

TEMPRATURE FORECAST MODEL



Submitted by:
UJJWAL PRATIK

Contents

TEMPRATURE FORECAST MODEL	1
ACKNOWLEDGMENT	3
INTRODUCTION	4
ABOUT TEMPRATURE FORECAST	4
PROBLEM STATEMENT	4
ABOUT DATASET	4
DATA ANALYSIS	5
COLUMNS IN THE DATAFRAME	5
OBJECTIVE	5
DATA DESCRIPTION	5
METHEDOLOGY	5
METRIC USAGE	6
SYSTEM REQUIREMENTS	7
APPROACH	8
IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES	8
TESTING OF IDENTIFIED APPROACH(Algorithms)	9
KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION	10
CONCLUSION	11
TARGET 1 (Next_Tmax)	11
TARGET 2(Next_Tmin)	11
LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE	12
REFERENCES	13

ACKNOWLEDGMENT

I sincerely thanks to the Data Trained Faculty for the guidance. They have covered the topics like Machine Language, Python & SQL. Under their guidance I learned a lot about this project. I had also taken help from YouTube & online videos.

INTRODUCTION

ABOUT TEMPRATURE FORECAST

A prediction of future weather, often for a specific locality, in a newspaper or on the radio or television. It may be sunny today, but the weather forecast is predicting rain for tomorrow.

PROBLEM STATEMENT

This data is for the purpose of bias correction of next-day maximum and minimum air temperatures forecast of the LDAPS 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. Hindcast validation was conducted for the period from 2015 to 2017.

ABOUT DATASET

We have two target variables in the dataset. Next_Tmax and Next_Tmin.

DATA ANALYSIS

COLUMNS IN THE DATAFRAME

- station
- Present_Tmax
- Present_Tmin
- LDAPS_RHmin
- LDAPS_RHmax
- LDAPS_Tmax_lapse
- LDAPS_Tmin_lapse
- LDAPS_WS
- LDAPS_LH
- LDAPS_CC1
- LDAPS_CC2
- LDAPS_CC3
- LDAPS_CC4
- LDAPS_PPT1
- LDAPS_PPT2
- LDAPS_PPT3
- LDAPS_PPT4
- Lat
- Lon
- DEM
- Slope
- Solar radiation
- Next_Tmax (Target1)
- Next_Tmin (Target 2)

OBJECTIVE

- Our main objective is to predict bias correction of next-day maximum and minimum air temperatures forecast.
- All the parameters will be analysed through Machine Learning algorithms like Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression etc which will help to predict.

DATA DESCRIPTION

- The dataset contains two target variables Next_Tmax and Next_Tmin.
- The source of data is taken from GitHub.
- (<https://github.com/dsrscientist/Dataset2/blob/main/temperature.csv>)

METHEDOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learnings techniques to determine the temperature because it gives the best outcome.

METRIC USAGE

- a. Linear Regression.
- b. Lasso Regression.
- c. Ridge Regression.
- d. Elastic Net Regression.
- e. K Nearest Regressor.
- f. Random Forest Regressor.

SYSTEM REQUIREMENTS

Hardware and Software Requirements and Tools Used

- a) Hardware Requirement:
 - i. Intel core i5
 - ii. 8 GB Ram
- b) Software Requirement:
 - i. Python 3.x with packages:
 - 1. Pandas: Data analysis and manipulation tool
 - 2. NumPy: Provide support for mathematical functions, random number etc.
 - 3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
 - 4. Seaborn: is a library mostly used for statistical plotting in python.
 - 5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

APPROACH

IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES

- R2 score: is used to evaluate the performance of a linear regression mode.
- Linear Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
- Lasso: The Lasso is a linear model that estimates sparse coefficients with l1 regularization.
- Ridge: Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model.
- Elastic Net: is a linear regression model trained with both l1 and l2 -norm regularization of the coefficients.
- Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
- Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
- Mean Squared Error: this metric gives an indication of how good a model fits a given dataset.
- Root Mean Squared error : is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.
- Z-score: Z-score is also known as standard score gives us an idea of how far a data point is from the mean.
- Label Encoder: Label Encoding refers to converting the labels into numeric form.
- K Nearest Regressor: It observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.
- Standard Scaler: Standard Scaler. Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance).
- Random Forest Regressor: A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

TESTING OF IDENTIFIED APPROACH(Algorithms)

- a. Train Test Split
- b. Linear Regression
- c. Lasso Regression
- d. Ridge Regression
- e. Elastic Net Regression
- f. Grid Search CV
- g. Cross Validation
- h. Decision Tree Regressor
- i. Random Forest Regressor

KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

- Analysed data for any unique values.
- Analysed data for distribution.
- Checked outliers through z-score method.
- checked skewness present in the dataset.
- Done Standard Scaling.
- Cross validate the r^2 score from overfitting.
- Hyper Parameter tuning using Grid Search CV

CONCLUSION

TARGET 1 (Next_Tmax)

After checking r2 score, cross validation, Ensemble Techniques, we came to the conclusion Random Forest Regressor 99% accuracy is best suited model for our purpose of predicting.

TARGET 2(Next_Tmin)

After checking r2 score, cross validation, Ensemble Techniques, we came to the conclusion Random Forest Regressor 99% accuracy is best suited model for our purpose of predicting

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc. It also helped me to learn how to apply various model techniques on data and enable predications.

REFERENCES

- Data trained course videos.
- Google Search
- YouTube
- GitHub
- UCI Machine learning repository