Group Project Guidelines

STATS 415

Winter 2022

This project provides an opportunity for you to apply the statistical learning techniques covered in class to analyzing a real dataset. You will work in groups to analyze data from a large national health survey conducted by the Center for Disease Control (CDC). The project has two parts. The first part is a prediction challenge. You will participate in an in-class Kaggle competition to test your ability to build powerful predictive models. The second part is more open-ended. You will pose and answer two questions about the CDC's dataset. You will submit a final report detailing your analysis as well as R (or R Markdown) files with any code you used for the project.

Due dates

- Form groups: Feb. 25, 2022.
- Proposal: Mar. 20, 2022, 8:59 pm. No more than one page, submitted by each group via Canvas.
- Final project reports: Apr. 17, 2022, 8:59 pm. No more than 10 pages including figures and tables, turned in by each group via Canvas as a pdf. You will also submit all code used for your project.

Team

The first step is to form a group—part of the point of this project is to learn to do data mining as a team. The teams should comprise three members (strongly preferred), but two or four is also acceptable. If you aren't sure who to work with, you may use this Google sheet to form a group. Once you have formed a group, your group must sign up as a group here on Canvas.

The data

For this project, you will be using data from the National Health and Nutrition Examination Survey (NHANES) to pose and answer your own questions about human health. You will download the data for the open-ended portion of the project from the CDC's website (your primary source of the data should be the section called "Continuous NHANES"). You can use the R package `haven` to read in the `.XPT` files. For the prediction challenge (Kaggle) part of the project, we will be providing `.csv` files containing the unique IDs of the subjects used for the data generation.

Project proposal

Each team submits one copy of their project proposal via Canvas. The proposal should be no longer than one page. The proposal must include the team members' names, the questions you intend to answer with the data, and the methods you intend to use to answer the questions. You should include a rough timeline for project completion and specify the proposed responsibilities of each team member. Finally, as part of your proposal your team must join the Kaggle competition, make one Kaggle submission (the accuracy of this submission does not matter), and report the name of the account who will be submitting entries for the competition.

Kaggle

To assess your ability to build predictive models, we will be holding a class competition on Kaggle. You will sign up for the Kaggle competition at this link (one entry per team). ***Your Kaggle account name should be included at the top of your report.*** More details on the Kaggle competition can be found in the appendix.

We will provide two files, `train.csv` and `test.csv`, for the prediction challenge. The training file will contain the unique IDs of the individuals in the training set as well as the responses corresponding to these individuals, while the test file will only contain the unique IDs for individuals in the testing set. You will have to upload your predictions to Kaggle to see your test R^2 (squared correlation between the predicted response and the true response). You will be limited to five Kaggle submissions per day.

Final report

You will write up your results in a formal report. Unlike homework, the report should NOT include R code or raw outputs from R (e.g., you should not copy and paste the output from `summary(lm(y~x))` in your report). R outputs should be included in the form of tables and figures; models should be written as equations, e.g., $y = \beta_0 + \beta_1 x$. Each report must have a title, a summary/conclusions section, and a paragraph describing the individual contributions of each of the team members. You should dedicate approximately one page of your report to explaining how you approached the Kaggle challenge (models you tried, how you did hyperparameter tuning, how you chose the final model, etc.). Each team will submit one pdf copy of the report through Canvas. In addition to the pdf report, each team should submit `.R` and `.Rmd` files containing all of the code required to reproduce the analyses in the written report. The report should include an appendix (not counted towards the 10 page limit) which provides sufficient instructions for us to run your code to reproduce the analysis.

Grading rubric

Project proposal: 10%

Graded on following instructions (including all the information requested) and questions and methods making sense for the data.

Kaggle: 20%

If your submission achieves a test set R^2 of 0.30 or higher, your team will receive 5 points. The remaining 15 points will be awarded based on your rank among the class. If your test set R^2 is at least 0.70 you will receive the full 20 points, regardless of your class rank.

Project report: 70%

Grade consists of

- Questions posed (make sense for the data, can be answered with data at hand, are something people might care about in real life): 5%. You must pose **two** questions (one classification focused and one regression focused) and answer them in your report.
- Methods chosen for analysis: 15%. A minimum of **three** separate methods must be compared. These methods must match the questions and be appropriate for the data set (explain why).
- Analysis: 20%. Correctness and completeness (i.e., exploratory data analysis included, at least three methods compared across the two questions, tuning parameters selected appropriately, and appropriate plots included).
- Conclusions and interpretation: 15%. Answer the questions you posed and interpret your results.
- Writing: 10%. Having proper introduction, summary, sufficient data description, clear and informative figures and tables, using appropriate level of formality for a report. Your target audience is someone who knows about all the methods we studied, so you do not have to introduce them, but who does not know R. *No R code should be included in the written report. No raw/unformatted R output should be included in the written report.*
- Reproducibility: 5%. Your submission should include all code necessary to reproduce your analysis, and the written appendix should contain sufficient instructions for how to run the code.

Appendix: Kaggle details

Sign-up instructions

You can sign up for our class Kaggle competition at this link. There should be one account signed up for the competition per team.

Data description

The train.csv file contains two columns: SEQN and y. The SEQN column contains the unique IDs of the individuals in the training set and the y column contains the corresponding value of the repsonse variable for each individual. The test.csv file only contains the unique IDs (SEQN) and does not contain the response variable. To build the training and test data sets, you should download the following files for the years 2009, 2011, 2013, 2015, 2017 (under the section "Continuous NHANES").

- Demographic Variables and Sample Weights (DEMO)
- Body Measures (BMX)
- Blood Pressure (BPX)
- Smoking - Cigarette Use (SMQ)
- Cholesterol - Total (TCHOL)
- Dietary Interview - Total Nutrient Intakes, First Day (DR1TOT)

Each of the files can be joined on the SEQN variable. You can then use the data set you've constructed along with the response variable provided in train.csv as the training set. You will use the same procedure for the test set except you will not have access to the outcome variable.

The responses (y) were simulated from a selection of some of the variables in these files.

Preprocessing

Once you have loaded and joined all the data frames, you should drop the following variables: ("SMDUPCA", "SMD100BR", "DR1DRSTZ", "DRABF", "RIDSTATR"). You should then standardize (center and scale) all variables before fitting any models.

Kaggle submission instructions

To have a submission scored, you must upload a .csv file to Kaggle with two columns: SEQN and y, where y is the predicted value for the corresponding value of SEQN. Kaggle will score your submission using a subset of the test data (corresponding to the public leaderboard), but the final score will be computed on all of the test data (corresponding to the private leaderboard).

Write-up instructions

In the write-up for the Kaggle portion of the competition, you should include a table with all of the methods you tried and the train/test error for each method. Additionally, you should discuss any kind of model tuning/hyperparameter selection and/or variable selection you did.