Microsoft

OFFICIAL MICROSOFT LEARNING PRODUCT

20773A

Analyzing Big Data with Microsoft R

Product Number: 20773A

Part Number (if applicable):

Released: 10/2018

Module 1

Microsoft R Server and Microsoft R Client

Contents:

Lesson 1

Introduction to Microsoft R Server

Contents:

Question and Answers

Question: Which of the following is not an advantage of using R Server over open source R?

( ) The ScaleR functions in R Server improve distribution and parallelization of analysis operations across nodes in a cluster.

( ) R Server can run advanced machine learning algorithms implemented by the MicrosoftML package.

( ) R Server runs exclusively on Windows

( ) R Server supports the remote execution of R code.

( ) R Server enables you to package and deploy code as a web service.

Answer:

( ) The ScaleR functions in R Server improve distribution and parallelization of analysis operations across nodes in a cluster.

( ) R Server can run advanced machine learning algorithms implemented by the MicrosoftML package.

(√) R Server runs exclusively on Windows

( ) R Server supports the remote execution of R code.

( ) R Server enables you to package and deploy code as a web service.

Lesson 2

Using Microsoft R Client

Contents:

Question and Answers

Question: How can you run interactive code remotely in R Server from R Client?

( ) Use the remoteExecute function and specify the code to run on the remote R Server.

( ) Specify the name of the server on which to run the code as the remoteServer parameter to the ScaleR functions

( ) Deploy the code to the remote R Server.

( ) Use the remoteLogin function to connect to the remote R Server and start an interactive session on that server.

( ) You can't. You must log in to the remote server manually and start an interactive R session there.

Answer:

( ) Use the remoteExecute function and specify the code to run on the remote R Server.

( ) Specify the name of the server on which to run the code as the remoteServer parameter to the ScaleR functions

( ) Deploy the code to the remote R Server.

(√) Use the remoteLogin function to connect to the remote R Server and start an interactive session on that server.

( ) You can't. You must log in to the remote server manually and start an interactive R session there.

Demonstration: Using R Client with Visual Studio and RStudio

Lesson 3

The ScaleR functions

Contents:

Question and Answers

Question: You run the rxSummary function over a very large dataset in a session running in R Client. The rxSummary function will attempt to read all of the data in the dataset into memory first. True or False?

( ) True

( ) False

Answer:

(√) True

( ) False

Lab Review Questions and Answers

Lab: Exploring Microsoft R Server and Microsoft R Client

Question and Answers

Exercise 1: Using R Client in RTVS and RStudio

Question: How many columns are there in the data frame, including the new **MonthName** column?

Answer: 31

Exercise 2: Using R Client in RTVS and RStudio

Question: How many columns are there in the data frame?

Answer: 1,135,221

Exercise 3: Using R Client in RTVS and RStudio

Question: What are the minimum and maximum arrival delay times recorded in the data frame?

Answer: The correct answer is -1 298 minutes (a flight appears to have arrived 21 hours and 38 minutes early), and 1,423 minutes (23 hours and 43 minutes late).

Exercise 4: Using R Client in RTVS and RStudio

Question: How many flights were cancelled in June?

Answer: 3,715

Module 2

Exploring Big Data

Contents:

Lesson 1

Understanding ScaleR data sources

Contents:

Question and Answers

Question: You want to use the **rxImport** function to transfer data from a local CSV file into a SQL Server table. You use an **RxTextData** data source to read the data from the CSV file and an **RxSqlServerData** data source to write to the SQL Server database. You should perform this operation in the **RxInSqlServer** compute context. True or False?

( ) True

( ) False

Answer:

( ) True

(√) False

Demonstration: Reading data from SQL Server and HDFS

Lesson 2

Reading and writing XDF data

Contents:

Question and Answers

Question: Which argument to the **rxImport** function does not enable you to filter data?

