Medicon Engineering Themes
Volume 2 Issue 6 June 2022
Research Article

Predicting Global Average Land Temperatures Using Microsoft Azure Machine Learning Studio

Sagar Tewari[1], Sarthak Agarwal[2] and Madhulika Bhatia[3]*

[1]B.Tech, Artificial Intelligence, Amity University, India

[2]B.Tech ,Artificial Intelligence, Amity University, India

[3]Associate Professor, Computer Science and Engineering, Amity University, India

***Corresponding Author:** Madhulika Bhatia, Associate Professor, Computer Science and Engineering, Amity University, India.

Received: April 29, 2022; **Published:** May 30, 2022

Abstract

To accurately calculate the global average temperature has proven to be an arduous task since the 19th century, the major reasons are maintaining accurate records of the same locations over a lengthy period which has strained the meteorologists, especially in places located remotely, like mountains or deserts. This is the reason meteorologists use global averages which generally span over 3 decades to give perspective and context to information.

We decided to use the Linear Regression Model to achieve this task at hand. Linear regression is an algorithm which is used for determining the relationship between an dependent variable and some independent variables which are also known as scalar response explanatory variables respectively. The relationship between dependent and independent variables are calculated using predictor functions which are linear in nature whose parameters which are not known are found out through the means of the data provided. Hence, it is paramount that we give our model a clean and digestible data so that it learns well and predicts accurately. It's equally important that we maintain the dimensions of the data and be careful that it is not too high or not too low, it should be just right. We also employed the Pearson's Correlation method for Feature Selection. After which we also tuned the hyperparameters among other things to gain the most suitable parameters for the model. Finally, we compared our results from normal model generation to hyperparameter tuned model.

Keywords: AZURE; Machine Learning; Linear Regression; Pearson's Correlation Coefficient

Abbreviations

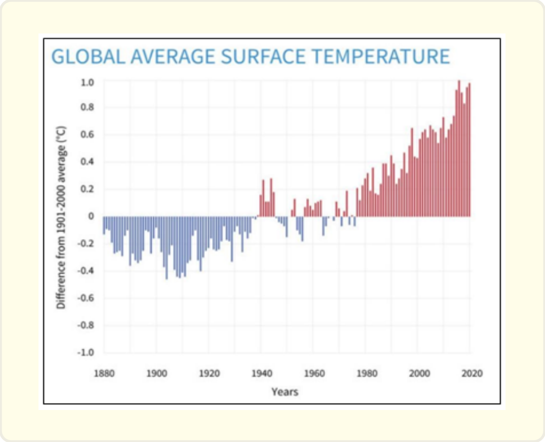
ML	- Machine Learning
MAE	- Means the total error
RMSE	- Root mean square error
RAE	- Relative absolute error
RSE	- Relative squared error

Introduction

Looking back at the history of Earth's temperature we see that it has increased by approximately 0.08°C every decade all the way since the 80's, also the factor with which global warming in the past 40 years has multiplied is more than 2 i.e., almost 0.18°C per decade since 1981.

Citation: Madhulika Bhatia., et al. "Predicting Global Average Land Temperatures Using Microsoft Azure Machine Learning Studio". Medicon Engineering Themes 2.6 (2022): 66-76.

A rise or drop of a couple of degrees may seem insignificant at first especially when we look at the changes to weather that we have a look at daily, but it is paramount to note that changes to the ocean and land temperatures play a much more significant and dire consequence for the entire world. For example, a few degrees drop during the Middle Ages is what paved the path to what came to be known as the 'Little Ice Age', which in turn caused the winters in the Northern Hemisphere to be much colder for over six centuries.



Materials and Methods

AZURE ML Studio

Microsoft's Azure ML Studio is one of central points of contact when it comes to Machine Learning Computation using cloud services. It is the heir apparent to the Microsoft Machine Learning Studio Classic retiring in 2024. Over the years it has seen many expansions like an increase in the number of features and possibilities.

Some of its key highlights include:

- Ability to building and training models rapidly
- By utilizing the development experience that of a studio we can access best-in-class support for libraries and open-source frameworks.
- Operationalize at scale
- Deploying models with the simple push of a button, managing, and governing them using MLOPs.
- Delivering responsible solutions to the problems
- Understanding and protecting data and improving model.
- Innovating on a secure and sturdy hybrid platform Running ML workloads anywhere in the world with a built-in security, compliance and governance.

