between two data file.

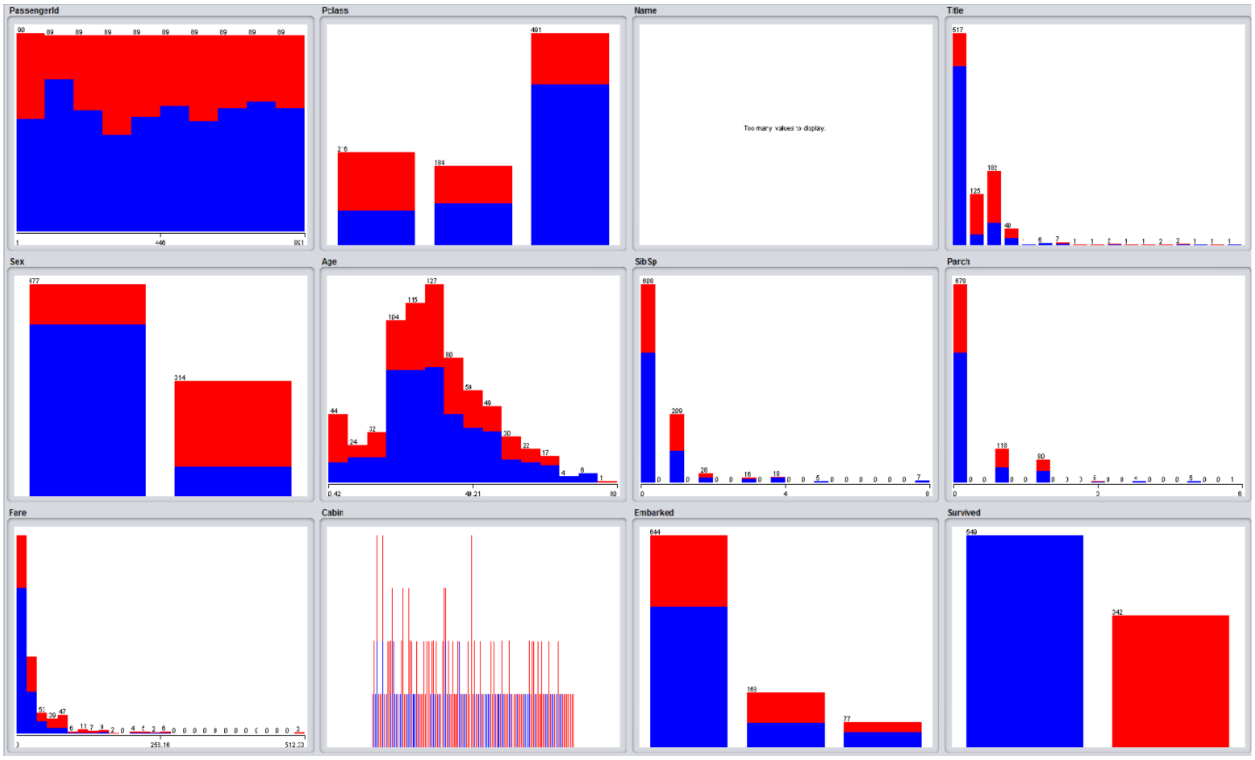
   So far the attributes left are like this.

Attributes	Description	Type
Survived	Survival	Nominal, Class
Pclass	Ticket class	Nominal
Name	Passenger name	Nominal
Sex	Sex	Nominal
Age	Age in years	Numeric
SibSp	# of siblings / spouses aboard the Titanic	Numeric
Parch	# of parents / children aboard the Titanic	Numeric
Fare	Passenger fare	Numeric
Cabin	Cabin number	Nominal
Embarked	Port of Embarkation	Nominal

