Predicting Climate Model Outcomes Using Logistic Regression and Support Vector Machines

—

Lakota Nguyen
Dr. Lozinski; TA Dominique Stumbaugh
AOS C111/C204: Introduction to Machine Learning for Physical Sciences
December 8, 2023

Introduction

The scarcity of empirical observations calls for the need of climate models to better understand and predict the Earth's physical, chemical, and biological processes. In the presence of rapid, human-induced climate change, these models are crucial in simulating future climate trends. One example is the Community Climate System Model (CCSM4); it is used in Intergovernmental Panel on Climate Change (IPCC) reports. The CCSM4 is composed of 4 separate sub-models (atmosphere, ocean, land surface, and sea-ice) that work together to reproduce the Earth's climate system (NSF, 2020). These models are very complex, for they take into account equations of state, conservation laws, energy budgets, and biogeochemical cycles all across space (locations) and time (Lucas et al., 2013). As such, one faulty component or a small perturbation of any of the 4 sub-model parameters will thwart the entire climate model as a whole, sometimes causing the entire simulation to crash (Lucas et al., 2013).

Uncertainty quantifications are conducted to assess the capability of climate models. One such uncertainty quantification study records a series of simulation crashes within the Parallel Ocean Program (POP2), the ocean component of the CCSM4. Lucas et al. (2013) explains how 8.5% of CCSM4 simulation failures are due to the numerical values of the model parameters. Thus, a machine learning approach to the dataset could be implemented in order to predict the simulation outcomes under the given model parameter values. Doing so would ascertain which parameters are most influential to climate model success and failure, which could then provide insight into understanding and improve climate models (Lucas et al., 2013).

Methodology

The dataset was obtained from the UCI Machine Learning Repository. It documented the climate model simulation outcomes and the associated model parameters of the Parallel Ocean Program (POP2) within the Community Climate System Model (CCSM4). The dataset contains 3 studies with 180 runs each (540 total instances). In order to prepare the data for the machine learning models, the data was split into 4 groups. Studies 1-3 corresponded to the individual study number and Study 4 corresponded to the entire dataset (studies 1-3). The target variable was column 21, titled 'outcome,' which had discrete values of 0 and 1 (0 = failure, 1 = success). The feature variables were columns 3-20, which denoted numerical values of the climate model parameters normalized on the interval [0,1].

The nature of the dataset offered a supervised classification problem. It was supervised learning because the dataset contained baseline inputs (climate model parameters) and their desired outputs (climate model success or failure) with which an algorithm/computer could learn a general rule that mapped inputs to outputs (Bortnik). It was a classification problem since the goal was to take the given climate model parameter values (input) and map it to 2 discrete classes, success or failure (output). Therefore, classification models of logistic regression (LR) and support vector machines (SVM) were used.

4 logistic regression models were trained simultaneously using their respective study group. Then, each model's accuracy,