Proposed Methodology

Opening the ML Studio

1. Open web browser (google chrome\explore).
2. And go to this link and login your account
3. First go to https://studio.azureml.netor simply google Microsoft Machine Learning Studio Classic.

Upload Dataset

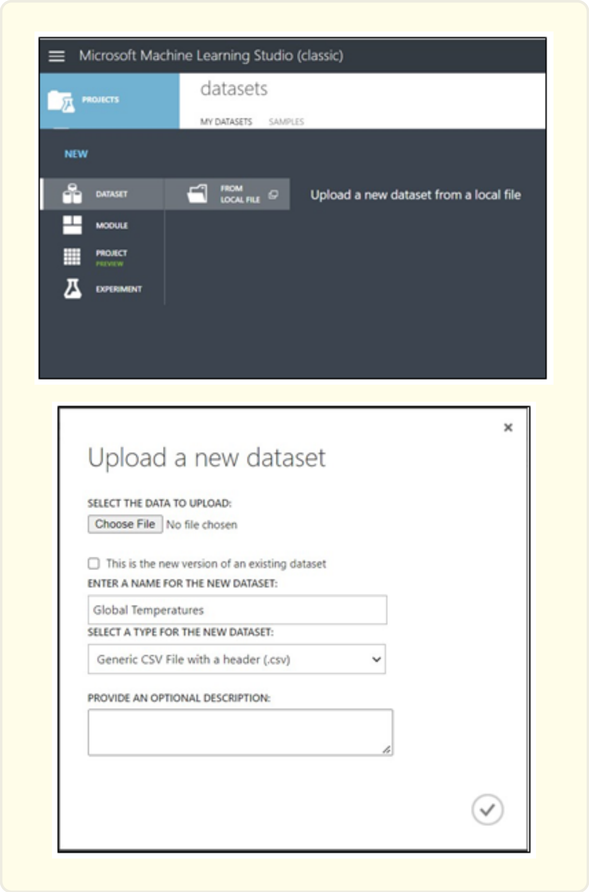
There can be more than one sources from which one can upload the data set.

Citation: Madhulika Bhatia., et al. "Predicting Global Average Land Temperatures Using Microsoft Azure Machine Learning Studio". Medicon Engineering Themes 2.6 (2022): 66-76.

Steps to import a dataset to ML studio

- Select the NEW button situated at the bottom of the page and create experiment, then blank experiment.
- Then to Upload dataset Click on NEW Tab
- Click DATASET option.
- Upload dataset from the local machine or by using the dataset URL.

We Uploaded GlobalTemperatures.csv data set from our local system.



Upload a new dataset

SELECT THE DATA TO UPLOAD:

No file chosen

□ This is the new version of an existing dataset

ENTER A NAME FOR THE NEW DATASET:

Global Temperatures

SELECT A TYPE FOR THE NEW DATASET:

Generic CSV File with a header (.csv)

PROVIDE AN OPTIONAL DESCRIPTION:

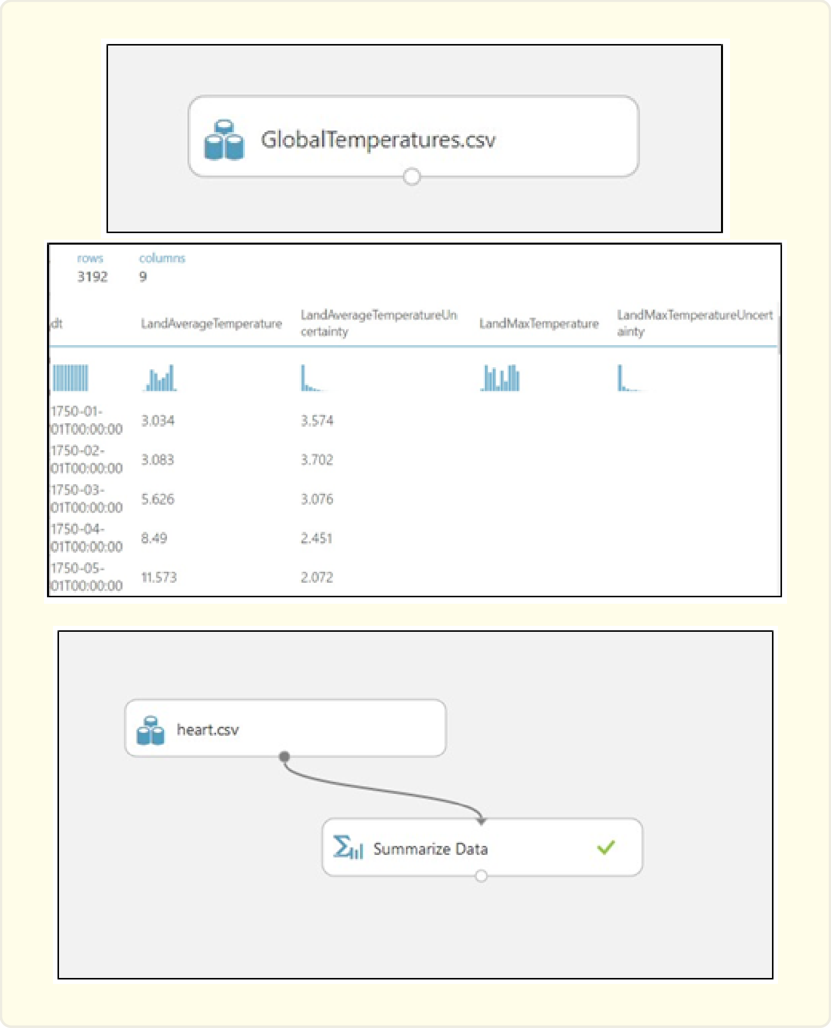
Exploring the Dataset

After importing you should explore the dataset to get valuable insights.

This can be done by using:

- Summarize data module
- Right clicking on the dataset and selecting the visualize option.

These modules will show some basic statics of each column like mean, mode, unique count, standard deviation.

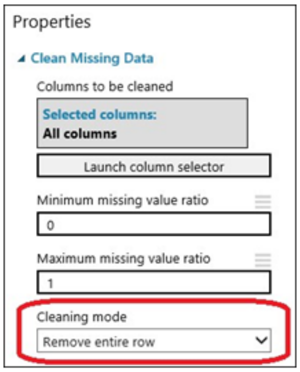


rows 3192 columns 9

dt	LandAverageTemperature	LandAverageTemperatureUncertainty	LandMaxTemperature	LandMaxTemperatureUncertainty
1750-01-01T00:00:00	3.034	3.574		
1750-02-01T00:00:00	3.083	3.702		
1750-03-01T00:00:00	5.626	3.076		
1750-04-01T00:00:00	8.49	2.451		
1750-05-01T00:00:00	11.573	2.072		

Data Preprocessing

- Before analyzing data, we need to clean the dataset or perform some preprocessing techniques on it. Missing values need to be dealt with by deleting rows, filling them etc.
- To get an accurate analysis of the data, the above step should be done carefully and properly.
- For the missing values, we deleted the entire row.
- Then we selected the required columns using the select column module.
- After that we normalized all numeric values of the selected columns and changed it to common scale. We shrank all numeric values between 0-1.



