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

Climate change represents an existential threat to humanity. Its effects have the potential to dramatically shape life as we know it, creating climate refugees, resource wars, and submerging major cities around the globe. However, unlike previous challenges to our way of life, the threat of climate change will not manifest in a single event, rather in a series of natural disasters that will eventually escalate to a point where we no longer have the resources to manage them. This is being seen already in California as wildfires sweep the state, forcing people to relocate, causing issues with access and use of electricity and causing an estimated $10 billion in damages. Thus, in our project, we plan to determine what are the strongest environmental predictors of forest fires by looking at data from California.

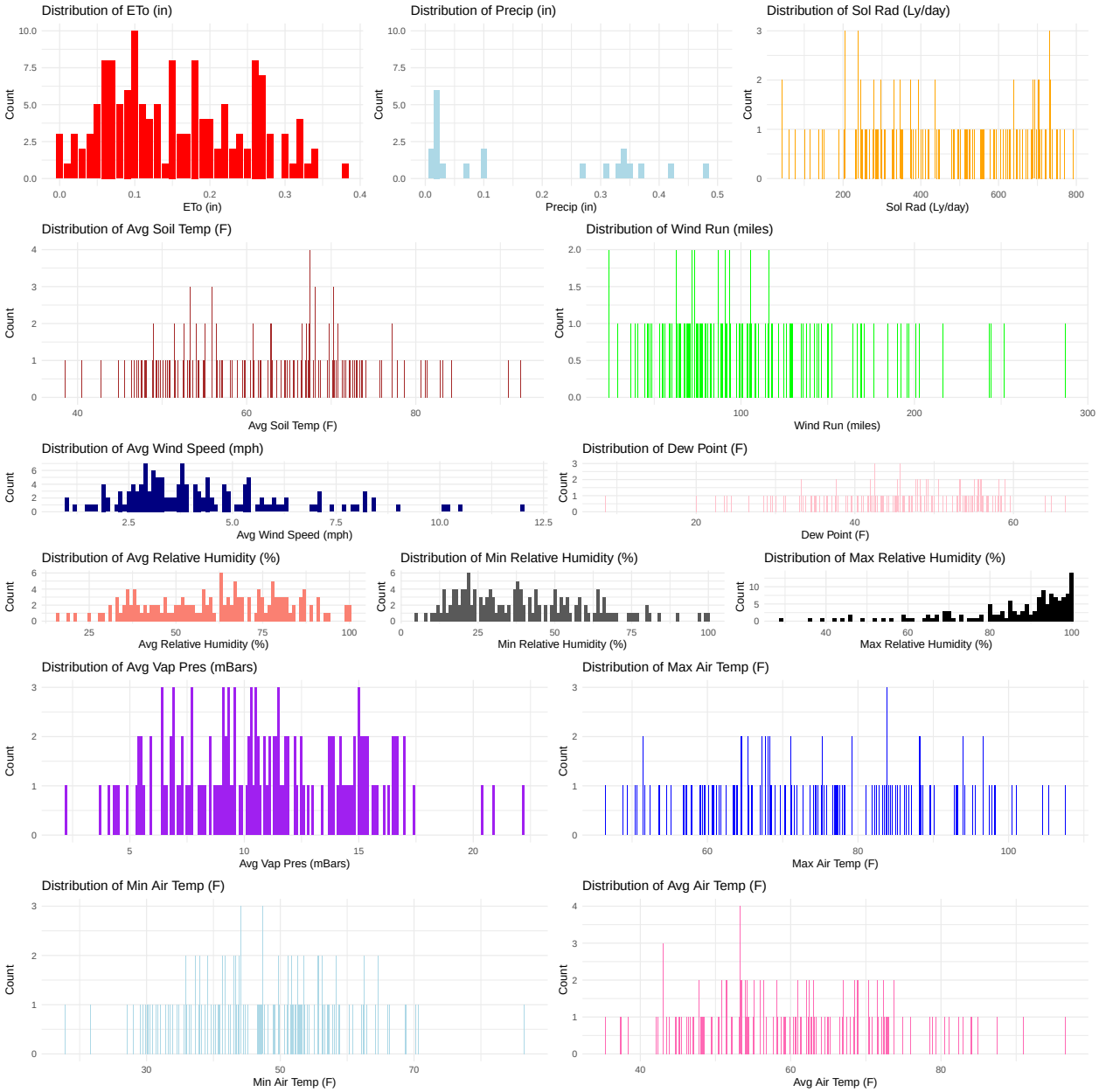
It is important to understand what these predictors are to determine how to prevent short term fires from escalating and to correct these conditions in the long run to see if it is possible to mitigate the threat of these fires going forward. If we know what predictors play a role in a fire, it can help authorities better understand what days or seasons present a higher risk, and thus prepare accordingly. While climate change will continue to affect our way of life, insights into how to manage its consequences and effects will help us plan for both the short term and long term future, at least until policy and research catch up to the severity of the issue.

