Machine Learning for Numerical Weather Prediction

Ali Cem Çakmak, Diego Garces, Deepak Sorout, Muhammad Fakhar, Yuqi Fang

Linear Regression

Linear Regression

Dataset

- **27,513 days** of historical weather measurements (over 75 years)
- **Features**: Daily minimum temperature, maximum temperature, and mean temperature
- **Data Split**:
  - 80% Training data - 10% Validation data -10% Test data

Models Compared

1. **Simple Linear Regression**
   - Uses only yesterday's mean temperature
2. **Multiple Linear Regression**
   - Uses yesterday's minimum, maximum, and mean temperatures

Model Equations & Coefficients

Simple Linear Regression Model

mean_temp = 0.62 + 0.94 × mean_temp_lag1

- **Intercept**: 0.62 °C , **Coefficient**: 0.94
- **Interpretation**: Tomorrow's mean temperature is approximately 94% of yesterday's mean temperature plus 0.6 °C

Multiple Linear Regression Model

mean_temp = 0.07 - 0.12 × min_temp_lag1 + 0.07 × max_temp_lag1 + 0.96 × mean_temp_lag1

- **Intercept**: 0.07 °C
- **Minimum temperature coefficient**: -0.12 (negative due to high correlation)
- **Maximum temperature coefficient**: 0.07
- **Mean temperature coefficient**: 0.96 (dominant predictor)

Key Insight: Yesterday's mean temperature carries the majority of predictive power in both models

Performance Interpretation

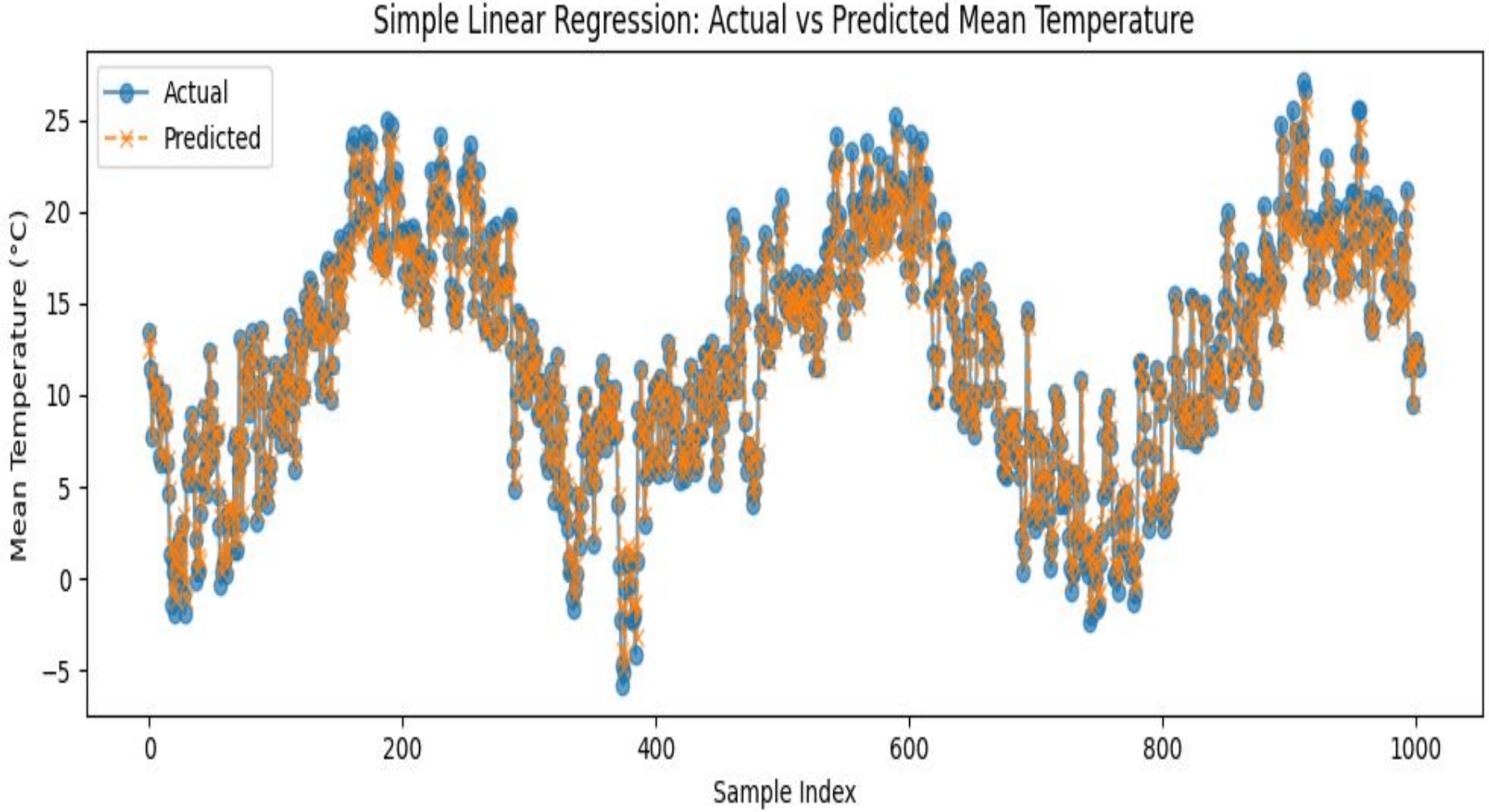
- **R-squared = 0.885**: The model explains 88.5% of the variance in daily mean temperature
- **Mean Absolute Error = 1.72 °C**: On average, predictions are within ±1.7 °C of the actual value
- Performance is strong for short-term, day-ahead temperature forecasting