Properties

Clean Missing Data

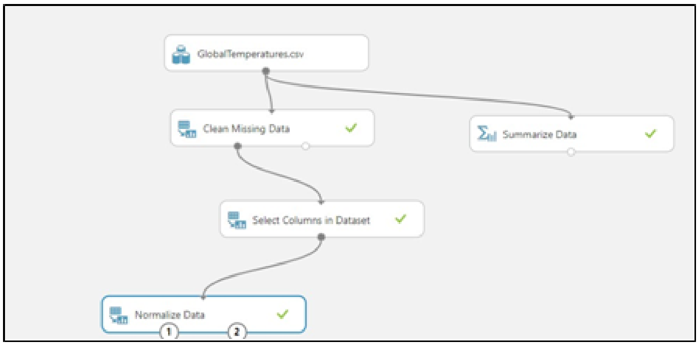
Columns to be cleaned

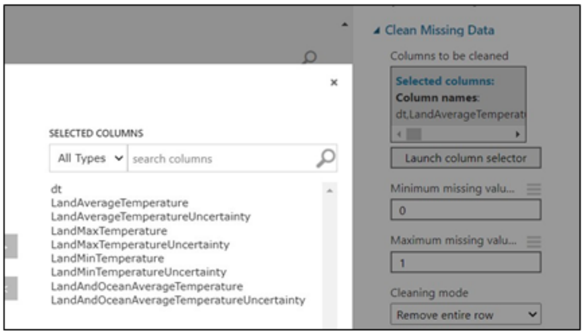
Selected columns:
All columns

Minimum missing value ratio: 0

Maximum missing value ratio: 1

Cleaning mode: Remove entire row





SELECTED COLUMNS

dt
LandAverageTemperature
LandAverageTemperatureUncertainty
LandMaxTemperature
LandMaxTemperatureUncertainty
LandMinTemperature
LandMinTemperatureUncertainty
LandAndOceanAverageTemperature
LandAndOceanAverageTemperatureUncertainty

Clean Missing Data

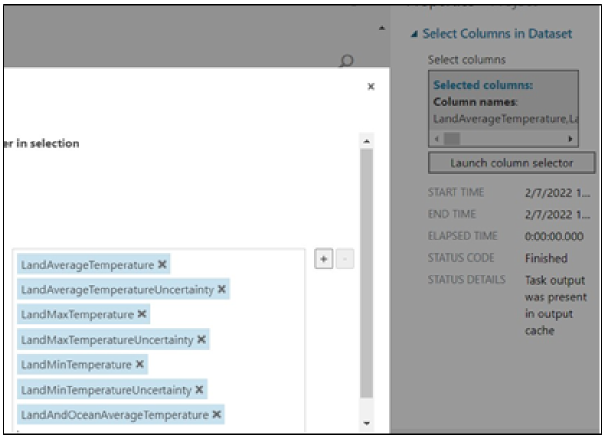
Columns to be cleaned

Selected columns:
Column names:
dt,LandAverageTemperat

Minimum missing valu... 0

Maximum missing valu... 1

Cleaning mode: Remove entire row



er in selection

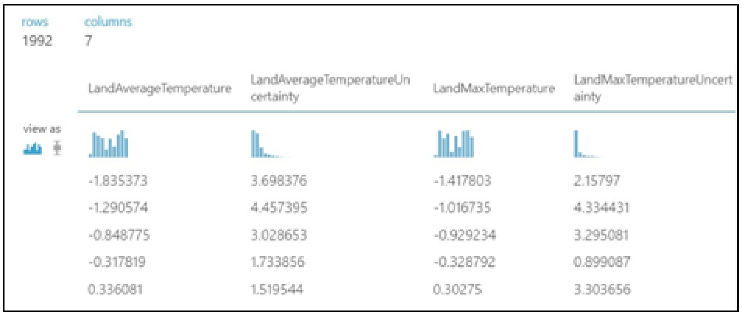
LandAverageTemperature ✕
LandAverageTemperatureUncertainty ✕
LandMaxTemperature ✕
LandMaxTemperatureUncertainty ✕
LandMinTemperature ✕
LandMinTemperatureUncertainty ✕
LandAndOceanAverageTemperature ✕

Select Columns in Dataset

Select columns

Selected columns:
Column names:
LandAverageTemperature,La

START TIME	2/7/2022 1...
END TIME	2/7/2022 1...
ELAPSED TIME	0:00:00.000
STATUS CODE	Finished
STATUS DETAILS	Task output was present in output cache



rows 1992 columns 7

view as	LandAverageTemperature	LandAverageTemperatureUncertainty	LandMaxTemperature	LandMaxTemperatureUncertainty
	-1.835373	3.698376	-1.417803	2.15797
	-1.290574	4.457395	-1.016735	4.334431
	-0.848775	3.028653	-0.929234	3.295081
	-0.317819	1.733856	-0.328792	0.899087
	0.336081	1.519544	0.30275	3.303656

Feature Selection (Filter-Based)

- Filter-Based feature selection uses the selected metric to find non-essential attributes then filter the missing columns in your model.
- Select one mathematical value that fits your data, and it will compute the score for each feature column.
- Feature score of each column is returned and are ranked by it.
- To improve the efficiency and accuracy of model select a suitable feature for the model.
- Columns with low score should not be selected or should be left out and only high score columns should be used in the analysis.
- We used Pearson's correlation for feature selection in our dataset.
- Pearson's correlation coefficient which gives back a value that signifies the correlation strength between any two variables. It is sometimes referred as r value in statistical model.
- Change in scale of variables doesn't affect the coefficient.

