Project Toto:
Real Time Tornado Risk Prediction and Assessment in Iowa

Steve Veldman, Dharti Seagraves, Forough Mofidi, Michael Goodman
University of Chicago | MS Applied Data Science | Spring 2024

ADSP 32019 ON01 Real-Time Intelligent Systems

Table of contents

01 **Problem Statement**

Objective

02 **Data Source**

Historical Data, Data Preprocess

03 **Modeling**

Machine learning model; XGboost

04 **Real Time System**

Server, Client Set up

05 **Results**

User interface for live alerts

06 **Source Code**

Github Repository and background

01

Problem Statement

Problem Statement

- Studying tornado occurrences and predicting their probability using machine learning techniques is crucial for improving preparedness and reducing potential harm and loss of life.

- Machine learning algorithms can leverage historical data and weather patterns to identify **factors contributing to tornado formation**, aiding in the development of early warning systems.

- This empowers communities and emergency responders to **take timely actions**, such as issuing alerts and implementing evacuation procedures, ultimately saving lives and mitigating the impact of natural disasters.

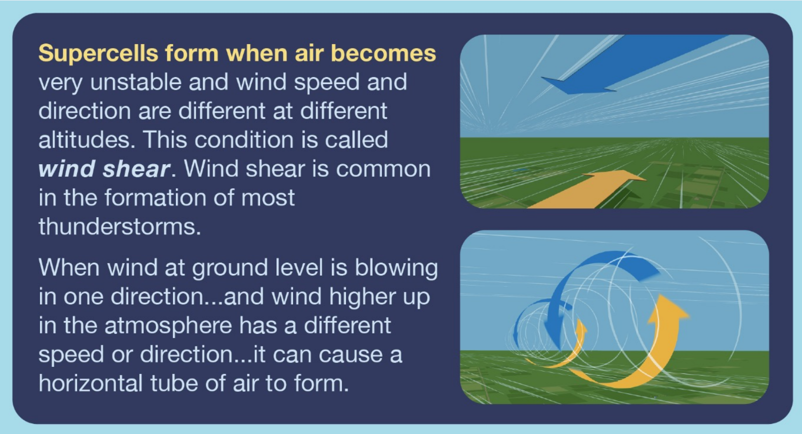
- The objective of this project is to develop and implement **a framework that enables tornado risk assessment in real time**, reducing public safety authorities’ response time during a disaster.

Background

❑ The diagrams to the right explain the fundamental components of a tornado formation based on scientific studies. **We will attempt to use a combination of these metrics to assign a tornado formation risk**

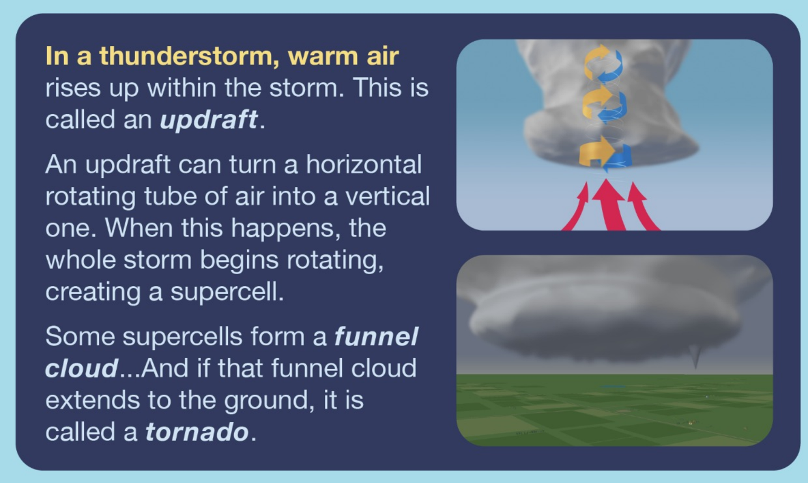
❑ As **Iowa is mostly topographically consistent**, the initial risk is constant across the state. This would not be the case in larger states like Illinois or Texas.

