Final Exam

Srijesh Reddy Yarram

AA-5300-12: Advanced Analytics

Prof. Michael Fisher

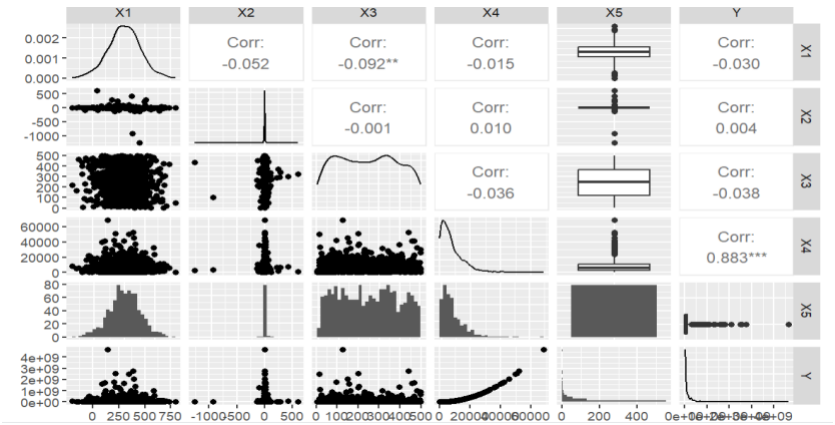
March 13, 2023

1. **Import the dataset associated with this exam Download dataset associated with this exam into R (it is a .csv file, with commas as field separators, and the first row representing the names of the columns). Variables X1 – X4 are continuous, while X5 is categorical. The outcome is also continuous, labeled Y. X5 has two categories, representing the population from which an observation is drawn. Compute a summary of this column and verify that 80% of the observations are drawn from one population and the remaining (20%) of the observations are drawn from another population.**

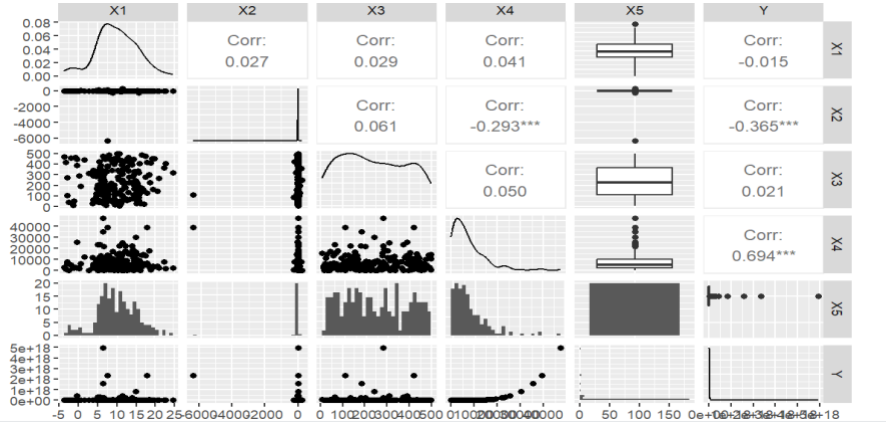
Done.

2. **Next, compute correlations in the subset of columns X1 – X4 and Y. Do this for each subset of rows (i.e., compute one correlation matrix using observations where X5 == 1 and another correlation matrix using observations where X5 == 2. Note whether the correlation matrices are the same or different. Hint: remember this finding as you address the following questions.**

Subset1: df1 <- subset(dff, X5 == 1)



Subset 2 : df2 <- subset(dff, X5 == 2)



The correlation matrices for df1 and df2 are different. This can be seen by comparing the values in each matrix.

In df1, there is a strong positive correlation between X4 and Y (r = 0.882955059), and a weak negative correlation between X1 and Y (r = -0.029723724). The other correlations are either weak or negligible.

In df2, there is a moderate positive correlation between X4 and Y (r = 0.69396589), and a moderate negative correlation between X2 and Y (r = -0.36472343). The other correlations are either weak or negligible.

3. **Next, divide the dataset into two subsets, based on the value of X5: the first subset should consist only of observations from the majority group (value of X5) and the second subset should consist of the remaining observations. The first subset would be used as the training-testing subset and the second one as the holdout set. Exclude X5 from both the training-testing and the holdout subsets (i.e., X5 should not be used as a predictor).**

Done.

4. **Next, fit the following models on the training-testing subset, where the model is fit to minimize the test-set prediction error (RMSE):**
    a. A multiple linear regression model.
    b. A GAM, with appropriate choice of basis functions applied on appropriate predictors. Explain the rationale you have used for determining which basis functions to apply.
    c. A boosted tree model.

Done.

