Forecasting PFAS Data

MSAI Final Project

CS6463 AI Practicum

Kyle Bailey

Introduction

What are PFAS?

- PFAS (Per- and Polyfluoroalkyl Substances) are harmful man-made chemicals that are highly persistent in the environment and human body.

Why Forecast?

- Forecasting future PFAS levels can help authorities plan for future contamination risks, implement protective measures, and inform the public about potential exposure hazards.

Goals and Objectives

- The goal of this project is to train a model to forecast future levels of PFAS contamination in different regions using historical PFAS data.
- The objective is to explore how well models can predict future growth or changes in PFAS levels based on historical trends.

The UCMR Dataset

The Dataset

- The UCMR dataset used contained water samples collected from across the US from 2001 to 2024 that included:
  - Public Water System (PWS) information
  - Reading levels for various contaminants
  - Water system type e.g groundwater, surface water
  - Geolocation data for each public water system.
  - Additional data included water treatment info from each PWS

Challenges

- Irregular time series - non-constant spacing of observations times, missing data, relations between observations may change over time
- Left-censored data - measurements may fall below a detection limit so you only know that the values is than a certain threshold rather than the exact value.
- Sparse data for PFAS contamination – while there were many samples collected that measured dozens of contaminants, there are relatively few samples of PFAS contamination

Methodology

- Focus on UCMR 5 PFAS data
  - UCMR 3 only other UCMR data with PFAS readings, but occurred 8 years prior to UCMR 5

Data Preprocessing

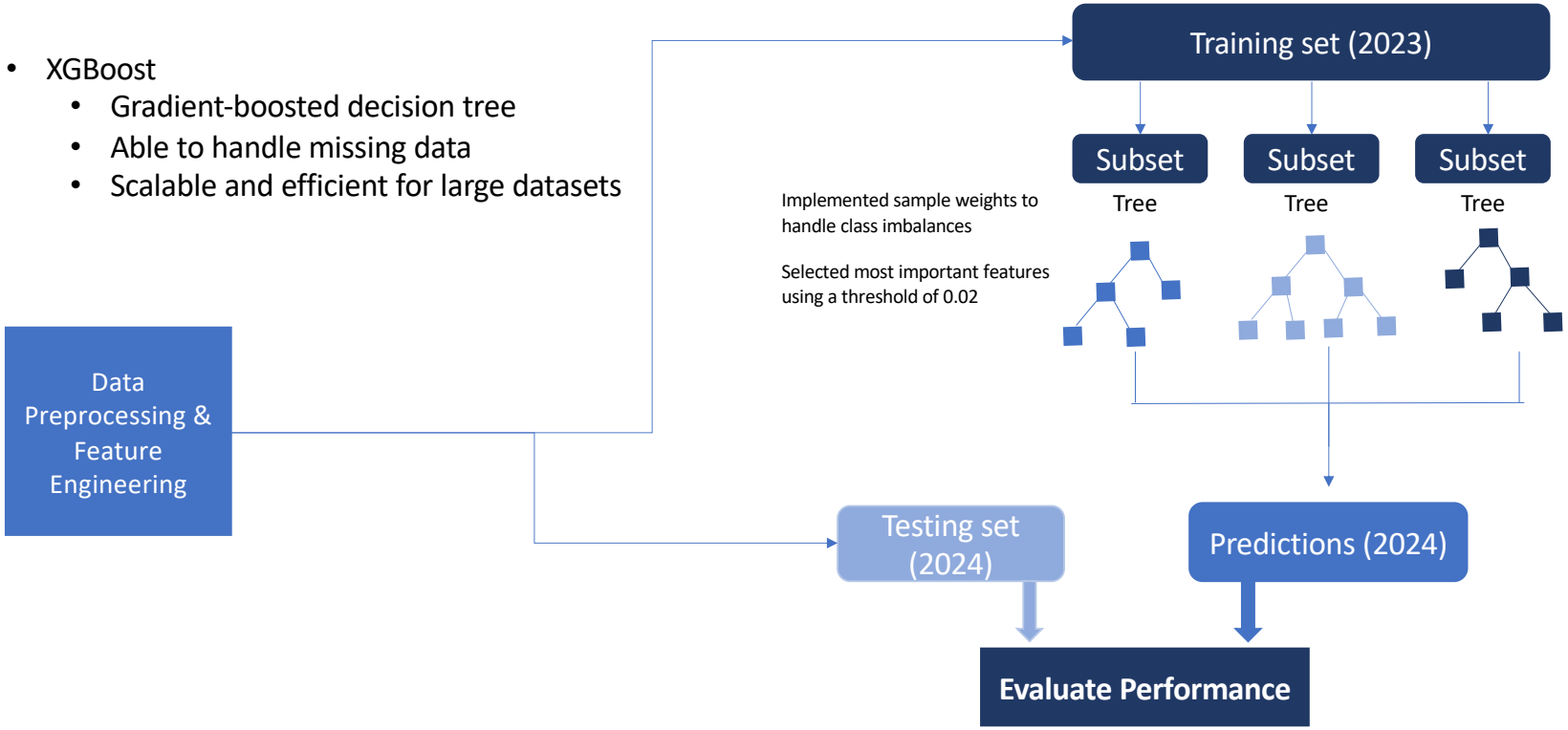
- Exclude Lithium readings since does not correlate to PFAS
- Readings under MRL set to 0
- Added missing records through forward/backward filling methods
- Used log transform of readings to address skewness

