JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Dataset: Forest Fires Caused By Lightning Final Discussion

Team Beast:
Jack Moody
Maddie Braye
Jonathan Thomas
Joshua Hernandez

Agenda

- Agenda
- BLUF
- Introduction
- Construction of the Data Set
- Classification Methods
- Regression Methods
- Dimensionality Reduction
- Conclusions
- Next Steps
- Citations

BLUF

Our constructed dataset contains variables pertaining to wildfires started by lightning as well as climate in the United States. In this project we used machine learning models to investigate the connection between the severity of these wildfires and the climate of the area.

Introduction

- 44% of wildfires in the Western US are started by lightning
- Lightning ignited wildfires caused 71% of the area burned between 1992 and 2015 (NYT)
- Increased lightning strikes have been linked to global warming and are predicted to increase by up to 50% over this century (Romps et al.)
- As lightning increases and temperatures rise, it it easier for wildfires to start and becoming harder to contain
    - 90 percent of land in the Western states was experiencing moderate to severe drought during 2020 (III)

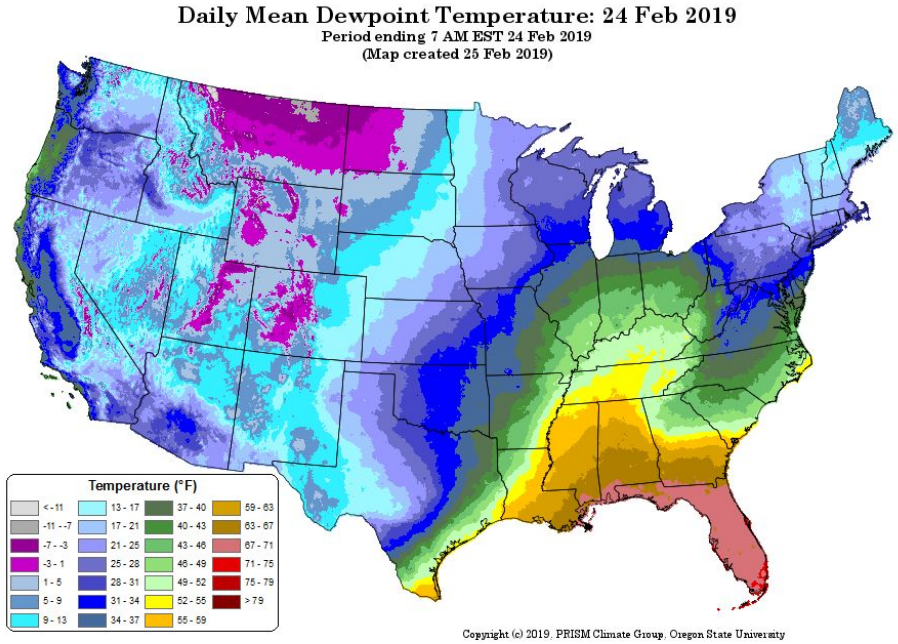
Wikipedia

Construction of Data Set

- The original Lighting Wildfires data set consisted of the variables: Date of Fire (1992 - 2015), Time of Discovery, Time of Containment, Fire Size and Class, Fire Location (LAT/LON), State/Code
- We were curious about the connection between climate and wildfires, so combined this data set with PRISM climate data including precipitation, max/min/mean temperature, mean dew point temperature, and max/min vapor pressure deficit
- To gain a clear understanding of the climate, we included the variables on the date of fire as well as the mean of the variable for the week leading up to the date of the fire
- Not all the lightning wildfire data had climate data available, which did cut down the number of samples available for training/testing to about 115,000 samples
- **Specifically: Alaska was entirely cut from the dataset**
- **2014-07-21 tdmean was climate data that was corrupt**

