Project Report Update - Models

Team 74

Regarding the business question ”Is there a relationship between the issuance of carbon credits and greenhouse gas emissions?” we could relate the total credits (issued and retired) by country with the total CO2 emitted by the country each year. As a result, we have a time series from 1996 until 2018 with these variables.

	country	year	credits_issued	credits_retired	registry_issued_credits	credits_remaining	CO2_emitted
0	Argentina	1996	0	0	0	0	126560000.0
1	Argentina	1997	0	0	0	0	127320000.0
2	Argentina	1998	0	0	0	0	133170000.0

With the previous dataset, we could create a categorization model (possibly a machine learning regression model) to obtain the amount of CO2 reduced (or increased) by year determined by the number of credits issued or retired the previous year, not only for the years available in the dataset but for the future.

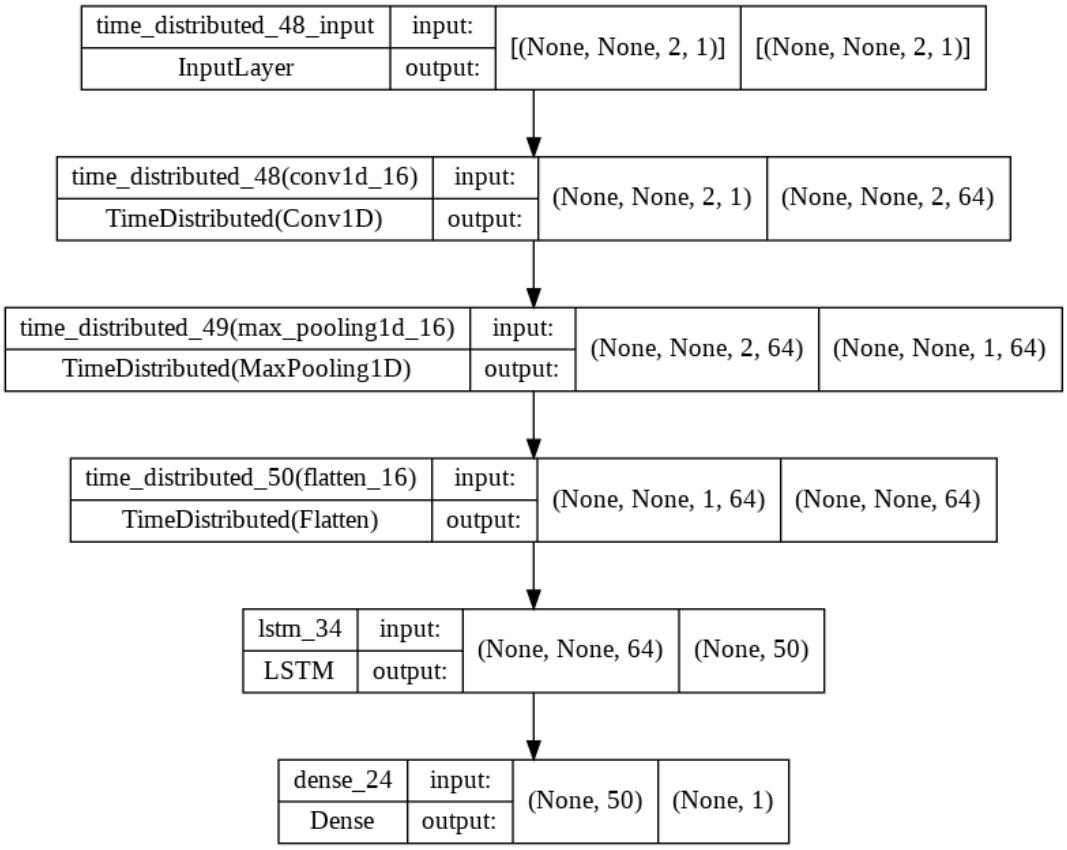
We start creating a projection model for the variables involved:

Projection CO2_emitted: Option 1 CNN LSTM

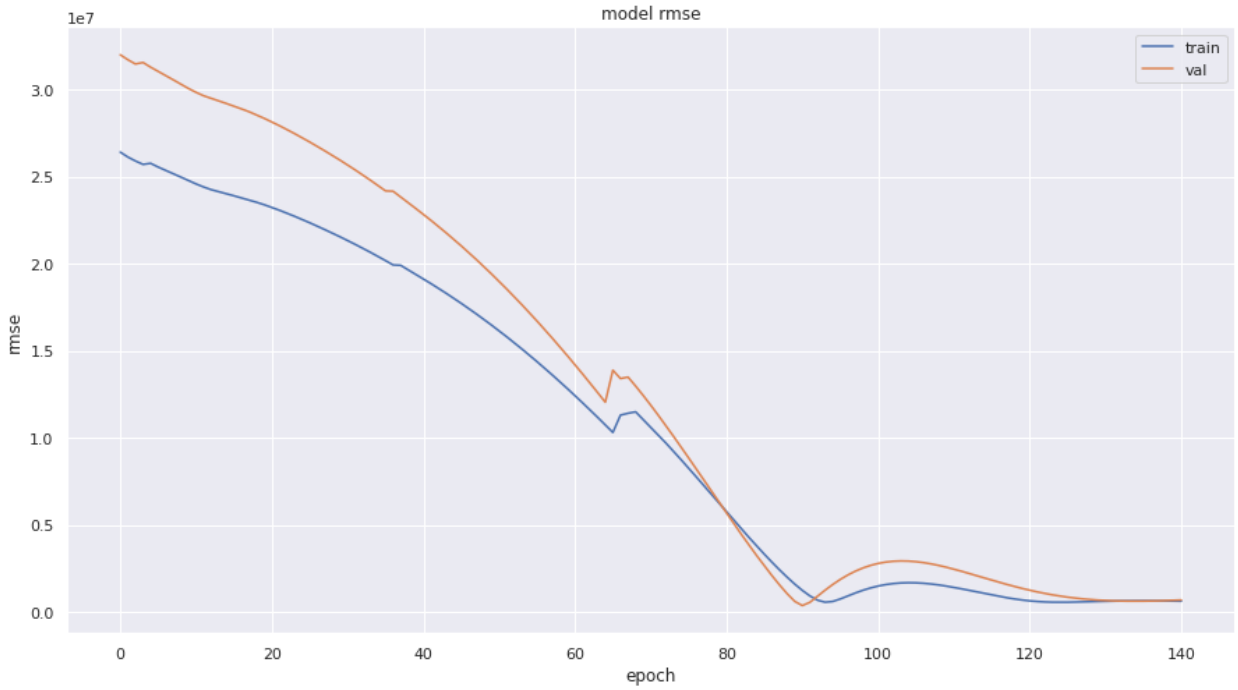
The first proposed model is a Convolutional Neural Network combined with a Long Short Term Memory model to predict the total CO2 emitted yearly.

