UNIVERSITATEA TEHNICĂ
A MOLDOVEI

Predicting Covid-19 pandemic behaviour to prevent deaths increase

Prezentatori: Dodon Ion
Verebceanu Mirela
Speianu Dana
Tîmbur Ștefan

Grupa: IS-211M
Coordonator: Beșliu Corina

Agenda

- Problem & Proposed solution
- Objectives
- Dataset description
- Data cleaning/processing
- Multiple Linear Regression
- Random Forest
- Auto ARIMA
- VARMAX
- Conclusions

Problem

The increasing number of deaths caused by Covid - 19

Proposed solution

Prevent the number of deaths in the future by
forecasting the pandemic behaviour

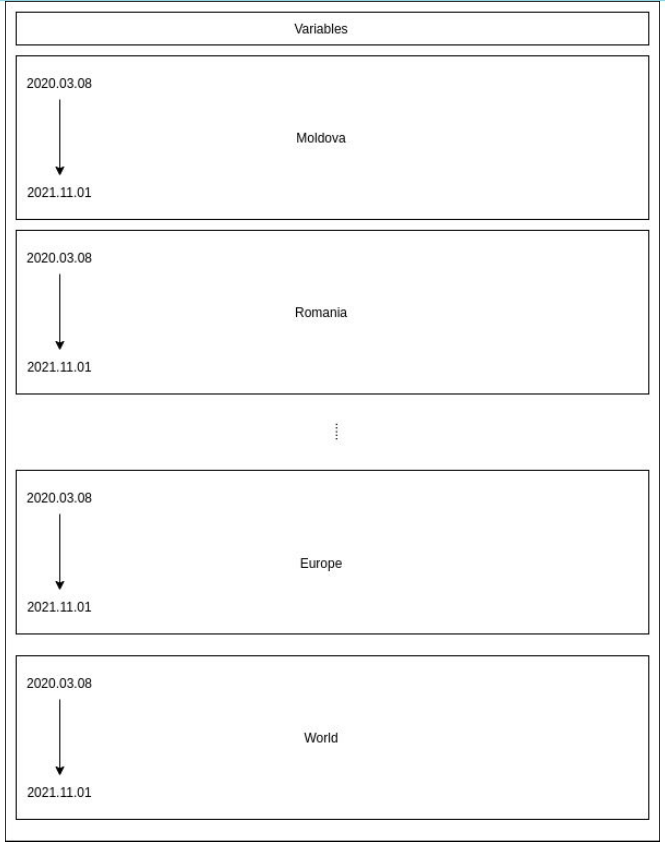
Objectives

- Find and train a model with the best accuracy for a better forecasting.
- Warn people about the need to follow the rules against the pandemic according to the forecasting results.
- Cooperation with the authorities to adjust the stringency according to the forecasting results.
- Decreasing the number of deaths caused by pandemic.

Data description

65 variables x 127 817 rows

- **Total_cases, new_cases**, new_cases_smoothed, total_cases_per_million, new_cases_per_million, new_cases_smoothed_per_million
- Total_deaths, **new_deaths**, new_deaths_smoothed, total_deaths_per_million, new_deaths_per_million, new_deaths_smoothed_per_million
- Excess_mortality, excess_mortality_cumulative, excess_mortality_cumulative_absolute, excess_mortality_cumulative_per_million
- **Icu_patients**, icu_patients_per_million, hosp_patients, hosp_patients_per_million, weekly_icu_admissions, weekly_icu_admissions_per_million, weekly_hosp_admissions, weekly_hosp_admissions_per_million
- **Stringency_index**
- Reproduction_rate
- Total_tests, **new_tests**, total_tests_per_thousand, new_tests_per_thousand, new_tests_smoothed, new_tests_smoothed_per_thousand, **positive_rate**, tests_per_case, tests_units
- Total_vaccinations, **people_vaccinated**, people_fully_vaccinated, **total_boosters**, **new_vaccinations**, new_vaccinations_smoothed, total_vaccinations_per_hundred, people_vaccinated_per_hundred, people_fully_vaccinated_per_hundred, total_boosters_per_hundred, new_vaccinations_smoothed_per_million, new_people_vaccinated_smoothed, new_people_vaccinated_smoothed_per_hundred
- Iso_code, **continent, location, date, population, population_density**, median_age, aged_65_older, aged_70_older, gdp_per_capita, extreme_poverty, **cardiovasc_death_rate, diabetes_prevalence,** female_smokers, male_smokers, handwashing_facilities, hospital_beds_per_thousand, life_expectancy, **human_development_index**



Variables

2020.03.08 → 2021.11.01 Moldova

2020.03.08 → 2021.11.01 Romania

⋮

2020.03.08 → 2021.11.01 Europe

2020.03.08 → 2021.11.01 World

Data cleaning/processing

Observations

1. To many variables and most of them have the same information
2. Missing data for: icu_patients
3. Microcontries have no significant information for the model because the number of population is too small
4. There are negative values
5. There are NaN values
6. There are anomalies (values that increase/decrease in an unexpected manner)
7. There are countries that have no data for some variables

