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 (Denchak, 2017). 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 (Louie, 2020).

In our project, we aim to determine the strongest environmental predictors of forest fires by looking at data from various weather stations in California.

It is important to understand what the predictors for fires are for multiple reasons. First, in the short run, we will be able to identify the conditions that increase the likelihood for forest fires. With this in mind, we can try to correct these conditions through policy and practice to see if it is at all feasible to mitigate the threat of forest fires in our future. Second, if we know what predictors play a role in predicting fires, fire fighting personnel may better understand which environmental or meteorological conditions present a greater fire risk and hopefully prepare accordingly. Thus, while climate change will continue to affect our way of life, insights into how to manage its immediate consequences are evidently necessary.

Thus, the research question we wish 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." The 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 (Zaloumis, 2020).

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 investigated 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. ETo is measured in inches.

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. Precipitation is measured in inches.

solrad - The solrad variable measures the average amount of solar radiation found in the each station's region in the prior days. Solar radiation is measured in Langley/day, which can be understood as about half a Watt per square meter.

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. Average vapor pressure is measured in mBars.

avgsoiltemp - The avgsoiltemp variable measures the average soil temperature found in the each station's region. Average soil temperature is measured in degrees Fahrenheit.

windrun - The windrun variable measures the sum of wind speed over a month long period. Windrun is measured in miles.

avgwindspeed - The avgwindspeed variable measures the average wind speed found in the each station's region. Average wind speed is measured in miles per hour (mph).

dewpoint - The dewpoint variable measures the average temperature of the dew on the grass in each station over a month long period. Dewpoint is measured in degrees Fahrenheit.

avgrelhum - The avgrelhum variable measures the average relative humidity found in the each station's region. Average relative humidity is represented as a percentage (%).

avgairtemp - The avgairtemp variable measures the monthly average of the air temperature found in the each station's region. Average air temperature is measured in degrees Fahrenheit.

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 grouped observations by station id number. Next, one day was chosen at random from each of the stations. This gave us 153 observations instead of the original dataset of 128,126 observations.

By reducing the data set to a small training set, we were able to make a model that is a more realistic approximation of the conditions that might cause a fire, sans the heavy independence violation 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 separate sample of data to assess its predictive power. The sampling procedule also allows us to ensure that randomness and independence is satisfied for our analysis. Whether a fire is reported or not is no longer conditional on surrounding stations and surrounding environmental variables, as each station observation is from different days and from different times of the year. Thus, while the dataset in its entirety does not satisfy independence or randomness, our random sample meets this criteria.

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

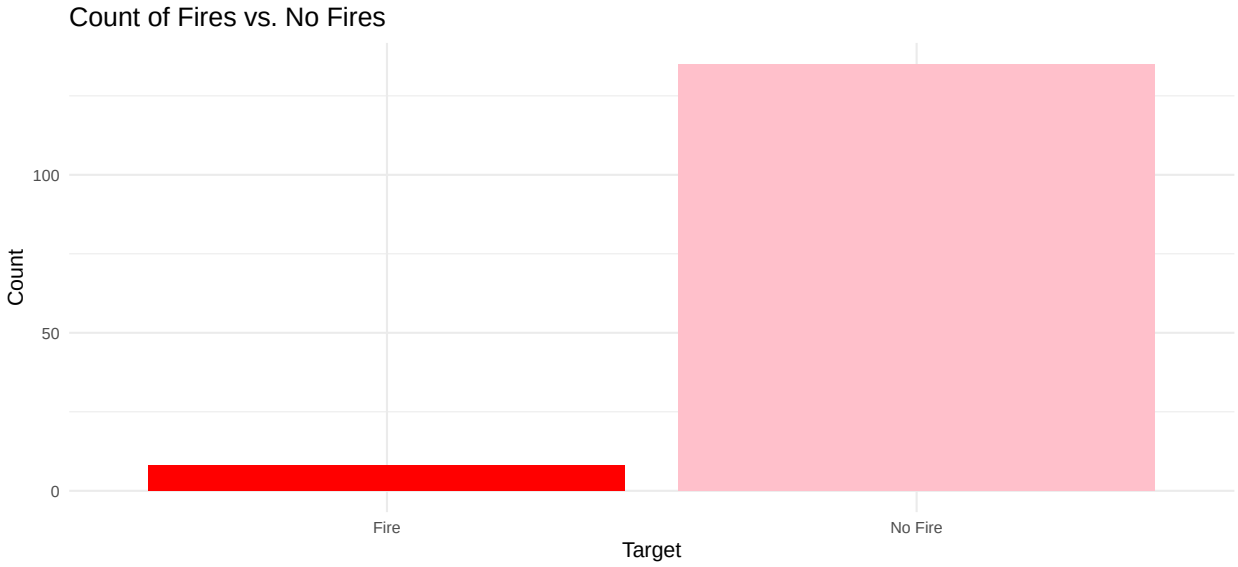
We looked at the missing observations to see if they differed at all systematically from the rest of the data.

