Predikcija emisije ugljen-dioksida(CO_2)

Problem regresije – Mašinsko učenje – Projekat 1.

Skup podataka: Agri-food CO2 emission

- **Cilj projekta:** Predvidjanje **ukupne emisije C0$_2$**(*Total emission*) po **teritoriji,** pomoću **regresije** koristeći različite *demografske*, *poljoprivredne* i *podatke o korišćenju zemljišta*.

- **Merenje CO2**: Sve emisije se beleže u kilotonama (kt), gde je 1 kt = 1 000 000 kg CO2.

- **Smanjenje ugljen-dioksida:** Šumsko zemljište(*Forestland)* je jedina karakteristika koja pokazuje negativne emisije zbog svoje uloge kao ponora ugljenika, apsorbujući CO2 iz atmosfere.

Eksploratorna analiza podataka

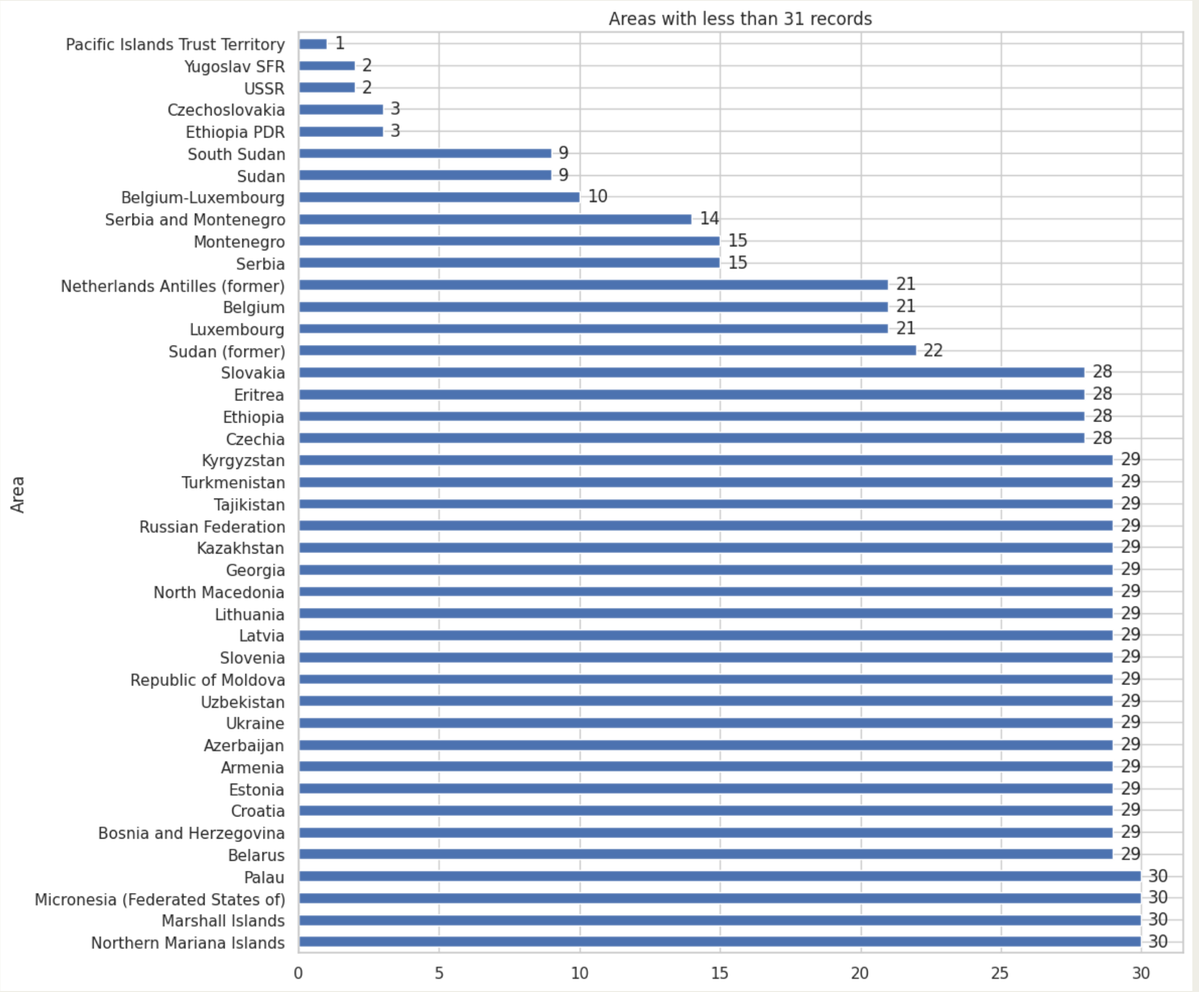
	Nedostajuce vrednosti	Tip
Area	0	object
Year	0	int64
Savanna fires	**31**	float64
Forest fires	**93**	float64
Crop Residues	**1389**	float64
Rice Cultivation	0	float64
Drained organic soils (CO2)	0	float64
Pesticides Manufacturing	0	float64
Food Transport	0	float64
Forestland	**493**	float64
Net Forest conversion	**493**	float64
Food Household Consumption	**473**	float64
Food Retail	0	float64
On-farm Electricity Use	0	float64
Food Packaging	0	float64
Agrifood Systems Waste Disposal	0	float64
Food Processing	0	float64
Fertilizers Manufacturing	0	float64
IPPU	**743**	float64
Manure applied to Soils	**928**	float64
Manure left on Pasture	0	float64
Manure Management	**928**	float64
Fires in organic soils	0	float64
Fires in humid tropical forests	**155**	float64
On-farm energy use	**956**	float64
Rural population	0	float64
Urban population	0	float64
Total Population - Male	0	float64
Total Population - Female	0	float64
total_emission	0	float64
Average Temperature °C	0	float64

- *Broj kolona*: **31**
- *Broj opservacija (redova)*: **6965**

Pojedine teritorije imaju manji broj merenja od 42 (maksimalan broj)