❑ **Wind Shear** is often not a metric included in weather APIs, we will calculate our own measure by understanding the difference in magnitude of wind vectors at different altitudes.



Supercells form when air becomes very unstable and wind speed and direction are different at different altitudes. This condition is called ***wind shear***. Wind shear is common in the formation of most thunderstorms.

When wind at ground level is blowing in one direction...and wind higher up in the atmosphere has a different speed or direction...it can cause a horizontal tube of air to form.



In a thunderstorm, warm air rises up within the storm. This is called an ***updraft***.

An updraft can turn a horizontal rotating tube of air into a vertical one. When this happens, the whole storm begins rotating, creating a supercell.

Some supercells form a ***funnel cloud***...And if that funnel cloud extends to the ground, it is called a ***tornado***.

Source: NOAA SciJinks

02

Data Source

Data Sources

Historical Tornado Information

- NOAA: National Oceanic and Atmospheric Administration
- https://www.ncdc.noaa.gov/stormevents/choosedates.jsp?statefips=19%2CIOWA#
- Storm Events database to request CSV of Tornado events in Iowa, their severity, and location from 1/1/2018 through 12/31/2023

Historical Weather Sensor Data

- https://open-meteo.com/en/docs/historical-weather-api#location_mode=csv_coordinates
- This is a free API that allows us to request sensor information going back 6 years (5 years of training, 1 year of test). To avoid API throttle limits, the GUI was used with the centroid file as our list of locations

Data Sources

Data Integration

- Hourly weather sensor data was matched with records of tornado occurrences within 3 hours
- Resulting dataset includes a binary response variable indicating whether a tornado occurred during that window, suitable for training a supervised machine learning model

Adaptable Framework

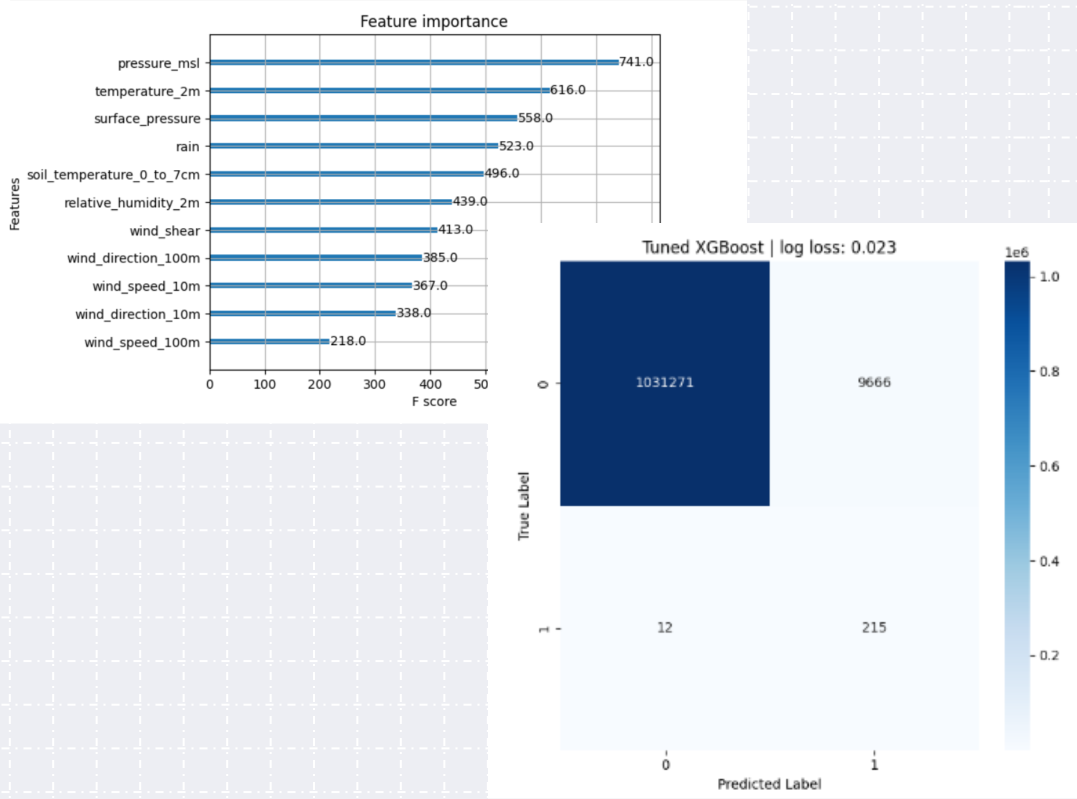
- For this proof-of-concept collected 5 years of data at one-hour intervals
- Future models can be fine-tuned based on the desired frequency of sensor updates

