ADSP 31014 Statistical Models for Data Science

Course Project Part 2

Business Problem

The Chicago Department of Transportation (CDOT) is interested in understanding how Chicago Divvy bike trip duration is related to other factors of the user, the trip, and the weather in Chicago.

Linear Regression Model

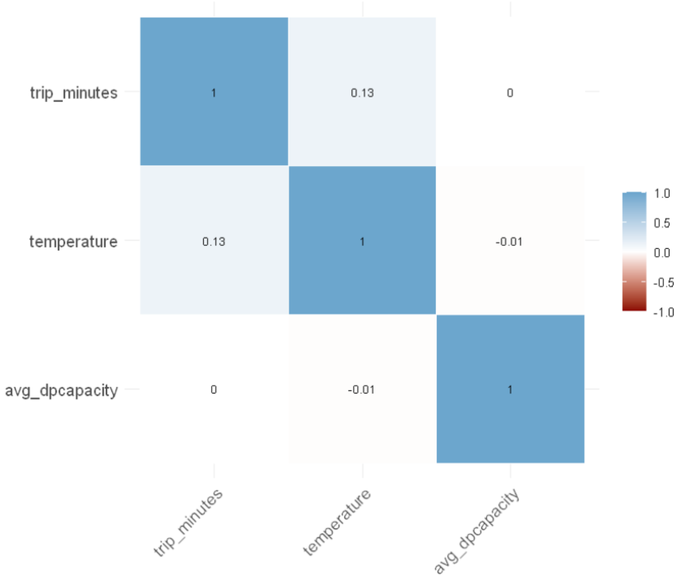
We build a linear regression model using a cleaned random sample of 100,000 historical Chicago Divvy bike data from year 2014-2017. We use trip_minutes as the response variable and other useful information as explanatory variables. The model formula is

```
trip_minutes ~ 0 + temperature * factor(hour) + factor(month) + factor(day) +
area_start * area_end + factor(gender) + factor(events) + avg_dpcapacity
```

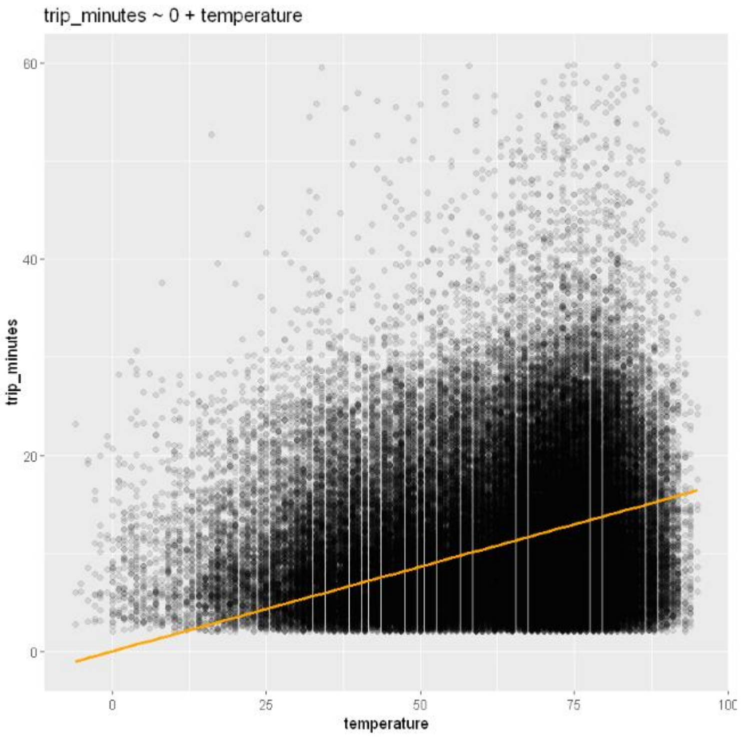
This model has 87 parameters with $R^2 = 0.7767$. See Appendix 1 for model summary and estimated coefficients.