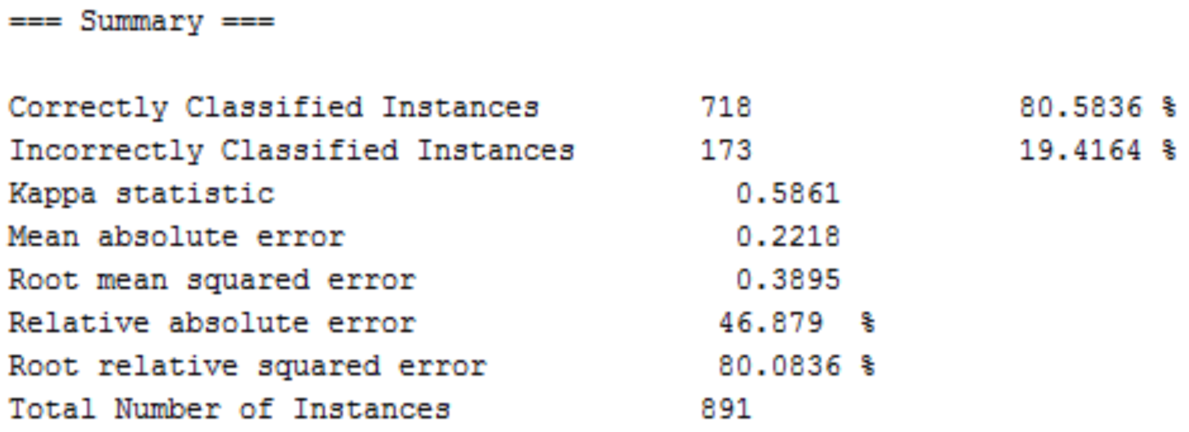
Data cleaning is critically essential because there are attributes that are neither necessarily be taken into consideration nor put in the data set to be mined. For example, the title of a passenger, which we believe has little to do with the survival information. Another example is the attributes cabin and embarked, which barely has influence over survival data. That is because these two attributes do not impact the survival data since people are running for survival during the event, it does not matter which cabin you live in and embarked you get board on.

4 Data Mining

After cleaning the data, we make advantage of Weka to do the data mining. First, we pay attention to the relationship between death/survival and many attributes. Second, there are many pre-build-in algorithms that we can rely on to do the research. The following are the details about the result of algorithms that are applied to the data.



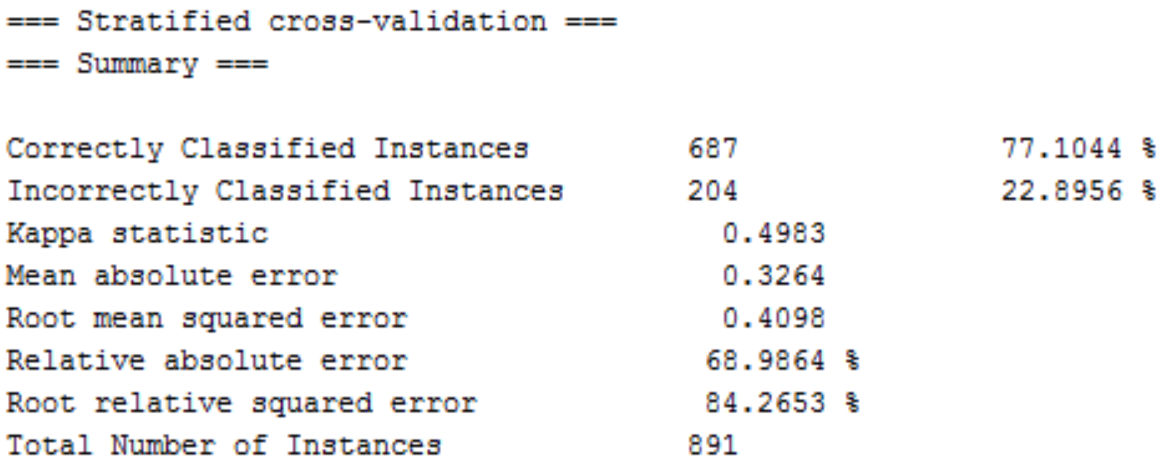
Naïve Bayes



```
=== Summary ===

Correctly Classified Instances         718               80.5836 %
Incorrectly Classified Instances       173               19.4164 %
Kappa statistic                          0.5861
Mean absolute error                      0.2218
Root mean squared error                  0.3895
Relative absolute error                 46.879  %
Root relative squared error             80.0836 %
Total Number of Instances              891
```