( ) rowSelection

( ) varsToDrop

( ) varsToKeep

( ) colClasses

( ) numRows

Answer:

( ) rowSelection

( ) varsToDrop

( ) varsToKeep

(√) colClasses

( ) numRows

Resources

Transforming data on import

 **Best Practice:** Test transformations over a small subset of the data first. This gives you a way to quickly test that the transformation is correct. When you are satisfied that the transformation is working correctly, you can perform it over the entire data.

 **Best Practice:** The **transforms** argument contains a list of transformations. Use this list to batch together transformations over different fields, rather than performing separate runs of **rxImport** over the same data, each with its own single transformation.

Lesson 3

Summarizing data in an XDF object

Contents:

Question and Answers

Question: When you perform a base R function, such as **summary** or **head**, over XDF data, the XDF data is cast into a data frame first. True or False?

( ) True

( ) False

Answer:

( ) True

(√) False

Demonstration: Transforming, summarizing, and cross-tabulating XDF data

Lab Review Questions and Answers

Lab: Exploring big data

Question and Answers

Lab Review

Question: From the summaries that you developed, were you able to perceive any relationship between airport or state and flight delay times?

> **Answer:** No. The analysis is inconclusive at this stage, but there does not appear to be any relationship between airport and/or state and flight delay times. However, it is difficult to be definitive without examining the data further—at this stage, it would appear that other factors might be more important.

Question: Given your answer to the first question, is the effort performing these tasks justified so far?

> **Answer:** Yes. It is important to be able to disprove possible relationships so that you can eliminate them from the analysis. You then know to focus your efforts elsewhere.

Module 3

Visualizing Big Data

Contents:

Lesson 1

Visualizing in-memory data

Contents:

Question and Answers

Question: The data source for a ggplot2 graph must be a data frame. True or False.

( ) True

( ) False

Answer:

(√) True

( ) False

Demonstration: Creating a faceted plot with overlays using ggplot

Lesson 2

Visualizing big data

Contents:

Question and Answers

Question: How can you control the number of bins used by the rxHistogram function if the data being plotted is continuous rather than categorical?

( ) Convert the data into a factor.

( ) Specify the numBreaks argument of the rxHistogram function.

( ) Filter the data to remove any non-factor items.

( ) Use the transforms argument of the rxHistogram function to round the data up or down a set of discrete values.

( ) Set the xNumTicks argument of the rxHistgram function to the number required.

Answer:

( ) Convert the data into a factor.

(√) Specify the numBreaks argument of the rxHistogram function.

( ) Filter the data to remove any non-factor items.

( ) Use the transforms argument of the rxHistogram function to round the data up or down a set of discrete values.

( ) Set the xNumTicks argument of the rxHistgram function to the number required.

Demonstration: Generating a histogram with rxHistogram

Module Review and Takeaways

Tools

The ggplot2 package is huge and has a bewildering array of different options, plot types and transformations. See these resources for further information:

The official ggplot2 documentation covers every option in fine detail—see: http://docs.ggplot2.org/current/index.html.

RStudio produces a very useful cheat sheet—see: https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf.

Winston Chang's *Cookbook for R* has an excellent chapter on ggplot2—see: http://www.cookbook-r.com/Graphs/.

Hadley Wickham, the author of ggplot2, wrote a paper on the *Layered Grammar of Graphics* used in the package—see: http://vita.had.co.nz/papers/layered-grammar.pdf.

Whatever ggplot2 problem you have, it is very likely someone else has experienced something similar and has found a fix on stackoverflow.com.

Lab Review Questions and Answers

Lab: Visualizing data

Question and Answers

Exercise 1: Visualizing data using the ggplot2 package

Question: Does the regression indicate that there is any relationship between flight distance and flight delay times?

> **Answer:** Not really. The regression line across all states is flat and shows little variation with distance. When zooming in to the state level, the same is true for most states although some states, such as South Dakota (SD), show some deviation. However, it's likely that is because there is only a small number of flights in the sample data for these states, resulting in a high signal-to-noise ratio of the data.

Exercise 2: Examining relationships in big data using the rxLinePlot function

Question: What do you observe about the graphs showing flight delay as a proportion of travel time against distance? Does this bear out your theory that there is little relationship between these two variables? If not, how do you account for any discrepancy between your theory and the observed data?

