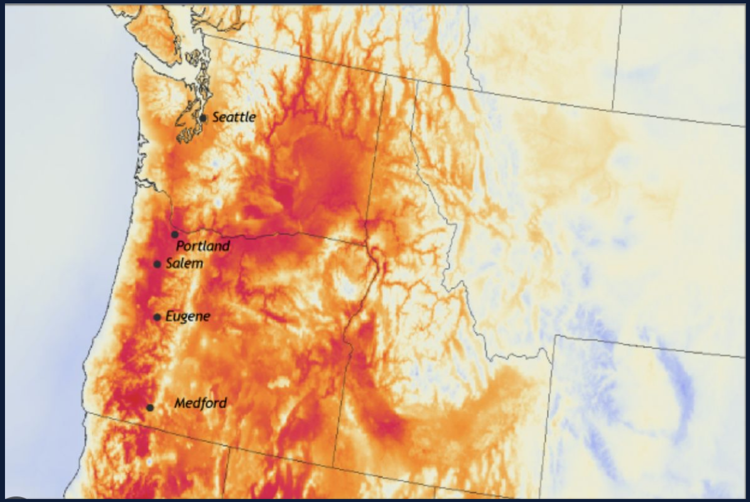
Unraveling the Nexus of Illnesses and Heatwaves: Predictive Modeling for Early Warning Systems

Rigved Goyal, Suchet Sapre, Sai Nikhil Vangala

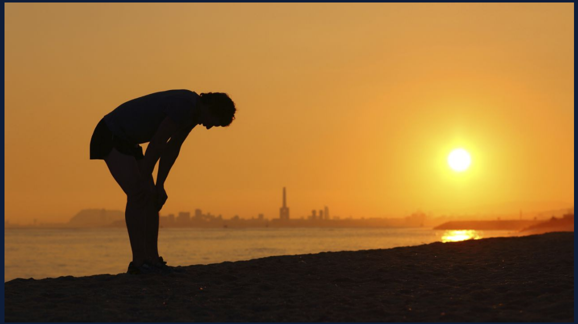
Motivation/Introduction

- In recent years, a surge in heat waves has raised concerns globally, attributed to climate change.
- Despite its mild climate, the Pacific Northwest is not immune to climate change impacts.



Introduction

- Understanding the complex dynamics of temperature and health is a crucial step towards preserving the well-being of our communities
- Our research delves into correlations between temperature variations and heat-related illnesses (HRIs) in this region.
  - A foundation for evidence-based decision-making
  - Implementation of Adaptive Strategies
  - Protection of those most at risk from rising temperatures.

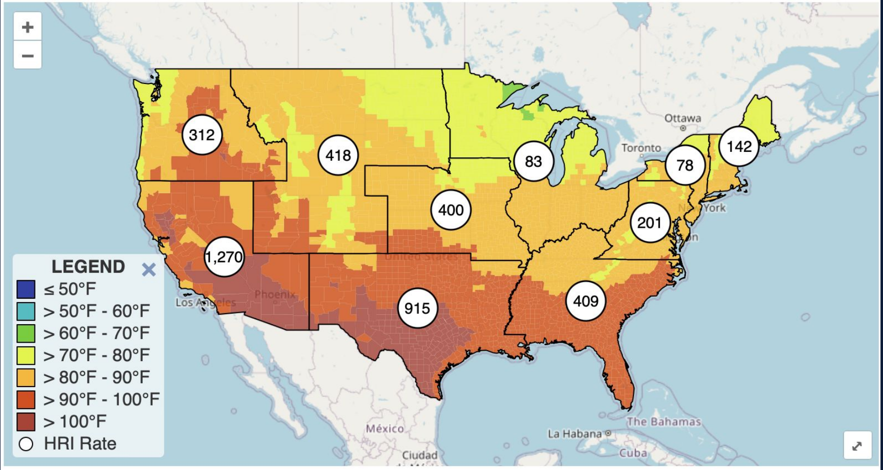


Problem Definition

- Comparing correlations between Temperature and Heat-Related Illnesses (HRIs) in different Regions.
    - Aimed to identify regional variations in the relationship.
- Using Temperature data to predict Heat-Related Illnesses in the Pacific Northwest Region.
    - Analyze historical temperature data and correlate it with instances of HRIs.
    - Models: ARIMA, XGBoost, and LSTM