Key Considerations in Modeling

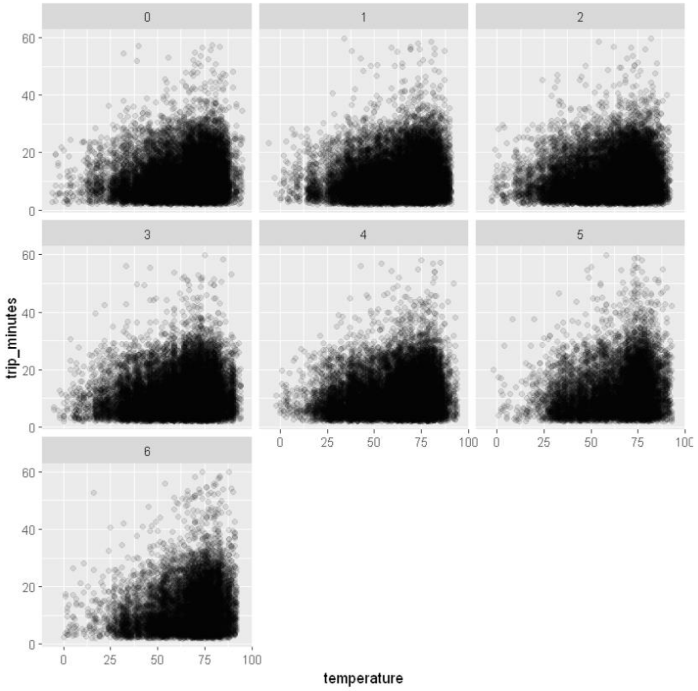
- If adding intercept, the model has very low $R^2 = 0.2127$, therefore we exclude the intercept (Excluding the intercept might not be a sensible thing to do here, and it makes sense that temperature is not highly correlated with trip_minutes intuitively)
- The distribution of temperature is roughly uniform and temperature has negative value, log transformation is not applicable
- avg_dpcapacity is derived from dpcapacity_start and dpcapacity_end
- temperature and avg_dpcapacity are the only two numerical explanatory variables. They are not highly collinear with other predictors in the model since their VIF are both close to 1



- This baseline model (trip_minutes ~ 0 + temperature) has $R^2 = 0.6964$



- Categorical day interacts with temperature. Month and hour also seem useful



- The model includes interaction between area_start and area_end so that all 4×4=16 combinations are represented in the model

- Among other categorical explanatory variables that are not highly correlated with trip_minutes, usertype is not useful (p-value = 0.5845) while some weather events might be. This final model has $R^2 = 0.7767$

Logistic Regression Model

Since the dataset has most categorical variables, we also build a logistic regression model using the same dataset as above. We separate trip_minutes with binary classification and use them as response variables: if trip_minutes > 10, return True; if trip_minutes <= 10, return False. We use all other useful categorical information as explanatory variables. The model formula is

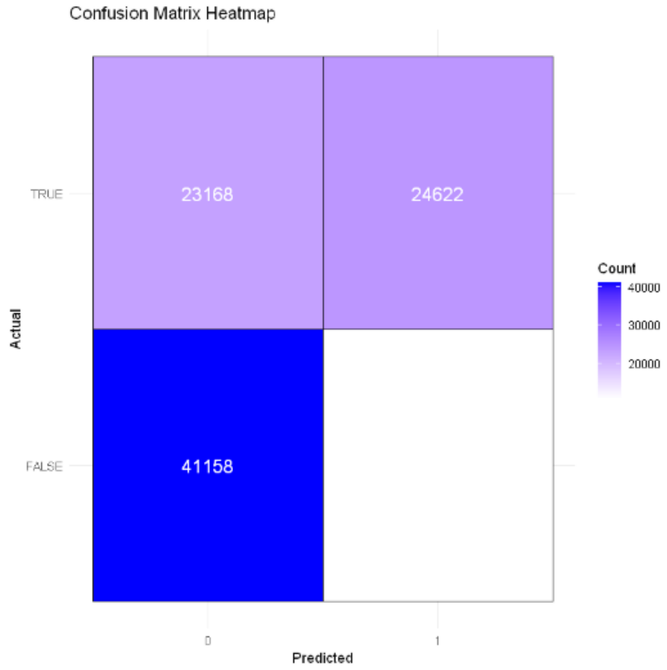
```
factor(trip_minutes > 10) ~ temperature * factor(year) + factor(month) * factor(day) +
factor(area_start) * factor(area_end) + factor(usertype) + factor(gender) +
factor(events) + avg_dpcapacity
```