The data we will be using was scraped from CIMIS (California Irrigation Management Information System) weather stations by github user czaloumi using a selenium chromedriver. The dataset was combined with Wikipedia tables listing California fires by county and city to create the Target column, which indicates whether or not there was a fire on a particular day. Additionally, the curator of the dataset adds that this dataset was "used in conjunction to building an XGBoost Classifier to accurately predict probability for fire given environmental condition feature." This user's data contains a mixture of environmental and geospatial data to understand the size and the scope of the forest fires, as well as where the fires seem to be most frequent.

However, due to the size and scope of our data set, we decided to adjust accordingly. Each of the stations in our data contains data between 2018 and summer 2020. To correct for this, we will be randomly selecting a day from each of the sections within this range, and focusing more so on what each station reported from its respective day. This will give us 153 observations instead of 128,126. We decided that trimming it down would not only reduce the data, but also reduce the noise.

From day-to-day in a station it is unlikely that there is substantial variation, however when comparing station to station over different days, we are able to see different variables from all times of the year. However, there was still some missing data, so our total number of observations has been further reduced to 143, from stations all across the state of California.

Our first step of our exploratory data analysis was to look at the shape of the distributions of each variable. This would give us a sense about which transformations might be necessary.



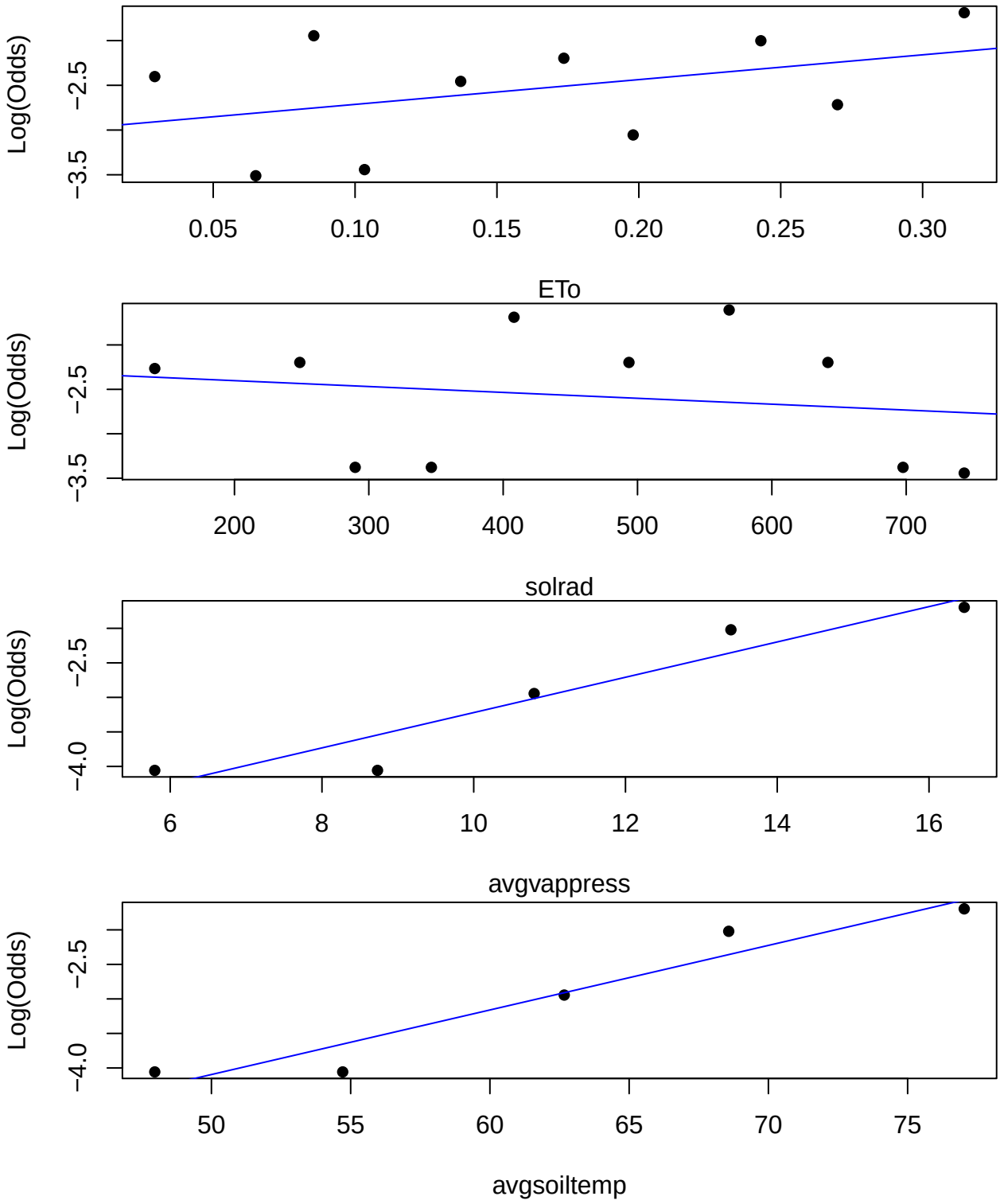
In general it appears as thought the data points tend to follow a normal distribution. However, not all normal distributions are made equal. Maximum relative humidity appears to be left skewed, while precipitation appears to be right skewed. While it appears that some have multiple peaks and unique spreads and distributions this is likely due to limited data points.