> **Answer:** There appears to be a significant relationship between distance and flight delay as a proportion of travel time—the delay percentage time drops with distance. This contradicts the theory posed earlier. However, the original theory assumed that travel time itself was linearly related to distance, and this is unlikely to be the case. Short flights take comparably more time than long flights, due to take-off and landing overheads, delays on approach, and other factors. The longer the flight, the smaller these overheads become as a proportion of the travel time.
>
> This shows the danger of making assumptions about data that can lead to false conclusions. If you have time, plot a line graph of travel time against distance. You should see that the relationship is not linear.

Exercise 3: Examining relationships in big data using the rxLinePlot function

Question: Using the graph showing delay times against departure time for each day of the week, which time of day generally suffers the worst flight delays, and which day of the week has the longest delays in this period?

> **Answer:** The graph shows that delays build up during the afternoon of each day, and the evening tends to be the period that has the longest delays. Friday evening looks to be the worst time to travel.

Exercise 4: Creating histograms over big data

Question: What is the most common arrival delay (in minutes), and how frequently does this delay occur?

> **Answer:** The most frequent arrival delay is between 5 and 10 minutes. This delay occurs nearly 18 percent of the time.

Exercise 5: Creating histograms over big data

Question: Which month has the most delays caused by poor weather? Which months have the least delays caused by poor weather?

Answer: December is the worst month for delays caused by bad weather. The most benign weather appears to occur in April and September.

Module 4

Processing Big Data

Contents:

Lesson 1

Transforming big data

Contents:

Question and Answers

Question: You create a transformation function for use with rxDataStep. The transformation function runs once for every row retrieved by rxDataStep. True or False?

( ) True

( ) False

Answer:

( ) True

(√) False

Resources

Using the rxDataStep function

Best Practice: Remember that the **transforms** argument is a list that can contain any number of transformations. Add all the transformations that you require to this list to process the data in a single pass. Do not perform multiple runs of **rxDataStep**, each implementing a single transformation.

Best Practice: Avoid defining transformations that require access to all observations in the dataset simultaneously, such as the **poly** and **solve** matrix operations. These operations can be expensive because they can involve repeatedly reading the dataset, and they will be performed for every row in the dataset.
Also, when sampling data in a transform, remember that the sampling algorithm only has access to the current chunk unless you reread the entire dataset.

Using custom transformation functions

Best Practice: Milliseconds matter
Make sure that your transformations are as efficient as possible. It is worth spending time tuning and analyzing their performance on small subsets of your data. A large dataset might consist of 100 million rows. If each iteration in a transformation function takes 1 millisecond, then it will require 100,000 seconds, or nearly 28 hours, to process your data. Needless to say, if your code is only slightly less efficient and takes 1.5 milliseconds per iteration, this will add another 14 hours to the processing time.
This is also a situation where you should consider the size of the platform on which you are running your code. Add as much memory and processing power to your computing environment as possible. It might even be worth creating a temporary cluster of large VMs in Azure® especially to perform the task. You can remove these VMs once you have finished.

Reblocking an XDF file

Best Practice: If you change the name of a variable in an XDF file, you might need to reblock the file afterwards to ensure that the metadata recording the new variable name is updated in every block.

 **Best Practice:** To reduce the chances of fragmentation, avoid transformations that change the length of a variable.

Demonstration: Using a transformation function to calculate a running total

Lesson 2

Managing big datasets

Contents:

Question and Answers

Question: Which option for the rxMerge function enables you to combine data horizontally from two different datasets that have a different number of rows?

( ) oneToOne

( ) union

( ) combine

( ) lookup

( ) inner

Answer:

( ) oneToOne

( ) union

( ) combine

( ) lookup

(√) inner

Demonstration: Sorting data with rxSort

Module 5

Parallelizing Analysis Operations

Contents:

Lesson 1

Using the RxLocalParallel compute context with rxExec

Contents:

Question and Answers

Question: The RxLocalParallel compute context enables you to parallelize all ScaleR functions running in the local compute context when running on R Client. True or False.