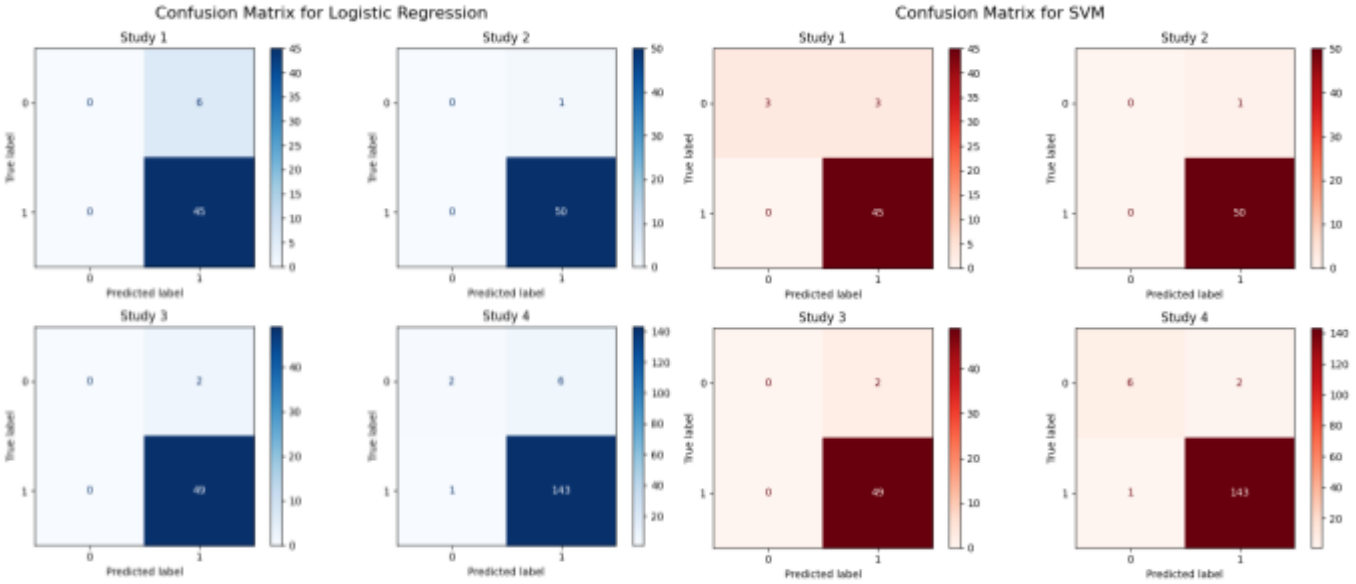

Figure 1: Comparisons of confusion matrices for logistic regression and linear support vector machines

confusion matrix, and ROC curve was calculated. Lastly, feature ranking using both recursive feature elimination (RFE) and absolute coefficient value (ACV) methods was conducted. This process was repeated using the support vector machine model.

As shown in Table 1, the linear kernel from the SVM model provided the greatest model accuracy. Therefore, the linear kernel was used for the SVM model instead of the sigmoid or rbf kernel.

Table 1: SVM kernel accuracy comparisons

SVM Kernel	Study 1	Study 2	Study 3	Study 4: entire data (studies 1-3)
Linear	0.9412	0.9804	0.9608	0.9803
Sigmoid	0.8824	0.9804	0.9608	0.9408
RBF	0.8824	0.9804	0.9608	0.9474

Table 2: Model accuracy comparisons

	Study 1	Study 2	Study 3	Study 4: entire data (studies 1-3)
Logistic Regression	0.8824	0.9804	0.9608	0.9539
SVM (kernel= 'Linear')	0.9412	0.9804	0.9608	0.9803

Results

Evaluation of model accuracy, confusion matrix, and ROC curves showed that support vector machines (SVM) were better at predicting climate model outcomes than logistic regression (LR). As displayed in the model accuracy comparisons in Table 2, the SVM model possessed greater accuracy values in every study. This signalled that the SVM model was better at correctly classifying the dataset than the LR model.

Comparisons of LR and SVM confusions matrices (Figure 1) further show that SVM was the better model. The confusion matrices for LR and SVM were the exact same for studies 2 and 3, but they differed slightly in studies 1 and 4. In study 1, although both LR and SVM had 45 instances of true positives, the LR confusion matrix had 6 instances of false positives and 0 instances of true negatives, while the SVM confusion matrix had 3 instances of false positives and 3 instances of true negatives. In study 4, LR contained 6 false positives and 2 true negatives whereas SVM contained 2 false positives and 6 true negatives.

In the context of the study, a false positive meant that the algorithm (LR, SVM) predicted climate model success even though the climate model failed in reality. A true negative meant that the algorithm correctly identified a climate model failure.

