Predicting Wildfire Spread Risk Using Institutional Wildfire Scores and Meteorological Data

Module 7

Mark Aaron Sansgaard

30 Aug 2020

Dr. Steve Chung

MIS581-1: Capstone - Business Intelligence and Data Analytics

# Introduction for Critical Thinking 6

In this rough draft you will find a few sections unfinished. I focused my effort on the project this week on ensuring the analytics for my simple neural network was completed.

# Importance of Data Analytics in Wildfire Prevention

“Forest fires are a major environmental issue, creating [economic] and ecological damage while endangering human lives” (Cortez & Morais, 2007). Montesinho National Park is one of the largest and most beloved parks in Portugal. Wildfire scientists have been recording fire observations for over 100 years with ties to daily readings about wind, temperature and other meteorological factors that I plan to use for analysis. The data present is already in third normal form, which allows for relational database modeling, data analysis, and simple importing of the data to tools I used for analytics on this project, SAS, Python, and Tableau.

In 2020, we’ve seen fires destroy the bulk of California and Australia so it is important to consider the impact on forest management. Preventing fires means fires wont have the chance to spread, which is where fires cause the most damage, overall.

# End Users of Wildfire Data Analytics Insights

The US Forest Service is a branch of the US Department of Agriculture and has 30,000 person federal force and represent the core that unites the many fire-related organizations that contribute to the bulk of fire fighting and fire prevention in the US. I really wanted to pursue wildfire information because of the rapid emergence of global fires in 2020 in California, Australia, Canada, Southeast Asia, and more. I am using the aforementioned organizations as the potential stakeholders and organizations which stand to benefit from the analytics of the underlying data.

The data makes it clear that fire is a detrimental issue in the United States, as the allocation of the National Forest Service’s budget into firefighting/prevention has shifted from 20% in 2008 to over 50% in 2018. According to the USDA’s website, there is an allocation of “$4.4 billion to mitigate wildfire risk” and another $2 billion to combat ongoing fires. With this amount of money, we need strategic decisions about which projects will yield the most benefit, so we can determine what direction offers the best opportunity-cost ratio. Centralizing, transforming, and normalizing fire data is valuable across many different areas, but will specifically help teams assess potential high risk areas, mitigate high risk factors, and provide information about data to the teams in an easy to digest and consistent manner.

# Hypothesis 1

**Null Hypothesis 1:** The Fine Fuel Moisture Code (FFMC) assesses risk of wildfire spread at the same level of efficacy as the Initial Spread Index (ISI).

**Alternate Hypothesis 1:** The Fine Fuel Moisture Code (FFMC) assesses risk of wildfire spread more accurately than the Initial Spread Index (ISI).

## Tests and Visualizations

## Linear Regression.

So far I have used a matrix, as seen below:

### Pearson Rho.

I wanted to assess how well the wildfire intensity scores DMC, ISI, FFMC, and DC. I will assess how this test works, and the stats involved.

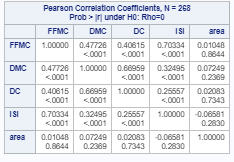


Figure 1. *Pearson Rho Correlation for Four Scores*

# Hypothesis 2

**Null Hypothesis 1:** The combination of all of the FFMC, DMC, DC, ISI scores into a new score does not produce a better scoring mechanism.

**Alternate Hypothesis 1:** The Fine Fuel Moisture Code (FFMC) assesses risk of wildfire spread more accurately than the Initial Spread Index (ISI).

## Tests and Visualizations

### T-Test.

This is the hypothesis I am most excited about. I will get to demonstrate that while each of these underlying scores assesses wildfire risk at a different rate of success, there is even more accurate approach to be found by combining them. This is effectively the same goal as a forest of trees in data model creation. All four of these scores are a “tree” and each attempt to model a behavior. By weighing each of the 4 trees based on which ones are more effective, we can generate a score that is more accurate than any of the other four scores. That’s the goal!

# Data Conversion and Transformation

I had to manipulate the data in a couple different ways to get my desired result. Below you will find some snippets of the code, as well as images of the outputs of correlation tests of the raw inputs, and later you will see the correlation test for our new score.

## Code

PROC SQL;

DROP TABLE WORK.AREA\_DATA\_ONLY\_MINUS\_OUTLIERS;

CREATE TABLE WORK.AREA\_DATA\_ONLY\_MINUS\_OUTLIERS AS

SELECT \* FROM WORK.RAW\_FIRE\_DATA WHERE AREA>0 and area<300 ;

QUIT;

proc sql;

create table work.FIRES\_DATA\_1 as

select \*, (ffmc/(select max(ffmc) from WORK.AREA\_DATA\_ONLY\_MINUS\_OUTLIERS)) as FFMC\_Ratio

, (dmc/(select max(dmc) from WORK.AREA\_DATA\_ONLY\_MINUS\_OUTLIERS)) as DMC\_Ratio

, (DC/(select max(DC) from WORK.AREA\_DATA\_ONLY\_MINUS\_OUTLIERS)) as DC\_Ratio

,(ISI/(select max(ISI) from WORK.AREA\_DATA\_ONLY\_MINUS\_OUTLIERS)) as ISI\_Ratio

from WORK.AREA\_DATA\_ONLY\_MINUS\_OUTLIERS;

quit;

proc sql;

DROP TABLE WORK.FIRES\_DATA\_FINAL;

quit;

proc sql;

