MSA 8190 Statistical Foundations

Mini-Project for Final Exam

From: Friday, December 1st, 2023 11:59AM (EDT)

To: Monday, December 4th, 2023 11:59AM (EDT)

NOTES: Please carefully read the following instructions and notes.

- Deadline to submit the solutions is Monday December 4, 2023 11:59 AM (EDT). No submission will be accepted after the deadline.
- This project is open book and open note. But, you are not allowed to get help from any source including the students in your class (except your teammates), other people, internet, etc. For more details about exam rules and the Honor Code, please review the course syllabus. Any cheating and violation of the honor code will result a grade of F for the course.
- You will need to do all computations in R and submit your R (or R Markdown) file. Team leaders need to submit the project report (R, R Markdown, pdf,...) and presentation files via iCollege, before the deadline, December 4th, 2023 11:59AM (EDT).
- On Monday, December 4th, during class hours, each group will present in 8 minutes followed by a short Q&A.

Project Description

Consider the "Bias correction of numerical prediction model temperature forecast Data Set" in `https://archive.ics.uci.edu/ml/datasets/Bias+correction+of+numerical+prediction+model+temperature+forecast`. Please read the data description in the link. You can also download its data as a csv file from the "Data Folder" in this link.

This data set includes 7750 observations and 25 attributes. The last two attributes (Next_Tmax and Next_Tmin) are the response variables that we want to predict. We will only build models to predict **Next_Tmax**

(a) Pre-processing: Check the data to see if there is any missing value. The simplest approach to deal with the missing values is to remove the rows with any missing values. Use are free to pick other methods from literature to deal with missing data. Remember that the final data provided to the regression models should not have any missing value. Also, make sure that the variables have the correct type. For example, "station" number doesn't really have a numerical meaning! Or the date should not be included as a predictor.

(b) Split the data into three parts. Based on the "date", consider the first 60% for training (**Train**), the next 20% for validation (**Valid**), and the last 20% for testing (**Test**). We will not touch the test set until the very end step.

(c) Initial Model: Fit a multiple linear regression model on the training data. Use this model to make prediction on the validation set. Then, calculate RMSE on the validation set using the following formula

$$\text{RMSE}_{\text{Valid}} = \sqrt{\frac{\sum_{i \in \textbf{Valid}} (y_i - \hat{y}_i)^2}{\text{size of } \textbf{Valid} \text{ set}}}$$

Explore and discuss this model, its quality, significance of variables, residuals, ...

(d) Improved Model: Using the model in the previous part as your base model, you may remove or transform some of the features or create and add some new feature. Try in a systematic way to improve your model and come up with you best model. You can use $\text{RMSE}_{\text{Valid}}$ to measure the models' performances and pick the one with the highest $\text{RMSE}_{\text{Valid}}$ as your best model.

Explore and discuss this model, its quality, significance of variables, residuals, ...

(e) Test your initial model (from part c) and improved model (from part d) on the test set. Report $\text{RMSE}_{\text{Test}}$ for each of these models.

$$\text{RMSE}_{\text{Test}} = \sqrt{\frac{\sum_{i \in \textbf{Test}} (y_i - \hat{y}_i)^2}{\text{size of } \textbf{Test} \text{ set}}}.$$