Peer-graded Assignment: Kaggle Competition Peer Review

Data Science on the Titanic Passenger Survival Kaggle Competition

by Nathaniel Evans

June 3, 2017

like Flag this submission

PROMPT

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?

RUBRIC

Is the problem description clear and comprehensible?

- 0 pts
  The submission makes no real attempt to describe the problem.
- 2 pts
  The submission makes an attempt to describe the problem, I don't really understand what is being described.

Example: "The task is to predict whether a given passenger survived the sinking of the Titanic based on various attributes including age, location of the passenger's cabin on the ship, family members, the fare they paid, and other information. Solutions are evaluated by comparing the percentage of correct answers on a test dataset."

I worked on the Titanic: Machine Learning Through Disaster competition, in order to follow a competition that was in close correspondence to materials from the course. Using information obtained from passengers aboard the Titanic, the competition asks us to to predict whether or not a given passenger would survive. The data is organized into attributes for each passenger, including the ticket class, sex, age, number of siblings and parents aboard, fare price, cabin number, port of embarkation, and ticket number. The training set includes "ground truth" labels for whether or not a given passenger survived. The objective is to create a model that can looks at a separated test data set and assigns a binary prediction for each passenger (survived, not survived). The evaluation is taken to be the percentage of passengers you correctly predict on the test set (accuracy).

- (•) 5 pts
  The problem description is clear enough; I understand what is going on.

PROMPT

Part 2: Analysis Approach. Write a few sentences describing how you approached the problem. What techniques did you use?

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."

To begin, I looked carefully over all of the Kaggle descriptions on the dataset, and downloaded the data. I loaded the training and test data, and I began by poking around the data to better understand what I was looking at. I tried to understand what ranges values took for different attributes, what the data types were (e.g. categorical, continuous), and assessed how "clean" the data were (e.g. missing data, formatting, etc).

I wrote down a list of all attributes I could use as predictors, and I created a few simple confusion matrices (counts of survival/non-survival for each attribute). I also created scatter plots for individual

RUBRIC

Is the approach to the problem described clearly? Do you have some idea how you might employ these techniques to solve the problem?

- 0 pts
  No real attempt was made to describe the approach to the problem.
- 2 pts
  There's a description, but I don't fully understand what it's saying.
- (•) 5 pts
  The description is clear enough; I understand the approach.

attributes (separating survival/non-survival passengers by color). After this initial exploratory analysis, I had a vague idea of which attributes might be the most meaningful to begin my model creation. For those attributes which weren't immediately obvious to me how I might use them (e.g. sibling count, parent count, name), I tried to imagine how I might transform or combine attributes into new features for my model.

PROMPT

Part 3: Initial Solution. Write a few sentences describing how you implemented your approach. What languages and libraries did you use? What challenges did you run into?

Example: "I partitioned the data by gender manually using Excel. I used Weka to build the random forest."

Using R, I probed I first probed into the data to have a better understanding of the content (as described above). For my exploratory analysis, I used the ggplot2 library with an additional shortcut library for visualization (scales), plotting several variables against one another as a function of survival. Using the subset() command in R, I reduced the training set, eliminating

RUBRIC

Is the initial solution implementation clearly described? Are the tools, languages, and libraries used reasonable?

- ○ 0 pts
  No significant attempt was made to answer the question.
- ○ 2 pts
  The implementation is described, but I don't fully understand how it works.
- ◉ 5 pts
  The implementation is well-described; I understand how it works.

"PassengerId", "Ticket" (number), and "Cabin" attributes from my model. I did not find these fields to contain information relevant to predicting survival. Next, I looked for missing values and thought of schemes to interpolate missing information. For my first attempt, I replaced all missing continuous values with the mean of the remaining values of the attribute (applied to "Fare" and "Age". I then created categorical factors out of the Sex, and Port of Embarkment variables using R's as.factor().

Before delving into deeper machine learning techniques, I wanted to first use a simple logistical regression model. Thus, using the training data and R's function for a general linear model glm() with a binomial logit function, I created a model. I analyzed the coefficients of this model and tried to understand what it learned from the data before applying it to the test set for evaluation. Using the summary(model) from R, I saw that the model learned that "Price class", "Sex", "Age" and "Number of Siblings" were each significant predictors of survival. In particular, the lowest p-value for a coefficient in the model was for "Sex", and the negativity of its coefficient (considering the factorization of the Male/Female split), indicated to me that being a female was

advantageous to survival. I also looked at the reduction in residual deviance explained by the addition of each variable into the model using R's anova() function on the model. This showed me that the addition of Sex, PClass, and Age led to significant reductions in model variance, but the remaining variables did not contribute much.

