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`

```
## # A tibble: 128,125 x 19
##    `Stn Id` `Stn Name` `CIMIS Region` Date  `ETo (in)` `Precip (in)`
##       <dbl> <chr>      <chr>          <chr>      <dbl>         <dbl>
##  1        2 FivePoints San Joaquin V~ 1/1/~       0.06          0
##  2        2 FivePoints San Joaquin V~ 1/2/~       0.04          0
##  3        2 FivePoints San Joaquin V~ 1/3/~       0.04          0
##  4        2 FivePoints San Joaquin V~ 1/4/~       0.07          0.01
##  5        2 FivePoints San Joaquin V~ 1/5/~       0.07          0
##  6        2 FivePoints San Joaquin V~ 1/6/~       0.02          0.04
##  7        2 FivePoints San Joaquin V~ 1/7/~       0.06          0
##  8        2 FivePoints San Joaquin V~ 1/8/~       0.01          0.38
##  9        2 FivePoints San Joaquin V~ 1/10~       0.05          0
## 10        2 FivePoints San Joaquin V~ 1/11~       0.02          0
## # ... with 128,115 more rows, and 13 more variables: `Sol Rad (Ly/day)` <dbl>,
## #   `Avg Vap Pres (mBars)` <dbl>, `Max Air Temp (F)` <dbl>, `Min Air Temp
## #   (F)` <dbl>, `Avg Air Temp (F)` <dbl>, `Max Rel Hum (%)` <dbl>, `Min Rel Hum
## #   (%)` <dbl>, `Avg Rel Hum (%)` <dbl>, `Dew Point (F)` <dbl>, `Avg Wind Speed
## #   (mph)` <dbl>, `Wind Run (miles)` <dbl>, `Avg Soil Temp (F)` <dbl>,
## #   Target <fct>
```

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.

Thus the research question we want to answer is: what are the strongest environmental predictors of forest fires in California?

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.

Our Data

There are 128,126 observations in the data set. Each observation represents information on the weather conditions at a given weather station on a specific date.

The response variable we will be investigating is Target, which corresponds to fires on the respective observation date, in the observation region. The Target variable is a binary indicator, with a value of 1 indicating there was a fire and a value of 0 indicating there was not a fire.

Our potential predictor variables are:

`ETo` - The ETo variable measures the average amount of evapotranspiration present in the soil in each of the regions. This means that it is the amount of water transferred to the land by means of plants.

`precip` - The precip variables measure the long term monthly average amount of precpitation found in the each station's region.

`solrad` - The solrad variables measure the average amount of solar radiation found in the each station's region.

`avgvappress` - The avgvappress variables measure the average amount of vapor pressure found in the each station's region.

`avgsoiltemp` - The avgsoiltemp variables measure the average soil temperature found in the each station's region.

`windrun` - The windrun variables measures the sum of wind speed over a month.

`avgwindspeed` - The avgwindspeed variables measure the average wind speed found in the each station's region.

`dewpoint` - The dewpoint variables measure the average temperature of the dew on the grass in each station over a month long period.

`avgrelhum` - The avgrelhum variables measure the average relative humidity found in the each station's region.

`avgairtemp` - The avgairtemp variables measure the monthly average of the air temperature found in the each station's region.

