Jianming Dong
Haiyang Yu

**Conclusion: The highest chance of an on time flight is on Tuesday-Saturday before 15pm, on Delta, Comair, United, or USAirways. (And of course a good weather).**

**Logistic regression:**

The probability of delay =

$$\frac{1}{1+e^{-(-2.8025+17.9209*\text{weather}+0.6256*\text{peaktime}+0.6743*\text{mon\_sun}+1.1145*\text{carrier\_co\_mq\_dh\_ru})}}$$

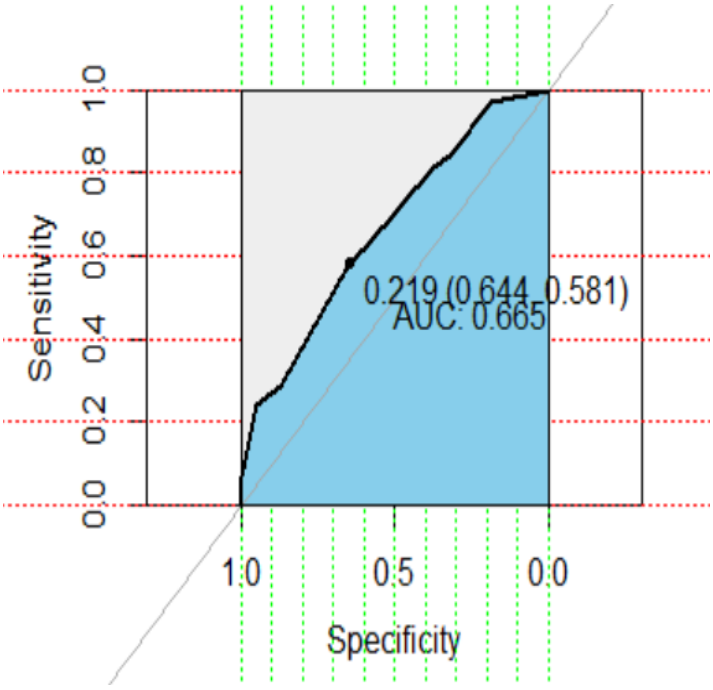
$$\text{Log(odds)} = -2.8025 + 17.9209 * \text{weather} + 0.6256 * \text{peaktime} + 0.6743 * \text{mon\_sun} + 1.1145 * \text{carrier\_co\_mq\_dh\_ru}$$

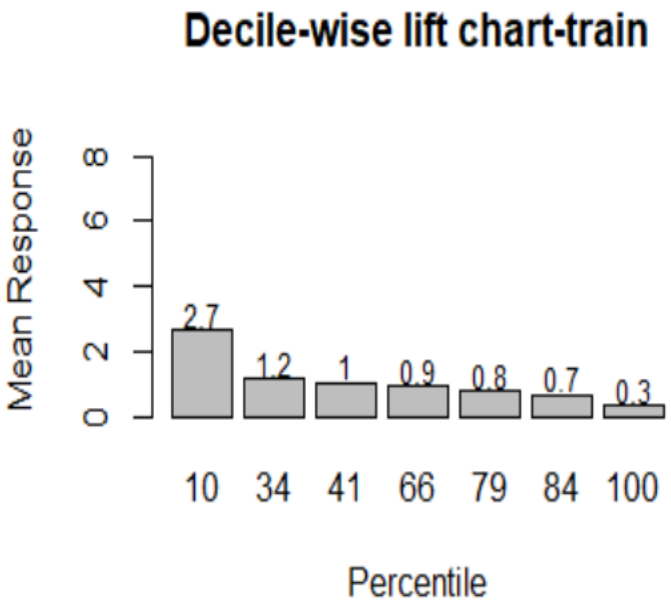
Weather: If bad weather, 1; else, 0.
Peaktime: If 15pm – 21pm, 1; else,0.
Mon_sun: If in Monday or Sunday, 1; else,0.
Carrier_CO_MQ_DH_RU: If in carrier CO,MQ,DH,RU, 1;else, 0.

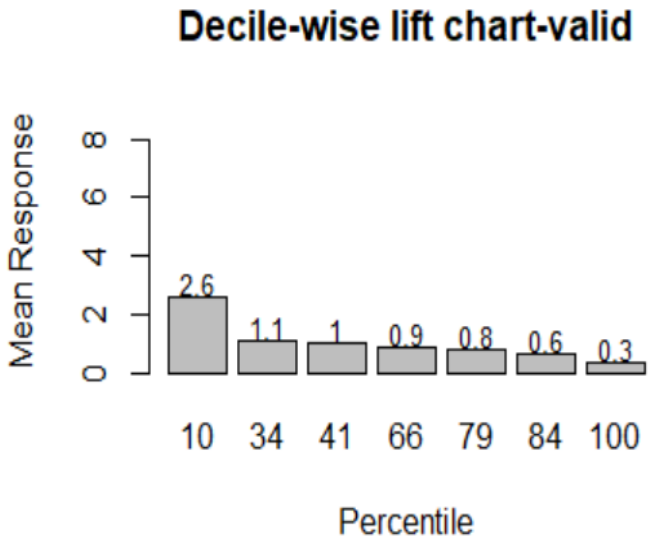
Coefficients:

| | Estimate | Std. Error | z-value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -2.8025 | 0.1865 | -15.031 | < 0.0000000000000002 *** |
| weather | 17.9209 | 503.1677 | 0.036 | 0.972 |
| peaktime | 0.6256 | 0.1513 | 4.135 | 0.000035485705 *** |
| Mon_Sun | 0.6743 | 0.1581 | 4.266 | 0.000019910414 *** |
| carrier_CO_MQ_DH_RU | 1.1145 | 0.1764 | 6.320 | 0.000000000262 *** |

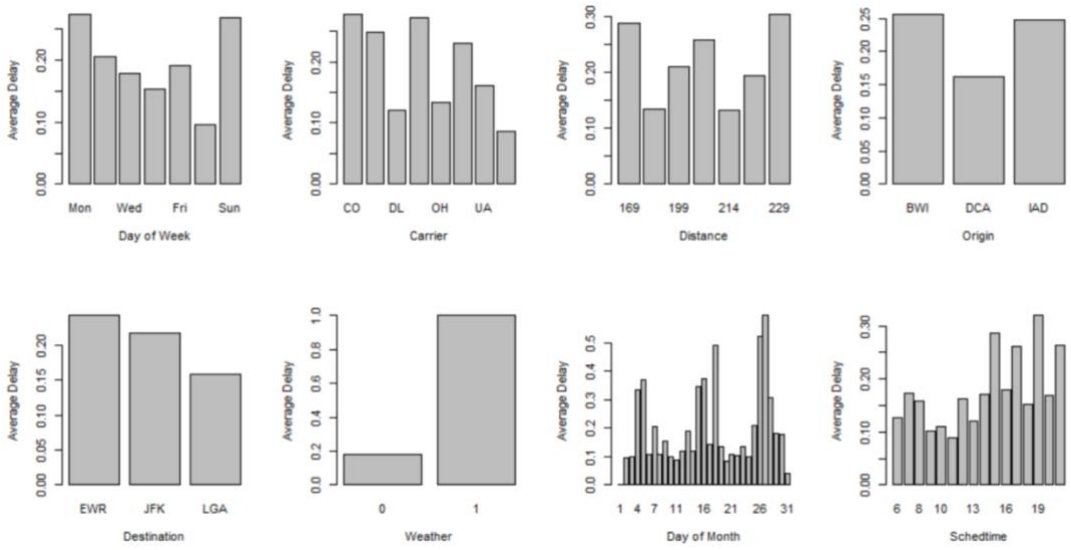
AIC: 1167.8 Null deviance = 1312.7 Residual deviance = 1157.8
1-pchisq = 0.54 1-pchisq = 0.99






```
Confusion Matrix for logistic regression - valid
          Reference
Prediction   0   1
         0 662 129
         1  44  46
              Accuracy : 0.8036
```




| | BWI | DCA | IAD | Total |
|---|---|---|---|---|
| CO | | 94 | | 94 |
| DH | | 27 | 524 | 551 |
| DL | | 388 | | 388 |
| MQ | | 295 | | 295 |
| OH | 30 | | | 30 |
| RU | 115 | 162 | 131 | 408 |
| UA | | | 31 | 31 |
| US | | 404 | | 404 |
| Total | 145 | 1370 | 686 | 2201 |

| RU | ontime | delayed | Total | Rate of delayed |
|---|---|---|---|---|
| BWI | 82 | 33 | 115 | 0.29 |
| DCA | 134 | 28 | 162 | 0.17 |
| IAD | 98 | 33 | 131 | 0.25 |
| Total | 314 | 94 | 408 | 0.23 |

Figure4. Origin analysis

Remove "distance","deptime","dest","date","flight_number","origin","tailnu"

**Model 1**

Input variables: Shedtime(7am-21pm),carrier(7 in total),weather,dayweek(Monday-Sunday),daymonth(1-30)

AIC:1131.8

Remove “daymonth”

**Model 2(no daymonth)**

Input variables: Shedtime,carrier,weather,dayweek

AIC:1192.7

Grouping “schedtime”,”dayweek”,”carrier”

