TAMU 2023 Data Science Competition

Blaze Abaters

April 18, 2023

Agenda

- Introduction
  - Problem Statement
- Method
  - Data
- Models
- Visualization
  - Data Warehouse Architecture
  - Dashboards
- Conclusions
  - Future Research
  - Recommendations

Introduction

Complexity

Academia and Startups

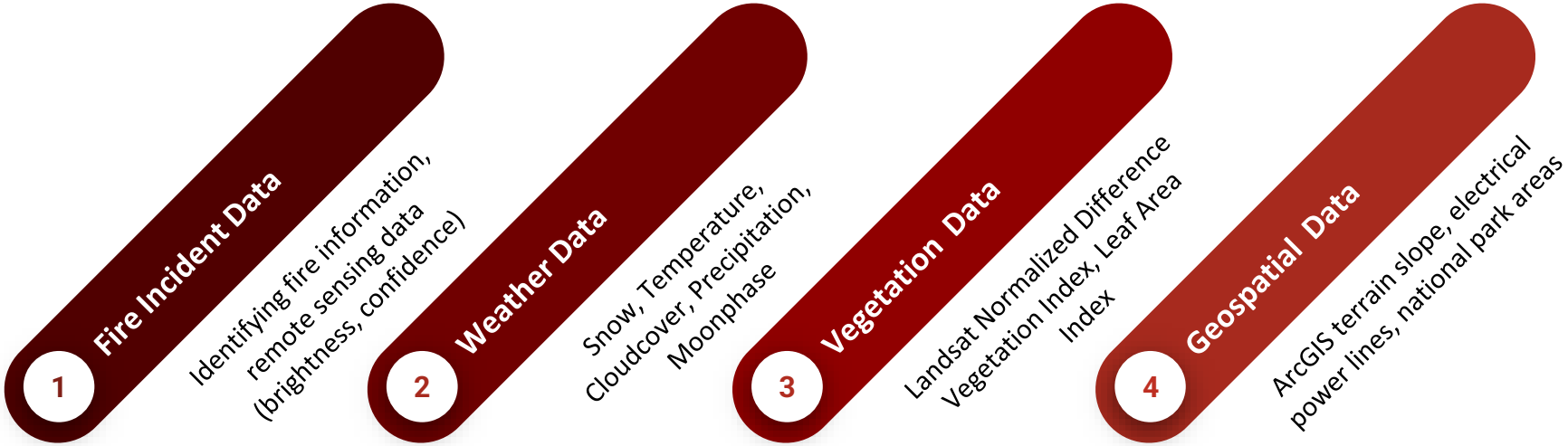
Time Constraint

Problem Statement

What features lead to increased wildfire occurrences in the Denver, CO region compared to the Columbus, OH region?

Are there any peculiar variables affecting wildfires?

Data



Data pulled for:
January 2014 - December 2022

Data sourced from NASA Giovanni, NASA FIRMS, Visual Crossing, ArcGIS, and the National Interagency Fire Center

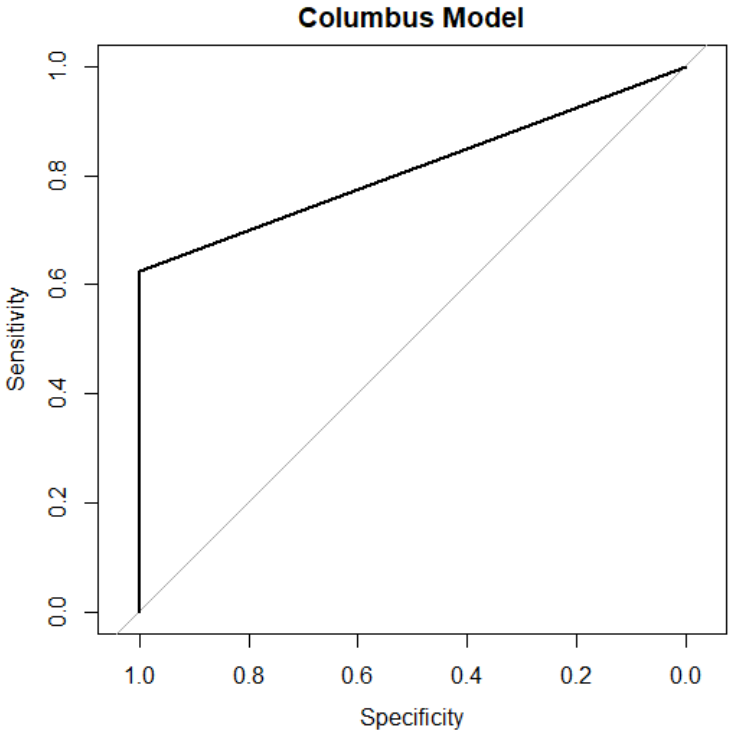
Methodology

Research and Data Collection	Data Wrangling	EDA and Visualization	Modelling and Performance	Results & Recommendations
Spend 70% of effort here	Merge datasets on Latitude, Longitude, Year and Month columns	Perform variable selection (lasso and forward/backward)	Utilize brightness temperature and coded confidence as the response	Gather facts and intuitions
Conduct literature study on wildfire modelling	Obtain final dataset with 108 observations and 40 variables for each region	Understand variables like NDVI, Precipitation, Soil Moisture, etc.	Implement 1 Classification and 3 Regression models for each region	Observe interaction of variables like NDVI, Wind, and Soil Composition
Gaining industry knowledge	Average adjacent rows to fill missing values	Analyze trends in variables	Test model accuracy on unseen test data	Analyze importance of variables like Moonphase and Transpire
Obtain ready to download data	Drop “severe risk” column	Evaluate differences between two regions	Ensure models have reliable outputs and validate problem statement	Assess the reliability of NN, LR, SVM and RF models for wildfire prediction
Obtain remaining data via API pulls		Communicate relationship between variables to users via Power BI dashboard and ArcGIS map		