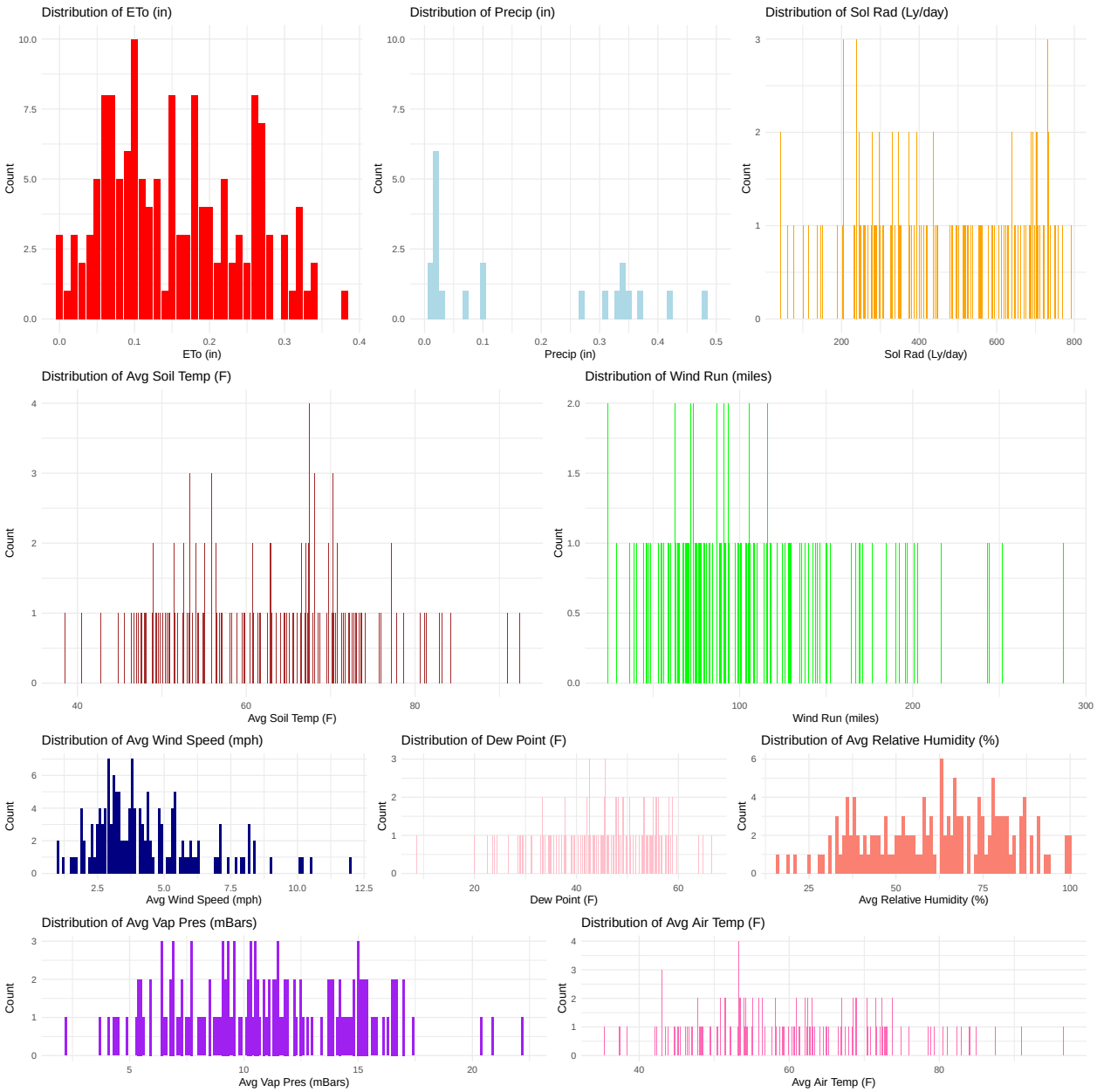
Exploratory Data Analysis

Each of the stations in our data contains data between 2018 and summer 2020. To account for this, we randomly selected a day from each of the sections within this range and focused more so on what each station reported from its respective day. This gave us 153 observations instead of 128,126. Using a random sample of the data is an effective strategy to handle this data set for a myriad of reasons. Firstly, my reducing the data set to a small training set, we are able to make a model that is a more realistic approximation of the conditions that might cause a fire. This is a good approach to this data because due to the vast number of data sets, and differing climates across the state of California, a random sample will most likely allow us to separate the signal from the noise. Secondly, taking a smaller, random sample of the data allows us to ensure that randomness is satisfied for the logistic regression. This is important because the data has to be randomly distributed for a logistic regression model to be applicable. Finally, taking a random sample would

ensure that the data satisfies independence, because whether a fire is reported or not is no longer conditional on surrounding stations, as each station is from different days and from different times of the year. Thus, while the entire dataset does not satisfy independence, a random sample would meet this requirement.

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. We decided to reduce further because the missing data was missing not at random. Since the data was not at random, it did not make sense to mean-impute the values for the missing observations. Thus by removing these entries, we can ensure that our model is an accurate reflection of our random sample.