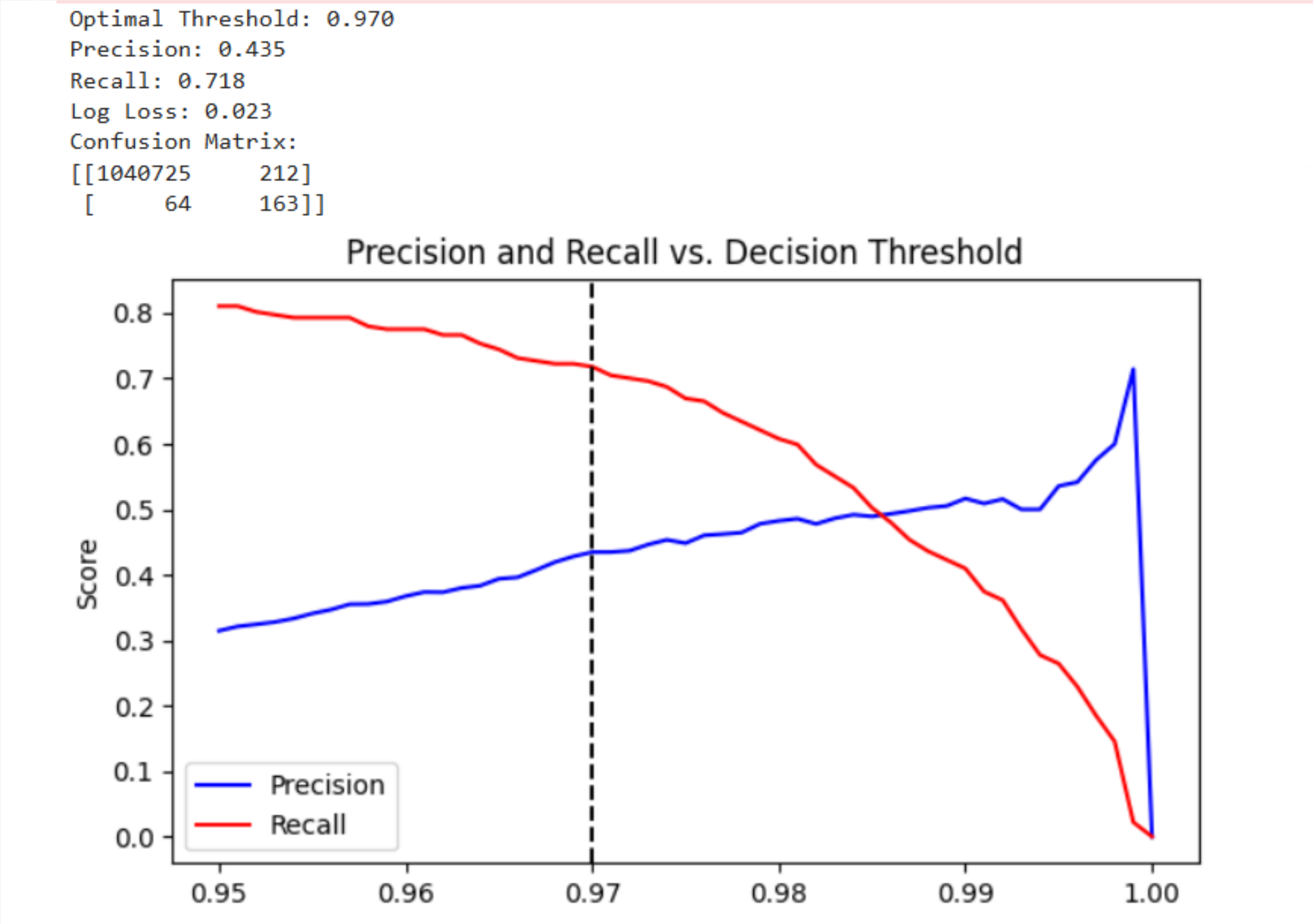
03

Modeling

Model Selection and Tuning

- ➔ Due to HIGH class imbalance **scale_pos_weight** parameter was utilized in all models
- ➔ Defined as negative events/positive events
- ➔ **Log Loss** was used to evaluate and train