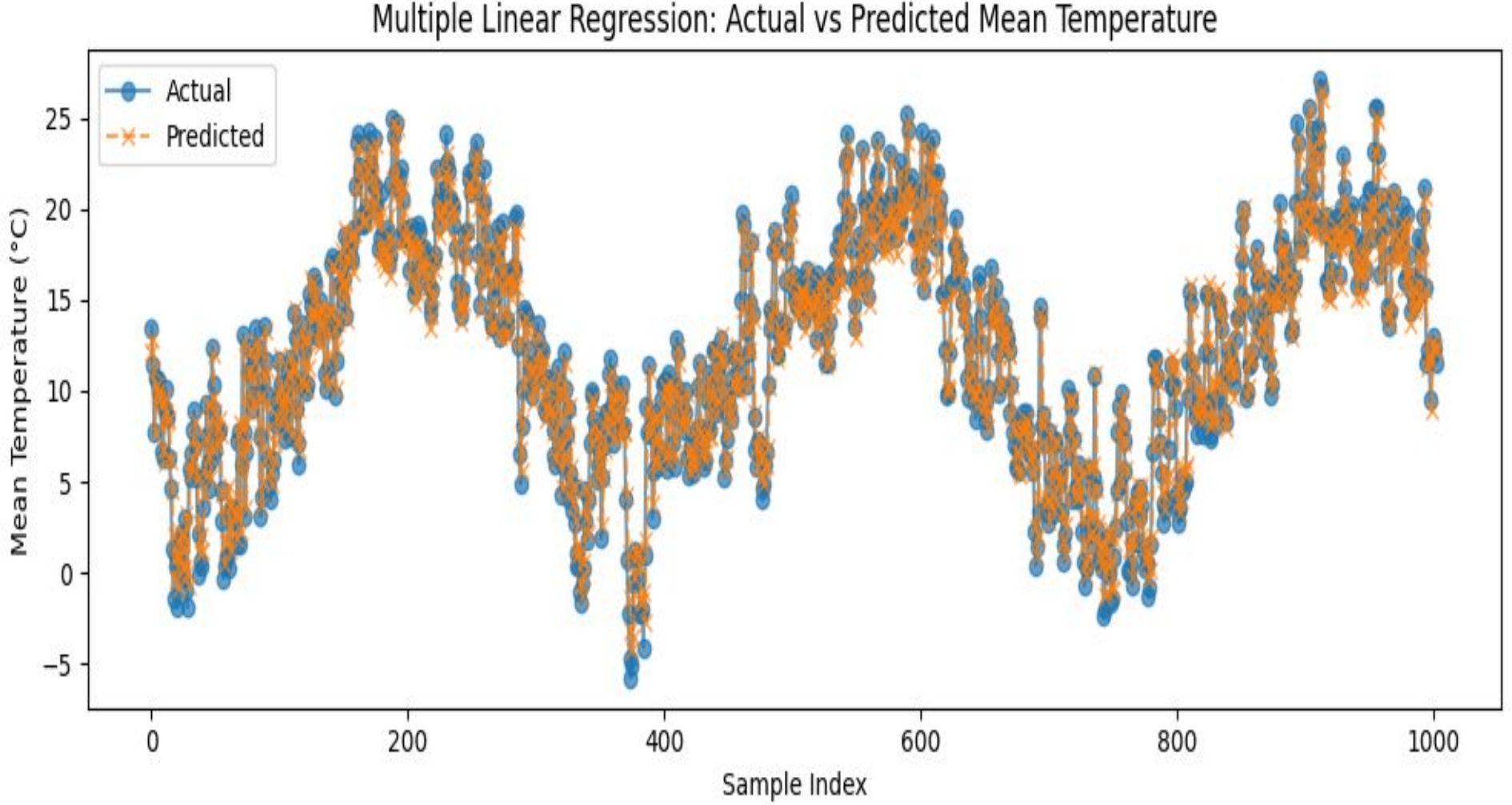
Model Comparison

- The simple model performs very well using only one feature
- Adding minimum and maximum temperatures results in only marginal improvement
- Indicates strong correlation between minimum, maximum, and mean temperatures

Results

Metric	Simple Model	Multiple Model	Improvement
Mean Absolute Error	1.76 °C	1.72 °C	2.3% decrease
Root Mean Squared Error	2.26 °C	2.22 °C	1.8% decrease
R-squared	0.880	0.885	0.6% increase


Simple Linear Regression: Actual vs Predicted Mean Temperature
Actual
Predicted
Mean Temperature (°C)
25
20
15
10
5
0
−5
0
200
400
600
800
1000
Sample Index


Multiple Linear Regression: Actual vs Predicted Mean Temperature
Actual
Predicted
Mean Temperature (°C)
25
20
15
10
5
0
−5
0
200
400
600
800
1000
Sample Index

Insights & Future Work

Key Findings

- **Strong temporal dependency**: Yesterday's temperature is highly predictive of tomorrow's temperature
- **Diminishing returns from additional features**: Highly correlated inputs add limited new information
- **Linear relationship is effective**: A simple linear model captures most of the temperature dynamics

Limitations

- Uses only a one-day lookback window
- Assumes linear relationships between variables
- Does not include other meteorological factors such as humidity, pressure, wind, or seasonality

Future Improvements

Add temporal features such as last 7-day data instead of just yesterday.

