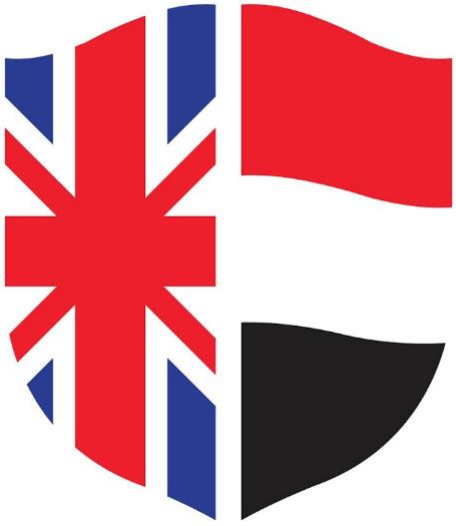


24CSAI03H
Machine Learning

Project 1 Report

Temperature Prediction for The City of Seoul

Sameh218767 Merihan226392

Overview:

This dataset focuses on predicting next-day maximum and minimum air temperatures using data from the LDAPS (Local Data Assimilation and Prediction System) model operated by the Korea Meteorological Administration over Seoul, South Korea. This data consists of summer data from 2013 to 2017. The input data is largely composed of the LDAPS model's next-day forecast data, in-situ maximum and minimum temperatures of present-day, and geographic auxiliary variables. There are two outputs (i.e. next-day maximum and minimum air temperatures) in this data. The primary objective is to improve bias correction in temperature predictions by utilizing historical weather data alongside model forecasts. This is crucial for applications in agriculture, energy management, and public safety, where accurate temperature predictions can significantly impact decision-making. Hindcast validation was conducted for the period from 2015 to 2017.

Source: https://www.kaggle.com/datasets/smokingkrils/temperature-forecast-project-using-ml

Features:

1. **station**: Weather station number (1 to 25).
2. **Date**: Present day in yyyy-mm-dd format ('2013-06-30' to '2017-08-30').
3. **Present_Tmax**: Maximum air temperature between 0 and 21 h on the present day (°C), ranging from 20 to 37.6.
4. **Present_Tmin**: Minimum air temperature between 0 and 21 h on the present day (°C), ranging from 11.3 to 29.9.
5. **LDAPS_RHmin**: LDAPS model forecast of next-day minimum relative humidity (%), ranging from 19.8 to 98.5.
6. **LDAPS_RHmax**: LDAPS model forecast of next-day maximum relative humidity (%), ranging from 58.9 to 100.
7. **LDAPS_Tmax_lapse**: LDAPS model forecast of next-day maximum air temperature with applied lapse rate (°C), ranging from 17.6 to 38.5.
8. **LDAPS_Tmin_lapse**: LDAPS model forecast of next-day minimum air temperature with applied lapse rate (°C), ranging from 14.3 to 29.6.

9. LDAPS_WS: LDAPS model forecast of next-day average wind speed (m/s), ranging from 2.9 to 21.9.
10. LDAPS_LH: LDAPS model forecast of next-day average latent heat flux (W/m²), ranging from -13.6 to 213.4.
11. LDAPS_CC1: LDAPS model forecast of next-day average cloud cover for the 1st 6-hour period (0-5 h) (%), ranging from 0 to 0.97.
12. LDAPS_CC2: LDAPS model forecast of next-day average cloud cover for the 2nd 6-hour period (6-11 h) (%), ranging from 0 to 0.97.
13. LDAPS_CC3: LDAPS model forecast of next-day average cloud cover for the 3rd 6-hour period (12-17 h) (%), ranging from 0 to 0.98.
14. LDAPS_CC4: LDAPS model forecast of next-day average cloud cover for the 4th 6-hour period (18-23 h) (%), ranging from 0 to 0.97.
15. LDAPS_PPT1: LDAPS model forecast of next-day precipitation for the 1st 6-hour period (0-5 h) (%), ranging from 0 to 23.7.
16. LDAPS_PPT2: LDAPS model forecast of next-day precipitation for the 2nd 6-hour period (6-11 h) (%), ranging from 0 to 21.6.
17. LDAPS_PPT3: LDAPS model forecast of next-day precipitation for the 3rd 6-hour period (12-17 h) (%), ranging from 0 to 15.8.
18. LDAPS_PPT4: LDAPS model forecast of next-day precipitation for the 4th 6-hour period (18-23 h) (%), ranging from 0 to 16.7.
19. lat: Latitude (°), ranging from 37.456 to 37.645.
20. lon: Longitude (°), ranging from 126.826 to 127.135.
21. DEM: Elevation (m), ranging from 12.4 to 212.3.
22. Slope: Slope (°), ranging from 0.1 to 5.2.
23. Solar radiation: Daily incoming solar radiation (Wh/m²), ranging from 4329.5 to 5992.9.

Target Variables:

1. Next_Tmax: Forecast of the next-day maximum air temperature (°C), ranging from 17.4 to 38.9.
2. Next_Tmin: Forecast of the next-day minimum air temperature (°C), ranging from 11.3 to 29.8.