Actions

1. Chose only the most relevant variables
2. Left only Europe and United States (they have the most complete data)
3. Removed countries with populations < 500000, from Europe
4. Replaced negative values with prev. Non negative value
5. Replaced NaN values with mean of prev non missing and next non missing
6. Replaced anomalies with mean values from a the window where the anomalies is met
7. Removed countries that have no data for some variables

Removed Micro Countries

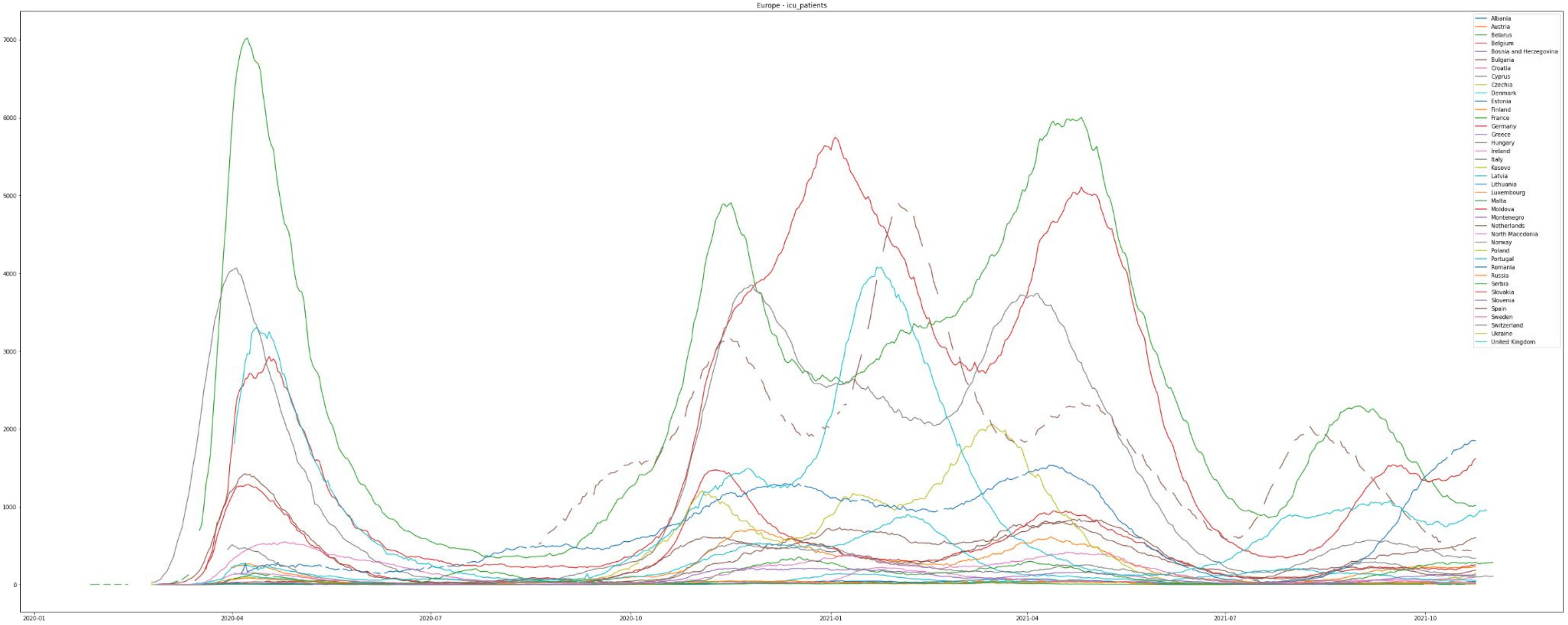
Total 11 microcountries

[Andorra, Faeroe Islands, Gibraltar, Guernsey, Iceland, Isle of Man, Jersey, Liechtenstein, Monaco, San Marino, Vatican]

