forcasting forest fires

fantastic fouR: Eli Levine, Hannah Long, Preetha Ramachandran

Oct. 26, 2020

Your written report goes here! Before you submit, make sure your code chunks are turned off with `echo = FALSE` and there are no warnings or messages with `warning = FALSE` and `message = FALSE`

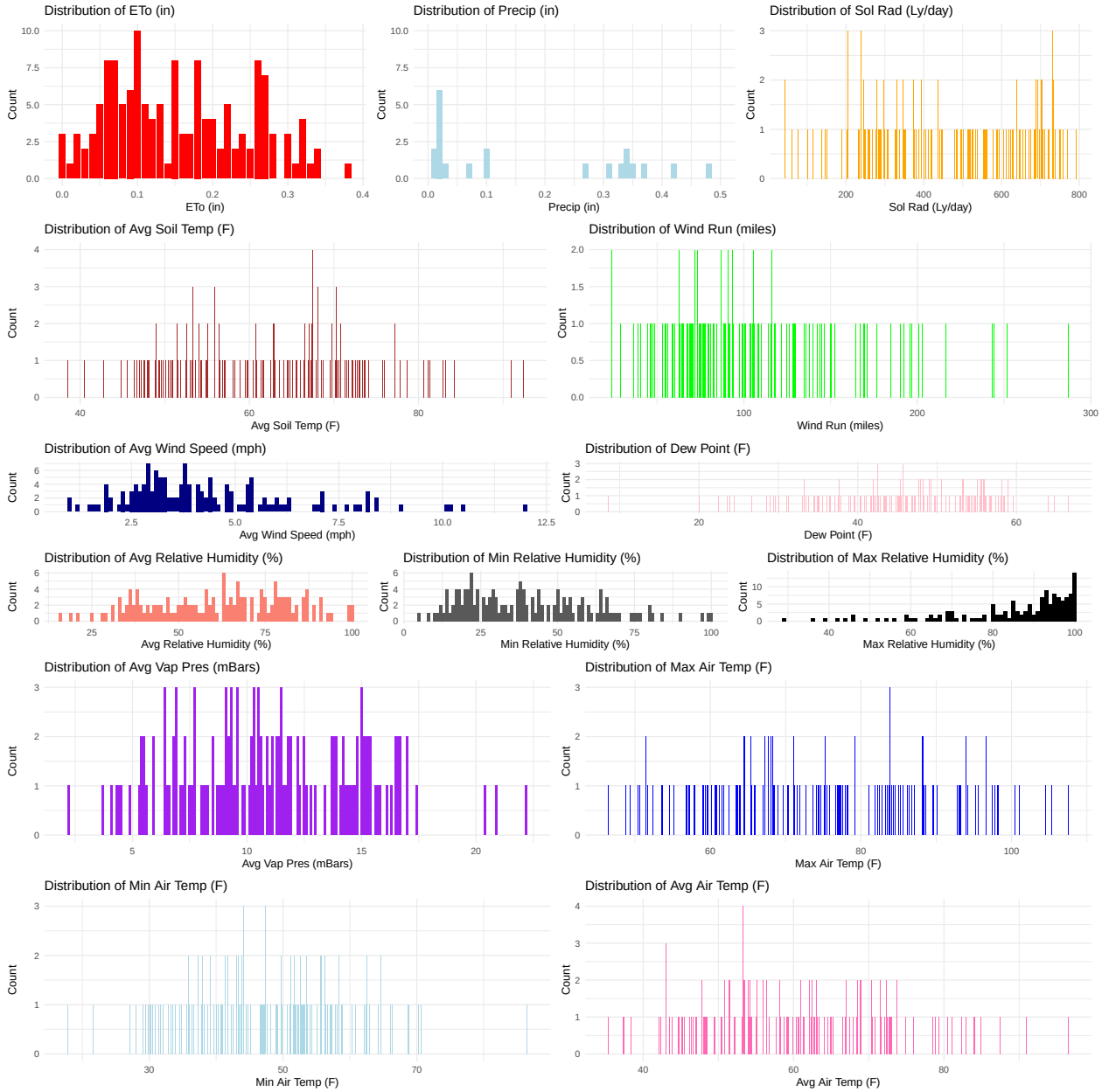
###Introduction

We started by creating distribution plots for each of the continuous variables. Because some of the predictor names have parentheses that aren't compatible with R, we renamed some of the variables.