Area

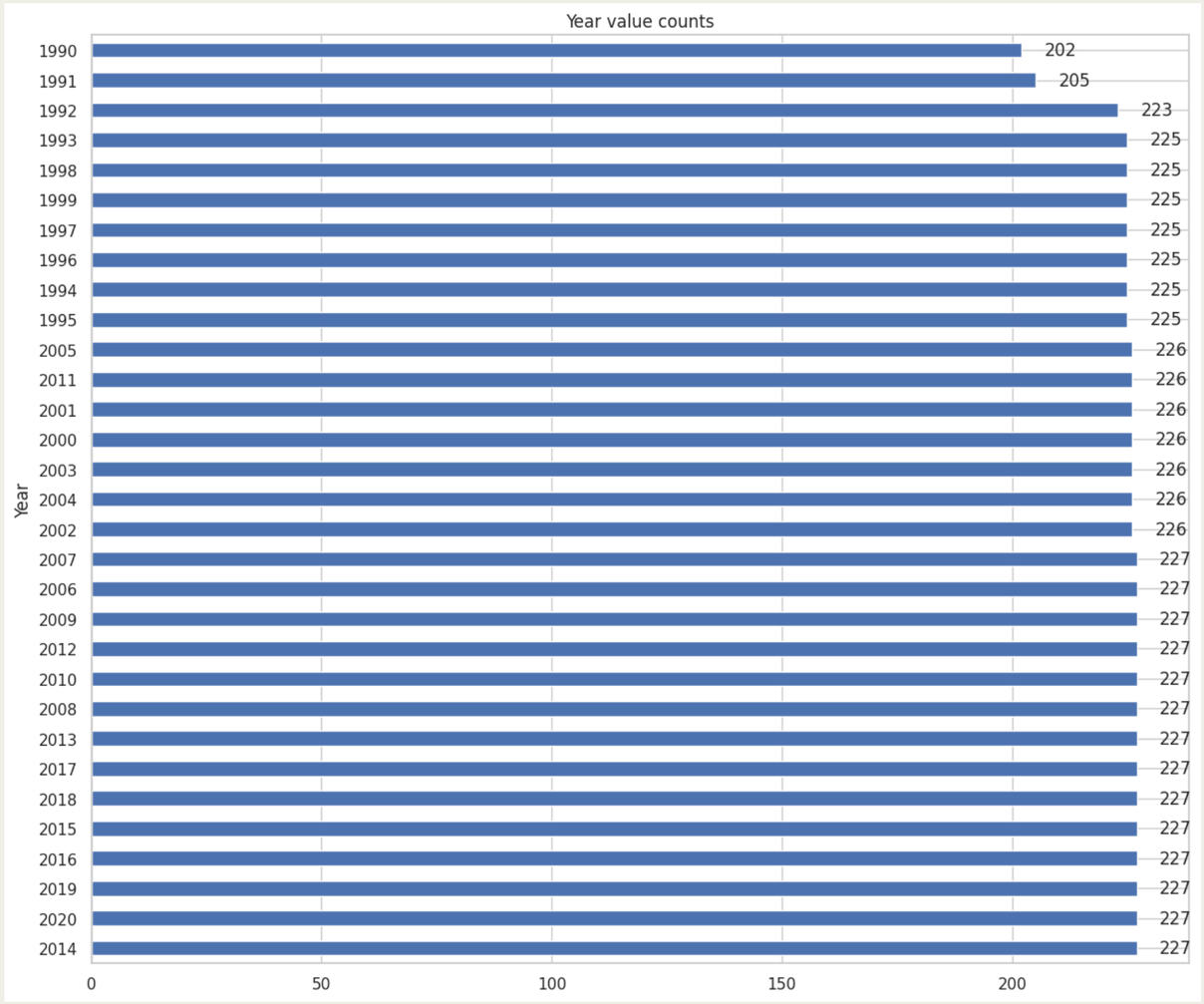
- Kategorička - nominalna varijabla
- Procenat nedostajućih vrednosti: **0.0%**
- Broj jedinstvenih teritorija (kategorija): **236**



Ne postoji za sve godine isti broj merenja

Year

- Numerička – diskretna varijabla
- Procenat nedostajućih vrednosti: **0.0%**
- Završna godina (*max*): **2020**
- Početna godina (*min*): **1990**



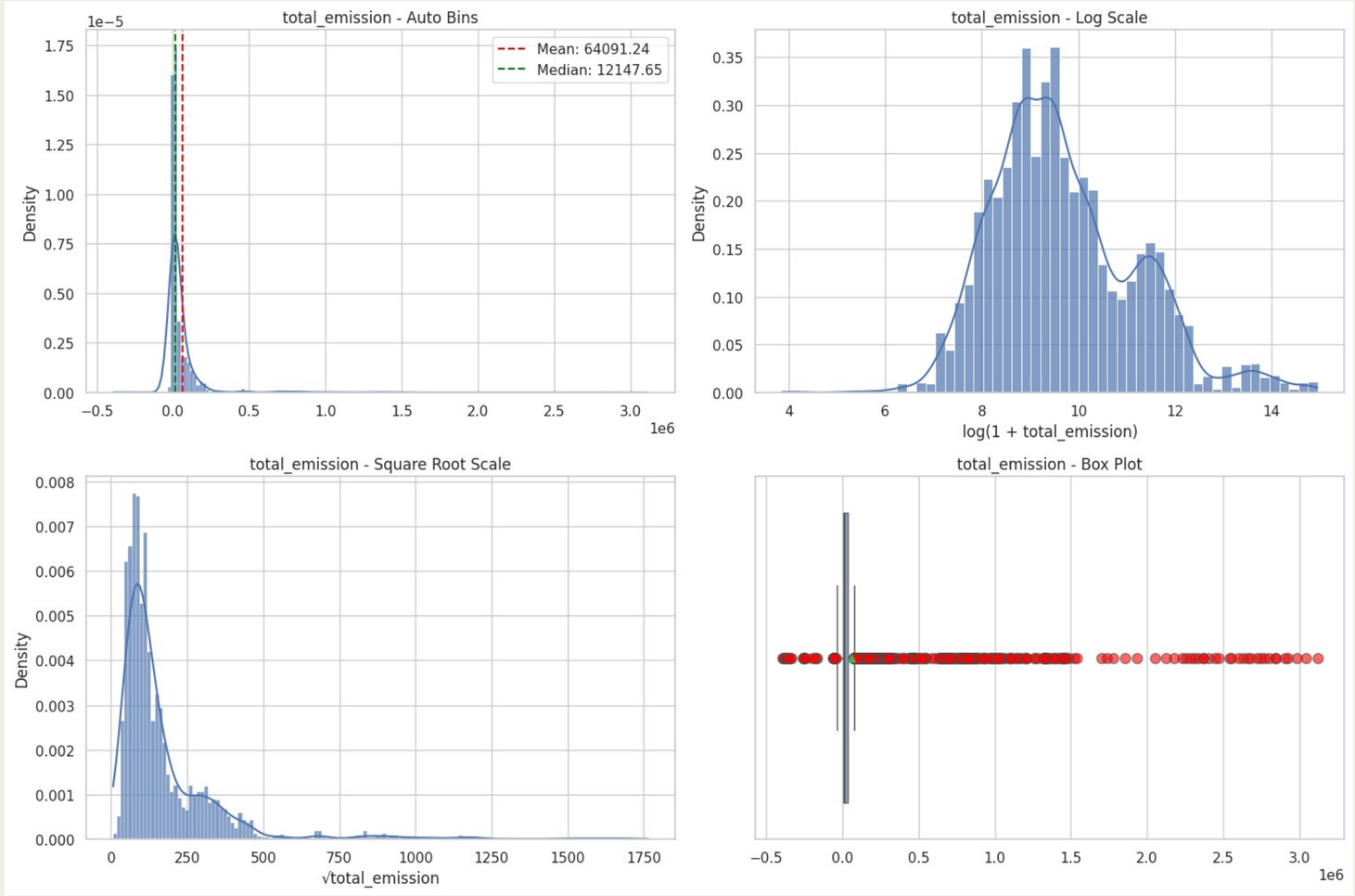
Ne možemo tvrditi da su outlier-i stvarne greške u merenjima, stoga ih **ne odstranjujemo**

Total emission

- Numerička – kontinualna varijabla
- Procenat nedostajućih vrednosti: **0.0%**

- Statistički podaci:
  - **mean** 64091.24
  - **std** 228312.96
  - **min** -391884.06
  - **25%** 5221.24
  - **50%** 12147.65
  - **75%** 35139.73
  - **max** 3115113.75

- Izveštaj o outlier-ima:
  - Q1: **5221,24**
  - Q3: **35139,73**
  - IQR: **29918,49**
  - Donja granica: **-39656,49**
  - Gornja granica: **80017,46**
  - Broj kandidata za statističke outlier-e: **1142**
  - Procenat kandidata za statističke izuzetke: **16,4%**
  - Prvih 5 ispod donje granice: [**-391884,06 -387630,12 -378502,97 -375376,67 -365474,32**]
  - Prvih 5 iznad gornje granice: [**2902693,11 2919286,26 2978585,26 3039089,09 3115113,75**]



