Untitled

Bohan Yang

07/09/2019

train <- read.csv('train.csv')  
data <- read.csv('final.csv')  
test <- read.csv('test.csv')

t\_model0 <- lm(tip\_amount ~ 1, data = train)  
t\_model1 <- step(t\_model0, scope = ~ . + avg\_speed + pickup\_hour + Minimum.Temperature + Maximum.Temperature + Temperature)

## Start: AIC=28608.46  
## tip\_amount ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + avg\_speed 1 22327.9 73564 24384  
## + Maximum.Temperature 1 32.9 95859 28605  
## + Temperature 1 31.5 95860 28605  
## + Minimum.Temperature 1 27.0 95864 28606  
## + pickup\_hour 1 12.3 95879 28608  
## <none> 95891 28609  
##   
## Step: AIC=24383.97  
## tip\_amount ~ avg\_speed  
##   
## Df Sum of Sq RSS AIC  
## + pickup\_hour 1 318.0 73246 24317  
## + Maximum.Temperature 1 18.8 73545 24382  
## + Temperature 1 17.8 73546 24382  
## + Minimum.Temperature 1 14.6 73549 24383  
## <none> 73564 24384  
## - avg\_speed 1 22327.9 95891 28609  
##   
## Step: AIC=24316.89  
## tip\_amount ~ avg\_speed + pickup\_hour  
##   
## Df Sum of Sq RSS AIC  
## + Maximum.Temperature 1 19.6 73226 24315  
## + Temperature 1 18.6 73227 24315  
## + Minimum.Temperature 1 15.4 73230 24316  
## <none> 73246 24317  
## - pickup\_hour 1 318.0 73564 24384  
## - avg\_speed 1 22633.6 95879 28608  
##   
## Step: AIC=24314.63  
## tip\_amount ~ avg\_speed + pickup\_hour + Maximum.Temperature  
##   
## Df Sum of Sq RSS AIC  
## <none> 73226 24315  
## + Minimum.Temperature 1 2.7 73223 24316  
## + Temperature 1 0.4 73226 24317  
## - Maximum.Temperature 1 19.6 73246 24317  
## - pickup\_hour 1 318.8 73545 24382  
## - avg\_speed 1 22620.0 95846 28605

t\_model2 <- lm(tip\_amount ~ (avg\_speed + pickup\_hour + Maximum.Temperature)^2, data = train)  
anova(t\_model1, t\_model2)

## Analysis of Variance Table  
##   
## Model 1: tip\_amount ~ avg\_speed + pickup\_hour + Maximum.Temperature  
## Model 2: tip\_amount ~ (avg\_speed + pickup\_hour + Maximum.Temperature)^2  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 15941 73226   
## 2 15938 73092 3 134.33 9.7638 1.969e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

t\_model3 <- step(t\_model2)

## Start: AIC=24291.35  
## tip\_amount ~ (avg\_speed + pickup\_hour + Maximum.Temperature)^2  
##   
## Df Sum of Sq RSS AIC  
## - pickup\_hour:Maximum.Temperature 1 3.986 73096 24290  
## - avg\_speed:Maximum.Temperature 1 6.438 73098 24291  
## <none> 73092 24291  
## - avg\_speed:pickup\_hour 1 126.262 73218 24317  
##   
## Step: AIC=24290.22  
## tip\_amount ~ avg\_speed + pickup\_hour + Maximum.Temperature +   
## avg\_speed:pickup\_hour + avg\_speed:Maximum.Temperature  
##   
## Df Sum of Sq RSS AIC  
## - avg\_speed:Maximum.Temperature 1 5.585 73101 24289  
## <none> 73096 24290  
## - avg\_speed:pickup\_hour 1 127.045 73223 24316  
##   
## Step: AIC=24289.44  
## tip\_amount ~ avg\_speed + pickup\_hour + Maximum.Temperature +   
## avg\_speed:pickup\_hour  
##   
## Df Sum of Sq RSS AIC  
## <none> 73101 24289  
## - Maximum.Temperature 1 20.119 73121 24292  
## - avg\_speed:pickup\_hour 1 124.760 73226 24315

summary(t\_model3)

