forcasting forest fires

fantastic fouR: Eli Levine, Hannah Long, Preetha Ramachandran

Oct. 26, 2020

Introduction

Climate change represents an existential threat to humanity. Its effects have the potential to dramatically shape life as we know it by creating climate refugees, resource wars, and submerging major cities around the globe (citation). 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 these crises. 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 (citation). Thus, in our project, we are trying 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 for multiple reasons. Firstly, in the short run, we will be able to identify the conditions that make forest fires more likely. Thus, we can try to correct these conditions in the long run and see if it is possible to mitigate the threat of forest fires in the future. If we know what predictors play a role in predicting fires, 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 immediate 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 are 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 are 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 variable measures the monthly average amount of precpitation found in the each station's region in the days prior to the recording.

`solrad` - The solrad variable measures the average amount of solar radiation found in the each station's region in the prior days.

`avgvappress` - The avgvappress variable measures the average amount of vapor pressure found in the each station's region in the days leading up to the recorded day.

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

`windrun` - The windrun variable measures the sum of wind speed over a month long period.

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

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

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

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

Exploratory Data Analysis

As previously stated, the dataset contains observations of weather conditions and indicates the presence of a fire on a specific date at a certain weather station in California. Each of the stations in our dataset has recorded observations of these weather conditions between 2018 and summer 2020. Because the conditions recorded by one station will likely be similar to those in a nearby station and similar to the recordings the day before, we simulate independence by filtering our data.

First, we group observations by station id number. Next, one day is chosen at random from each of the stations. This gives us 153 observations instead of the original dataset of 128,126 observations.

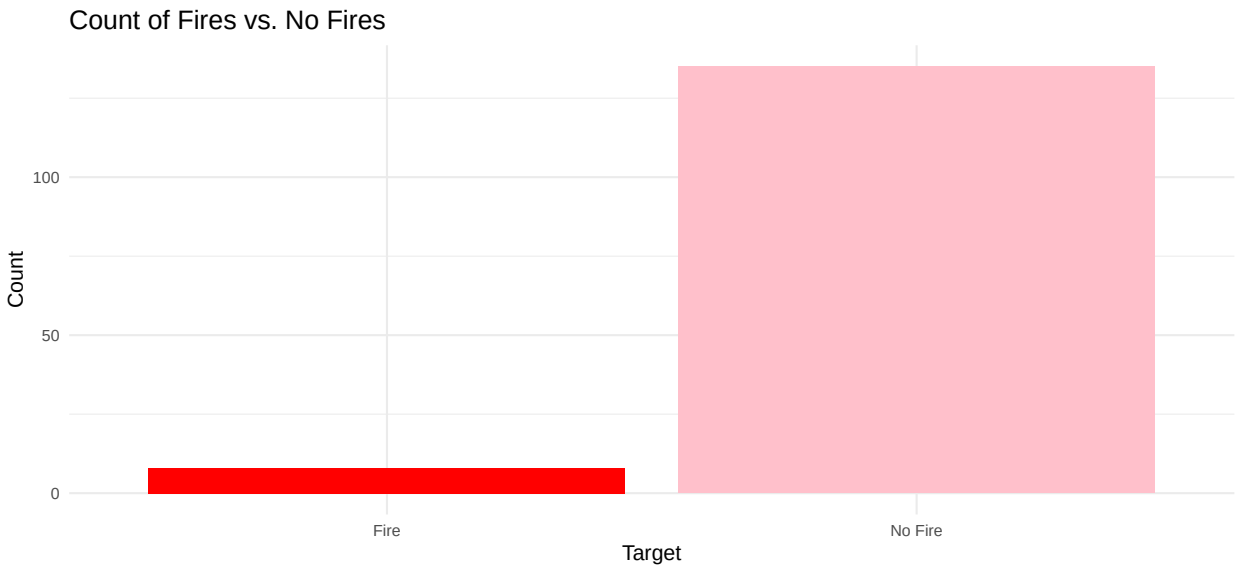
By reducing the data set to a small training set, we are able to make a model that is a more realistic approximation of the conditions that might cause a fire, sans the correlation that comes with keeping all of the data. Additionally, constructing a model with a random sample of the data allows us to test our final model on a smaller sample of data to assess its predictive power. In addition, taking a smaller, random sample of the data allows us to ensure that randomness is satisfied for our analysis.