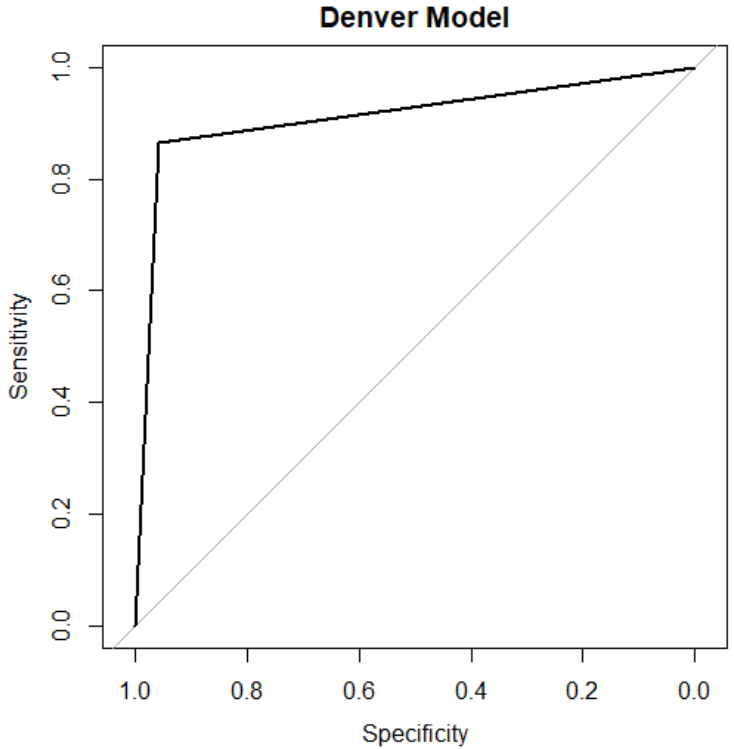
Models: Details

Model	Type	Parameters	Response Variable	Test Error Metric	Improving Accuracy
Logistic Regression	Classification	0.5 threshold for accuracy calculation	Confidence	ROC	Perform variable transformation
XG Boosting	Regression	max_depth, alpha, gamma, lambda	Brightness Temp	RMSE	Test wider range of values for hyper parameters and use nested CV
SVM	Regression	C, Gamma, Kernel	Brightness Temp	RMSE	Reduce feature correlations
Neural Networks	Regression	ReLu, adams, 3 layers, 25 epochs	Brightness Temp	MSE/MAE	Obtain more data and perform variable selection