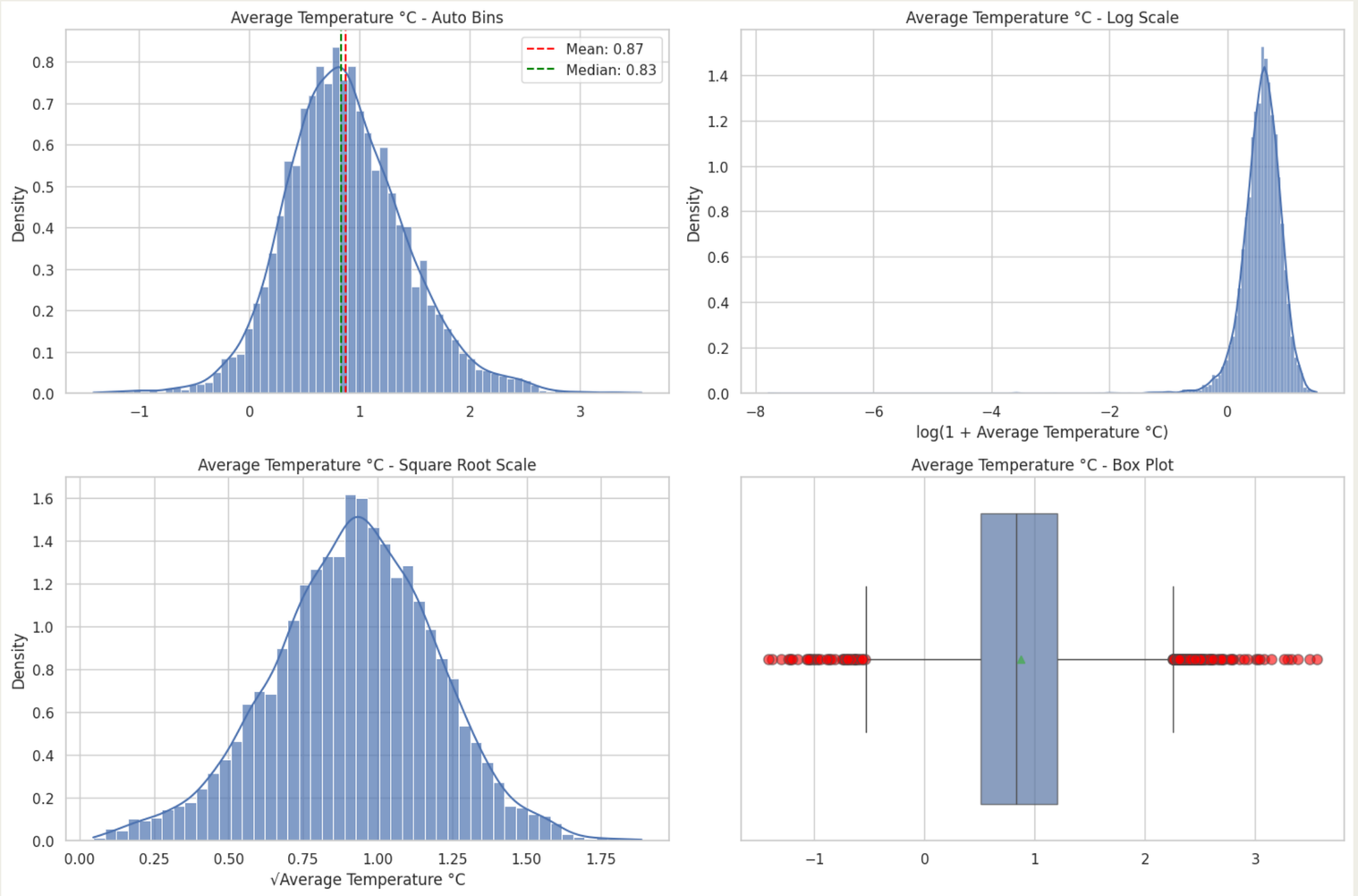
Jedina varijabla koja ima distribuciju približno normaloj (Gausova)

Average temperature °C

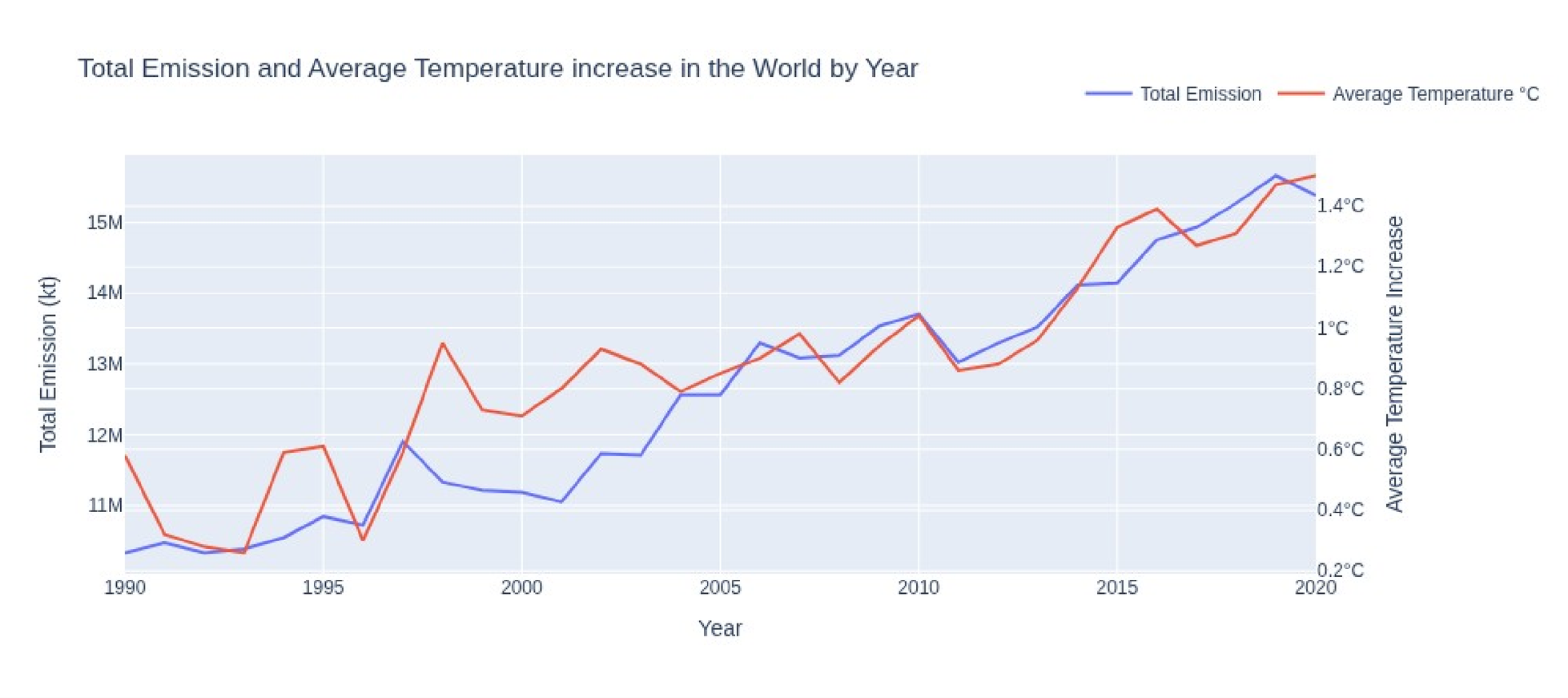
- Numerička – kontinualna varijabla
- Procenat nedostajućih vrednosti: **0.0%**

- Statistički podaci:
  - **mean** 0.87
  - **std** 0.56
  - **min** -1.42
  - **25%** 0.51
  - **50%** 0.83
  - **75%** 1.21
  - **Max** 3.56