Now that I understood the model that was created on the training data, I sought out to test its predictions on the test set. To do so, I used R's function predict with the type='response', R will use the trained model to generate a probability of survival on the test data (between 0 and 1). Then, I created a simple classifier which split the outcome probabilities in two: if it is below 0.5, I classify the passenger as not surviving, and above 0.5 as surviving. Using this binary class, I assigned the labels to a result vector and uploaded my result to Kaggle, where I found out that my classifier obtained an accuracy of 0.77 on the test set.

Help Center

PROMPT

Part 4: Initial Solution Analysis. Write a few sentences assessing your approach. Did it

RUBRIC

Is the initial solution analysis comprehensible and reasonable?

work? What do you think the problems were?

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."

I believe that my first approach was elegant in its simplicity. I successfully familiarized myself with R (before I had only used python and Matlab), I was able to apply a familiar model (binomial logistic regression) to the data, and I created a prediction that resulted in a decent accuracy. Because I wanted to really understand what I was looking at, I approached the problem in a simple manner, and this paid off for my learning.

The only challenge for me was to think of how to treat missing values appropriately. My choice of replacing missing values by the mean value is questionable, and I would prefer to think of interpolating results with a little more finesse. I have read about statistical imputation methods that can do this, and perhaps I'll improve on this for my revised solution. I also would like to use more sophisticated non-linear prediction models, rather than the simple logistic function on straightforward features that I used.

- ○ 0 pts
  No significant attempt was made to assess the approach.
- ○ 2 pts
  There's an assessment, but I don't really understand what is being described.
- ◉ 5 pts
  The assessment is adequately described; I understand what's going on.

PROMPT

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 began by adding an additional feature to the classifier, namely the size of the family of a passenger. This information could be computed as the sum of the siblings and parents aboard, and allowed me to reduce dimensionality of the feature vector by removing the individual Sibling and Parent count attributes. Next, I improved the complexity of my model and sophistication of the analysis, I elected to use the mice R library to perform multivariate data imputation for missing values in Age and Fare, rather than replacing NA values by the mean. Once I assembled the new feature vector with sophisticated missing value replacement, I was ready to train a model.

This time, I wished to use the random forest classifier algorithm (R library randomForest()). More

RUBRIC

Is the revised solution and analysis presented clearly?

- ○ 0 pts
  No significant attempt was made to describe an improved solution.
- ○ 2 pts
  There's a description, but I don't fully understand it.
- ◉ 5 pts
  The improvement is adequately described; I understand it.

specifically, I trained a random forest to predict the Survived label as a function of PClass, Sex, Age, Fare, Embarkment place, and Family Size (new feature). After the model was trained, again I tried to analyze its predictions on the training set. Specifically, I looked at the model error as a function of the number of trees in the forest. I also looked at the variable importance (Gini coefficient) of the forest. This analysis showed me that, as with my first approach, Sex was the most important variable, followed by Fare, Age, Pclass, FSize, and Embarked. Finally, I applied my random forest model to the test set and created predictions for the Survived label using the R function predict(). I submitted my new attempt to Kaggle and found that my new analysis resulted in an accuracy of 0.78. It looks like my usage of more sophisticated methods, coupled with more work massaging the feature vectors did indeed increase performance, but only marginally (only 1 percent)!

OVERALL ASSIGNMENT RUBRIC

Overall evaluation-Given the description provided, do you feel you would be able to reproduce the result?

- 0 pts

No, the description is too sketchy, incomplete, or unclear for me to be able to reproduce this result.

- ○ 2 pts
  No, but primarily because I don't have a strong enough background in the area; there is a fair amount of jargon used.
- ○ 4 pts
  Yes, I think I could reproduce the result with some effort.
- ◉ 5 pts
  Yes, the solution is clearly and thoroughly described -- I could follow them easily.

Evaluating the analyses of others' is a great way to pick up new insights and viewpoints. What is one helpful thing you got out of reviewing this submission? (e.g. maybe ideas on analyzing solutions, incentive to go try this competition, etc.)?

Ideas on analyzing solutions.

(This question will not affect the submitter's score.) You're not expected to actually try and reproduce the submitter's result, but you may choose to do so. Did you try to reproduce the result?

- ◉ No
- ○ Yes, but my result was different
- ○ Yes, and my result was the same.

Submit Review

Comments

Comments left for the learner are visible only to that learner and the person who left the comment.

share your thoughts...