Data Collection & Dataset

- Weekly heat-related illnesses, county-level temperature data from CDC
- Called CDC-provided APIs, merged and processed data
- Weekly time granularity of data
- 10 regions composed of various continental states



Data Collection & Dataset (contd.)

Regional Temperature/HRI Data				
County	**Avg. Temp.**	**Region**	**Illnesses**	**Week Of**
Autauga, AL	82.3	4	30	10/28/2023
Baldwin, AL	82.6	4	30	10/28/2023
Barbour, AL	81.2	4	30	10/28/2023
Bibb, AL	82.1	4	30	10/28/2023
Blount, AL	79.3	4	30	10/28/2023
Bullock, AL	81.1	4	30	10/28/2023
Butler, AL	82.7	4	30	10/28/2023
Calhoun, AL	78.7	4	30	10/28/2023
Chambers, AL	77.8	4	30	10/28/2023
Cherokee, AL	88.2	4	30	10/28/2023
...	...	...	...	...

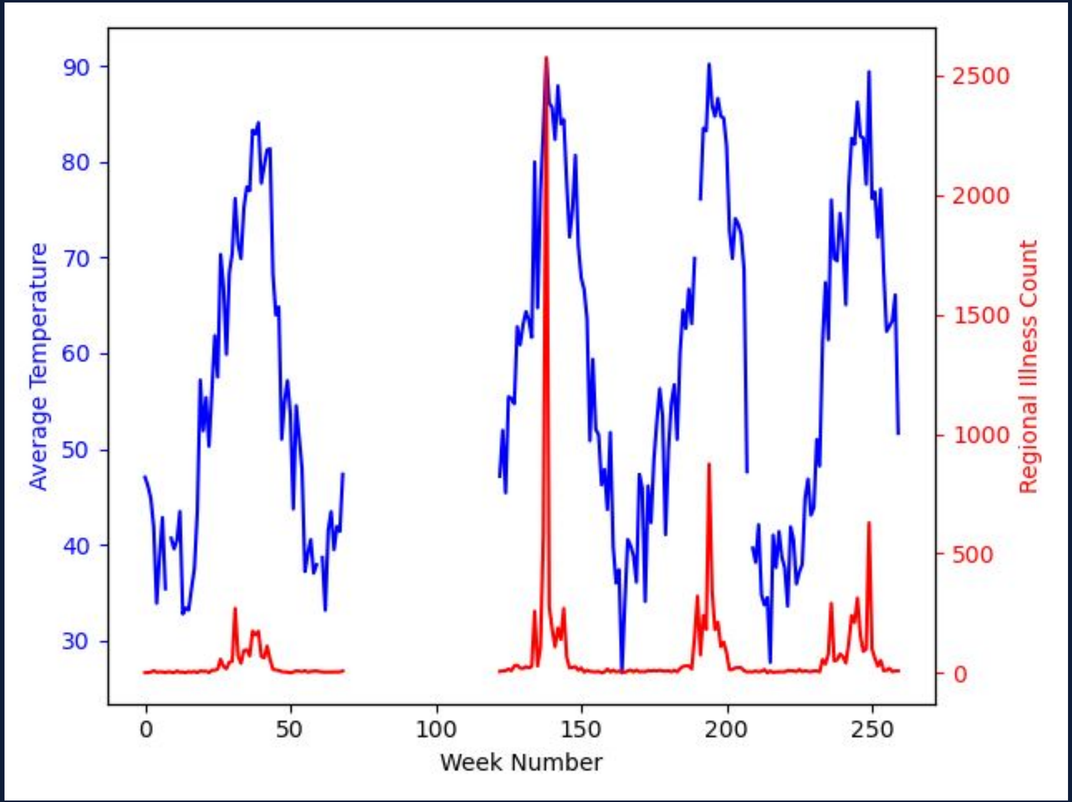
Temperature/HRI data after preprocessing				
Region	**Avg. Temp.**	**Illnesses**	**Week Of**	**Week #**
1	61.855	16	10/28/2023	259
2	64.652	9	10/28/2023	259
3	69.195	13	10/28/2023	259
4	77.844	30	10/28/2023	259
5	64.359	6	10/28/2023	259
6	81.997	16	10/28/2023	259
7	69.820	0	10/28/2023	259
8	51.643	7	10/28/2023	259
9	68.514	18	10/28/2023	259
10	51.643	8	10/28/2023	259
1	58.146	4	10/21/2023	258
2	58.431	5	10/21/2023	258
...	...	...	...	...

