A5 - Routing

Jiangtao, Joyal and Bhanu

October 12, 2017

Contents

Prediction

After generating test/train data through the MR job we use KNN classifier to predict labels for the test routes. (Please note that label=1 is an invalid route and label=2 is a valid route.)

Train Data Set

The complete train dataset contains 23K observations, but due to compute bottleneck in R, we sampled the training to 10,000.Thus there are 10,000 rows, 8 features, and 1 label.

```
dim(model.train.data)
```

```
## [1] 10000     9
```

Below are the features we selected for the model. These features basically predict on the seasonal delay trend.

```
str(model.train.data)
```

```
## Classes 'data.table' and 'data.frame':   10000 obs. of  9 variables:
##  $ l1.month     : int  7 12 11 10 6 2 12 12 11 9 ...
##  $ l1.dayOfWeek : int  2 4 7 2 3 2 3 3 2 6 ...
##  $ l1.dayOfMonth: int  20 2 28 12 23 9 8 22 9 18 ...
##  $ l1.hourOfDay : int  9 10 8 8 15 17 7 12 7 21 ...
##  $ l2.month     : int  7 12 11 10 6 2 12 12 11 9 ...
##  $ l2.dayOfWeek : int  2 4 7 2 3 2 3 3 2 6 ...
##  $ l2.dayOfMonth: int  20 2 28 12 23 9 8 22 9 18 ...
##  $ l2.hourOfDay : int  17 10 18 10 12 17 15 14 12 9 ...
##  $ label        : Factor w/ 2 levels "1","2": 2 1 2 1 1 1 2 1 2 2 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Test Data Set

Training set has 877 observations distributed over 8 similar features. These are all the possible two hop routes for the given input query.

```
dim(model.test.data)
```

```
## [1] 877   8
```

```
str(model.test.data)
```

```
## Classes 'data.table' and 'data.frame':   877 obs. of  8 variables:
##  $ l1.month     : int  10 10 10 10 10 10 10 2 2 2 ...
##  $ l1.dayOfWeek : int  6 6 6 6 6 6 6 1 1 1 ...
##  $ l1.dayOfMonth: int  1 1 1 1 1 1 1 21 21 21 ...
##  $ l1.hourOfDay : int  17 18 17 7 13 8 12 17 17 17 ...
##  $ l2.month     : int  10 10 10 10 10 10 10 2 2 2 ...
##  $ l2.dayOfWeek : int  6 6 6 6 6 7 7 1 1 2 ...
##  $ l2.dayOfMonth: int  1 1 1 1 1 2 2 21 21 22 ...
##  $ l2.hourOfDay : int  9 9 9 9 9 9 9 8 7 7 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