Why the combination of a CNN with an LSTM? The LSTM is very useful with the seq2seq kind of data (as text or time series) and the CNN adds a component of feature extraction (mostly used for image data).

Following is the network structure:

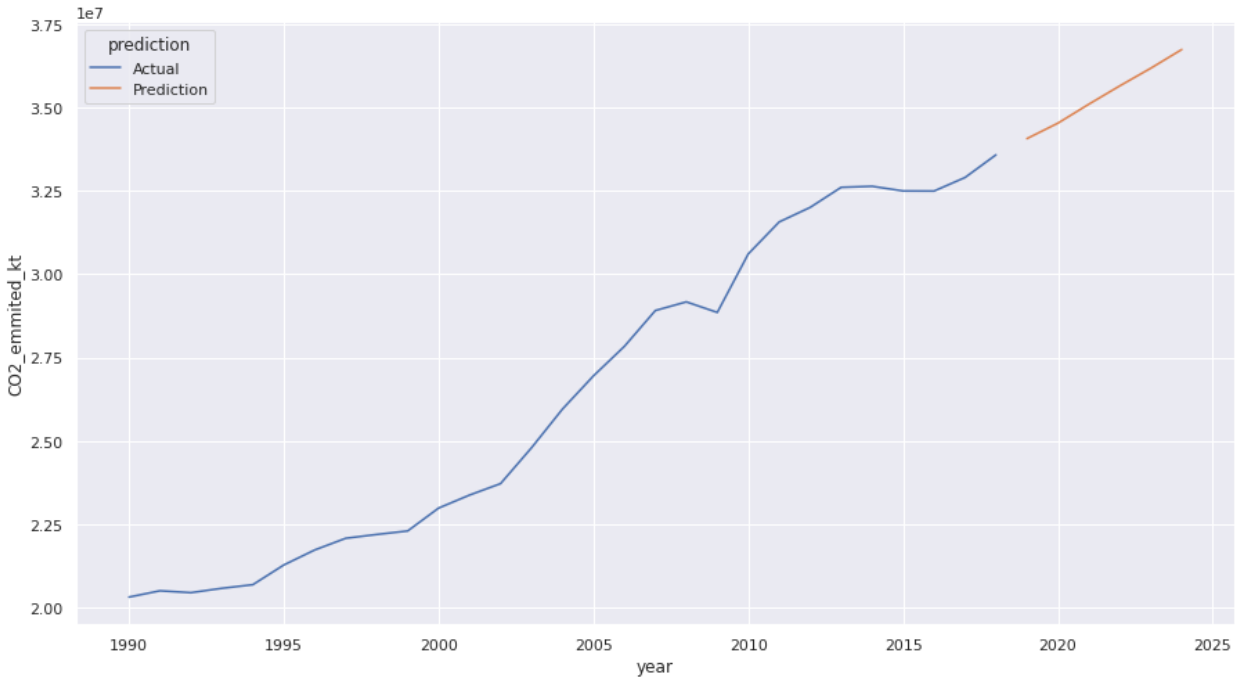


We train the model with a sequence of 4 periods, a subsequence of 2 periods, and 1 feature (CO2 emitted), we use the Keras sequential model adding each layer. The parameters used were the Adam optimizer, the MSE loss function, the RMSE metric to compare with other models, and a validation split of 10% (Due to the reduced amount of data available).



We use an early stop call to avoid overfitting the model with a result of 140 epochs to train, with an RSME of 629.204,38.

We keep the model and proceed to make a forecast for the next periods:



Projection CO2_emitted: Option 2 ARIMA (1, 1, 1)

We built an ARIMA model to compare the results. The best model splitting the data with a test set of 10% is an ARIMA (1,1,1). The RMSE is 323.180,94 (beating the CNN LSTM model), and the general results are the following:

ARIMA Model Results

Dep. Variable:	D2.CO2_emmited_kt	No. Observations:	24
Model:	ARIMA(1, 2, 1)	Log Likelihood	-348.617
Method:	css-mle	S.D. of innovations	465487.594
Date:	Thu, 16 Jun 2022	AIC	705.233
Time:	02:00:28	BIC	709.946
Sample:	2	HQIC	706.483

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.532e+04	1.71e+04	0.898	0.379	-1.81e+04	4.88e+04
ar.L1.D2.CO2_emmited_kt	0.2736	0.217	1.260	0.222	-0.152	0.699
ma.L1.D2.CO2_emmited_kt	-0.9999	0.166	-6.030	0.000	-1.325	-0.675

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	3.6551	+0.0000j	3.6551	0.0000
MA.1	1.0001	+0.0000j	1.0001	0.0000