Stn Name	Target	Stn Id	CIMIS Region	ETo
Kesterson	1	92	San Joaquin Valley	NA
Green Valley Road	0	111	Monterey Bay	NA
Blythe NE	0	135	Imperial/Coachella Valley	NA
Winters	0	139	Sacramento Valley	NA
Bryte (experimental)	0	155	Sacramento Valley	NA

Stn Name	Target	Stn Id	CIMIS Region	ETo
Owens Lake North	0	183	Bishop	NA
Owens Lake South	0	189	Bishop	NA
Lompoc	0	231	Central Coast Valleys	NA
Santa Maria II	0	232	Central Coast Valleys	NA
Joshua Tree	1	233	San Bernardino	NA

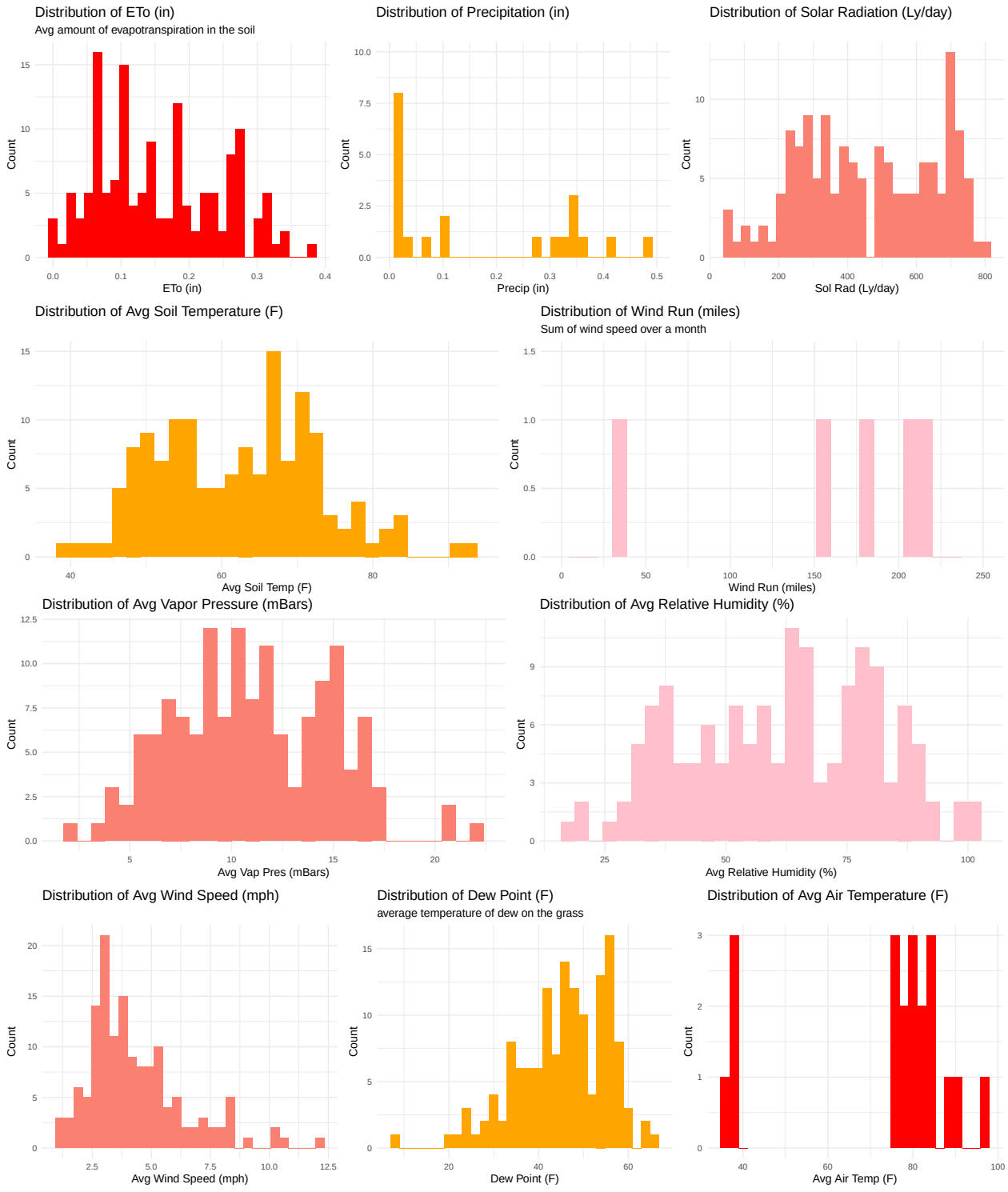
It appears as though the stations with missing observations are completely random, indicated by the varied names and station id numbers; it is therefore unlikely that our resulting analysis is biased by the decision to remove the data. We then proceeded with analysis.