Furthermore, taking a random sample ensures that the data satisfies independence, because whether a fire is reported or not is no longer conditional on surrounding stations and surrounding environmental variables, since each station is from different days and from different times of the year. Thus, while the dataset in its entirety does not satisfy independence, our random sample meets this criteria.

Even after reducing our dataset to 153 random observations, it is clear that there are still some stations with missing observations in some variable categories. We have decided to use only complete observations in the analysis, and thus, the total number of observations is further reduced to 143.

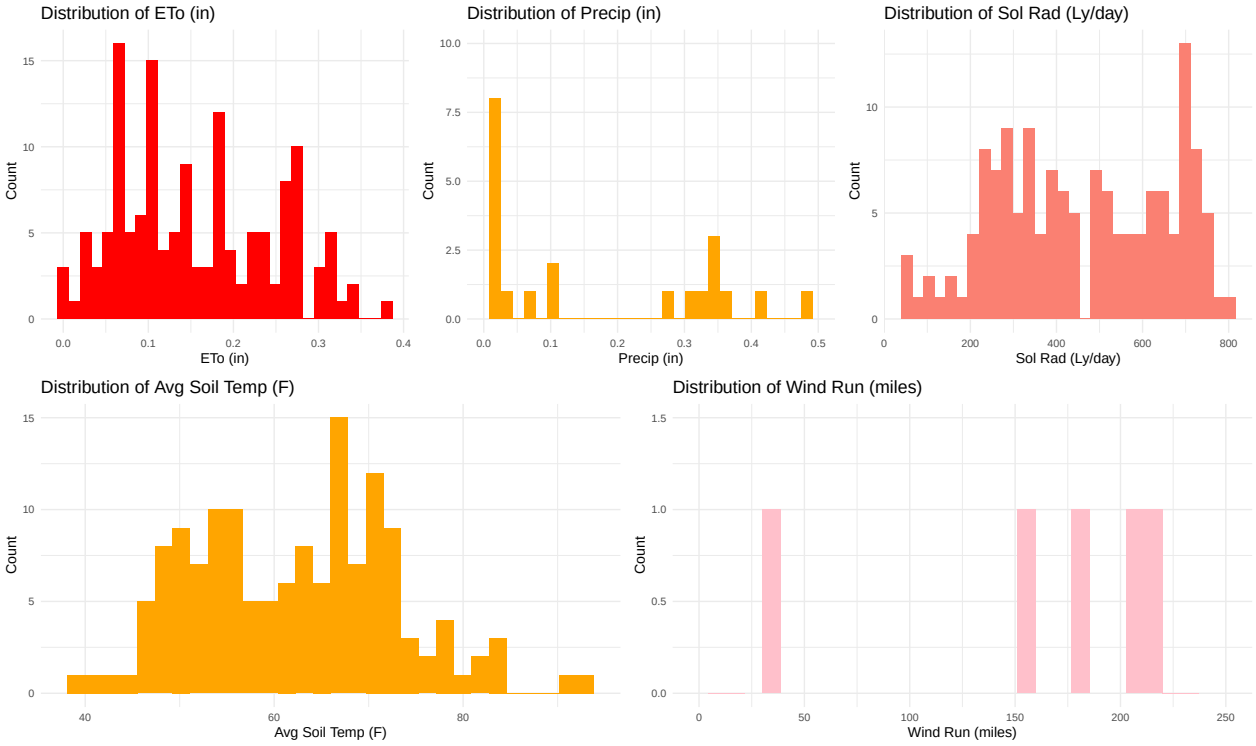
Glimpsing the data, it does not appear that the observations with missingness differ systematically from the complete observations; it is thus unlikely that our resulting analysis is biased by the decision to remove the data.