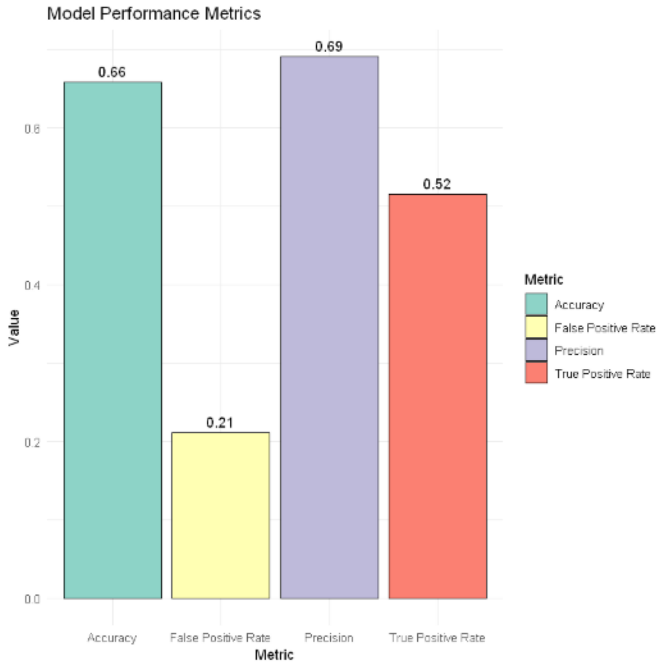
Then we did a backward elimination to exclude unnecessary features and get final model:

```
factor(trip_minutes > 10) ~ temperature + factor(year) + factor(month) + factor(day) +
factor(area_start) + factor(area_end) + factor(usertype) + factor(gender) +
factor(events) + avg_dpcapacity + temperature:factor(year) +
factor(area_start):factor(area_end)
```

Key Considerations in Modeling

- The factors contributing significantly to trip_minutes ($p < 0.05$) include: temperature, usertype, events, avg_dpcapacity, area_start, area_end (See Appendix 2 for variable selection logic)
- Confusion matrix highlights an imbalance between trip_minutes longer and shorter than 10 minutes
- The model has 66% accuracy and 69% precision, showing moderate overall prediction success





