Exercise 2.1 Solution

Part 1. Concepts

$$\textbf{RMSE}\ (\text{Root Mean Square Error}) = \sqrt{\frac{\sum_{i=1}^{i=n} (\hat{y}_i - y_i)^2}{n}}$$

Q2.1. Explain why RMSE is a good metric of model predictive accuracy for continuous Y.

It can consider all the data points from 1st observation to the last n-th observation.
It can also consider the distance between predicted and actual value of Y.
And it has a same unit with Y by combining both Root and Square.

Q2.2. Can we use RMSE for categorical Y? Explain.

No. It differs depending on how we assign numerical values to each categorical value.

Q2.3. Netflix used RMSE in their US$1 million prize. What is the implication?

Netflix's Y is categorical, i.e., 1 star to 5 star grading.

Confusion Matrix

		Actual	
		Not Fraud	Fraud
Model Prediction	Not Fraud	10	17
	Fraud	3	20

Q2.4: Is this a good result? Comment.

Bad. It can predict only 20/(20+17) = 54% of actual fraud.

Q2.5: What is the (a) true positive, (b) false positive?

(a) 20, (b) 3

Q2.6: What is the (a) true negative, (b) false negative?

(a) 10, (b) 17

Q2.7: What is the overall error rate? Some algorithm and researcher reported only the overall error rate. Ok or not?

Overall Error Rate = (17 + 3)/50 = 0.4 = 40%. Not ok. False negative is especially high.

Q2.8 If Model Prediction Error = 0, it means the model is excellent for use. True/False? Explain.

False. Model Prediction Error = 0 can be obtained only on historical data. We are predicting future data, not historical data.

Q2.9 We must always do Train-Test split in every analytics model. True/False? Explain.

False. Train-test split is necessary to get a more reliable estimate of model (future) prediction error, today. Some models are built not for prediction purpose, e.g. identifing high risk factors. In this case, there is no need to sacrifice data for test set , e.g., Unsupervised learning.

Part 2. Run R Script

R script file: BA1w2 baby.R

Dataset: baby.csv

Objectives

- Learn how to use the comments operator #
  - Annotate and explain your code to a human
  - Record your results [optional]
- Learn how to create data within R.
- Learn how to create a dataset within R.
- Learn how to calculate simple statistics in R.
- Learn how to do simple plot in R.
- Learn how to export the final dataset in R to CSV format.

Run BA1w2 baby.R

- Run code one line at a time to check if there is any error in that line.