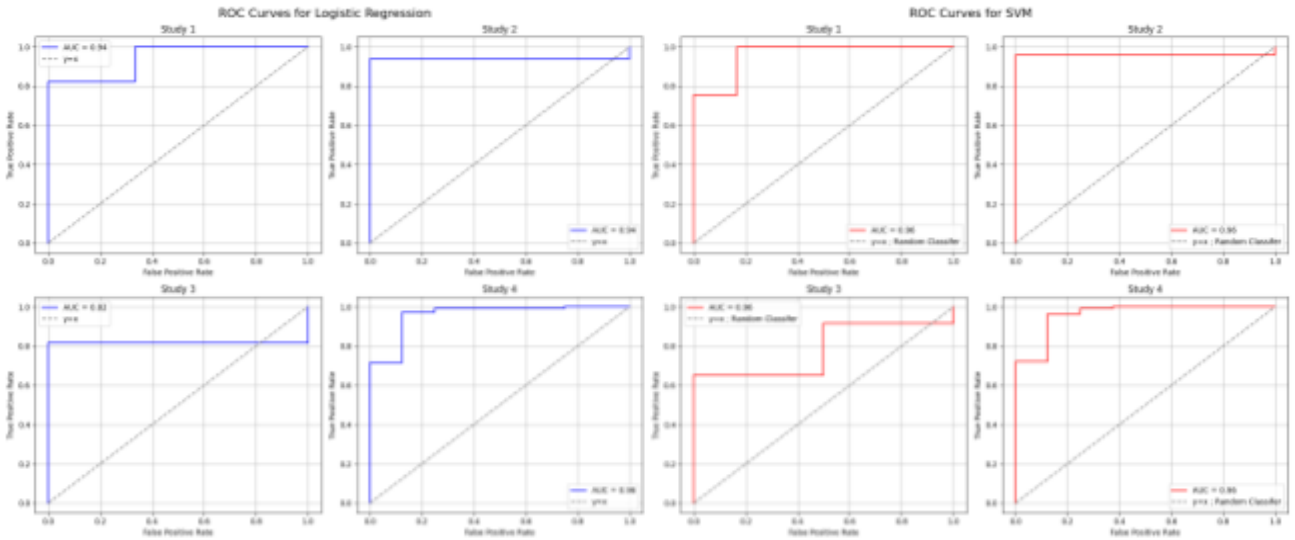

Figure 2: Comparison of ROC curves and AUC values for logistic regression and linear support vector machines

The SVM model was a better predictor since it gave a smaller number of false positives than the LR model. In other words, the SVM model made fewer mistakes in predicting the climate model simulation failures.

The receiver operating characteristic (ROC) curves of the LR and SVM models were almost identical (Figure 2). Both LR and SVM had the same area under the curve (AUC = 0.96) for study 4, implying that both models classified the overall data (3 studies combined) the same. The differences arose in studies 1-3, as the SVM model possessessed greater AUC values. This revealed that the SVM model had more precise classification in each individual study.

Tables 3 and 4 presented the top 3 features of the LR and SVM models using the recursive feature elimination (RFE) and absolute coefficient value (ACV) methods. With the exception of SCM-ACV-Study 3, all models and methods recorded `vconst_corr` as the most important feature, followed by `vconst_2`. This indicated that the climate model parameters of `vconst_corr` and `vconst_2` contributed the most to the climate models' outcome.

Figure 3 graphed the feature coefficient values of each model and study. In all studies and models, the top 2 features `vconst_corr` and `vconst_2` had negative coefficients. A negative coefficient signified a negative correlation between the feature and target variables: as the feature variable increased, the target variable outcome being 1 (climate model success) decreased. In essence, the variables `vconst_corr` and `vconst_2` impacted model failure the most. Notably, the variables of `bckgrnd_vdc1`, `vconst_4`, and `vconst_5` (which were ranked either #2 or #3) possessed positive coefficients. This conveyed that these variables were the most influential in predicting climate model success.

Method	Study 1	Study 2	Study 3	Study 4: entire data (studies 1-3)
RFE	vconst_corr vconst_2 vertical_decay_scale	vconst_corr vconst_2 vconst_4	vconst_corr vconst_2 bckgrnd_vdc1	vconst_corr vconst_2 convect_corr
ACV	vconst_corr vconst_2 vertical_decay_scale	vconst_corr vconst_2 **vconst_5**	vconst_corr vconst_2 **bckgrnd_vdc1**	vconst_corr vconst_2 convect_corr

Table 3: Logistic regression feature ranking- recursive feature elimination vs. absolute coefficient value; parameters influencing climate model success in bold

Method	Study 1	Study 2	Study 3	Study 4: entire data (studies 1-3)
RFE	vconst_corr vconst_2 vertical_decay_scale	vconst_corr vconst_2 vconst_5	vconst_corr vconst_2 bckgrnd_vdc1	vconst_corr vconst_2 bckgrnd_vdc1
ACV	vconst_corr vconst_2 vertical_decay_scale	vconst_corr vconst_2 **vconst_4**	vconst_corr **bckgrnd_vdc1** vconst_2	vconst_corr vconst_2 convect_corr