Appendix 1

```
Call:
lm(formula = trip_minutes ~ 0 + temperature * factor(hour) + 
    factor(month) + factor(day) + area_start * area_end + factor(gender) + 
    factor(events) + avg_dpcapacity, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.187  -4.400  -1.217   3.089  51.030 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
temperature                   5.655e-02  1.692e-02   3.342 0.000833 
factor(hour)0                 3.254e+00  1.013e+00   3.210 0.001327 
factor(hour)1                 2.673e+00  1.269e+00   2.107 0.035079 
factor(hour)2                 4.107e+00  1.576e+00   2.606 0.009150 
factor(hour)3                 3.194e+00  1.981e+00   1.612 0.106982 
factor(hour)4                 1.903e+00  1.689e+00   1.127 0.259826 
factor(hour)5                 4.262e+00  7.711e-01   5.527 3.26e-08 
factor(hour)6                 4.626e+00  4.644e-01   9.962  < 2e-16 
factor(hour)7                 5.409e+00  3.313e-01  16.323  < 2e-16 
factor(hour)8                 5.908e+00  3.129e-01  18.879  < 2e-16 
factor(hour)9                 5.257e+00  4.037e-01  13.024  < 2e-16 
factor(hour)10                3.873e+00  4.639e-01   8.349  < 2e-16 
factor(hour)11                4.480e+00  4.485e-01   9.988  < 2e-16 
factor(hour)12                4.224e+00  4.200e-01  10.058  < 2e-16 
factor(hour)13                3.939e+00  4.329e-01   9.100  < 2e-16 
factor(hour)14                4.382e+00  4.478e-01   9.786  < 2e-16 
factor(hour)15                5.000e+00  3.932e-01  12.716  < 2e-16 
factor(hour)16                5.694e+00  3.227e-01  17.646  < 2e-16 
factor(hour)17                5.080e+00  2.942e-01  17.265  < 2e-16 
factor(hour)18                4.550e+00  3.376e-01  13.480  < 2e-16 
factor(hour)19                4.512e+00  3.984e-01  11.326  < 2e-16 
factor(hour)20                3.572e+00  4.775e-01   7.481 7.44e-14 
factor(hour)21                3.451e+00  5.524e-01   6.248 4.19e-10 
factor(hour)22                2.891e+00  6.782e-01   4.263 2.02e-05 
factor(hour)23                4.233e+00  7.789e-01   5.435 5.50e-08 
factor(month)2                1.668e-01  1.648e-01   1.012 0.311344 
factor(month)3               -1.253e-01  1.560e-01  -0.803 0.422003 
factor(month)4                1.203e-01  1.552e-01   0.775 0.438243 
factor(month)5                4.416e-01  1.592e-01   2.774 0.005540 
factor(month)6                1.382e-01  1.677e-01   0.824 0.410014 
factor(month)7                3.130e-01  1.706e-01   1.835 0.066576 
factor(month)8                1.136e-01  1.687e-01   0.673 0.500857 
factor(month)9               -1.376e-01  1.638e-01  -0.840 0.400807 
factor(month)10              -1.136e-01  1.511e-01  -0.752 0.452067 
factor(month)11              -8.494e-02  1.490e-01  -0.570 0.568749 
factor(month)12              -1.003e-01  1.564e-01  -0.642 0.521142 
factor(day)1                 -1.162e-01  7.157e-02  -1.624 0.104319 
factor(day)2                 -8.127e-02  7.208e-02  -1.127 0.259587 
factor(day)3                 -3.161e-02  7.208e-02  -0.439 0.660969 
factor(day)4                  3.007e-02  7.318e-02   0.411 0.681094 
factor(day)5                  1.052e+00  8.369e-02  12.565  < 2e-16 
factor(day)6                     9.205e-01  8.459e-02  10.882  < 2e-16
area_startHyde Park              3.991e+01  3.200e+00  12.470  < 2e-16
area_startLakefront              2.852e+00  9.207e-02  30.972  < 2e-16
area_startOther                  6.880e+00  9.825e-02  70.029  < 2e-16
area_endHyde Park                4.203e+01  4.533e+00   9.274  < 2e-16
area_endLakefront                3.236e+00  9.093e-02  35.587  < 2e-16
area_endOther                    7.247e+00  9.654e-02  75.062  < 2e-16
factor(gender)Male              -1.289e+00  4.713e-02 -27.359  < 2e-16
factor(events)cloudy            -1.466e-01  9.265e-02  -1.583 0.113488
factor(events)not clear         -1.687e-01  2.307e-01  -0.731 0.464641
factor(events)rain or snow      -8.220e-01  1.308e-01  -6.285 3.29e-10
factor(events)tstorms           -1.306e+00  2.606e-01  -5.010 5.45e-07
factor(events)unknown           -1.362e+00  3.697e+00  -0.368 0.712626
avg_dpcapacity                   6.722e-02  4.378e-03  15.354  < 2e-16
temperature:factor(hour)1        7.928e-03  2.737e-02   0.290 0.772094
temperature:factor(hour)2       -2.257e-02  3.217e-02  -0.702 0.482936
temperature:factor(hour)3       -3.875e-03  3.846e-02  -0.101 0.919735
temperature:factor(hour)4        6.943e-03  3.477e-02   0.200 0.841711
temperature:factor(hour)5       -2.761e-02  2.145e-02  -1.287 0.198126
temperature:factor(hour)6       -2.935e-02  1.832e-02  -1.602 0.109059