( ) True

( ) False

Answer:

( ) True

(√) False

Demonstration: Using rxExec to perform tasks in parallel

Demonstration: Creating waiting and non-waiting jobs in Hadoop

Lesson 2

Using the RevoPemaR package

Contents:

Question and Answers

Question: You have written a PEMA class that performs a complex analysis in parallel. You decide to test the class on a cluster with a single compute node. The data is divided into 50 chunks. How many times does the **updateResults** method of the PEMA object run?

( ) 50

( ) 1

( ) 0

( ) It varies, depending on how the master node decides to distribute the work, but it could be anywhere between 1 and 50.

( ) 2 (once at the start of the operation and once at the end)

Answer:

( ) 50

( ) 1

(√) 0

( ) It varies, depending on how the master node decides to distribute the work, but it could be anywhere between 1 and 50.

( ) 2 (once at the start of the operation and once at the end)

Demonstration: Creating and running a PEMA object

Lab Review Questions and Answers

Lab: Parallelizing analysis operations

Question and Answers

Exercise 1: Capturing flight delay times and frequencies

Question: How many flights were made from LAX to JFK by DL, and how many were delayed? What was the longest delay?

Answer: There were 2,931 flights, of which 1,628 were delayed. The longest delay was 1,235 minutes (20 hours and 35 minutes).

Exercise 2: Capturing flight delay times and frequencies

Question: How could you verify that the results produced by the **PemaFlightDelays** object are correct?

Answer: Use the **rxSummary** function to generate a summary of the delay information using the same parameters. For example:

```
rxSummary(~Delay, flightDelayData, rowSelection = (Origin == "ABE") & (Dest ==
"PIT") & (UniqueCarrier == "US") & (Delay > 0))
```

This tells you the number of delayed flights. If you omit the **Delay** clause in the **rowSelection** expression, you can find the total number of flights. These figures should match the **totalDelays** and **totalFlights** values in the **PemaFlightDelays** object.

Module 6

Creating and Evaluating Regression Models

Contents:

Lesson 1

Clustering big data

Contents:

Question and Answers

Question: You have used the rxKmeans function to cluster data. The ration of the "between cluster sum of squares" and the "total sum of squares" across the cluster is very high (99.8%). What does this indicate?

( ) Clustering has been ineffective as each cluster contains vastly differing data.

( ) Clustering has been effective as most of the clusters contain highly homogenous data.

( ) You cannot draw any conclusions about the effectiveness of the clustering using this measure.

( ) The data values are too disparate to be clustered effectively.

( ) All the data values in the entire dataset are nearly identical.

Answer:

( ) Clustering has been ineffective as each cluster contains vastly differing data.

(√) Clustering has been effective as most of the clusters contain highly homogenous data.

( ) You cannot draw any conclusions about the effectiveness of the clustering using this measure.

( ) The data values are too disparate to be clustered effectively.

( ) All the data values in the entire dataset are nearly identical.

Demonstration: Creating and examining a cluster

Lesson 2

Generating regression models and making predictions

Contents:

Question and Answers

Question: You have generated a linear model using the rxLinMod function over a large dataset. You have tested the model by making predictions using this model and comparing them to a set of known results. You have plotted a ROC curve displaying the accuracy of the known results to the predicted values. The ROC curve shows a diagonal straight line from the point (0.0) to the point (1,1). This indicates that the model is making very accurate predictions. True or False?

( ) True

( ) False

Answer:

( ) True

(√) False

Demonstration: Fitting a linear model and making predictions

Demonstration: Modeling using logistic regression

Lab Review Questions and Answers

Lab: Creating and using a regression model

Question and Answers

Exercise 1: Clustering flight delay data

Question: What do the graphs you created in this exercise tell you about flights made from 6:01 PM onwards?

> **Answer:** Although there are fewer flights at this time than earlier in the day, these flights appear to be more prone to delay.