Pearson's correlation coefficient Formula

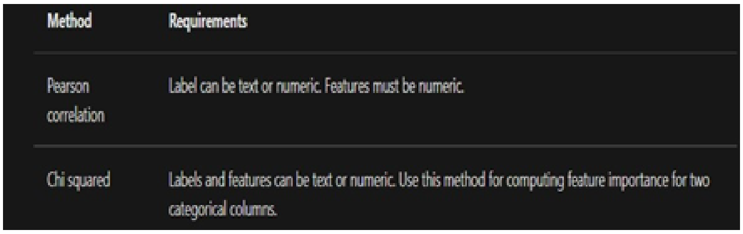
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

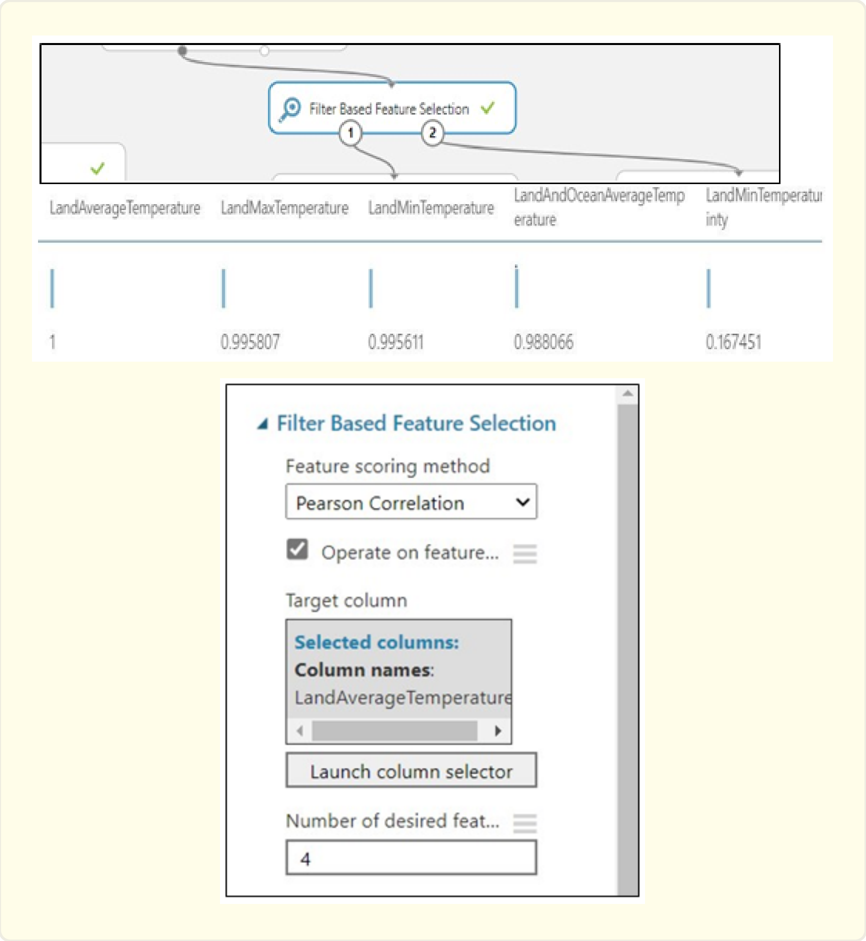
r = Pearson Correlation Coefficient

x_i = x variable samples $\quad$ y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\quad$ $\bar{y}$ =mean of values in y variable



Method	Requirements
Pearson correlation	Label can be text or numeric. Features must be numeric.
Chi squared	Labels and features can be text or numeric. Use this method for computing feature importance for two categorical columns.



