Midterm Project

Florin Andrei

3/26/2021

Purpose

The purpose of this analysis is to identify the main factors that may produce high ozone levels in the Upland, CA area near Los Angeles. Many environmental variables could be tracked, but not all are equally important in predicting when ozone levels exceed acceptable limits. Eliminating irrelevant factors may simplify prediction, resulting in a process that's easier and cheaper to implement.

Data

The variables investigated in this analysis are:

- weekday: a factor with levels Monday, Tueday, Wednesday, Thursday, Friday, Saturday, Sunday
- pressure500Height: 500 millibar pressure height [meters]
- windSpeed: wind speed [mph]
- humidity: relative humidity [%]
- tempSandburg: (typo) temperature at Sandberg, CA [degrees F]
- inversionBaseHeight: inversion base height [feet]
- pressureGradientDaggett: pressure gradient from LAX to Daggett, CA [mm Hg]
- inversionBaseTemp: inversion base temperature [degrees F]
- visibility: visibility [miles]

They are all environmental variables collected in or near the area of interest.

Analysis

Since the outcome is binary (high ozone level: yes or no), some form of logistic regression must be used. Additionally, we are interested in techniques that can automatically identify and eliminate predictors that are not relevant.

Logistic regression constrained with Elastic Net, and **boosted decision trees**, are techniques that fit these criteria. Several models were created using these two techniques; their performance was assessed with double cross validation. Models were ranked by performance, outliers and models with various issues (pending further analysis) were eliminated, and only the best one model was kept for the final results. Some of the other top models were also used to make sure the importance of the predictors is consistent across all models.

Findings

The analysis revealed that not all variables are equally important for predicting high ozone levels. Some variables such as **windSpeed** do not matter at all. Other variables such as **weekday** are at best very marginal, and in practice may be ignored.

The most important variable (confirmed by most models) is **tempSandburg**. For ranks #2 and #3, the variables **inversionBaseTemp** and **inversionBaseHeight** are consistently highlighted by most models. This is a visual representation of the relative importance of predictors:

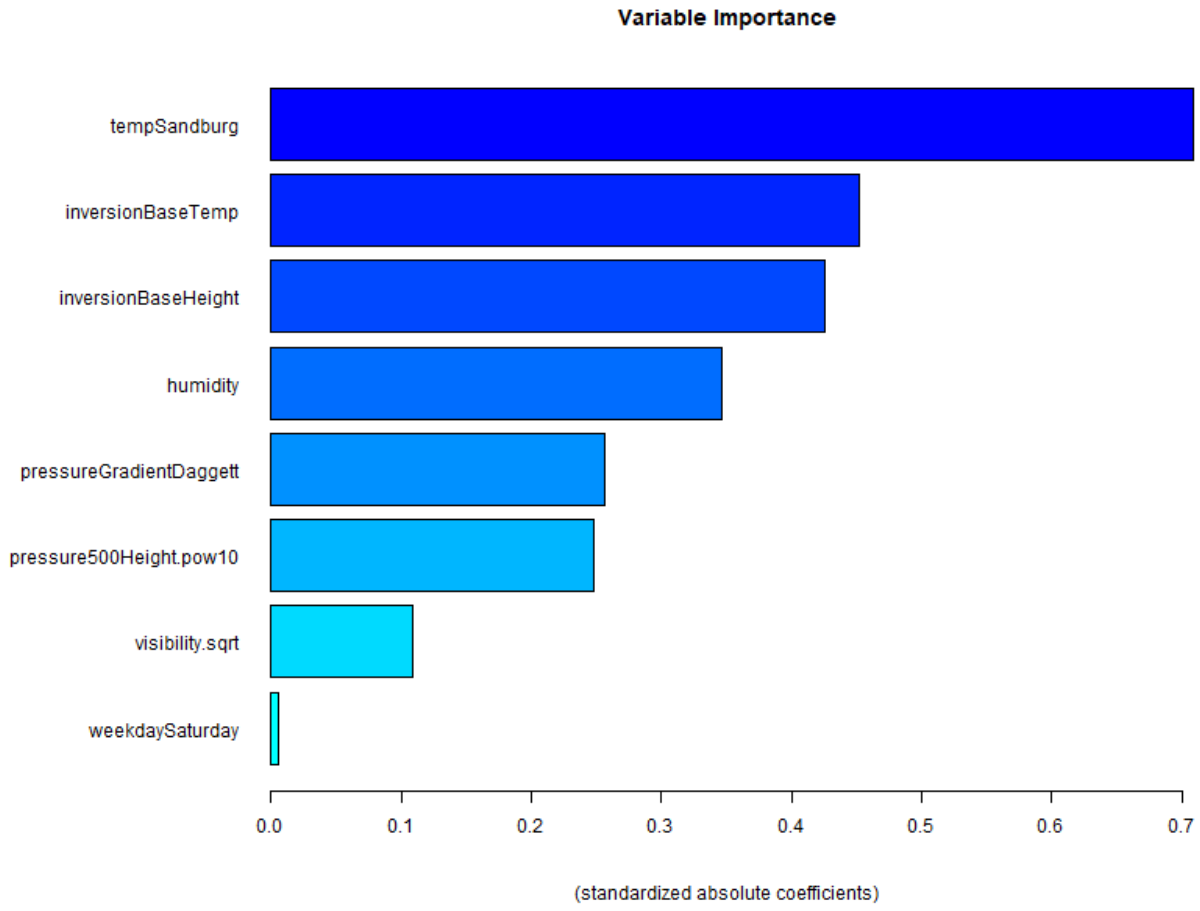

Figure 1: variable importance

This is the full, final fitted model (variables closer to the end are lesser in terms of importance):

```
mod.out =  -6.688 + 0.049 * tempSandburg + 0.0328 * inversionBaseTemp - 0.000236 * inversionBaseHeight
+ 0.0175 * humidity + 0.00719 * pressureGradientDaggett + 0.361 * pressure500Height.pow10
- 0.0296 * visibility.sqrt + 0.0162 * weekdaySaturday
```

```
p = exp(mod.out) / (1 + exp(mod.out))
```

where:

```
visibility.sqrt = visibility ^ 0.5
```

```
pressure500Height.pow10 = pressure500Height ^ 10 / 10 ^ 37
```

`weekdaySaturday = 1` when `weekday = Saturday`

To make predictions: Enter measured values in the formula. Variables near the end of the formula may be ignored if a high precision is not required. Calculate `mod.out`, then calculate `p`. When `p > 0.5`, high ozone conditions are likely.

Appendix

Coefficients for all variables (column `coef`) along with their relative importance (column `abs.std.coef`):

```
##                          X          coef abs.std.coef
##              tempSandburg  0.0490563377  0.709292675
##         inversionBaseTemp  0.0328029826  0.452756484
##       inversionBaseHeight -0.0002359874  0.425694395
##                  humidity  0.0174631061  0.346904604
##  pressureGradientDaggett  0.0071944182  0.256964335
##  pressure500Height.pow10  0.3614007841  0.248091027
##           visibility.sqrt -0.0295606834  0.108479645
##           weekdaySaturday  0.0161890101  0.005765229
##               (Intercept) -6.6880862904  0.000000000
##             weekdayMonday  0.0000000000  0.000000000
##             weekdaySunday  0.0000000000  0.000000000
##           weekdayThursday  0.0000000000  0.000000000
##             weekdayTueday  0.0000000000  0.000000000
##          weekdayWednesday  0.0000000000  0.000000000
##                 windSpeed  0.0000000000  0.000000000
```