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 though the data points tend to follow a normal distribution. However, not all normal

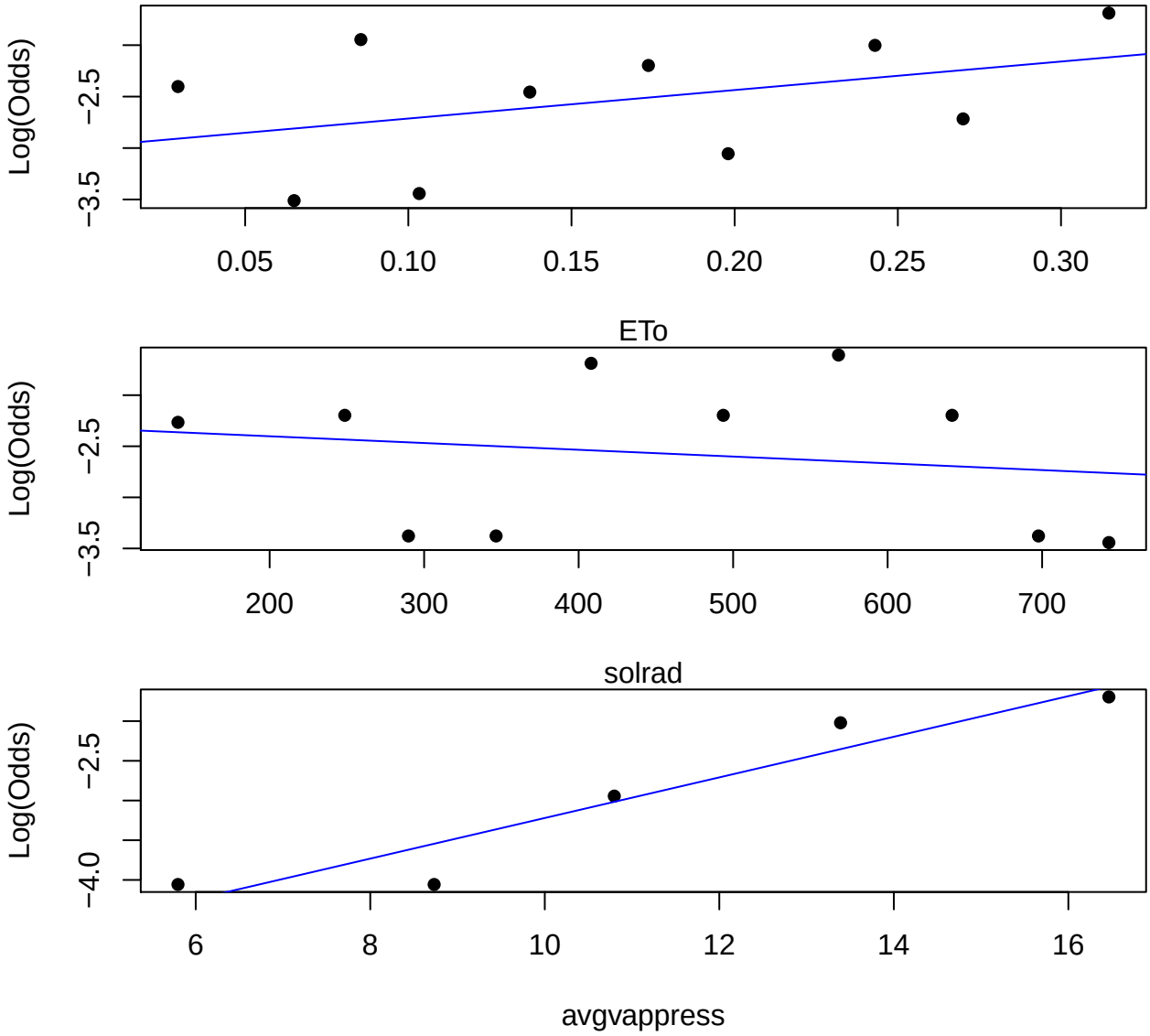
distributions are made equal. Maximum relative humidity appears to be left skewed, while precipitation appears to be right skewed. While it appears that some have multiple peaks and unique spreads and distributions this is likely due to limited data points.