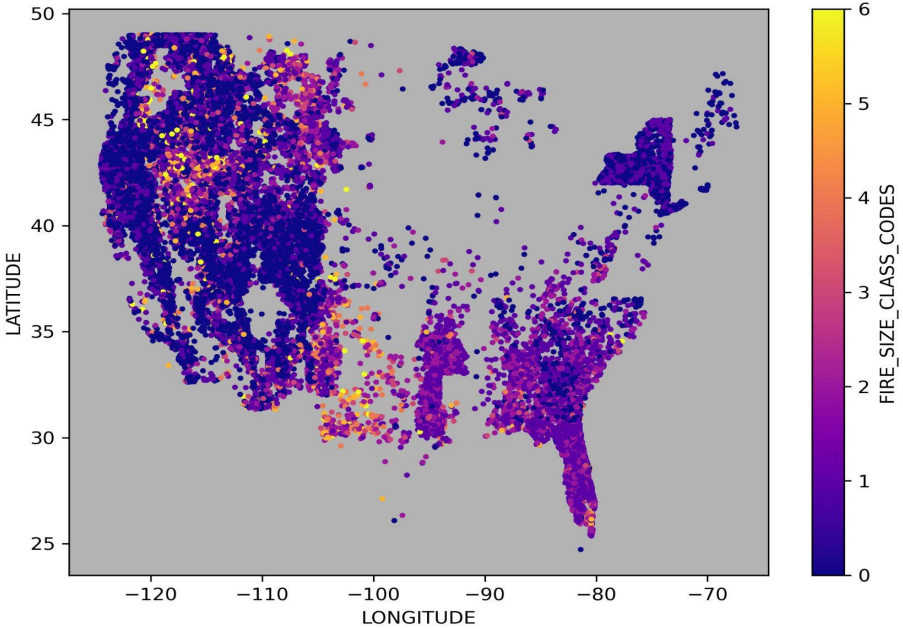
PRISM Climate Data

- The PRISM Climate Group gathers climate observations from a wide range of monitoring networks and develops spatial climate datasets for use in analysis of climate patterns
- The data is available in multiple formats, here we used daily BIL images which contain pixel information of the United States in a 4km^2 grid pattern
- The data was then filtered to days containing data and the weeks leading up to lightning-based wildfires, and aggregated to the Lightning dataset



Classification Methods

- We used several different classification methods with the target variable as Class of Fire, which varied from A for least severe to G for most severe
- The methods used were:
  - Linear Discriminant Analysis, with svd, lsqr, and eigen solvers
  - Reduced Rank Linear Discriminant Analysis, with the same solvers
  - Quadratic Discriminant Analysis
  - Logistic Regression
  - Support Vector Machines
  - K-Nearest Neighbors

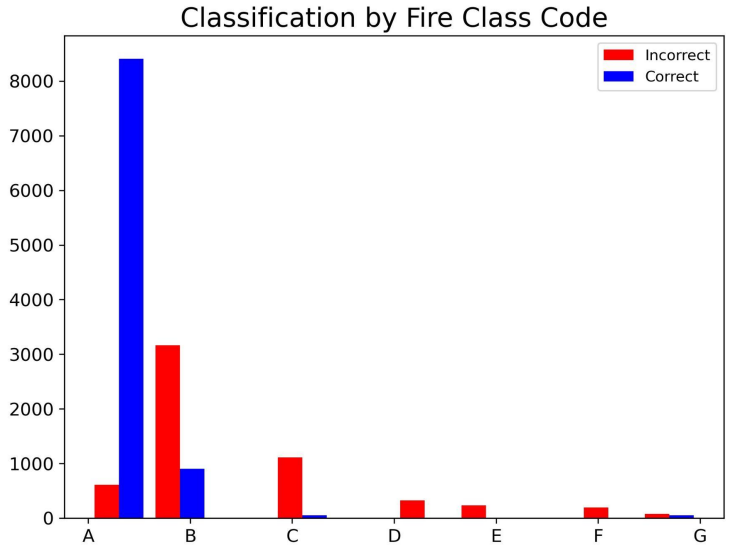


Method Analysis

- Many of the classification methods use separation surfaces and therefore require some level of separability of the data
- To explore parameters that may help the classifier, for each method we looked at the performance of the classifier across each parameter, to see if there were any discernible patterns

Linear Discriminant Analysis (LDA)

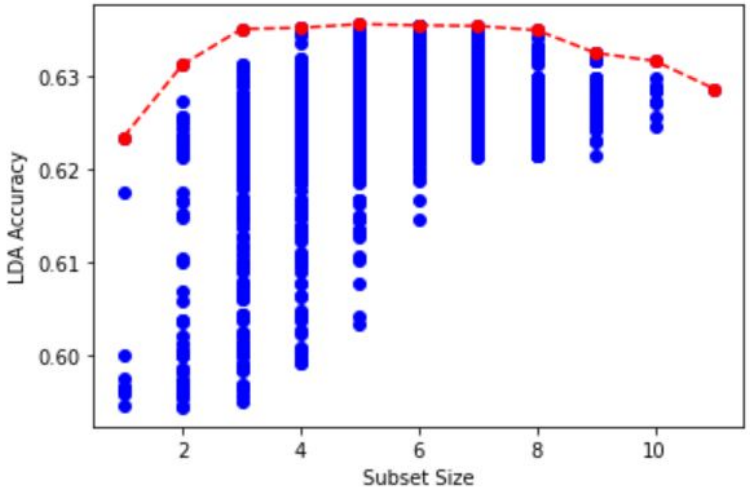
- LDA could shine in this data set if the data is linearly separable
- It is also a popular method because it is very fast and easy to implement
- Weaknesses to LDA however, include lack of data normality, which is present here
- Some data transformations were attempted to achieve normality but were not very successful



