The World Temperature Changes Related to Atmospheric Concentrations in the 21st Century Based on Machine Learning

Zixian Gong

Mathematics and Applied Mathematics
Xiamen University Malaysia, 43900 Sepang, Selangor Darul Ehsan, Malaysia
**Corresponding author. E-mail: MAT1909400@xmu.edu.my*

ABSTRACT
Since the 21st century, the industry has developed rapidly, and emissions of various greenhouse gases have reached new highs. Global warming has also led to more natural disasters. Choosing a suitable model to analyze the influence of several gas concentrations in the atmosphere on the overall temperature change since the 21st century can predict climate trends to a certain extent. This article compares the MAE, RMSE, and R-Square of the machine learning models of Linear Regression, ElasticNet Regression, Random Forest, Extra Trees, SVR, Gradient Boosting Regression Tree, XGBoost to predict temperature changes. Finally, as the atmospheric concentrations of the four greenhouse gases rise, the temperature deviation from 1951 - 1980 mean also gradually increased may even be exceeding 1.5°C in the future. It is found that among these methods, Gradient Boosted Tree has better results than R-square of it comes to 0.72 which is the largest one and a relatively good value, so it can be used as a model for predicting temperature changes.

Keywords: *Greenhouse Gases, Temperature, Environmental Protection, Machine Learning, Regression.*

1. INTRODUCTION

Since the beginning of the new century, the rapid development of global industry has not only brought convenience to people, but also brought considerable impact on the natural environment. The frequent occurrence of extreme weather around the world and sea-level rise can be said to be related to global warming to a certain extent. Countries and international organizations are paying more and more attention to climate issues, on November 13, the 26th United Nations Climate Change Conference in Glasgow (COP26), which lasted for half a month, came to an end, although the conference was postponed for a year due to the epidemic, in general, it still achieved rare results in global climate governance in recent years, achieved the “Glasgow Climate Pact” and affirmed the necessity and urgency of the goal of “achieving a temperature control of 1.5°C”[1], which will clarify the development direction for the implementation of the “The Paris Agreement”.

This article uses the global monthly mean trend concentration data of CO_2, CH_4, N_2O, and SF_6 (Kyoto Protocol)[2] recorded by the National Oceanic and Atmospheric Administration (NOAA)[3][4][5][6], and the combined Land-Surface Air and Sea-Surface Water Temperature (Land-Ocean Temperature Index, L-OTI) data[7][8] provided by National Aeronautics and Space Administration (NASA) and Goddard Institute for Space Studies (GISS). Regression analysis and comparison between different machine learning models are used to select a more appropriate model which can make help to achieve the purpose of temperature change prediction. The above data are from January 2001 to August 2021, a total of 248 months. These Greenhouse Gases are reported as a “dry air mole fraction”. This concept can be interpreted as molecules’ number of the gas divided by the total number of molecules in the air. Among them, CO_2 concentration is measured in parts per million (ppm), CH_4 and N_2O concentrations are measured in parts per billion (ppb) and SF_6 concentration is measured in parts per trillion (ppt). As to the temperature data, NOAA GHCN v4 (meteorological stations) and ERSST v5 (ocean areas), combined as described in Hansen et al. (2010)[9] and Lenssen et al. (2019)[10] and given as the deviation from 1951-1980 mean. Analyzing and predicting the global temperature change in the 21st century combined with atmospheric concentration can help to better limit some industrial emissions to protect the climate.

2. METHOD

This article uses a total of seven machine learning models: Linear Regression, ElasticNet Regression, Random Forest, Extra Trees, Support Vector Regression, Gradient Boosting Regression Tree, XGBoost, and uses the median in the temperature change training set as the baseline and calculates its MAE, RMSE, R-Square to compare with different machine learning models.