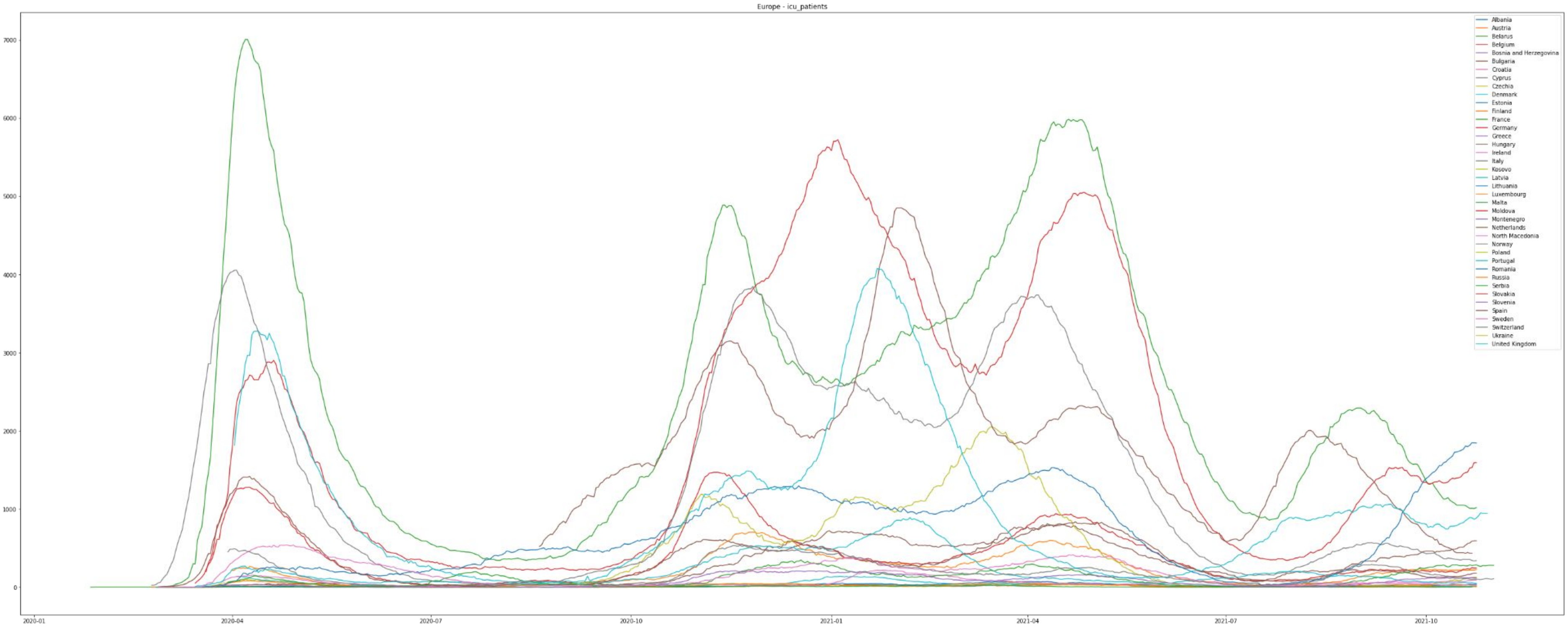
Remaining countries after cleaning

['Austria', 'Belgium', 'Bulgaria', 'Cyprus', 'Czechia', 'Denmark', 'Estonia', 'Finland', 'France', 'Germany', 'Ireland', 'Italy', 'Luxembourg', 'Malta', 'Netherlands', 'Portugal', 'Romania', 'Serbia', 'Slovenia', 'Spain', 'Sweden', 'Switzerland', 'United Kingdom', 'United States']

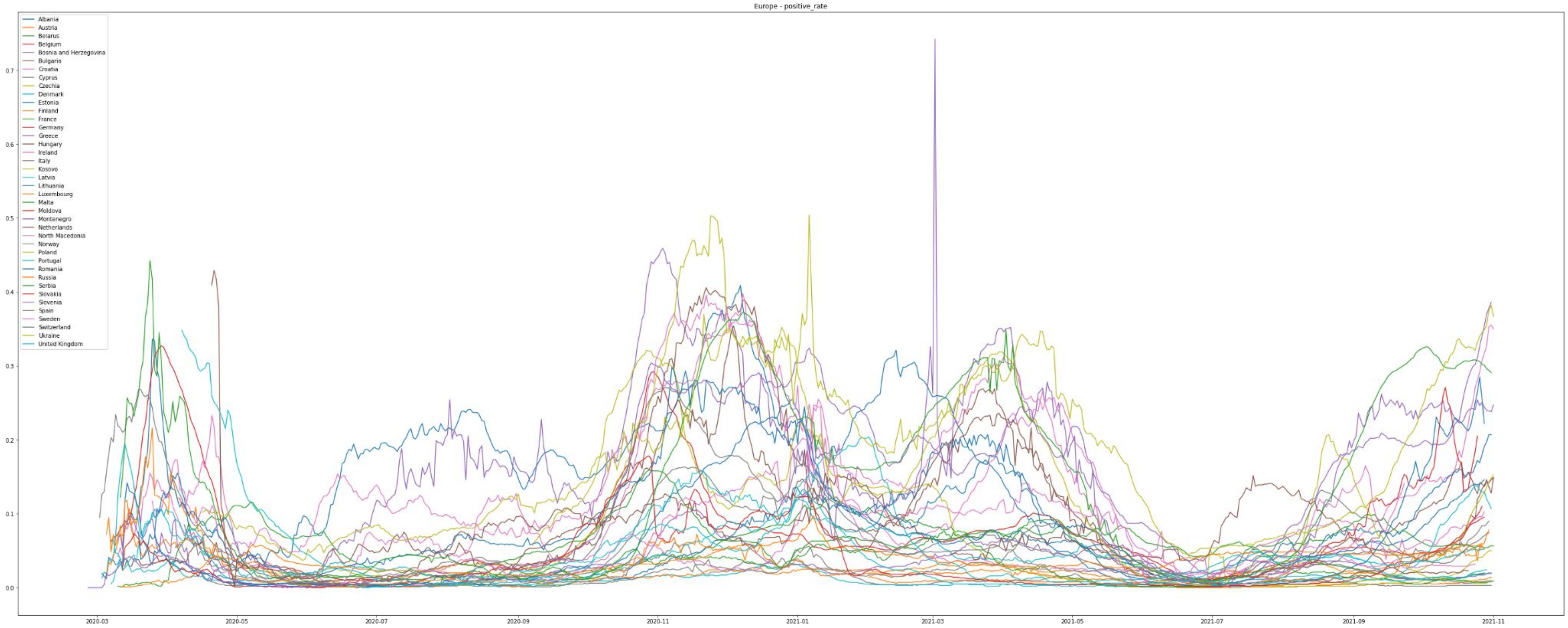
Unprocessed dataset (icu_patients)