Table 4: Linear support vector machines feature ranking- recursive feature elimination vs. absolute coefficient value; parameters influencing climate model success in bold

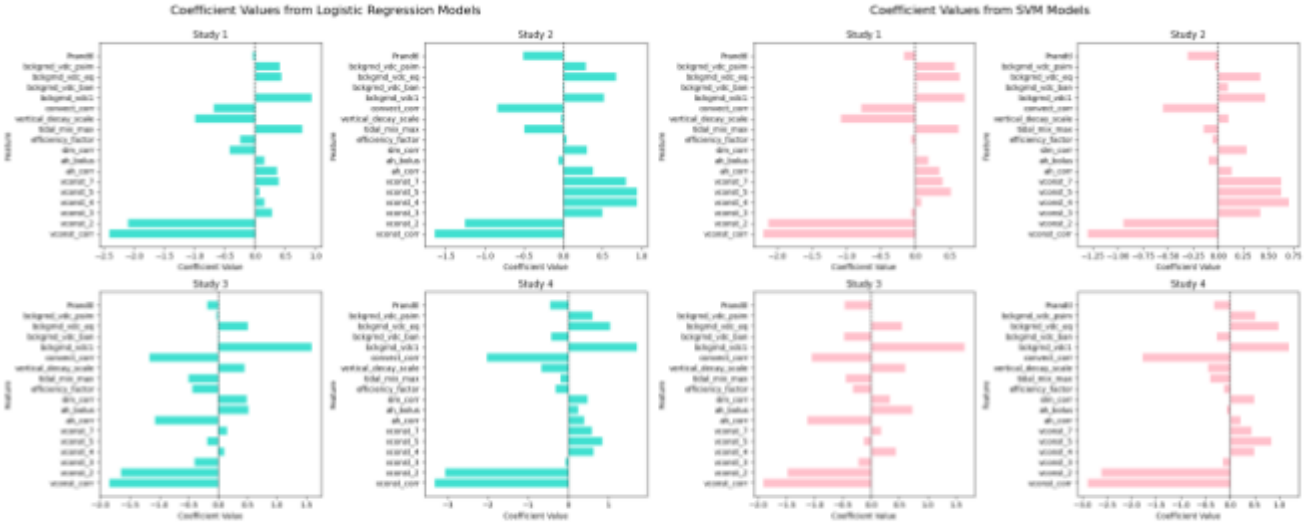

Figure 3: Comparison of logistic regression and linear support vector machines feature importances (coefficient values)

Discussion

Both linear regression (LR) and support vector machines (SVM) were adequate in predicting climate model simulation outcomes, for both had accuracies above 0.8 across all studies and had overall AUCs of 0.96. The SVM model was slightly more accurate than the LR model. This difference could be attributed to the nature of SVMs. The objective of SVMs was to maximize the margins between the decision boundary hyperplane and the support vectors, while the objective of LR was to maximize the likelihood of observed data under the sigmoid function (Sriashi0397). Thus, the SVM model was less sensitive to data outliers than the LR model, which led to better model performance and subsequent higher accuracies.

The top 2 features that influenced LR and SVM model outcome were `vconst_corr` and `vconst_2`, as determined by their highly negative coefficient values. Other notable features influencing climate model outcomes included `bckgrnd_vdc1` and `convect_corr`. `Vconst_corr`, `vconst_2`, and `convect_corr` had negative coefficient values, signalling that the features greatly influenced climate model failure. Conversely, `bckgrnd_vdc1` had positive coefficient values, signalling contribution to climate model success. These findings were in agreement with Lucas et al. (2013), for they determined that the probability of climate model failure increased with increasing values of `vconst_corr`, `vconst_2`, and `convect_corr`, and the probability of climate model failure decreased with increasing values of `bckgrnd_vdc1`.