Initial Findings

Correlative Task

PCC and temperature values for each region			
Region	**Avg. Temperature**	**PCC**	**Avg. HRIs**
1	57.281056	0.603907	39.14
2	59.710965	0.596408	28.80
3	65.553054	0.677653	61.67
4	73.732904	0.761961	93.74
5	59.148623	0.591785	44.79
6	76.285816	0.713821	135.23
7	64.171449	0.639717	116.47
8	56.544747	0.703789	57.55
9	70.703420	0.742930	122.34
10	58.070153	0.424426	58.41

Focus: Region 10



Proposed Approaches

ARIMA

- Based on literature review, ARIMA family is one of the most common groups of classical time series models
- Two main components: autoregressive (AR) & moving average (MA)
  - AR: Computes linear combination of past values
  - MA: Computes linear combination of past forecasting errors
- Parameterized by three variables: p, d, and q
  - p: number of past values included in AR component
  - d: number of differences of time series data to be taken (ensures that data is stationary)
  - q: number of past forecasting errors included in MA component
- Designed for single-view time-series data → can't include temperature data

XGBoost

- Based on literature review, XGBoost was chosen due to its efficacy through machine learning and commonly used in time-series classification tasks.
- Similar to the Random Forest model discussed by many of the sources, but final classification is made on collection of weak learners.
- Leverages gradient boosting framework for consistently better results.
- Referred to xgboost python library to implement the classifier.
- Used grid search to find best hyperparameters:
  - Number of estimators
  - Learning Rate
  - Max Depth
- Designed for multi-view time-series data (Avg. Temperature & Regional Illnesses)

