**Phase 4 Blogs / Article**

**Temperature Forecasting Using Machine Learning: A Case Study in Seoul, South Korea**

**Prepare by Bhimrao Pawar**

Introduction:

**Predicting weather accurately is crucial for various sectors, including agriculture, transportation, and event planning. In this blog post, we delve into a machine-learning project to improve the forecast of next-day maximum and minimum air temperatures using data from the Local Data Assimilation and Prediction System (LDAPS) operated by the Korea Meteorological Administration.**

**Problem Statement**

**The goal of this project is to correct the bias in the next day maximum (Next-Tmax) and Minimum (next Tmin) air temperature forecasts provided by the Local Data Assimilation and Prediction System operated by the Korea Meteorological Administration.**

Dataset Description

**The dataset comprises multiple attributes, including:**

* **Present Tmax: Maximum air temperature on the present day.**
* **Present Tmin: Minimum air temperature on the present day.**
* **LDAPS RHmin: Forecast of next-day minimum relative humidity.**
* **LDAPS RHmax: Forecast of next-day maximum relative humidity.**
* **LDAPS Tmax lapse: Forecast of next-day maximum air temperature with lapse rate applied.**
* **LDAPS Tmin lapse: Forecast of next-day minimum air temperature with lapse rate applied.**
* **LDAPS WS: Forecast of next-day average wind speed.**
* **LDAPS LH: Forecast of next-day average latent heat flux.**
* **LDAPS CC1 to CC4: Forecast of next-day cloud cover in four 6-hour splits.**
* **LDAPS PPT1 to PPT4: Forecast of next-day precipitation in four 6-hour splits.**
* **Latitude, Longitude, Elevation, Slope: Geographic variables.**
* **Solar radiation: Daily incoming solar radiation**.

The Target Variables are

* Next Tmax: Actual next-day maximum air temperature (target variable).
* Next Tmin: Actual next-day minimum air temperature (target variable).

Methodology

Data Preprocessing

1. **Importing all required libraries for data injection, basic computation, data Visualization and plotting also importing warnings for different versions of packages**
2. **Loading Datasets with the help of pandas and making data into a Data Frame for better computation and feasibility**
3. **Checking rows and columns of the dataset contained 7752 rows and 25 no columns in Dataset**
4. **Verifying the column names also checking more information about the dataset. In this Dataset, all variables are numerical and Next Tmax and Next Tmin are the two target variables.**
5. **Checking Missing value, we found missing value by checking its percentage As missing values present are less than 1 %. So we can directly drop those missing values finally, with no missing value present in the dataset**
6. **In Dataset we see the data type of the date are different so we can change it date, month, and year format.**
7. **For more insight into data, we have created additional columns with the location of stations in terms of city and respective state by importing reverse geocoder**
8. **We have state and city as corresponding to its latitude and longitude available incorporate state and city columns in our dataset**

Encoding the categorical data

**We are two variables are categorical i.e city and state both are encoded by using label encoding**

**Dropping the unnecessary columns' lite date**

Feature Selection and Engineering:

**1 Outliers Detection in the data set we see many columns containing the missing value checked with the boxplot method**

**2 With the help of the Zscore method we removed outliers and we could see the data set contained 7639 rows and 29 columns after removing outliers. Almost We are 11.18 % of data losing but it's ok we can afford it because we have enough data set for building the model**

**3 Skewness is an important feature for continuous data. There is no relevance of skewness for discrete numerical features like month and categorical features. So we will ignore the skewness present in discrete numerical and categorical feature.**

**4 Checking the correlation with target feature wise also checking Multicollinearity between features using variance inflation factor**

Checking Principle component Analysis explained the variance ratio

**As per the graph, we see those 15 principal components contribute to 90% of the variation in the data. We shall pick the first 15 components for our prediction.**

**Task-1**

**I am separating the Dependent and independent features using the train test split method for the target of Next Tmax.**