Data Exploration

- Numerical Features:

Weather-related features: `Present_Tmax`, `Present_Tmin`, `LDAPS_RHmin`, `LDAPS_RHmax`, etc.

Geographical and environmental features: lat, lon, DEM (elevation), Slope, and Solar radiation.

- Categorical Feature:

Date: Represents the date in a non-standard format (`dd-mm-yyyy`) and will be converted to a time stamp.

- Missing Values:

Some columns have missing values, `Present_Tmax`, `Present_Tmin`, `LDAPS_RHmin`, and other LDAPS features, 75 missing values each but still not a large percentage.

- Outliers:

Potential outliers might be present in temperature and humidity values based on their distributions. Identifying and handling these outliers will be crucial.

- Dimensionality: The dataset has 7,752 entries and 25 columns.

Data Cleaning

1. **Null percentage:**

Feature	Missing Values	Percentage
LDAPS_CC1	75	0.97
LDAPS_PPT4	75	0.97
LDAPS_PPT2	75	0.97
LDAPS_PPT1	75	0.97
LDAPS_CC4	75	0.97
LDAPS_CC3	75	0.97
LDAPS_CC2	75	0.97
LDAPS_LH	75	0.97
LDAPS_WS	75	0.97
LDAPS_Tmin_lapse	75	0.97
LDAPS_Tmax_lapse	75	0.97
LDAPS_RHmax	75	0.97
LDAPS_RHmin	75	0.97
LDAPS_PPT3	75	0.97
Present_Tmin	70	0.90
Present_Tmax	70	0.90
Next_Tmax	27	0.35
Next_Tmin	27	0.35
Date	2	0.03
station	2	0.03

2. Filling nulls

Median Imputation for Temperature Columns: We first identified the columns related to temperature (i.e., those containing 'Tmax' or 'Tmin' in their names). For these columns, we used median imputation because median is less sensitive to outliers, which can be common in temperature data.

Mean Imputation for Other Numeric Columns: For the remaining numeric columns, we applied mean imputation.

Handling Missing Values in the 'Date' Column: For the Date column, we replaced any missing values with the most recent date in the dataset as they were the last 2 rows in the dataset. This ensures that no missing dates are left, while keeping the data consistent.

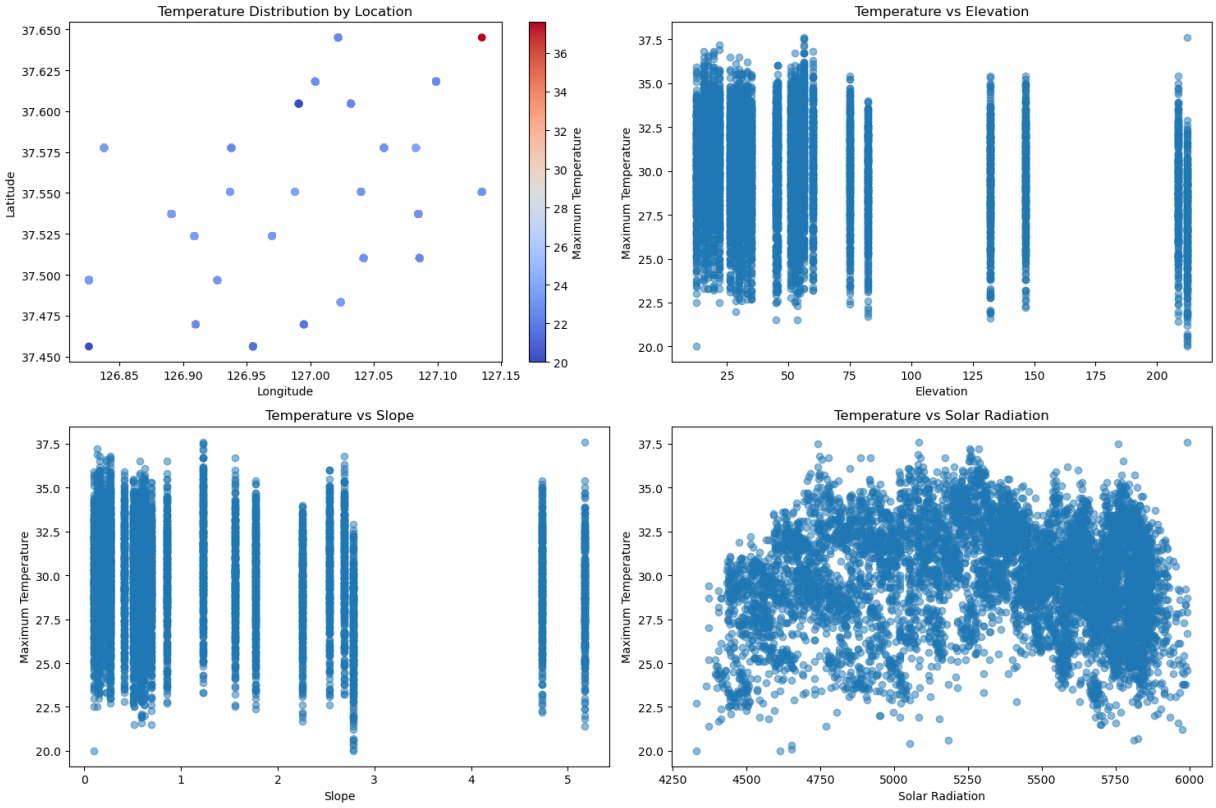
3. Outlier Handling

Handling Outliers in Temperature Columns: We focused on four temperature-related columns: 'Present_Tmax', 'Present_Tmin', 'Next_Tmax', and 'Next_Tmin'. For these columns, we used a two-step process to identify and handle outliers:

Interquartile Range (IQR) Method: We also used the IQR method to detect outliers. Any value falling below the first quartile (Q1) minus 1.5 times the IQR, or above the third quartile (Q3) plus 1.5 times the IQR, was flagged as an outlier.

After identifying outliers using both methods, we created a final mask to keep values that passed both checks. For values flagged as outliers, we replaced them with the median of the respective column.

Data Visualization:



- **Temperature Distribution by Location**:

  The map shows temperature spread across different locations. Most of the area has moderate temperatures (in blue), but there's one noticeable hot spot (in red), which could mean this specific location has unique factors making it warmer, like urban heat or specific environmental features.

- **Temperature vs Elevation**:

  This plot checks if elevation affects temperature. There's no strong trend here—temperatures seem to appear across a wide range of elevations. It does seem slightly warmer at lower elevations, which is typical, but overall, elevation doesn't seem to have a big impact on temperature here.

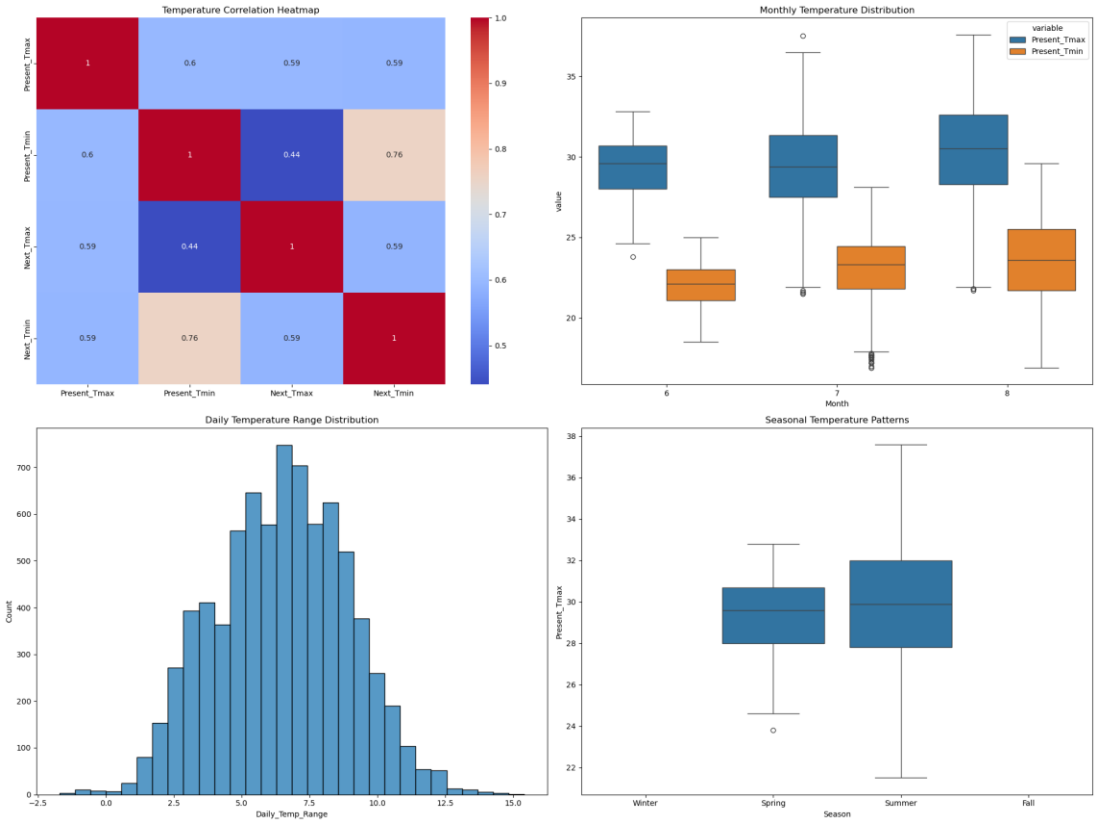
- **Temperature vs Slope**:

  Here, we're looking at whether slope influences temperature. The data shows that temperature values are scattered across different slopes without any clear pattern. It suggests that slope doesn't play a major role in temperature variation in this area.

- **Temperature vs Solar Radiation**:

  This plot shows a clearer relationship as solar radiation increases, so does temperature. This makes sense because more sunlight generally means higher temperatures. This link is stronger than what we saw with elevation or slope.