LSTM

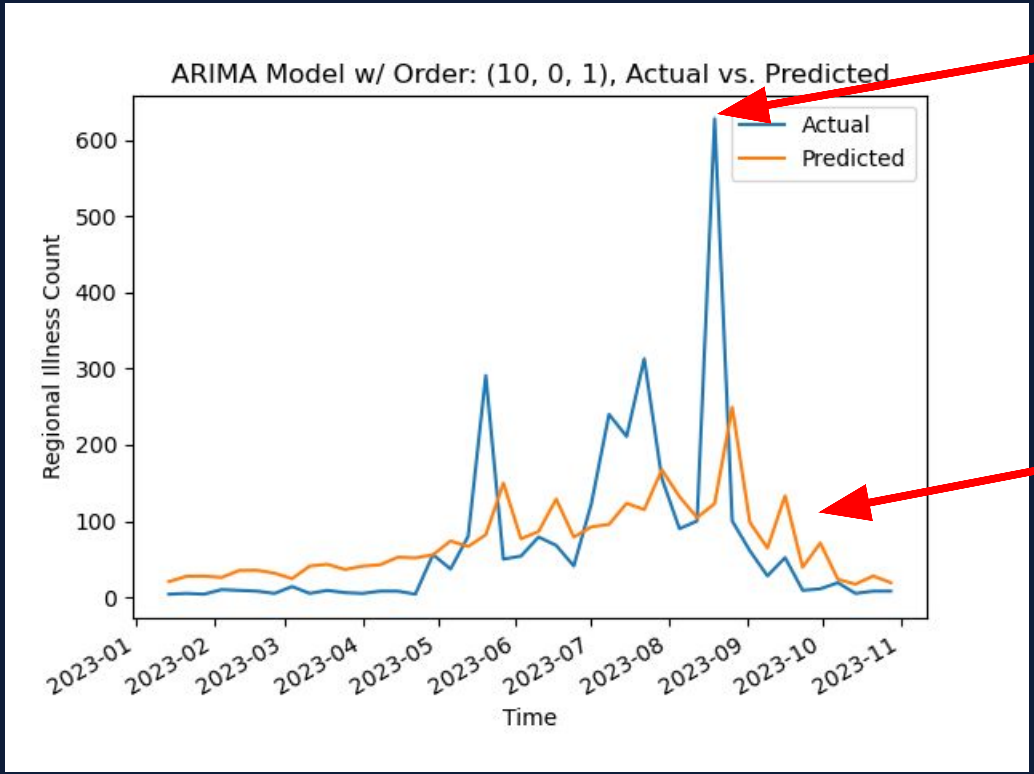
- Based on literature review, LSTM is ideal for working with time-series data and predictions
- Uses historic, sequential time-series data as a feature to make target predictions at any given time-step
- Implemented with Tensorflow using LSTM layer
    - Used grid search to optimize:
        - Units
        - Dropout rate
        - Learning rate
        - Activation

Experiments
&
Results

ARIMA/XGBoost Model - Overview

- Served as our baseline models of comparison
- Focused on Region 10
- Used 80/20 train/test split
- Forecast window of size 1 (forecasted 1 week into the future)
- Implemented rolling-forecast approach