Neural Network

Random Forest

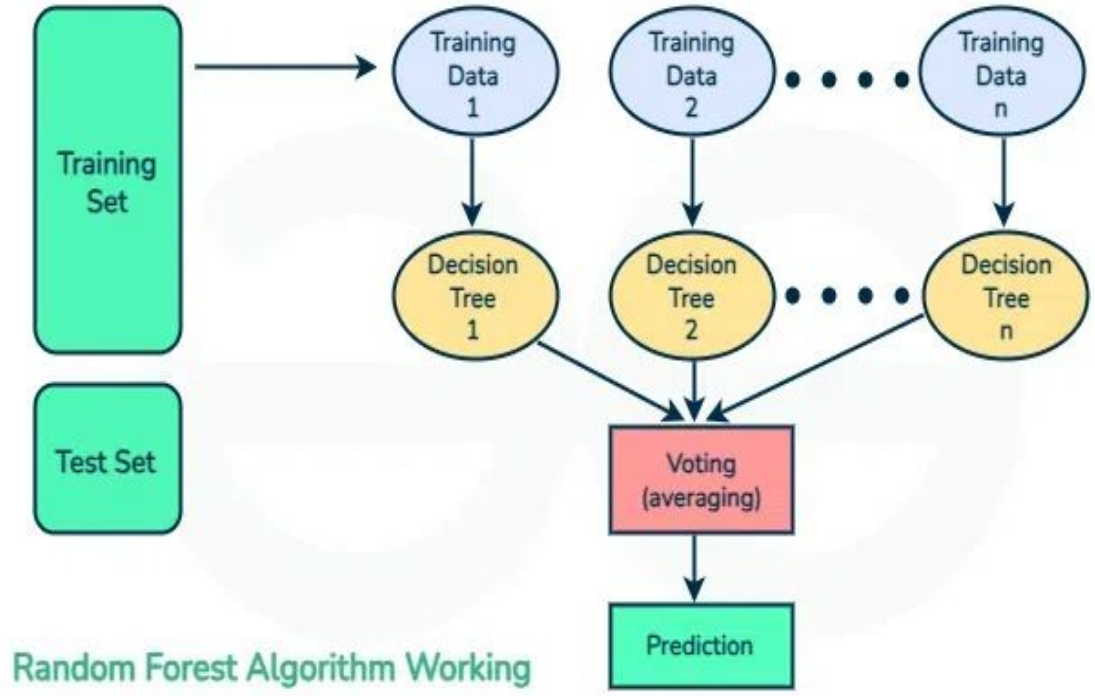
Quick Recap - What is Random Forest?

Core Concept: "The Wisdom of the Crowds"

Process:

- **Many Trees:** We build 100 independent decision trees.
- **Randomness:** Each tree sees a random subset of data & features.
- **Averaging:** Final Prediction = Average of all 100 trees.

Why: Reduces overfitting and variance.



Source: GeeksforGeeks

Method: Autoregression & Lag Features (Past 15 days -> Predict Today)

Original data:

DATE	TN	TX	RR	SS	HU	FG	FX	CC	SD	TG
1957-10-01	19.0	134.0	0.0	49.0	81.0	18.0	80.0	3.0	0.0	56.0
1957-10-02	14.0	123.0	0.0	30.0	87.0	15.0	42.0	5.0	0.0	57.0
1957-10-03	9.0	115.0	3.0	0.0	89.0	27.0	109.0	6.0	0.0	83.0
1957-10-04	23.0	117.0	0.0	56.0	78.0	21.0	86.0	5.0	0.0	58.0
1957-10-05	-8.0	140.0	9.0	14.0	79.0	24.0	66.0	6.0	0.0	93.0
1957-10-06	92.0	140.0	0.0	0.0	95.0	12.0	41.0	5.0	0.0	108.0
1957-10-07	75.0	173.0	0.0	76.0	86.0	28.0	79.0	2.0	0.0	106.0
1957-10-08	24.0	205.0	0.0	92.0	81.0	16.0	39.0	0.0	0.0	105.0

Method: Autoregression & Lag Features (Past 15 days -> Predict Today)