KNN Classifier

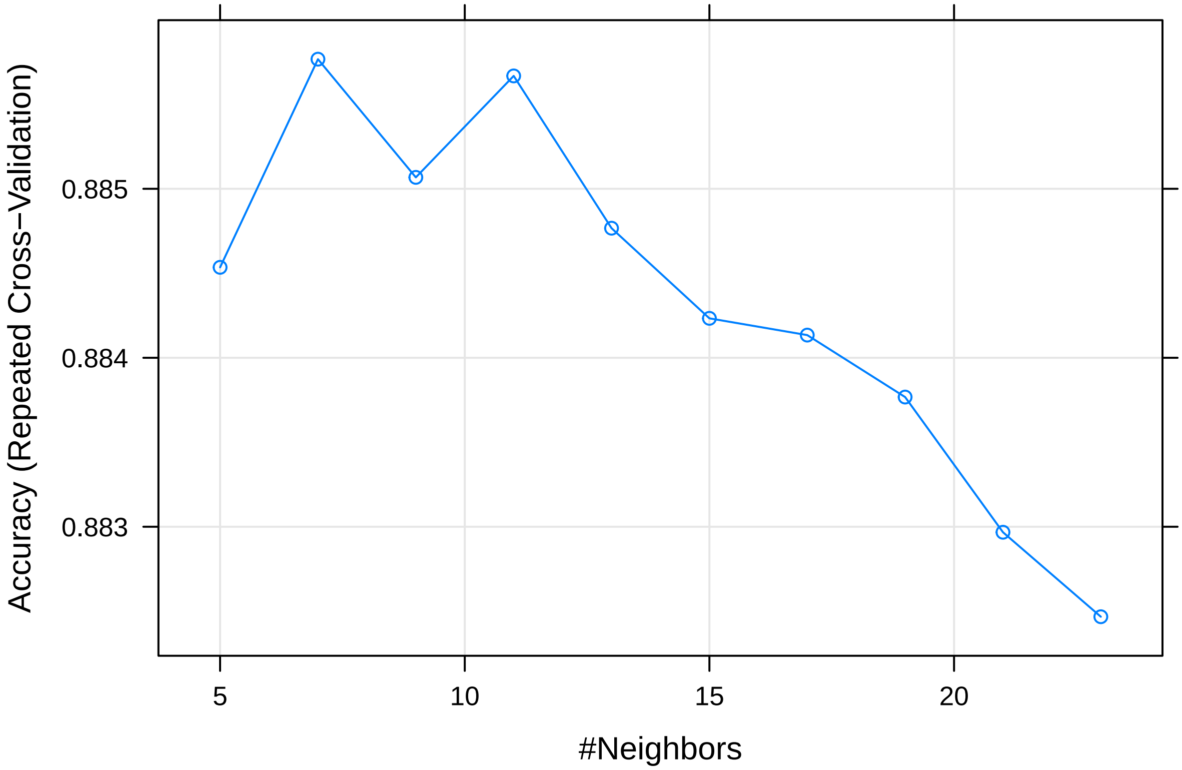
While training a classifier we are using reapeated cross validation to optimize the model parameters such as k(no of neighbors). Additionally there is a data normalization and scalling for features before feeding to the classifier.

```
# Model
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
model.knn <- train(label~., data = model.train.data, method = "knn"
 ,trControl=trctrl
 ,preProcess = c("center", "scale")
 ,tuneLength = 10
 )
model.knn
```

```
## k-Nearest Neighbors
##
## 10000 samples
##     8 predictors
##     2 classes: '1', '2'
##
## Pre-processing: centered (8), scaled (8)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 9000, 9000, 9001, 9000, 8999, 9000, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    5  0.8845352  0.7045393
##    7  0.8857669  0.7058633
##    9  0.8850676  0.7024658
##   11  0.8856679  0.7029536
##   13  0.8847668  0.6999338
##   15  0.8842334  0.6974951
##   17  0.8841341  0.6964547
##   19  0.8837675  0.6949189
##   21  0.8829675  0.6920803
##   23  0.8824675  0.6902165
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was k = 7.
```

Plot1. KNN Neighbors vs Accuracy

```
plot(model.knn)
```



`Plot 1.` shows accuracy per k. With parameter optimization we found the that `k=7` is giving best accuracy of `0.8857669`.

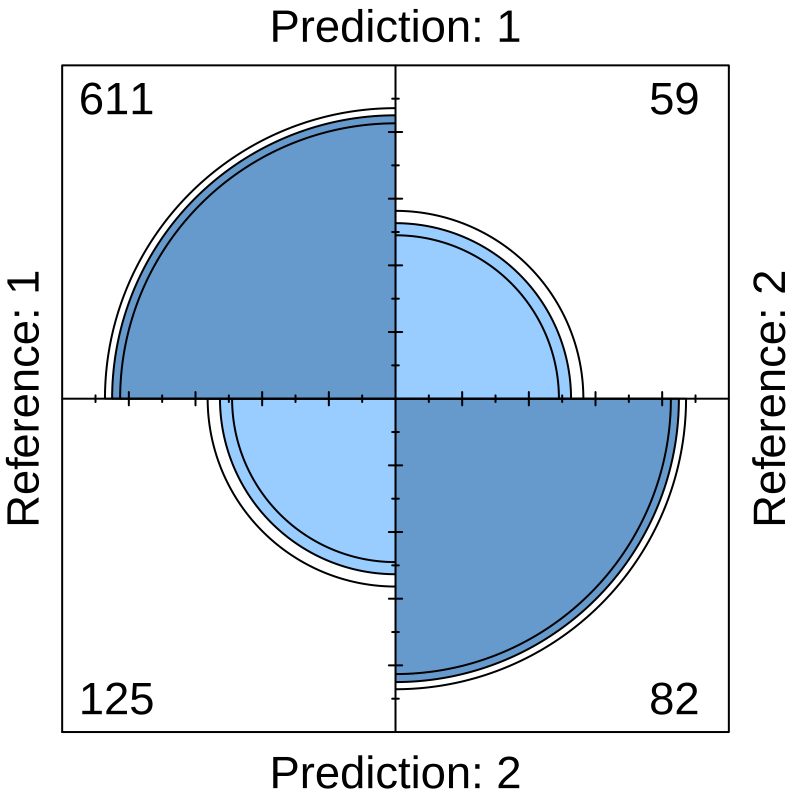
Confusion Matrix

```
model.test.cf
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2
##          1 611  59
##          2 125  82
##
##                Accuracy : 0.7902
##                  95% CI : (0.7617, 0.8167)
##     No Information Rate : 0.8392
##     P-Value [Acc > NIR] : 0.9999
##
```

```
##                    Kappa : 0.3462
##  Mcnemar's Test P-Value : 1.652e-06
##
##             Sensitivity : 0.8302
##             Specificity : 0.5816
##          Pos Pred Value : 0.9119
##          Neg Pred Value : 0.3961
##              Prevalence : 0.8392
##          Detection Rate : 0.6967
##    Detection Prevalence : 0.7640
##       Balanced Accuracy : 0.7059
##
##        'Positive' Class : 1
##
```

```
fourfoldplot(model.test.cf$table)
```