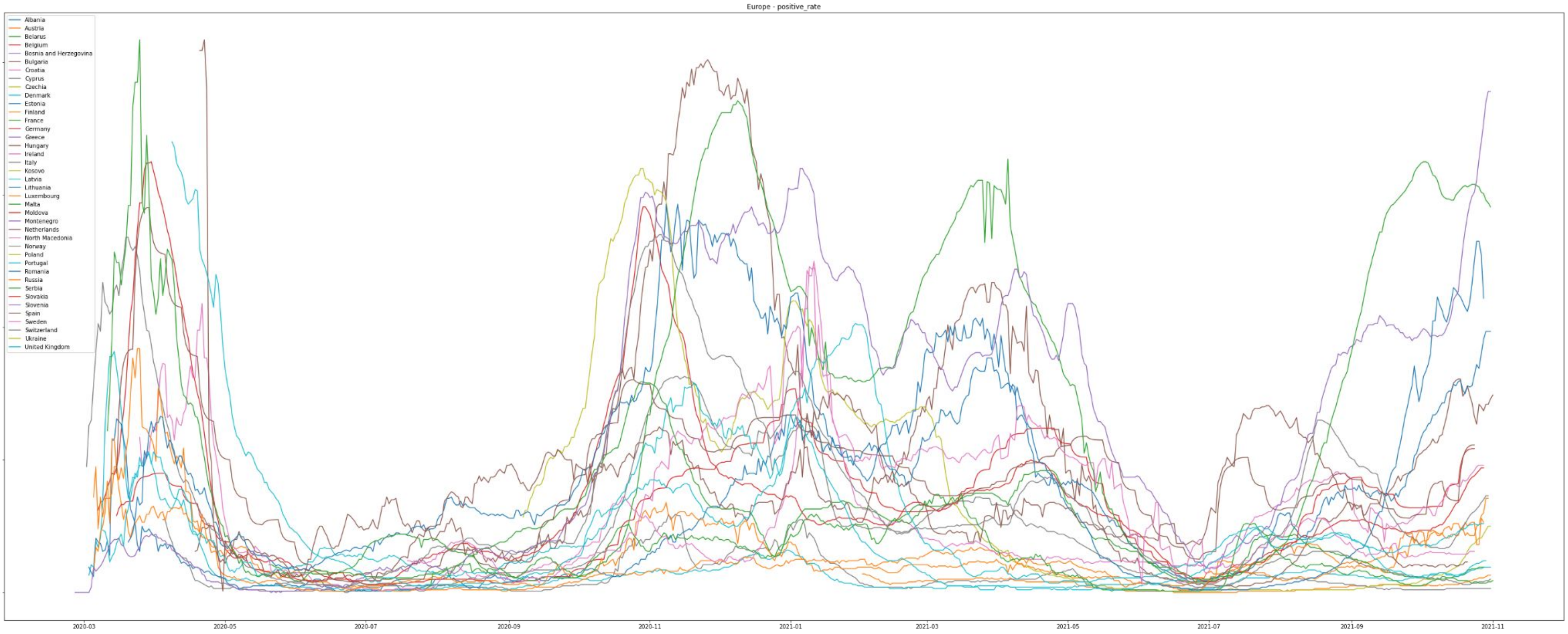
Processed dataset (icu_patients)



