A6 Prediction Report

Pankaj Tripathi, Kartik Mahaley

February 07, 2016

This is a report generated from R for our assignment A6 Prediction.

For A6 we are resuing some part of our code in A2 like AirlineDetails.java, InsaneInputException.java and InvalidFormatException.java to gather the flight details from the files provided as input.

Machine used:

i5-8gb Mac machine and i5-8gb Linux machine

Code Analysis:

Using 3 jobs of mapper reducer to run in parellel:

In order to make a prediction on the test data we created a 3 jobs of map reduce to run in parellel. The first job included a map reduce which the read the history data and wrote the reducer data to the files. The key for the mapper is **{year, month}**. Hence we had 36 files created for 3 year data. For the second job running in parellel we considered the test data and reduced it to 12 files each based on year and month. In this case we comsidered the key{year, month}. For the third job we had read the validate data and wrote the output of the reducer to files. Here again our key was year and month. Once these files are created we fetch the output from these mapper-reducer jobs based on months. In this case we have 3 files for month of Janauary from 1995-1997 for history files and one file for validate data and one for test data. We have written a R scirpt which runs after the jobs are executed. It simply reads data from the 3 output folder for each mapper reducer and comapres them to give a result of false and true count along with confusion matrix.

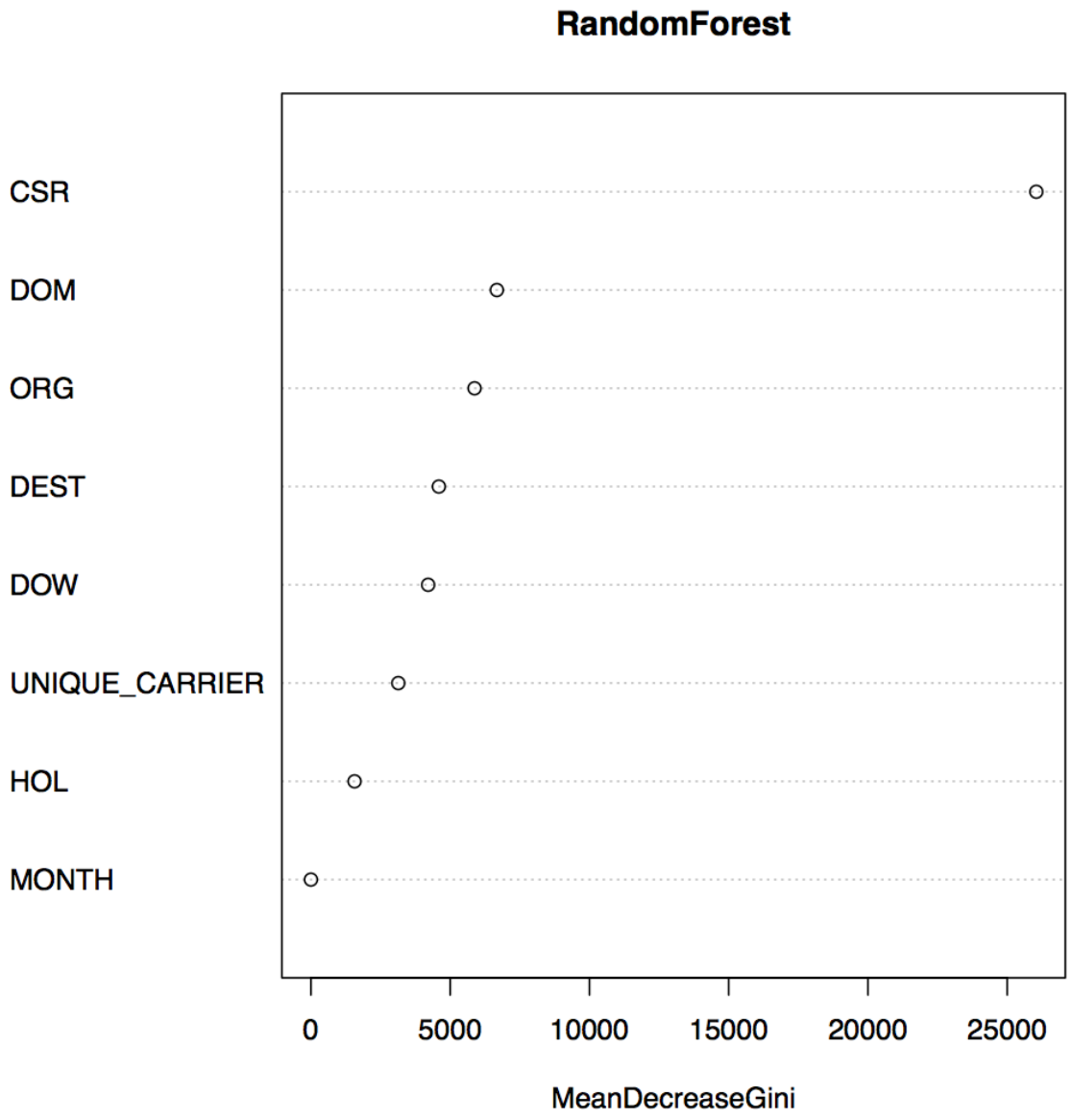
R Script:

In R we have used random forest to make prediction based on the model in place from the historical data. The columns that we have considered for the prediction are Year, Month which are key followed by Carrier, crsDepartureTime, Origin, Destination, Fldate, FlNum, DayOfMonth, DayOfWeek, Delay and DaysTillNextHoliday. For the holidays we considered some federal holidays and created al list of it. We compare each date with the list of holidays and get the nearest holiday to a particular date.