  There's a hot spot in one specific location, possibly due to local factors. Elevation and slope don't seem to influence temperature much. Solar radiation appears to be the main factor affecting temperature, with higher sunlight leading to warmer conditions. So, it's likely that sunlight exposure drives temperature differences in this area, more than elevation or slope.



Temperature Correlation Heatmap:

There is a **strong positive correlation** between the current and next month's maximum temperatures (Present Tmax and Next Tmax, with a correlation of 0.76). and between the current and next month's minimum temperatures (Present Tmin and Next Tmin, also 0.76).

Monthly Temperature Distribution:

Boxplots of the monthly temperatures show clear variations in both Present Tmax and Present Tmin.

Some months have a wider temperature range than others, reflecting fluctuations in temperature patterns throughout the year.

Daily Temperature Range Distribution:

The histogram of the daily temperature range reveals a peak around 10 degrees, indicating that on average daily temperature changes are around 10 degrees. However, the range is wide, showing that there can be significant variability in daily temperature changes.

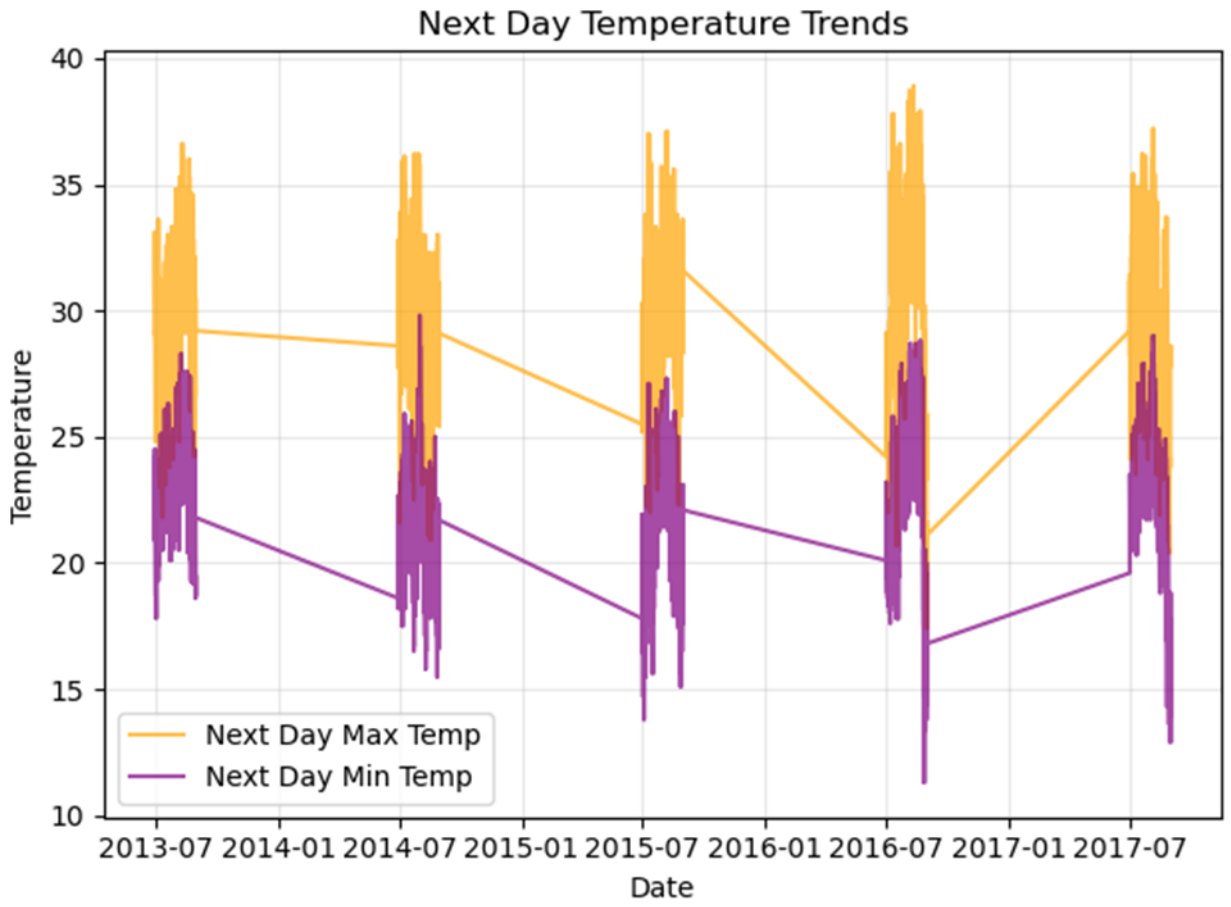
Seasonal Temperature Patterns:

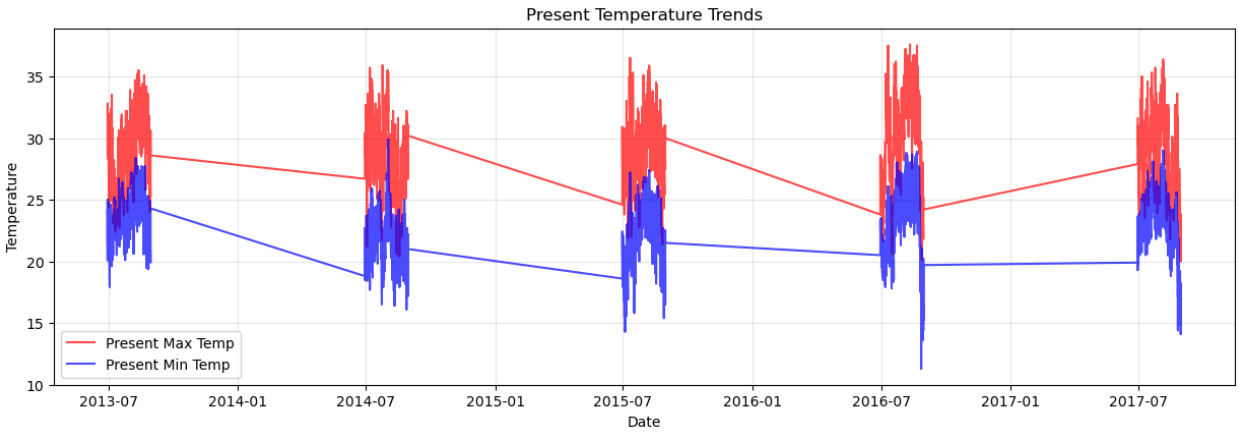
Summer has the highest median temperatures and the largest temperature range for both maximum and minimum temperatures.

Spring has moderate ranges.

Which concludes that there is a strong correlation between the current and next month's temperatures, which suggests some predictability in temperature patterns.

Daily temperature changes show more variability than monthly changes, indicating less consistency day-to-day.





This also prove the point that There is a **strong positive correlation** between the min and maximum temperatures for present and nest days.

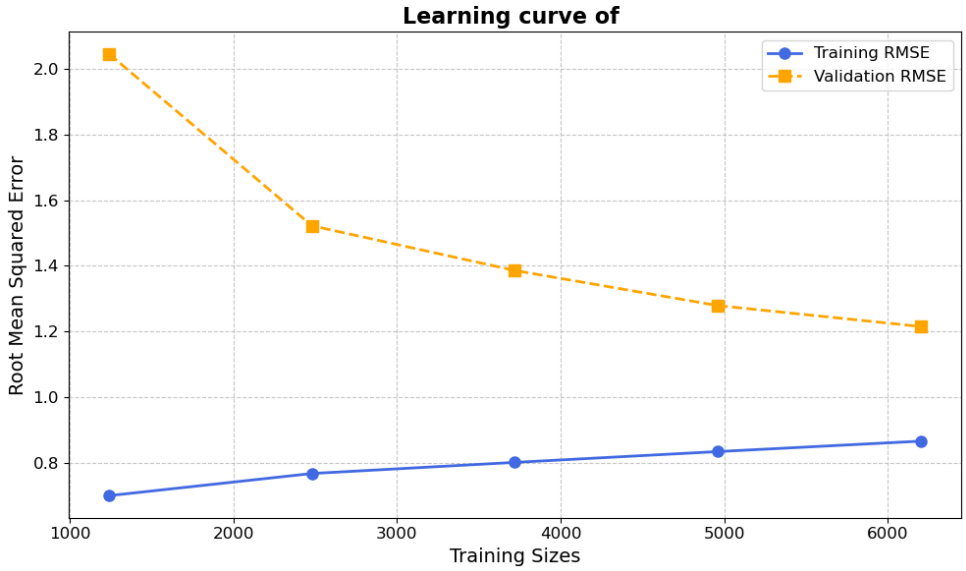
Model Training and Evaluation

Supervised Learning Algorithms:

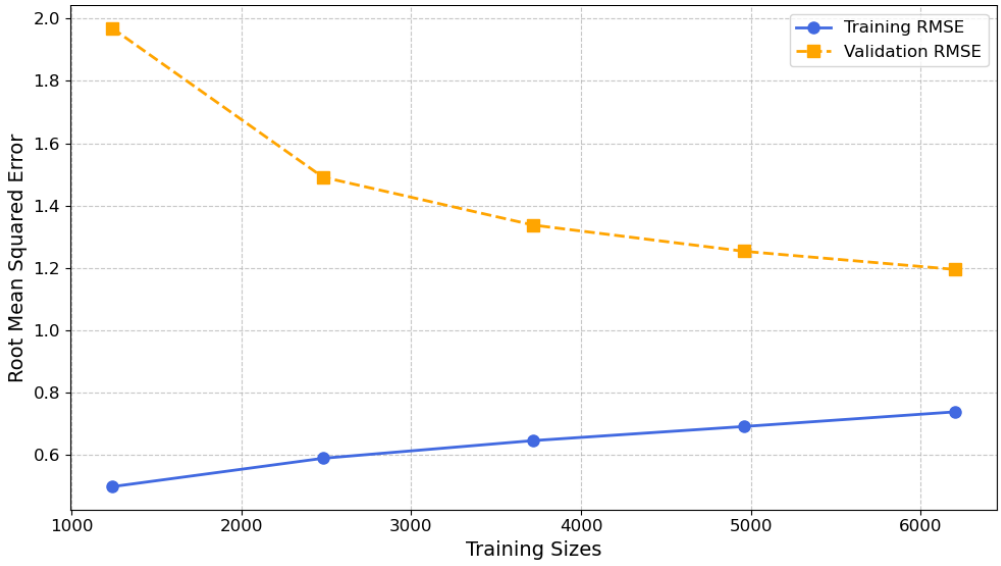
- Three models are covered: **Random Forest Regressor**, **Linear Regression**, and **LightGBM**.