2.1. Data overview

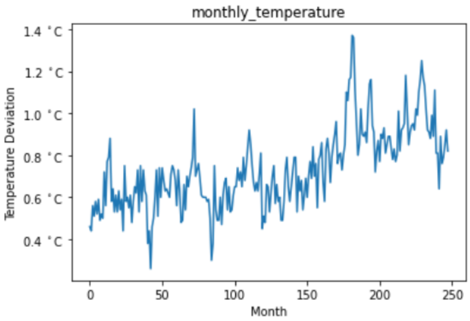


Figure 1 Monthly temperature against the month starts from January 2001

The overall temperature has an upward trend (figure 1). (The x-axis represents January 2001 to August 2021, and the y-axis represents temperature deviation from the 1951-1980 mean)

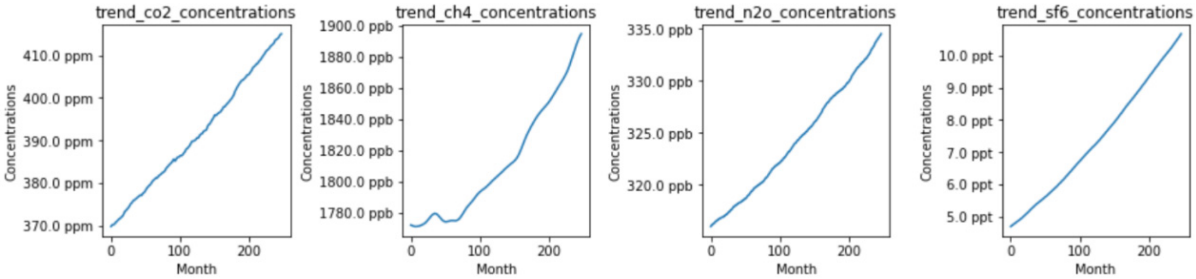


Figure 2 Trend Concentrations against the month starts from January 2001 of four gases

The atmospheric concentrations of these four Greenhouse gases are also on the rise in new century. (figure 2)

2.2. Data preprocessing

Different greenhouse gas concentrations use different measurement units, so there are differences in magnitude. Before model training and parameter optimization, all factors need to be normalized and limited to between 0 and 1 as follows:

$$X_n = \frac{X_n - \min(X)}{\max(X) - \min(X)}$$

To evaluate the predictive ability of different algorithm models and avoid over-fitting of the model, this article uses the open-source toolkit scikit-learn to randomly segment the normalized samples to ensure the independence of the test samples and randomly select the sample data 30% are used as test data, and 70% are used as training data for model training and parameter optimization.

2.3. Machine learning methods

2.3.1 Linear Regression

Linear Regression is the most common method in regression analysis. Usually there is only one independent variable, and the simple linear regression means that the relationship between the independent variable and the dependent variable can be represented by a straight line. Multiple linear regression refers to the existence of two or more independent variables, and there is a linear relationship between dependent variable and independent variable.

There are four greenhouse gas atmospheric concentration data. All have a certain linear relationship with temperature changes. The atmospheric concentration of the four gases can be regarded as an independent variable. The temperature deviation data is regarded as a dependent variable and the model is obtained: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$. y represents temperature deviation, x represents the atmospheric concentration of four different gases, and ε represents random error.

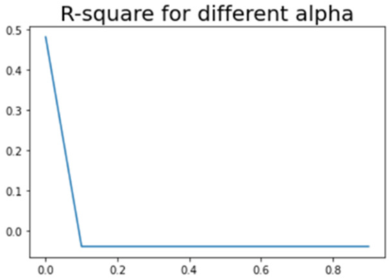
2.3.2 ElasticNet Regression

ElasticNet regression is a combination of ridge regression and Lasso technology. Elastic Network (ENET) is a linear regression model trained using L1, L2 norms as a priori regular term and can be formulated as the minimizer of:

$$\hat{\beta}(\text{ENET}) = \underset{\beta}{\operatorname{argmin}} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \text{subject to } P_\alpha(\boldsymbol{\beta}) =$$

$$(1-\alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|_2^2 \le s \text{ some } s \text{ [11]}$$

Where $P_\alpha(\boldsymbol{\beta})$ is the ENET penalty [12]. The ENET simplifies to simple ridge regression when α=1 and to the lasso when α=0. The advantage of doing this is that it can inherit Ridge's stability during the cycle. (under rotate).