Our first step of our exploratory data analysis was 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 include significantly more "No Fire" observations as opposed to "Fire" observations. This is to be expected, as the forests are not on fire more often than they are.

The next step was to look at the shape of the distributions of each potential predictor variable. This gave us a better understanding of our data and hinted 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.

ETo appears to follow a trimodal distribution with a slight right skew. Precipitation follows a unimodal right skewed distribution. It also appears to contain some outlier observations on the right side of the distribution. Solar radiation appears to follow a bimodal distribution that has a slight left skew. Average soil temperature is a bimodal distribution. Wind run does not have a clear pattern, however the data points are more left

skewed than central. There is a data point that is much smaller than the others, however it is unclear if this is a true outlier or a consequence of a small sample size. Average vapor pressure follows a bimodal distribution. Average relative humidity is a multimodal distribution. Average wind speed follows a right skewed unimodal distribution. Dew point follows a left skewed unimodal distribution. Average air temperature does not have a clear pattern on the data, however the data points are more left skewed than center. Furthermore, there appears to be potential outlier points on the left side of the distribution.

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. However, these plots are still useful in their current form to show the distribution of each of the variables to help inform our analysis.

Methodology

Our aim was 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 used a logistic regression for our analysis.

After identifying our regression method, we began to construct prediction models. We started first with a main effects model, **fire_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 terms in this model are the term for average soil temperature and solar radiation (the only terms with an associated p.value of less than 0.05), we suspected multicollinearity amongst variables. We used 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 removed 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 were removed as well due to multicollinarity with average vapor pressure

and average air temperature, respectively. We then constructed a new model,**full_fire_model**, without the aforementioned variables.

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 safely concluded that we removed all highly correlated variables from analysis.

Next, using backwards selection from **full_fire_model**, we constructed **reduced_fire_model**. This was done to remove unnecessary variables and 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
avgsoiltemp	0.2580629	0.0646067	3.994368	0.0000649	0.1472155	0.4076018

AIC	BIC
40.678	49.566

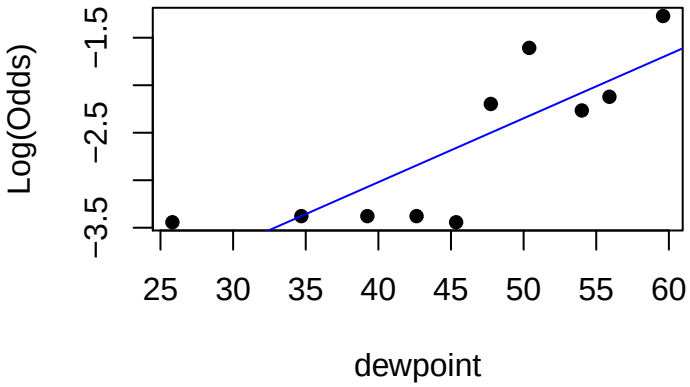
This gave us **reduced_fire_model**:

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

We then tried out various 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.



While not exactly quadratic, the relationship appears distinctly non-linear.

With this in mind, we tested a quadratic transformation of dewpoint as a predictor to our main effects model and fit a new model, **main_fire_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(dewpoint^2)	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

We then performed backwards selection on **main_fire_model** to construct **final_fire_model**.

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(dewpoint^2)	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 **reduced_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 **final_fire_model**:

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

Next, we explored potentially significant interaction terms. We ultimately chose to test the only one interaction term: ETo*avgwindspeed. We inferred that large amounts of water transferred to the land by plants (ETo) and high wind speed together would significantly decrease the log odds of a forest fire. We hypothesized that fast-moving wind may increase water spread and thus make it harder for a fire to develop in a particular area.