**Model 3(grouping)**

Input variables: Weather,morning,noon,afternoon,Mon_sun,carrier_co_mq_dh_ru

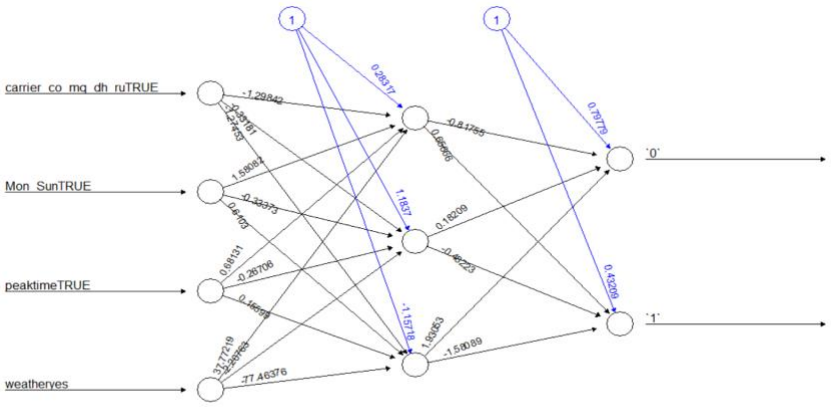
AIC:1174

Re-Grouping “shedtime”

**Model4(current model)**

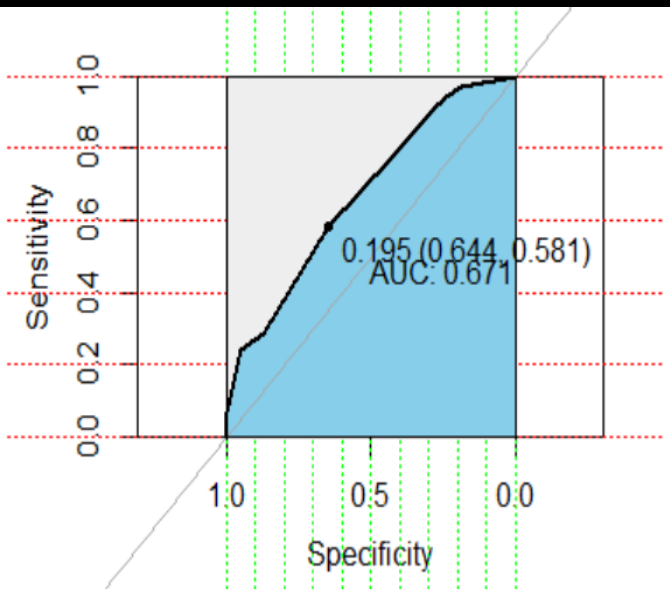
AIC:1167.8

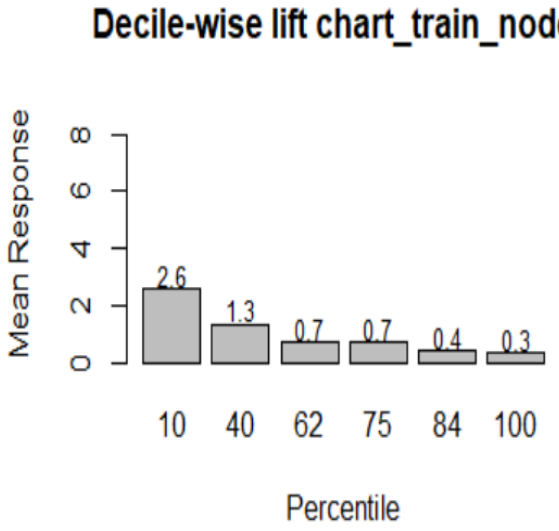
**Neural net (nodes = 3):**

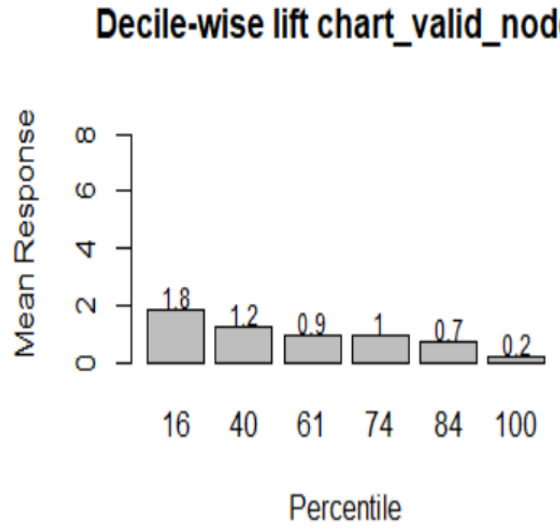



Error: 180.577485 Steps: 892

```
Confusion Matrix for nn_nodes = 3 - train
          Reference
Prediction    0    1
         0 1067  233
         1    0   20
              Accuracy : 0.8235
Confusion Matrix for nn_nodes = 3 - valid
          Reference
Prediction   0   1
         0 706 163
         1   0  12
              Accuracy : 0.815
```