"

Figure 3 R-square of ElasticNet Regression under different alphas

According to the figure 3, in this project, it is better to choose the case of α = 0, which is lasso.

2.3.3 Random Forest Regression (RF)

Random Forest Regression is a kind of supervised learning algorithms using ensemble learning method. RF is based on bagging regression, and multiple sample combinations are obtained by random and repeated observations of sample data with replacement, and through node splitting and random extraction of random feature variables to form multiple decision trees to form a "forest". The weighted average of the results of each decision tree is used as the regression result of the sample [13]. In this way, the prediction accuracy and generalization ability can have improvement to a certain extent, and the phenomenon of overfitting can be avoided. This article will choose the number of trees = 100 for analysis.

2.3.4 Extra Trees

Extra Trees and Random Forest have many similarities. They are ensemble methods, both of which consist of numbers of decision trees. The predictions of each tree were taken into account by the final decision. Extra trees, as an extension of the random forest algorithm, was developed. It adopts the same principle that of random forest, and trains each base estimate using the random subset of features [14]. In random forests, the observations are chosen with replacement, so it may repeat observations in a random forest. But in the extra tree, the observations are chosen without replacement. Thus, it does not have revisited observations like random forests.

Another difference is that in extra trees nodes are split based on random splits among a random subset of the features selected at every node while the random forest selects the best. However, once the split points are selected, it will still choose the best one in the subset of features. Therefore, extra trees adds the randomization but still has optimization and it tends to have a smaller variance. This article will also choose the number of trees = 100 for analysis.

2.3.5 Support Vector Regression (SVR)

Support Vector Regression (SVR) is an application of Support Vector Machine (SVM) to regression problems. Support vector machine is a supervised learning method. Through nonlinear mapping, the principle of structural risk minimization is adopted to map the low-dimensional space and linearly inseparable data into the high-dimensional space, so that can make it linearly separable and do classification and prediction in the high-dimensional space [15]. SVR can also be used as a regression method. SVR uses the same principles as SVM classification, but there are still some differences between them. Because the target value is a real number which is difficult to predict due to the infinite possibilities. The algorithm is more complex, so it needs to be considered. However, the main idea is always the same: minimizing error; individualizing the hyperplane which maximizes the margin; keeping in mind that part of the error is tolerated [16]. In general, SVR has two types: linear and non-linear. The choice of model is more inclined to consider its accuracy, so non-linear is considered, which is gaussian kernel for radial base function (rbf):

$$k(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(-\frac{\|\mathrm{x}_i - \mathrm{x}_j\|^2}{2\sigma^2}\right)"$$

2.3.6 Gradient Boosting Regression Tree

Gradient boosting is one of the most powerful machine learning methods which concatenates multiple simple learning algorithms (weak learners) to improve the accuracy and generalization ability of the model. The learner generated in the next round of iteration is trained based on the previous round. Each additional learner is a correction to the previous model, so the model can be expressed as:

$$F_m(x) = F_{m-1}(x) + \gamma_m \varphi_m(x)$$

where the $\mathrm{F_m}$ represents the integrated model obtained after the m-th iteration, φ_m means the mth learner, and γ_m is the weight of this learner. The main idea is to gradually reduce the prediction deviation and improve the prediction accuracy through iterative processes.

Gradient Boosting Regression Tree (GBRT) is a combination of gradient boosting and regression trees for regression problems [17]. The tree is continuously built, and each tree tries to correct the error of the previous tree, and finally gets a more accurate result.

2.3.7 XGBoost

XGBoost is similar to GBRT, both of which follow the principle of gradient boosting. But XGBoost stands for Extreme Gradient Boosting which uses a more regularized model formalization to control over-fitting. First of all, XGBoost considers the second-order gradients which provides more information about the direction of gradients, and another is that XGBoost uses advanced regularization to further improve model generalization. In this article, both Gradient Boosting Regression Tree and XGBoost will choose number of trees = 50 to analyze the atmospheric concentrations of the four gases and temperature changes.