Recall

```
model.test.recall
```

```
## [1] 0.830163
```

Precision

```
model.test.precision
```

```
## [1] 0.9119403
```

Plot for KNN Labels

```
# p <- ggplot(classify.results.sample, aes(logpr.a, logpr.b))
# p + geom_point(aes(colour = factor(item.class))) + geom_abline(intercept=0, slope=1)
```

Results

Input Queries

year	month	day	origin	destination
2011	10	1	BOS	SEA
2011	16	11	SEA	LAX
2011	12	24	LAX	BOS
2011	11	21	DEN	JFK
2011	90	1	DEN	DCA
2011	8	4	DCA	LAX
2011	6	15	DCA	BOS
2011	4	11	BOS	DEN
2011	2	21	BOS	DCA
2011	1	7	LAX	DEN

Ouput Routes

flightDate	origin	des	l1.actDepTime	l1.actArrTime	l1.carrier	l1.origin	l1.dest	l2.actDepTime	l2.actArrTi
20111001	BOS	SEA	658	943	AA	BOS	DFW	1852	21
20111001	BOS	SEA	559	854	AA	BOS	DFW	1852	21
20110411	BOS	DEN	604	908	B6	BOS	FLL	1441	17
20110411	BOS	DEN	604	908	B6	BOS	FLL	1458	17
20111001	BOS	SEA	723	1036	CO	BOS	IAH	1441	17
20110411	BOS	DEN	820	1157	UA	BOS	LAX	2053	23
20110411	BOS	DEN	820	1157	UA	BOS	LAX	1530	18
20110411	BOS	DEN	820	1157	UA	BOS	LAX	2019	23
20110411	BOS	DEN	820	1157	UA	BOS	LAX	2011	23
20110411	BOS	DEN	820	1157	UA	BOS	LAX	1739	20
20110411	BOS	DEN	820	1157	UA	BOS	LAX	1727	20
20110411	BOS	DEN	820	1157	UA	BOS	LAX	1527	18
20110411	BOS	DEN	820	1157	UA	BOS	LAX	1825	21
20110411	BOS	DEN	820	1157	UA	BOS	LAX	1828	21
20110411	BOS	DEN	820	1157	UA	BOS	LAX	2050	23
20110411	BOS	DEN	820	1157	UA	BOS	LAX	1928	22
20110411	BOS	DEN	820	1157	UA	BOS	LAX	1556	18
20110411	BOS	DEN	820	1157	UA	BOS	LAX	1949	22
20110411	BOS	DEN	820	1157	UA	BOS	LAX	1540	18
20110411	BOS	DEN	909	1237	B6	BOS	LAX	2053	23
20110411	BOS	DEN	909	1237	B6	BOS	LAX	1530	18
20110411	BOS	DEN	909	1237	B6	BOS	LAX	2019	23

