

Assignment 1: Part D

(Report)

# Divyarashmi Gurram

# A1917395

**The University of Adelaide**

# 4436\_COMP\_SCI\_7209 Big Data Analysis & Project Lecturer: Bernard Evans

**Restatement**

This dataset has 63 columns with 11 variables, and it includes temperature data of different cities and countries starting from 1853 and has 48,470 entries. Incredibly, the necessary columns are as follows: AverageTemperatureFahr, AverageTemperatureUncertaintyFahr, City, Country, Latitude, and Longitude. It offers a history of maximum and minimum temperatures together with associated geographic information that may be useful in climate analysis and assessment of trends.

Climate change is associated with several direct biophysical effects, regional variations in the amount and intensity of precipitation, and the growing frequency of extreme weather events. (Janches, Henderson, & MacColman, 2014, pg.2).

I want to focus on how reliable the historical temperature differences across countries to the current forecasted temperatures is and to what extent can geographical location and historical variations in temperatures inform accurate future temperatures estimates. Following with, would it be possible to come up with a model using this correlation between countries and the difference in temperature they record compared to others.

**Analysis and Visualization**

One of the potential issues with the dataset is that there are some missing values, especially when it comes to the temperature columns. The data is preprocessed, and a few models were performed on the obtained dataset. The results of the models are as follows

* XGBoost Regressor: capable of calculating non-linear and complex factors influenced by geography, seasonality, and climatic conditions in structured / tabular datasets.

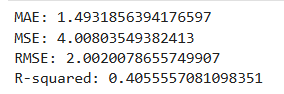
A number on a white background

Description automatically generated

A graph showing a number of blue and red lines

Description automatically generated

* RandomForest Regressor: It can work with datasets having missing values, it captures interactions between variables, and lastly, the method offers reliable and accurate results.

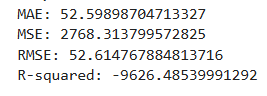


* SVR: It captures non-linearity and is not sensitive to small points to predict with and works effectively with high dimensional spaces.

A number on a white background

Description automatically generated

* LSTM: It captures temporary dependencies, help in model long-term patterns and trends in temperature variations overtime.



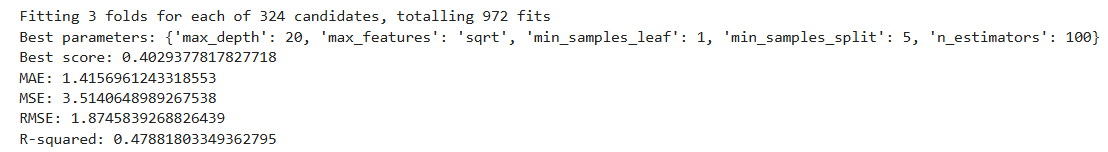
A graph showing a red line

Description automatically generated

LSTM is a complex model for my dataset as it is very less size and very limited features to train. It learns the data very well while training such that it also learns the noises during the training period because of less complex data and performs poorly while testing which makes LSTM not the best model for my dataset.

**Improvement of Situation**

As model refinement is done on the best performing model from the models represented above and improved the result of the model. Among all the models used on my dataset RandomForest Regressor stands out with pretty good accuracy. Now, model refinement steps are performed as follows: GridSearchCV is used for the hyperparameters optimization of Random Forest Regressor such as n\_estimators, max\_features, max\_depth, min\_samples\_split, and min\_samples\_leaf. This process helps in determining the correct structure of the model to be used in developing the model that will enhance the likelihood of making accurate predictions.



A graph showing the temperature of a person

Description automatically generated with medium confidence

**Conclusion and Future work**

The evaluation indicators used to assess the model’s performance are Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²). XGBoost Regressor, RandomForest Regressor, SVR, LSTM are the models used. The XGBoost model is not that efficient, where R-squared of 0. 111 which shows that the model can only explain 11%. We found its impact on the outcome or explanation variable to contribute to 1% of the variance in the data. This is evident by the RMSE of 2. 45 and MAE of 1. 79 meaning there is much that the model could be improved in error prediction.

The Random Forest model is more accurate with the R-squared of 0. 406, this shows that the variable gives account of 40% of the total variation. 6% of the variance. Compared with the XGBoost model, lower RMSE (2. 00) and MAE (1. 49) inform of better prediction whereas there still exists the potential of enhancement.

The SVR model gives an R-Squared of 0. RMSE=2. 24 and MAE = 1. 70 suggest that while our model has prediction errors well but less than XGBoost, the accuracy is reasonable but with room for improvement.

The RMSE for the LSTM model is very high, about 52. 61 and the R-squared value is negative of 9626. 49 meaning the model is worse than random and relies on just predicting the mean. As my dataset is very small for such a complex model to work on. It trains the model very well such that all the errors are neglected.  
  
 From the above models RandomForest regressor performs very well and to improve further, GridSearchCV is used. The GridSearchCV refinement improved the Random Forest model's performance, evident from an increased R-squared value (0.479), indicating better variance explanation. The lower RMSE (1.87) and MAE (1.42) reflect enhanced prediction accuracy.

For future research: The model can be improved on hyperparameter tuning beyond the current iteration; including more features that may be relevant and give more scope for better prediction; considering deploying deep learning models such as Transformers; and enhance data preprocessing. The use of ensemble methods, which combine multiple models, would also improve the accuracy of the predictions. The efficiency and robustness of the forecasts could also be boosted through refinement of data quality, especially for time series models.

**References**

data.gov.au. (n.d.). Search. [online] Available at:

<https://data.gov.au/search?q=climate+changes>

Janches, F., Henderson, H. and MacColman, L. (2014). Climate Change, Vulnerability, and Adaptation in Urban Areas. [online] JSTOR. Available at:<https://www.jstor.org/stable/resrep18470.5>

data.world. (n.d.). Search - temperature in major cities. [online] Available at:<https://data.world/search?q=temperature+in+major+cities>

figshare.com. (2016). temperature.csv. [online] doi:<https://figshare.com/articles/dataset/temperature_csv/3171766>