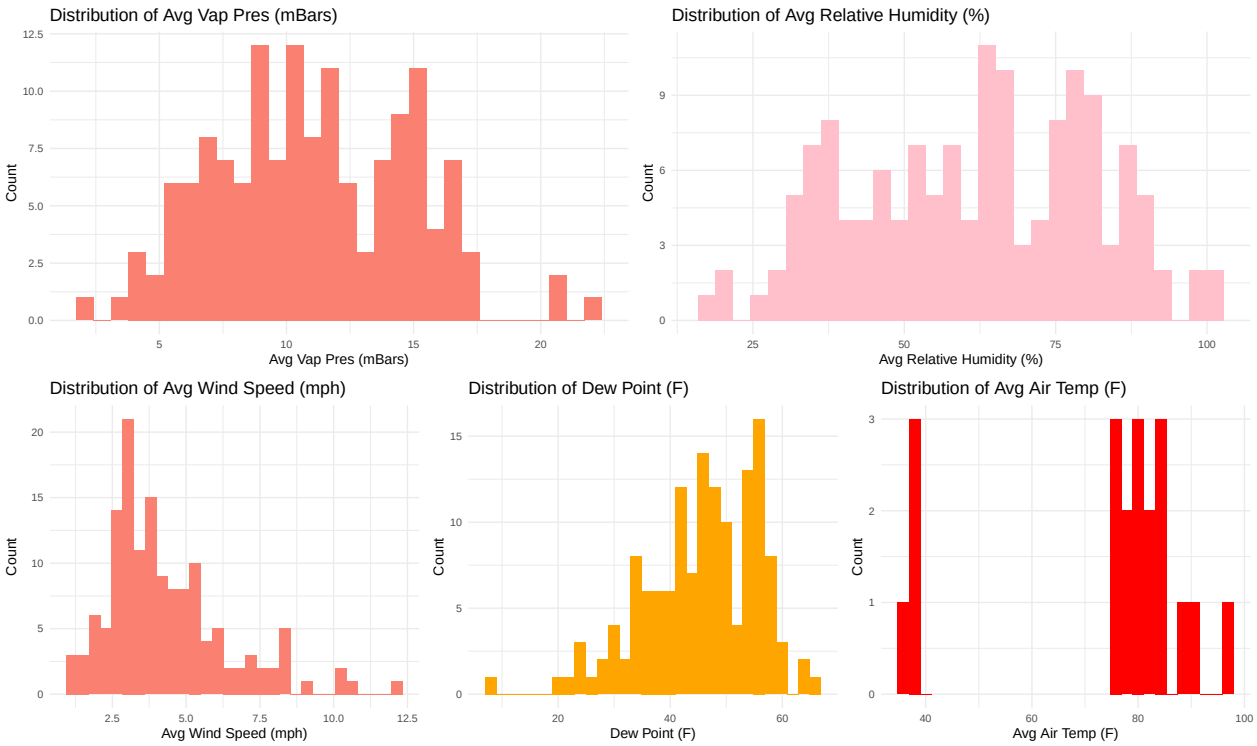
Our first step of our exploratory data analysis is to look at the distribution of the response variable, Target, where 1 indicates a fire and 0 indicates no fire.



It's clear that our sampled data includes significantly more "No Fire" observations as opposed to "Fire" observations.

The next step is to look at the shape of the distributions of each potential predictor variable. This gives us a better understanding of our data and hints at which distributions are exceptionally non-normal or otherwise grossly affected by our data sampling procedure.





Some of the variables in our dataset are not shown here as they were not used for analysis. The plots for these variables can be found in Appendix A along with an explanation for their elimination.

Precipitation and average wind speed appear to be right skewed, ETo is roughly trimodal, average relative humidity is bimodal, and dew point is roughly left skewed. Across the board, the histograms are far from normal, the multiple peaks and unique spreads and distributions are likely due to limited data points and California's vast geographic diversity, a hypothesis made clear by the low counts shown on each graph.

Methodology

We are trying to predict the presence of a fire with a binary response variable, Target (which indicates the presence or non presence of a fire); therefore we will use a logistic regression for our analysis.

After identifying our regression method, we begin to construct prediction models. We begin first with a main effects model containing all possible predictors.

term	estimate	std.error	statistic	p.value
(Intercept)	10.675	35.013	0.305	0.760
ETo	188.028	87.104	2.159	0.031
solrad	-0.084	0.032	-2.581	0.010
avgvappress	-1.638	2.454	-0.667	0.505
avgsoiltemp	0.372	0.158	2.359	0.018
windrun	5.004	2.377	2.106	0.035
avgwindspeed	-121.460	57.515	-2.112	0.035
avgrelhum	-0.430	0.305	-1.411	0.158
precip	2.630	7.298	0.360	0.719
dewpoint	1.961	1.527	1.284	0.199
avgairtemp	-1.239	0.678	-1.827	0.068

Because the only significant term in the model is the term for average soil temperature (the only term with

an associated p.value of less than 0.05), we suspect multicollinearity among variables. We use vif to further investigate this issue.