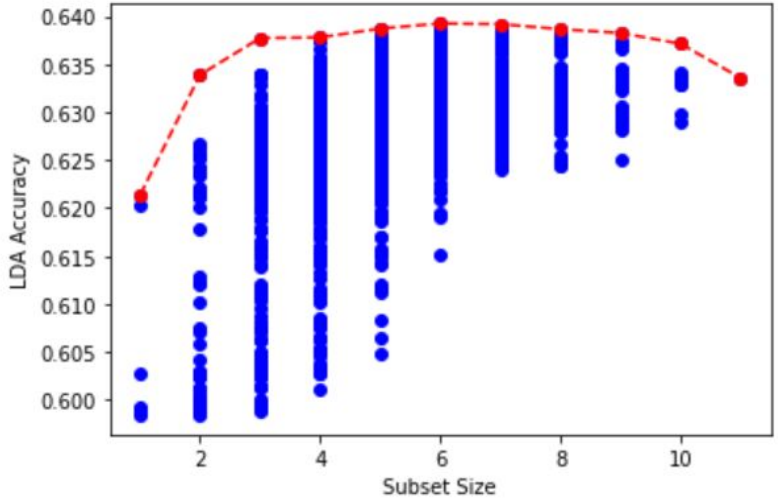
LDA Subset Selection

- As in Best Subset Selection, the best performance is achieved with only 5-7 variables

LDA with supplemental non-week-data removed

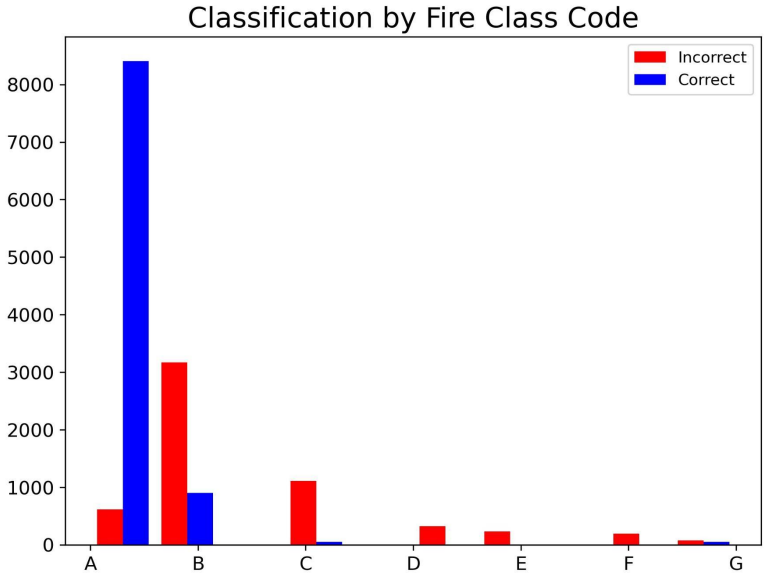


LDA with supplemental week-data removed



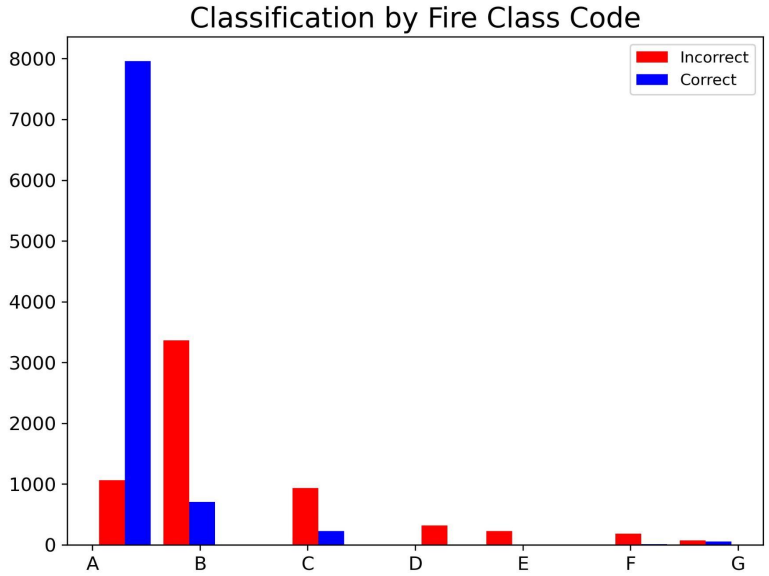
Reduced Rank LDA

- By reducing the dimensionality of the data, Reduced Rank LDA tries to make the data more linearly separable
- This method will have many of the same strengths and weaknesses of LDA
- It will perform better if there is one or two parameters cluttering up the separability of the data



