Temperature Prediction in Ho Chi Minh City

Tam-Ngoc-Bang Nguyen[1] • Tu-Van Huynh[2]

{1712747[1], 1712856[2]}@student.hcmus.edu.vn

Abstract

In this project, we explore the relationships between weather variables and use machine learning to predict next-day mean temperature with a case study in Ho Chi Minh City, Vietnam. We exploited historical data of the past $N = 3$ days before the prediction for our implementation. For final result, our best model achieved an R^2 score of 0.766 an RMSE of 0.896.

Introduction

Traditional weather forecasts have employed physical models to handle real-time and dynamic properties of climate patterns which demand high-performance computing systems. Our project aims to propose a more lightweight approach for short-range temperature prediction using machine learning techniques. With a case study in Ho Chi Minh City, given climate information of the previous N days, we expect to predict the mean temperature of the following day.

Method

The project pipeline involves four major steps.

1: Data collection
- HTML parsing from World Weather Online website with Selenium

2: Data preprocessing
- Handling null and duplicated values
- Handling skewed attributes with log transformation

3: EDA & visualization
- Checking correlations and feature distributions

4: Modelling & evaluation
- Benchmarking 5 models
- Hyperparameter tuning using GridSearch
- MSE, RMSE, MAE and R^2 score for evaluation

Data

We collected data for Ho Chi Minh City from World Weather Online from 01/01/2009 to 21/12/2020. For each day, the features obtained were the weather classification, maximum and minimum temperature, wind speed and direction, mean precipitation, percentage of humidity, cloudiness and mean atmospheric pressure. We then averaged the temperature information from 8 periods of each day to create the mean temperature feature.

Features

We selected 8 numerical features for our regression models: to predict the mean temperature the following day, we concatenated weather data of the previous 3 days ($N = 3$).

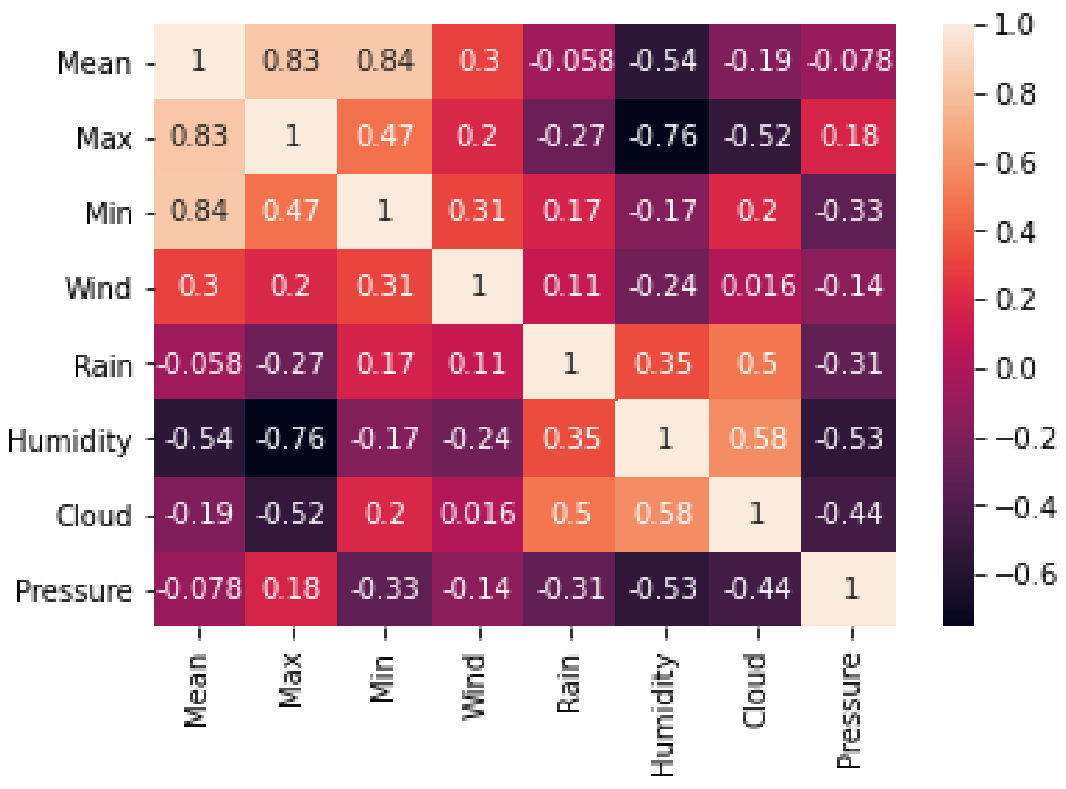


Figure 1. Heatmap of 8 numerical features

In total, we used 24 variables: maximum and minimum temperature, wind speed, mean precipitation, percentage of humidity, cloudiness and mean atmospheric pressure and mean temperature of the previous 3 days to build our models.

