Using Random Forest to Predict Survivals in the Titanic Disaster

* Part 1: Problem Description. Give the name of the competition you selected and write a few sentences describing the competition problem as you interpreted it. You want your writeup to be self-contained so your peer-reviewer does not need to go to Kaggle to study the competition description. Clarity is more important than detail. What's the overall goal? What does the data look like? How will the results be evaluated?

The problem I choose to work on is the Titanic Competition. This is a supervised learning problem: I am required to predict if one given passenger in the dataset survived or not (label) in the disaster by selecting and analyzing various attributes including sex, age, cabin, social class and so on.

One training dataset (named "train") and one testing dataset (named "test") are given, however since the test dataset is only for testing the submitted project, I only used the training dataset. The dataset has 12 variables (passenger id, social class, name, sex, age, number of siblings/spouses aboard, number of parents/children aboard, ticket number, passenger fare, cabin, port of embarkation, and survival). Among those variables, I initiated with 7 of them as features (excluding passenger id, name, ticket number, cabin and survival), because those excluded variables are not useful for prediction. There are 891 records in the dataset.

The test results (test set) can be evaluated by calculating its accuracy. Accuracy = “the number of correctly predicted labels (survival)” / “the number of total labels (survival)”. The higher the accuracy is, the better results I got.

* Part 2: Analysis Approach. Write a few sentences describing how you approached the problem. What techniques did you use? Clarity is more important than technical depth in this exercise.

I randomly split the dataset into two subsets (70% for training and 30% for testing) in order to validate the model based on the given data. I firstly trained a random forest classifier with 100 trees, and the accuracy was 82.3%. Then I consider maybe the number of trees will affect the result, so I trained several classifiers with different number of trees.

* HINT: Favor simple approaches over complex approaches. The idea is to get a good, simple, comprehensible solution rather than try to impress people with how smart you are. Start small, and improve your solution incrementally -- don't try to save the world, boil the ocean, or any other hyperbolic metaphors.
* Example: "I split the data by gender and handled each class separately. For the females, I trivially classified all of them as "survived." For the males, I trained a random forest as a classifier. I ignored the pclass atribute that indicated the location of the passenger's cabin because I didn't think it was relevant."
* Part 3: Initial Solution. Write a few sentences describing how you implemented your approach. Think of it as a whiteboard conversation or a descriptive forum post rather than a full technical report. Try to provide enough detail for someone with some experience to follow your recipe and reproduce your results: Describe how you prepared the data, the method(s) you applied, and any tools you used (not detailed code). What languages and libraries did you use? What challenges did you run into?

I used the “caret” package in R (introduced in the lecture) to split the data into two sets (training set and testing set), and used the “randomForest” function in the “caret” package to train the random forest classifiers. Then, I got the predicted values by feeding “predict” function with the classifier I built and the testing dataset. Finally I calculated the accuracy by using the formula introduced in Part 1.

* Example: "I partitioned the data by gender manually using Excel. I used Weka to build the random forest."
* Part 4: Initial Solution Analysis. Write a few sentences assessing your approach. Did it work? What do you think the problems were?

My results work OK, but definitely have much space to improve. The initial average accuracy I got was 75.07%. I suspect there might be an optimal number for the number of trees.

* Example: "My approach did not work so well, achieving a score of 0.65. This is less than the sample solution. I suspect I should not have ignored the pclass attribute."
* Part 5: Revised Solution and Analysis. Write a few sentences describing how you improved on your solution, and whether or not it worked.
* Example: "I included the pclass attribute and ignored the ticket number attribute. My score improved to 0.68."

I tested on different number of trees (from 1 to 100) and got 100 results (accuracy) by using some loop function. Then I plot them (100 data points, x-axis=”number of trees”, y-axis=”accuracy”) out in order to inspect some patterns, but unfortunately the plot shows no obvious effects generated by the number of trees.