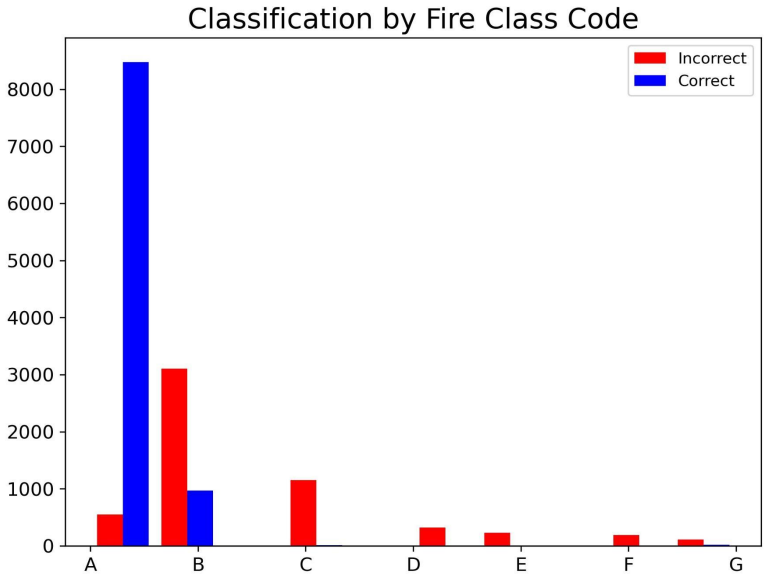
Quadratic Discriminant Analysis (QDA)

- This method is very similar to LDA, however it uses a quadratic decision boundary to attempt to separate the data
- Again, if the data is separable by some parameters, this method will do well, but it has the same weaknesses as LDA in that it generally requires well separated means and normality of data



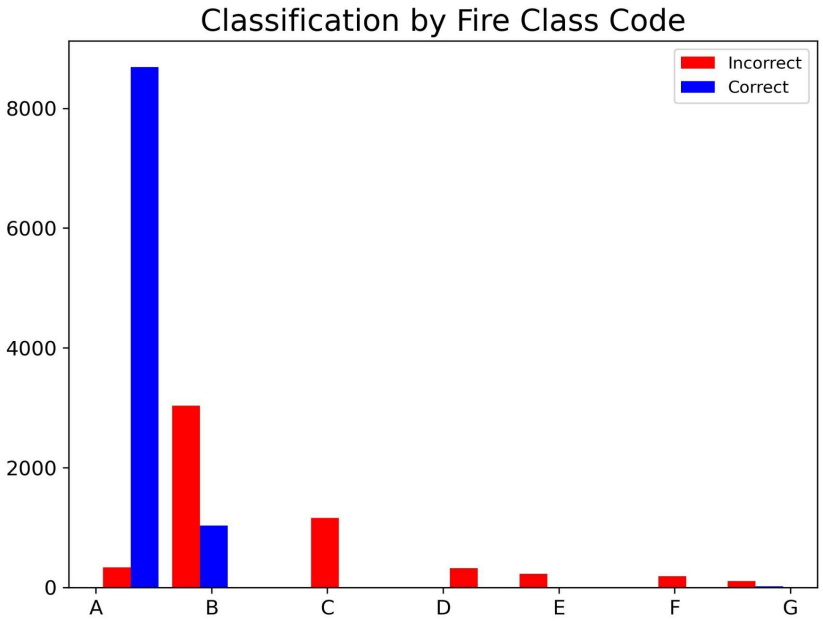
Logistic Regression

- Logistic regression is easy to implement, very efficient, and explainable
- This method has a decision boundary similar to LDA and QDA, but adds some nonlinearity using the sigmoid function
- It does work best when the data is easily separable



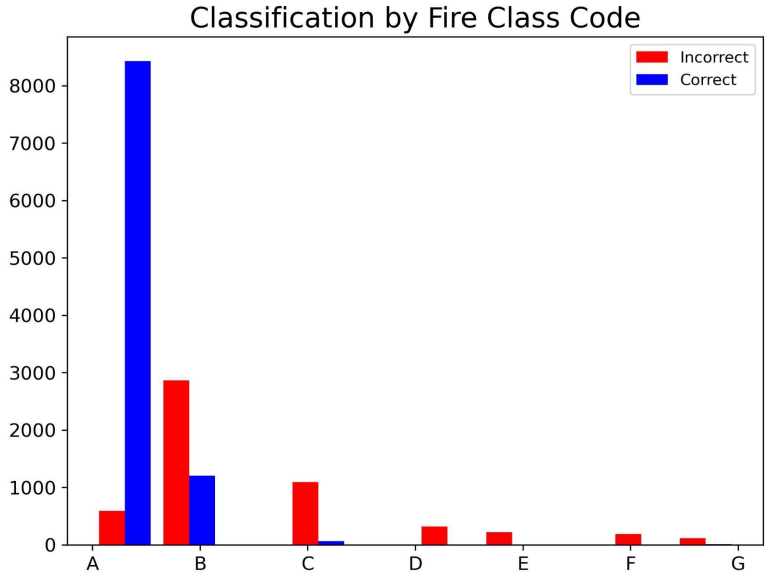
Support Vector Machines (SVM)

- The SVM method is very helpful in cases that are not separable in their input space, which is true here. We were hopeful that the kernel of the SVM would be able to find a separating hyper-plane in higher dimensions
- A known weakness of SVMs is their training time for large amounts of data
- SVMs also have trouble on problems with a large numbers of features



K-Nearest Neighbors (KNN)