Logistic Regression: ROC Plots

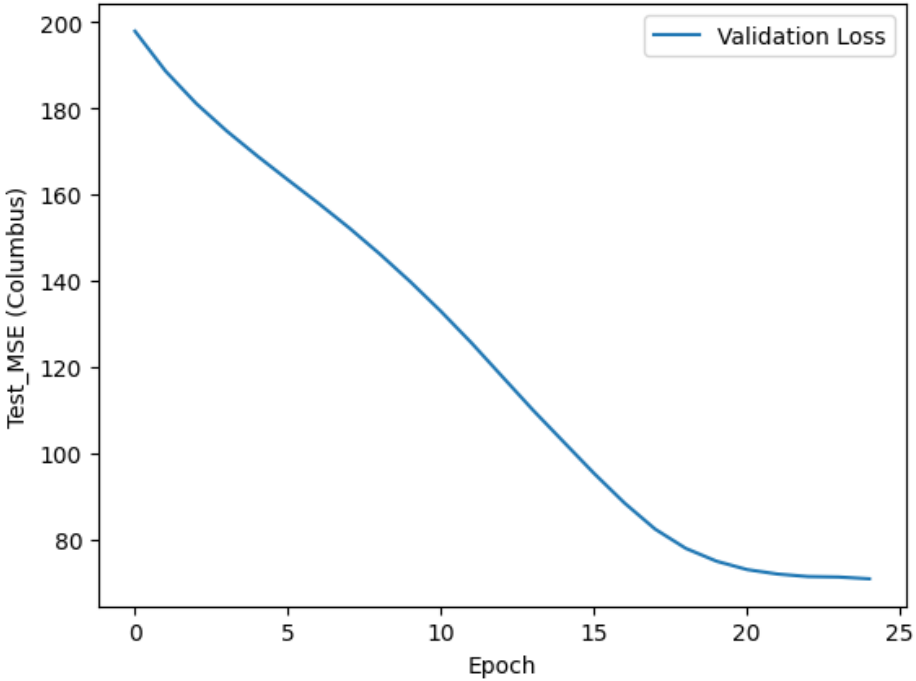


Test Denver, Train Columbus

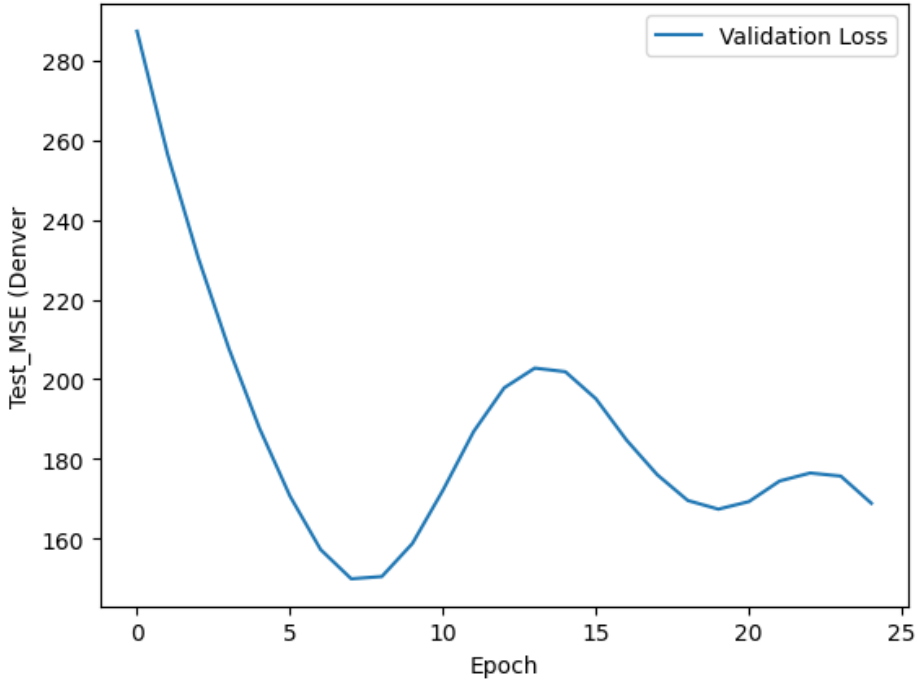


Test Columbus, Train Denver

NN: Test MSE



Test Denver, Train Columbus



Test Columbus, Train Denver

Models: Details

Model	Type	Parameters	Response Variable	Test Error Metric	Improving Accuracy
Logistic Regression	Classification	0.5 threshold for accuracy calculation	Confidence	ROC	Perform variable transformation
XG Boosting	Regression	max_depth, alpha, gamma, lambda	Brightness Temp	RMSE	Test wider range of values for hyper parameters and use nested CV
SVM	Regression	C, Gamma, Kernel	Brightness Temp	RMSE	Reduce feature correlations
Neural Networks	Regression	ReLu, adams, 3 layers, 25 epochs	Brightness Temp	MSE/MAE	Obtain more data and perform variable selection