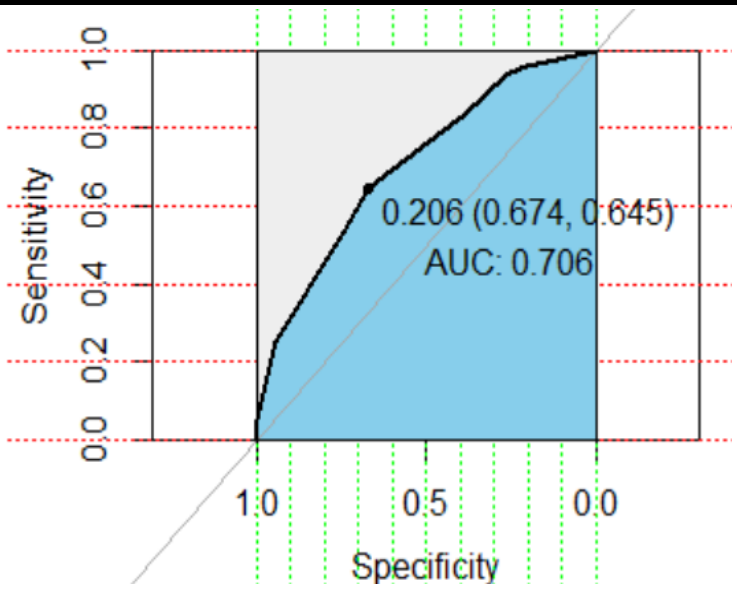


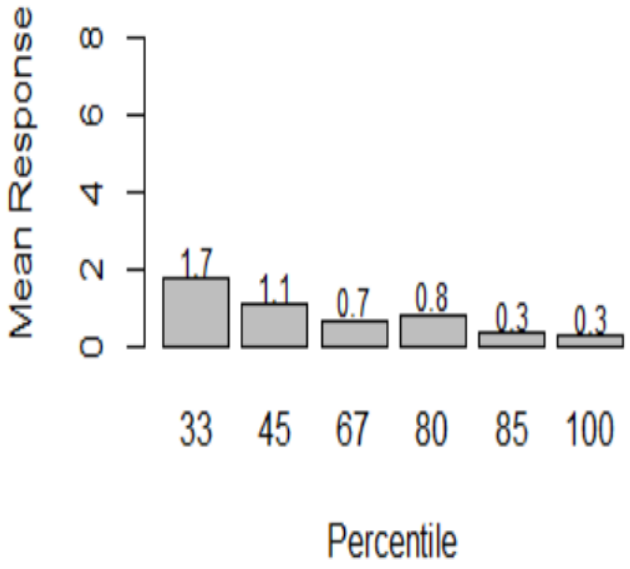



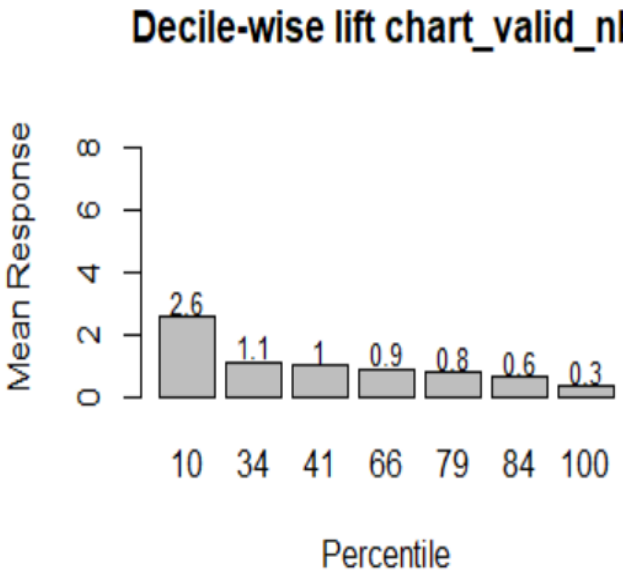

**Naïve bayes:**

```
Priori probabilities
Ontime(0)   delayed(1)
0.8022727 0.1977273
Confusion Matrix for naïve bayes - train
         Reference
Prediction    0    1
        0 1067  233
        1    0   20
             Accuracy : 0.8235
Confusion Matrix for naïve bayes - valid
         Reference
Prediction   0   1
        0 706 163
        1   0  12
```

```
Accuracy : 0.815
```




Decile-wise lift chart_valid_nb
Mean Response
0
2
4
6
8
2.6
1.1
1
0.9
0.8
0.6
0.3
10
34
41
66
79
84
100
Percentile