2.4 Inspection standards

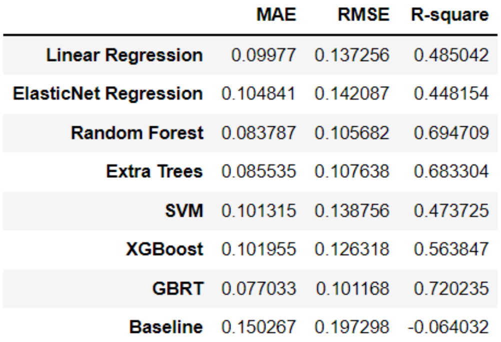
This article will use Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Coefficient of Determination (R-square) as standards to compare the ability of different models to predict temperature changes through the atmospheric concentration of four greenhouse gases.

$$MAE = \frac{\sum|(y_i - \hat{y}_i)|}{n}''$$

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}''}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}''$$

3. RESULT AND DISCUSSION



	MAE	RMSE	R-square
Linear Regression	0.09977	0.137256	0.485042
ElasticNet Regression	0.104841	0.142087	0.448154
Random Forest	0.083787	0.105682	0.694709
Extra Trees	0.085535	0.107638	0.683304
SVM	0.101315	0.138756	0.473725
XGBoost	0.101955	0.126318	0.563847
GBRT	0.077033	0.101168	0.720235
Baseline	0.150267	0.197298	-0.064032

Figure 4 MAE, RMSE, R-square result for different methods

As can be seen from figure 4, compared to Baseline, several machine learning methods have been optimized to a certain extent, resulting in smaller MAE and RMSE. Among them, the Gradient Boosting Regression Tree performs better. It can be checked more direct by the comprehensive bar charts:

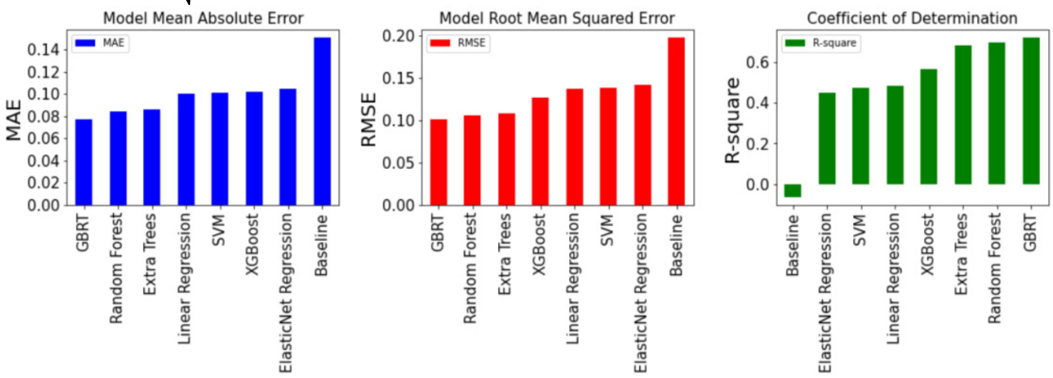


Figure 5 Comprehensive bar charts

From the bar chart of MAE (as shown in figure 5), it is not difficult to find that in the case of analyzing the difference between atmospheric concentration and temperature of greenhouse gases, GRBT has a smaller MAE, which means that the model has higher accuracy. As for the bar chart of RMSE, similar results were obtained. As for R-square, which represents the variation in the temperature deviation can be explained by the regression relation of the atmospheric concentration of four greenhouse gases, GRBT also performed the best. One thing to note here is that R-square of Baseline is less than 0. This is because r2_score is directly called, and the baseline leads to a worse fit than the horizontal line. Based on the inspection standards selected in this article and specific comparisons, it is a good choice to use Gradient Boosting Regression Tree to analyze the atmospheric concentration of the four gases and temperature changes.

Now that it turns out the most suitable model, further study of the temperature changes can be investigated by this method. It can not only predict the approximate temperature changes through the atmospheric concentration but also “achieve a temperature control of 1.5°C” in the Glasgow Climate Pact. It also can be used to impose specific restrictions on greenhouse gas emissions.