Models

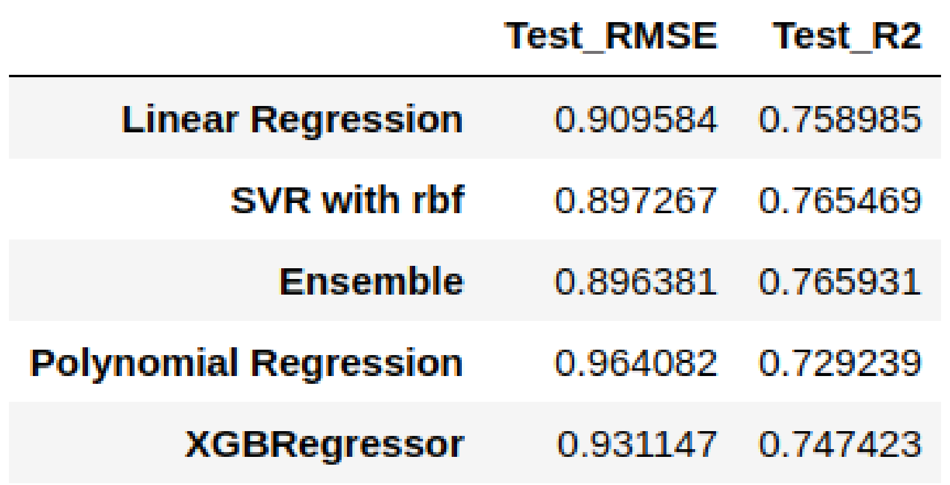
We implemented 5 types of machine learning models for regression problems: linear regression, SVR with rbf and polynomial kernels, polynomial regression, XGBoost regressor and ensemble of the two best performing models.

Our data consists of 4369 samples and 11 features. The train - test ratio is around 85 : 15, with training set ranges from 01/01/2009 to 31/03/2019 and testing set ranges from 01/04/2019 to 21/12/2020.

For hyperparameter tuning, we applied GridSearch and found out $C = 100$ and $\gamma = 0.0001$ yield the most optimal SVR.

Results

Below is the benchmark (i.e the RMSE and R^2 score) on testing set



	Test_RMSE	Test_R2
Linear Regression	0.909584	0.758985
SVR with rbf	0.897267	0.765469
Ensemble	0.896381	0.765931
Polynomial Regression	0.964082	0.729239
XGBRegressor	0.931147	0.747423

Figure 2. RMSE and R^2 score

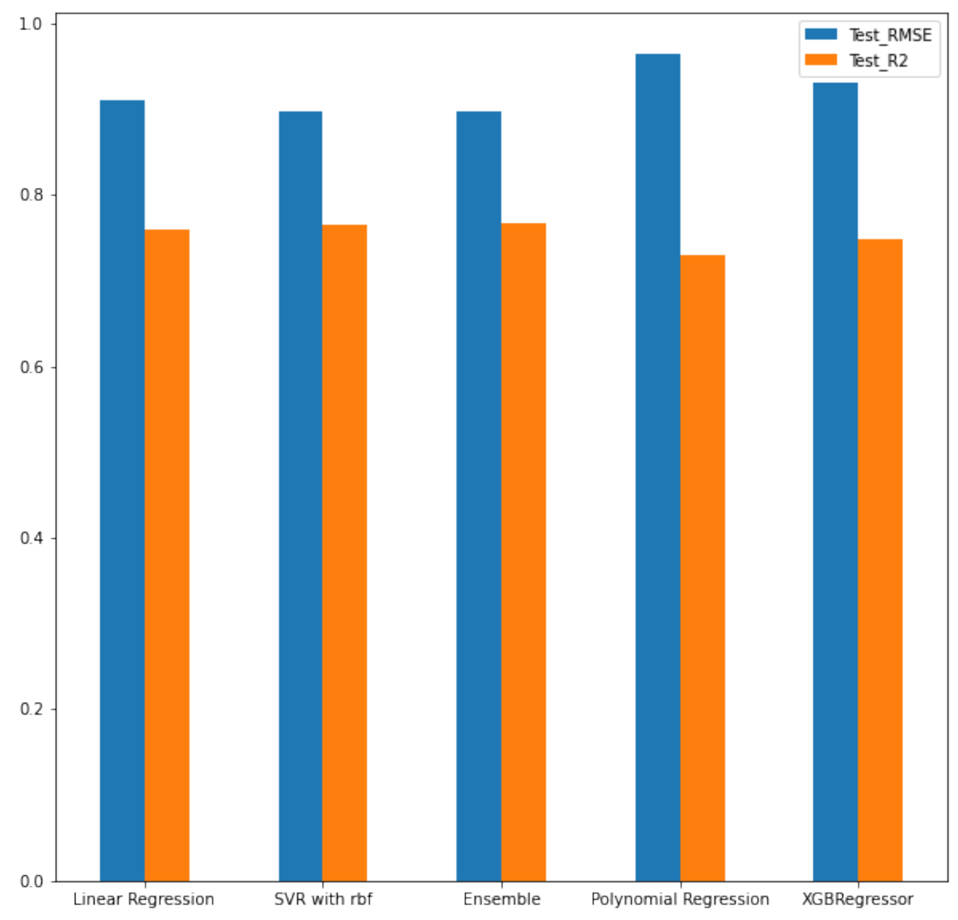


Figure 3. RMSE and R^2 score in visualization

Conclusion

We gained some interesting insights regarding the weather information of Ho Chi Minh City:

- Rain distribution of Ho Chi Minh City is heavily left-skewed and the models achieved better performance (R^2 score was improved by a margin of 0.02 and MSE was reduced by 0.07 after we applied log transformation for this attribute.
- The difference between maximum and minimum temperature of a day in HCMC is fairly high and mostly ranges from 6 to 10°C.
- Higher percentage of cloudiness strongly correlates the overall temperature.

Results also showed that ensemble model achieved the highest R^2 score and lowest RMSE on this dataset.

Future

We plan to extend the project by increasing data volume and employing categorical attributes including the weather classification and wind direction for model building.

References

[1] Holmstrom, M. and D. Liu. “Machine Learning Applied to Weather Forecasting.“ (2016).