Statistical Machine Learning

Prediction Assignment: Regression

Oct. 29th, 2022

Due

Wednesday, Nov. 30th, 11pm

Email me your final report (results): ab.safari.w@gmail.com

Policy

1. This project is to be completed *independently*, with no outside help. You may use whatever class materials you wish in completing this assignment. **BUT DO NOT DISCUSS QUESTIONS OR RESULTS WITH ANYONE ELSE, WITHIN OR OUTSIDE OF THE CLASS**. Failure to follow this directive will result in a failing grade.
2. Late projects will be accepted at a penalty of 5 points/hour (it's a 100-point project), *strictly enforced*.

Assignment

The data for this project are available on LMS. This file is a comma-separated file (csv) that contains a timeseries with the length of *~4700* observations (*1* explanatory variables (DATE) and one response (AMOUNT)). This dataset is obtained from a large international company (a real dataset - not allowed to reveal its name), which shows its sale amount at unequal time intervals since 2004 until end of 2020. This is a regression task and you need to predict the company's sale amount in the first half of 2021 (first 6 months). I will compare your predictions with the observed sales of the company during that sis-month period (I have the data until end of 2021!).

Deliverables

You will produce two required items and one optional item.

1. The test set is the first six months of 2021. I will leave to you to decide how you breakdown this six-month period and make prediction for each time unit. For instance, if

you breakdown this period to months, you need to have sale prediction for the following dates:

i. Jan. 31st, 2021
ii. Feb. 28th, 2021
iii. Mar. 31st, 2021
iv. May 30th, 2021
v. Jun. 31st, 2021
vi. Jul. 31st, 2021

Another option is to have one single prediction for the entire 6 months. The objective is to have as many accurate predictions as possible. That is, I will give the highest mark to the one who can have more predictions (shorter time intervals, e.g., weeks) with high accuracy. In other words, my evaluation scheme is to consider both number of predictions as well as their accuracy simultaneously. I will come up with a formula for this.

2. You will return a list of sales prediction with their corresponding dates. The list should be *two columns labelled (date & amount – same structure as the training set) with no row numbers and no column header*.

Look at your file before you submit it to make sure that the format is correct (and also to make sure that the predicted values are sensible!).

3. You will supply a written report of the steps you took to create your model and predictions. Details are given below.

Report

Your report should answer these questions, as numbered below:

1. *What models or machines did you attempt to fit?* For each one, paste the code (R or python) from your program for the initial successful model fit. I want to see what you tried. For example, "fit1 = lm(y~., data=train)" is what I would list if I used multiple linear regression on all of the variables and my training data were called "train". **The answer should be a list (e.g., with bullets) of nothing but the code for each of these model fits.** Don't list code that did not run. If tuning was involved in the initial fitting process, you can paste the function with variable names for the tuning parameters (e.g., your function might have "mtry=mm" if you looped over a variable called "mm").
2. *What process(es) did you use to evaluate and compare models and to select your final model?* I am thinking of Lecture 4, specifically: **Give 1-2 sentences explaining the method, the quantity, how results were turned into decisions.** For example, "I used

50,000 bootstrap resamples, fit all models to each resample, and used largest training error from last resample as my best model." (This example answer is complete, but represents something rather stupid to do...)

3. *Did you tune any methods?* If so, (a) what process(es) did you use to evaluate and compare models and to select your final model (i.e., **I want to see an answer like to the previous question, but relating to how TUNING was done**), and (b) **for each method list all parameter values that were considered** (e.g., "For "Blasting" I used a grid of values with A=(1,2,3,...,60) and B=(0.00317, sqrt(3.14159)). For "Blooming" I used combinations of (*z,*)=(0.1,3), (0.5,6), and (1.1, 12)).
4. *What was your chosen prediction machine?* **Paste the code that produced your predicted values, including all values of tuning parameters if any, random number seeds, and explaining any variable names that are not obvious.** I should be able to run your code and produce the same results (or extremely similar if randomization is used). If I try and it doesn't work, there will be a major deduction.

The main thing here is that I should be able to see what your thought process was and whether you considered (or failed to consider) important ideas.

Submission

You need to email me all the above materials by the project deadline.

Grades

Your grade will be based partly on how well your model performs and partly on the steps you took to get there. I will compute a form of agreement between your predicted values and the test set responses (e.g., accuracy). *I will scale these against the best model produced by a member of the class, so this is a competition!* If your accuracy is only 80% as large as the best, your mark for this part will be 80%.

Your report as described above. This portion will count for 30% of your grade and the remaining 70% will come from your model's performance.

Final Comments

GOOD LUCK, HAVE FUN, and remember: in real life an employer will take action based on the results you provide them. These may be million-dollar decisions which rely extensively on YOUR expertise. This is practice...