	x
ETo	213.555
solrad	137.988
avgvappress	96.436
avgsoiltemp	6.590
windrun	27969.818
avgwindspeed	28326.883
avgrelhum	102.907
precip	1.497
dewpoint	156.484
avgairtemp	184.489

We remove the variable representing the sum of wind speed over the month (windrun) due to its multicollinarity with average wind speed, as indicated by the large and similar vif values for both. Average relative humidity and dewpoint are removed as well due to multicollinarity with average vapor pressure and average air temperature, respectively.We then construct a new model (main_fire_model) without the aforementioned variable.

term	estimate	std.error	statistic	p.value
(Intercept)	-5.392	7.582	-0.711	0.477
ETo	66.411	47.197	1.407	0.159
solrad	-0.031	0.015	-2.075	0.038
avgvappress	0.432	0.316	1.365	0.172
avgsoiltemp	0.218	0.082	2.650	0.008
avgwindspeed	-0.430	0.429	-1.003	0.316
precip	1.430	4.442	0.322	0.747
avgairtemp	-0.203	0.180	-1.128	0.259

	x
ETo	104.092
solrad	49.728
avgvappress	4.094
avgsoiltemp	4.308
avgwindspeed	2.944
precip	1.245
avgairtemp	23.040

Because the remaining vif 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 construct new_fire_model. This was done to remove unnecessary variables and thus improve the model's predictive ability.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-15.2290556	3.5724080	-4.262967	0.0000202	-23.5649146	-9.1682137
solrad	-0.0118340	0.0037360	-3.167531	0.0015374	-0.0205131	-0.0053647

term	estimate	std.error	statistic	p.value	conf.low	conf.high
avgsoiltemp	0.2580629	0.0646067	3.994368	0.0000649	0.1472155	0.4076018