flightDate	origin	des	l1.actDepTime	l1.actArrTime	l1.carrier	l1.origin	l1.dest	l2.actDepTime	l2.actArrTi
20110411	BOS	DEN	909	1237	B6	BOS	LAX	2011	23
20110411	BOS	DEN	909	1237	B6	BOS	LAX	1739	20
20110411	BOS	DEN	909	1237	B6	BOS	LAX	1727	20
20110411	BOS	DEN	909	1237	B6	BOS	LAX	1527	18
20110411	BOS	DEN	909	1237	B6	BOS	LAX	1825	21
20110411	BOS	DEN	909	1237	B6	BOS	LAX	1828	21
20110411	BOS	DEN	909	1237	B6	BOS	LAX	2050	23
20110411	BOS	DEN	909	1237	B6	BOS	LAX	1928	22
20110411	BOS	DEN	909	1237	B6	BOS	LAX	1556	18
20110411	BOS	DEN	909	1237	B6	BOS	LAX	1949	22
20110411	BOS	DEN	909	1237	B6	BOS	LAX	1540	18
20111001	BOS	SEA	726	1006	AA	BOS	LAX	1945	22
20111001	BOS	SEA	726	1006	AA	BOS	LAX	1555	18
20111001	BOS	SEA	724	1027	B6	BOS	LAX	1945	22
20111001	BOS	SEA	724	1027	B6	BOS	LAX	1555	18
20110411	BOS	DEN	559	838	B6	BOS	MCO	1755	20
20110411	BOS	DEN	1124	1425	FL	BOS	MCO	1755	20
20110411	BOS	DEN	717	1009	DL	BOS	MCO	1755	20
20110411	BOS	DEN	959	1255	B6	BOS	MCO	1755	20
20111001	BOS	SEA	638	940	B6	BOS	MCO	1808	21
20111001	BOS	SEA	638	940	B6	BOS	MCO	1823	21
20111001	BOS	SEA	1121	1438	FL	BOS	MCO	1808	21
20111001	BOS	SEA	1121	1438	FL	BOS	MCO	1823	21
20111001	BOS	SEA	830	1148	B6	BOS	MCO	1808	21
20111001	BOS	SEA	830	1148	B6	BOS	MCO	1823	21
20111001	BOS	SEA	1324	1629	B6	BOS	MCO	1808	21
20110411	BOS	DEN	559	857	AA	BOS	MIA	1656	19
20110411	BOS	DEN	559	857	AA	BOS	MIA	1655	19
20110411	BOS	DEN	900	1210	AA	BOS	MIA	1656	19
20110411	BOS	DEN	900	1210	AA	BOS	MIA	1655	19
20110411	BOS	DEN	703	1004	AA	BOS	MIA	1656	19
20110411	BOS	DEN	703	1004	AA	BOS	MIA	1655	19
20110411	BOS	DEN	607	915	US	BOS	PHX	1353	16
20110411	BOS	DEN	607	915	US	BOS	PHX	1338	16
20110411	BOS	DEN	607	915	US	BOS	PHX	1450	17
20110411	BOS	DEN	607	915	US	BOS	PHX	1846	21
20110411	BOS	DEN	607	915	US	BOS	PHX	2151	
20110411	BOS	DEN	607	915	US	BOS	PHX	1846	21
20110411	BOS	DEN	607	915	US	BOS	PHX	1451	17
20110411	BOS	DEN	607	915	US	BOS	PHX	1731	20
20110411	BOS	DEN	607	915	US	BOS	PHX	1731	20
20110411	BOS	DEN	607	915	US	BOS	PHX	1443	17
20110411	BOS	DEN	607	915	US	BOS	PHX	2156	
20110411	BOS	DEN	607	915	US	BOS	PHX	1738	20
20110411	BOS	DEN	607	915	US	BOS	PHX	2153	
20110411	BOS	DEN	607	915	US	BOS	PHX	1528	18
20110411	BOS	DEN	607	915	US	BOS	PHX	1750	20
20110411	BOS	DEN	607	915	US	BOS	PHX	1455	17
20110411	BOS	DEN	607	915	US	BOS	PHX	2142	
20110411	BOS	DEN	607	915	US	BOS	PHX	1328	16
20110411	BOS	DEN	607	915	US	BOS	PHX	1544	18
20110411	BOS	DEN	600	937	UA	BOS	SFO	1931	22