5. **Rank the three models fit, in response to the question above, in increasing order of prediction error. For each model, specify the relative importance of the predictors.**

The order of models in increasing order of prediction error (RMSE) is:

1. BOOSTED TREE FIT (RMSE = 0.6423540)
2. GAM FIT (RMSE = 0.8012199)
3. MLR FIT (RMSE = 1.128875)

The relative importance of predictors in each model is as follows:

1. GAM Model:
    - Variable importance (in decreasing order): X4, X3, X1, X2
    - X4 is the most important predictor with 100% importance, followed by X3 with 17.720240% and X1 with 7.674085%. X2 has 0% importance.
2. Multiple Linear Regression Model: 1.128875
    - Variable importance (in decreasing order): X2, X4, X1, X3

- X2 is the most important predictor with 100% importance, followed by X4 with 82.44699%, X1 with 26.44080%, and X3 with 0% importance.

3. Boosted Tree Model:
    - Variable importance (in decreasing order): X4, X1, X2, X3.
    - X4 is the most important predictor with 100% importance, followed by X1 with 8.705125%, X2 with 2.464967%, and X3 with 0% importance.

For the GAM model, X4 is the most important predictor, followed by X3 and X1. For the Multiple Linear Regression model, X2 is the most important predictor, followed by X4 and X1. For the Boosted Tree model, X4 is the most important predictor, followed by X1 and X2.

6. **Using the three predictive model objects, make predictions of the outcome variable in the holdout subset. Determine the relative ranking of the models, on the basis of the prediction error.**

In increasing order of prediction error:

1. GAM Model: RMSE 0.2600176
2. Multiple Linear Regression Model: RMSE 0.2919405
3. Boosted Tree Model: RMSE 0.5180165

Based on the given results, the GAM model has the lowest prediction error, followed by the Multiple Linear Regression model and then the Boosted Tree model.

7. **On the holdout subset. Comment on the relative predictive performance of the models, comparing the ranking with the ranking found in step 4. Do the ranks match? Why/why not? Discuss based on:**

d. **how the bivariate relationship between each predictor and the outcome is same/different between the two subsets.**
e. **the specific assumptions and approaches used in by the model-fitting algorithms used in multiple linear regression, GAM, and boosted tree approaches.**

On the holdout subset:

Based on the Root Mean Squared Error (RMSE) values of holdout subset, the Boosted tree model has the lowest prediction error (RMSE = 0.00961707), followed by the GAM model (RMSE = 0.0244132) and the Multiple Linear Regression model (RMSE = 0.02646454). Therefore, the Boosted tree model is the best-performing model in terms of predictive accuracy among the three models.

In terms of variable importance, the rankings for the most important predictors are consistent across all three models:

1. X1 is the most important predictor in the multiple linear regression and boosted tree models, with 100% importance in both.
2. X4 is the second-most important predictor of all three models, with importance values ranging from 57.4% to 67.2%.
3. X3 is the third-most important predictor in the GAM model, with an importance value of 33.4%. In the multiple linear regression model, X3 is the third-most important predictor with an importance value of 21.2%. However, in the boosted tree model, X2 has higher importance than X3.
4. X2 has the least importance of all three models, with an importance value of 0%.

COMPARISION OF RANKING BETWEEN TWO SUBSETS:

The ranking of the models based on their predictive accuracy differs between the training and holdout subsets. In the training subset, the GAM model had the lowest RMSE (0.2600176), followed by the Multiple Linear Regression model (0.2919405) and the Boosted Tree model (0.5180165).

However, in the holdout subset, the Boosted Tree model had the lowest RMSE (0.00961707), followed by the GAM model (0.0244132) and the Multiple Linear Regression model (0.02646454).

Comparing the rankings of predictor importance between the holdout subset and training subset:

There are some differences in the rankings of predictor importance between the holdout subset and training subset.

In the GAM model, X4 has the highest importance in both subsets, but in the training subset, X3 is the second most important predictor whereas in the holdout subset, X1 is the second most important predictor. Similarly, in the boosted tree model, X1 is the second most important predictor in the holdout subset whereas in the training subset, X2 is the second most important predictor. In the multiple linear regression model, the most important predictor is X2 in the training subset but X4 in the holdout subset. Additionally, in the training subset, X4 has a higher importance value than X1, while in the holdout subset, X1 has a higher importance value than X4.

Overall, while there are some differences in the rankings of predictor importance between the holdout subset and training subset, there is still some consistency in the relative importance of predictors across the two subsets.

8. **Next, shuffle the dataset 3 times (hint: shuffling was first covered in the week when we looked at k-fold cross-validation – refer to the demonstration document and R code from that week's learning materials). Then, excluding the variable X5 (X5 is not to be used as a predictor), split the dataset into two subsets:**
    f. **an 80% subset, to be used as the training-testing subset**
    g. **a 20% subset, to be used as the holdout subset.**

Done.

9. **Repeat steps 3 – 5 above.**

Done.