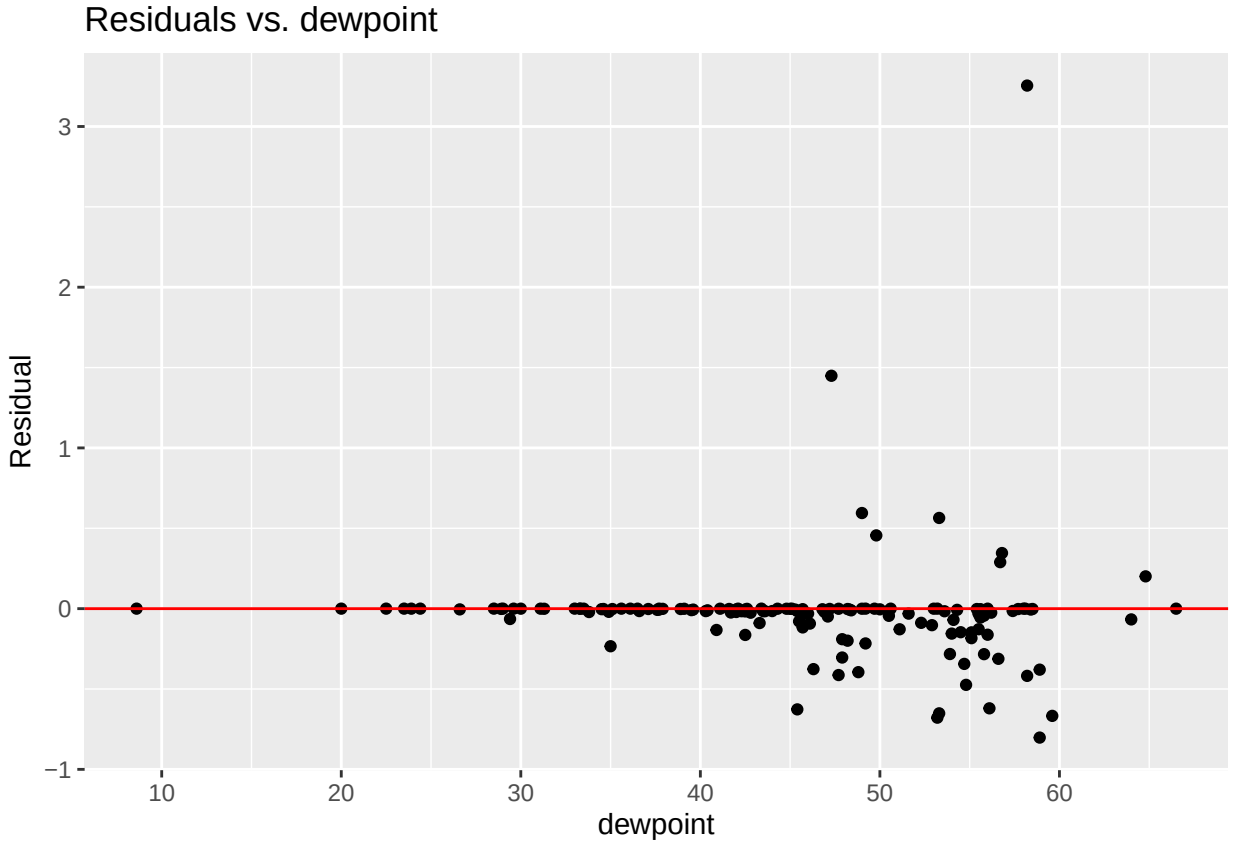
AIC	BIC
40.678	49.566

$$log - odds(\hat{Target}) = -15.229 - 0.0118340 solrad + 0.2580629 avgsoiltemp$$

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

From a theoretical perspective, it is likely that our response variable, the log likelihood of a fire, and one of our predictors, average temperature of the dew on the grass (dewpoint), have a curvilinear relationship. A low value for dewpoint could be recorded by a particular station as a result of firefighting efforts while a high value for dewpoint could be the result of a fire.

To see if this hypothesis is supported by our data, we graphed the relationship between the log odds of our response variable, Target, and dewpoint.



****

To determine if this might be the case, we test a quadratic transformation of dewpoint as a predictor to our main effects model and fit a new model.

term	estimate	std.error	statistic	p.value
(Intercept)	56.625	46.019	1.230	0.219
ETo	186.740	84.868	2.200	0.028
solrad	-0.083	0.032	-2.602	0.009
avgvappress	-7.723	6.523	-1.184	0.236
avgsoiltemp	0.374	0.159	2.350	0.019
windrun	4.937	2.384	2.071	0.038
precip	2.378	7.374	0.323	0.747
I(dewpointˆ2)	0.047	0.035	1.327	0.185
avgairtemp	-1.142	0.643	-1.775	0.076
avgwindspeed	-119.824	57.703	-2.077	0.038
avgrelhum	-0.379	0.286	-1.324	0.185

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-6.727	8.645	-0.778	0.437	-25.007	10.901
ETo	157.331	75.853	2.074	0.038	39.359	352.290
solrad	-0.070	0.029	-2.400	0.016	-0.147	-0.026
avgsoiltemp	0.434	0.180	2.414	0.016	0.169	0.924
windrun	5.255	2.506	2.097	0.036	1.418	11.731
I(dewpointˆ2)	0.002	0.002	1.312	0.189	0.000	0.007
avgairtemp	-0.374	0.233	-1.606	0.108	-0.932	0.009
avgwindspeed	-127.386	60.670	-2.100	0.036	-284.273	-34.577

AIC	BIC
37.967	61.669

Adding the quadratic transformed variable, AIC has a three point improvement over new_fire_model, but BIC is larger. Because we have no preference for a parsimonious model (what is indicated by a lower value of BIC), we keep the quadratic term for dewpoint. Thus, our current model is:

$$\hat{log - odds(Target)} = -6.727 + 157.331ETo - 0.070solrad$$
$$+0.434avgsoiltemp + 5.255windrun + 0.002(dewpoint^2) - 0.374avgairtemp$$
$$-127.386avgwindspeed$$

Next, we explore potentially significant interaction terms. We ultimately choose to test the only interaction term which seemed meaningful: ETo*avgwindspeed. We infer that large amounts of water transferred to the land by means of plants (ETo) and high wind speed together would significantly decrease the log odds of a forest fire happening because a lot of fast-moving wind may spread the water more and thus make it harder for a fire to develop in the area. On the other end of the spectrum, the log odds of a forest fire happening probably increases in regions with less plant based transpiration and low average wind speed because there is much less spreading of moisture in the air.