Data Warehouse Architecture

Data Source

- NASA Giovanni
- NASA FIRMS
- NIFC
- Visual Crossing

ETL Process

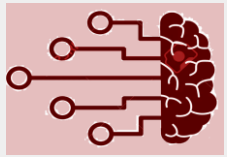
Data Pipelines

Data Transformation

Storage

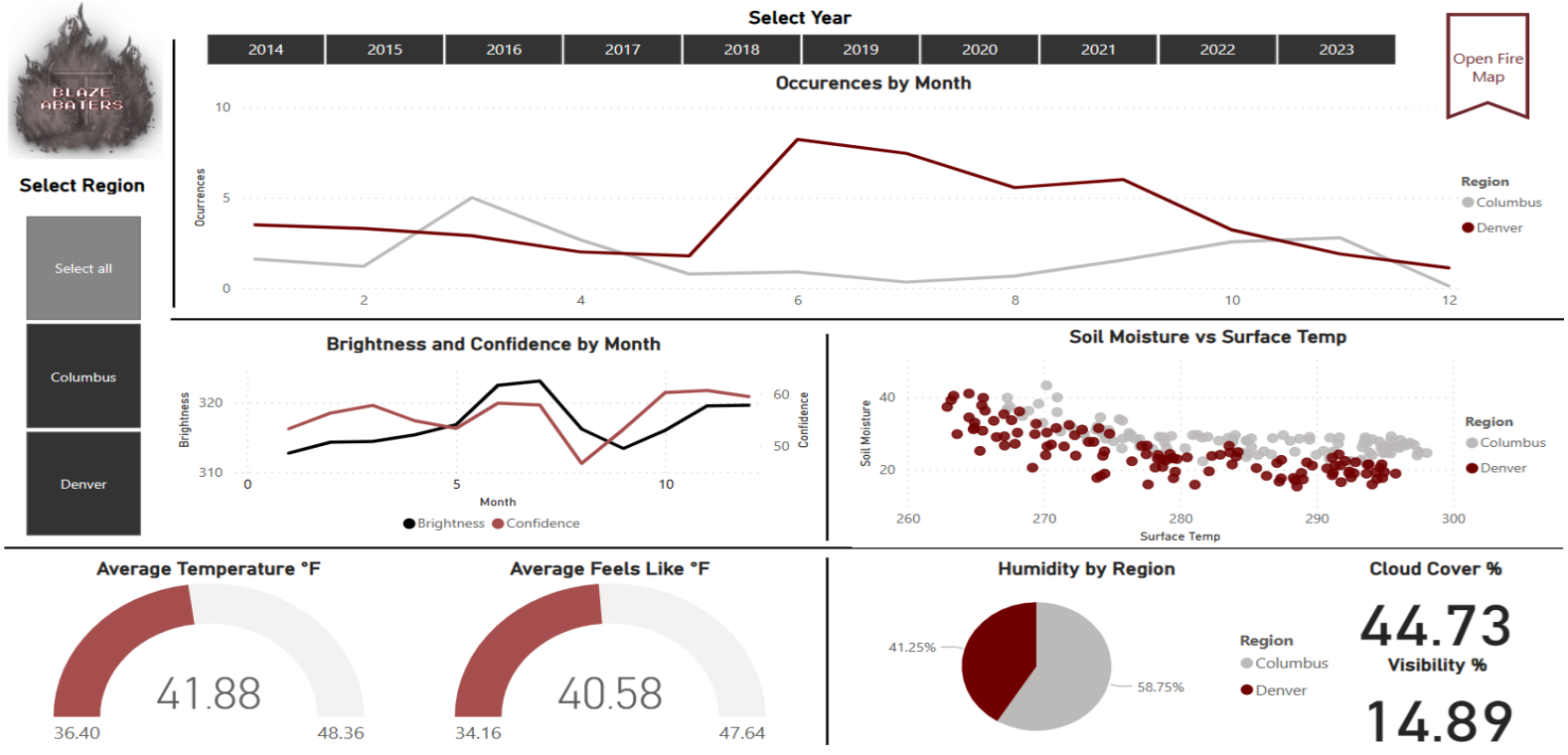
Azure Synapse Analytics

Data Visualization

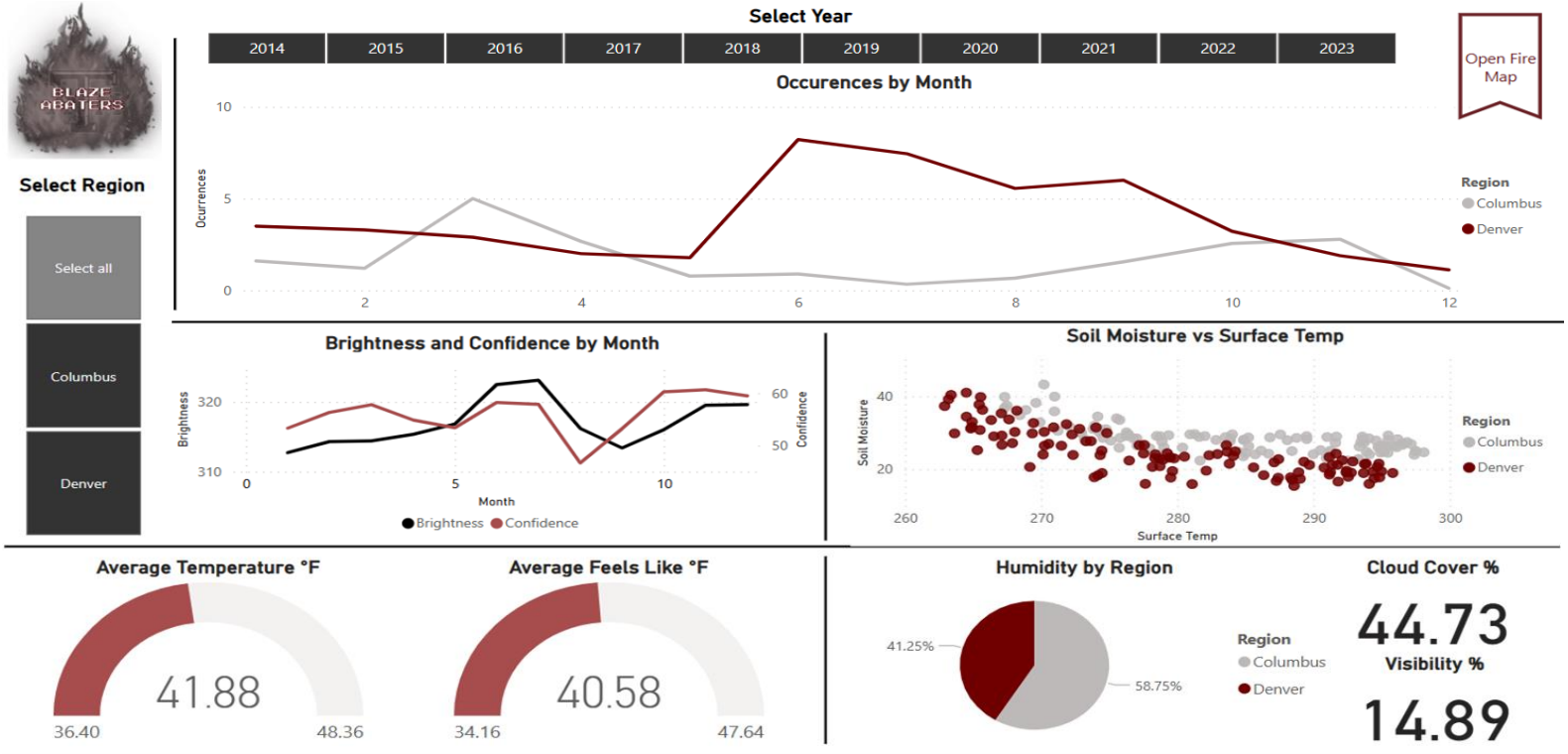


ML Models

Data Visualization: Power BI



Data Visualization: Power BI

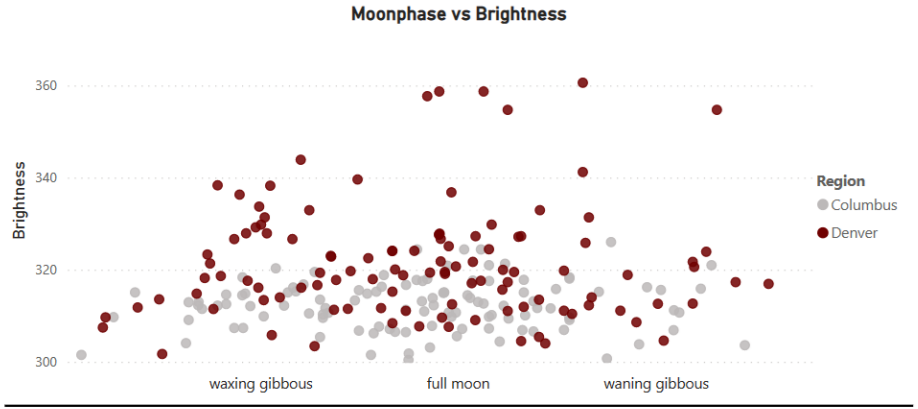


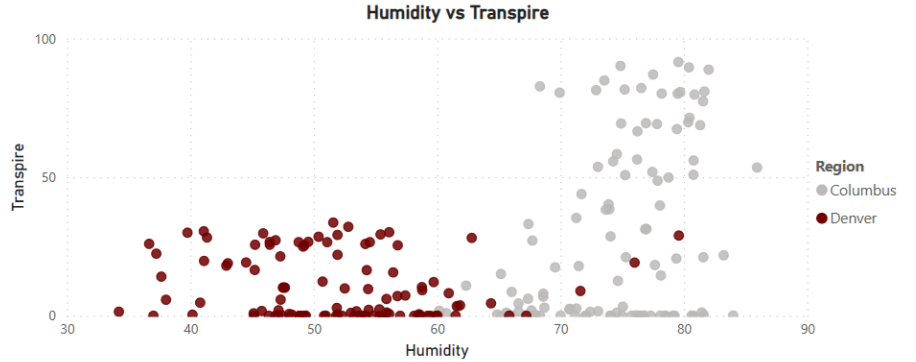
Data Visualization: Power BI

Future Research

- Lunar phase impacts on the occurrence of wildfires.
- How natural fluctuations in transpiration can affect the occurrence and severity of wildfires.
- Targeted studies on how airflow in ridges and draws may impact fires.
- How the use of prescribed fires affects naturally occurring wildfires.
- How time of day impacts fires.