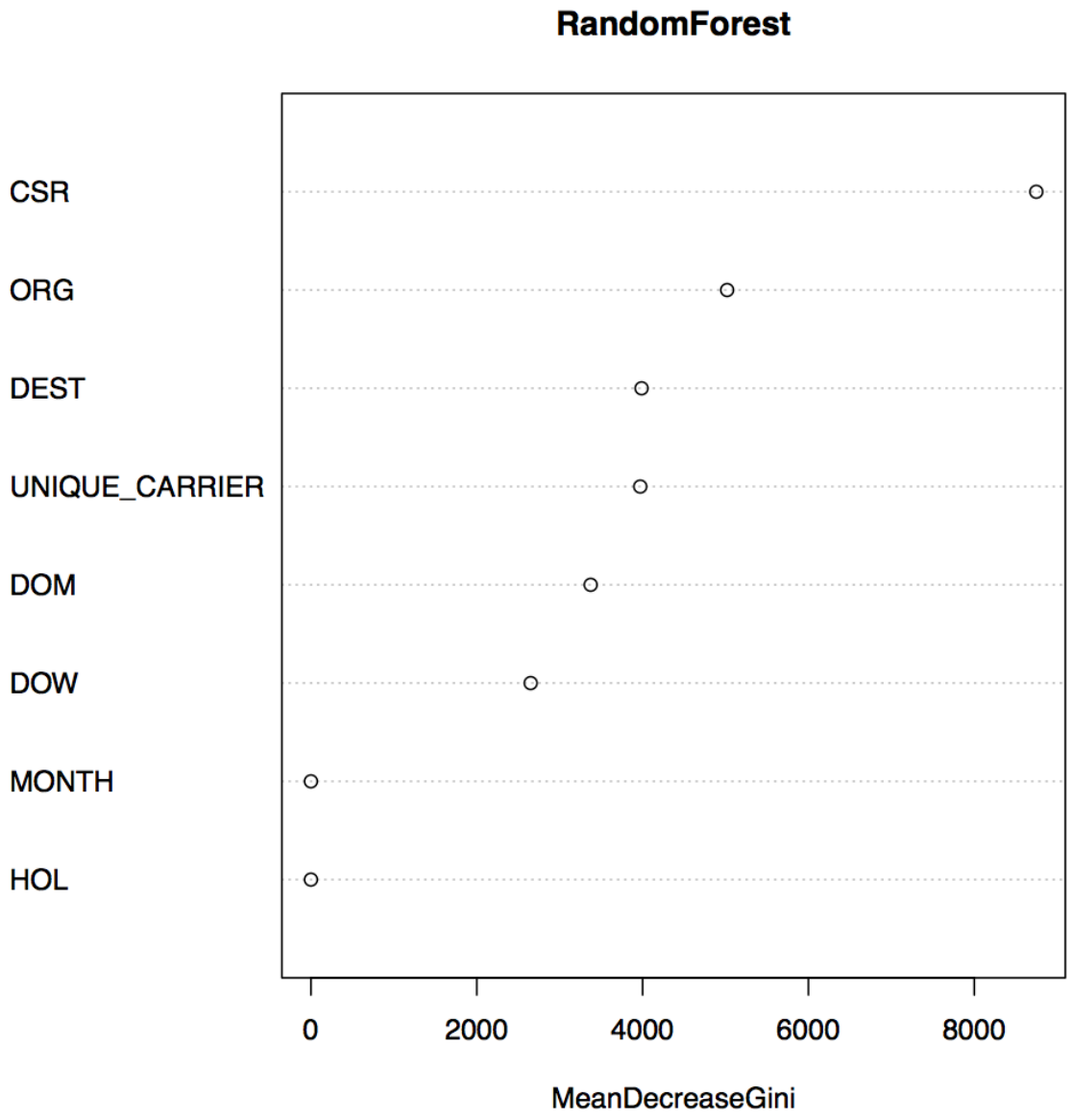
Execution time:

The execution time taken for this project is 8 mins in pseudo clustered mode.