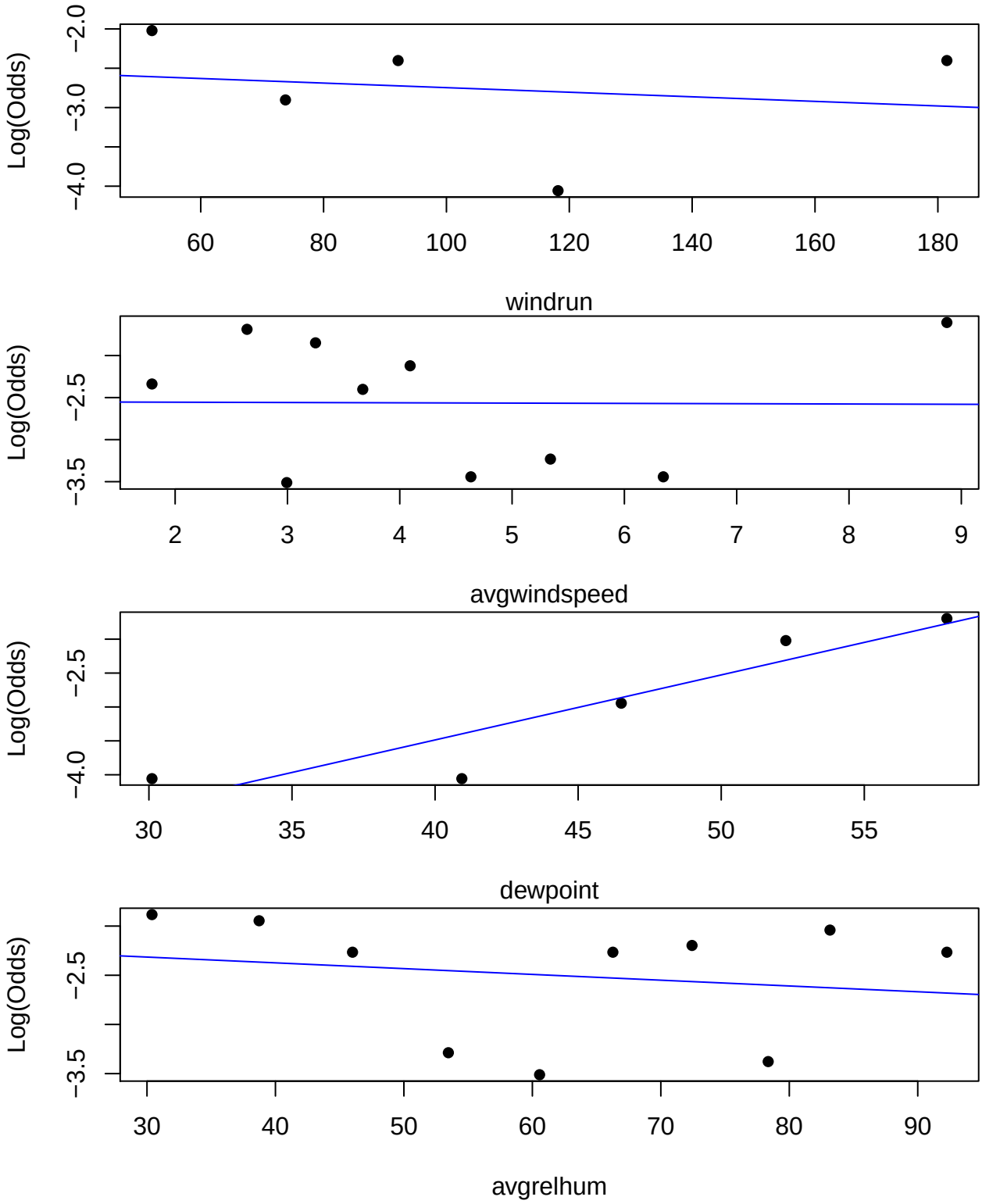
Methodology

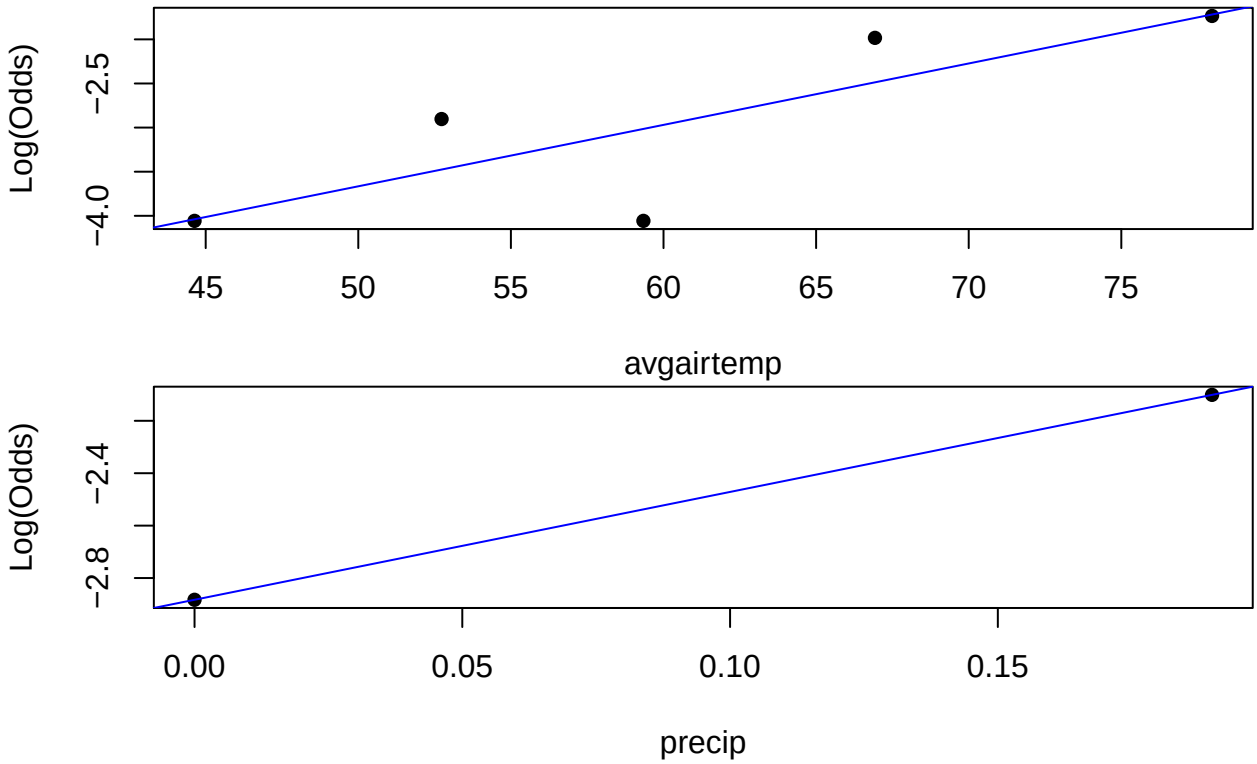
To reduce the multicollinearity, we decided to focus on only specific variables. For example, the dataset included the average, minimum and maximum for a number of variables. We determined that average was likely most relevant to the conditions recorded by each station throughout the day.

Since we are trying to predict with a binary response variable, we will use a logisitic regression. However, to be able to use a logisitic regression we need to check the conditions, and ensure that the data satsifies linearity, randomness and independence.

