To: Danielle Sherman

Subject: Customer Buying Patterns Analysis

Danielle,

Here are my notes from my analysis per your request:

1. **My predictions concerning how far a certain car can travel based on speed.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TestRecord | name.of.car | speed.of.car | distance.of.car | Predicted Distance |
| 1 | Ford | 4 | 2 | -14.9541490857947 |
| 2 | Jeep | 4 | 4 | -14.9541490857947 |
| 6 | BMW | 9 | 16 | 10.4132911392405 |
| 16 | GMC | 13 | 26 | 30.7072433192686 |
| 18 | Chrysler | 13 | 28 | 30.7072433192686 |
| 20 | Acura | 14 | 32 | 35.7807313642756 |
| 22 | Chevrolet | 14 | 34 | 35.7807313642756 |
| 23 | Buick | 14 | 34 | 35.7807313642756 |
| 34 | Land Rover | 18 | 52 | 56.0746835443038 |
| 35 | Lexus | 18 | 54 | 56.0746835443038 |
| 38 | Nissan | 19 | 56 | 61.1481715893108 |
| 39 | GMC | 20 | 60 | 66.2216596343178 |
| 44 | Audi | 22 | 76 | 76.3686357243319 |
| 46 | Buick | 24 | 84 | 86.515611814346 |
| 47 | Jeep | 24 | 85 | 86.515611814346 |

1. **My predictions concerning the petal length through using the petal’s width.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TestRecord | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species | Predicted Petal Length |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | 1 | 1.48243059262792 |
| 2 | 4.9 | 3 | 1.4 | 0.2 | 1 | 1.48243059262792 |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | 1 | 1.48243059262792 |
| 5 | 5 | 3.6 | 1.4 | 0.2 | 1 | 1.48243059262792 |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 | 1 | 1.48243059262792 |
| 18 | 5.1 | 3.5 | 1.4 | 0.3 | 1 | 1.7100397064816 |
| 19 | 5.7 | 3.8 | 1.7 | 0.3 | 1 | 1.7100397064816 |
| 28 | 5.2 | 3.5 | 1.5 | 0.2 | 1 | 1.48243059262792 |
| 29 | 5.2 | 3.4 | 1.4 | 0.2 | 1 | 1.48243059262792 |
| 33 | 5.2 | 4.1 | 1.5 | 0.1 | 1 | 1.25482147877425 |
| 36 | 5 | 3.2 | 1.2 | 0.2 | 1 | 1.48243059262792 |
| 45 | 5.1 | 3.8 | 1.9 | 0.4 | 1 | 1.93764882033528 |
| 48 | 4.6 | 3.2 | 1.4 | 0.2 | 1 | 1.48243059262792 |
| 49 | 5.3 | 3.7 | 1.5 | 0.2 | 1 | 1.48243059262792 |
| 55 | 6.5 | 2.8 | 4.6 | 1.5 | 2 | 4.44134907272572 |
| 56 | 5.7 | 2.8 | 4.5 | 1.3 | 2 | 3.98613084501837 |
| 57 | 6.3 | 3.3 | 4.7 | 1.6 | 2 | 4.6689581865794 |
| 58 | 4.9 | 2.4 | 3.3 | 1 | 2 | 3.30330350345734 |
| 59 | 6.6 | 2.9 | 4.6 | 1.3 | 2 | 3.98613084501837 |
| 61 | 5 | 2 | 3.5 | 1 | 2 | 3.30330350345734 |
| 62 | 5.9 | 3 | 4.2 | 1.5 | 2 | 4.44134907272572 |
| 65 | 5.6 | 2.9 | 3.6 | 1.3 | 2 | 3.98613084501837 |
| 66 | 6.7 | 3.1 | 4.4 | 1.4 | 2 | 4.21373995887204 |
| 68 | 5.8 | 2.7 | 4.1 | 1 | 2 | 3.30330350345734 |
| 70 | 5.6 | 2.5 | 3.9 | 1.1 | 2 | 3.53091261731101 |
| 77 | 6.8 | 2.8 | 4.8 | 1.4 | 2 | 4.21373995887204 |
| 83 | 5.8 | 2.7 | 3.9 | 1.2 | 2 | 3.75852173116469 |
| 84 | 6 | 2.7 | 5.1 | 1.6 | 2 | 4.6689581865794 |
| 94 | 5 | 2.3 | 3.3 | 1 | 2 | 3.30330350345734 |
| 95 | 5.6 | 2.7 | 4.2 | 1.3 | 2 | 3.98613084501837 |
| 98 | 6.2 | 2.9 | 4.3 | 1.3 | 2 | 3.98613084501837 |
| 100 | 5.7 | 2.8 | 4.1 | 1.3 | 2 | 3.98613084501837 |
| 101 | 6.3 | 3.3 | 6 | 2.5 | 3 | 6.71744021126249 |
| 104 | 6.3 | 2.9 | 5.6 | 1.8 | 3 | 5.12417641428675 |
| 105 | 6.5 | 3 | 5.8 | 2.2 | 3 | 6.03461286970146 |
| 111 | 6.5 | 3.2 | 5.1 | 2 | 3 | 5.5793946419941 |
| 113 | 6.8 | 3 | 5.5 | 2.1 | 3 | 5.80700375584778 |
| 116 | 6.4 | 3.2 | 5.3 | 2.3 | 3 | 6.26222198355513 |
| 125 | 6.7 | 3.3 | 5.7 | 2.1 | 3 | 5.80700375584778 |
| 131 | 7.4 | 2.8 | 6.1 | 1.9 | 3 | 5.35178552814043 |
| 133 | 6.4 | 2.8 | 5.6 | 2.2 | 3 | 6.03461286970146 |
| 135 | 6.1 | 2.6 | 5.6 | 1.4 | 3 | 4.21373995887204 |
| 140 | 6.9 | 3.1 | 5.4 | 2.1 | 3 | 5.80700375584778 |
| 141 | 6.7 | 3.1 | 5.6 | 2.4 | 3 | 6.48983109740881 |
| 145 | 6.7 | 3.3 | 5.7 | 2.5 | 3 | 6.71744021126249 |