Learning Curves:

1. RFregressor



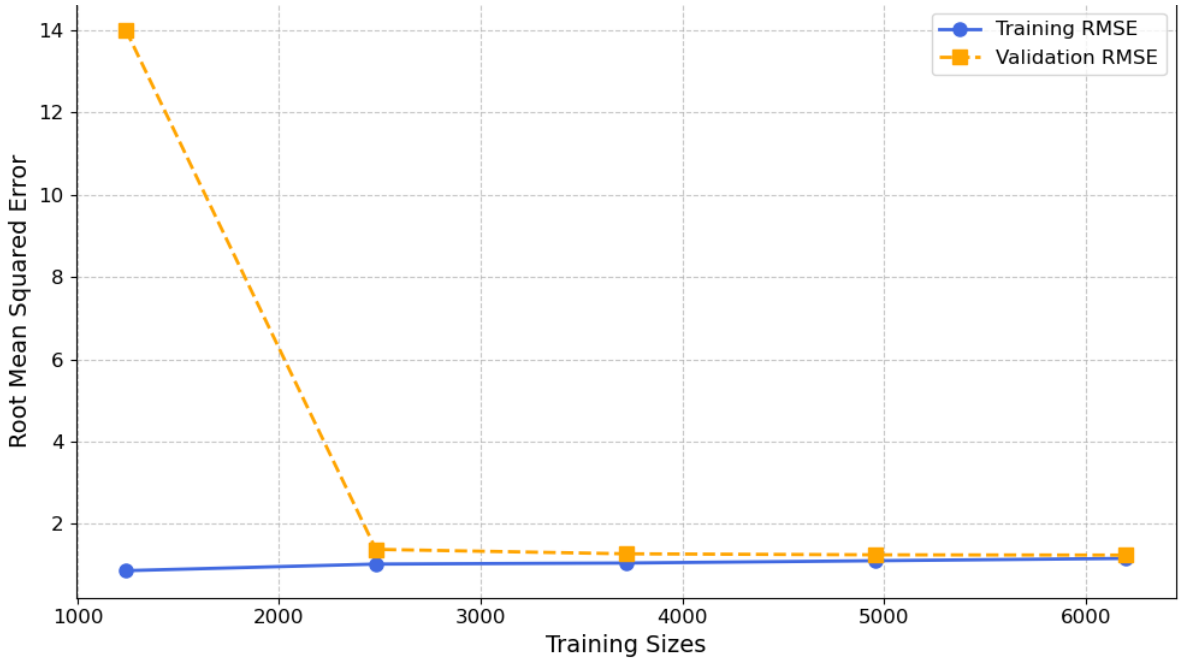
2. RFregressor with feature selection



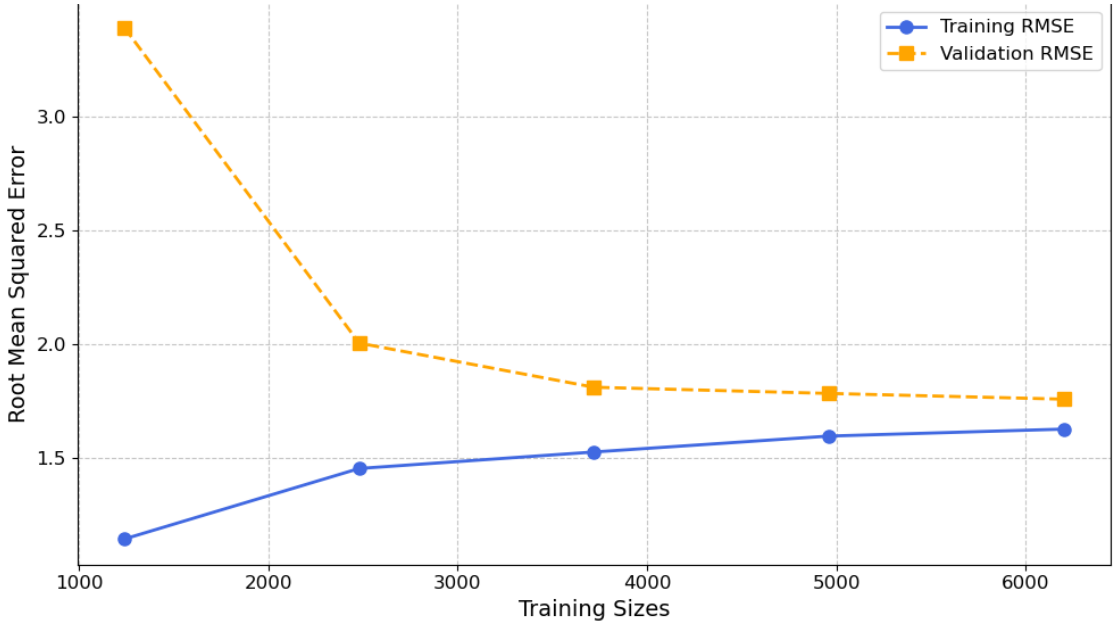
Without feature selection and tuning, the model does pretty well on the training data, but it has a hard time generalizing to new data. The validation error starts off high and only improves a bit

