PS4

Siddharth Koppaku

1. Analysis I: Run simulations with series, neural nets, and random forests (2 points)

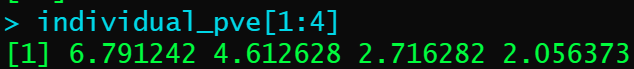
A screenshot of a computer program

Description automatically generated

i. For specification 1, the neural network model has the lowest MSE, making it the best performer in this case. This suggests that the neural network is able to capture the underlying patterns in the data more effectively compared to polynomial regression and random forest. All of these modle preform better in this case they do with specification\_2.

ii. For specification 2, the neural network model still has the lowest MSE, indicating that it performs the best among the three models. This reinforces the idea that neural networks are well-suited for capturing complex relationships and patterns in the data, which is advantageous in this scenario with non-linear transformations and interactions.

2.a





2.b

A black and green background

Description automatically generated

For the most part the MSE makes sense as the most complex model as the losest MSE. This also indicates that we do not have an overfitting problem because in that case a more complex model would have a worst test MSE. The model also shows that Principal Components model performed the worst. This is to be expected because the cumulative\_pve shows that the first four PC only account for 73% of the variance of all of the variables. Overall, it seems in this case the Ridge performed the best.

3. a.

A black and green rectangle with a black background

Description automatically generated

As can be seen above, the Random Forest Model performs significantly better than the Logit Model. This can be due to the fact that Random Forest is better at dealing with more complex non-linear relationships. Random forests can automatically select important features, and this might lead to its superior performances.

3.c.i.

A screen shot of a computer

Description automatically generated

The main difference between the pairwise correlation in the datasets is that is summary dataset there is a much lower correlation between host\_experience and review\_scores\_rating.

3.c.ii.

A black background with numbers and letters

Description automatically generated

The summary model is worst at predicting but it is order of magnitudes for efficient computational. That being said, in this case the add predictive vale seem to be worth due to huge differences.