Titanic – Machine Learning From Disaster

Gowrishankar Vindhiyavarman
Anand Sampathkumar

Warsaw University of Technology

Abstract

In this project, we see how we can use machine-learning techniques to predict survivors of the Titanic. With a dataset of 891 individuals containing features like sex, age, and class, we attempt to predict the survivors of a small test group of 418. In particular, compare different machine learning techniques like Decision Tree, SVM, Neural Networks and Random Forest analysis.

1. Introduction

Using data provided by www.kaggle.com, our goal is to apply machine-learning techniques to successfully predict which passengers survived the sinking of the Titanic. Features like ticket price, age, sex, and class will be used to make the predictions.

I take several approaches to this problem in order to compare and contrast the different machine learning techniques. By looking at the results of each technique we can make some insights about the problem. The methods used in the project include Simple Class Model, Random Forest, SVM, Neural Networks and decision tree. Using these methods, we try to predict the survival of passengers using different combinations of features. The challenge boils down to a classification problem given a set of features. One way to make predictions would be to use Decision Tree. Another would be to use SVM to map the features to a higher dimensional space. Once this is complete, we use random forest analysis on our data to see if we can achieve better results. Lastly we use Neural Network analysis.

2.Data Sets

The data we used for our project was provided on the Kaggle website. We were given 891 passenger samples for our training set and their associated labels of whether or not the passenger survived. For each passenger, we were given his/her passenger class, name, sex, age, number of siblings/spouses aboard, number of parents/children aboard, ticket number, fare, cabin embarked, and port of embarkation. For the test data, we had 418 samples in the same format. The dataset is not complete, meaning that for several samples, one or many of fields were not available and marked empty (especially in the latter fields – age, fare, cabin, and port). However, all sample points contained at least information about gender and passenger class.

3. Data preparation

In order to prepare our data for training in our classifier, we have to take a simple look at the data set. At first, let us start with a principle we all know: save children and women first in the disaster. So let us take a look at the Sex and Age variables to see if any patterns are evident. We'll start with the gender of the passengers.

	0	1
Female	0.258	0.742
Male	0.811	0.189

We can see that the majority of females abroad survived, and a very low percentage of males did. Now, we can dig into the age variables:

Min	1st Qu	Median	Mean	3rd Qu	Max	NA’s
0.42	20.12	28.00	29.70	38.00	80.00	177

Table 1: The age distribution of passengers in training data

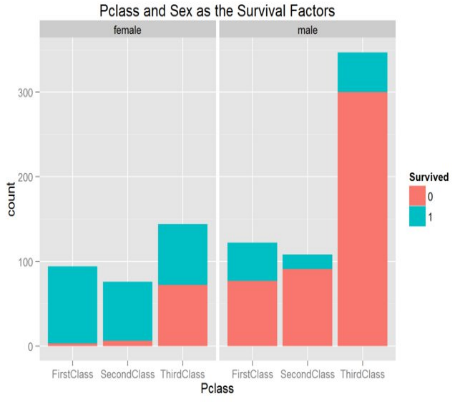
With this table, we can define the passengers who under 18 years old are belong to “children”, otherwise belongs to adults. We can get a form about age and sex pattern:

	Child	Sex	Survived
1	No	Female	0.7529
2	Yes	Female	0.6909
3	No	Male	0.1657
4	Yes	Male	0.3965

Table 2: The Age and Sex distribution of passengers in training data

It seems that if the passenger is female most survive, and if they were male most don’t, regardless of whether they were children or not. Now, let’s look at a couple of other potentially interesting variables to see if we can find anything more the class they were riding in, and what they paid for their tickets.