temperature:factor(hour)7       -3.985e-02  1.739e-02  -2.291 0.021969
temperature:factor(hour)8       -4.303e-02  1.726e-02  -2.493 0.012668
temperature:factor(hour)9       -3.602e-02  1.769e-02  -2.035 0.041804
temperature:factor(hour)10      -1.496e-02  1.796e-02  -0.833 0.404842
temperature:factor(hour)11      -2.319e-02  1.780e-02  -1.303 0.192657
temperature:factor(hour)12      -1.696e-02  1.764e-02  -0.961 0.336329
temperature:factor(hour)13      -1.434e-02  1.768e-02  -0.811 0.417242
temperature:factor(hour)14      -1.732e-02  1.774e-02  -0.976 0.328828
temperature:factor(hour)15      -2.299e-02  1.749e-02  -1.314 0.188684
temperature:factor(hour)16      -3.121e-02  1.721e-02  -1.814 0.069679
temperature:factor(hour)17      -2.240e-02  1.710e-02  -1.310 0.190257
temperature:factor(hour)18      -1.393e-02  1.729e-02  -0.805 0.420556
temperature:factor(hour)19      -1.496e-02  1.761e-02  -0.850 0.395523
temperature:factor(hour)20      -1.584e-04  1.810e-02  -0.009 0.993020
temperature:factor(hour)21       1.126e-03  1.868e-02   0.060 0.951943
temperature:factor(hour)22       9.802e-03  1.972e-02   0.497 0.619063
temperature:factor(hour)23      -1.263e-02  2.088e-02  -0.605 0.545111
area_startHyde Park:area_endHyde Park -8.126e+01  5.551e+00 -14.639  < 2e-16
area_startLakefront:area_endHyde Park -2.104e+01  4.613e+00  -4.561 5.11e-06
area_startOther:area_endHyde Park     -4.130e+01  4.559e+00  -9.059  < 2e-16
area_startHyde Park:area_endLakefront -1.912e+01  3.293e+00  -5.807 6.38e-09
area_startLakefront:area_endLakefront -3.465e+00  1.332e-01 -26.020  < 2e-16
area_startOther:area_endLakefront      1.404e+00  1.575e-01   8.911  < 2e-16
area_startHyde Park:area_endOther     -3.856e+01  3.236e+00 -11.916  < 2e-16
area_startLakefront:area_endOther      1.517e+00  1.523e-01   9.963  < 2e-16
area_startOther:area_endOther         -1.096e+01  1.256e-01 -87.290  < 2e-16

temperature                      ***
factor(hour)0                    **
factor(hour)1                    *
factor(hour)2                    **
factor(hour)3
factor(hour)4
factor(hour)5                    ***
factor(hour)6                    ***
factor(hour)7            ***
factor(hour)8            ***
factor(hour)9            ***
factor(hour)10            ***
factor(hour)11            ***
factor(hour)12            ***
factor(hour)13            ***
factor(hour)14            ***
factor(hour)15            ***
factor(hour)16            ***
factor(hour)17            ***
factor(hour)18            ***
factor(hour)19            ***
factor(hour)20            ***
factor(hour)21            ***
factor(hour)22            ***
factor(hour)23            ***
factor(month)2
factor(month)3
factor(month)4
factor(month)5             **
factor(month)6
factor(month)7             .
factor(month)8
factor(month)9
factor(month)10
factor(month)11
factor(month)12
factor(day)1
factor(day)2
factor(day)3
factor(day)4
factor(day)5             ***
factor(day)6             ***
area_startHyde Park          ***
area_startLakefront         ***
area_startOther           ***
area_endHyde Park            ***
area_endLakefront           ***
area_endOther             ***
factor(gender)Male          ***
factor(events)cloudy
factor(events)not clear
factor(events)rain or snow      ***
factor(events)tstorms        ***
factor(events)unknown
avg_dpcapacity            ***
temperature:factor(hour)1
temperature:factor(hour)2
temperature:factor(hour)3
temperature:factor(hour)4
temperature:factor(hour)5
temperature:factor(hour)6
temperature:factor(hour)7       *
temperature:factor(hour)8       *
```

```
temperature:factor(hour)9           *
temperature:factor(hour)10
temperature:factor(hour)11
temperature:factor(hour)12
temperature:factor(hour)13
temperature:factor(hour)14
temperature:factor(hour)15
temperature:factor(hour)16          .
temperature:factor(hour)17
temperature:factor(hour)18
temperature:factor(hour)19
temperature:factor(hour)20
temperature:factor(hour)21
temperature:factor(hour)22
temperature:factor(hour)23
area_startHyde Park:area_endHyde Park ***
area_startLakefront:area_endHyde Park ***
area_startOther:area_endHyde Park     ***
area_startHyde Park:area_endLakefront ***
area_startLakefront:area_endLakefront ***
area_startOther:area_endLakefront     ***
area_startHyde Park:area_endOther     ***
area_startLakefront:area_endOther     ***
area_startOther:area_endOther         ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.396 on 99913 degrees of freedom
Multiple R-squared:  0.7767,	Adjusted R-squared:  0.7765 
F-statistic:  3995 on 87 and 99913 DF,  p-value: < 2.2e-16
```