CREATE TABLE WORK.FIRES\_DATA\_FINAL AS

select t.\*, WILDFIRE\_SPREAD\_SCORE/max(WILDFIRE\_SPREAD\_SCORE) as Wildfire\_Comparative\_Metric

from (

SELECT F.\*,

((FFMC\*(SELECT WEIGHT FROM WORK.TEMP\_DATA WHERE TYPE='FFMC'))+(DMC\*(SELECT WEIGHT FROM WORK.TEMP\_DATA WHERE TYPE='DMC'))+(DC\*(SELECT WEIGHT FROM WORK.TEMP\_DATA WHERE TYPE='DC'))/3) AS WILDFIRE\_SPREAD\_SCORE

FROM WORK.FIRES\_DATA\_1 F, WORK.TEMP\_DATA A) t

;

QUIT;

PROC TTEST DATA=WORK.AREA\_DATA\_ONLY\_MINUS\_OUTLIERS;

PAIRED FFMC\*AREA;

RUN;

PROC TTEST DATA=WORK.AREA\_DATA\_ONLY\_MINUS\_OUTLIERS;

PAIRED DMC\*AREA;

RUN;

PROC TTEST DATA=WORK.AREA\_DATA\_ONLY\_MINUS\_OUTLIERS;

PAIRED DC\*AREA;

RUN;

PROC TTEST DATA=WORK.AREA\_DATA\_ONLY\_MINUS\_OUTLIERS;

PAIRED ISI\*AREA;

RUN;

proc corr data = WORK.AREA\_DATA\_ONLY\_MINUS\_OUTLIERS PLOTS=MATRIX;

VAR FFMC DMC DC ISI AREA;

run;

data temp\_data;

input type $ weight;

datalines;

FFMC 0.050963391

DMC 0.848362235

DC 0.100674374

;

run;

proc corr data = FIRES\_DATA\_FINAL plots = matrix ;

VAR Wildfire\_Comparative\_Metric area ;

run;

## Result

I first conducted manual analytics on the data, and, as you see at the bottom of the code, I used DATALINES to include the weighted scores for each input after analysis. ISI suffered a poor performance, so did not contribute to our score.

# Conclusion

While there are some local and national data governance concerns surrounding geospatial data in general, there are no opportunities for data privacy or security violations in this project since the data is public-facing and freely available. In my case there is no risk that the data collection is riskier than the value from the analytics, like in the example above. As I continue to review the data and seek new inputs to enhance the model, I will consider the security and privacy of the data as well as the local and national governance related to the data.

So far in my project I am finding messing with the weights of my inputs to be the most difficult component. Regardless, my results are performing better than 75% of the competing models in play in the experiment. I will continue to shore up the final project, but I feel that the bulk of the underlying queries are ready.

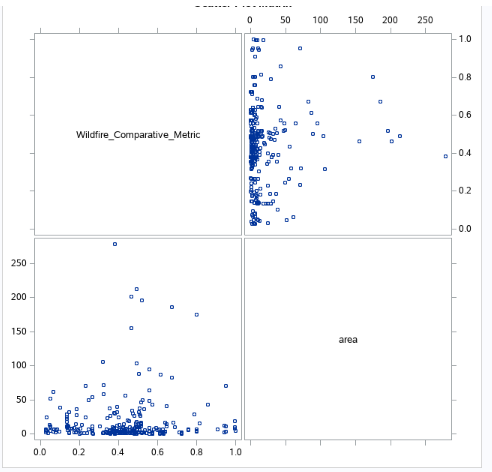


Figure 2. *Wildfire Score/Area Burned Matrix*

As you see in the image above, the potential certainly exists for a linear relationship between this new score and the data. Like the other inputs, the correlation scores are on the lower end, but this method is, so far, proving equally as effective as the most effective alternative.

# References

Congressional Research Service (CRS). (2020, July). Wildfire Statistics. Retrieved August 02, 2020, from https://fas.org/sgp/crs/misc/IF10244.pdf

Cortez, P., Morais, A., & Department of Information Systems, University of Minho, Portugal. (2007). UCI: Repository for Machine Learning, Wildfires. Retrieved July 19, 2020, from https://archive.ics.uci.edu/ml/datasets/forest fires

Graham, M. (2018, October 24). How Big Data Is Helping Predict and Fight Wildfires. Retrieved July 19, 2020, from https://www.delltechnologies.com/en-us/perspectives/how-big-data-is-helping-predict-and-fight-wildfires/

Menardiere, M. (2007). Successful response starts with a map improving geospatial support for disaster management. Washington, D.C.: National Academies Press.

USDA. (n.d.). USDA Forest Data Registry. Retrieved July 19, 2020, from https://www.fs.usda.gov/rds/archive/

Quinton, S. (2019, February 5). Logging, Money Battles Delay Wildfire Prevention Work. Retrieved July 19, 2020, from https://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2019/02/05/logging-money-battles-delay-wildfire-prevention-work

Safi, Y., & Université Hassan II de Casablanca. (2013, January). Retrieved July 19, 2020, from https://www.researchgate.net/figure/The-map-of-Montesinho-natural-park\_fig4\_290160237

Wenzl, K. (2019, March 03). Database Third Normal Form Explained in Simple English. Retrieved July 19, 2020, from https://www.essentialsql.com/get-ready-to-learn-sql-11-database-third-normal-form-explained-in-simple-english/