Fig 1 Gender Class Model



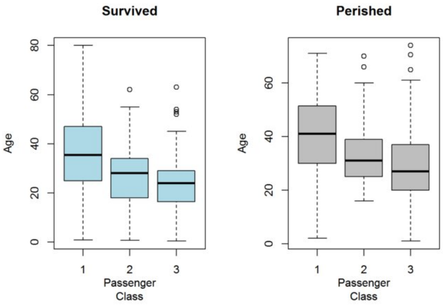


Fig 2 The class and Age distribution of passengers in training data

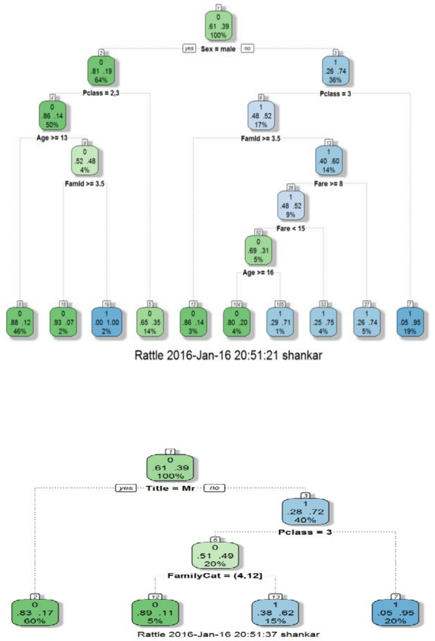
Interestingly, we can simply refer some new hypotheses: people in the first class have more chances survived than the lower classes. People who paid more will get more chances of surviving and females in the upper class have higher rate of survival.

4. Modeling

Decision Tree:

We built our decision starting with gender model. We split all the train dataset into male and female. Because it was most correlated with the chance of survival. After that, I use the function of Rpart in R library, which will generate custom features automatically of dataset and output a decision tree model. Next, we look at the feature age. Since the domain of age is continuous, we have to find a good decision boundary to split our data. After plotting the age and survival of passengers in each gender and passenger class, we decided to use a binary decision because in most cases, older passengers were more likely to die than younger ones. Instead of using the same age boundary for each gender and passenger class, we considered each gender and passenger class, case by case and found different boundary thresholds for each. To find our boundary threshold, we tried to minimize the classification error on our training set. This means that we chose the age boundary for each gender and passenger class such that if we classify all samples below the age boundary as survived and all above as died, we minimize the classification error on the training set. After including age in the decision tree, we achieve accuracy of 84.82%

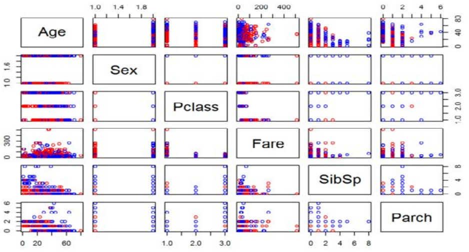
Fig 3 Decision Trees



With Feature Engineering we have categorized Title, Class and Family Category in to consideration and the accuracy of this model is 85.86% ,which is greater than the previous result.

SVM

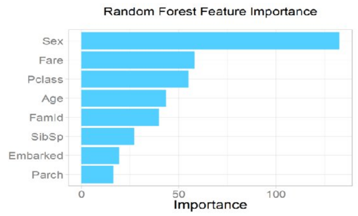
To improve our classification, we used support vector machines [2]. We considered the following features: 1) passenger class, 2) sex, 3) age, 4) number of siblings, 5) patriarchal status, 6) fare.



Using passenger class, sex, age, number of siblings, patriarchal status, fare. resulted in the accuracy of 84.82%. With Feature Engineering we have Title, Class , ChildorWoman and Family Category in to consideration and the accuracy of this model is 84.82%. which is same as the previous result..The training data has reached its asymptotic value of 84.82% and any additional sample does not improve the accuracy.

Random Forest:

Previously, we found that decision tree has over-fitting sometimes when dealing with terrible parameters. But if we grow a whole lot of them and have them vote on the outcome, we can get passed this limitation. That’s why we use random forests. This suggests that perhaps class and sex are strong indicators of survival whereas age and fare are weaker indicators of survival decisions based on different variables. So let’s imagine a female passenger from Southampton who rode in first class. Tree one and two would vote that she survived, but tree three votes that she perishes. If we take a vote, it’s 2 to 1 in favor of her survival, so we would classify this passenger as a survivor. Random Forest models grow trees much deeper than the decision stumps above, in fact the default behavior is to grow each tree out as far as possible, But since the formulas for building a single decision tree are the same every time, some source of randomness is required to make these trees different from one another. Through these sources of randomness, the ensemble contains a collection of totally unique trees which all make their classifications



As with our simple example, each tree is called to make a classification for a given passenger, the votes are tallied (with perhaps many hundreds, or thousands of trees) and the majority decision is chosen. Since each tree is grown out fully, they each over-fit, but in different ways. Thus the mistakes one makes will be averaged out over them all.