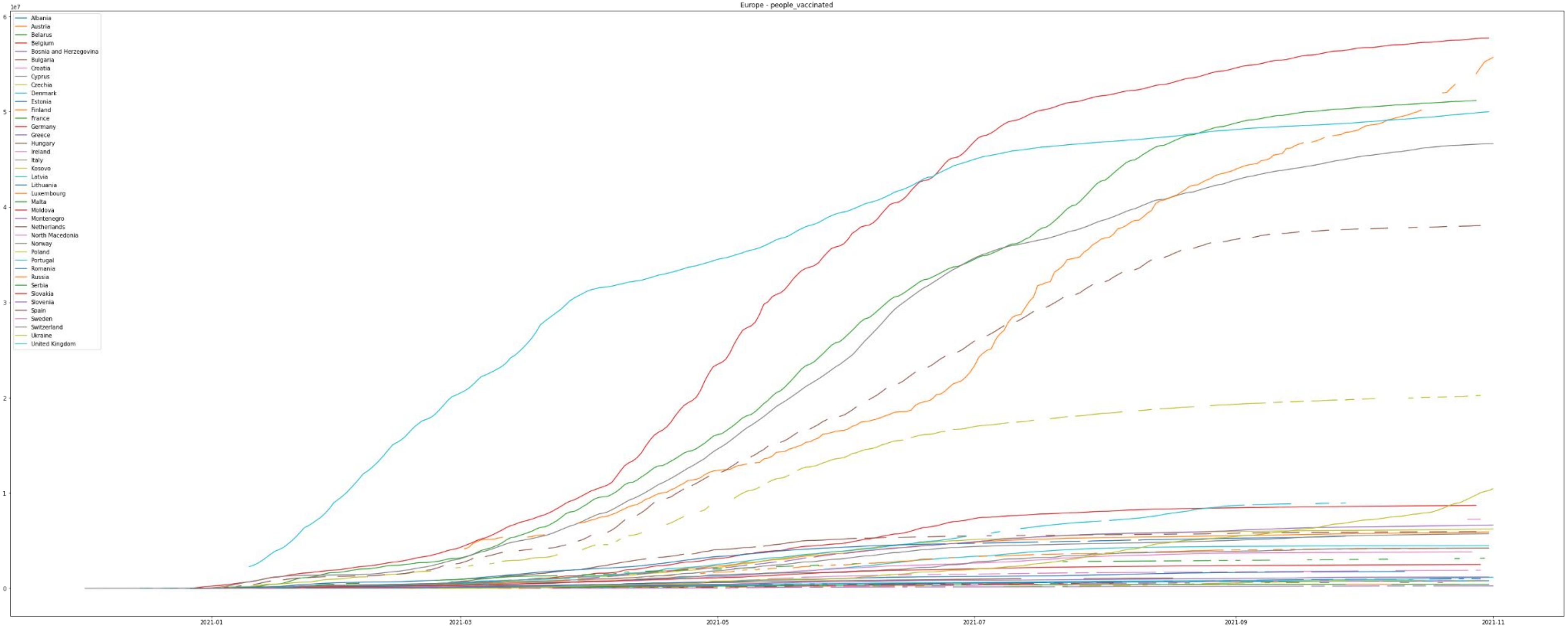
Unprocessed dataset (positive_rate)



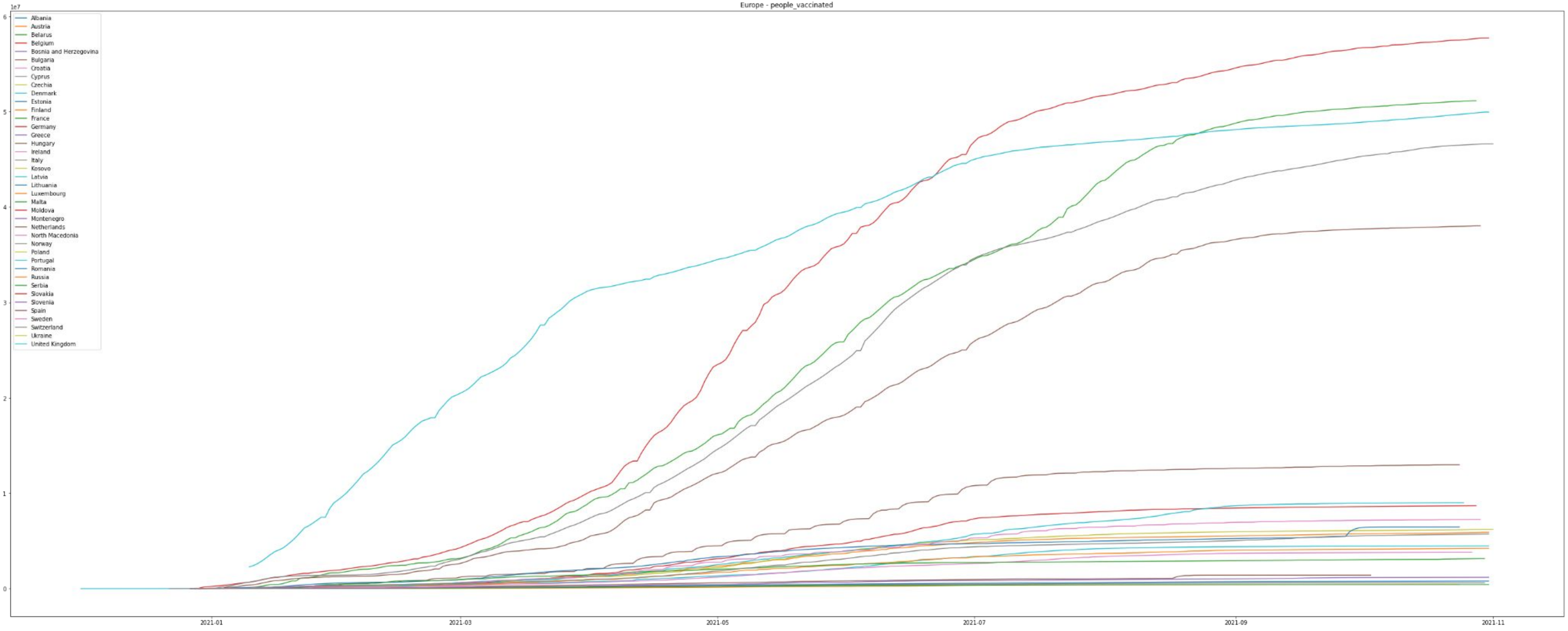
Processed dataset (positive_rate)



Unprocessed dataset (people_vaccinated)