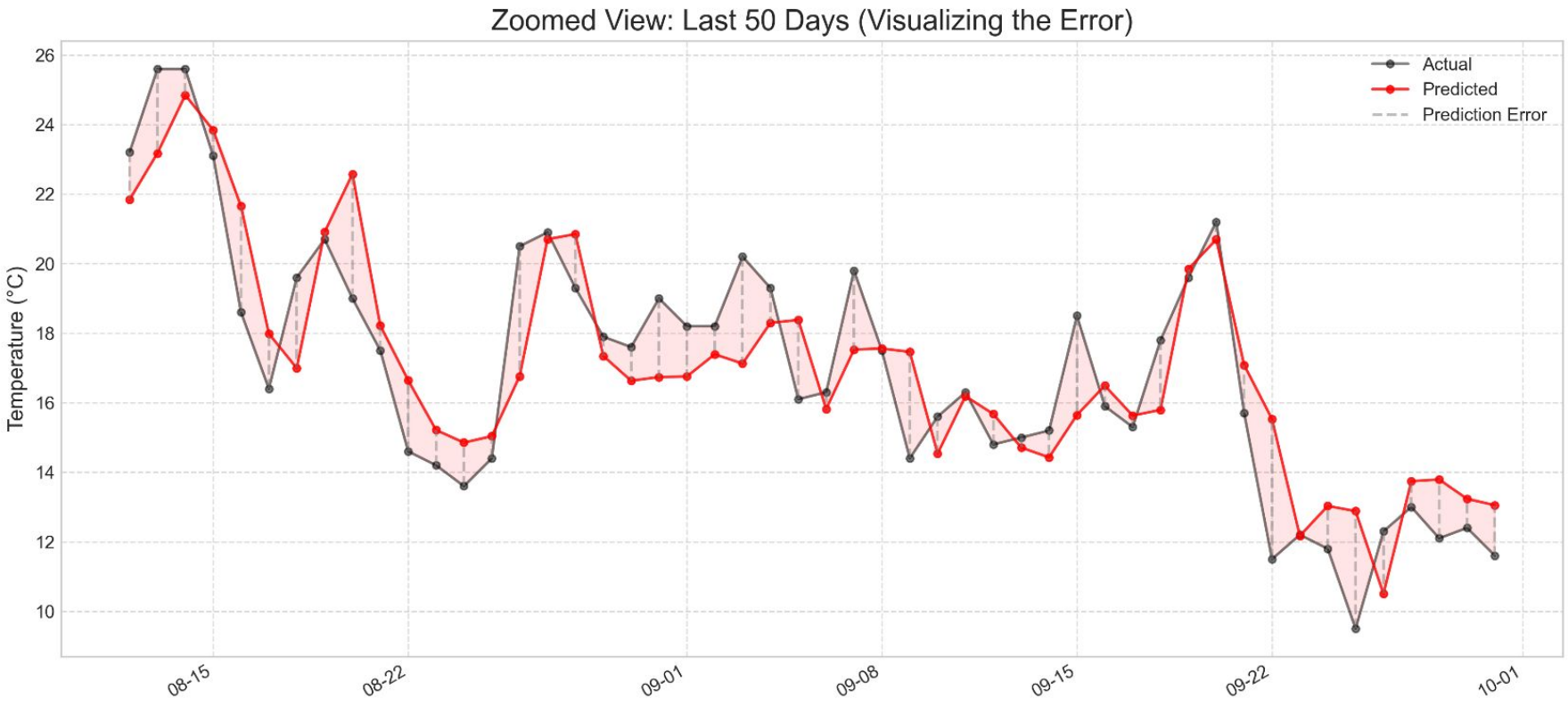
Processed data:

DATE	TG_lag_1	TG_lag_2	TG_lag_3	TG_lag_4	...	TG_lag_15	Other Vars		Target_TG (Today)
1957-10-16	108.0	102.0	97.0	106.0	...	56.0	...	==>	139.0
1957-10-17	139.0	108.0	102.0	97.0	...	57.0	...	==>	139.0
1957-10-18	139.0	139.0	108.0	102.0	...	83.0	...	==>	174.0
1957-10-19	174.0	139.0	139.0	108.0	...	58.0	...	==>	105.0
1957-10-20	105.0	174.0	139.0	139.0	...	93.0	...	==>	77.0
1957-10-21	77.0	105.0	174.0	139.0	...	108.0	...	==>	100.0
1957-10-22	100.0	77.0	105.0	174.0	...	106.0	...	==>	77.0
1957-10-23	77.0	100.0	77.0	105.0	...	105.0	...	==>	84.0

Total features: 10*15 = 150 columns

Preliminary Results & Visualizations

MAE: 1.58 °C RMSE: 2.06 °C R^2 : 0.902



Feature Importance & Next Steps

Top 5 Most Important Features:

TG_lag_1	0.8814
RR_lag_1	0.0059
FX_lag_1	0.0040
TX_lag_1	0.0037
FG_lag_1	0.0029

Possible Next Steps:

- **Integrate Validation Set:**
  Train -> Test to Train -> Val -> Test.
- **Optimize Model Architecture:**
  Tune max_depth, n_estimators, and lag_window (e.g. 15 days vs. 7 days)
- **Feature Selection:**
  Reduce 150 features to ~Top 20 to remove noise.