Estimating prevalence of undernourishment using conflict, climate, and economic data

PoU

Time Range: 2001-2020
Number of Countries: 155
Data Type: continuous
Factors: Climate, Economy, Conflict
Cross Validation: Time Series Split (sklearn)

Fig. 1: Correlation matrix

	GDP (MUSD)	GNI (MUSD)	Food Prod. Index	Temperature	Precipitation	NDVI	Casualties	Undernourishment (%)
GDP (MUSD)	1	1	0.11	-0.22	0.26	-0.025	0.0035	-0.25
GNI (MUSD)	1	1	0.11	-0.22	0.25	-0.026	0.0023	-0.25
Food Prod. Index	0.11	0.11	1	-0.053	-0.083	-0.17	0.021	-0.21
Temperature	-0.22	-0.22	-0.053	1	0.1	0.27	0.026	0.41
Precipitation	0.26	0.25	-0.083	0.1	1	0.18	0.035	0.025
NDVI	-0.025	-0.026	-0.17	0.27	0.18	1	-0.1	0.17
Casualties	0.0035	0.0023	0.021	0.026	0.035	-0.1	1	0.13
Undernourishment (%)	-0.25	-0.25	-0.21	0.41	0.025	0.17	0.13	1

Highlight

A forecast for the Prevalence of Undernourishment (PoU) is provided for the year 2020 using a random forest regressor (R^2=0.78) that is based on conflict, climate, and economic data.

Primary Author
Nils Marthiensen

Contributing Authors
Neelesh Bhalla Jerome Chang

Background: "Zero Hunger" is the second Sustainable Development Goal (SDG) of the United Nations (UN) [1]. One indicator for this SDG Goal is the PoU, defined [2] as an estimate of the proportion (%) of the population whose habitual food consumption is insufficient to provide the dietary energy levels that are required to maintain a normal active and healthy life. In a 2021 report on world hunger, the Food & Agriculture Organization (FAO) pinpoints three major factors contributing to PoU - conflict, economic shocks and weather extremes [3]. In this work, we collect data on these factors to generate yearly country-level PoU forecasts.

Results: The PoU data for different years is not independent and identically distributed. A five-fold time series split was used to cross validate the findings. A range of different models were exploited and it was found that random forest regressor performed the best [Figure 3] with an R^2-value of 0.80.

The final dataset had 18 years (X = 2001-18; y = 2002-19) worth of independent variable data for 155 countries [Figure 4]. With the random forest regressor model, predictions were made with a root mean squared error of 5.65 and a R^2-value of 0.78 [Figure 5]. However there was an observed overfitting (bias) as the R^2-value on the training data amounted to 0.98.

Constructing dataset:

Different datasets are combined to construct a new, unique dataset that is tailored to this specific forecasting problem.

For **conflicts**, casualties corresponding to events of organized violence [Data 1] were considered. Only the recorded incidents are included, and they may vary from estimates of real total casualties, especially for wars. For missing fields, 0 casualties were assumed for a given country and year.

For **weather**, the total precipitation per year, the average temperature [Data 2] and the Normalized Difference Vegetation Index (NDVI) [Data 3] were considered. NDVI [4] is an indicator for vegetation density. The yearly temperature and precipitation data is not sensitive to seasonal variations.

For **economic** data the Gross Domestic Product (GDP), the Gross National Income (GNI) and the Food Production Index (FPI) were considered [Data 4]. FPI and the GDP were excluded from the final model. GDP has a correlation of 1 to GNI [Figure 1], however, the 'feature importance' of the model classified it as much less relevant for the output. FPI was excluded as it ranked lowest [Figure 2].

Future work: Our work can be extended in various directions:- 1) granular datasets - days/months instead of years; 2) population, drought, flood data as variables; and 3) stronger neural network architectures.