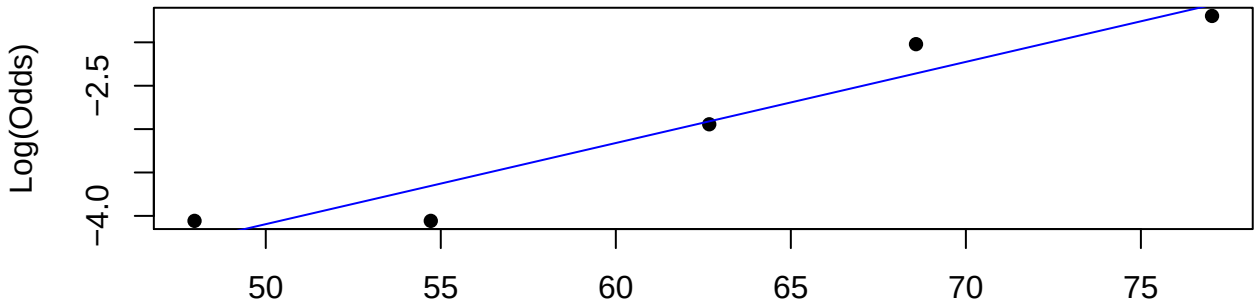
Methodology

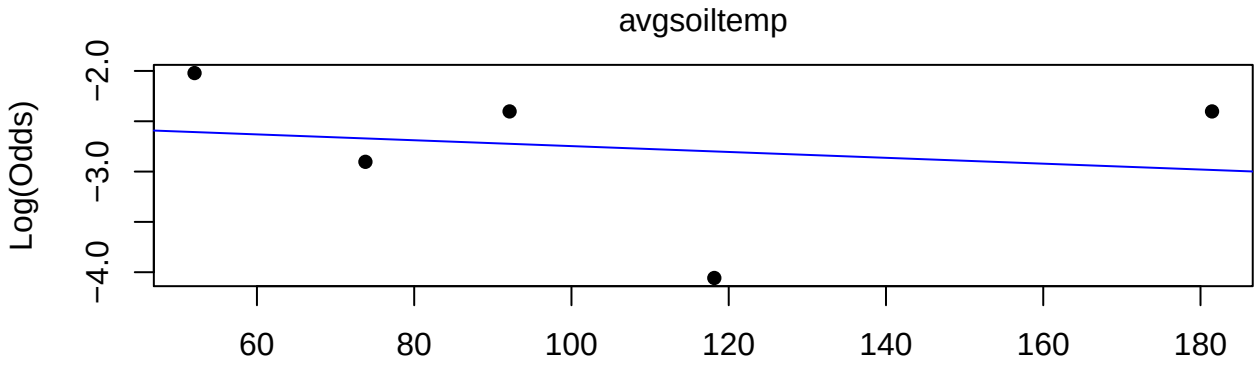
To reduce multicollinearity, we decided to focus on only specific variables. For example, the dataset included the average, minimum and maximum for a number of variables. We determined that average was likely most relevant to the conditions recorded by each station throughout the day.

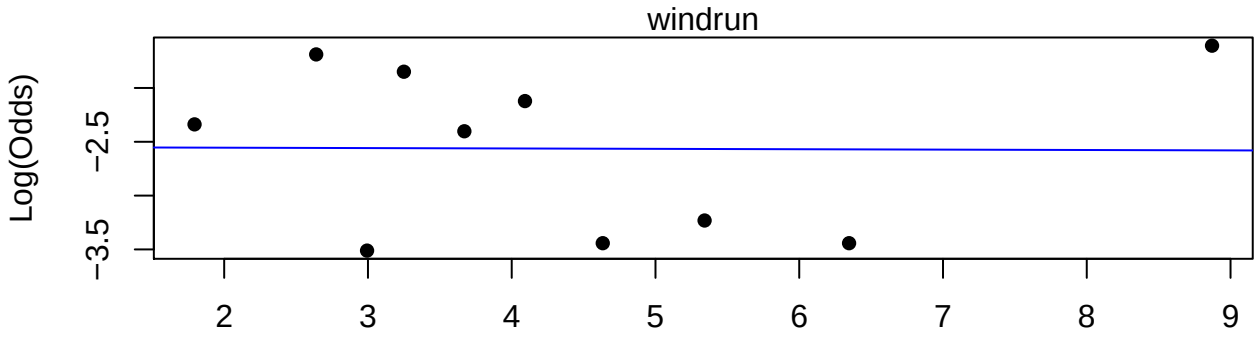
Since we are trying to predict with a binary response variable, we will use
logistic regression. However, to be able to use logistic regression we need to check the conditions, and ensure that the data satisfies linearity, randomness and independence.