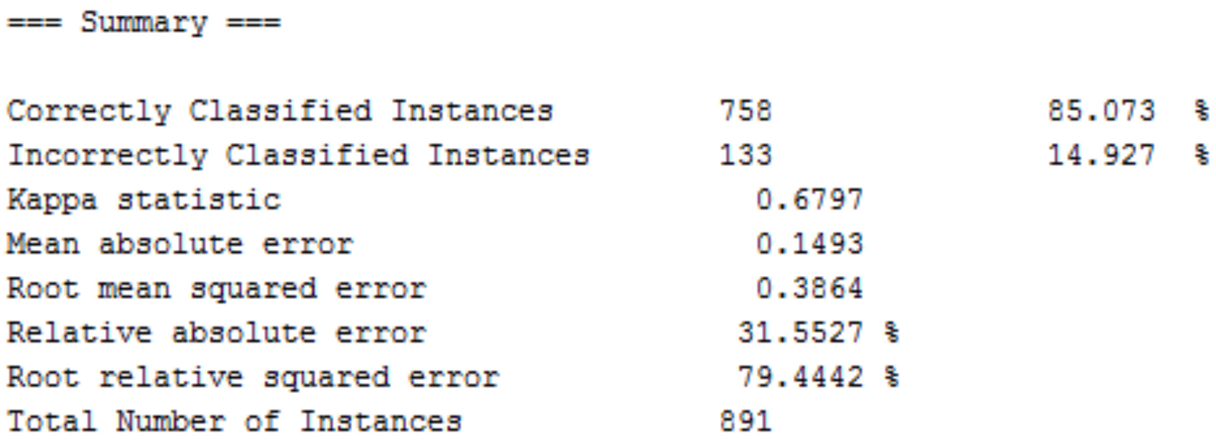
J48



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         687               77.1044 %
Incorrectly Classified Instances       204               22.8956 %
Kappa statistic                          0.4983
Mean absolute error                      0.3264
Root mean squared error                  0.4098
Relative absolute error                 68.9864 %
Root relative squared error             84.2653 %
Total Number of Instances              891
```

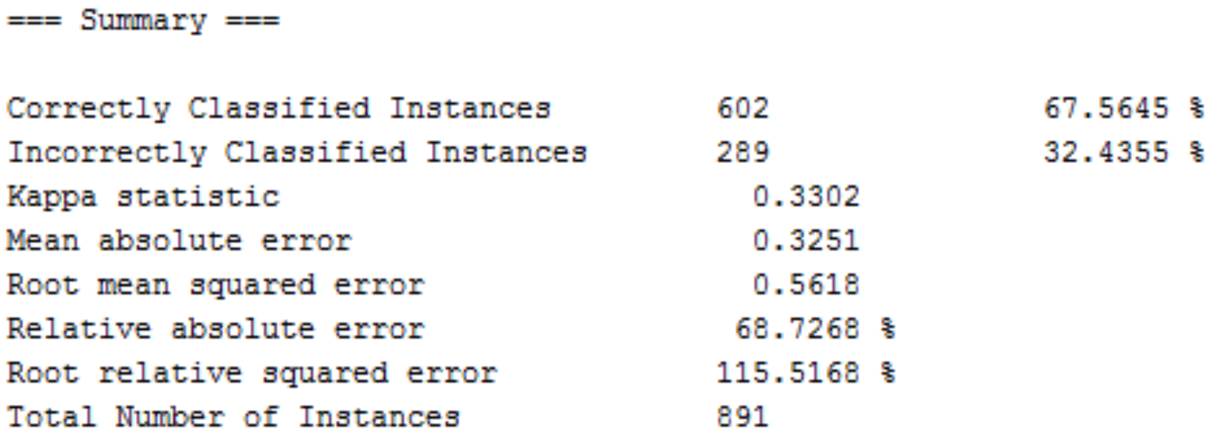
SMO