To determine if this interaction term is statistically significant, we add it to the model with the quadratic dewpoint term (shown above) and conduct a drop-in-deviance test between the model with and without the interaction term.

term	estimate	std.error	statistic	p.value
(Intercept)	-7.071	8.901	-0.794	0.427
ETo	141.203	90.262	1.564	0.118
avgwindspeed	-123.252	59.778	-2.062	0.039
solrad	-0.066	0.031	-2.119	0.034
avgsoiltemp	0.422	0.177	2.387	0.017
windrun	5.076	2.474	2.052	0.040
I(dewpointˆ2)	0.002	0.002	1.223	0.221
avgairtemp	-0.341	0.257	-1.325	0.185
ETo:avgwindspeed	1.244	3.990	0.312	0.755

Resid..Df	Resid..Dev	df	Deviance	p.value
135	21.967	NA	NA	NA
134	21.874	1	0.092	0.761

The p-value of the drop-in-deviance test is 0.761, much greater than our alpha level of 0.05, which suggests that the data do not provide sufficient evidence to suggest that the interaction term is statistically significant. Thus, we do not include the interaction term in our final model.

After all these analyses/tests, our final model is:

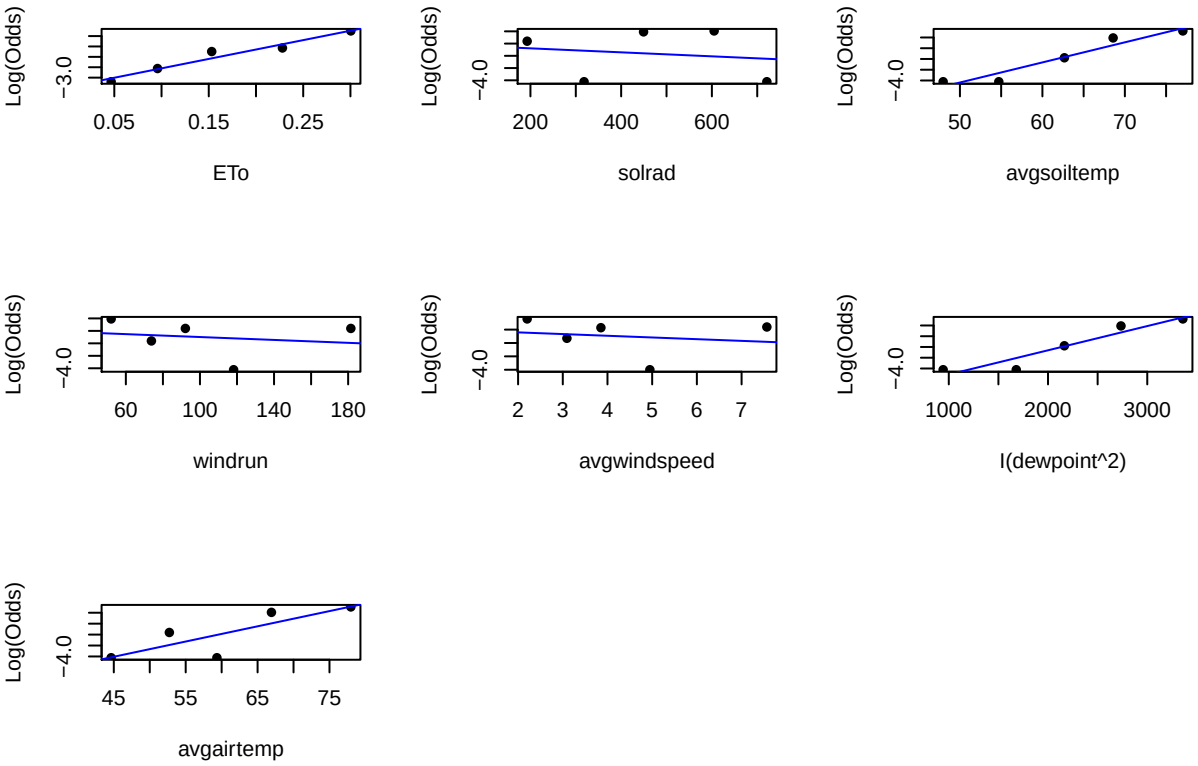
term	estimate	std.error	statistic	p.value
(Intercept)	-6.727	8.645	-0.778	0.437
ETo	157.331	75.853	2.074	0.038
solrad	-0.070	0.029	-2.400	0.016
avgsoiltemp	0.434	0.180	2.414	0.016
windrun	5.255	2.506	2.097	0.036
I(dewpointˆ2)	0.002	0.002	1.312	0.189
avgairtemp	-0.374	0.233	-1.606	0.108
avgwindspeed	-127.386	60.670	-2.100	0.036

$$\hat{log - odds(Target)} = -6.727 + 157.331ETo - 0.070solrad$$
$$+0.434avgsoiltemp + 5.255windrun + 0.002(dewpoint^2) - 0.374avgairtemp$$
$$-127.386avgwindspeed$$

Conclusion

With a final model identified, logistic model conditions (linearity, randomness and independence) can be assessed.