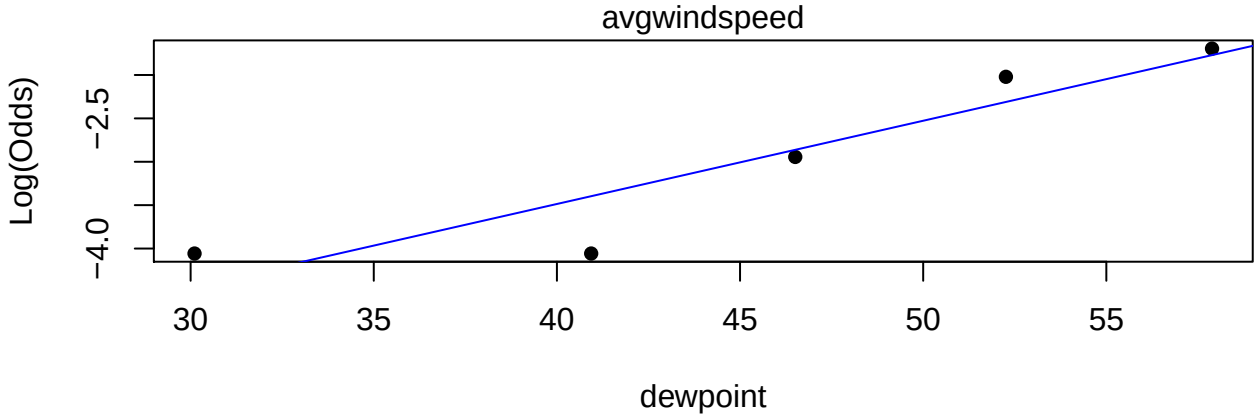
To check linearity, we will make an empirical logistic regression plot for each of the predictor variables.

