Course3 T1 – Informal Report: Priya Gaddam

Predictions - R Tutorial

In the R Tutorial, I was able to successfully create a model for predicting how far a car can travel based on speed. Predictions in this model were accurate, with a multiple R-Squared value of 0.92. The model also showed that there is a significant impact of Speed on Distance with a p-value of <2.2e-16.

Chart, scatter chart

Description automatically generated

Iris data set:

I created a linear regression model for this data set to predict Petal width based on petal length.

Model produced Multiple r squared value of 0.927 and p-value: < 2.2e-16. This model proved that there is a significant impact of petal length in predicting petal length.

Chart, scatter chart

Description automatically generated

Also created another linear regression model for predicting sepal width based on sepal length.

This model produced a Multiple R squared value of 0.01148 and p-value: 0.2766.

This model does not fit well for predicting sepal length based on sepal width, the p value does not show a very significant relationship found between the 2 variables.

Chart, scatter chart

Description automatically generated

Errors and warnings:

1. Error importing CSV file:

*The file name should be in double quotes and also it should be available in the same directory as the project we are working on.*

1. Error due to syntax/spelling

*summary (IrisDataset) instead of summary(risDataset)*

*str(IrisDataset) instead of str(IrisDatasets)*

*trainSize instead of trainSizes*

*prediction instead of predictions*

*plot(IrisDataset$Sepal.Length) instead of plot(IrisDataset$Sepal.Length*

1. Error in creation of a histogram of Species:

*Species column is character while histograms require numerical data. I was able to change the data into numerical values by using as.factor() and then using as.numeric() conversion on species column.*

1. Error trying to create a qqnorm for the whole Iris dataset.

*This was fixed by creating normal quantile plots for each numerical column.*

1. Error trying to change the species values from characters to numerical values.

*Initially all the values got converted into NAs as the system would not know what value to be converting the chars to. I had to re-import the whole dataset and use as.factor() to automatically convert the chars to numbers.*

1. Error in creating the first train set.

*When creating the trainset at 20% train size was much less than the test size.in reality it should be the opposite. Train size should be more than the test size. So I changed the 0.2 to 0.7 which created the right model.*

1. Linear model definition.

*○ LinearModel<- lm(trainSet$Petal.Width ~ testingSet$Petal.Length)*

*We should use the columns from the same data set(trainset)*

*LinearModel<-lm(Petal.Width~ Petal.Length, trainSet)*

1. Prediction function call

*prediction<-predict(LinearModel,testSet) instead of prediction<-predict(LinearModeltestSet)*

*Predictions was called instead of Prediction object*

Takeaways:

* Installing R and R studio was easy following the documentation.
* Took some time to understand the differences between python and R code.
* Lot of google help to understand the R function call syntaxes and default values also interpreting the results.