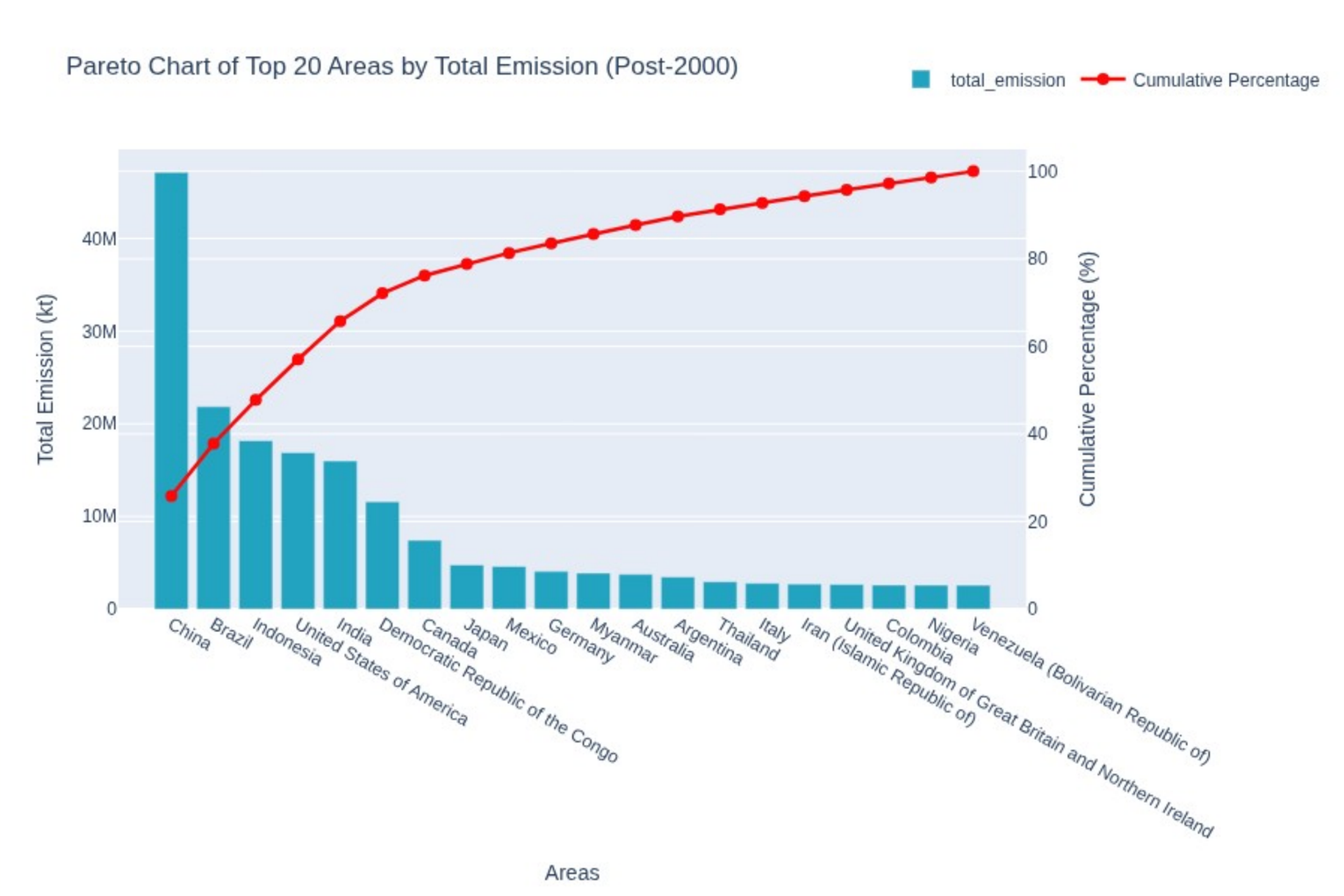
- Izveštaj o outlier-ima:
  - Q1: **0.51**
  - Q3: **1,21**
  - IQR: **0.7**
  - Donja granica: **-0.53**
  - Gornja granica: **2.25**
  - Broj kandidata za statističke outlier-e: **155**
  - Procenat kandidata za statističke izuzetke: **2.23%**
  - Prvih 5 ispod donje granice: **[-1.42 -1.38 -1.3 -1.24 -1.22]**
  - Prvih 5 iznad gornje granice: **[3.3 3.33 3.38 3.5 3.56]**



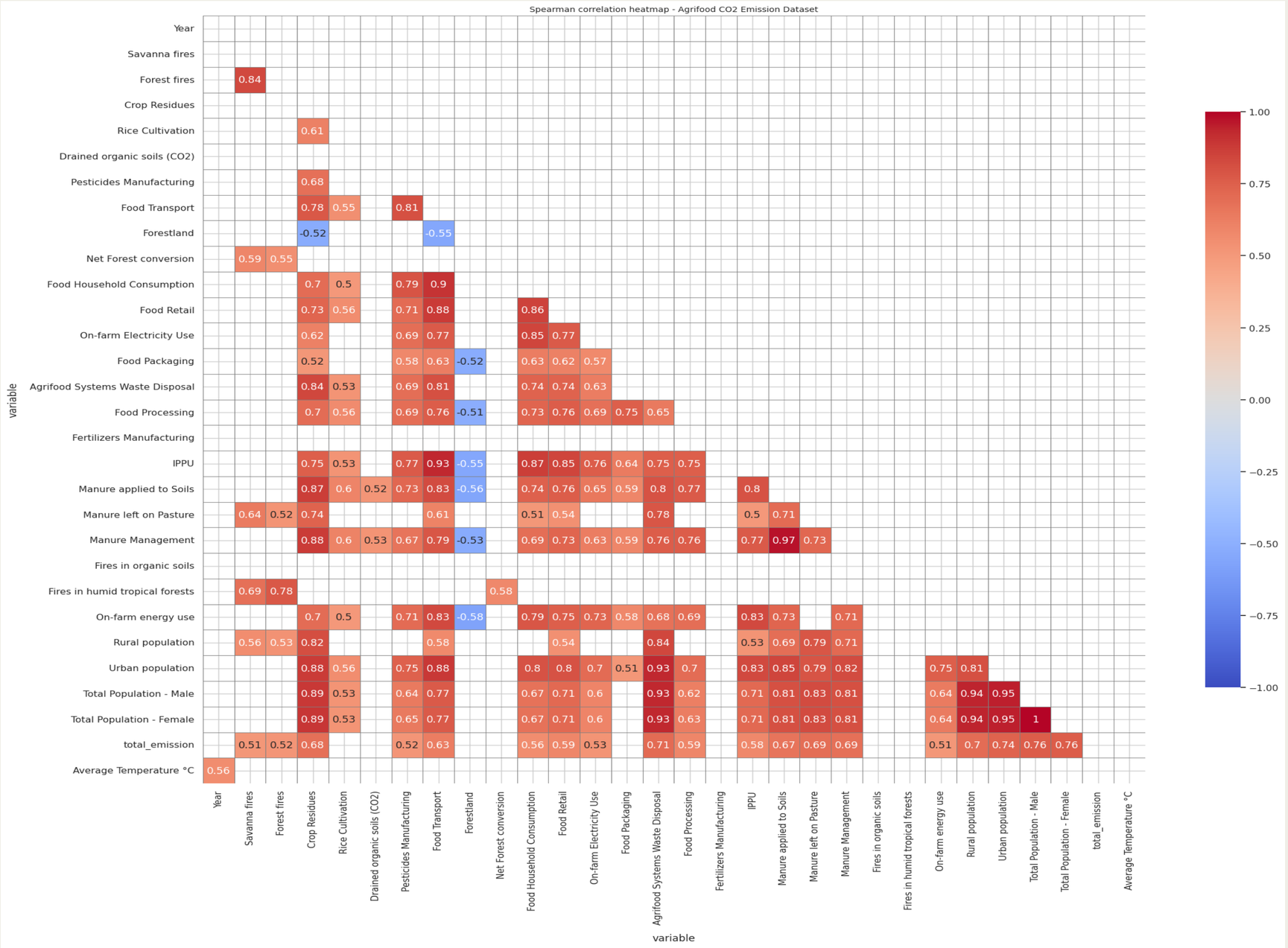
Totalna emisija CO_2 prati prosečno povećanje temperature kroz vreme