10. **Comparing the results from ~~3-5 with those from 6-7~~ the two sets of analyses -- the one where the raw data in their original row arrangement were used versus the one where the data rows were shuffled prior to splitting the dataset and fitting models and making predictions -- comment on the differences/similarities in results, by drawing on the change in the training-set error and in the hold-out set prediction error, which resulted from shuffling of the data prior to partitioning it into training-testing and holdout subsets. What lessons can you draw by looking at the results, taken as a whole (from the two sets of models and their use in prediction on two different types of holdout sets)? Provide your answer by applying the concept of bias-variance trade-off in the context of building and evaluating the performance of predictive models.**

ANALYSIS AFTER SHUFFLING:

Ranking the three models in increasing order of prediction error:

1. GAM Model with RMSE of 0.8867985

2. Boost Model with RMSE of 0.8888142
3. Multiple Linear Regression Model with RMSE of 0.9125247

Relative importance of the predictors in each model:

- For all three models, X4 has the highest relative importance with a score of 100.00.
- In the GAM model, X1 has a relative importance of 12.11, followed by X3 with a relative importance of 5.19. X2 has a relative importance of 0.00.
- In the Boost model, X1 has a relative importance of 14.70, followed by X3 with a relative importance of 1.77. X2 has a relative importance of 0.00.
- In the Multiple Linear Regression model, X1 has a relative importance of 33.64, followed by X3 with a relative importance of 7.34. X2 has a relative importance of 0.00.

COMPARING THE ANLAYSIS OF SUBSETS BEFORE SHUFFLING AND AFTER SHUFFLING:

Similarities:

1. X4 is the most important predictor of all models for both the before and after shuffling set, with a relative importance score of 100% in each model.
2. The GAM model has the lowest prediction error in both sets, while the Boosted Tree model has the highest prediction error in both sets.

Differences:

Before shuffling the data set, the GAM model has the lowest prediction error with an RMSE value of 0.2600176, which means that its predicted values are closer to the actual values. The Multiple Linear Regression model has a slightly higher prediction error with an RMSE value of 0.2919405, indicating that it is not as accurate as the GAM model but still performs reasonably well. The

Boosted Tree model has the highest prediction error with an RMSE value of 0.5180165, indicating that it is the least accurate model among the three.

After shuffling the set prior to partition of training and holdout subsets, the GAM model still has the lowest prediction error with an RMSE value of 0.8867985, indicating that it is still the most accurate model. However, the Boost model has a slightly higher prediction error with an RMSE value of 0.8888142, while the Multiple Linear Regression model has the highest prediction error with an RMSE value of 0.9125247. This shows that the performance of the models can be affected by shuffling the input variables, and the ranking of the models based on their prediction error can change.

The relative importance of the predictors in each model is different between the before and after shuffling sets. For example, in the Multiple Linear Regression model, X2 has 100% importance in the before shuffling set, but 0% importance in the after shuffling set. Additionally, the order of importance of the predictors can change between the before and after shuffling sets.

The relative importance of the predictors can change between models for the same set. For example, in the before shuffling set, X2 is the most important predictor in the Multiple Linear Regression model but has 0% importance in the GAM and Boosted Tree models. However, in the after shuffling set, X2 has 0% importance in all models.

Lessons that I can draw by looking at the results, taken as a whole (from the two sets of models and their use in prediction on two different types of holdout sets) are:

The results of the analysis show that the performance of predictive models can be affected by various factors, including the shuffling of input variables, the choice of modeling techniques, and the size and composition of the data sets. The concept of bias-variance trade-off is relevant in this

context as it refers to the trade-off between underfitting and overfitting, which are two types of errors that can occur in predictive models.

In the context of the current analysis, underfitting can occur if the model is too simple and fails to capture the underlying patterns in the data, leading to poor performance on both training and holdout sets. On the other hand, overfitting can occur if the model is too complex and fits the noise in the data rather than the underlying patterns, leading to high performance on the training set but poor performance on the holdout set.

Before the shuffling, the GAM model performs the best, while the Boosted Tree model performs the worst. The Multiple Linear Regression model has intermediate performance. This suggests that the GAM model is well-suited to capturing the underlying patterns in the data, while the Boosted Tree model may be overfitting to noise in the data. The differences in relative importance of predictors between models suggest that the models are capturing different aspects of the data and may have different biases.

After shuffling, prior to partition of training and holdout subsets, the performance of the models' changes, with the GAM model still performing the best, but the Boosted Tree model performing slightly worse than before shuffling. This suggests that shuffling the input variables can affect the performance of models and that the ranking of models based on their prediction error can change.

Overall, the results highlight the importance of considering the bias-variance trade-off when building and evaluating predictive models. It is important to strike a balance between underfitting and overfitting, and to consider the biases that may be inherent in different modeling techniques. Additionally, it is important to evaluate models on multiple holdouts sets and to consider the impact of shuffling and other factors that may affect model performance.