##   
## Call:  
## lm(formula = tip\_amount ~ avg\_speed + pickup\_hour + Maximum.Temperature +   
## avg\_speed:pickup\_hour, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.6401 -1.2043 -0.3768 0.6563 29.7792   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.5438963 0.0965799 5.632 1.82e-08 \*\*\*  
## avg\_speed 0.1048986 0.0037786 27.761 < 2e-16 \*\*\*  
## pickup\_hour -0.0065915 0.0062738 -1.051 0.2934   
## Maximum.Temperature 0.0031051 0.0014825 2.095 0.0362 \*   
## avg\_speed:pickup\_hour 0.0013558 0.0002599 5.216 1.85e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.141 on 15940 degrees of freedom  
## Multiple R-squared: 0.2377, Adjusted R-squared: 0.2375   
## F-statistic: 1242 on 4 and 15940 DF, p-value: < 2.2e-16

t\_model4 <- step(t\_model0, scope = ~ . + (avg\_speed + pickup\_hour + Minimum.Temperature + Maximum.Temperature + Temperature)^2)

## Start: AIC=28608.46  
## tip\_amount ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + avg\_speed 1 22327.9 73564 24384  
## + Maximum.Temperature 1 32.9 95859 28605  
## + Temperature 1 31.5 95860 28605  
## + Minimum.Temperature 1 27.0 95864 28606  
## + pickup\_hour 1 12.3 95879 28608  
## <none> 95891 28609  
##   
## Step: AIC=24383.97  
## tip\_amount ~ avg\_speed  
##   
## Df Sum of Sq RSS AIC  
## + pickup\_hour 1 318.0 73246 24317  
## + Maximum.Temperature 1 18.8 73545 24382  
## + Temperature 1 17.8 73546 24382  
## + Minimum.Temperature 1 14.6 73549 24383  
## <none> 73564 24384  
## - avg\_speed 1 22327.9 95891 28609  
##   
## Step: AIC=24316.89  
## tip\_amount ~ avg\_speed + pickup\_hour  
##   
## Df Sum of Sq RSS AIC  
## + avg\_speed:pickup\_hour 1 124.2 73121 24292  
## + Maximum.Temperature 1 19.6 73226 24315  
## + Temperature 1 18.6 73227 24315  
## + Minimum.Temperature 1 15.4 73230 24316  
## <none> 73246 24317  
## - pickup\_hour 1 318.0 73564 24384  
## - avg\_speed 1 22633.6 95879 28608  
##   
## Step: AIC=24291.82  
## tip\_amount ~ avg\_speed + pickup\_hour + avg\_speed:pickup\_hour  
##   
## Df Sum of Sq RSS AIC  
## + Maximum.Temperature 1 20.119 73101 24289  
## + Temperature 1 19.088 73102 24290  
## + Minimum.Temperature 1 15.843 73105 24290  
## <none> 73121 24292  
## - avg\_speed:pickup\_hour 1 124.225 73246 24317  
##   
## Step: AIC=24289.44  
## tip\_amount ~ avg\_speed + pickup\_hour + Maximum.Temperature +   
## avg\_speed:pickup\_hour  
##   
## Df Sum of Sq RSS AIC  
## <none> 73101 24289  
## + avg\_speed:Maximum.Temperature 1 5.585 73096 24290  
## + pickup\_hour:Maximum.Temperature 1 3.133 73098 24291  
## + Minimum.Temperature 1 2.757 73098 24291  
## + Temperature 1 0.390 73101 24291  
## - Maximum.Temperature 1 20.119 73121 24292  
## - avg\_speed:pickup\_hour 1 124.760 73226 24315

deviance(t\_model3) # residual sum of squares

## [1] 73101.16

deviance(t\_model3)/t\_model3$df.residual # sample variance

## [1] 4.58602

confint(t\_model3)

## 2.5 % 97.5 %  
## (Intercept) 0.3545888303 0.733203815  
## avg\_speed 0.0974921072 0.112305081  
## pickup\_hour -0.0188887912 0.005705771  
## Maximum.Temperature 0.0001992483 0.006011039  
## avg\_speed:pickup\_hour 0.0008462842 0.001865314

plot(t\_model3)