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

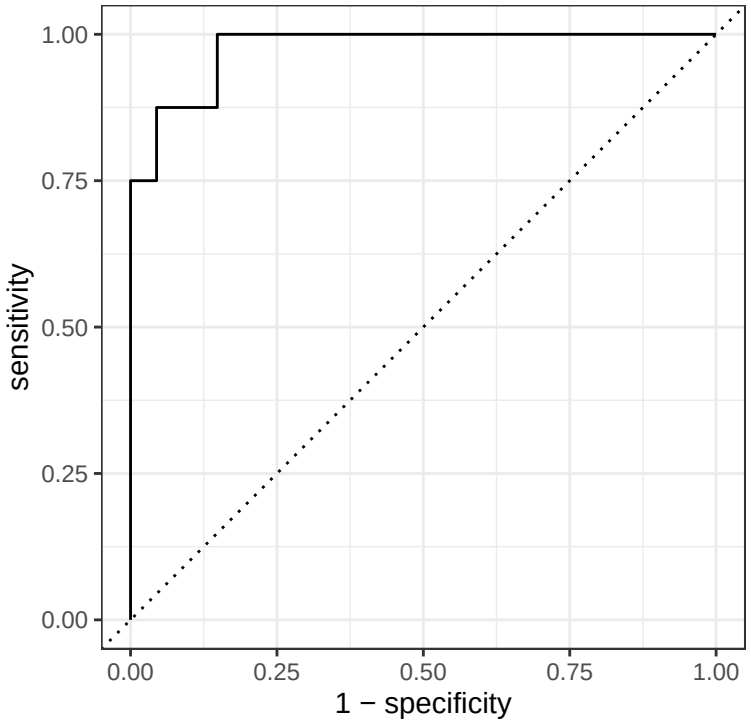


Potential violations of linearity are apparent in the plots for solar radiation (solrad), (windrun), and average wind speed (avgwindspeed). However, the violations are not egregious enough to suggest that there is no linear relationship between the empirical logit and the predictor variables.

We check randomness based on the context of the data and how the observations were collected. The dataset itself does not satisfy this condition, as nearby stations and observations on close dates are subject to a lack of randomness. That said, the way in which we filtered data (grouping by station id and randomly selecting an observation) serves to satisfy this condition as well as the independence condition.

With all logistic regression conditions satisfied by our final model, we can go ahead with assessing its predictive power.

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



From this ROC curve, we select a prediction threshold of 0.193, identified by minimizing 1-specificity while maximizing sensitivity. Because there is greater risk in failing to predict a fire (type 2 error), we are less interested in the false positive rate as opposed to high sensitivity.

.threshold	specificity	sensitivity	false_rate
0.193	0.956	0.875	0.044

Target	pred_resp	n
1	fire	6
1	no fire	2
0	fire	6
0	no fire	129

The confusion matrix indicates that the model correctly predicts the presence or non presence of a fire in 135/143 cases or 94.41% of the time at a threshold level of 0.193.

To truly test our model, we randomly select a new set of observations from our original dataset and assess the model's predictive power on the new data points.

Target	pred_resp	n
1	fire	5
1	no fire	3
0	fire	8
0	no fire	126

On the new, randomly selected test dataset, the confusion matrix indicates that the model correctly predicts

the presence or non presence of a fire in 131/142 cases or 92.25% at a threshold level of 0.193.

Discussion

Based on our final model, ETo, solrad, avgsoiltemp, windrun, I(dewpointˆ2), avgairtemp, avgwindspeed are the most significant environmental predictors of forest fires in California.