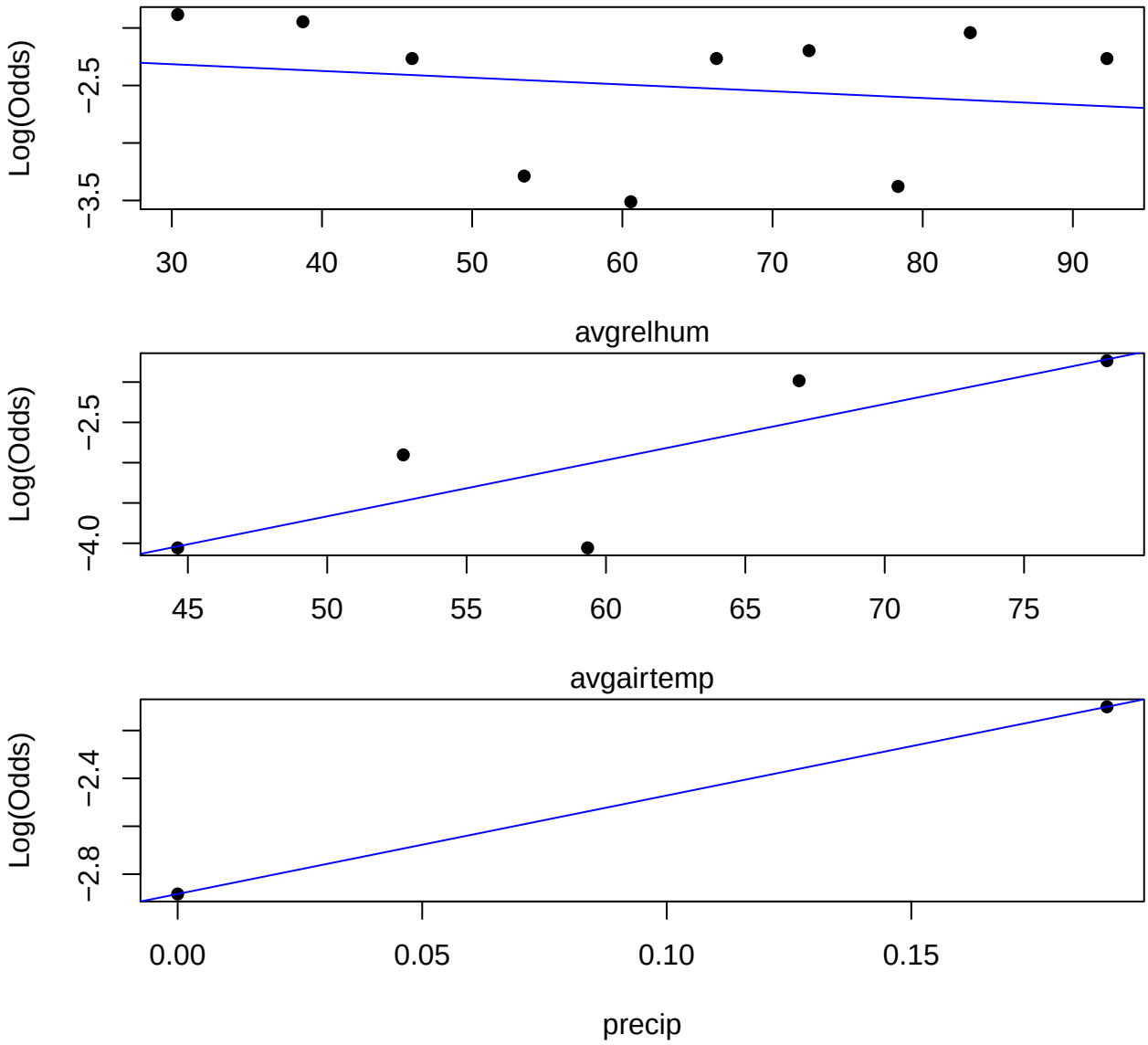



Log(Odds)
−4.0
−2.5
50
55
60
65
70
75


avgsoiltemp
Log(Odds)
−4.0
−3.0
−2.0
60
80
100
120
140
160
180


windrun
Log(Odds)
−3.5
−2.5
2
3
4
5
6
7
8
9


avgwindspeed
Log(Odds)
−4.0
−2.5
30
35
40
45
50
55
dewpoint



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.

#maybe move this to EDA where we actually do this process Much 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.

We then began to construct prediction models. We first started with a main effects model containing all possible predictors.

term	estimate	std.error	statistic	p.value
(Intercept)	-23.620	28.495	-0.829	0.407
ETo	76.380	51.373	1.487	0.137
solrad	-0.033	0.016	-2.109	0.035
avgvappress	-1.154	2.195	-0.526	0.599
avgsoiltemp	0.210	0.082	2.572	0.010

term	estimate	std.error	statistic	p.value
windrun	-0.020	0.019	-1.066	0.286
dewpoint	0.828	1.183	0.700	0.484
avgairtemp	-0.253	0.212	-1.190	0.234

Because the only significant term in the model is avgsoiltemp (the only term with an associated p.value of less than 0.05), we suspected multicollinearity. We used vif to further investigate this idea.

	x
ETo	126.467
solrad	56.605
avgvappress	115.670
avgsoiltemp	3.992
windrun	3.303
dewpoint	129.184
avgairtemp	30.659

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 multicollinearity with average air temperature and ETo respectively. We then constructed a new model (main_fire_model) without the variables with high multicollinarity.

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

	x
ETo	7.073
avgvappress	1.401
avgsoiltemp	3.218
windrun	1.446
avgairtemp	5.996

Because these values are 1) dissimilar from each other or 2) generally small, we can conclude that we have removed highly correlated variables from analysis.

Next, using backwards selection from main_fire_model, we constructed new_fire_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

AIC	BIC
48.074	59.925

write out new_fire_model here

We then decided to try out various variable transformations to potentially bolster the predictive power of our model.

Looking at the shapes of the empirical logit plots, average wind speed and average relative humidity have slightly parabolic distributions. So, we added quadratic transformations of variables avgwindspeed and avgrelhum as predictors and fit a new model.

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

term	estimate	std.error	statistic	p.value	conf.low	conf.high
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

Adding the quadratic transformed variables, AIC has marginal improvement, but BIC is larger. So, we will not add the quadratic transformed variables in our final model.

Next, we explored potentially meaningful interaction terms. We ultimately chose to test the only interaction term which seemed meaningful: ETo*avgwindspeed. We inferred that large amounts of ETo in the soil and high wind speed would significantly increase the predicted probability of a forest fire because a lot of fast-moving wind might spread the water and thus make it harder for a fire to develop in the area that the water was spread, but possibly easier for a fire to begin in a region with less plamt based transpiration.

To determine if this interaction term is statistically significant, we added it to the main_model with the main effects of all the initial predictor variables and then conducted a drop-in-deviance test between the two models.

term	estimate	std.error	statistic	p.value
(Intercept)	-19.073	14.261	-1.337	0.181
ETo	-65.856	23.964	-2.748	0.006
avgwindspeed	-43.128	25.569	-1.687	0.092
avgsoiltemp	0.235	0.089	2.654	0.008
avgairtemp	0.160	0.186	0.861	0.389
avgvappress	0.056	0.422	0.132	0.895
windrun	1.753	1.053	1.664	0.096
avgrelhum	-0.020	0.110	-0.186	0.852
ETo:avgwindspeed	5.981	2.661	2.248	0.025

Resid..Df	Resid..Dev	df	Deviance	p.value
137	38.257	NA	NA	NA
134	30.528	3	7.729	0.052

The p-value of the drop-in-deviance test is slightly higher than the conventional alpha level of 0.05, which means the data do not provide sufficient evidence to suggest that the interaction term is statistically significant. So, we will not include the interaction term in our final model.