- KNN is a very useful method for a number of classes that may be clustered in certain parameters
- Imbalanced data can heavily skew KNN method, and is very sensitive to outliers
- KNN is expected to be one of the stronger method for this data set, where clusters of fires may have very similar parameters



Classification Performance

- All the classification methods we used had an accuracy performance ~60-65%
- We did see KNN and SVM perform better, most likely due to their improved performance with non-linear data
- Results could be skewed by imbalance of data towards smaller fires in certain locations

	accuracy
LDA	0.622102
RRLDA	0.623951
QDA	0.592774
Logistic	0.626594
KNN	0.642909
SVM	0.64456

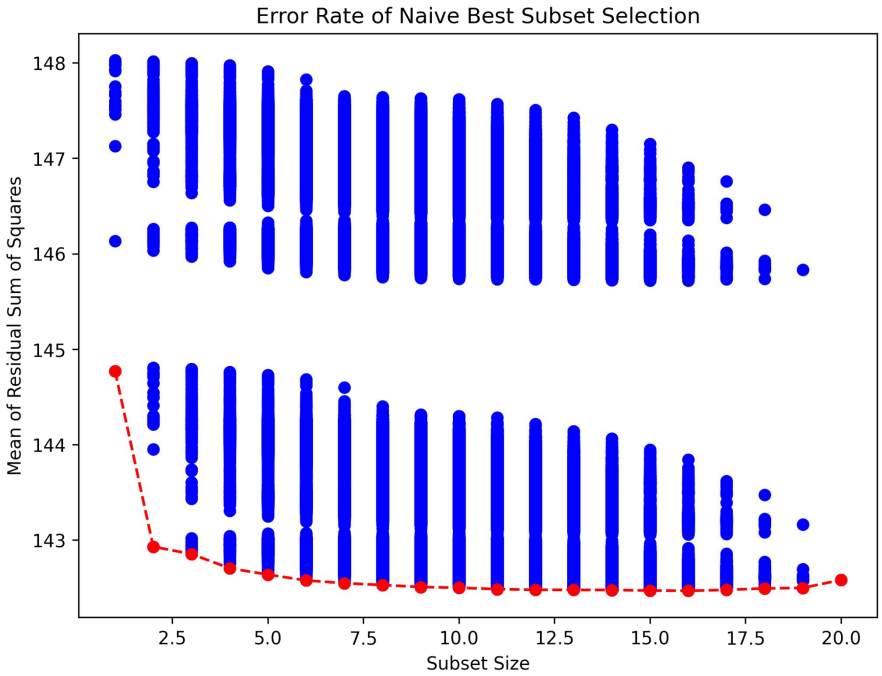
Regression Methods

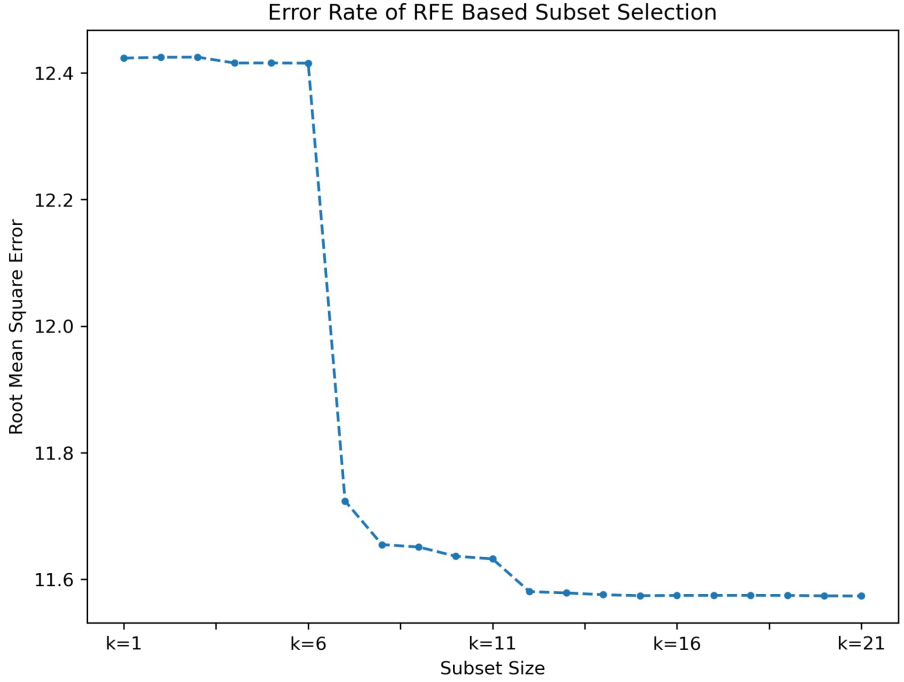
- We also explored several different regression methods with the target variable as the number of days it took to put out the wildfire
- The methods used here were:
  - Linear Regression
  - Ridge Regression
  - Lasso Regression
  - RFE Based Best Subset Selection
  - Principal Component Analysis