To determine if this interaction term is statistically significant, we added it to the model with the quadratic dewpoint term (shown above) and conducted 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 did not include the interaction term in our final model.

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

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

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

$$-127.386 avgwindspeed$$

We can now interpret the coefficients of our model with our original research question in mind.

Our model indicates that the average amount of evapotranspiration present in the soil (ETo), average soil temperature, the sum of wind speed over a month (windrun), and the average temperature of the dew on the grass (dewpoint) should all be taken into consideration. A one unit increase in any of these predictors, to varied degrees, increases the odds of a fire existing at a particular weather station. Conversely, our model indicates that a one unit increase in average air temperature (avgairtemp), solar radiation (solrad), or average wind speed (avgwindspeed) decreases the odds of a fire existing at a particular weather station. Our aim was to identify the most significant predictors of forest fires in California, and all of the aforementioned variables would fall under such an umbrella.

More specifically, our model shows that:

Holding all other variables constant, for every one inch increase in the average amount of evapotranspiration present in the soil (ETo), the odds that a fire occurs multiply by a factor of 2.128 (exp(157.331)).

Holding all other variables constant, for every one Langley/day increase in solar radiation (can be understood as about half a Watt per square meter), the odds that a fire occurs multiply by a factor of 0.932 (exp(-0.070)).

Holding all other variables constant, for every one unit increase in the average soil temperature, the odds that a fire occurs multiply by a factor of 1.543 (exp(0.434)).

Holding all other variables constant, for every one mile increase in the sum of wind speed over a month (windrun), the odds a fire occurs multiply by a factor of 191.521 (exp(5.255)).

Holding all other variables constant, for every one squared degree (Fahrenheit) increase in the average temperature of the dew on the grass (dewpoint), the odds a fire occurs are multiplied by a factor of 1.002 (exp(0.002)).

Holding all other variables constant, for every one degree Fahrenheit increase in average air temperature, the odds a fire occurs are multiplied by a factor of 0.688 (exp(-0.374)).

Holding all other variables constant, for every one mile per hour increase in average wind speed, the odds a fire occurs are multiplied by a factor of 4.752949e-56 (exp(-127.386)).

Conclusion

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

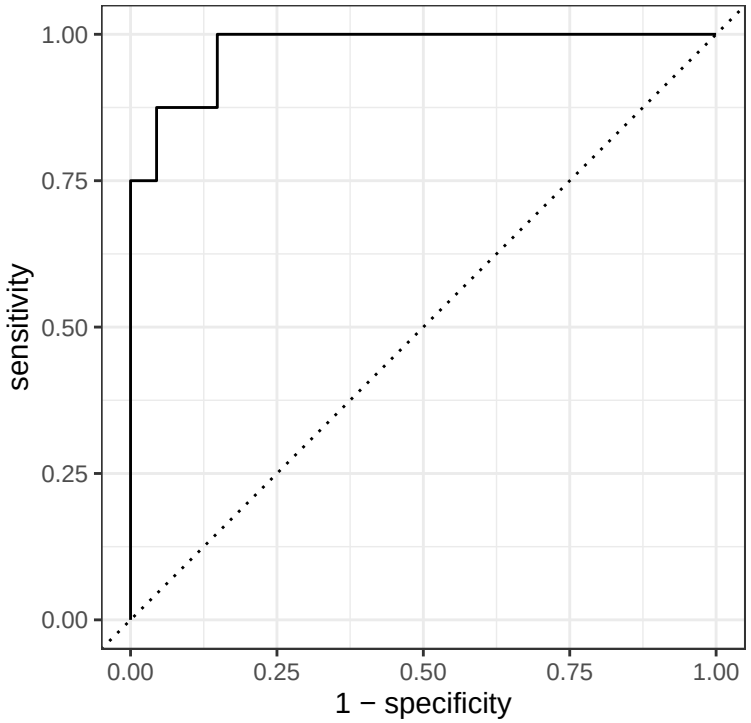
To check linearity, we calculated an empirical logistic regression plot for each of the predictor variables. The empirical logit plots can be found in Appendix B.

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 checked 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, went ahead with assessing its predictive power.

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 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 were 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 selected a new set of observations from our original dataset and assessed the model's predictive power on the new data points.

x
0.915

The test dataset has 142 total observations, 8 of which indicate fires and 134 of which indicate non fires.

On the new randomly selected test dataset, the model correctly predicts the presence or non presence of a fire 91.55% of the time at a threshold level of 0.193. The sampling and prediction procedure for the test data set are included in Appendix C.