Processed dataset (people_vaccinated)



Multiple Linear Regression

Used variables:

Predicted variable:

- new_deaths

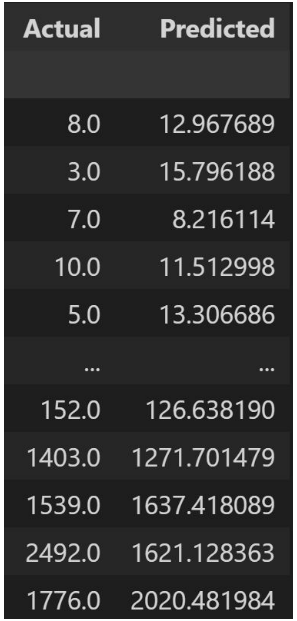
Predictors variables:

- new_cases
- positive_rate
- people_vaccinated
- stringency_index
- human_development_index

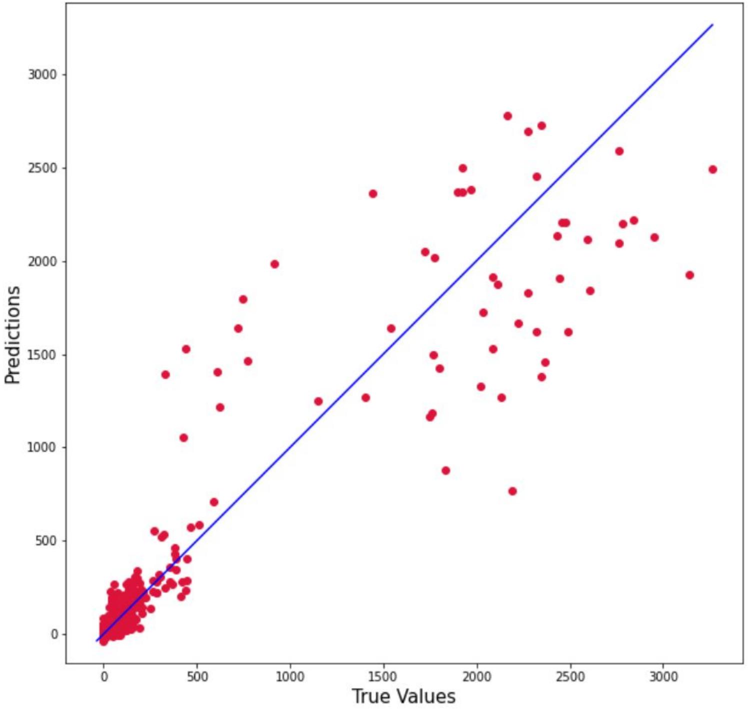
R2 Score = 0.88

Conclusion:

The accuracy score is greater than 0.8 it means we can use this model. But also should check overfitting.



Actual	Predicted
8.0	12.967689
3.0	15.796188
7.0	8.216114
10.0	11.512998
5.0	13.306686
...	...
152.0	126.638190
1403.0	1271.701479
1539.0	1637.418089
2492.0	1621.128363
1776.0	2020.481984

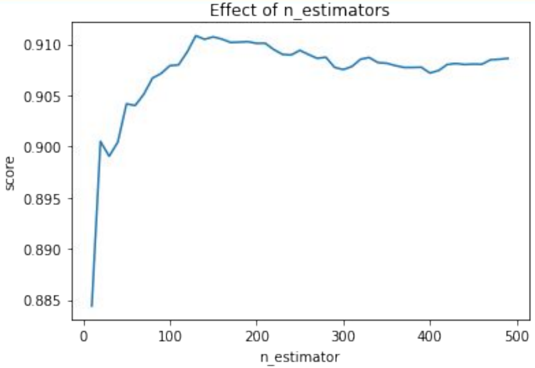


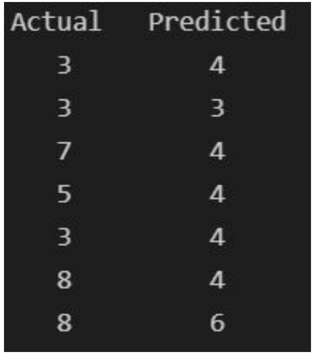
Random Forest Regression

Used variables:

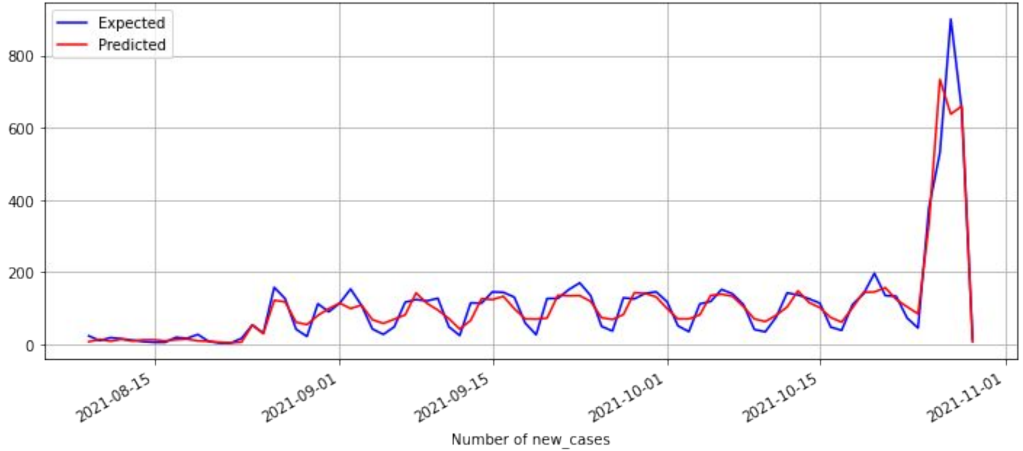
- Population
- New_cases
- Icu_patients
- People_vaccinated
- Stringency_index
- Positive_rate
- Human_development_index
- Diabetes_prevalence
- + 6 lags of new_deaths