((narrative about why we had to create a new data frame with only one observation per station. something about how it's down to 153 observations but how that's still okay))

((narrative about getting rid of observations with missing values. something about how it's down to 143 observations but that's still okay!!))

((narrative about how we wanted to look at the distributions of all of the response variables to get an idea of what they look like post data cleaning))



((general descriptions of what each of the plots look like (shape, center, spread, distribution), attribute all weird distributions to fewer data points))

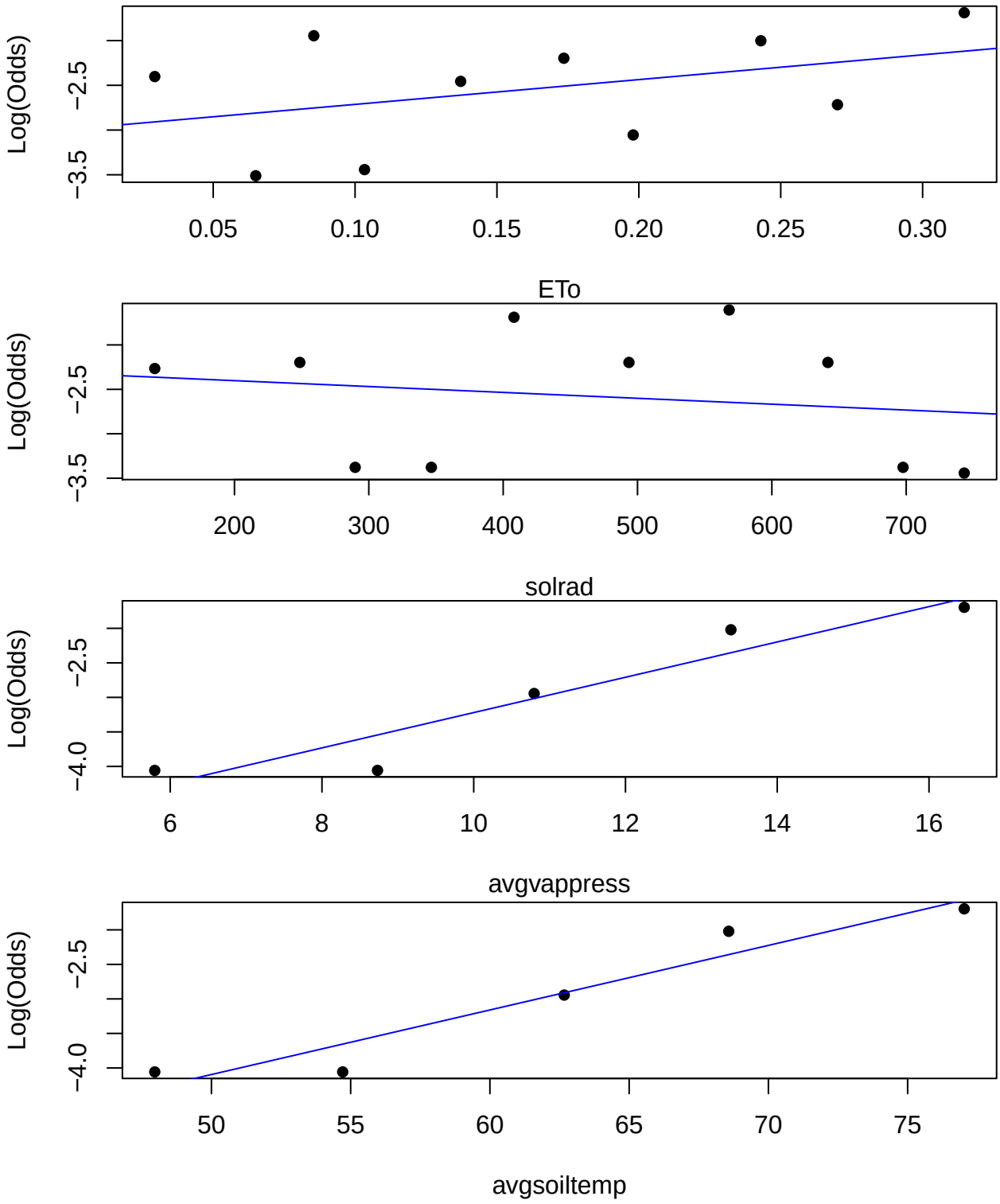
Methodology

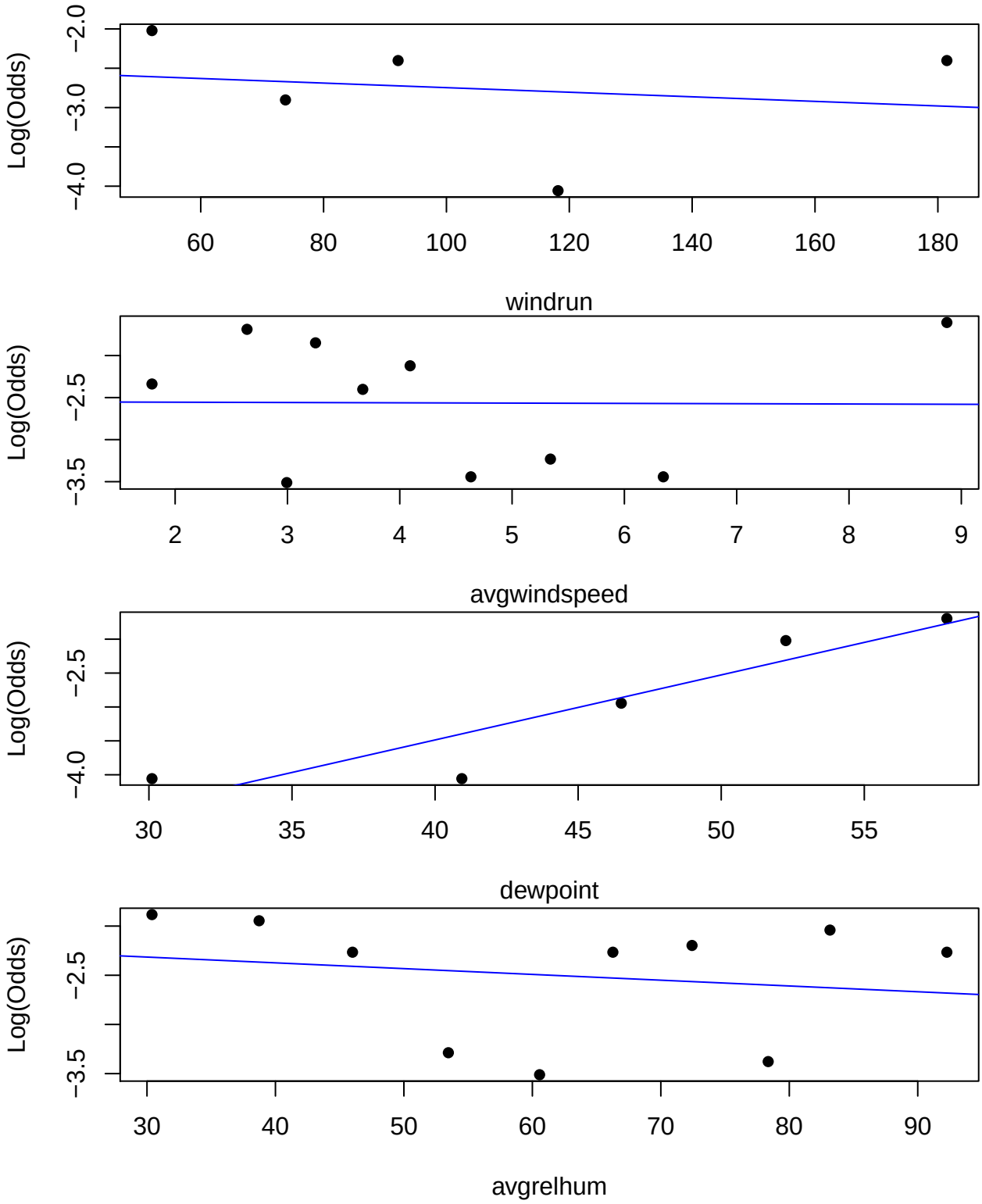
((narrative about how we are tossing out min and max distributions and keeping averages because of multicollinarity (take from tackett's email)))