as more data is added, which suggests it's trying to learn from too many unnecessary details. Once we apply feature selection and tune the model, things improve a lot. The validation error drops a lot , and the gap between training and validation errors gets smaller, meaning the model is now better at focusing on what really matters and not the noisy data . With more thanks to the fine tuning

3. Linear Regression (Without Feature Selection)

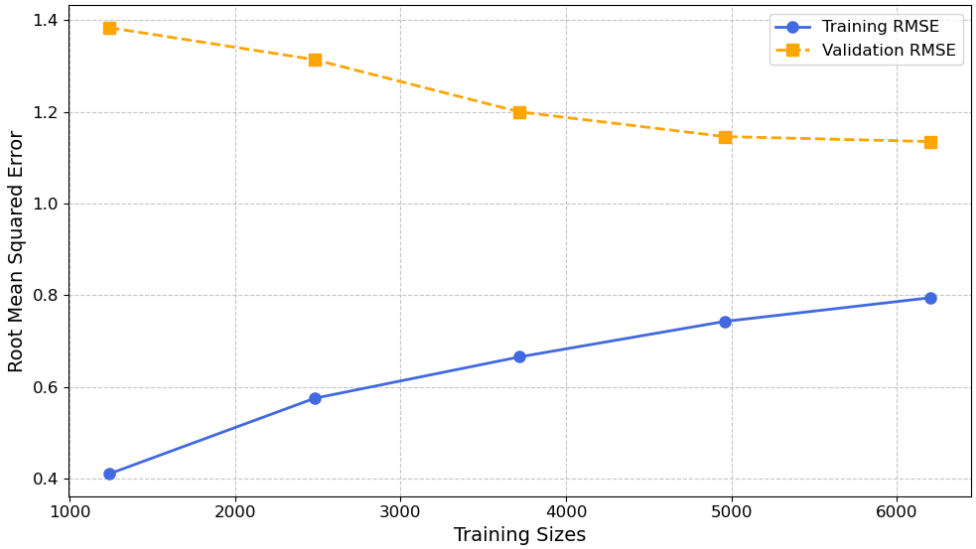


4. Linear Regression (With Feature Selection)

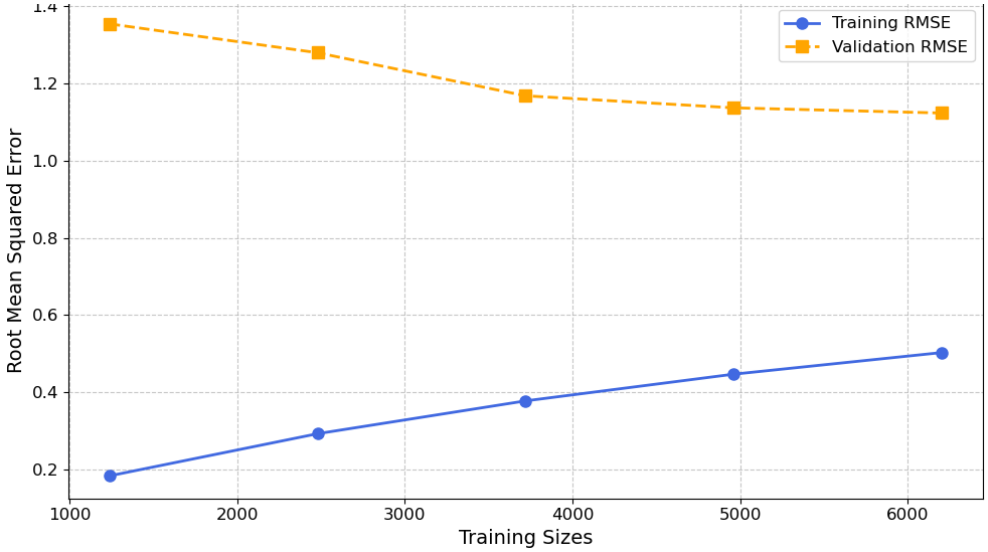


The model without feature selection validation error starts off high, but as we add more data, it quickly drops and aligns with the training error. This big drop suggests the model is initially thrown off by irrelevant features aka noise but starts to stabilize as it sees more data. When we use feature selection, the model behaves much better from the start. The validation error is lower and stays closer to the training error, even with smaller data sizes. This means features selection helps it stay focused and balanced, reducing overfitting and making it more consistent.

