4/21/2021: I have finished the select and training a model section, I have found out that my data somehow produce a model not as good as the book’s. The book’s best model was Random Forest but mine was a Linear Regression! Perhaps due to the data collection or my data are relatively small to theirs could be a reason. Or Perhaps due to the nature of the data itself? Maybe I have to constraint it (regularize it) or get more training data. Also trying out more models might work as well. I also tried out some sort of things: theoretically, stratified sample can produce a better model or show a better representation of the population. However, when I used the random sample, it produces overall better models for all of the algorithms. (I have tried Linear Regression, Decision Tree Regression, Random Forest Regression).

**The decision tree** seems to overfit the data. Both my data and the data used in the book are being overfitted. As the tree is trained, the RMSE, when tested with the training data, is 0. Whereas the RMSE when tested with the testing data, is much higher than the Linear Regression. This means it has low bias, however, very high variance. Thus, it is overfitted. The internet said the tree tends to be overfitted especially when the tree is ‘deep’. So, I would have tune and change some parameters of the decision tree in order for it to work.

5/5/2021: In the chapter 3 – classification, the main part of these learnings is measuring the performance of the algorithms. It seems to be very hard to get a ‘perfect’ model. It really depends on what is the task or problem you are trying to solve. We learned that accuracy is generally an unreliable way to measure performance of a classification model. Instead, using the confusion matrix to calculate the precision and recall measure. Those 2 have a trade-off which is why you can’t really make a perfect model with really high precision and recall.