STAT 561 - Homework 4

Instructions

1. *Due Monday, April 28th at the 11:59pm. Any submission after that (24hrs after) will be graded out of 50%.*

2. *Your submission should include a* ***pdf*** *from your generated R markdown, and your R markdown file.*

3. *Make sure to highlight the part of the output that have the information you will be providing as answers.*

4. *This work should be done as a group. Only one person in the group should submit the work with the names of all the people in the group on the document*

Using Absenteesim data from HW 3

For the questions in 1-9 below,

A. **Phase 1:** Use multiple linear regression with the response variable being "Absenteeism in hours" variable.

B. **Phase 2:** Use logistic regression with the Absenteeism categorization done in Homework 3 as your response.

The features are: Month of absence, Day of the week, Seasons, Transportation expense, Distance from Residence to Work, Service time, Age, Work load Average/day, Education, Son, Pet, Weight, Height, Body mass index. Be sure to make the categorical variables factors in R.

---

1. Split the data into training and test set. How did you do your data split?

2. Fit a Lasso regression model in R using the glmnet package using one choice of alpha. Report the test error.

3. Perform Ridge regression on the same dataset using one choice of alpha. Report the test error.

4. Now fit an Elastic Net model to the data using your own choice of hyper parameters. Report the test error.

5. Use cross-validation to select optimal values of alpha and or lambda in each of the methods in 2-4. Report the optimal hyper parameter values you used for these methods in a Table.

6. Tabulate the test error for each of these models in 5 and compare with their corresponding models you fit in 2,3,4. What does this tell you about the model's performance?

7. For the models in 5, which one will you choose as the final model based on the test errors?

8. Describe your next steps in the modeling process now that you have selected your final model from 7.

9. Based on the final model's output, which factors are most predictive of absenteeism in the workplace? How did you decide on those features?

10. **Phase 1 only:** Discuss the potential implications of your findings for the management of the company.

11. **Phase 1 only:** How might the company use the insights from your final model to reduce absenteeism rates?

12. **Phase 2 only:** For the regularization methods with logistic regression, you can compare these 3 models under AUC, F1 score, etc... Tabulate these metrics under the 3 models and comment on which one you would choose based on these metrics.