One of the toughest things is replacing missing value

We can use a Random Forest to fill in those values instead. We should pick up where we left off last lesson, and take a look at the combined data frame’s age variable to see what we’re up against:

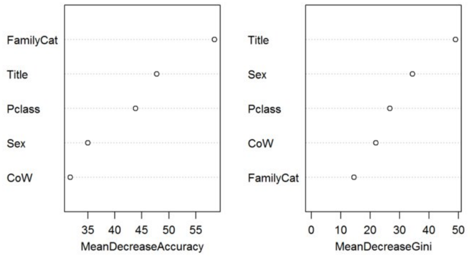
```
->summary(combi$Age)
Min. 1st Qu. Median Mean 3rd Qu. Max.        NA's

0.17 21.00 28.00    29.88     39.00     80.00        263
```

After this, all missing values are predicted by decision tree model we use in the first section. At this time, we can use the random forest tools to solve the problem.

```
fit2_rf<-        randomForest(as.factor(Survived)        ~
Pclass+Sex+CoW+FamilyCat+Title, data=dataTrainfe,
importance=TRUE, ntree=2000)
```

The result is as following.

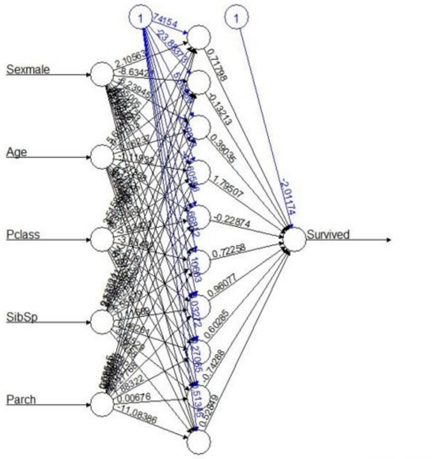


There’s two types of importance measures shown above. The accuracy one tests to see how worse the model performs without each variable, so a high decrease in accuracy would be expected for very predictive variables. The Gini one digs into the mathematics behind decision trees, but essentially measures how pure the nodes are at the end of the tree. Again it tests to see the result if each variable is taken out and a high score means the variable was important. Unsurprisingly, our Title variable was at the top for both measures. We should be pretty happy to see that the remaining engineered variables are doing quite nicely too. Finally, this model get 86.39% accuracy.

Neural Networks

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. They are used to estimate or approximate functions that can depend on large number of inputs and are generally unknown. They can compute values from inputs by changing weightages accordingly, for desired output. A simple neural network consists of Input nodes, Hidden layers and Output node. These nodes are connected by respective weightages.

Data cleaning can be done initially removing variables not used for the model, at this point we can remove Passenger Id, Ticket, Fare, Cabin, Embarked attributes. Qualitative variables like Male / Female are converted into 0 and 1 (quantitative variables). Next inference is made on missing values of age entries. We can infer them based on relation between name and age attributes, i.e. using prefixes in name to infer the age. For this the name variable is changed to corresponding shortcut or prefix and so grouped together. Next we may use the siblings and parents abroad data and create new variables like Child – if passenger is below age of 12, Family – Family size of passengers aboard, Mother – If the passenger is mother (prefix being Mrs. and kids greater than 0). These variables are created indicating that there might be a close relation between these variables and survival rate. Creating such variables can vary the performance of the model which can be checked during model fitting. Test data will also undergo the same process of data cleaning except it misses the survived column, which needs to be predicted by the model. Now the train data is ready to be modelled. The accuracy for this model is 82.72%



6. Conclusion

After implementing four methods, we got some conclusions: Of all the four methods, ANN performs worst with 83.77% but Random Forest performs best with 86.39% accuracy. So only the difference is nearly 3 percent. This is probably because there was one feature that was strongly correlated with whether a passenger survives. Decision Tree and Neural Networks only combine features but didn't give them specific weights and correlations. So this shows that assuming that features are independent is not necessarily a bad assumption for our problem. Table 4 offers a summary of the achievable accuracy using Decision Tree, SVM, Neural Networks and Random Forest analysis.

Table 4 Comparison of Performance with Feature Engineering

Decision Trees	0.8486387
Random Forest	0.8638743
C Forest	0.8586387
SVM	0.8481675
Neural Networks	0.8376963

As for the future work, I suggest that I can pay attention to find the internal correlations between other attributes and optimize the parameter of Random Forest.

In terms of the question that who will survive from the disaster, there are some inferences:

1. Woman and Children have the best chance of surviving because human always have to protect the vulnerable groups first. Most adult males have to survive by their own faith ,physical conditions or luck.
2. Wealth and upper –class people may be more possible to get survived mainly because they paid expensive tickets, which contributes to the result that they might live closer to safeboat than people who live under the deck.
3. People who are despicable and shameless may survive from disaster partially because they can obey their core values in order to get a seat in safe boat even those who comes from upper class and well-educated family.

   I believe there are more interesting conclusions to be found when I go further and collect more data about sinking ship events.