flightDate	origin	des	l1.actDepTime	l1.actArrTime	l1.carrier	l1.origin	l1.dest	l2.actDepTime	l2.actArrTi
20110411	BOS	DEN	600	937	UA	BOS	SFO	1304	16
20110411	BOS	DEN	600	937	UA	BOS	SFO	1929	22
20110411	BOS	DEN	600	937	UA	BOS	SFO	1300	16
20110411	BOS	DEN	1114	1438	UA	BOS	SFO	1931	22
20110411	BOS	DEN	1114	1438	UA	BOS	SFO	1929	22
20110411	BOS	DEN	816	1148	UA	BOS	SFO	1931	22
20110411	BOS	DEN	816	1148	UA	BOS	SFO	1929	22
20110615	DCA	BOS	1153	1409	US	DCA	MCO	1854	21
20110615	DCA	BOS	1153	1409	US	DCA	MCO	1805	21
20110615	DCA	BOS	1153	1409	US	DCA	MCO	1823	21
20110615	DCA	BOS	1526	1824	AA	DCA	MIA	2109	
20110615	DCA	BOS	1526	1824	AA	DCA	MIA	2118	
20110615	DCA	BOS	841	1104	AA	DCA	MIA	1730	20
20110615	DCA	BOS	841	1104	AA	DCA	MIA	1752	20
20110615	DCA	BOS	841	1104	AA	DCA	MIA	2109	
20110615	DCA	BOS	841	1104	AA	DCA	MIA	2118	
20110615	DCA	BOS	841	1104	AA	DCA	MIA	1326	16
20110615	DCA	BOS	1146	1410	AA	DCA	MIA	1730	20
20110615	DCA	BOS	1146	1410	AA	DCA	MIA	1752	20
20110615	DCA	BOS	1146	1410	AA	DCA	MIA	2109	
20110615	DCA	BOS	1146	1410	AA	DCA	MIA	2118	
20110615	DCA	BOS	1249	1518	AA	DCA	MIA	1730	20
20110615	DCA	BOS	1249	1518	AA	DCA	MIA	1752	20
20110615	DCA	BOS	1249	1518	AA	DCA	MIA	2109	
20110615	DCA	BOS	1249	1518	AA	DCA	MIA	2118	
20111121	DEN	JFK	1638	2215	WN	DEN	FLL	2143	
20111121	DEN	JFK	600	923	AA	DEN	ORD	1300	16
20111121	DEN	JFK	600	923	AA	DEN	ORD	1927	22
20111121	DEN	JFK	855	1156	UA	DEN	ORD	1300	16
20111121	DEN	JFK	855	1156	UA	DEN	ORD	1927	22
20110107	LAX	DEN	718	1022	OO	LAX	ASE	1201	12
20110107	LAX	DEN	2231	700	UA	LAX	BOS	954	12
20110107	LAX	DEN	2340	827	B6	LAX	BOS	756	10
20110107	LAX	DEN	2340	827	B6	LAX	BOS	954	12
20110107	LAX	DEN	2340	827	B6	LAX	BOS	834	11
20111224	LAX	BOS	1508	1839	WN	LAX	DEN	1749	23
20111224	LAX	BOS	702	1041	UA	LAX	DEN	1627	22
20111224	LAX	BOS	702	1041	UA	LAX	DEN	1431	20
20111224	LAX	BOS	702	1041	UA	LAX	DEN	1434	20
20111224	LAX	BOS	702	1041	UA	LAX	DEN	1749	23
20111224	LAX	BOS	1509	1837	UA	LAX	DEN	1749	23
20111224	LAX	BOS	2100	23	MQ	LAX	DEN	1035	16
20111224	LAX	BOS	2100	23	MQ	LAX	DEN	1031	16
20111224	LAX	BOS	2100	23	MQ	LAX	DEN	1031	16
20111224	LAX	BOS	2100	23	MQ	LAX	DEN	1057	16
20111224	LAX	BOS	1117	1717	WN	LAX	MDW	1622	19
20111224	LAX	BOS	1117	1717	WN	LAX	MDW	2014	23
20111224	LAX	BOS	845	1444	WN	LAX	MDW	1622	19
20111224	LAX	BOS	845	1444	WN	LAX	MDW	2014	23
20111224	LAX	BOS	845	1444	WN	LAX	MDW	1423	17
20111224	LAX	BOS	1336	1941	WN	LAX	MDW	2014	23
20111224	LAX	BOS	704	1302	WN	LAX	MDW	1622	19