LandAverageTemperature	LandMaxTemperature	LandMinTemperature	LandAndOceanAverageTemperature	LandMinTemperatur inty
1	0.995807	0.995611	0.988066	0.167451

◢ Filter Based Feature Selection

Feature scoring method

Pearson Correlation

☑ Operate on feature...

Target column

Selected columns:
Column names:
LandAverageTemperature

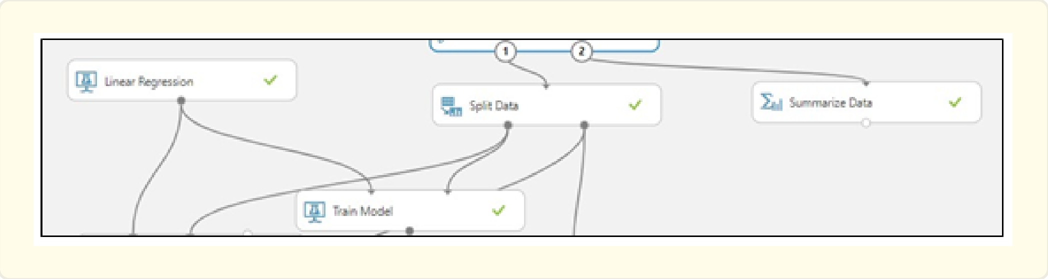
Launch column selector

Number of desired feat...

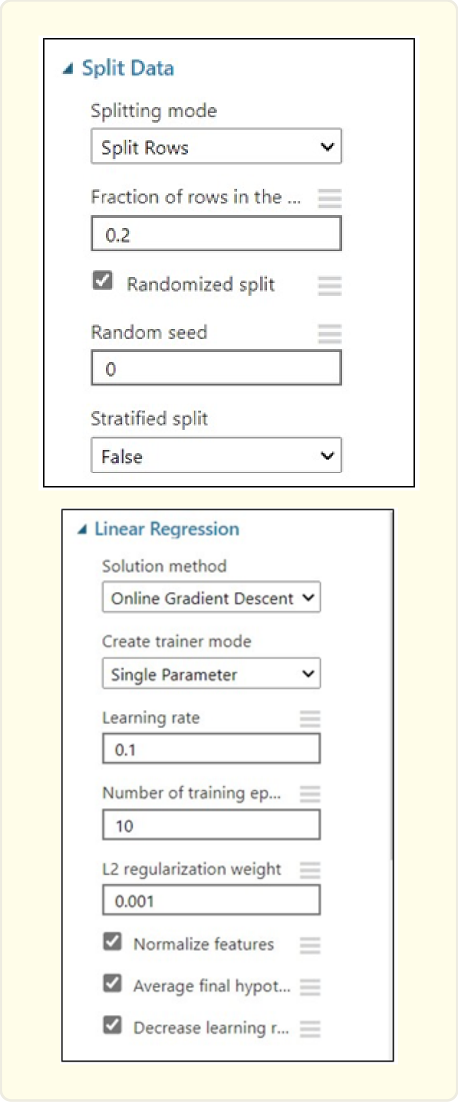
4

Splitting Dataset and Training Model

- Split dataset by dragging the module into the canvas and making the connections.
- Open the module and set the Fraction of rows to 0.8.
- This means 80% of the data will be used in training the model and 20% to test the model.
- After successfully running the experiment the model, the data would get split randomly and 80% will be used for training the model.
- Then we drag the train model module and selected an algorithm which we will used to predict.
- We used Online gradient decent learning regression for predicting values with a learning rate of 0.1%.
- To optimize the parameters, we used L2 regularization.



Citation: Madhulika Bhatia., et al. "Predicting Global Average Land Temperatures Using Microsoft Azure Machine Learning Studio". Medicon Engineering Themes 2.6 (2022): 66-76.



