Mesonet Summer Internship Conclusive Report

Cy Dixon

August 15, 2025

Introduction

Machine Learning Applied to Thermal Inversion Forecasting

During this summer internship, our main goal was to forecast and predict surface-level climate attributes within the state of Kentucky using the *Micro Macro* (MiMa) machine learning model. One of our first assigned tasks was to gap-fill station data to ensure it could be used effectively in modeling. After this, we would need to run some fearture significance tests on the data to figure out what attributes of these datasets would be the best predictor for our specific target. To do so we had to delve into feature engineering. Finnaly we were tasked with running different forecasting methods using this data to predict future points.

Methodology

Kentucky Mesonet and Data Description

The primary data source for this project was the Kentucky Mesonet, focusing on two stations: `CRRL` and `RFSM`. These stations recorded a wide range of attributes, including:

> `NetSiteAbbrev, County, UTCTimestampCollected, TAIR, DWPT, PRCP, PRES, RELH, SRAD, WDIR, WSPD, WDSD, WSSD, SM02, SM04, ST02, ST04, VT05, VT20, VT90, VR05, VR20, VR90`.

Gap Filling Methods

To prepare the datasets for modeling, we addressed missing data using several gap-filling techniques:

- Linear interpolation
- Circular interpolation
- Forward/backward fill
- Random forest imputation
- Gradient boosting imputation

The method applied depended on both the size of the gap and the specific variable being analyzed. This ensured that each dataset was consistent and ready for feature analysis.

Feature Significance Methods

Feature significance was determined using Random Forest and SHAP (SHapley Additive exPlanations) techniques. The goal was to identify which features in the two-station datasets were the strongest predictors for:

1. TAIR – VT20
2. VT90 – VT20

The results of this analysis for both CRRL and RFSM stations are shown in Figure 1.

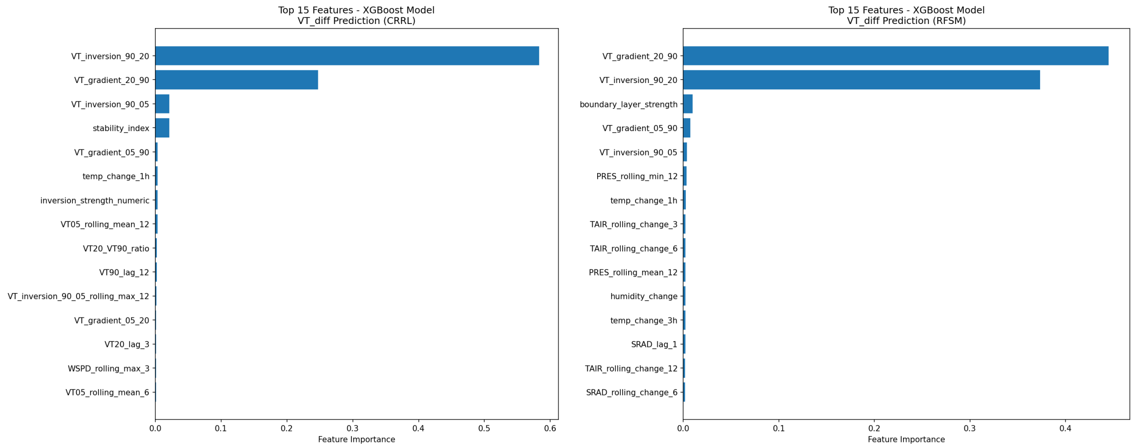

Figure 1: Feature importance analysis for CRRL and RFSM stations (VT differences).

Analysis: According to the figures some of the major predictors of the VT differences are those such as the VT gradient and the 1 hour tempature change. These are both features that were engineered from the original dataset. This appears to be the case for both CRRl and RFSM.

Forecasting Methods

Several forecasting approaches were explored:

1. Decision Trees / Random Forest
2. Boosting Methods
3. Deep Learning (Neural Networks): LSTM / Micro Macro

Prediction results for VT20, VT90, and VT differences at CRRL are shown in Figure 2, while RFSM predictions are shown in Figure 3.

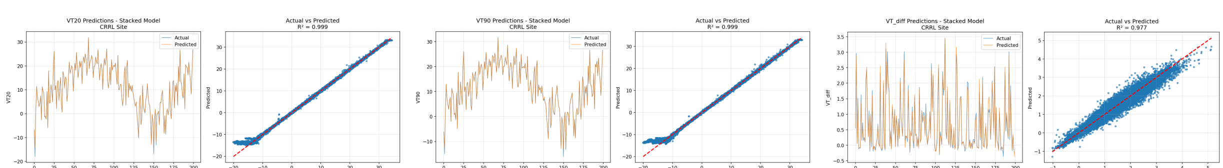

Figure 2: Predictions for VT20, VT90, and VT differences at CRRL station.

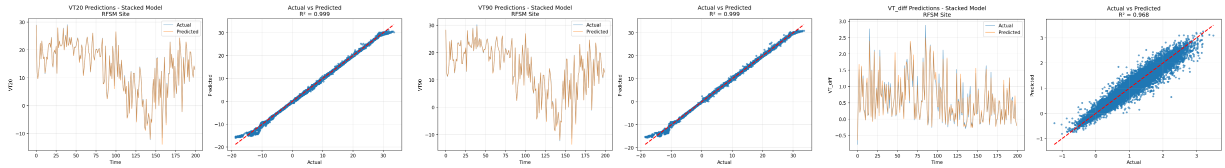

Figure 3: Predictions for VT20, VT90, and VT differences at RFSM station.

Results

Summary of Findings

We find that in general the forecasting methods seemed to work relativly well. Unfortunately I was unable to get around to working very in depth within the GPU intensive server to use the MiMa model, but the prepared data would lead me to the unsumption that we could get the data into and receive great results.

Discussion and Conclusion

Scientific Meaning of Results

The feature importance plots (Figure 1) highlight which climate variables are most influential in predicting VT differences. The prediction results for CRRL (Figure 2) and RFSM (Figure 3) indicate model accuracy and the ability to capture inversion patterns. Finally, TAIR comparison plots (Figure **??**) demonstrate how well the model matches observed surface temperatures.

Personal Takeaways

This internship with the Mesonet allowed me to go deeper into forecasting and machine learning with surface level meterological data and has provided me with excellent experiance with machine learning with datasets similar. If I were to have some more time with the project I would have spun up the MiMa model within the server and tried to get an instance of that working with my engineered data. Additionally, all codes, figures and any other data from this internship can be found on my github page linked here: github.com/cylissama/mesonet-machine-learning.