**Applying the Standard Scalar method to normalise the data set**

Model Selection or Finding the Best Model

**1 Random Forest Regressor**

**2 Decision Tree Regressor**

**3 Extra Tree Regressor**

**4 Bagging Boosting Regressor**

**5 Gradient Boosting Regressor**

**6 XGB Regression**

Cross Validation

**Checking all model cross-validation scores we found the XGB Regressor is best model for this one**

Saving model

**Saving the Best model using the job lib method**

**Prediction of new data according to the final model check works well for new data with better accuracy**

**Task - 2**

**I am separating the Dependent and independent features using the train test split method for the target of Next Tmin.**

**Applying the Standard Scalar method for normalise the data set**

Finding the Best Model

**1 Random Forest Regressor**

**2 Decision Tree Regressor**

**3 Extra Tree Regressor**

**4 Bagging Boosting Regressor**

**5 Gradient Boosting Regressor**

**6 XGB Regression**

**7 Ada Boosting regressor**

Cross Validation

**Checking all model cross-validation scores we found the XGB Regressor is best model for this one**

Saving model

**Saving the Best model using the job lib method**

**Prediction of new data according to the final model check works well for new data with better accuracy**

**Results and Analysis**

**Present the results of your models, comparing their performance metrics.**

|  |  |  |
| --- | --- | --- |
| Models | Next Tmax | Next Tmin |
| Random Forest Regressor Score | 79 | 82 |
| DT Regressor Score | 45 | 77 |
| Bagging Boosting Regressor Score | 73 | 81 |
| Gradient Boosting Regressor Score | 76 | 79 |
| XGB Regressor Score | 89 | 86 |
| Ada Boosting Score | 69 | 66 |

**Analysis**

1. **RFR got a score 79 but its R2 score is less compared to another model**
2. **DT Regression model score is low**
3. **Bagging & Gradient boosting having a low score**
4. XGB showing Nice Score with other models

Challenges and Solutions

**1 Data Collection and Understanding the independent variable against the target variable**

**2 Handling Missing value**

**3 Handling outliers**

**4 Data Imbalance: Address any imbalance in the dataset and its impact on model performance.**

**6 Overfitting: Techniques used to prevent overfitting, such as cross-validation and regularization.**

**7 Feature Selection: How you determined the most relevant features for the models.**

**Conclusion**

**But here considering all Features and observations the model gives great accuracy with the minimum error with a nice R2 Score we achieve the goal of problems**

**The goal of this project is to correct the bias in the next day maximum (Next-Tmax) and Minimum (next Tmin) air temperature forecasts provided by the Local Data Assimilation and Prediction System operated by the Korea Meteorological Administration because of findings below points**

**Temperature and cloud cover-related features are moderately or highly correlated with the Next Tmax and Next Tmin**

**Tmax and Tmin for applied lapse rate are almost normally distributed.**

**Tmax\_lapse for the majority of days its values lies in the range 27 to 33 (Â°C).**

**Tmin\_lapse for the majority of days its values lies in the range 23 to 26 (Â°C).**

**We can say that for cloud cover greater than 0.7 high-value solar radiation contributed the most.**

**Irrespective of 6 hr quarter in 24 hr day time, Next-day Temperature Maximum decreases as cloud cove**

**We can notice that station 18 has the highest temperature collection for both Next\_Tmax and Next\_Tmin targets.**

**The maximum temperature difference at the same station occurs in station 17. r increases beyond 0.6.**

**For the majority of readings, the Next Temperature Maximum varies between 27.5 to 32.5 degrees.**

**For the majority of readings Next Temperature Minimum varies in between 22.5 to 26 degrees.**

**GitHub Link Project**

[**https://github.com/TVsony/Evaluation-project-/blob/main/Temperature\_Forecast%20(1).ipynb**](https://github.com/TVsony/Evaluation-project-/blob/main/Temperature_Forecast%20(1).ipynb)

**Thanks**

**Regards**

**Bhimrao Pawar**