- ➔ **GridSearch** used to tune hyperparameters
- ➔ Adjusted classification threshold to have **optimized recall/precision**



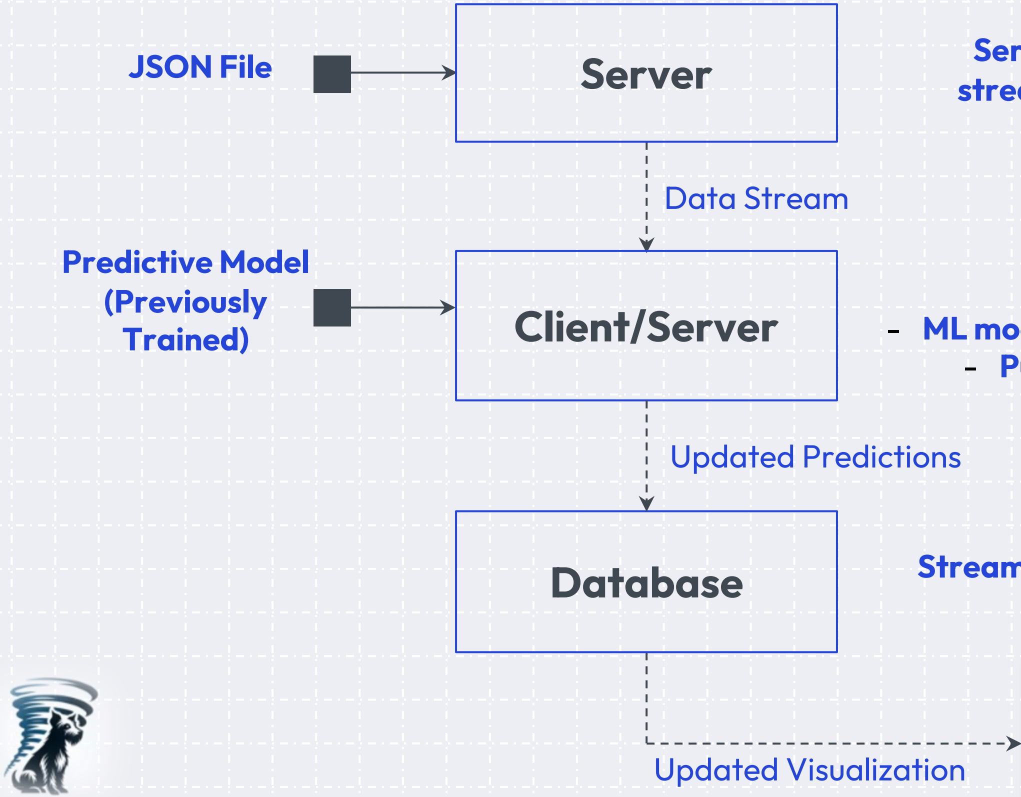


```
Optimal Threshold: 0.970
Precision: 0.435
Recall: 0.718
Log Loss: 0.023
Confusion Matrix:
[[1040725     212]
 [      64     163]]
```