TEXAS A&M UNIVERSITY





Open Fire Map

Select Region

Select all

Columbus

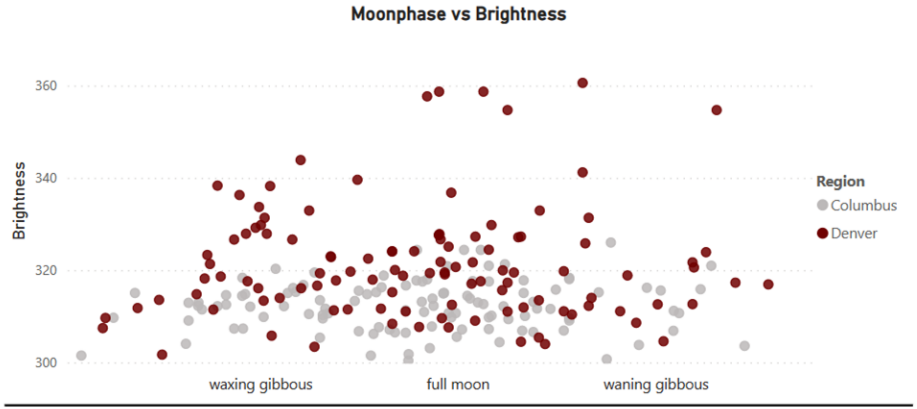
Denver

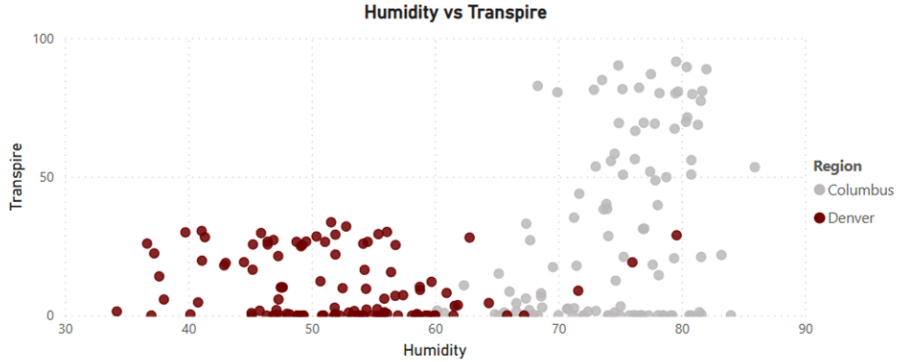
Data Visualization: Power BI

Future Research

- Lunar phase impacts on the occurrence of wildfires.
- How natural fluctuations in transpiration can affect the occurrence and severity of wildfires.
- Targeted studies on how airflow in ridges and draws may impact fires.
- How the use of prescribed fires affects naturally occurring wildfires.
- How time of day impacts fires.

TEXAS A&M UNIVERSITY





Open Fire Map

Select Region

Select all

Columbus

Denver

Conclusions

What features lead to increased wildfire occurrences in the Denver, CO region compared to the Columbus, OH region?

NDVI, LAI, Humidity, Windspeed, and Visibility

Are there any peculiar variables affecting wildfires?

Moon Phase and Transpiration

Future Research

Lunar Phase

Lunar phase is known to impact surface temperature and humidity.

How does this impact wildfires?

Transpiration

Current literature is focused primarily on how wildfires impact transpiration.

Can natural fluctuations in transpiration lead to the occurrence of wildfires?

Turbulent Air Flow

Do lines of demarcation impact wildfire occurrence?

Are fires more likely to occur along draws or ridges?

Prescribed Fires

Does the utilization of prescribed fires result in less wildfire occurrence?

Time of Day

How does time of day impact the occurrence and severity of wildfires?

Does lunar phase affect the occurrence of night wildfires?

Recommendations

Research grants for future research topics

Comprehensive data repository

Money to maintain/support data warehouse, cloud storage and BI platform

Increase data science positions in related industry

Emphasis on synergy of variables

Using ensembling of models

Using remote sensing data for real time prediction

QUESTIONS

Backup Slides

Expanded Methodology

1- Research

- **Previous Models**
  - Discover what models have been successful in the past.
- **Fire Science**
  - Identify potential dependent and independent variables.
- **Industry Knowledge**
  - Learn what is important to industry experts.

2- Data Collection

- **Fire Incident Data**
- **Weather Data**
- **Vegetation Data**
- **Geospatial Data**

3- Data Wrangling

- **Handle Missing Values**
  - Imputation (averaging out values in adjacent rows) or Drop
- **Merging Data**
  - Merge fire incident, weather, and vegetation data using the Year, Month, Lat, and Long columns

4- EDA and Visualization

- **Correlation Analysis**
  - Pearson/Spearman
  - Pairplots
- **Distribution Analysis**
  - Skewness and Outliers
- **Additional Visualization**
  - Strip Plots, Scatter Plots, Cross Tab