True Best Subset Selection Vs Recursive Feature Elimination Algorithm

- True best subset selection can get computationally time consuming for large number of columns.
  - For 20 variables we have to attempt 1,048,575 combinations of variables
- RFE starts with the full set of variables, then recursively considers smaller and smaller sets of variables.
  - The importance of a variable was determined by the magnitude of the coefficient during the linear fit.
- True best subset selection did only slightly better at the cost of tremendously more compute time

True Best Subset Selection Vs Recursive Feature Elimination Algorithm





True Best Subset Selection Vs Recursive Feature Elimination Algorithm

- Top 5 Variables Chosen By Each Algorithm

True Best Subset	Recursive Feature Elimination
FIRE_SIZE	tmean
LATITUDE	tmax
LONGITUDE	tmin
vpdmax	tmean-7
vpdmin	tmax-7

Regression Performance

- All the regression methods we used had an RMSE around 11.5 to 12.2.
- We did see best subset selection and PCA regression perform better.
  - Ridge/Lasso might have underperformed as there might be data columns antithetical to the predictions.
- We unfortunately did fairly poorly.

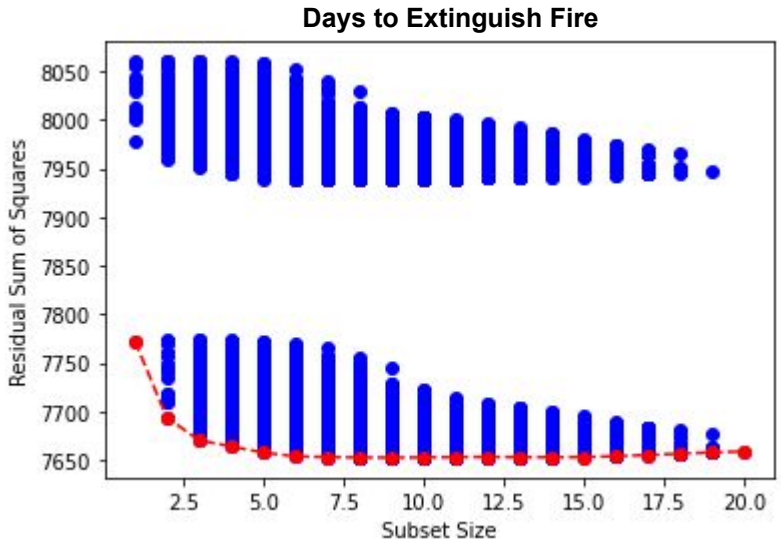
	score	rmse
Linear	0.038648	12.219802
Ridge	0.03853	12.220549
Lasso	0.006472	12.422618
RFE Based Best Subset	0.137647	11.573523
PCA	0.137504	11.57448

Dimensionality Reduction

- Dimensionality reduction was important for us for two main reasons, the first being that there are many variables that are unrelated to the target variable, and the second being that part of the purpose of this project is to determine which variables may be the most important in predicting lightning caused wildfires
- Here we compared two methods for regression:
  - 1) Best subset selection: Allows us to explain which parameters have the largest effect on the target variable, may lose relevant parameter interaction
  - 2) Principal Component Analysis (PCA): Captures parameter interaction, possibly improving model performance, reduces explainability of results
- In this case we think best subset selection will be most useful, since we are in the early stages of analysis and are trying to evaluate which parameters are the most significant, but in further development of models PCA would likely be the most useful

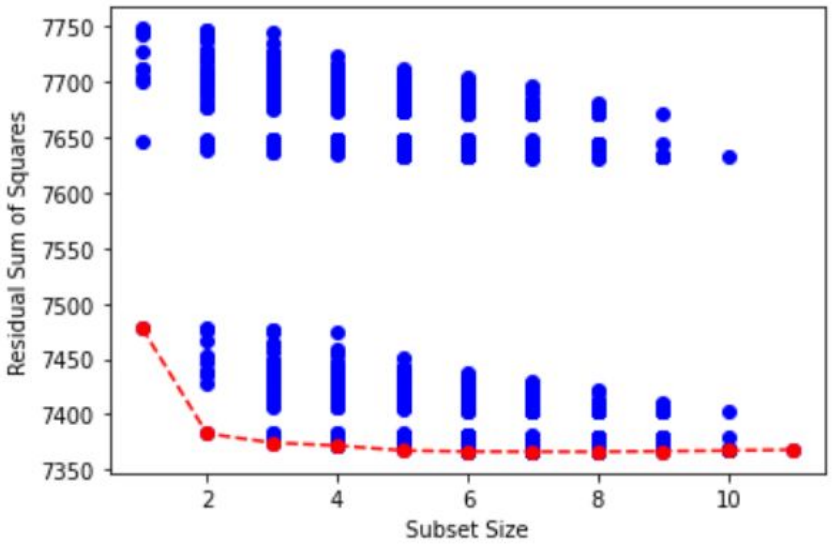
Best Subset Selection Results

- Linear regression applied to the number of days to extinguish the fire
- Subset sizes in the range 5-7 yield lowest residual sum of squares
- Variables most often selected
  - Fire Size
  - Latitude
  - Longitude
  - Temperatures