4. CONCLUSION

The theme of this article is to predict and limit the temperature changes by finding the most suitable machine learning model as to the Glasgow Climate Pact. The article gives a brief introduction to seven different machine learning methods and compares the ability of different machine learning methods to predict temperature changes using atmospheric concentrations of four greenhouse gases by the chosen inspection standards. After the analysis of this project and the specific situation of different machine learning models, it can be found that it is a better choice to choose Gradient Boosting Regression Tree as the model to predict temperature deviation from 1951-1980 mean in the 21st century. According to the R-square of GBRT, it can be seen that its predictive ability is also relatively good. So, when atmospheric concentrations of CO_2, CH_4, N_2O, and SF_6 are used to predict temperature changes, it is relative better to choose GBRT for analysis. Because the collected data is monthly data, and weathers are different in different seasons of one year, it is allowed to add labels to the data of different months to distinguish different months during data preprocessing. The advantage of this is that the performance of different machine learning models in each season or even month of the year can be analyzed in the final analysis, and a more accurate result can be obtained.

REFERENCES

[1] UNFCCC Glasgow Climate Pact, 2021. Part IV. Mitigation Article 20 - Article 22

[2] The six main greenhouse gases, Protocol, K. (1997). Kyoto protocol. UNFCCC Website. http://unfccc. int/kyoto_protocol/items/2830. php.

[3] Trend concentration data of CO_2, Global Monitoring Laboratory Website. Ed Dlugokencky, NOAA/GML: https://gml.noaa.gov/ccgg/trends/global.html

[4] Trend concentration data of CH_4, Global Monitoring Laboratory Website. Ed Dlugokencky, NOAA/GML: https://gml.noaa.gov/ccgg/trends_ch4/

[5] Trend concentration data of N_2O, Global Monitoring Laboratory Website. Ed Dlugokencky, NOAA/GML: https://gml.noaa.gov/ccgg/trends_n2o/

[6] Trend concentration data of SF_6, Global Monitoring Laboratory Website. Ed Dlugokencky, NOAA/GML: https://gml.noaa.gov/ccgg/trends_sf6/

[7] GISTEMP Team, 2021: GISS Surface Temperature Analysis (GISTEMP), version 4. NASA Goddard Institute for Space Studies. Dataset accessed 20YY-MM-DD at https://data.giss.nasa.gov/gistemp/.

[8] Lenssen, N., G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss, Improvements in the GISTEMP uncertainty model. J. Geophys. Res. Atmos., 124(12), (2019), 6307-6326, doi:10.1029/2018JD029522.

[9] Hansen, J., R. Ruedy, M. Sato, and K. Lo, Global surface temperature change. Rev. Geophys., 48(4) (2010) RG4004, doi:10.1029/2010RG000345.

[10] Lenssen, N., G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss, Improvements in the GISTEMP uncertainty model. J. Geophys. Res. Atmos., 124(12) (2019) 6307-6326, doi:10.1029/2018JD029522.

[11] Ogutu, J. O., Schulz-Streeck, T., & Piepho, H. P. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC proceedings, Vol. 6, No. 2 (2012) 1-6.

[12] Zou H, Hastie T: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society B. 67, (2005), 301-320. 10.1111/j.1467-9868.2005.00503.x.

[13] Cao Z F. Study on optimization of random forests algorithm[D]. Capital University of Economics and Business, 2014. DOI: CNKI:CDMD:1.1014.220587.

[14] John, V., Liu, Z., Guo, C., Mita, S., & Kidono, K. Real-time lane estimation using deep features and extra trees regression. In Image and Video Technology. Springer, Cham. 2015, pp. 721-733.

[15] Zhao, G, Tu, X, Wang, T. Drought prediction based on artificial neural network and support vector regression machine [J]. People's Pearl River, 42(4) (2021) 1-9.

[16] Support Vector Machine - Regression (SVR): http://www.saedsayad.com/ support_ vector_machine_reg.htm

[17] Huang, Y., Liu, Y., Li, C., & Wang, C. GBRTVis: online analysis of gradient boosting regression tree. Journal of Visualization, 22(1) (2019) 125-140.