Accuracy: 91.08%

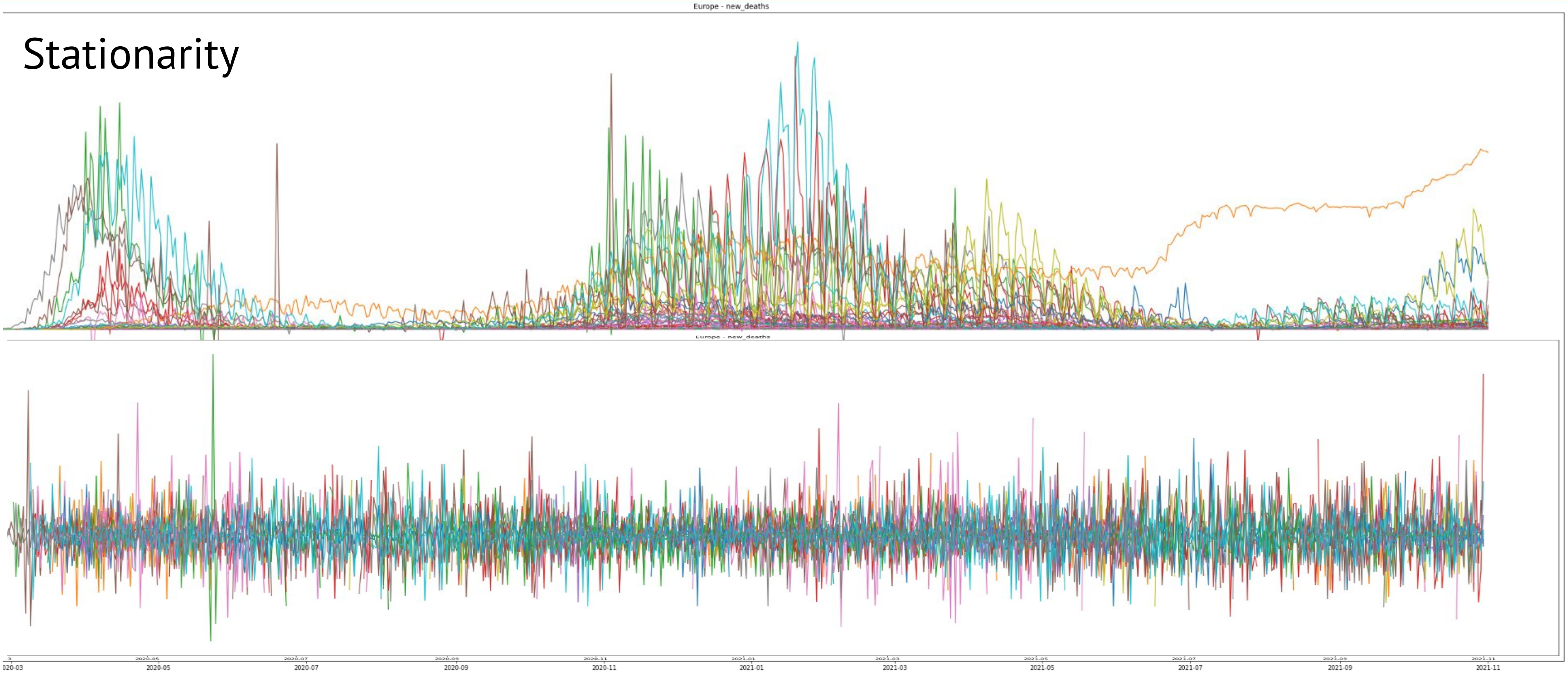




```
Actual    Predicted
  3          4
  3          3
  7          4
  5          4
  3          4
  8          4
  8          6
```



Stationarity



Augmented Dickey-Fuller test (ex: France)