5- Modeling and Performance

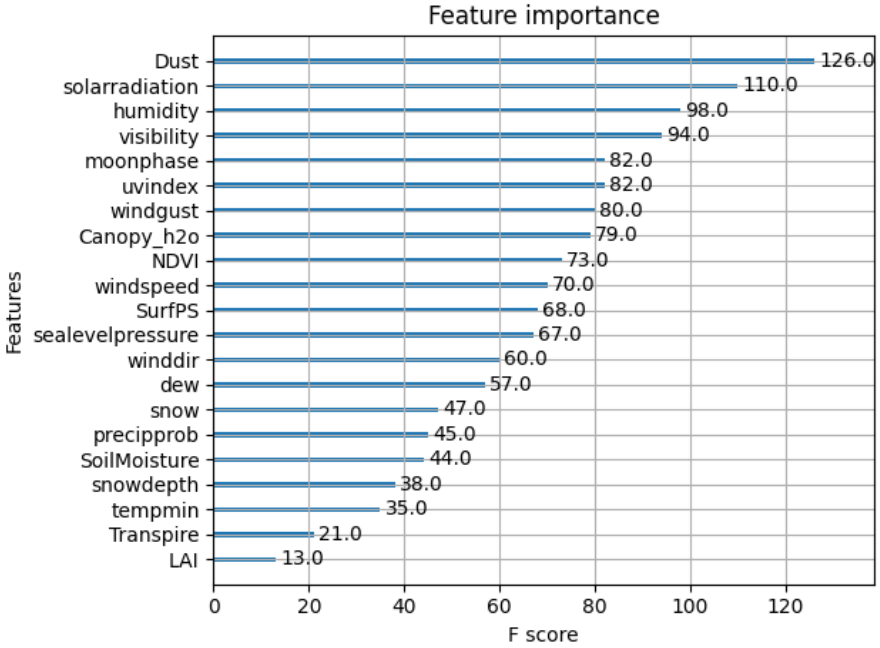
- **Forward/Backward Variable Selection**
- **Lasso Regression**
- **Support Vector Machine**
- **Gradient Boosting**
- **Neural Network**
- **Logistic Regression**

6-Analysis and Recommendations

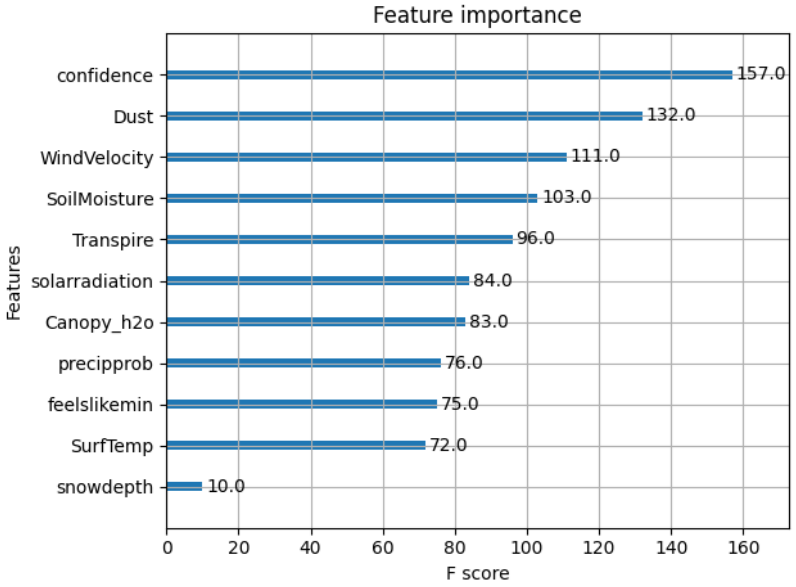
- **Describe insights**
- **Answer research questions**
- **Provide recommendations for model improvements**
- **Provide topics for future study**

Models: Gradient Boosting

Feature Importance Plot: Model trained using Denver data, tested on Columbus data

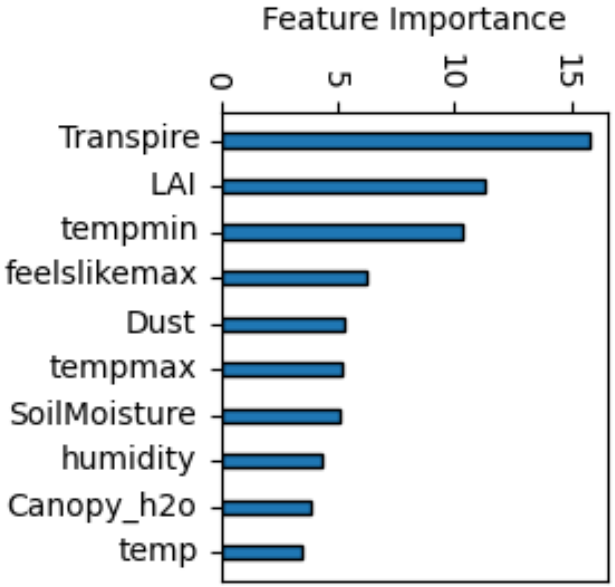


Feature Importance Plot: Model trained using Columbus data, tested on Denver data

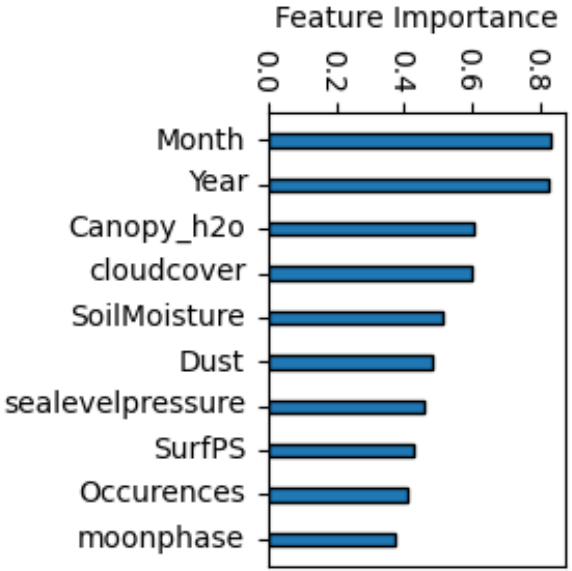


Models: SVM

Feature Importance Plot: Model trained using Denver data, tested on Columbus data



Feature Importance Plot: Model trained using Columbus data, tested on Denver data



Models: F/B Selection & Lasso Regression

- Variables chosen by Forward/Backward (F/B) selection were used as the input for the gradient boosting and neural network models, and those chosen by lasso regression for logistic regression and SVM.