Split Data

Splitting mode: Split Rows

Fraction of rows in the ...: 0.2

☑ Randomized split

Random seed: 0

Stratified split: False

Linear Regression

Solution method: Online Gradient Descent

Create trainer mode: Single Parameter

Learning rate: 0.1

Number of training ep...: 10

L2 regularization weight: 0.001

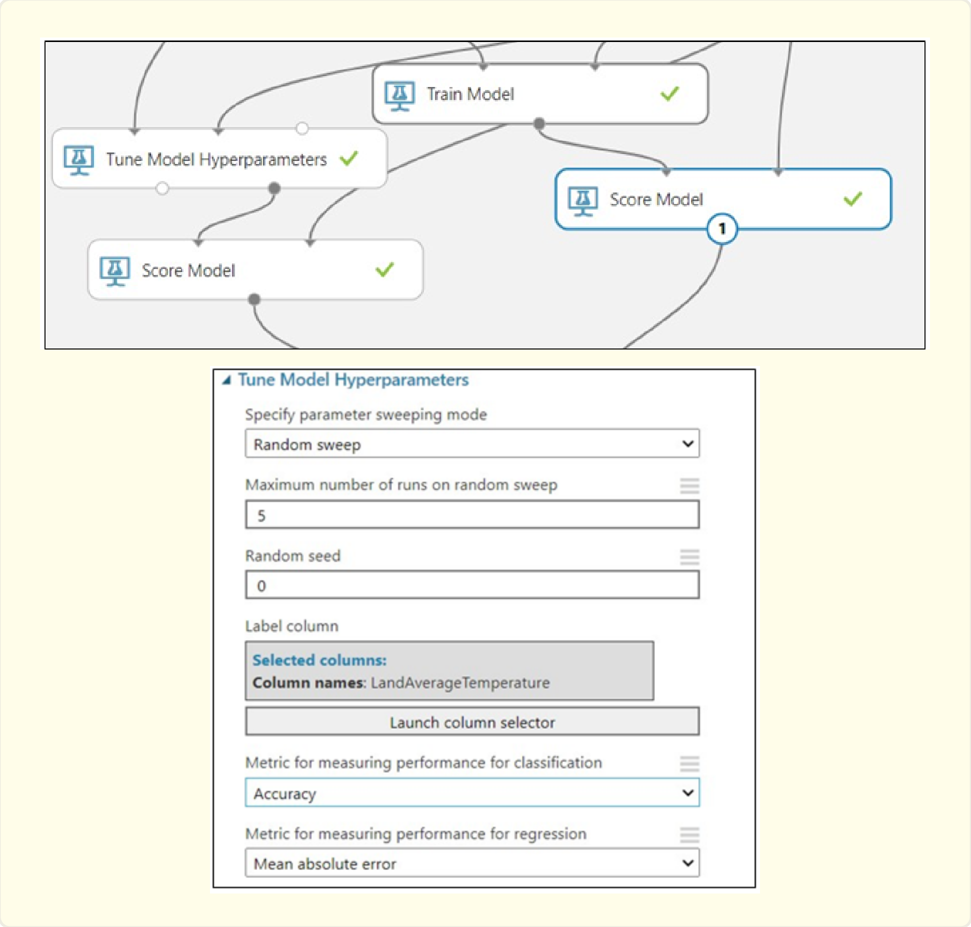
☑ Normalize features

☑ Average final hypot...

☑ Decrease learning r...

Hyperparameter Tuning

- Hyperparameters are parameters which can be changed to control the model training process. For example- In CNN the number of neurons for each layer, activation function, etc.
- should be decided before.
- Model performance is linearly dependent on hyperparameters.
- Towards the right-hand side of the experiment, we can see the properties panel and select the "Tune Model Hyperparameters "module.
- In the below picture, first we have chosen the "Random Sweep" option which performs a given number of iterations by randomly choosing the parameters' values. So, this component will try out various parameters' value randomly and train the model based on them.