Each feature variable was a parameterization of certain ocean dynamics within the POP2 model. The variables of `vconst_corr` and `vconst_2` captured horizontal ocean mixing under anisotropic viscosity (varying viscosity value depending on the direction of measurement)(Large et al., 2001). Lucas et al. (2013) explained how the upper bounds of these parameters are constrained by the linear diffusion stability requirement. They suggested that high values of these parameters triggered the upper limit and subsequently caused model failure (Lucas et al., 2013). Since these 2 variables contributed the most to failure prediction (outcome = 0), it is imperative for POP2 model developers to improve the parameterizations of horizontal ocean mixing in order to reduce overall climate model failure.

The variables of `bckgrnd_vdc1` and `convect_corr` corresponded to vertical convection and mixing of the ocean column under the K-Profile-Parameterization scheme (KPP) (Large et al., 1994). `Bckgrnd_vdc1` , with a positive coefficient, influenced the success prediction (outcome = 1). Thus, these findings expressed that `bckgrnd_vdc1` was an adequate parameterization and gave rise to climate model success. On the contrary, these findings exposed that `convect_corr`, with a negative coefficient, was an inadequate parameterization and must be improved. These contrasting influences of climate model outcomes from `bckgrnd_vdc1` and `convect_corr` showcased that the parameter group of vertical convection and mixing both helped and hurt model success.

The LR and SVM models used machine learning to classify climate model outcomes (success or failure) for given values of 18 feature parameters. They provided insight into which parameters impacted climate model failure and success the most. However, LR and SVM were limited in the correlation between the numerical values of 2 parameters and their influence on the climate model outcome. For instance, the model could not tell which combination of `const_corr` and `bckgrnd_vdc1` values caused model failure. Lucas et al. (2013) further explained how having `vconst_corr` and `bckgrnd_vdc1` in different parameter groups (horizontal mixing and vertical mixing) made attribution of model failure difficult and complicated. Determination of which combination of parameter values caused model failure required analysis beyond the scope of this course. Thus, future research pertaining to the relationships between model parameter values can be explored.

Conclusions

Taking a machine learning approach to the climate model success/failure dataset brought about classification predictions of outcomes and analysis of most impactful feature variables. Evaluation and comparison of the accuracies, confusion matrices, and ROC curves of the logistic regression (LR) and linear support vector machines (SVM) model showed that the SVM model was better at predicting climate model success and failure.

Both LR and SVM calculated that the variables of `vconst_corr` and `vconst_2` were most significant in predicting model failure, while the variable `bckgrnd_vdc1` contributed to predicting model success. Attribution of these features to their respective parameter groups will allow climate model developers to pinpoint which parameterizations of the ocean's processes need improvement.

The limitations of LR and SVM models highlight an opportunity for further investigation on the correlations between the numerical values of multiple parameters and their combined impact on model outcome.

This study provides insights to better understand the relationship between a singular model parameter and its impact on the climate model outcome. Through this, Model developers can improve these parameterization and such improve climate models.

Citations/Acknowledgements

Bortnik, J. (n.d.). Lecture 0.2: What is machine learning?. Bruinlearn. https://bruinlearn.ucla.edu/courses/167864/pages/lecture-0-dot-2-what-is-machine-learning?module_item_id=6238955.

Large, W. et al. (1994). Oceanic vertical mixing: A review and a model with a nonlocal boundary-layer parameterization, Reviews Of Geophysics, 32, 363-403. https://doi.org/10.1029/94RG01872.

Large, W. et al. (2001). Equatorial circulation of a global ocean climate model with aniostropic horizontal viscosity. Journal Of Physical Oceanography, 31, 518-536. http://n2t.net/ark:/85065/d7fb53g1.

Lucas, D. et al. (2013). Failure analysis of parameter-induced simulation crashes in climate models, Geosci. Model Dev., 6, 1157–1171. https://doi.org/10.5194/gmd-6-1157-2013.

NOAA. (n.d.). Climate modeling. GFDL. https://www.gfdl.noaa.gov/climate-modeling/#:~:text=Climate%20models%20are%20important%20tools,or%20a%20combination%20of%20both.

NSF. (2020). CESM models. CESM Models | CCSM4.0 Public Release. https://www2.cesm.ucar.edu/models/ccsm4.0/

Sriashi0397. (n.d.). Differentiate between Support Vector Machine and Logistic Regression, geeksforgeeks. https://www.geeksforgeeks.org/differeNtiate-between-support-vector-machine-and-logistic-regression/.