Feature Engineering

- Created temporal features (month, season)
- Added lagged features (prior analytical result values)
- Performed one-hot encoding for categorical variables
- Split data into training (2023) and testing (2024) sets

Methodology

- XGBoost
  - Gradient-boosted decision tree
  - Able to handle missing data
  - Scalable and efficient for large datasets



Evaluation Metrics

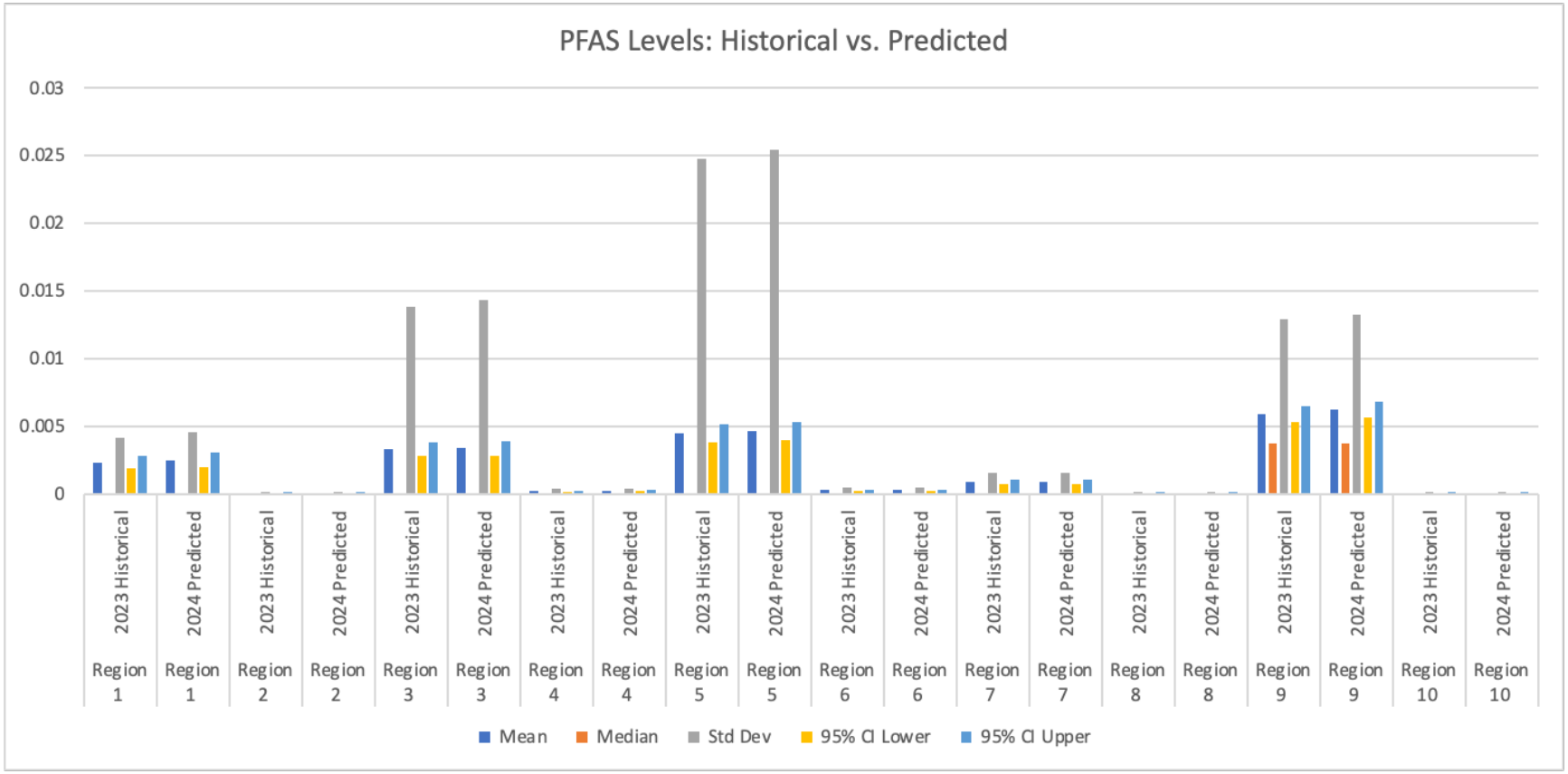
Training Set Performance:

- MSE: 3.955e-07
- RMSE: 6.289e-04
- MAE: 2.035e-05
- R2: 0.940 (94.0%)
- Pearson R: 0.972

Testing Set Performance:

- MSE: 4.086e-07
- RMSE: 6.393e-04
- MAE: 2.030e-05
- R2: 0.934 (93.4%)
- Pearson R: 0.968

- The model shows strong performance with high R2 values on both training and test sets, indicating good predictive capability.
- The similar performance metrics between training and test sets suggest the model is not overfitting.


PFAS Levels: Historical vs. Predicted
0.03
0.025
0.02
0.015
0.01
0.005
0
2023 Historical
2024 Predicted
Region 1
Region 2
Region 3
Region 4
Region 5
Region 6
Region 7
Region 8
Region 9
Region 10
Mean
Median
Std Dev
95% CI Lower
95% CI Upper

Discussion – Interpretation of results

- Highest Predicted PFAS Levels:
  - Region 9: Shows highest predicted concentrations
    - Driven by high levels in Guam and Northern Mariana Islands
    - Consistent pattern across prediction period
  - Region 5: Second highest predicted levels
    - Minnesota contributing significantly to regional levels
    - Shows more variance across different states within region
  - Region 3: Third highest predicted levels
    - Delaware showing elevated concentrations
    - More uniform distribution across states

- Lowest Predicted PFAS Levels:
  - Region 8: Consistently lower predictions
  - Region 10: Lower than average predictions
  - Region 2: Relatively stable, lower concentrations

Conclusion

Limitations

- Public Water Sources' (PWS) coordinates hard to extract from EPA geographical data.
  - Including PWSID in data would allow easily joining to PWS coordinates
  - Benefit would be to be able to identify potential PFAS issues with shared water sources between PWS's
- More data collection can improve the model's accuracy

Further Research

- Examining if there are correlations between other non PFAS reading levels and PFAS results
- Predicting PFAS levels for more granular locations, i.e. by city or by county
- Adjusting the model with future data as more PFAS readings are measured

Questions?

Region-wise Statistical Summary

Region	Time Period	Mean	Median	Std Dev	95% CI Lower	95% CI Upper
Region 1	Historical	0.00234	0.00003	0.00412	0.00189	0.00279
	Predicted	0.0025	0.00003	0.00456	0.00198	0.00302
Region 2	Historical	0.00007	0.00003	0.00012	0.00005	0.00009
	Predicted	0.00007	0.00003	0.00012	0.00005	0.00009
Region 3	Historical	0.00328	0.00003	0.01385	0.00278	0.00378
	Predicted	0.00336	0.00003	0.01432	0.00282	0.0039
Region 4	Historical	0.00021	0.00003	0.00038	0.00017	0.00025
	Predicted	0.00022	0.00003	0.00039	0.00018	0.00026
Region 5	Historical	0.00448	0.00003	0.02478	0.00383	0.00513
	Predicted	0.00464	0.00003	0.0254	0.00395	0.00533
Region 6	Historical	0.00026	0.00003	0.00045	0.00021	0.00031
	Predicted	0.00026	0.00003	0.00045	0.00021	0.00031
Region 7	Historical	0.00089	0.00003	0.00156	0.00074	0.00104
	Predicted	0.00089	0.00003	0.00156	0.00074	0.00104
Region 8	Historical	0.00007	0.00003	0.00012	0.00005	0.00009
	Predicted	0.00007	0.00003	0.00012	0.00005	0.00009
Region 9	Historical	0.00589	0.00369	0.01289	0.00534	0.00644
	Predicted	0.00621	0.00369	0.01323	0.00563	0.00679
Region 10	Historical	0.00007	0.00003	0.00012	0.00005	0.00009
	Predicted	0.00007	0.00003	0.00012	0.00005	0.00009