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

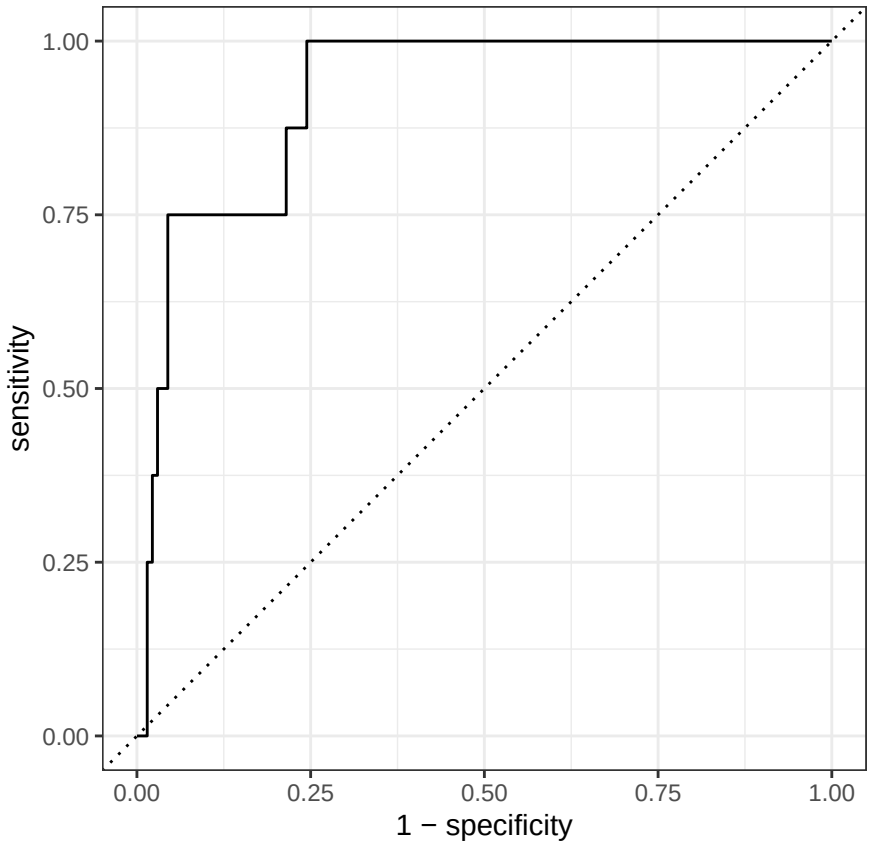
do we have to show it again? or is just writing it out good enough?

term	estimate	std.error	statistic	p.value
(Intercept)	-18.365	4.458	-4.120	0.000
ETo	-26.276	10.039	-2.618	0.009
avgsoiltemp	0.169	0.070	2.409	0.016
avgairtemp	0.128	0.076	1.690	0.091

$$\hat{Target} = -18.365 - 26.276ETo + 0.169avgsoiltemp + 0.128avgairtemp$$

Conclusion

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. To test our model, we first constructed an ROC curve to identify a prediction threshold.



From this ROC curve, we selected a prediction threshold of X, identified by minimizing 1-specificity while maximizing sensitivity. Because there is greater risk in failing to predict a fire (type 2 error), we were less interested in the false positive rate as opposed to the sensitivity.

```
## # A tibble: 5 x 4
##   .threshold specificity sensitivity false_rate
##        <dbl>       <dbl>       <dbl>      <dbl>
## 1    -Inf        0                1      1
## 2      -9.82     0                1      1
## 3      -9.56     0.00741          1      0.993
## 4      -9.18     0.0148           1      0.985
## 5      -9.02     0.0222           1      0.978
```

```
## # A tibble: 7 x 4
##   .threshold specificity sensitivity false_rate
##        <dbl>       <dbl>       <dbl>      <dbl>
## 1      -3.05       0.756       0.875      0.244
## 2      -3.05       0.763       0.875      0.237
## 3      -3.04       0.770       0.875      0.230
## 4      -2.98       0.778       0.875      0.222
```

```
## 5      -2.97       0.785       0.875       0.215
## 6      -2.97       0.785       0.75        0.215
## 7      -2.91       0.793       0.75        0.207
```

((ROC curve to identify prediction threshold))

((try model on a test dataset?))