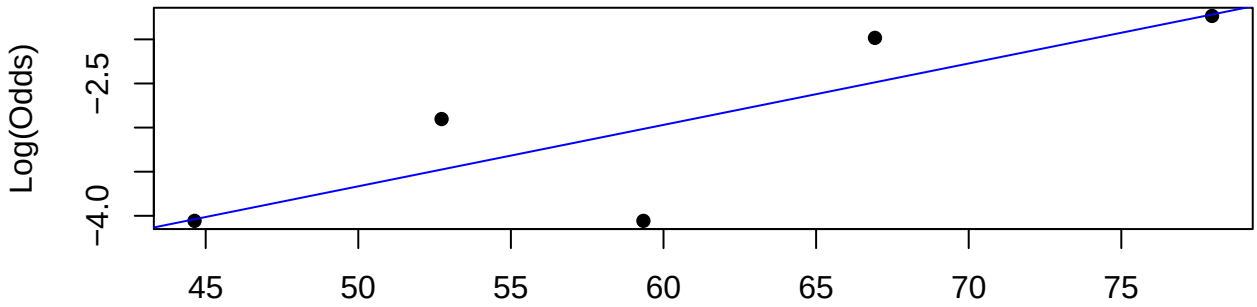
((narrative about what we're trying to do here (predict) and explain why a logistic regression is used for that (binary prediction - fire vs no fire)))

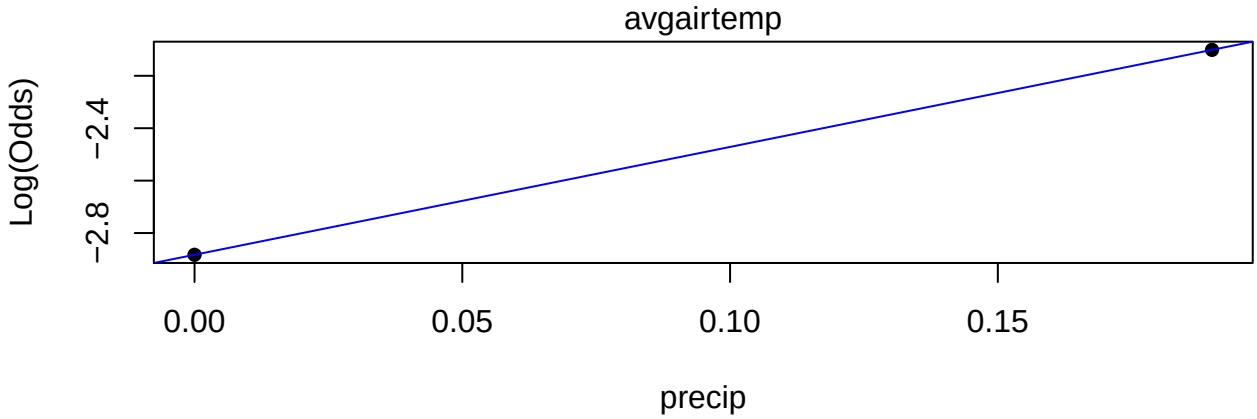
((narrative about the conditions that need to be checked, linearity, randomness, independence))

((LINEARITY))


Log(Odds)
−3.5
−2.5
0.05
0.10
0.15
0.20
0.25
0.30
ETo
Log(Odds)
−3.5
−2.5
200
300
400
500
600
700
solrad
Log(Odds)
−4.0
−2.5
6
8
10
12
14
16
avgvappress
Log(Odds)
−4.0
−2.5
50
55
60
65
70
75
avgsoiltemp


Log(Odds)
−2.0
−3.0
−4.0
60
80
100
120
140
160
180
windrun
Log(Odds)
−2.5
−3.5
2
3
4
5
6
7
8
9
avgwindspeed
Log(Odds)
−2.5
−4.0
30
35
40
45
50
55
dewpoint
Log(Odds)
−2.5
−3.5
30
40
50
60
70
80
90
avgrelhum





((we are looking for a linear relationship and we don't see that in avgwindspeed, avgrelhum. also plot for precip only calculated if ngroups is set to 2 and that doesn't tell us much))

((RANDOMNESS))

((narrative along the lines of the way the data itself was collected is not random but the way we cleaned it and created a new data frame helps to account for that))

((INDEPENDENCE))

((narrative along the lines of the way the data itself was collected does not lend itself to independent observations but the way we cleaned it and created a new data frame helps to account for that))