- new_cases
  ADF Statistic: -4.744447
  p-value: 0.000069

- new_deaths
  ADF Statistic: -3.063384
  p-value: 0.029384

- Icu_patients
  ADF Statistic: -7.595366
  p-value: 0.000000

- people_vaccinated
  ADF Statistic: -4.520706
  p-value: 0.000180

- new_vaccinations
  ADF Statistic: -7.620452
  p-value: 0.000000

Auto ARIMA

```
"Austria":{
   "new_cases":{
     "p":3,
     "q":0
   },
   "new_deaths":{
     "p":3,
     "q":0
   },
   "icu_patients":{
     "p":1,
     "q":0
   },
   "new_tests":{
     "p":3,
     "q":0
   },
   "positive_rate":{
     "p":0,
     "q":0
   },
   "people_vaccinated":{
     "p":2,
     "q":0
   },
   "new_vaccinations":{
     "p":3,
     "q":0
   },
   "total_boosters":{
     "p":2,
     "q":0
   },
   "stringency_index":{
     "p":2,
     "q":0
   }
 }
```

```
"Belgium":{
   "new_cases":{
     "p":2,
     "q":0
   },
   "new_deaths":{
     "p":3,
     "q":0
   },
   "icu_patients":{
     "p":1,
     "q":0
   },
   "new_tests":{
     "p":3,
     "q":0
   },
   "positive_rate":{
     "p":0,
     "q":0
   },
   "people_vaccinated":{
     "p":1,
     "q":0
   },
   "new_vaccinations":{
     "p":3,
     "q":0
   },
   "total_boosters":{
     "p":0,
     "q":0
   },
   "stringency_index":{
     "p":1,
     "q":0
   }
 }
```

```
"United States":{
   "new_cases":{
     "p":3,
     "q":0
   },
   "new_deaths":{
     "p":0,
     "q":0
   },
   "icu_patients":{
     "p":3,
     "q":0
   },
   "new_tests":{
     "p":3,
     "q":0
   },
   "positive_rate":{
     "p":1,
     "q":0
   },
   "people_vaccinated":{
     "p":3,
     "q":0
   },
   "new_vaccinations":{
     "p":3,
     "q":0
   },
   "total_boosters":{
     "p":0,
     "q":0
   },
   "stringency_index":{
     "p":2,
     "q":0
   }
 }
```

VARMAX

```
mod = sm.tsa.VARMAX(
     np.asarray(varmax_train_dataset[endogeneous_variables]),
     np.asarray(varmax_train_dataset[exogeneous_variables]
), order=(1, 0))
```

```
exogeneous_variables = [
   'population',
   'population_density',
   'diabetes_prevalence',
  'human_development_index',
   'new_tests',
   'stringency_index',
   'icu_patients',
   'cardiovasc_death_rate',
   'people_vaccinated',
   'new_vaccinations',
   'total_boosters'
]
```

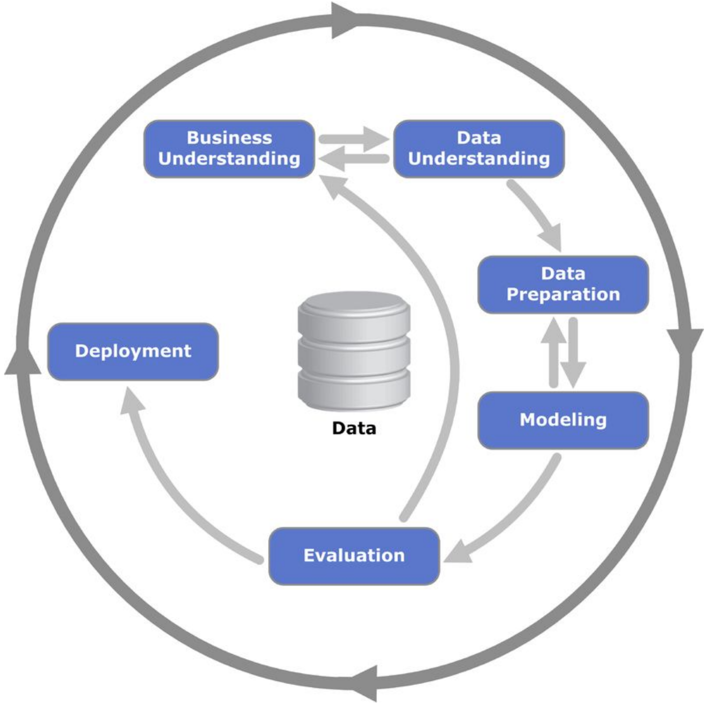
```
endogeneous_variables = [
   'new_cases',
   'new_deaths',
   'positive_rate'
]
```

```
              Real                                          Forecasted
 New_cases,  new_deaths,   positive_rate        New_cases,   new_deaths,  positive_rate
…
[-0.46419841 -0.8206747  -0.11778304]        [0.55200313 0.21145555 0.06910153]
[ 0.4638342   0.26876337 -0.11778304]       [0.54222414 0.22742225 0.0692272 ]
[-1.42084716 -1.66887848 -0.11778304]        [0.55274807 0.13892331 0.0666618 ]
[ 0.69095626  1.16538023 -0.11778304]       [0.53853355 0.14110688 0.06648889]
[ 2.43891026  2.56450023 -0.11778304]       [0.53574616 0.14206916 0.06657076]
…
```

Root mean square error: 0.6750305025536002
R2 = 0.975630916147559

Challenges & Conclusions

- The more complete the dataset is, the easiest it is to work on it.
- When there are many variables it is more difficult to choose which are the best and we should work with just a few of them.
- Sometimes it is better to make a VAR model for each entity from the dataset, than to make a VARMAX that would work for all entities.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data