References

1. UN Sustainable Development Goals
2. PoU-definition
3. Main drivers for food insecurity
4. NDVI-definition

Protocols

1. Source code
2. Visualization

Data sets

1. Global Violent Conflicts-data
2. Precipitation and Temperature-data
3. NDVI-data
4. PoU, GDP, GNI, FPI-data

Fig. 1: Correlation matrix

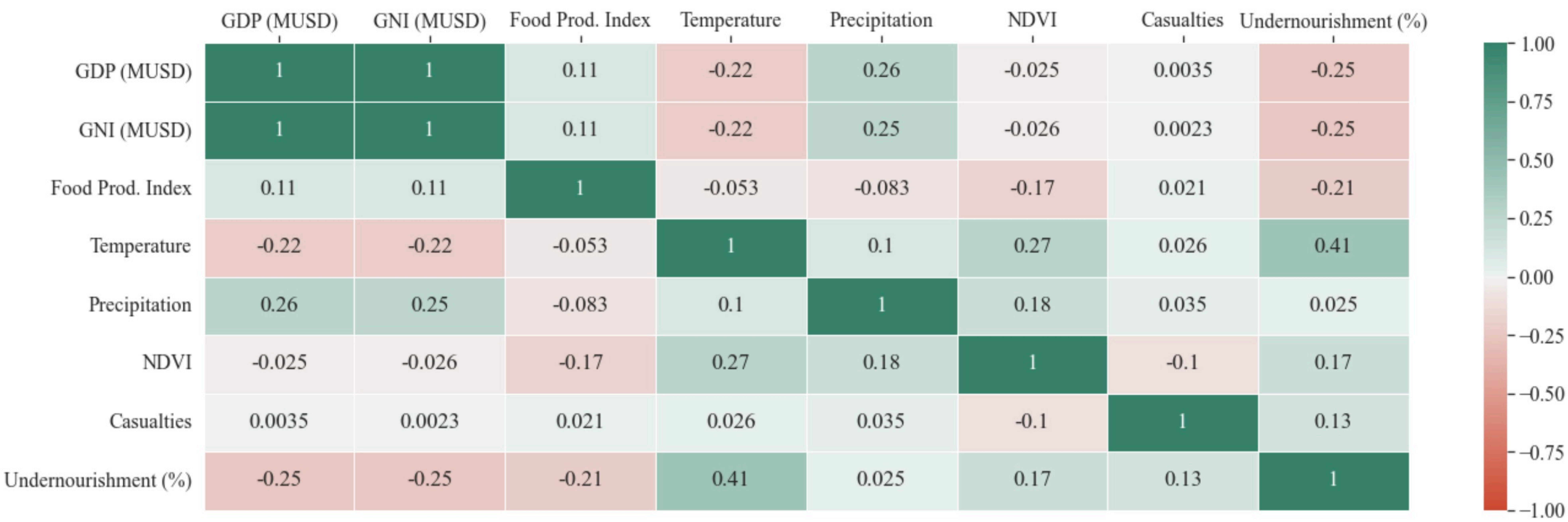

Fig. 2: Relative feature importances

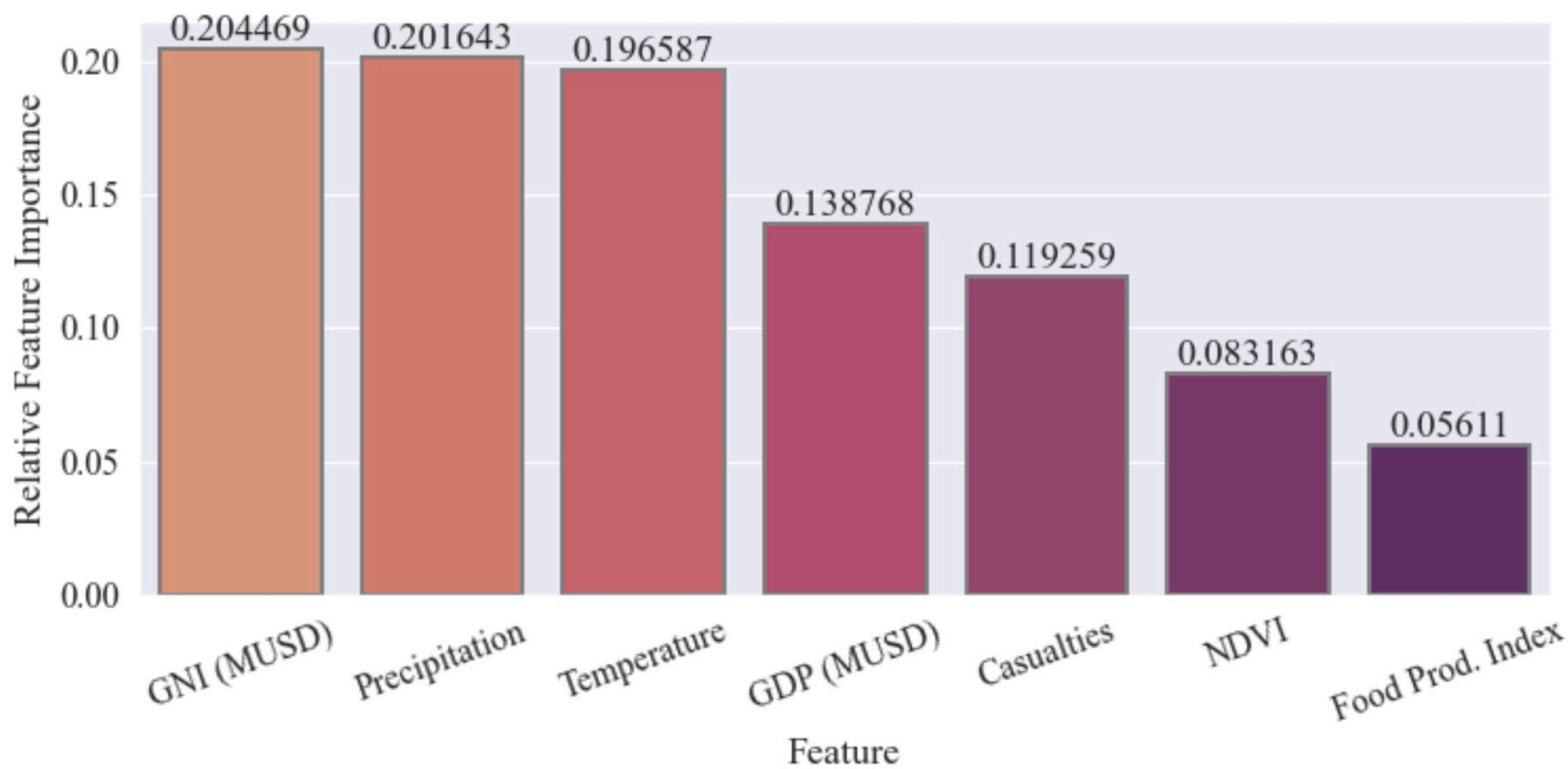

Fig. 3: Different model performances