Prva četiri zagađivača predstavljaju 60% od ukupne zagađenosti među top 20 zagađivača



Spearmanova korelacija – jer gotovo sve varijable nemaju normalnu distribuciju



Korelacioni racio (eta koeficijent η) – meri jačinu nelinearne veze između nominalne i kontinualne variable

Correlation Ratio η

	Var_1	Var_2	Eta
0	Area	Rice Cultivation	0.998
1	Area	Drained organic soils (CO2)	0.996
2	Area	Manure Management	0.996
3	Area	Agrifood Systems Waste Disposal	0.996
4	Area	Total Population - Female	0.995
5	Area	Total Population - Male	0.995
6	Area	Manure applied to Soils	0.993
7	Area	Rural population	0.992
8	Area	Manure left on Pasture	0.991
9	Area	Crop Residues	0.988
10	Area	On-farm energy use	0.985
11	Area	Food Transport	0.972
12	Area	Urban population	0.963
13	Area	Pesticides Manufacturing	0.962
14	Area	Food Processing	0.961
15	Area	Fertilizers Manufacturing	0.955
16	Area	On-farm Electricity Use	0.953
17	Area	Net Forest conversion	0.943
18	Area	Fires in humid tropical forests	0.935
19	Area	total_emission	0.935
20	Area	Forestland	0.934
21	Area	Savanna fires	0.932
22	Area	Forest fires	0.923
23	Area	Food Packaging	0.914
24	Area	Food Household Consumption	0.885
25	Area	Food Retail	0.875
26	Area	IPPU	0.870
27	Area	Fires in organic soils	0.759
28	Area	Average Temperature °C	0.486

- **Area** je u jakoj korelaciji sa većinom varijabli, zbog toga ćemo je odstraniti iz skupa podataka u daljim koracima

Inženjering varijabli i priprema podataka

1. Deljenje skupa podataka na trening i test skup

- Uzete reprezentativne godine: **["1993", "1996", "2000", "2005", "2008", "2013", "2018"]**
- Procentualni udeo testnog skupa podataka: **22,73%**

2. Imputacija nedostajućih vrednosti --- korišćenjem **Random Forest Regressor-a**

- Treniranje Imputora samo na trening skupu podataka radi izbegavanja data leakage-a

3. **Odbacivanje karakteristike:** zbog visoke korelacije **(≥ ±0.9)** i/ili velikog broja nedostajućih vrednosti:

1) Total population – Female + Total population – Male = **Total population**

2) Odbacivanje zbog visoke korelacije:

(a) *Manure Managemen*t

(b) *Food Transport*

(c) *Area*

(d) *Crop Residues* – i zbog velikog broja nedostajućih vrednosti – **19.94%** (najveći u skupu podataka)

4. Normalizacija – korištćenjem *Robust scaler-a* – robustan na outlier-e, koristi **medijanu** i **IQR**

Podešavanje modela i Optimizacija hiperparametara

Korišćeni modeli u analizi:

- Linear Regression
- Linear RegressionRidge
- RegressionLasso
- RegressionElasticNet
- RegressionKNN
- RegressorDecision Tree
- RegressorBagging
- RegressorRandom Forest
- RegressorAdaBoost
- RegressorXGBoost
- SupportVectorMachines Regressor

Optimizacija hiperparametara:

- Korišćena metoda Random Search
- 20 iteracija (slučajnih kombinacija)
- Sa Cross Validacijom – K-Fold=5

Najbolji nađeni hiperparametri za 2 najbolja modela:

1) Bagging Regressor:
   a) 'n_estimators': 500,
   b) 'max_samples': 1.0,
   c) 'max_features': 0.6

2) Random Forest Regressor:
   a) 'n_estimators': 500,
   b) 'min_samples_split': 5,
   c) 'min_samples_leaf': 1,
   d) 'max_features': 'log2',
   e) 'max_depth': 20,
   f) 'bootstrap': False

Redukcija dimenzionalnosti – PLS se pokazala kao korisna

Principal component analiza (**PCA**):

- Podešena pomoću Greed Search-a
- Dobijena jedna komponenta od svih varijabli
- Komponenta objašnjava 99% varijanse
- Podešavanje svih modela regresije na komponentu
- **Zaključak:** Pristup se pokazao kao loš (loši rezultati)

Parcijalni najmanji kvadrati (**PLS**):

- Uzima u obzir multikolinearnost između nezavisnih varijabli i zavisne varijable, za razliku od PCA, koja posmatra samo varijansu između nezavisnih varijabli

- **Zaključak**: Koristeći prag **0.6** kod VIP skorova (*Variable Importance in Projection*= za redukciju dimenzionalnosti, dobili smo približne rezultate kao i kod *podešenih modela bez optimizacije hiperparametra*

Feature Importance - Sorted by VIP Scores

	feature	VIP
17	Fires in humid tropical forests	2.582
11	Food Packaging	2.202
8	Net Forest conversion	1.558
7	Forestland	1.185
16	Fires in organic soils	1.161
12	Food Processing	1.006
2	Forest fires	0.945
10	On-farm Electricity Use	0.695
5	Pesticides Manufacturing	0.629
1	Savanna fires	0.601
3	Rice Cultivation	0.549
18	On-farm energy use	0.442
20	Total Population	0.406
4	Drained organic soils (CO2)	0.344
9	Food Retail	0.314
15	Manure Management	0.304
13	Fertilizers Manufacturing	0.286
14	Manure left on Pasture	0.245
6	Food Transport	0.218
0	Year	0.010
19	Average Temperature °C	0.004

Evaluacija modela – *Bagging regressor* se pokazao kao najbolji sa optimizacijom hiperparametra

Model Performance Comparison - Sorted by Test RMSE

	Model	Train_MAE	Train_RMSE	Train_MSE	Train_R2	Train_R2_Adj	Test_MAE	Test_RMSE	Test_MSE	Test_R2	Test_R2_Adj
0	Random Forest Regressor	1405.84	7838.68	61444938.33	1.00	1.00	2369.03	11170.91	124789283.35	1.00	1.00
1	Bagging Regressor	2061.27	16423.37	269727178.95	0.99	0.99	3020.63	14927.56	222832093.96	1.00	1.00
2	XGBoost Regressor	788.53	1410.74	1990194.21	1.00	1.00	3329.33	15585.54	242908996.89	1.00	1.00
3	Lasso Regression	7445.84	17244.27	297365009.70	0.99	0.99	7410.80	15954.59	254549020.59	0.99	0.99
4	Ridge Regression	7446.02	17244.27	297364998.35	0.99	0.99	7411.23	15954.67	254551569.61	0.99	0.99
5	Linear Regression	7446.19	17244.27	297364997.30	0.99	0.99	7411.50	15954.72	254553127.18	0.99	0.99
6	ElasticNet Regression	7172.77	17343.59	300800059.76	0.99	0.99	7040.07	15984.31	255498214.91	0.99	0.99
7	Decision Tree Regressor	0.00	0.00	0.00	1.00	1.00	3233.95	17193.76	295625246.37	0.99	0.99
8	KNN Regressor	5306.88	25276.71	638912142.30	0.99	0.99	5894.57	26038.36	677995931.46	0.99	0.99
9	AdaBoost Regressor	48914.50	57561.65	3313343812.88	0.94	0.94	49473.84	58304.74	3399442362.47	0.93	0.93
10	Support Vector Regressor	61467.02	235814.33	55608400116.56	-0.05	-0.06	59451.70	228250.77	52098414180.71	-0.05	-0.07