((we then started to build models. for our first model, we tossed out the predictor variables that do not have a linear relationship with response used backwards selection to build a model with everything else))

```
## # A tibble: 8 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) -23.6      28.5       -0.829  0.407
## 2 ETo          76.4      51.4        1.49   0.137
## 3 solrad       -0.0332    0.0158    -2.11   0.0349
## 4 avgvappress  -1.15      2.20      -0.526  0.599
## 5 avgsoiltemp   0.210     0.0816     2.57   0.0101
## 6 windrun      -0.0203    0.0190    -1.07   0.286
## 7 dewpoint      0.828     1.18       0.700  0.484
## 8 avgairtemp   -0.253     0.212     -1.19   0.234
```

```
##         ETo      solrad avgvappress avgsoiltemp     windrun    dewpoint
##  126.466677   56.604725  115.669844    3.992473    3.303266  129.184455
##  avgairtemp
##   30.659148
```

((multicollinearity first, then see if add square terms and interactions. mean center variables))

```
## # A tibble: 6 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) -20.7      5.07      -4.07   0.0000464
## 2 ETo         -32.4     12.2       -2.66   0.00778
## 3 avgvappress   0.0148   0.152      0.0974 0.922
## 4 avgsoiltemp   0.177    0.0721     2.45   0.0143
## 5 windrun       0.0125   0.00981    1.27   0.204
## 6 avgairtemp    0.147    0.0833     1.76   0.0785
```

```
##         ETo avgvappress avgsoiltemp     windrun  avgairtemp
##    7.073298    1.401153    3.218497    1.445667    5.995711
```

((drop-in deviance to see what variables are significant)) ((ROC curve to find threshold for prediciton "final model"))

```
## Start:  AIC=50.26
## Target ~ ETo + avgvappress + avgsoiltemp + windrun + avgairtemp
##
##               Df Deviance    AIC
## - avgvappress  1   38.267 48.267
## - windrun      1   39.933 49.933
## <none>             38.257 50.257
## - avgairtemp   1   41.660 51.660
## - avgsoiltemp  1   44.628 54.628
## - ETo          1   47.911 57.911
##
## Step:  AIC=48.27
## Target ~ ETo + avgsoiltemp + windrun + avgairtemp
##
##               Df Deviance    AIC
## - windrun      1   40.074 48.074
## <none>             38.267 48.267
## - avgairtemp   1   42.423 50.423
## - avgsoiltemp  1   45.482 53.482
## - ETo          1   49.259 57.259
##
## Step:  AIC=48.07
## Target ~ ETo + avgsoiltemp + avgairtemp
##
##               Df Deviance    AIC
## <none>             40.074 48.074
## - avgairtemp   1   43.297 49.297
## - avgsoiltemp  1   46.379 52.379
## - ETo          1   49.448 55.448
```

```
## # A tibble: 4 x 7
##   term        estimate std.error statistic   p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)  -18.4      4.46       -4.12 0.0000379 -28.7      -10.8
## 2 ETo          -26.3     10.0        -2.62 0.00886   -49.0       -8.67
## 3 avgsoiltemp    0.169    0.0701      2.41 0.0160      0.0366     0.320
## 4 avgairtemp     0.128    0.0755      1.69 0.0911     -0.0112     0.294
```

((This model only has three predictors, so we transformed some variables to see if that would give further insight. quadratic trans formation for wind speed and humidity, by looking at empirical log plot and observing

shape))

```
## # A tibble: 1 x 2
##     AIC   BIC
##   <dbl> <dbl>
## 1  48.1  59.9
```