```
=== Summary ===

Correctly Classified Instances         758               85.073  %
Incorrectly Classified Instances       133               14.927  %
Kappa statistic                          0.6797
Mean absolute error                      0.1493
Root mean squared error                  0.3864
Relative absolute error                 31.5527 %
Root relative squared error             79.4442 %
Total Number of Instances              891
```

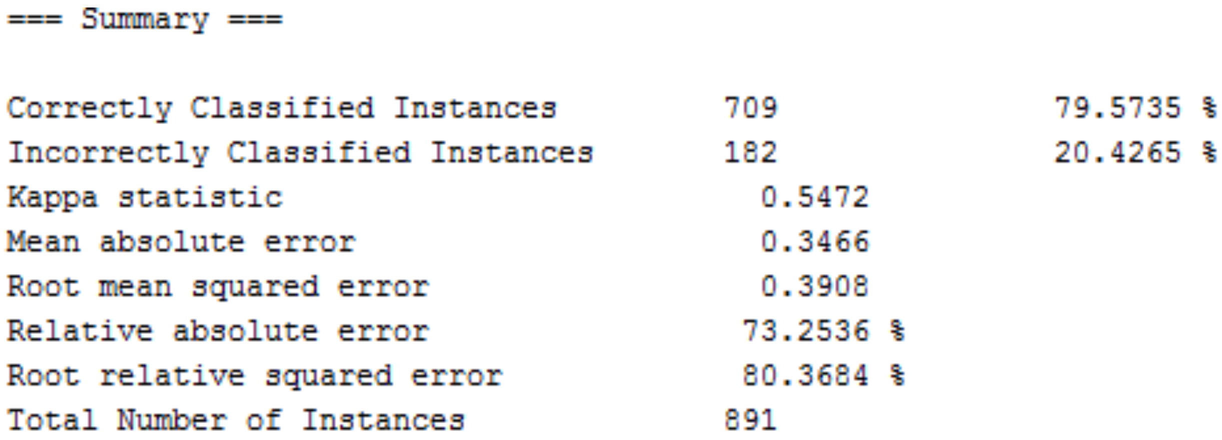
Logistic



```
=== Summary ===

Correctly Classified Instances         602               67.5645 %
Incorrectly Classified Instances       289               32.4355 %
Kappa statistic                          0.3302
Mean absolute error                      0.3251
Root mean squared error                  0.5618
Relative absolute error                 68.7268 %
Root relative squared error            115.5168 %
Total Number of Instances              891
```

RandomForest



```
=== Summary ===

Correctly Classified Instances         709               79.5735 %
Incorrectly Classified Instances       182               20.4265 %
Kappa statistic                          0.5472
Mean absolute error                      0.3466
Root mean squared error                  0.3908
Relative absolute error                 73.2536 %
Root relative squared error             80.3684 %
Total Number of Instances              891
```

5 Evaluation

We have two stage in evaluation. The first stage is to run the test file and the second stage is to upload to the Kaggle website to check out the result and rankings, which we can compare with other competitors.

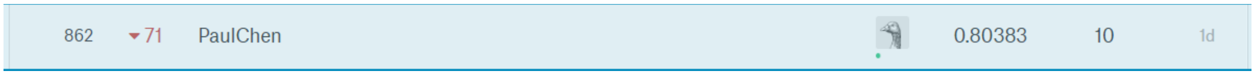
First stage: test file

we pick SMO algorithm as the best one as it gives the highest accuracy.

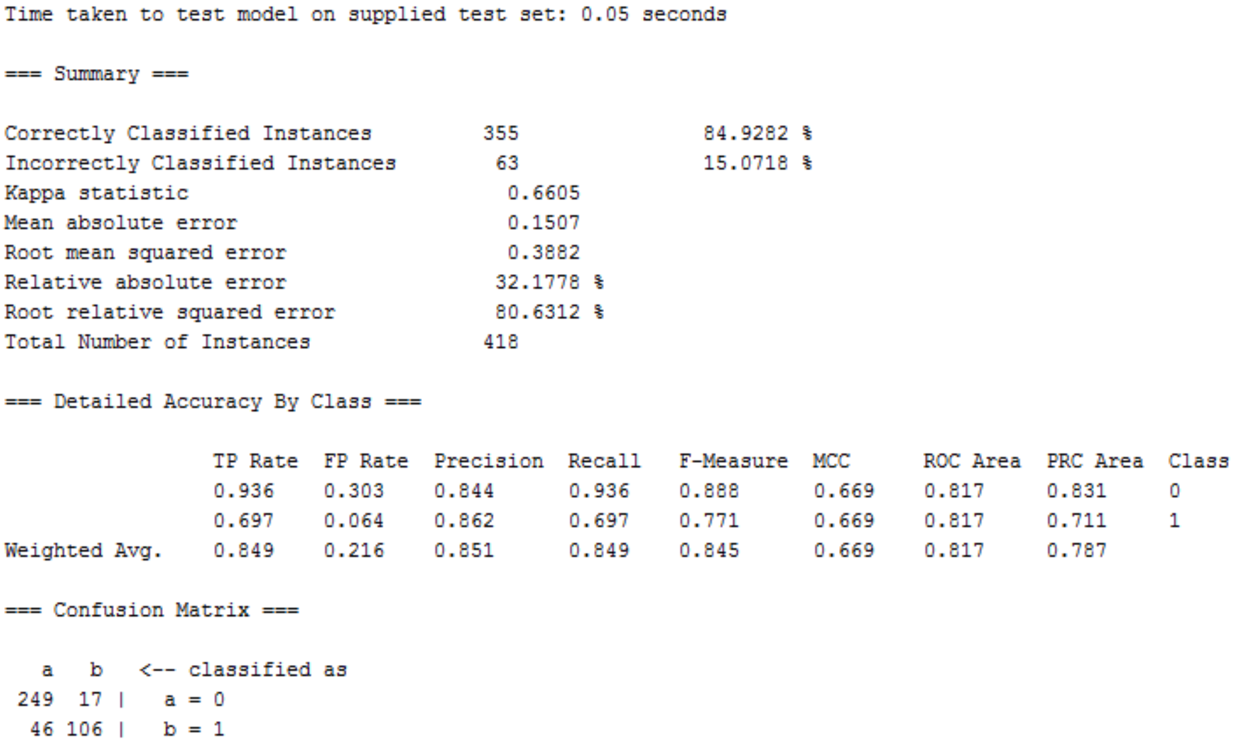
Second stage: Kaggle website

There are 6236 teams on this topic, therefore our score is in the top 14%(roughly 13.82%).

6 Conclusion



862 ▾71 PaulChen 0.80383 10 1d