Exercise 2: Fitting a linear model to clustered data

Question: What conclusions can you draw about the predictions made by the linear model using the clustered data?

> **Answer:** The individual predictions were not overly accurate. However, that's not surprising because flight delay time is unlikely to be a function solely of departure time (if it was, then airlines would simply reschedule the arrival times and no flights would be late). The main purpose of this model was to assess the relationship between delay and departure time across an entire cohort of observations—the overall shape of the graph of predicted delays was similar to that of the real data. The regression line displayed was an almost exact match in both cases, so the predictions made on the macro scale show some promise. However, the model seems to underestimate delays at the individual level, and it is worth investigating this further.

Lab Review

Question: This lab analyses the flight delay data to try and predict the answer to the question, "How long will my flight be delayed if it leaves at 'N' o'clock?" The linear regression analysis shows that, although it is nearly impossible to answer this question accurately for a specific flight (the departure time is clearly not the only predictor variable involved in determining delays), it is possible to generalize across all flights. What might be a better question to ask about flight delays, and how could you model this to determine a possible answer?

> **Answer:** A better question might be, "What are the chances that my flight will be delayed?" You could create a logit model to analyze the data to help answer this question.

Module 7

Creating and Evaluating Partitioning Models

Contents:

Lesson 1

Creating partitioning models based on decision trees

Contents:

Question and Answers

Question: When might you consider constructing a partitioning model rather than a linear model, to make predictions?

( ) If the dataset is too small to build a linear model.

( ) If the dataset is too large to build a linear model.

( ) If the relationship between the predictor variables and the dependent variable are non-linear.

( ) To avoid the overhead of sorting a large dataset first.

( ) If the data in the dataset is not uniformly distributed.

Answer:

( ) If the dataset is too small to build a linear model.

( ) If the dataset is too large to build a linear model.

(√) If the relationship between the predictor variables and the dependent variable are non-linear.

( ) To avoid the overhead of sorting a large dataset first.

( ) If the data in the dataset is not uniformly distributed.

Demonstration: Building partitioning models

Lesson 2

Evaluating models

Contents:

Question and Answers

Question: You should always test the accuracy of predictions made by a model against the training dataset. True or False?

( ) True

( ) False

Answer:

( ) True

(√) False

Demonstration: Running predictions against partitioning models

Lesson 3

Using the MicrosoftML package

Contents:

Question and Answers

Question: The functions in the MicrosoftML package are only available to R Server running on VMs using Azure, due to the amount of processing power involved. True or False?

( ) True

( ) False

Answer:

( ) True

(√) False

Lab Review Questions and Answers

Lab: Creating a partitioning model to make predictions

Question and Answers

Exercise 1: Fitting a DTree model and making predictions

Question: How many predicted delays were within 10 minutes of the actual reported delays? What proportion of the observations is this?

> **Answer:** The answers shown here were generated during testing. Your results should be similar but are unlikely to be identical:
>
> A total of 10,344 predictions were within 10 minutes of the actual delay time. This is 17.7 percent.

Exercise 2: Fitting a DTree model and making predictions

Question: How many predicted delays were within 5 percent of the actual delays?

> **Answer:** A total of 31,664 predictions were within 5 percent of the actual delay time. This is 54.2 percent.

Exercise 3: Fitting a DTree model and making predictions

Question: How many predicted delays were within 10 percent of the actual delays?

> **Answer:** A total of 32,189 predictions were within 10 percent of the actual delay time. This is 55.1 percent.

Exercise 4: Fitting a DTree model and making predictions

Question: How many predicted delays were within 50 percent of the actual delays?

> **Answer:** A total of 37,325 predictions were within 50 percent of the actual delay time. This is 63.9 percent.

Exercise 5: Fitting a DForest model and making predictions

Question: How many predicted delays were within 10 minutes of the actual reported delays? What proportion of the observations is this? How does this compare to the predictions made using the DTree model?