Discussion

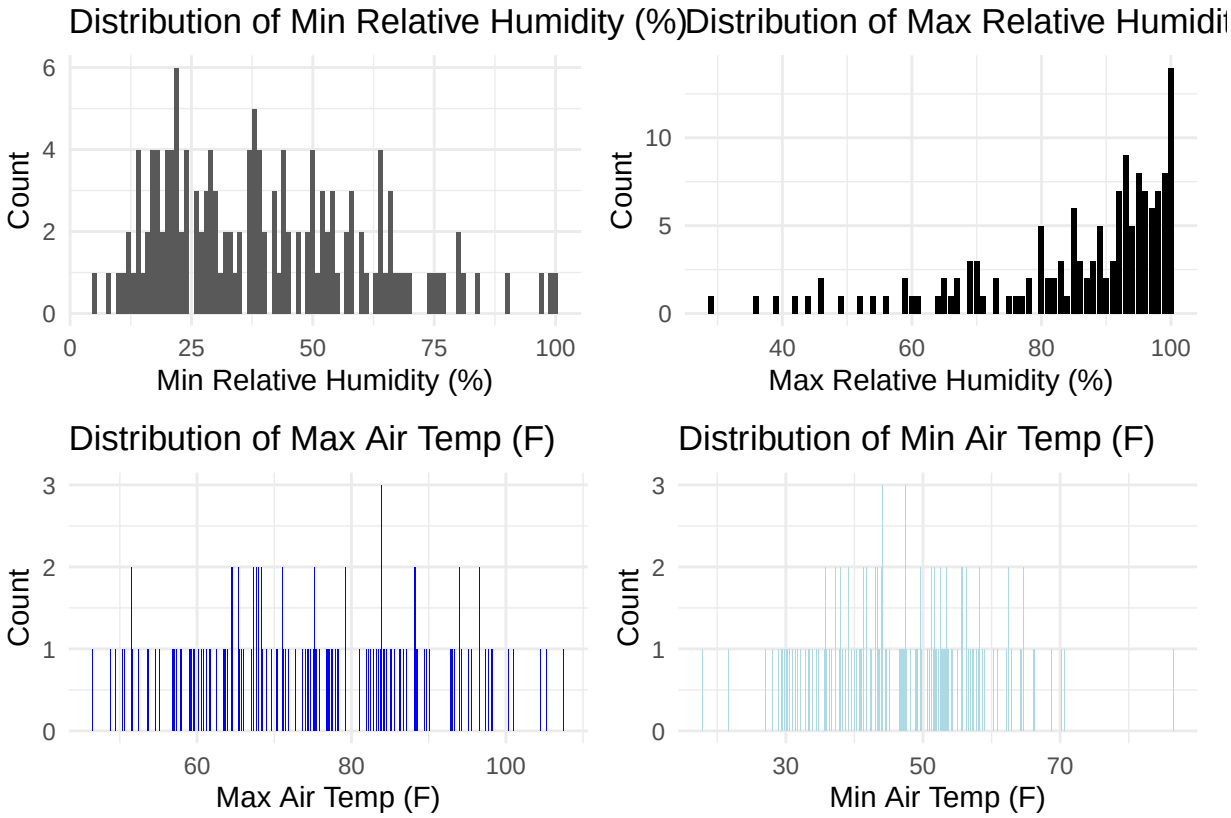
Since new_fire_model appears to be our strongest model, that means that ETo, average air temperature and average soil temperature are the most significant environmental predictors of forest fires in California. While this might mean that monitoring these three will provide a reduction strategy to fires, it creates some ethical questions. ETo, ethylene oxygen, is a flammable colorless gas that cannot 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, our model suggest that it is likely too late to be able to do anything effectively to prevent or can firefighting measures only be reactive from this point forward.

The reliability and validity of our data certainly comes into question. As previously stated, a single data point was randomly chosen from each station (as the data spans multiple years and the goal was to reduce multicollinearity as much as possible). However, this method is not foolproof. Stations that are spatially close together and whose randomly selected dates are close together are not screened for in our data selection process. With more time, this data selection process would be further refined to ensure data points are as independent as possible.

Additionally, we only considered one potentially meaningful interaction term, ETo*avgwindspeed, throughout our analysis. In an expanded version of this project, we would potentially explore more interaction terms, as this single term was ultimately left out of the model.

Though we considered both AIC and BIC throughout our analysis, we were partial to BIC, favoring a parsimonious model. The result is our final model with only three terms. With more time, we could construct 1) a model with AIC selection criterion and 2) a model with BIC selection criterion and compare the two on a new randomly selected set of data points to identify which has greater prediction accuracy.

Appendix



——— 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 multicollinearity w min and max air temp

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