```
Time taken to test model on supplied test set: 0.05 seconds

=== Summary ===

Correctly Classified Instances         355               84.9282 %
Incorrectly Classified Instances        63               15.0718 %
Kappa statistic                          0.6605
Mean absolute error                      0.1507
Root mean squared error                  0.3882
Relative absolute error                 32.1778 %
Root relative squared error             80.6312 %
Total Number of Instances              418

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.936    0.303    0.844      0.936    0.888      0.669    0.817     0.831     0
                 0.697    0.064    0.862      0.697    0.771      0.669    0.817     0.711     1
Weighted Avg.    0.849    0.216    0.851      0.849    0.845      0.669    0.817     0.787

=== Confusion Matrix ===

   a   b   <-- classified as
 249  17 |   a = 0
  46 106 |   b = 1
```

As mentioned before, it is very important to be prepared to encounter unexpected events. One way is to purchase insurance for your traveling.

We mined the data from the CSV file provided in the Kaggle website and used excel to perform some cleaning and transformation procedures to obtain the most suitable data for our approach. Then, we make use of Weka to import the CSV files to train and test.

We tried out different classifiers over the data and finally we choose SVM (named SMO in Weka) given the acceptable accuracy over 80%. As a result, the SVM revealed some interesting findings such as TP rate, FP rate, F-Measure and ROC area.

Next, we use the test file to justify our findings. The test result is not far away from what have theoretically acquired. We assume that the reasons are SVM takes every attribute into account to compute the result. This method explores all the possible relations between all the attributes.

Finally, we upload to Kaggle to compete with others to check our strategy. The outcome is exciting as we in the top 14%. That is to say, we figure out a competitive method to give a prediction with at least 80% accuracy rate.

Acknowledgments

This report is the project of 2017 Data Ming course SENG474, We thank Dr. Alex Thomo and Lab instructor Yudi Santoso for the guidance and assistance during the course.

References

[1] “Titanic: Machine Learning from Disaster.” [Online]. Available: https://www.kaggle.com/c/titanic

[2] Machine Learning Group at the University of Waikato, “WEKA: Waikato Environment for Knowledge Analysis.” [Online]. Available: http://www.cs.waikato.ac.nz/ml/index.html