Discussion

Based on our final model, the average amount of evapotranspiration present in the soil (ETo), average soil temperature, the sum of wind speed over a month (windrun), average air temperature, solar radiation, average wind speed, and the average temperature of the dew on the grass (dewpoint) 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, in practice, it raises some 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. That said, being able to identify these conditions can help to automate the process of notifying the proper authorities of the odds a fire will occur at a particular station. However, even once they are aware of the conditions for forest fires as suggested by our model, it is most likely too late to be able to do anything to effectively prevent the fire, and any firefighting measures can only be reactive.

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. It is likely that is a result of our dataset; the data contains very few observations with Target == 1, thus making it difficult to create a model that is able to predict this condition with accuracy. With more time and more data (or even different data), this would be further investigated.

Additionally, we only considered one potentially meaningful interaction term, ETo*avgwindspeed, throughout our analysis. In an expanded version of this project, we would 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.

References

CIMIS. (2020). CIMIS Overview. California Irrigation Management Information System. https://cimis.water.ca.gov/Default.aspx

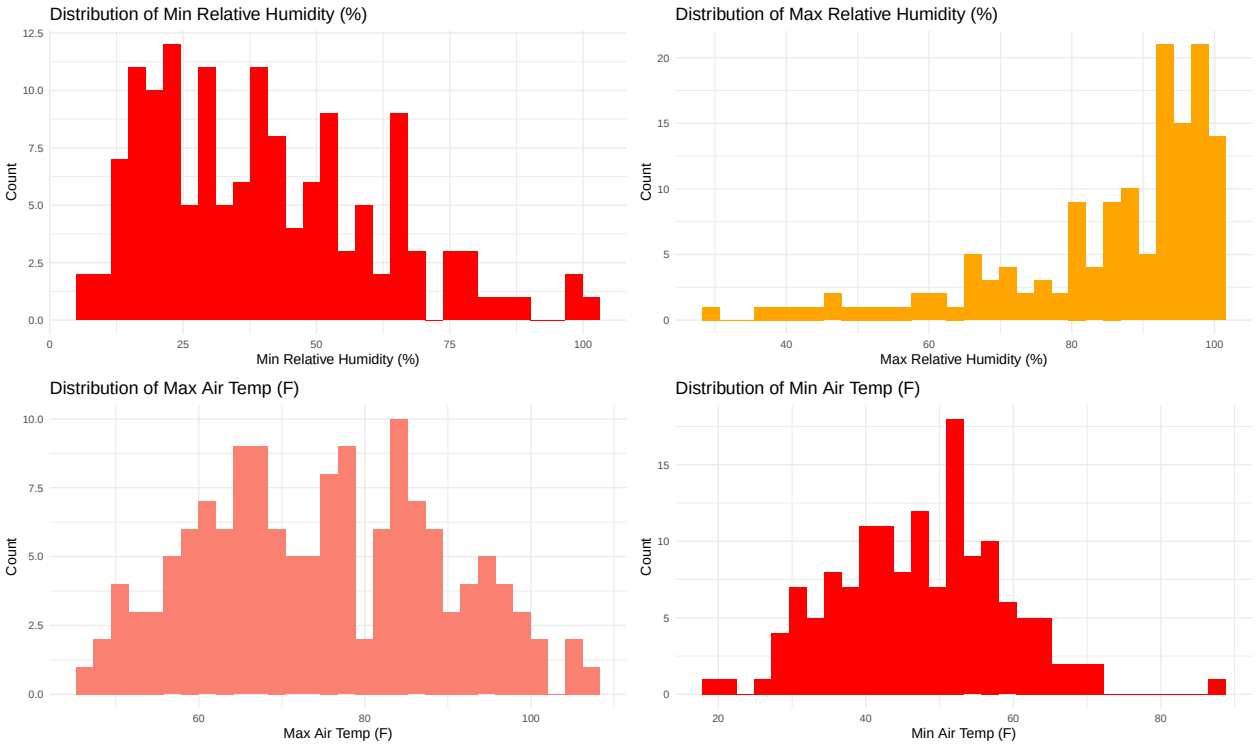
Denchak, M. (2017). Global Climate Change: What You Need to Know. NRDC.Org. https://www.nrdc.org/stories/global-climate-change-what-you-need-know

Louie, D. (2020, October 10). Damage from California's wildfires estimated at $10 billion, experts say; local, state, federal cooperation needed. ABC7 San Francisco. https://abc7news.com/california-wildfires-cost-of-cal-fire-stanford-wildfire-research/6897462/#:%7E:text=Damage%20from%20California\T1\textquoterights%20wildfires%20estimated%20at%20%2410%20billion%2C%20experts%20say,-KGO

Zaloumis, C. & CIMIS (2020, October). California Environmental Conditions Dataset (Version 2) [A collection of CIMIS recorded environmental conditions.]. https://www.kaggle.com/chelseazaloumis/cimis-dataset-with-fire-target

Appendix

FIGURE A



The dataset included the average, minimum and maximum value for a number of variables, including relative humidity and air temperature. To reduce multicollinearity, we decided to eliminate these variables from our analysis. 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.