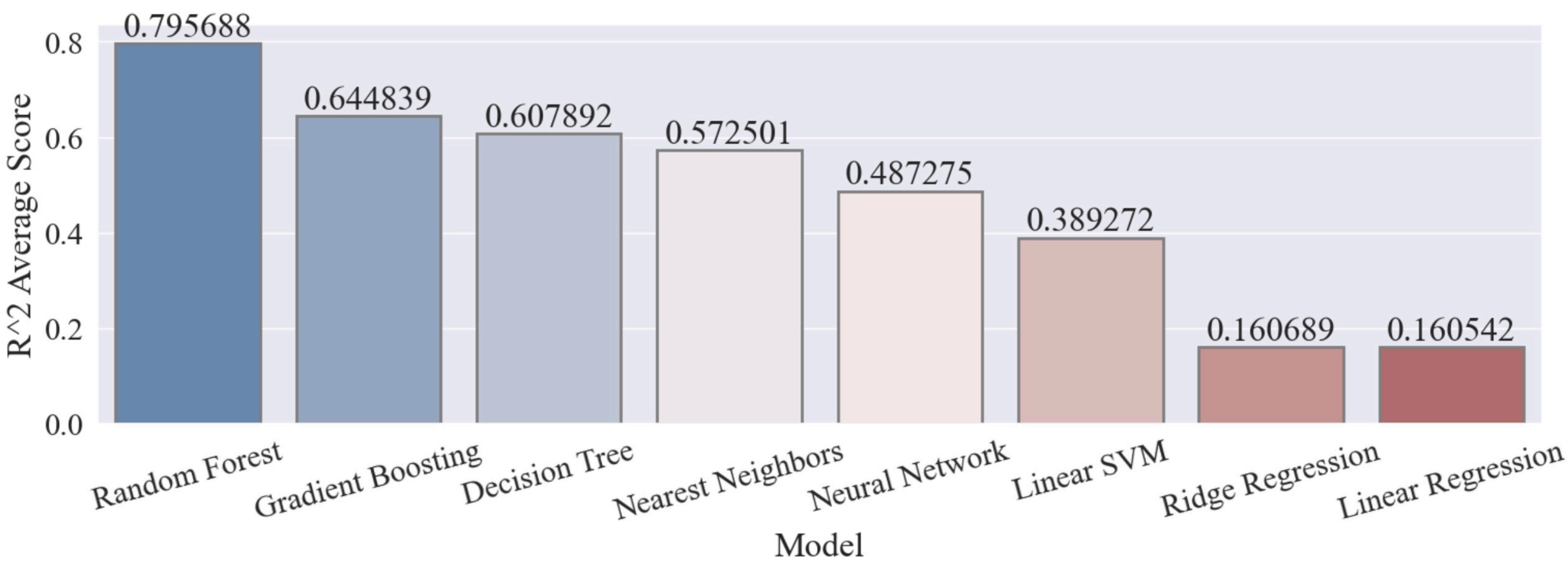

Fig. 4: Process summary - 2020 PoU prediction

	training data	test data
same year prediction	90% (18 years)	10% (2 years)
	cross validation	prediction
model selection [TimeSeriesSplit]	varies due to 5-fold [TimeSeriesSplit]	
	90% (18 years, X: 2001-18; y: 2002-19)	
future year prediction with model chosen in the previous step	90% (18 years)	5% (1 year)
	X: 2001-18; y: 2002-19	X’: 2019; y’: 2020

Fig. 5: Predicted and real PoU 2020