Data Mining Meteorological Data to Predict Forest Fires

Christopher de Freitas, Pranav Rajan

THE UNIVERSITY OF UTAH

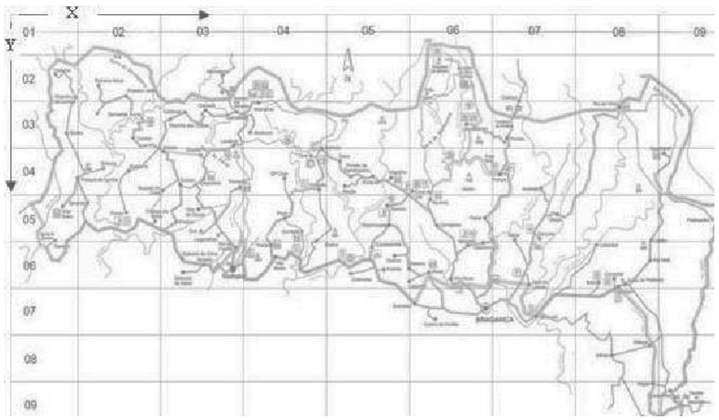
Background

- One major environmental cause of climate change is the rising frequency of forest fires
- Recent forest fires exacerbated by climate change include the 2019-2020 Australia Bushfires, the Amazon Fires of 2019-2020 and crop burning fires in Indonesia
- To understand the nature of such fires, we analyzed data from *Montesino Natural Park* located in the Tras-os-Montes region in Northeastern Portugal
- In our project we attempt to figure out which two factors are most important for determining forest fires using dimensionality reduction and regression
- With this information, firefighters and first responders can take precautionary measures earlier thus preserving resources and saving lives

Data

- The data was collected between January 2000 and December 2003 by researchers Paulo Cortez and Anibal Morais and published in 2007 alongside their paper *A Data Mining Approach to Predict Forest Fires using Meteorological Data*
- Portugal suffered particularly forest fires between 1980-2005 with over 2.7 million hectares destroyed making the region an important region for studying forest fires
- The data consists of 517 rows representing 517 distinct forest fire observations over time
- 13 columns represent 13 distinct features of each individual fire observation: location, time, natural weather features, burn area, and Canadian Fire Weather Index observations of the fire

Grid Based Map of fire data collected

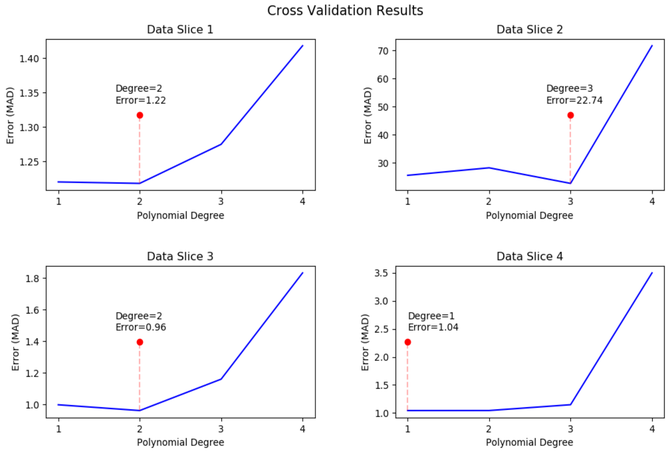


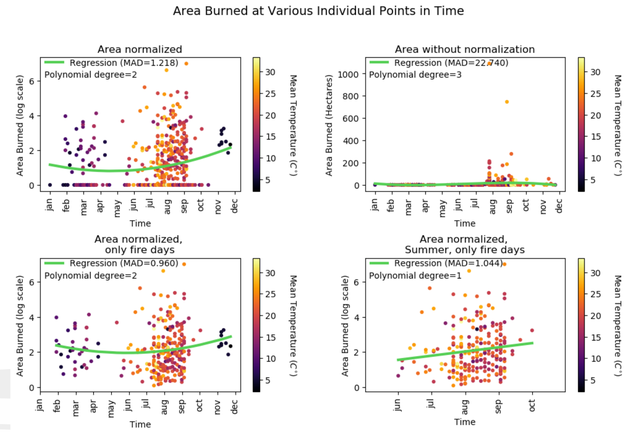
Preprocessing and Modeling

- For our project we used the natural weather features given their interpretability
- Given the right skew of the burn area we normalized the burn area with the function *ln(y + 1)* like Cortez et. al
- We excluded data points where no fires occurred when training our model
- We standardized the time [day, month] into scalar values ranging from 0 to 1

Technique I: Regression

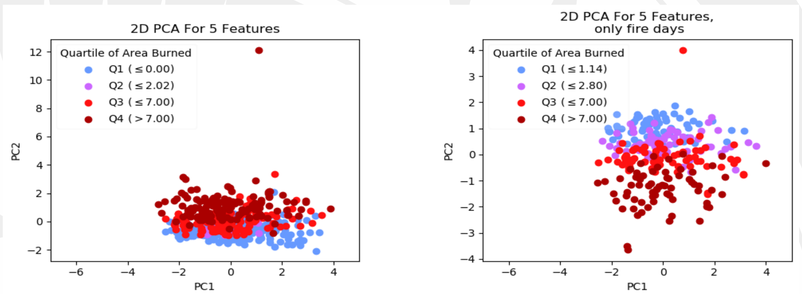
- Based on our work, we found that the relationship between natural weather phenomena and burn area are non-linear so we used polynomial regression
- Multiple models were built predicting the total burn area a fire will have given the month and week day a fire occurs
- The data was sliced in the following way:
  - Slice 1: The full dataset normalized using the preprocessing methods
  - Slice 2: The full dataset without normalized burn area
  - Slice 3: Normalized burn area containing on the days fire occurred
  - Slice 4: Sub-slice of Slice 3 analyzing fires between June-October
- Cross-Validation with 70-30 random split was used to determine the polynomial degree of best fit
- Mean average deviation (MAD) was used as the error metric since it was used by Cortez et. al in their paper and is sensitive to error





Technique II: Dimensionality Reduction

- We did dimensionality reduction to project 5 natural features: temperature, humidity, rain and burn area to 2 and 3 dimensions
- PCA was used on all the data and the days that fires occurred
- PCA revealed a highly negative correlation between principal component 2 and the amount of area burned



CONCLUSIONS

- Our work did not result in a robust model for predicting forest fire behaviors
- Cortez et. al skipped linear classifiers and used neural nets and random forest models for their paper and work
- Dimensionality Reduction was uninterpretable because the natural features had to be normalized for PCA and made it difficult to establish a relationship between principal component 2 and the size of the burn area
- Collinearity and non-linear relationships in the data affected the capacity of our models to predict forest fire behavior

REFERENCES

[Cortez and Morais, 2007] P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9