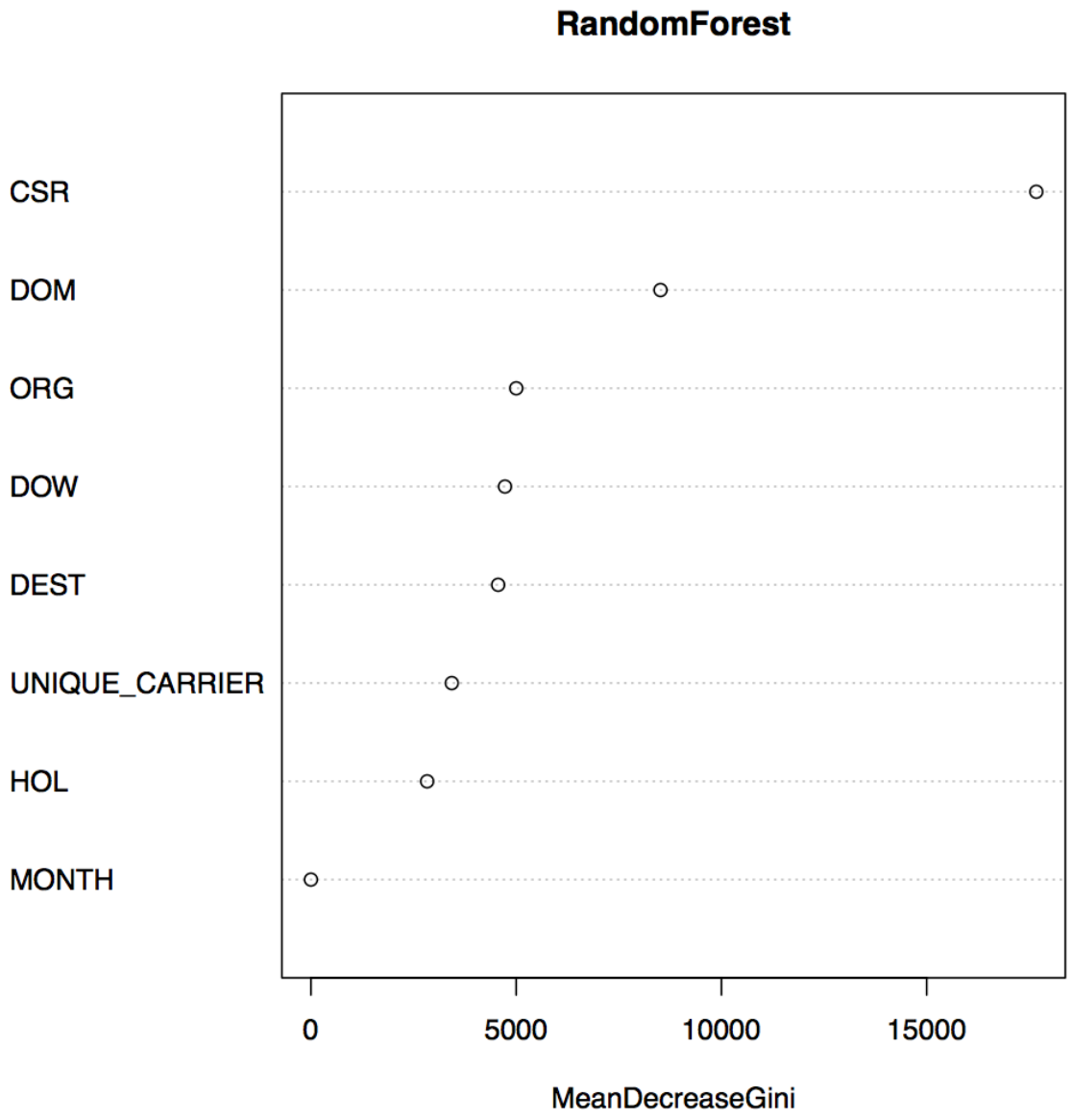
Graph for columns used in Janauary for modelling :



Graph for columns used in April for modelling :



Graph for columns used in December for modelling :



Monthwise Data for prediction count:

Month	Actual Count	True Count	False Count	T Pred for F	F Pred For T	pred Ontime	predLate	totalerror
Jan	188659	99289	89370	88969	333	0.9	0	0.47
Feb	200794	106052	94742	91375	2777	0.86	0.03	0.47
Mar	228631	120521	108110	104319	3567	0.87	0.03	0.47
Apr	221310	111236	110074	93671	12837	0.84	0.12	0.48
May	224157	117817	106340	102421	3036	0.87	0.03	0.47
June	217351	124357	92994	80947	12349	0.65	0.13	0.43
July	227828	112367	115461	98616	11390	0.88	0.1	0.48
Aug	232486	114626	117860	114398	2055	1	0.02	0.5
Sept	188659	111236	126233	121232	2511	1.31	0.02	0.57
Oct	218673	107290	117917	111900	4347	1.04	0.04	0.52
Nov	213765	94436	119329	111864	4358	1.18	0.04	0.54
Dec	203972	116921	87051	84673	1914	0.72	0.02	0.42