flightDate	origin	des	l1.actDepTime	l1.actArrTime	l1.carrier	l1.origin	l1.dest	l2.actDepTime	l2.actArrTi
20111224	LAX	BOS	704	1302	WN	LAX	MDW	2014	23
20111224	LAX	BOS	704	1302	WN	LAX	MDW	1423	17
20111224	LAX	BOS	1143	1742	FL	LAX	MKE	1841	21
20111224	LAX	BOS	1143	1742	FL	LAX	MKE	1836	21
20111224	LAX	BOS	745	1330	DL	LAX	MSP	1305	16
20111224	LAX	BOS	745	1330	DL	LAX	MSP	1301	16
20111224	LAX	BOS	1438	2030	AA	LAX	ORD	1811	21
20111224	LAX	BOS	1438	2030	AA	LAX	ORD	1818	21
20111224	LAX	BOS	1438	2030	AA	LAX	ORD	1941	22
20111224	LAX	BOS	759	1346	AA	LAX	ORD	1732	20
20111224	LAX	BOS	759	1346	AA	LAX	ORD	1732	20
20111224	LAX	BOS	759	1346	AA	LAX	ORD	1811	21
20111224	LAX	BOS	759	1346	AA	LAX	ORD	1818	21
20111224	LAX	BOS	759	1346	AA	LAX	ORD	1314	16
20111224	LAX	BOS	759	1346	AA	LAX	ORD	1602	19
20111224	LAX	BOS	759	1346	AA	LAX	ORD	1941	22
20111224	LAX	BOS	759	1346	AA	LAX	ORD	1319	16
20111224	LAX	BOS	759	1346	AA	LAX	ORD	1512	18
20111224	LAX	BOS	933	1524	AA	LAX	ORD	1732	20
20111224	LAX	BOS	933	1524	AA	LAX	ORD	1732	20
20111224	LAX	BOS	933	1524	AA	LAX	ORD	1811	21
20111224	LAX	BOS	933	1524	AA	LAX	ORD	1818	21
20111224	LAX	BOS	933	1524	AA	LAX	ORD	1602	19
20111224	LAX	BOS	933	1524	AA	LAX	ORD	1941	22
20111224	LAX	BOS	933	1524	AA	LAX	ORD	1512	18
20111224	LAX	BOS	2316	525	AA	LAX	ORD	1203	15
20111224	LAX	BOS	2316	525	AA	LAX	ORD	657	9
20111224	LAX	BOS	2316	525	AA	LAX	ORD	853	11
20111224	LAX	BOS	2316	525	AA	LAX	ORD	1314	16
20111224	LAX	BOS	2316	525	AA	LAX	ORD	1324	16
20111224	LAX	BOS	2316	525	AA	LAX	ORD	1014	13
20111224	LAX	BOS	2316	525	AA	LAX	ORD	1016	13
20111224	LAX	BOS	2316	525	AA	LAX	ORD	809	11
20111224	LAX	BOS	2316	525	AA	LAX	ORD	1602	19
20111224	LAX	BOS	2316	525	AA	LAX	ORD	1319	16
20111224	LAX	BOS	2316	525	AA	LAX	ORD	1329	16
20111224	LAX	BOS	2316	525	AA	LAX	ORD	1512	18
20111224	LAX	BOS	2316	525	AA	LAX	ORD	905	12
20111224	LAX	BOS	2316	525	AA	LAX	ORD	917	12
20111224	LAX	BOS	1231	1828	AA	LAX	ORD	1732	20
20111224	LAX	BOS	1231	1828	AA	LAX	ORD	1732	20
20111224	LAX	BOS	1231	1828	AA	LAX	ORD	1811	21
20111224	LAX	BOS	1231	1828	AA	LAX	ORD	1818	21
20111224	LAX	BOS	1231	1828	AA	LAX	ORD	1602	19
20111224	LAX	BOS	1231	1828	AA	LAX	ORD	1941	22
20111224	LAX	BOS	1020	1616	UA	LAX	ORD	1732	20
20111224	LAX	BOS	1020	1616	UA	LAX	ORD	1732	20
20111224	LAX	BOS	1020	1616	UA	LAX	ORD	1811	21
20111224	LAX	BOS	1020	1616	UA	LAX	ORD	1818	21
20111224	LAX	BOS	1020	1616	UA	LAX	ORD	1602	19
20111224	LAX	BOS	1020	1616	UA	LAX	ORD	1941	22
20111224	LAX	BOS	1020	1616	UA	LAX	ORD	1512	18