1. **The errors/warning messages that I encountered and how I overcame them.**

Here is a copy of the script you provided me. I added some comments at the end of each line prefaced by “#” that describes the error I found (if any otherwise “ok”) and in italics.

install.packages("readr") *#added "" around readr*

library(readr) *#removed "" around readr*

IrisDataset <- read.csv("iris.csv") *#added "" around filename*

attributes(IrisDataset) *#ok*

summary(IrisDataset) *# added I to risDataset mispelling*

str(IrisDataset) *#removed s at end of IrisDataset*

names(IrisDataset) *#ok*

NumericSpecies<-as.numeric(IrisDataset$Species) *#have to convert Species to numeric to utilize hist*

hist(NumericSpecies) *#ok*

plot(IrisDataset$Sepal.Length) *#added closing ‘)’*

qqnorm(IrisDataset$Sepal.Length) *#can't use the whole dataset - have to plot only one feature at time*

IrisDataset$Species<- as.numeric(IrisDataset$Species) *#ok*

set.seed(123) *#ok*

trainSize <- round(nrow(IrisDataset) \* 0.7) *#change % to .7 for training data set size*

testSize <- nrow(IrisDataset) - trainSize *#change trainSet to trainsize to calculate test data set size*

trainSize *# delete s misspelling*

testSize *#ok*

training\_indices<-sample(seq\_len(nrow(IrisDataset)),size =trainSize) *#need to define the training set indices*

trainSet <- IrisDataset[training\_indices, ] *#ok*

testSet <- IrisDataset[-training\_indices, ] *#ok*

set.seed(405) *# ok - no real reason to do this here though*

trainSet <- IrisDataset[training\_indices, ] *#ok*

testSet <- IrisDataset[-training\_indices, ] *#ok*

LinearModel<- lm(Petal.Length~ Petal.Width, trainSet) *#Label is Length predicted by Width on trainSet, also don't use dataset name IrisDataset$Petal.Width*

summary(LinearModel) *#ok*

d = data.frame(testSet, predict(LinearModel, testSet, interval="prediction")) *#put the prediction in a dataframe so can output easier*

#prediction←predict(LinearModel,testSet) *#added comma in predict argument list – this was the original line in the error script. I commented it out to utilize a data frame to grab the data in a better format for the report but I note the error in this line that I found*

df <- d[ -c(1) ] *#added to remove the first column of new data frame with is duplicate of second*

df *#outputs the dataframe like the statement “prediction” would have*

write.csv(df, 'predictions.csv') *#writes to csv so I can just grab the table already formatted correctly*

1. **Was it straightforward to install R and RStudio?**

I had no issues. This was helpful.

1. Was the tutorial useful? Would you recommend it to others?

This was really helpful. There were enough errors that it made me read closely. I was also a little challenged to figure out the error with the LinearRegression model where the dataset$column was in the statement. It took me a while to figure out to remove the dataset part. I also spend a little time trying to understand how to write back to a file, as well as format the output using a dataframe. So, I felt it was the right amount of effort required to allow me to spend a little time researching related topics and do a little more in the time we had.

1. What are the main lessons you've learned from this experience?

As with any of my experiences with writing code of any type: take my time and google is my friend.

1. What recommendations would you give to other employees who need to get started using R and doing predictive analytics in R instead of Rapidminer?

Give yourself more time to learn the complexities – which is really the flexibilities – of R. Rapid Miner is quick to learn, and limited in configurable options. R has a ton of options for each function. Run the function with defaults first and ensure you can produce an acceptable first pass without warnings or errors. Then spend a little time exploring the other options for that function that might improve the result of that function call – whatever the objective of the function might be.