- **Full** dataset – **Base** modeli

Model Performance Comparison - Sorted by Test RMSE - REDUCED DIMENSIONALITY WITH PLS

	Model	Train_MAE	Train_RMSE	Train_MSE	Train_R2	Train_R2_Adj	Test_MAE	Test_RMSE	Test_MSE	Test_R2	Test_R2_Adj
0	Random Forest Regressor PO	954.92	1520.81	2312851.44	1.00	1.00	3749.15	11452.84	131167622.15	1.00	1.00
1	Bagging Regressor PO	1972.22	8392.50	70433986.55	1.00	1.00	4179.57	15834.06	250717482.69	0.99	0.99
2	KNN Regressor PO	8.97	79.65	6343.51	1.00	1.00	5262.96	23455.88	550178249.21	0.99	0.99
3	Decision Tree Regressor PO	3487.34	13068.85	170794947.10	1.00	1.00	6079.64	23601.98	557053667.47	0.99	0.99
4	XGBoost Regressor PO	5782.51	10419.88	108573862.13	1.00	1.00	7779.97	25209.08	635497957.09	0.99	0.99
5	ElasticNet Regression PO	17198.12	40395.67	1631810503.21	0.97	0.97	16893.06	37987.07	1443017456.86	0.97	0.97
6	Ridge Regression PO	17198.12	40395.67	1631810503.21	0.97	0.97	16893.06	37987.07	1443017477.95	0.97	0.97
7	Lasso Regression PO	17198.12	40395.67	1631810503.21	0.97	0.97	16893.06	37987.07	1443017477.96	0.97	0.97
8	Linear Regression PO	17198.12	40395.67	1631810503.21	0.97	0.97	16893.06	37987.07	1443017477.98	0.97	0.97
9	AdaBoost Regressor PO	38691.93	59394.80	3527742558.39	0.93	0.93	38558.48	58753.00	3451914594.46	0.93	0.93
10	Support Vector Regressor PO	61136.16	234145.97	54824334437.74	-0.04	-0.04	59428.41	228163.02	52058363605.60	-0.05	-0.06

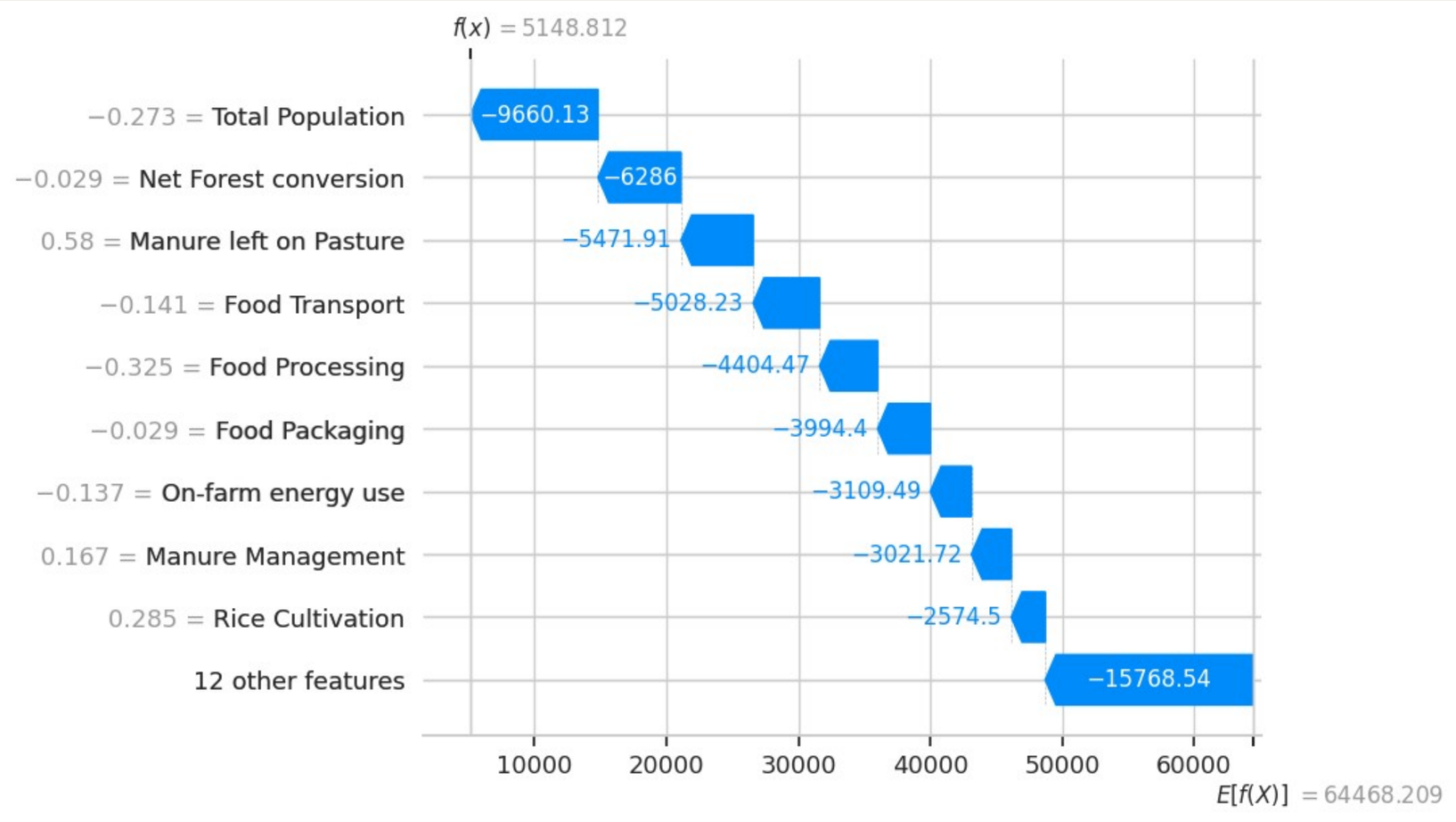
- **PLS** dataset – **Optimizovani** modeli

Model Performance Comparison - Sorted by Test RMSE

	Model	Train_MAE	Train_RMSE	Train_MSE	Train_R2	Train_R2_Adj	Test_MAE	Test_RMSE	Test_MSE	Test_R2	Test_R2_Adj
0	Bagging Regressor PO	1487.64	7453.85	55559907.37	1.00	1.00	2603.34	9690.98	93915009.74	1.00	1.00
1	Random Forest Regressor PO	1127.67	4156.39	17275585.75	1.00	1.00	2638.11	9986.00	99720111.70	1.00	1.00
2	XGBoost Regressor PO	2592.32	4618.29	21328582.01	1.00	1.00	4077.38	14348.18	205870180.30	1.00	1.00
3	ElasticNet Regression PO	7368.33	17250.51	297579976.76	0.99	0.99	7301.64	15940.33	254094140.39	0.99	0.99
4	Ridge Regression PO	7412.78	17245.39	297403400.48	0.99	0.99	7363.20	15946.75	254298917.36	0.99	0.99
5	Lasso Regression PO	7444.55	17244.28	297365285.57	0.99	0.99	7408.18	15954.10	254533449.59	0.99	0.99
6	Linear Regression PO	7446.19	17244.27	297364997.30	0.99	0.99	7411.50	15954.72	254553127.18	0.99	0.99
7	KNN Regressor PO	0.00	0.00	0.00	1.00	1.00	4144.03	21196.28	449282377.47	0.99	0.99
8	Decision Tree Regressor PO	1470.79	2928.33	8575134.81	1.00	1.00	6852.70	43040.16	1852455336.60	0.96	0.96
9	AdaBoost Regressor PO	36305.16	50177.41	2517772487.45	0.95	0.95	36183.79	49823.88	2482418892.67	0.95	0.95
10	Support Vector Regressor PO	61136.33	234146.66	54824656439.42	-0.04	-0.04	59428.42	228163.06	52058383855.24	-0.05	-0.07

