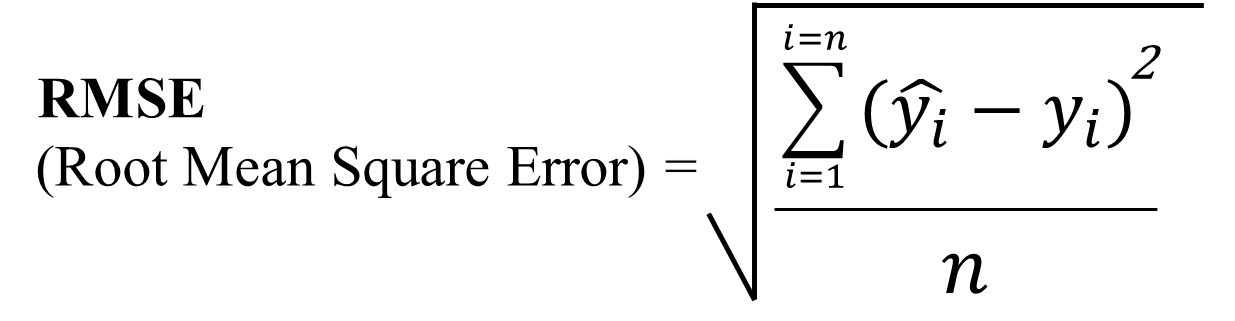
Exercise 2.1

# Part 1. Concepts



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

* Average of n data points, all data points
* Squaring, puts more weight on the error 🡪 Bigger the error 🡪 greater the penalty
* Units 🡪 Square root removes the square
* No negative values due to square

If don’t have

* + Model that have big positive and big negative error add together 🡪 cancel
* **Root Mean Square Error** (**RMSE**) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; **RMSE** is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.
* When Y is continuous, we are facing a regression problem and there are many different values

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

* Not favourable
* E.g Model predicts T and F 🡪 assign nos 1 and 0 🡪 no arbitrary someone might choose things like 20
  + Possible to manupilate data to get a v small RSME such as 1 and 1.000001
* Same case and contribute to the same to the error even though it overpredicts and underpredict
  + Cause is diff but according to RSME is the same

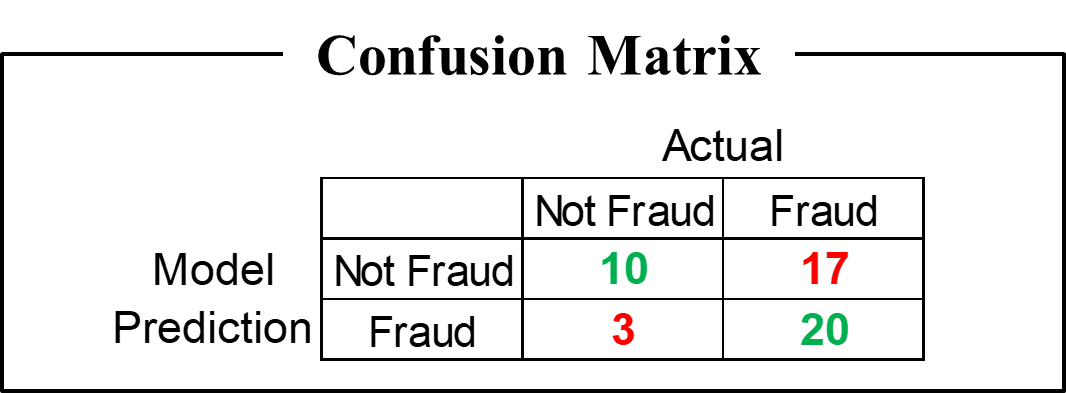
|  |  |  |  |
| --- | --- | --- | --- |
| Model Y^ | Actual Y | Diff | SE |
| 1 \* | 5\* | 4\* | 16units |
| 5\* | 1\* | 4\* | 16 units |

No. Categorical variables contain a finite number of categories or distinct groups. Categorical data might not have a logical order. For example, categorical predictors include gender, material type, and payment method

* When Y is a catrgorical variable, this is a classification problem
* In classification, you have (finite and countable) class labels, which do not correspond to numbers. Therefore you can not use RMSE because it is difficult to find difference between, say, label 'a' and 'b'. Therefore you try other measures such as accuracy, geometric mean, precision, recall, ROC and so on.
* E.g: Use of Logistic regression

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

* R^2 is only for linear regression
* RMSE can be used for anything 🡪 better



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

* + In this case trying to predict fraud
  + It predicts the results accurately half of the time

Actual

Predicted

|  |  |
| --- | --- |
| TN | FP |
| FN | TP |

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

* True 🡪 Model True
* False 🡪 Model false
* True (fact)
* Postive (model)

1. True Positive = 20
2. False Positive = 17

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

1. True Negative = 10
2. False Positive = 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

= FP+FN/TP+TN+FP+FN

= 20/40

= 0.5

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

* Model Prediction Error: Meaning error is zero base on historical data 🡪 unreliable
* Model is excellent for use: for prediction purpose
* What we truly need is an estimate of the accuracy of model prediction for future
* RMSE estimate for future 🡪 do a train test split
* False, often it is biased and optimistic

1. A zero prediction error can only occur when the train and test data set is the same 🡪 Too good to be true
2. Implications of Polynomial Interpolation Theorem
   * Use a very complex polynomial model 🡪 able to hit zero error
   * However, it will result in overfitting

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

False.

* Depends on whether u are using the model to predict a y
  + E.g you are trying to ascertain which of the x var is the most important
  + What are the factors affecting stock price? Which of the factor is the most important
    - GDP?
    - Country?
* Whenever you do a train test there will be a trade off
  + E.G 70 30 VS 50 50
  + Increase train set 🡪 Build better model
  + Test accuracy decreases
  + Can we find a way to decrease the trade off
    - Use 100% train and 100% test
    - Classification regression tree

# 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.