5. LightGBM (Without Feature Selection)



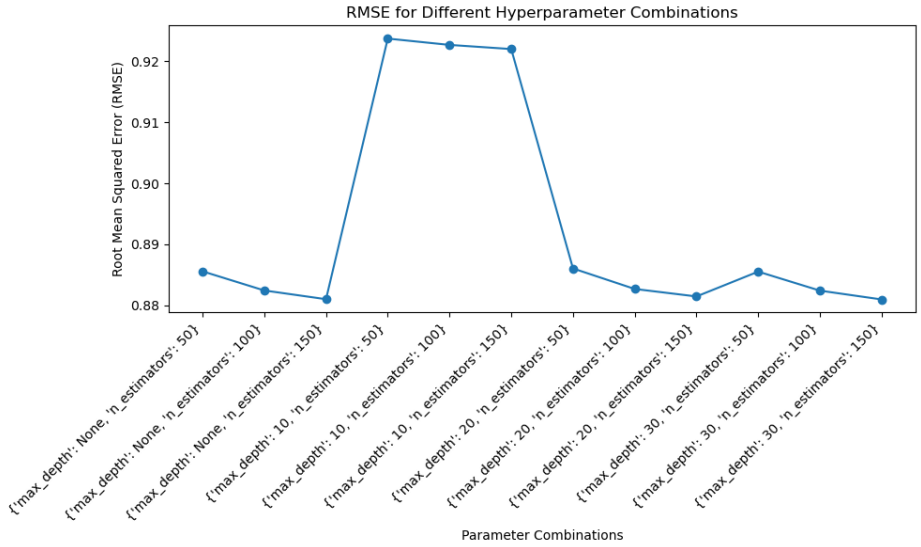
6. LightGBM (With Feature Selection)



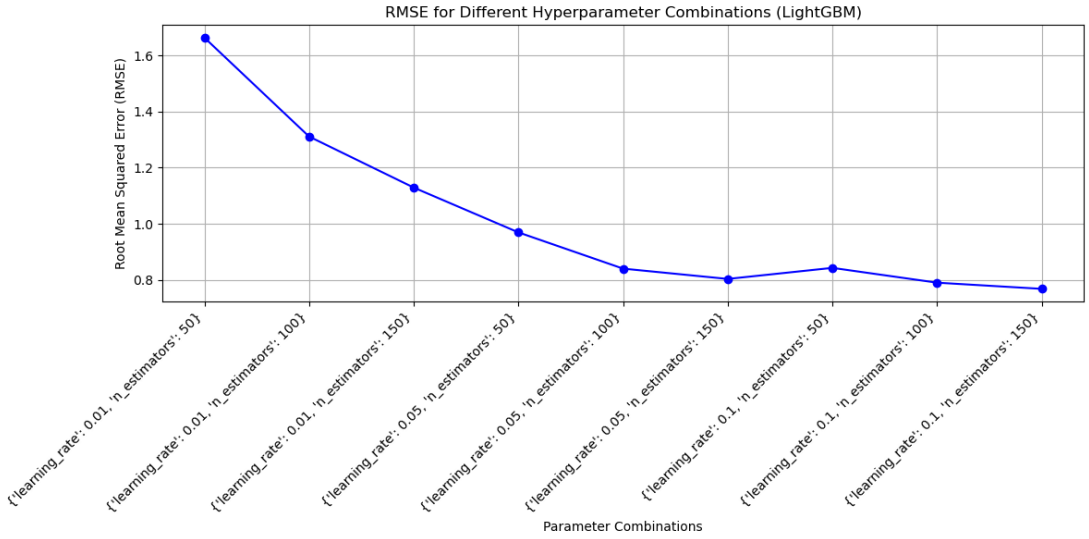
Without feature selection and tuning, the training error is low, but the validation error is higher and doesn't drop as much, indicating it's overfitting to some extent. After we fine-tune the model and remove irrelevant features, the validation error is much lower and is closer to the training error. This tells us that LightGBM really benefits from being fine-tuned.

RMSE Plot for different parameters :

1. RFregressor



2. Lightgbm



Model Performance Comparison:

- Comparing model performance using RMSE, MAE, and R2 score.

1. Best RFRegressor

Test RMSE: 0.9334571515629431

Test MAE: 0.6779546252836159

Test R^2: 0.8417526737541321

2. Best linear regressor

Test RMSE: 1.6588495683604956

Test MAE: 1.3074582501475074

Test R^2: 0.5002398601991447

3. Best light GBM

Test RMSE: 0.7110726610574573

Test MAE: 0.4867094486007157

Test R^2: 0.9081719062196303

LightGBM model outperforms the others, achieving the lowest RMSE (0.71) and MAE (0.49), along with the highest R^2 score (0.91).