ARIMA Model - Plot & RMSE

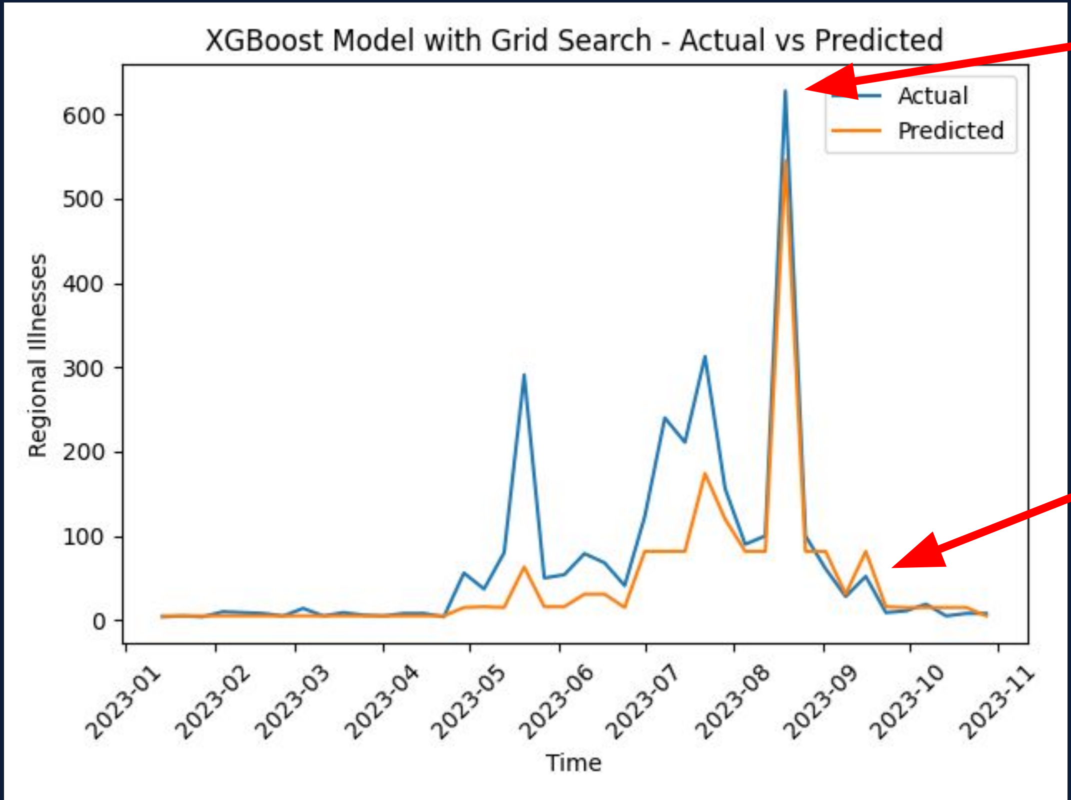


Unable to account for large deviations in regional illness count

Able to model general (seasonal) trends

RMSE = 102.0521

XGBoost Model - Plot & RMSE

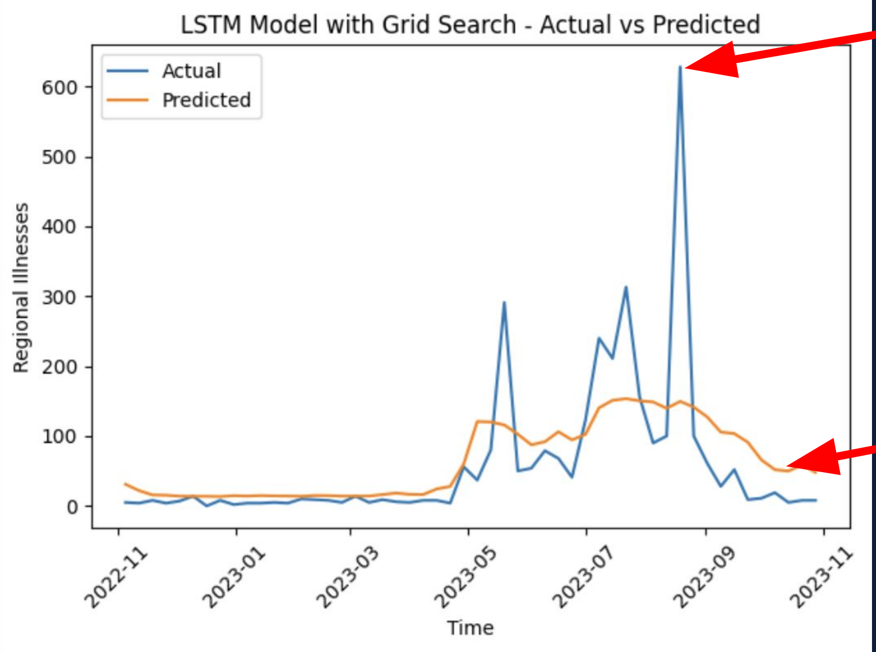


Able to account for large deviations in regional illness count

Able to model general (seasonal) trends

RMSE = 57.46

LSTM Model - Plot & RMSE



Not ideal for predicting on sudden outlier data

Able to model general (seasonal) trends

RMSE = 87.387

Conclusions & Future Work

- Task 1: Comparing Correlations Between Temperature and Heat-Related Illnesses (HRIs) in Different Regions
  - Region 4 strongest correlation & Region 10 weakest correlation
- Task 2: Using temperature data to predict HRIs in the Pacific Northwest region
  - **RMSEs** → ARIMA: 102.0521, XGBoost: 57.46, LSTM: 87.387
- Future Work
  - Exploring transformer-based models and ensemble models
  - Increasing forecasting window (currently 1)
  - Incorporating more weather-related data for predictions

Thank You!