FIGURE B

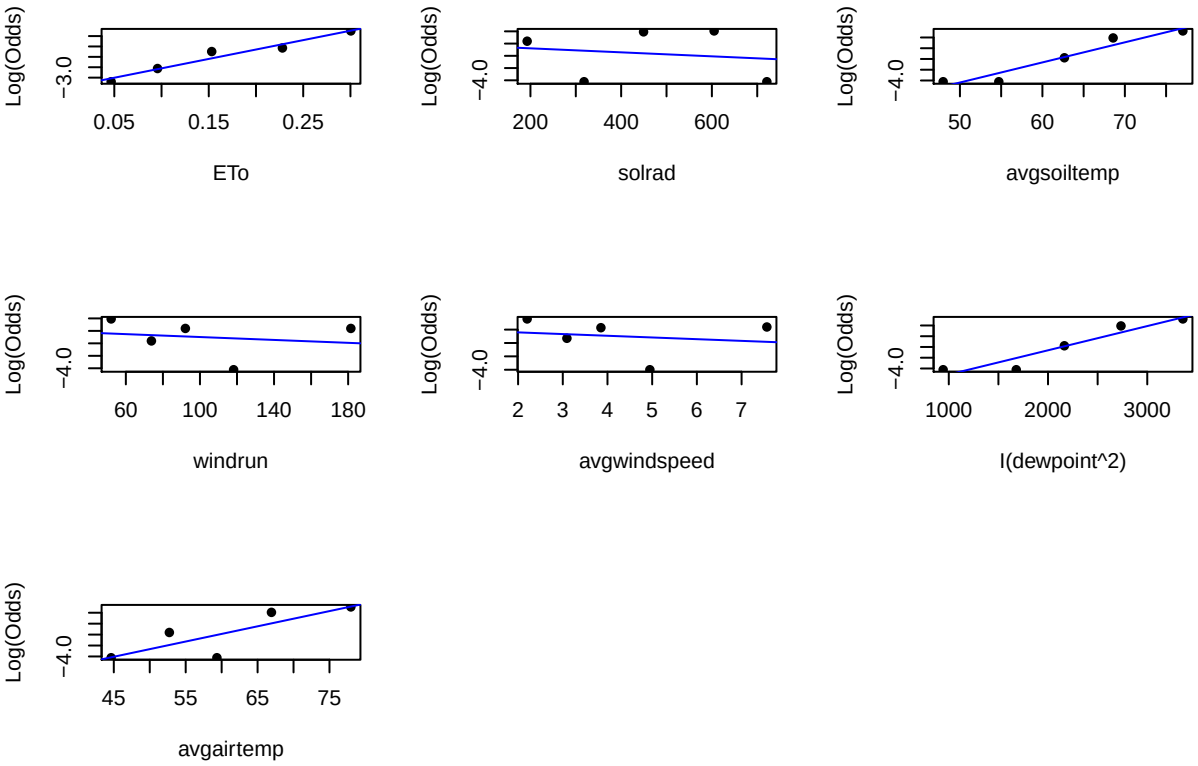


FIGURE C

```
set.seed(40)

newestfire_test <- newfire %>%
  group_by(`Stn Id`) %>%
  sample_n(1, replace = TRUE)
```

```
newestfire_test <- newestfire_test %>%
  drop_na()
```

```
true_type <- newestfire_test %>%
  pull(Target)

true_type
```

```
##   [1] 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0
##  [75] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0
## [112] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
```

```
pred_log_odds <- augment(final_fire_model, newdata = newestfire_test) %>%
  pull(.fitted)
```

```
pred_probs <- exp(pred_log_odds) / (1 + exp(pred_log_odds))
pred_probs <- round(pred_probs,3)
classified <- character(142)

for(i in 1:142){
  classified[i] <- if_else(pred_probs[i] > 0.193, 1, 0)
}
```

```r
mean(classified == true_type)
```

```
## [1] 0.915493
```