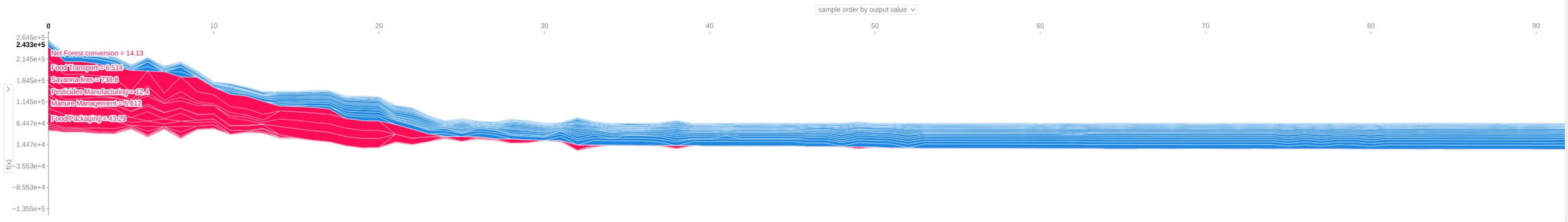
- **Full** dataset – **Optimizovani** modeli

SHAP waterfall – kod posmatrane opservacije *Total Population* najviše utiče za smanjenje predikcije u odnosu na prosečnu predikciju



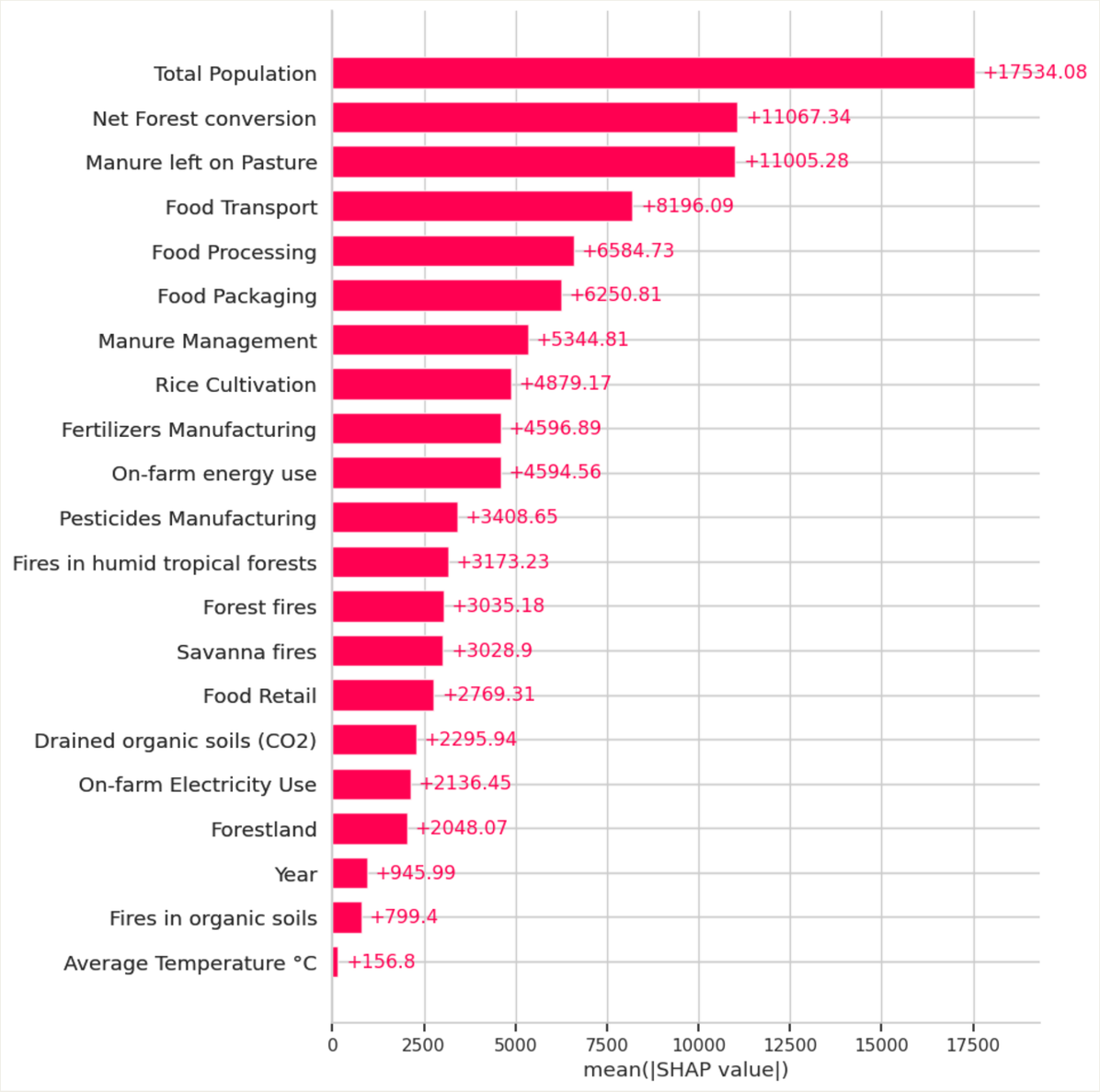
SHAP force plot – kod posmatrane opservacije *Net Forest* najviše utiče za smanjenje predikcije u odnosu na prosečnu predikciju