> **Answer:** The answers shown here were generated during testing. Your results should be similar but are unlikely to be identical:
>
> A total of 9,783 predictions were within 10 minutes of the actual delay time. This is 16.8 percent and 0.9 percent below that of the DTree model.

Exercise 6: Fitting a DForest model and making predictions

Question: How many predicted delays were within 5 percent of the actual delays? How does this compare to the predictions made using the DTree model?

> **Answer:** A total of 31,538 predictions were within 5 percent of the actual delay time. This is 54 percent and marginally below that of the DTree model (0.2 percent).

Exercise 7: Fitting a DForest model and making predictions

Question: How many predicted delays were within 10 percent of the actual delays? How does this compare to the predictions made using the DTree model?

> **Answer:** A total of 32,139 predictions were within 10 percent of the actual delay time. This is 55.0 percent and marginally below that of the DTree model (0.1 percent).

Exercise 8: Fitting a DForest model and making predictions

Question: How many predicted delays were within 50 percent of the actual delays? How does this compare to the predictions made using the DTree model?

> **Answer:** A total of 37,371 predictions were within 50 percent of the actual delay time. This is 64.0 percent and marginally above that of the DTree model (0.1 percent).

Exercise 9: Fitting a DForest model and making predictions

Question: Was the DForest model more accurate at predicting delays than the DTree model? What conclusions can you draw?

> **Answer:** The DForest model was slightly worse overall than the DTree model at predicting delays. This could be due to a number of reasons, including:
>
> - Overfitting of the individual trees in the forest to the data, so that when they are averaged out, they could be more rather than less biased to the training data, and consequently not work as well on the test data. You could try and reduce the **maxNumBins** argument to see if this is the case.
> - Delay time is probably not just a function of arrival time, departure time, month, and day of the week. If the model is missing significant factors (such as the weather due to the time of year, increased number of passengers due to holidays, and so on), then using a DForest rather than a DTree could exaggerate the effects of these missing parameters.

Exercise 10: Fitting a DForest model and making predictions

Question: Was the DForest model with a reduced depth more or less accurate than the previous model? What conclusion to you reach?

> **Answer:** The results are significantly worse than using the previous DForest model when predicting delays that are within 10 minutes of the actual values—although it is slightly better at the 5 percent, 10 percent, and 50 percent levels. Overfitting was probably not an issue, so perhaps the model should include other variables.

Exercise 11: Fitting a DTree model with different variables

Question: How do the predictions made using the new set of predictor variables compare to the previous set? What are your conclusions?

> **Answer:** The predictions made using this set of variables are comparable to those made using the arrival time, departure time, month, and day of the week variables at the 5 percent, 10 percent, and 50 percent levels—although accuracy is reduced when looking at predictions that lie within 10 minutes of the actual values. Perhaps the model should include all of these variables.

Exercise 12: Fitting a DTree model with a combined set of variables

Question: What do the results of this model show about the accuracy of the predictions?

Answer: The accuracy of predictions improved when predicting delays that were within 10 minutes of the actual delay value (up to 19.6 percent during a test run). The accuracy at the 5 percent, 10 percent, and 50 percent levels was roughly the same as before.

Module 8

Processing Big Data in SQL Server and Hadoop

Contents:

Lesson 1

Integrating R with SQL Server

Contents:

Question and Answers

Question: When you use the **sp_execute_external_script** stored procedure to run R code from SQL Server, the R code is executed by SQL Server. True or False?

( ) True

( ) False

Answer:

( ) True

(√) False

Demonstration: Storing and retrieving R objects from a database

Lesson 2

Using ScaleR functions with Hadoop on a Map/Reduce cluster

Contents:

Question and Answers

Question: All ScaleR operations running in the **RxHadoopMR** compute context are performed as Hadoop Map/Reduce jobs. True or False?

( ) True

( ) False

Answer:

( ) True

(√) False

Demonstration: Running an analysis as a Map/Reduce job

Lesson 3

Using ScaleR functions with Spark

Contents:

Question and Answers

Question: Which data source can you use to connect to a Hive database when using the **RxHadoopMR** compute context?