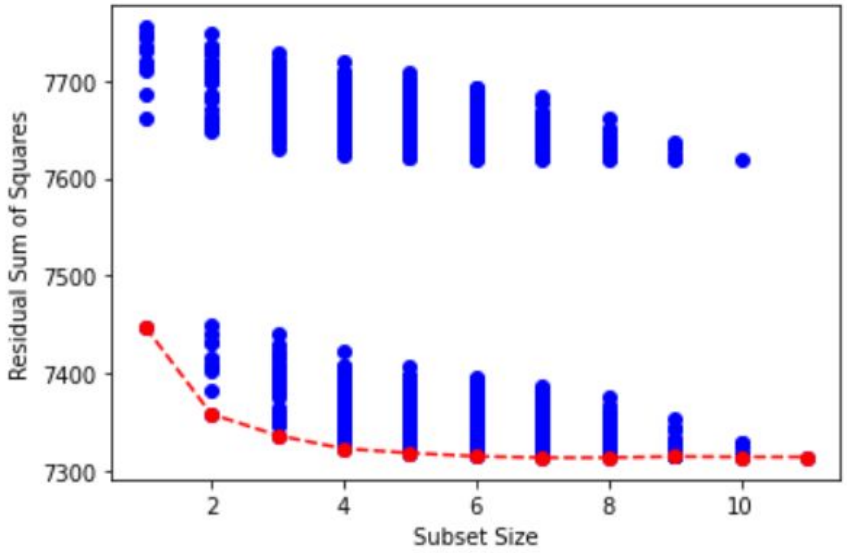


Best Subset Selection Results

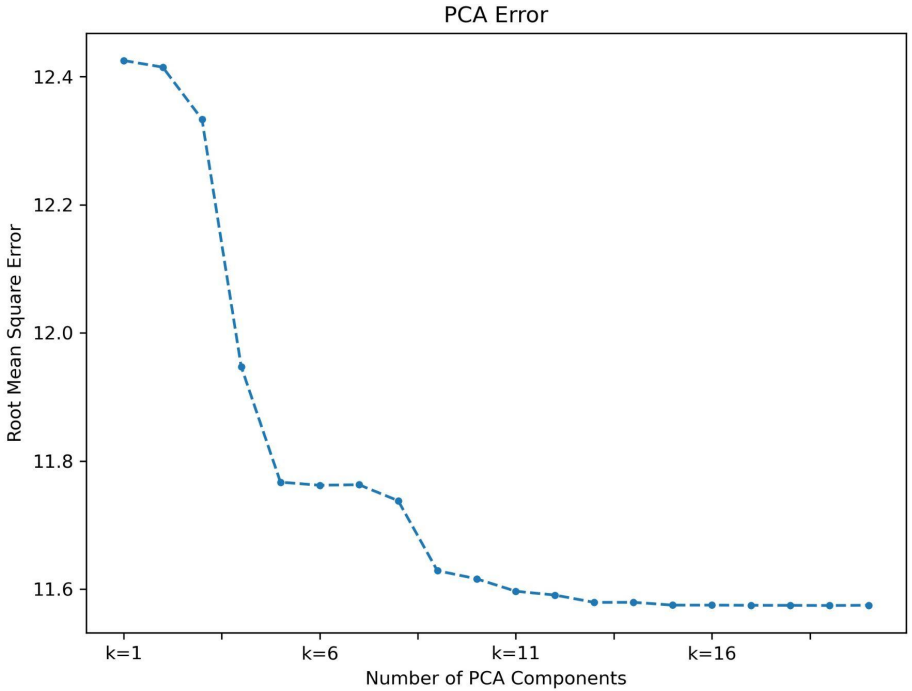
BSS with supplemental week-data removed



BSS with supplemental non-week-data removed



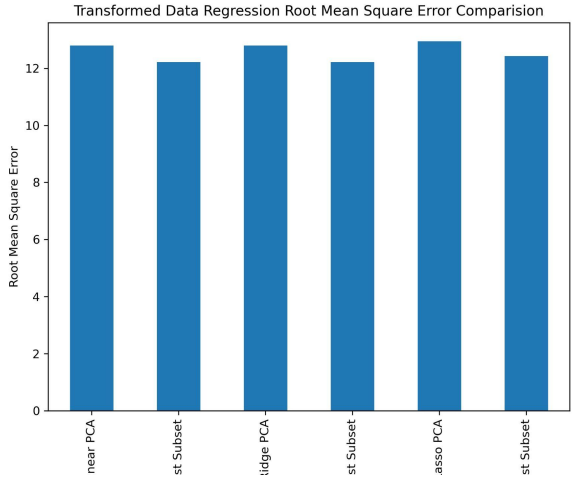
Principal Component Analysis Results



- The PCA method better captures parameter interaction and uses components of combined parameters to transform the data into lower dimensionality
- PCA has two clear "elbows" here, one at k=5 and another at k=9
- For our analysis of the method, we use 9 components to transform the data