Appendix 2

```
Start:  AIC=121596.3
factor(trip_minutes > 10) ~ temperature * factor(year) + factor(month) * 
    factor(day) + factor(area_start) * factor(area_end) + factor(usertype) + 
    factor(gender) + factor(events) + avg_dpcapacity

                                     Df Deviance    AIC     LRT  Pr(>Chi)    
- factor(month):factor(day)          66   121458 121556    91.9   0.01929 *  
<none>                                    121366 121596                      
- temperature:factor(year)            3   121376 121600     9.4   0.02447 *  
- factor(usertype)                    2   121374 121600     7.4   0.02413 *  
- factor(events)                      5   121428 121648    61.9  4.98e-12 ***
- avg_dpcapacity                      1   121463 121691    96.5 < 2.2e-16 ***
- factor(gender)                      1   121903 122131   536.2 < 2.2e-16 ***
- factor(area_start):factor(area_end) 9   134561 134773 13194.9 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Step:  AIC=121556.2
factor(trip_minutes > 10) ~ temperature + factor(year) + factor(month) + 
    factor(day) + factor(area_start) + factor(area_end) + factor(usertype) + 
    factor(gender) + factor(events) + avg_dpcapacity + temperature:factor(year) + 
    factor(area_start):factor(area_end)

                                    Df Deviance    AIC     LRT  Pr(>Chi)    
<none>                                  121458 121556                       
- temperature:factor(year)           3   121467 121559     8.6   0.03587 *  
- factor(usertype)                   2   121466 121560     7.3   0.02544 *  
- factor(month)                     11   121511 121587    52.7 2.040e-07 ***
- factor(events)                     5   121522 121610    63.5 2.303e-12 ***
- avg_dpcapacity                     1   121556 121652    97.7 < 2.2e-16 ***
- factor(day)                        6   121596 121682   138.0 < 2.2e-16 ***
- factor(gender)                     1   121994 122090   535.7 < 2.2e-16 ***
- factor(area_start):factor(area_end)  9   134677 134757 13219.0 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```