04

Data Streaming

Project Architecture



JSON File

Server

Data Stream

Predictive Model (Previously Trained)

Client/Server

Updated Predictions

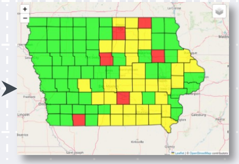
Database

Updated Visualization

Server loads weather data from JSON file and streams entries one line at a time, simulating live data stream

- Collects incoming weather data
- ML model generates predictions for each incoming record
- Pushes updated county predictions to database

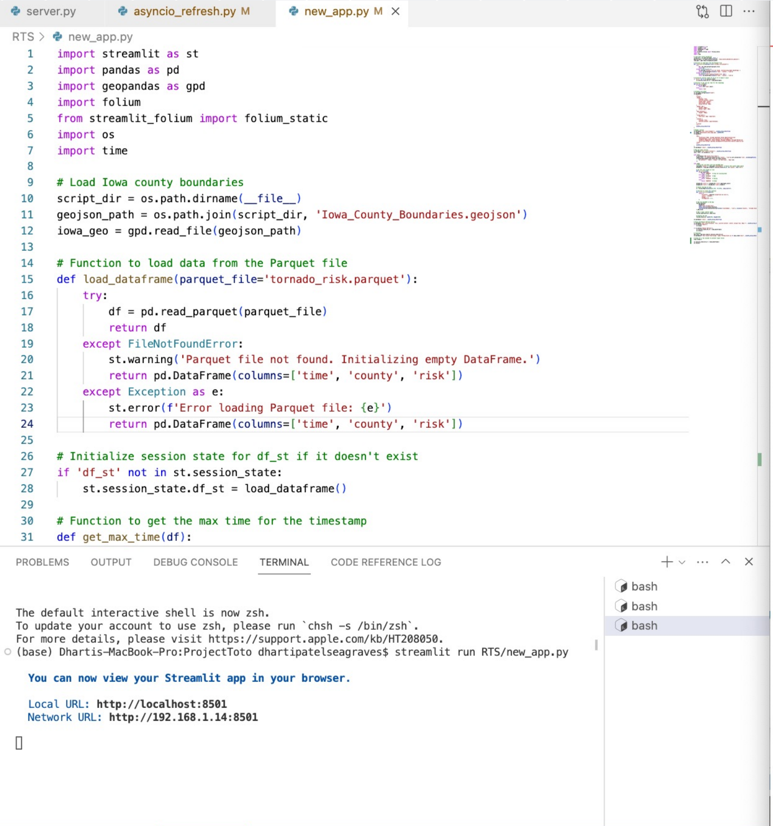
Streamlit application refreshed visualization based on current probabilities in database