((AIC, BIC, Adj R to determine))

```
## # A tibble: 8 x 5
##   term                estimate std.error statistic p.value
##   <chr>                  <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)       -17.2       7.62        -2.25   0.0242
## 2 ETo               -34.8      14.1         -2.47   0.0134
## 3 avgvappress         0.0710    0.295        0.241  0.810
## 4 avgsoiltemp         0.175     0.0732       2.39   0.0169
## 5 windrun            -0.0113    0.0285      -0.395  0.693
## 6 avgairtemp          0.119     0.124        0.956  0.339
## 7 I(avgwindspeed^2)   0.0558    0.0587       0.951  0.342
## 8 I(avgrelhum^2)     -0.000167  0.000570    -0.293  0.770
```

```
## Start:  AIC=53.23
## Target ~ ETo + avgvappress + avgsoiltemp + windrun + avgairtemp +
##     I(avgwindspeed^2) + I(avgrelhum^2)
##
##                     Df Deviance    AIC
## - avgvappress        1   37.293 51.293
## - I(avgrelhum^2)     1   37.321 51.321
## - windrun            1   37.383 51.383
## - I(avgwindspeed^2)  1   38.023 52.023
## - avgairtemp         1   38.064 52.064
## <none>                   37.232 53.232
## - avgsoiltemp        1   43.304 57.304
## - ETo                1   45.689 59.689
##
## Step:  AIC=51.29
## Target ~ ETo + avgsoiltemp + windrun + avgairtemp + I(avgwindspeed^2) +
##     I(avgrelhum^2)
##
##                     Df Deviance    AIC
## - I(avgrelhum^2)     1   37.321 49.321
## - windrun            1   37.503 49.503
## - I(avgwindspeed^2)  1   38.233 50.233
## <none>                   37.293 51.293
## - avgairtemp         1   40.510 52.510
## - avgsoiltemp        1   43.978 55.978
## - ETo                1   45.743 57.743
##
## Step:  AIC=49.32
## Target ~ ETo + avgsoiltemp + windrun + avgairtemp + I(avgwindspeed^2)
##
##                     Df Deviance    AIC
## - windrun            1   37.555 47.555
## - I(avgwindspeed^2)  1   38.267 48.267
## <none>                   37.321 49.321
## - avgairtemp         1   40.967 50.967
```

```
## - avgsoiltemp        1   44.261 54.261
## - ETo                1   47.912 57.912
##
## Step:  AIC=47.56
## Target ~ ETo + avgsoiltemp + avgairtemp + I(avgwindspeed^2)
##
##                     Df Deviance    AIC
## <none>                   37.555 47.555
## - I(avgwindspeed^2)  1   40.074 48.074
## - avgairtemp         1   41.607 49.607
## - avgsoiltemp        1   44.775 52.775
## - ETo                1   49.437 57.437
```

```
## # A tibble: 5 x 7
##   term              estimate std.error statistic   p.value  conf.low conf.high
##   <chr>                <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)       -20.0       4.77       -4.19 0.0000282 -31.1      -11.9
## 2 ETo               -33.7      11.7        -2.89 0.00389   -60.5      -13.4
## 3 avgsoiltemp         0.178     0.0701      2.54 0.0112      0.0476     0.333
## 4 avgairtemp          0.149     0.0792      1.88 0.0603      0.00372    0.323
## 5 I(avgwindspeed^2)   0.0321    0.0207      1.55 0.121      -0.00768    0.0761
```

```
## # A tibble: 1 x 8
##   null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##           <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1          61.7     142  -18.8  47.6  62.4     37.6         138   143
```

With quadratic transformed variables, AIC has marginal improvement, but BIC is larger. So, it is optimal to stick with the model without quadratic transformation of variables.

```
## # A tibble: 10 x 5
##    term                    estimate std.error statistic p.value
##    <chr>                      <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)             -36.2      34.6       -1.05    0.295
##  2 ETo                     -38.2      14.9       -2.56    0.0106
##  3 avgsoiltemp               0.416     0.373      1.12    0.264
##  4 avgairtemp                0.420     0.507      0.828   0.408
##  5 avgvappress               0.0374    0.450      0.0830  0.934
##  6 avgwindspeed            -18.7      20.1       -0.928   0.354
##  7 windrun                   0.764     0.837      0.914   0.361
##  8 avgrelhum                -0.0165    0.117     -0.142   0.887
##  9 avgsoiltemp:avgairtemp   -0.00352   0.00529   -0.665   0.506
## 10 avgwindspeed:windrun      0.00279   0.00244    1.15    0.252
```

```
## # A tibble: 2 x 5
##   Resid..Df Resid..Dev    df Deviance p.value
##       <dbl>      <dbl> <dbl>    <dbl>   <dbl>
## 1       137       38.3    NA    NA     NA
## 2       133       35.2     4     3.10   0.541
```

High p-value, so the data provide insufficient evidence to conclude the interaction terms are statistically significant.

After all these analyses/tests, our ultimate best model is new_fire_model (main effect model and backwards selection)

———— notes ————————————————————————- - picked variables that need transformation (rel humidity, sol rad, precip) - log transform? - build model, backward assumption

variables we want to keep: - average humidity multicollinearity w min and max, we think humidity prob has some effect on fire (water in air???) - avg air temp multicolinearity w min and max air temp

things to do after building model w everything: - log transform precipitaion?? it looks funny - sol radiation closely related to soil temperature (MAKE PLOT FOR THIS FIRST) - dew point//temperature//humidity - wind run//wind speed