APSTA-GE 2047 Fall 2018
MDML HW 5
Hongye Wu, James Wu, Yeonji Jung

Question 1

Part C

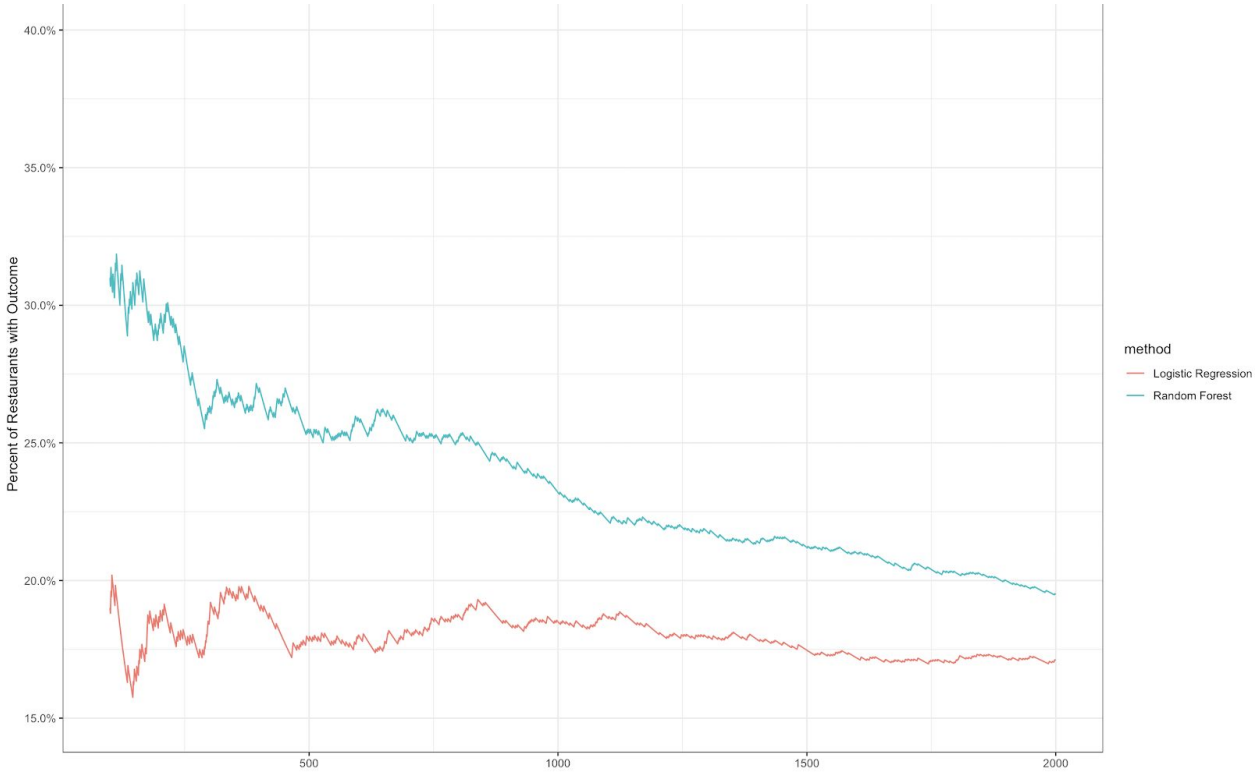
The AUC on test_half is 60.275, and the AUC on test_later is 57.85478. The AUC on test_half is higher because test_half contains data from the dataset and the same year as the training data used to build the random forest model. Test_later is probably a better estimate of the model's performance on unseen data because it contains data taken from a different year from the training set, despite both coming from the same dataset. Yes we should always shuffle the data and split randomly for training/test/validation sets. This helps disrupt the potential patterns in data entry and ordering, and therefore offers a more realistic and unbiased model.

Question 2

Part E

The AUC of a standard logistic regression model on the test dataset is 61.6564 . The AUC of the random forest model on the test dataset is 60.8596. The AUC for the RF model is slightly lower than the logit model. Considering how small the difference is, we don't think there is a substantial difference in terms of model performance.

Part F



APSTA-GE 2047 Fall 2018
MDML HW 5
Hongye Wu, James Wu, Yeonji Jung

Part G

Both the AUCs for the logistic regression and random tree method were relatively low. Based on their AUCs alone, we would be wary to choose either of them as they are both in the 60s. When considering the performance plot, however, it seems like the random forest model was able to predict positive outcomes more accurately, especially in the 250 highest rated restaurants. According to precision, random forest appears to be a better method when our priority is to identify which restaurant would have a score for the initial cycle inspection is 28 or above.

Given the outcome we are trying to predict, it is hard to imagine that the logit model (and its S shape) can accurately capture the underlying relationships between all the variables and the outcome scores. This might be the reason why the AUC is low. But for random forests, the overall performance might be hindered by its ability to overpredict positive cases, as a result of overfitting or the limited predictive power of the independent variables.

Part H

Several ethical issues can be raised related to using a predictive model to prioritize restaurant inspections. First, the way of defining the target variable "outcome" can be problematic. We created this outcome variable as a binary outcome with the criteria of whether the score for the initial cycle inspection was 28 or higher, or not. The prediction results, of course, vary based on the binary score (e.g. the score 20 and the score 35 could make a huge difference in the prediction). To deal with this issues, the theoretically and practically valid ways should be considered (e.g. reference to the restaurant grade policy based on the scores). Second, the way of prioritizing inspections based on whether or not they have a high score can make a problem to prioritize restaurant inspections. For a better decision to prioritize inspections, it can be better to do it based on violation code which can tell more about the critical need to look at the restaurant for customers' safety and health.

In addition to those problems involved in this homework, the relationships between the predictors should be considered to avoid redundancy of information when building a predictive model to prioritize inspections. For example, violation code and critical flag variable can be overlapped depending on how each variable consists of. Furthermore, the critical flag variable can raise an issue in the further prediction modeling since this variable can involve human bias implicitly. This variable was measured as either critical, not critical, or not applicable. Referring to the dataset description, critical violations indicated those most likely to contribute to foodborne illness, which might be prone to be inaccurately recorded depending on the inspector.