SHAP stacked force plot – za prvih 100 opservacija u skupu podataka

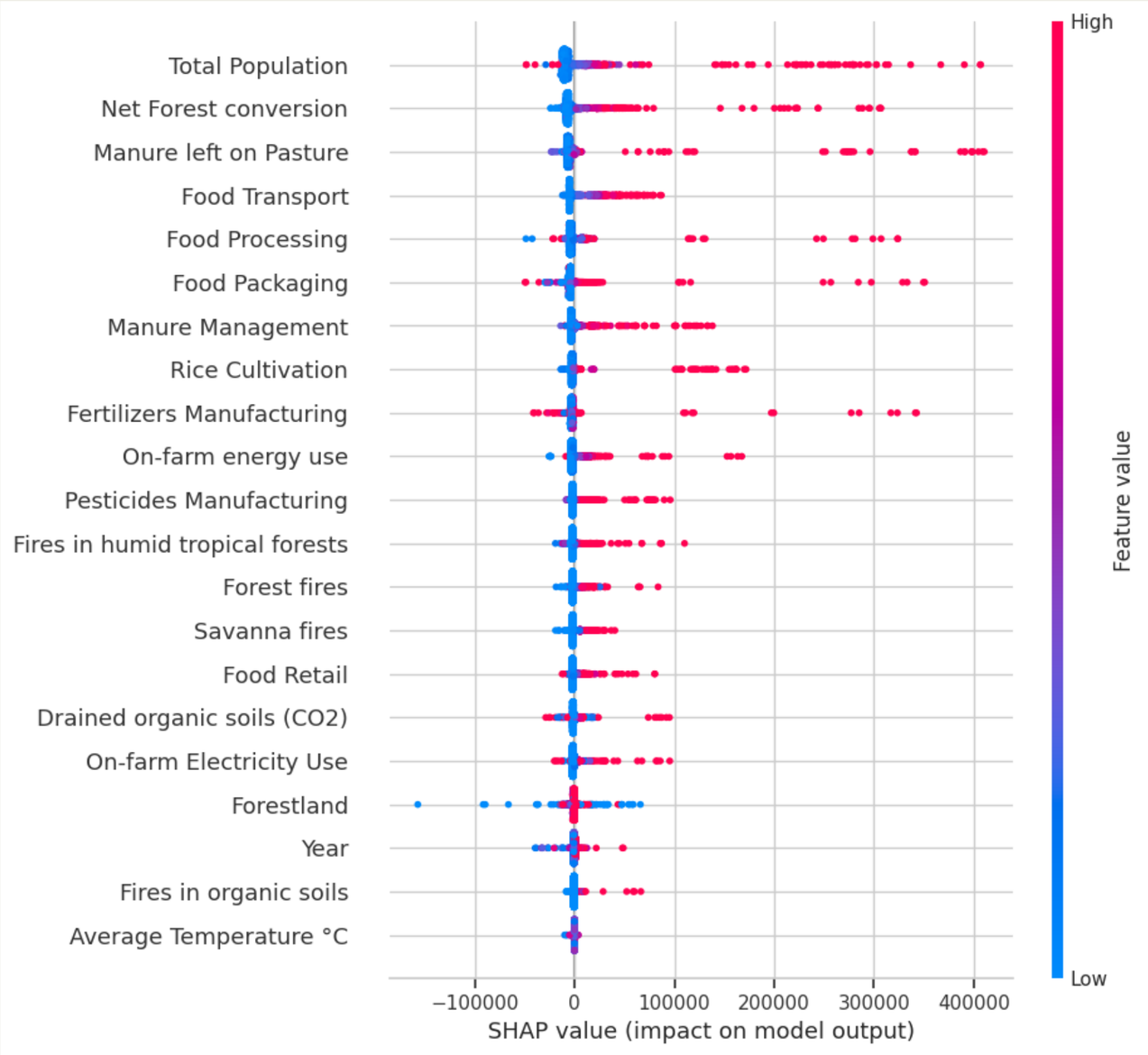


- Opservacija sa najvećom prediktovanom vrednošću, u posmatranih 100, ima velike realne vrednosti za:
  - *Net Forest conversion*
  - *Food Transport*
  - *Savanna fires*
  - *Pesticides Manufacturing*
  - *Manure Management*
  - *Food Packaging*

SHAP Absolute Mean – Varijabla sa najviše uticaja kod predikcije je *Total Population*

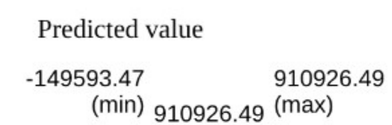


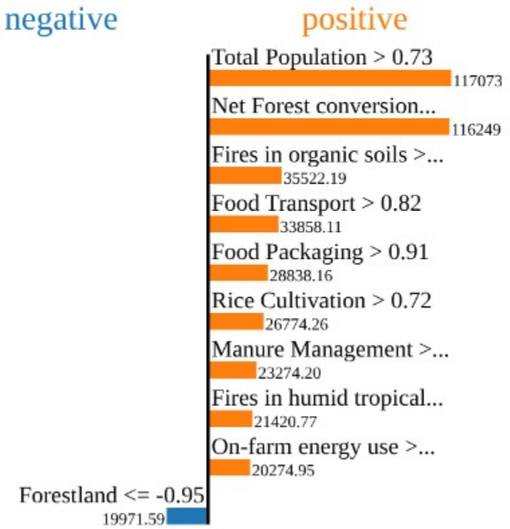
SHAP Beeswarm – Visoke **realne** vrednosti *Total Population* varijable imaju visoke ***SHAP*** vrednosti (*veliki uticaj na predikciju*)



LIME – *lokalna eksploracija pogrešnih predikcija* – Opservacija sa najvećom **kvadratnom greškom** čija predikcija je **precenjena** u odnosu na njenu **realnu** vrednost

```
Index                 680.00
Real Value         708778.78
Predicted Value    910926.49
Name: 0, dtype: float64
```





Feature	Value
Total Population	11.04
Net Forest conversion	175.12
Fires in organic soils	69157.31
Food Transport	4.24
Food Packaging	3.54
Rice Cultivation	50.00
Manure Management	5.13
Fires in humid tropical forests	138.44
On-farm energy use	5.14
Forestland	-184.06

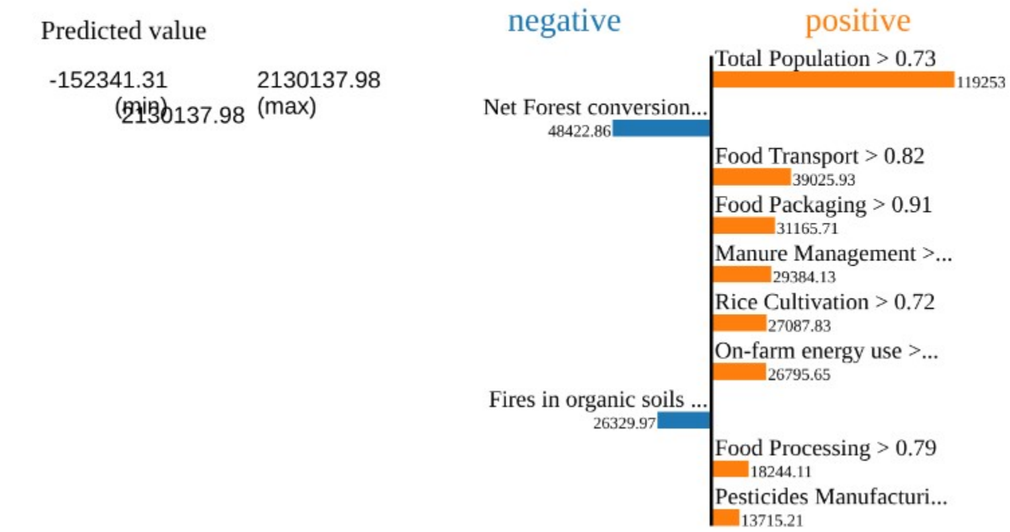
Najveći uticaj (POSITIVE) na precenjivanje predikcije u odnosu na realnu vrednost ove opservacije, imaju:

- *Total* Population
- *Net Forest Conversion*

kao varijable sa najvećim realnim vrednostima kod posmatrane opservacije koje imaju pozitivan uticaj.

LIME – *lokalna eksploracija pogrešnih predikcija* – Opservacija sa najvećom **kvadratnom greškom** čija predikcija je **potcenjena** u odnosu na njenu **realnu** vrednost

```
Index                319.00
Real Value       2229615.22
Predicted Value  2130137.98
Name: 3, dtype: float64
```



Feature	Value
Total Population	74.44
Net Forest conversion	-0.03
Food Transport	20.77
Food Packaging	617.27
Manure Management	58.33
Rice Cultivation	105.40
On-farm energy use	93.14
Fires in organic soils	0.00
Food Processing	236.13
Pesticides Manufacturing	100.49

Najveći uticaj (NEGATIVE) na potcenjivanje predikcije u odnosu na realnu vrednost ove opservacije, imaju:

- *Net Forest Conversion*
- *Fires in organic soils*

kao varijable sa najvećim realnim vrednostima kod posmatrane opservacije koje imaju negativan uticaj.

Hvala!

Pitanja?