To check linearity, we will make an empirical logisrtic regression plot for each of the predictor variables.


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



After looking at these graphs, it is apparent that average wind speed, and average relative humidity do not follow a linear relationship. Furthermore, linearity is only satisfied for precipitation when there are only 2 groups. This means that a linear model might not be an appropriate estimation for precipitation.

A lot of the data in the dataset is conditional instead of random. The conditions recorded by one station will likely be similar to those in a nearby station and similar to the recordings the day before. To correct for that, we took a random sample of the days in our data set. This also supports independence. Our cleaning of the data to include a random sample of days and stations.

Now that we have cleaned the data and satisfied the conditions of logistic regression, we began to build prediction models, and eliminated the variables that do not satisfy the conditions. Then we will use backward selection to reduce the data.

```
## # 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
```

Now that we see the variation inflation factors of each of the variables, we can reduce the multicollinearity by eliminating data points where one has a similar VIF to another. Thus we removed solar radiation and dew point, due to its multicollineaity with average air temperature and ETo respectively.

term	estimate	std.error	statistic	p.value
(Intercept)	-20.653	5.071	-4.073	0.000
ETo	-32.393	12.171	-2.661	0.008
avgvappress	0.015	0.152	0.097	0.922
avgsoiltemp	0.177	0.072	2.450	0.014
windrun	0.012	0.010	1.270	0.204
avgairtemp	0.147	0.083	1.759	0.079

```
##         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
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-18.3648054	4.4576757	-4.119816	0.0000379	-28.6854177	-10.7738913
ETo	-26.2762823	10.0385947	-2.617526	0.0088570	-48.9850407	-8.6686880
avgsoiltemp	0.1689266	0.0701338	2.408635	0.0160123	0.0365751	0.3196791
avgairtemp	0.1276049	0.0755239	1.689597	0.0911051	-0.0111881	0.2935839

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

AIC	BIC
48.074	59.925

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

term	estimate	std.error	statistic	p.value
(Intercept)	-17.172	7.616	-2.255	0.024
ETo	-34.759	14.058	-2.472	0.013
avgvappress	0.071	0.295	0.241	0.810
avgsoiltemp	0.175	0.073	2.389	0.017
windrun	-0.011	0.029	-0.395	0.693
avgairtemp	0.119	0.124	0.956	0.339
I(avgwindspeed^2)	0.056	0.059	0.951	0.342
I(avgrelhum^2)	0.000	0.001	-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
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-19.954	4.765	-4.187	0.000	-31.126	-11.900
ETo	-33.669	11.664	-2.887	0.004	-60.497	-13.382
avgsoiltemp	0.178	0.070	2.538	0.011	0.048	0.333
avgairtemp	0.149	0.079	1.879	0.060	0.004	0.323
I(avgwindspeed^2)	0.032	0.021	1.550	0.121	-0.008	0.076

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
61.678	142	-18.778	47.555	62.369	37.555	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.

term	estimate	std.error	statistic	p.value
(Intercept)	-36.232	34.607	-1.047	0.295
ETo	-38.212	14.943	-2.557	0.011
avgsoiltemp	0.416	0.373	1.116	0.264
avgairtemp	0.420	0.507	0.828	0.408
avgvappress	0.037	0.450	0.083	0.934
avgwindspeed	-18.677	20.135	-0.928	0.354
windrun	0.764	0.837	0.914	0.361
avgrelhum	-0.017	0.117	-0.142	0.887
avgsoiltemp:avgairtemp	-0.004	0.005	-0.665	0.506
avgwindspeed:windrun	0.003	0.002	1.146	0.252

Resid..Df	Resid..Dev	df	Deviance	p.value
137	38.257	NA	NA	NA
133	35.158	4	3.099	0.541

Due to a high p-value, the data provided is insufficient to provide evidence to conclude that the interaction terms are statistically significant.

After all these analyses/tests, our ultimate best model is new_fire_model.

Since new_fire_model appears to be our strongest model, that means that ETO, average air temperature and average soil temperature are the most significant environmental predictors of forest fires in California.

While this might mean that monitoring these three will provide a reduction strategy to fires, it creates some ethical questions. ETO is ethylene oxygen is a flammable colorless gas, that can not be simply removed from the atmosphere without introducing another agent. Furthermore, managing average temperature also provides a number of practical issues, as does managing soil temperature. Thus, is it too late to be able to do anything effectively to prevent or can fire fighting measures only be reactive from this point forward.

———— 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