Forward/Backward Selection

Denver Model

NDVI
SurfPS
Dust
SoilMoisture
LAI
Canopy_h2o
Transpire
tempmin
Humidity
precipprob
snow
snowdepth
windgust
windspeed
winddir
sealevelpressure
visibility
solarradiation
uvindex
moonphase
dew

Columbus Model

SurfTemp
Dust
SoilMoisture
WindVelocity
Canopy_h2o
Transpire
confidence
feelslikemin
precipprob
snowdepth
solarradiation

Lasso Regression

Denver Model

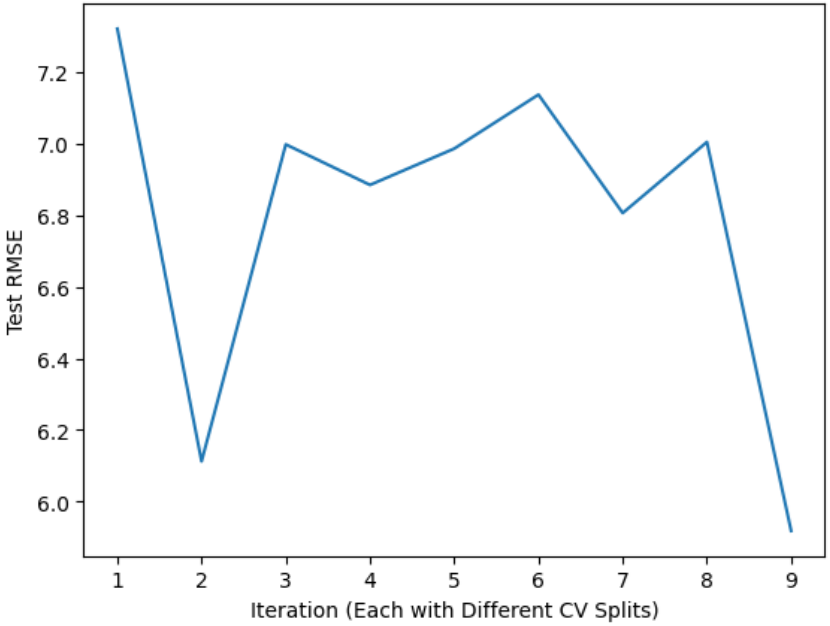
NDVI
SurfPS
Dust
SoilMoisture
LAI
Transpire
confidence
tempmax
tempmin
feelslikemax
dew
humidity
precipcover
snow
snowdepth
windspeed
sealevelpressure
solarradiation
solarenergy
uvindex
moonphase

Columbus Model

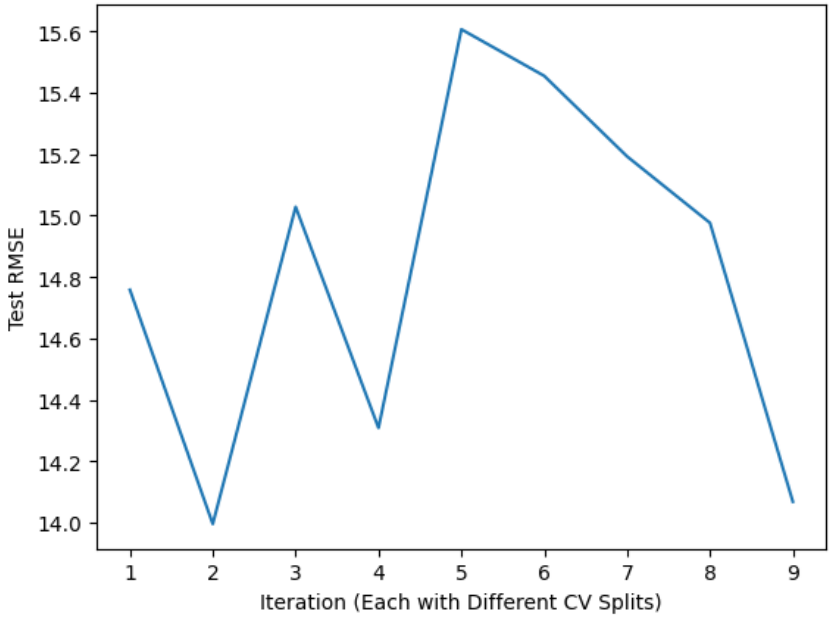
Dust
SoilMoisture
Transpire
humidity
precipprob
snowdepth
sealevelpressure
cloudcover
moonphase
confidence
WindVelocity

Legend:
Yellow: Common to both regions for both methods Green: Found only in Denver region for both methods