Tune Model Hyperparameters

Specify parameter sweeping mode
Random sweep

Maximum number of runs on random sweep
5

Random seed
0

Label column
Selected columns:
Column names: LandAverageTemperature

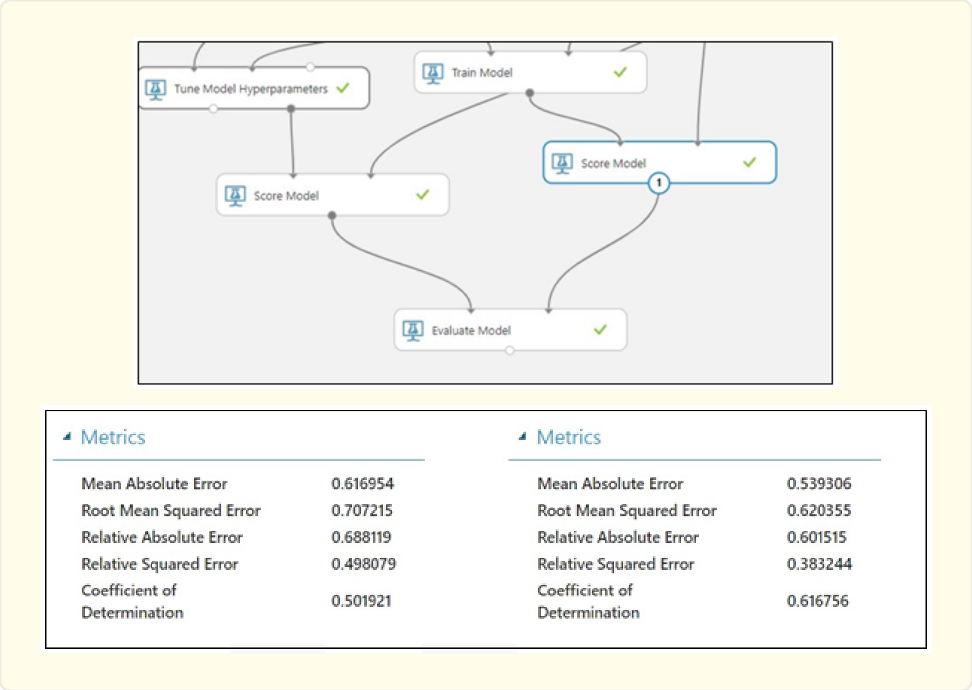
Launch column selector

Metric for measuring performance for classification
Accuracy

Metric for measuring performance for regression
Mean absolute error

Evaluating Model Before and After Hyperparameter Tuning

- In regression model an error matrix is returned to check the error rate. The model performs well if the difference between the current and estimated value is less.
- Residual pattern can also give lot of information about the potential bias and variance in the model.
- • Following metrics are returned-o Means the total error (MAE) o Root mean square error (RMSE).o Relative absolute error (RAE) o Relative squared error (RSE)
- Coefficient of determination normally referred as R^2
- We observed the prediction error decrease significantly after hyper tuning the model.



Metrics

Mean Absolute Error	0.616954
Root Mean Squared Error	0.707215
Relative Absolute Error	0.688119
Relative Squared Error	0.498079
Coefficient of Determination	0.501921

Metrics

Mean Absolute Error	0.539306
Root Mean Squared Error	0.620355
Relative Absolute Error	0.601515
Relative Squared Error	0.383244
Coefficient of Determination	0.616756

Results and Discussion

The given data was fit correctly and to some great extent as seen by the coefficient of determination. The tuning of hyperparameters greatly improved the efficiency of the model (almost 20%). Linear Regression turned out to be a good algorithm to predict the average land and ocean temperature.

Conclusion

It can be concluded that the average land and ocean temperature has seen an increase by a factor of 2 during the last couple of decades which is almost twice the couple of decades before them. The rise in temperature is alarming in nature and steps to counteract it should be taken hastily.

Acknowledgements

We would also like to show our gratitude to the Dr. Madhulika Bhatia, Professor, Computer Science and Technology for sharing their pearls of wisdom with us during the course of this research. She provided insight and expertise that greatly assisted the research.

Conflict of interest

None.

References

1. Min M., et al. “Estimating summertime precipitation from Himawari-8 and global forecast system based on machine learning”. IEEE transactions on geoscience and remote sensing 57.5 (2018): 2557-2570.
2. Yeh C., et al. “Using publicly available satellite imagery and deep learning to understand economic well-being in Africa”. Nature communications 11.1 (2020): 1-11.
3. Yuan Q., et al. “Deep learning in environmental remote sensing: Achievements and challenges”. Remote Sensing of Environment 241 (2020): 111716.

4. Maimaitijiang M., et al. "Soybean yield prediction from UAV using multimodal data fusion and deep learning". Remote sensing of environment 237 (2020): 111599.

Volume 2 Issue 6 June 2022