05

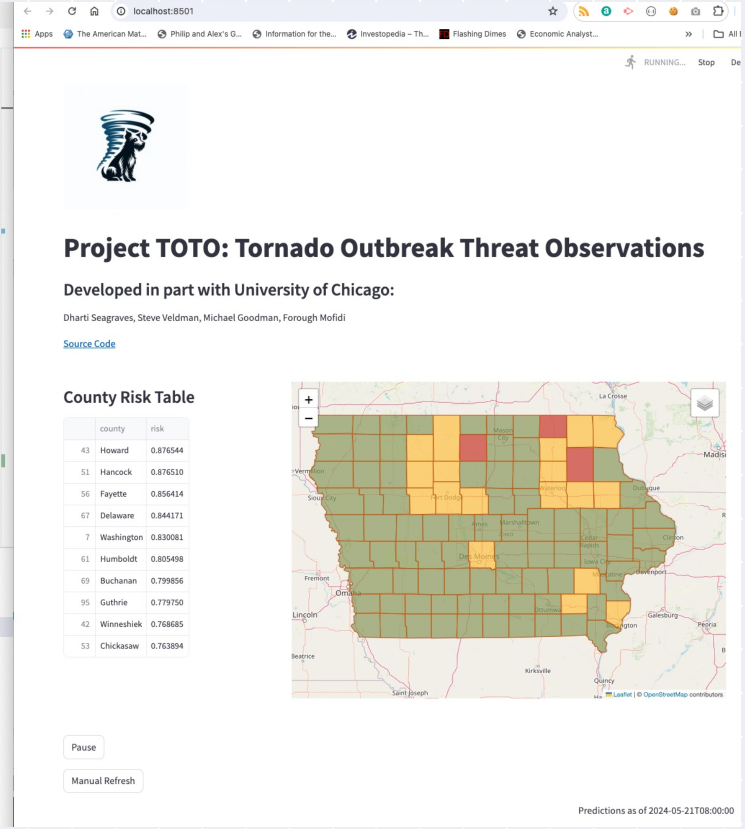
Demo

Streamlit App



```python
import streamlit as st
import pandas as pd
import geopandas as gpd
import folium
from streamlit_folium import folium_static
import os
import time

# Load Iowa county boundaries
script_dir = os.path.dirname(__file__)
geojson_path = os.path.join(script_dir, 'Iowa_County_Boundaries.geojson')
iowa_geo = gpd.read_file(geojson_path)

# Function to load data from the Parquet file
def load_dataframe(parquet_file='tornado_risk.parquet'):
    try:
        df = pd.read_parquet(parquet_file)
        return df
    except FileNotFoundError:
        st.warning('Parquet file not found. Initializing empty DataFrame.')
        return pd.DataFrame(columns=['time', 'county', 'risk'])
    except Exception as e:
        st.error(f'Error loading Parquet file: {e}')
        return pd.DataFrame(columns=['time', 'county', 'risk'])

# Initialize session state for df_st if it doesn't exist
if 'df_st' not in st.session_state:
    st.session_state.df_st = load_dataframe()

# Function to get the max time for the timestamp
def get_max_time(df):
```

```
The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
(base) Dhartis-MacBook-Pro:ProjectToto dhartipatelseagraves$ streamlit run RTS/new_app.py

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.1.14:8501
```



Project TOTO: Tornado Outbreak Threat Observations

Developed in part with University of Chicago:

Dharti Seagraves, Steve Veldman, Michael Goodman, Forough Mofidi

Source Code

County Risk Table

	county	risk
43	Howard	0.876544
51	Hancock	0.876510
56	Fayette	0.856414
67	Delaware	0.844171
7	Washington	0.830081
61	Humboldt	0.805498
69	Buchanan	0.799856
95	Guthrie	0.779750
42	Winneshiek	0.768685
53	Chickasaw	0.763894

Pause

Manual Refresh

Predictions as of 2024-05-21T08:00:00

06

Code + Resources

GitHub Repository: https://github.com/dpatel77/ProjectToto/tree/main

- Code for Server, Client/Server, and Client applications
- Code for data scraping and processing
- Notebook outlining development of predictive ML model

Real Time: Data Streaming and Processing

Data Stream from Server:

- We do not have access to the real-time sensor data from Iowa’s weather monitoring stations

- To simulate a real-time data, historical data representing sensor readings from individual weather stations across the state are streamed – one observation at a time - from server application

Data Receipt and Model Predictions:

- Client/Server receives each new observation and generates an updated prediction for the respective county

- While not strictly “real-time,” this approach functions well within the needs of this use case and allows reliable and efficient updates to model predictions and visualizations.

Real Time: Predictions and Visualization

Tornado Probability by County:

- Current probability of tornado occurrence by county is maintained in parquet format for easy dissemination

- Can be exported as a table/report, or integrated into other complementary applications

Data Visualization:

- Map of Iowa updated in real-time, using color to visualize risk of tornado for each county

- Can be hosted on website or smartphone app, or embedded in live television broadcast