flightDate	origin	des	l1.actDepTime	l1.actArrTime	l1.carrier	l1.origin	l1.dest	l2.actDepTime	l2.actArrTi
20111224	LAX	BOS	1419	2019	UA	LAX	ORD	1811	21
20111224	LAX	BOS	1419	2019	UA	LAX	ORD	1818	21
20111224	LAX	BOS	1419	2019	UA	LAX	ORD	1941	22
20111224	LAX	BOS	1215	1804	UA	LAX	ORD	1732	20
20111224	LAX	BOS	1215	1804	UA	LAX	ORD	1732	20
20111224	LAX	BOS	1215	1804	UA	LAX	ORD	1811	21
20111224	LAX	BOS	1215	1804	UA	LAX	ORD	1818	21
20111224	LAX	BOS	1215	1804	UA	LAX	ORD	1602	19
20111224	LAX	BOS	1215	1804	UA	LAX	ORD	1941	22
20110107	LAX	DEN	608	659	OO	LAX	SAN	1247	15
20110107	LAX	DEN	608	659	OO	LAX	SAN	1157	14
20110107	LAX	DEN	608	659	OO	LAX	SAN	1528	18
20110107	LAX	DEN	608	659	OO	LAX	SAN	1513	18
20110107	LAX	DEN	608	659	OO	LAX	SAN	1722	20
20110107	LAX	DEN	608	659	OO	LAX	SAN	1242	15
20110107	LAX	DEN	608	659	OO	LAX	SAN	1509	18
20110107	LAX	DEN	608	659	OO	LAX	SAN	2001	23
20110107	LAX	DEN	608	659	OO	LAX	SAN	1923	22
20110107	LAX	DEN	1006	1053	MQ	LAX	SAN	1513	18
20110107	LAX	DEN	1006	1053	MQ	LAX	SAN	1722	20
20110107	LAX	DEN	1006	1053	MQ	LAX	SAN	2001	23
20110107	LAX	DEN	1006	1053	MQ	LAX	SAN	1923	22
20111224	LAX	BOS	2141	6	AS	LAX	SEA	857	17
20111224	LAX	BOS	2141	6	AS	LAX	SEA	853	17
20110107	LAX	DEN	740	1018	OO	LAX	SEA	1303	16
20110107	LAX	DEN	740	1018	OO	LAX	SEA	1128	14
20110107	LAX	DEN	740	1018	OO	LAX	SEA	1109	14
20110107	LAX	DEN	740	1018	OO	LAX	SEA	1303	16
20110107	LAX	DEN	740	1018	OO	LAX	SEA	1302	16

Scores

Score	-10499
Total Valid Route Count	207
Delay Route Count	106
NonDelay Route Count	101

Job Execution

Psuedo Distributed

Used the below machine, to run the job in Psuedo Distributed mode.

```
OS: OSX
Processor Name: Intel Core i7
Processor Speed:    2.8 GHz
Number of Processors:   1
Total Number of Cores:  4
L2 Cache (per Core):    256 KB
```

```
L3 Cache:   6 MB
Memory: 16 GB
SSD: 256 GB
```

The job took 14.5m to run on the complete corpus. Below are some important observations,

- Input of 6.55 GB(for 337 items) was reduced to train-data(41.6 MB) and test-data(158.6 KB)
- Total input files to process : 337 and Number of splits:337. Per file per mapper which is expected default behavior.

AWS EMR

Ran the same job on 4 cluster m4.xlarge EMR. The entire corpus took 10.4m to run. There is not much improvement because the data not big enough to produce any noticeable results. Also there is network i/o between distributed mappers and reducer causing some delay.