Method Comparison

- Using a best subset transformation with 7 variables and a PCA transformation with 9 components, we compared 3 regression models
  - Linear Regression
  - Ridge Regression
  - Lasso Regression
- For these models, we get the following R^2 values and RMSE scores
- We see that the data transformed with the best subset method has higher R^2 values, and lower RMSE scores, and thus on these regression methods the best subset method outperforms PCA



Conclusions

- Dimensionality reduction methods indicate that only 5-7 variables are of primary importance for predicting the number of days to extinguish the fire
  - Location, Fire Size, and Temperature of the region seemed to be some of the most highly selected features
- Different classification methods yielded similar levels of accuracy
  - Further exploration of preprocessing methods may be necessary
  - There may be important features/data that are not included in the dataset
- Regression scores were near zero, which could suggest additional data preprocessing or features are required
- The sheer size of the dataset can make interpretability a challenge

Next Steps (If We Had More Time)

- There are still data sources we believe would be useful that we weren't able to compile, such as:
  - Lightning density
  - Terrain info
  - Specific information about wildfire mitigation/fighting at locations in data
- There are more questions we would like to explore or explore in greater depth
  - Can we predict regions at higher risk for rapid spreading of wildfires?
  - How many wildfires can a given location expect in subsequent years?
- Adapted sampling methods to reduce data imbalance
- Use transformed data in models to see effects
- Additional ML techniques
  - Neural networks, tree-based methods

Citations

- Hastie, T., Tibshirani, R.,, Friedman, J. (2001). *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc..
- Wikipedia contributors. (2021, September 24). August 2020 California lightning wildfires. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:03, December 9, 2021, from https://en.wikipedia.org/w/index.php?title=August_2020_California_lightning_wildfires&oldid=1046111209
- Rice, Doyle, et al. "California Fires: This Is How A Lightning Storm Can Start a Wildfire." *USA Today*, Gannett Satellite Information Network, 23 Aug. 2020, https://www.usatoday.com/in-depth/graphics/2020/08/21/california-fires-how-lightning-storm-can-start-wildfire/3412729001/.
- Schwartz, John, and Veronica Penney. "In the West, Lightning Grows as a Cause of Damaging Fires ..." *The New York Times*, 23 Oct. 2020, https://www.nytimes.com/interactive/2020/10/23/climate/west-lightning-wildfires.html.
- D. M. Romps, J. T. Seeley, D. Vollaro, J. Molinari, Projected increase in lightning strikes in the United States due to global warming. *Science* 346, 851–854 (2014).
- Iii.org. 2021. *Facts + Statistics: Wildfires | III*. [online] Available at: <https://www.iii.org/fact-statistic/facts-statistics-wildfires> [Accessed 9 December 2021].
- Nifc.gov. 2021. *Lightning-caused wildfires | National Interagency Fire Center*. [online] Available at: <https://www.nifc.gov/fire-information/statistics/lightning-caused> [Accessed 9 December 2021].
- Sande, S., 2021. *Pros and Cons of popular Supervised Learning Algorithms*. [online] Medium. Available at: <https://medium.com/analytics-vidhya/pros-and-cons-of-popular-supervised-learning-algorithms-d5b3b75d9218> [Accessed 9 December 2021].

Citations (cont.)

- Gonzalez P.L., Cléroux R., Rioux B. (1990) Selecting the Best Subset of Variables in Principal Component Analysis. In: Momirović K., Mildner V. (eds) Compstat. Physica-Verlag HD. https://doi.org/10.1007/978-3-642-50096-1_18
- GeeksforGeeks. (2020, June 4). *ML - Advantages and Disadvantages of Linear Regression*. Retrieved December 9, 2021, from https://www.geeksforgeeks.org/ml-advantages-and-disadvantages-of-linear-regression/
- GeeksforGeeks. (2020b, July 17). *Advantages and Disadvantages of different Regression models*. https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-regression-models/?ref=rp
- GeeksforGeeks. (2020c, July 17). *Differentiate between Support Vector Machine and Logistic Regression*. Retrieved December 9, 2021, from https://www.geeksforgeeks.org/differentiate-between-support-vector-machine-and-logistic-regression/?ref=rp
- GeeksforGeeks. (2020d, September 2). *Advantages and Disadvantages of Logistic Regression*. Retrieved December 9, 2021, from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/?ref=rp
- GeeksforGeeks. (2020e, September 28). *Advantages and Disadvantages of different Classification Models*. Retrieved December 9, 2021, from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-classification-models/?ref=rp
- GeeksforGeeks. (2020f, December 22). *Support vector machine in Machine Learning*. https://www.geeksforgeeks.org/support-vector-machine-in-machine-learning/?ref=rp