Gradient Boosting: Test RMSE

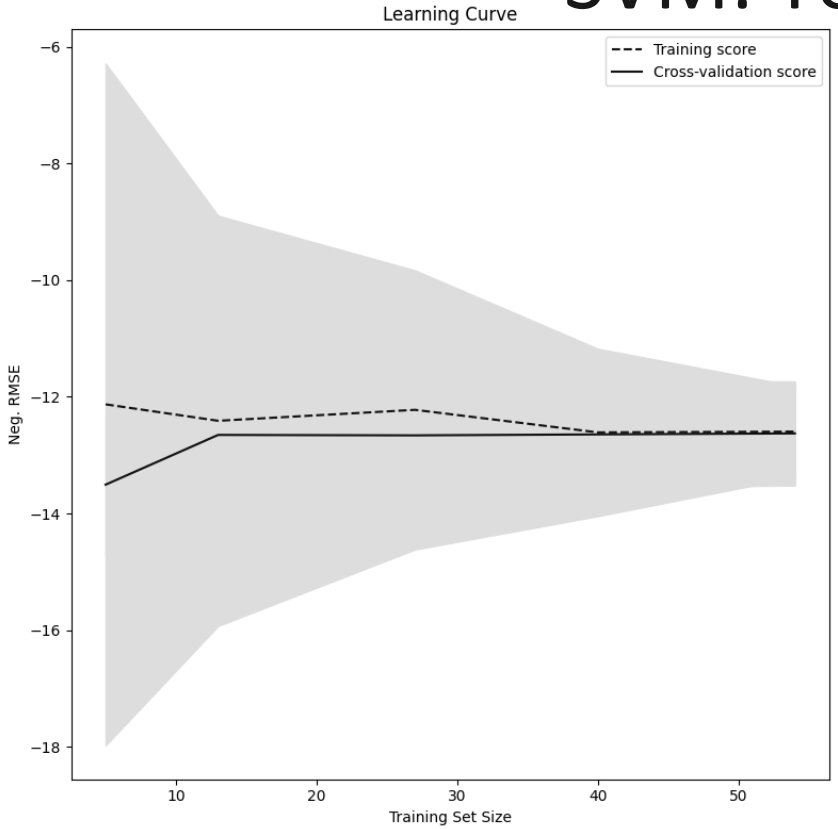


Test Columbus, Train Denver

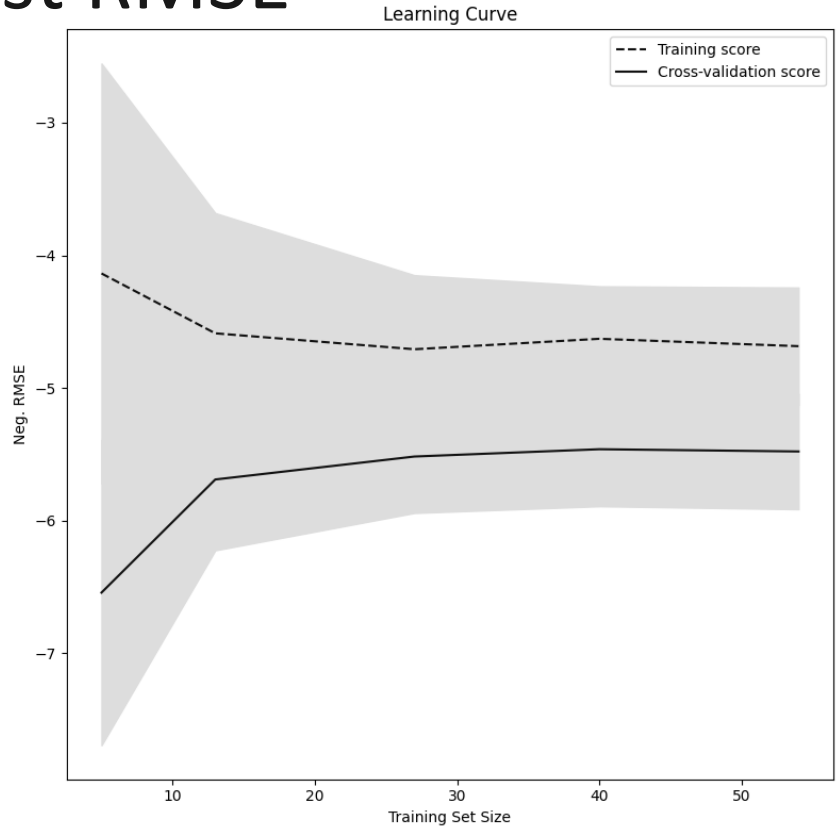


Test Denver, Train Columbus

SVM: Test RMSE



Test Columbus, Train Denver



Test Denver, Train Columbus

Insights

Facts/Intuition	Validated by model(s)	Comments
More fires in Denver area compared to Columbus area	Yes	Occurrences Variable, and Dashboard
Multiple factors are responsible for the wildfire	Yes	Variable Selection
Denver area is more susceptible to wildfires	Yes	Brightness and Confidence, Dashboard
More healthy vegetation in Columbus area	Yes	NDVI Index
Interaction of variables is crucial for understanding wildfire	Yes	Difference in models for the two regions, Optimal variable selection
Indirect variables might affect wildfire	Yes	Inclusion of moonphase in the models
Reliability of results	Yes	Actual test error
Importance of ensembling	Pending	Work in progress
Accuracy of results	Pending	Work in Progress
Going in the right direction	Yes	Our entire work.

References

Day, Joshua. 2016. “The Wind Zone Model.” *Risk Management Magazine.* U.S. Army Combat Readiness Center. https://safety.army.mil/MEDIA/Risk-Management-Magazine/ArtMID/7428/ArticleID/5514/The-Wind-Zone-Model.

Esri. 2023. “Terrain: Slope Map.” https://elevation.arcgis.com/arcgis/rest/services/WorldElevation/Terrain/ImageServer.

Esri Data and Maps. 2022. “USA Parks.” https://services.arcgis.com/P3ePLMYs2RVChkJx/arcgis/rest/services/USA_Parks/FeatureServer/0.

Federal_User_Community. 2022. “US Electric Power Transmission Lines OGC.”. https://services2.arcgis.com/FiaPA4ga0iQKduv3/arcgis/rest/services/US_Electric_Power_Transmission_Lines_OGC/OGCFeatureServer.

Halperin, Sandra and Oliver Heath. 2017. “Political Research: Methods and Practical Skills.” Oxford University Press: Oxford.

Kohyama, Tubasa, and John M. Wallace. 2016. “Rainfall Variations Induced by the Lunar Gravitational Atmospheric Tide and their Implications for the Relationship Between Tropical Rainfall and Humidity.” *Geophys. Res. Lett.*, 43, 918– 923, doi:10.1002/2015GL067342.

National Aeronautics and Space Administration. 2023a. “Giovanni:The Bridge Between Data and Science v 4.38.” *EarthData*. https://giovanni.gsfc.nasa.gov/giovanni/.

National Aeronautics and Space Administration. 2023b.”Fire Information and Resource Management System: FIRMS Fire Map.” https://firms.modaps.eosdis.nasa.gov/map/#d:24hrs;@0.0,0.0,3z.

National Interagency Fire Center. 2022. “Wildland Fire Incident Locations.” https://data-nifc.opendata.arcgis.com/datasets/nifc::wildland-fire-incident-locations/about.

Visual Crossing Corporation. 2023. “Total Weather Data.” https://www.visualcrossing.com/weather-data