( ) RxOdbcData

( ) RxHiveData

( ) RxHadoopData

( ) You don't need to use a specific data source. You can start a sparklyr session to read the Hive data.

( ) RxSpark

Answer:

(√) RxOdbcData

( ) RxHiveData

( ) RxHadoopData

( ) You don't need to use a specific data source. You can start a sparklyr session to read the Hive data.

( ) RxSpark

Lab Review Questions and Answers

Lab A: Deploying a predictive model to SQL Server

Question and Answers

Exercise 1: Upload the flight delay data

Question: According to the first histogram, is there a pattern to weather delays?

Answer: There appear to be fewer delays, proportionally, during the summer and early autumn, but apart from that there is no specific pattern.

Exercise 2: Upload the flight delay data

Question: Using the second histogram, which states appear to have the most delays as a proportion of the flights that depart from airports in those states? Proportionally, which state has the fewest delays?

Answer: Texas, Illinois, Florida, Georgia, New York, Pennsylvania, Colorado, New Jersey, Ohio, Michigan, Minnesota, Missouri, and Massachusetts have proportionally more delays than most other states. The majority of flights departing from airports in these states are delayed.

California reports the largest proportion of flights that are not delayed.

Exercise 3: Fit a DForest model to the weather delay data

Question: What is the Out-Of-Box (OOB) error rate for the DForest model?

Answer: Answers will vary, but the OOB error rate should be between 2 percent and 3 percent.

Exercise 4: Fit a DForest model to the weather delay data

Question: Are there any discrepancies between flights being forecast as delayed versus those being forecast as on-time? If so, how could you adjust for this?

Answer: The rate at which flights are incorrectly predicted to be on time is around 1-1.5 percent (answers will vary). The rate at which flights are incorrectly predicted to be delayed by weather is around 4-6 percent (again, answers will vary). The model seems to have a bias towards predicting delays. If you have time, you can try building another forest with a loss matrix that adjusts this bias (use the **parms** argument to the **rxDForest** function). However, you might need to experiment with different values for this matrix, and there is always the possibility of overfitting the model to the training data as a result.

Exercise 5: Fit a DForest model to the weather delay data

Question: Which predictor variable had the most influence on the decisions made by the model?

Answer: The **OriginState** variable has the most influence, followed closely by **Month**. **DestState** has significantly less importance, although it participates in much of the decision making.

Exercise 6: Fit a DForest model to the weather delay data

Question: What does the ROC curve tell you about the possible accuracy of weather delay predictions? Is this what you expected?

Answer: The ROC curve should indicate that the model has low to moderate accuracy (better than random guesswork). Depending upon your understanding of statistics, the ability to predict weather at the local level, and the available data, this should probably not be surprising. All this model can do is give a general guide as to whether a flight is likely to be delayed by bad weather rather than make a definitive statement.

Exercise 7: Store the model in SQL Server

Question: According to the DForest model, what is the probability of a flight from Georgia (GA) to New York (NY) in November being delayed by weather? What about a flight in June?

> **Answer:** Answers will vary. The model should show an 80-90 percent probability of a weather delay in November, but only a 10-15 percent probability of a weather delay in May.

Lab B: Incorporating Hadoop Map/Reduce and Spark functionality into the ScaleR workflow

Question and Answers

Exercise 1: Using Pig with ScaleR functions

Question: According to the histogram that displays delays against frequency, what is the most common delay period across all airlines?

> **Answer:** Between 15 and 20 minutes. This delay period occurred more than 15,000 times.

Exercise 2: Using Pig with ScaleR functions

Question: Using the second histogram, which airline has had the most delayed flights?

> **Answer:** The airline with code **WN**.

Exercise 3: Using Pig with ScaleR functions

Question: Which route has the most frequent airline delays? How long is the average airline delay on this route?

> **Answer:** The route form HOU (Houston) to DAL (Dallas) has the most delays. The average airline delay is just over 40 minutes. Interestingly, the reverse route from Dallas to Houston has the second highest number of airline delays, averaging just under 40 minutes.