While this might mean that monitoring these seven will provide a reduction strategy to fires, it raises some difficult policy questions. Each of these respective variables are naturally occurring meteorological features. While we can identify these conditions, not much can be done to alter these conditions in the short run. Therefore, our findings can only help identify the conditions that make a fire likely, but give little insight into what we can do to stop a forest fire. Being able to identify these conditions can help make the proper authorities aware of the possibility that a fire is likely. However, even once they are aware of the conditions for forest fires as suggested by our model, it is probably too late to be able to do anything to effectively prevent the fire at that point, and firefighting measures can only be reactive from there.

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 and as possible.

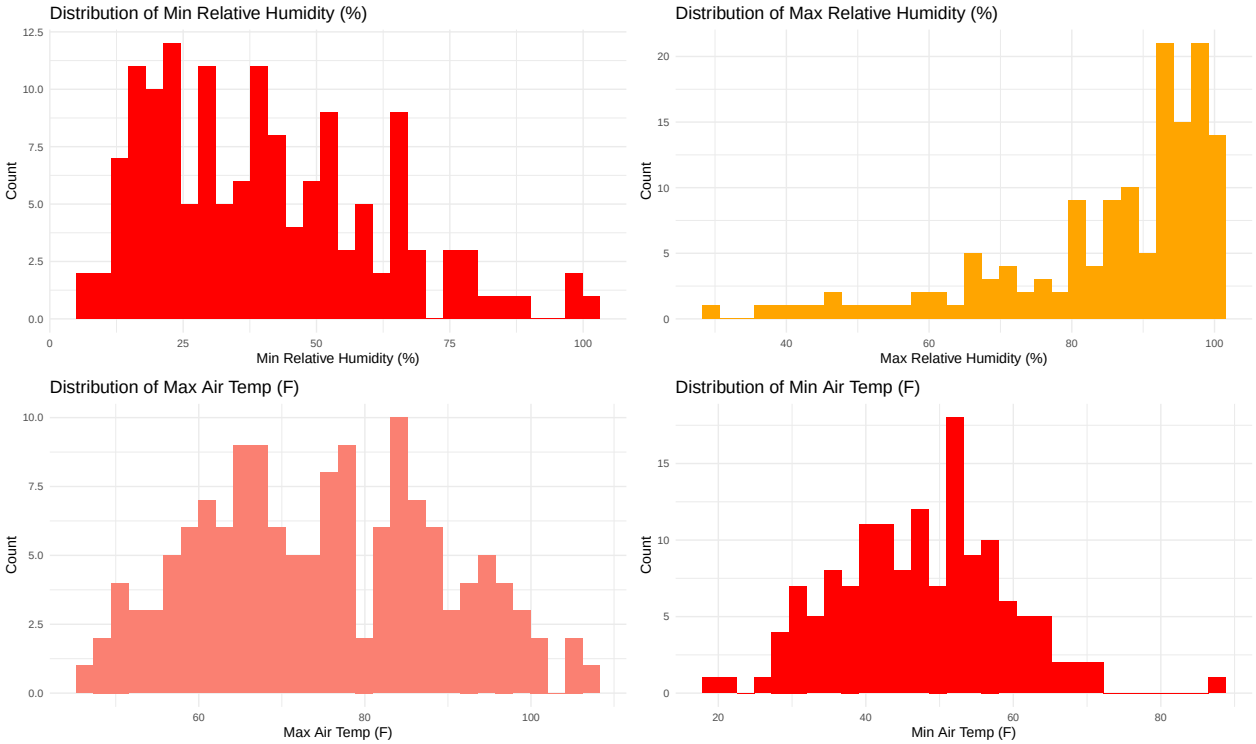
Looking at the confusion matrix, it is clear that the model is a stronger predictor of no fires than fires. Where there is no fire, the model is correct in 129/135 cases for an overall no fire prediction accuracy of 95.56%. On the other hand, when there is a fire, the model predicts its presence in 6/8 cases for an overall fire prediction accuracy of 75%. The stakes are higher when there is the potential for a fire, so a prediction accuracy of 75% is concerning, to say the least. With more time and more data, the reason for this would be further investigated, and the threshold for our final model potentially adjusted.

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 AIC, with no preference for a parsimonious model. With more time, we could construct 1) a model with AIC as our 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

FIGURE A



To reduce multicollinearity, we decided to eliminate these variables from our analysis. For example, the dataset included the average, minimum and maximum value for a number of variables, including relative humidity and air temperature. We determined that the daily averages for these variables were likely the most relevant value for each condition recorded by each station with regards to predicting fires.