Capstone 2 Project

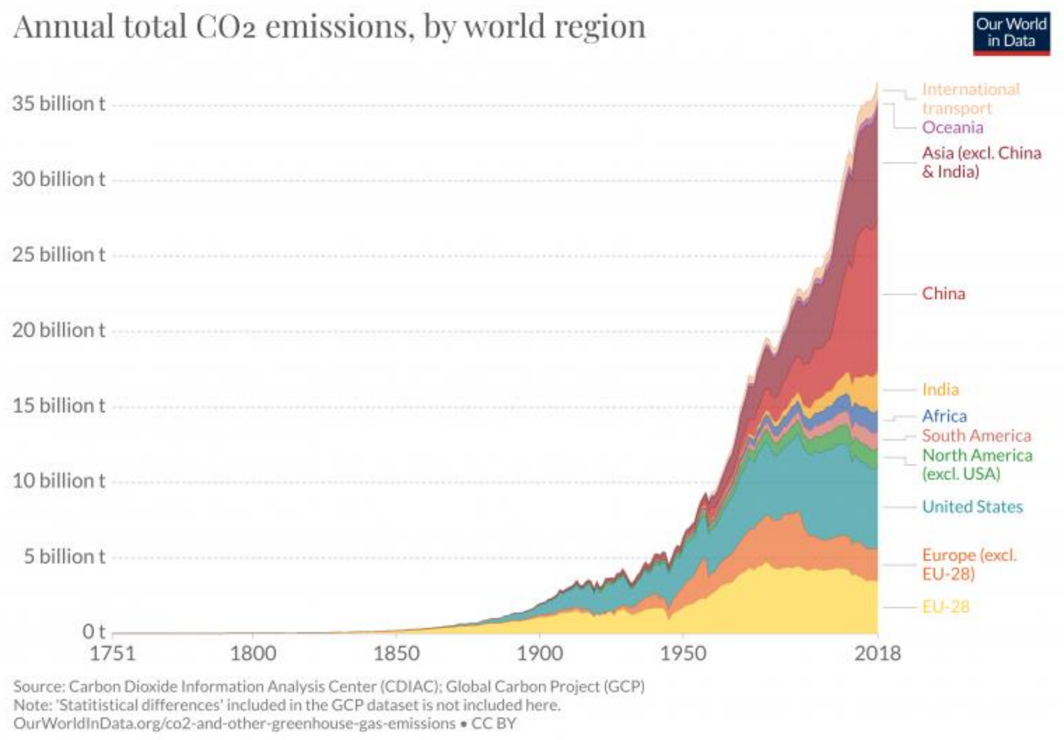
Global CO2 Emissions Prediction

Zizhong Liu
Springboard – May 2021

Introduction

❖ Problem Statement

- The growth of CO_2 emissions is a major contributing factor to the speed of climate change.
- Since 1970, CO_2 emissions have increased by about 90%,.
- In order to prioritize which initiative has the highest impact on reducing CO_2 emissions, governments and industries must be equipped with high-performing, predictive tools for future emissions.



❖ Key Stakeholders

- Politics and policies: ministries, departments, agencies, and directions of national governments;
- Research and education: universities, institutes, research centers, laboratories;
- Supply and demand: industrial companies related to energy, food, air, equipment manufacturing, etc.;
- Organizations, societies, and influencers related to energy, environment, health, etc.

Data Preprocessing

❖ Data Overview

- The CO2 and Greenhouse Gas Emissions dataset is a collection of key metrics maintained by Our World in Data. It includes data on CO2 emissions (annual, per capita, cumulative and consumption-based), other greenhouse gases, energy mix, and other relevant metrics of different countries from the year 1750 - 2019.

- The data set of agriculture and food production are sourced from UNDATA containing the information on agricultural land use and beef production of different countries from the year 1750 - 2019.

```
co2_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23708 entries, 0 to 23707
Data columns (total 7 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   iso_code                        20930 non-null  object
 1   country                         23708 non-null  object
 2   year                            23708 non-null  int64
 3   annual_co2_prod_Megaton         23170 non-null  float64
 4   primary_energy_consumption_10Gwh 6044 non-null  float64
 5   population                      21071 non-null  float64
 6   gdp                             13002 non-null  float64
dtypes: float64(4), int64(1), object(2)
memory usage: 1.3+ MB
```

```
agri_land_data.info()
```

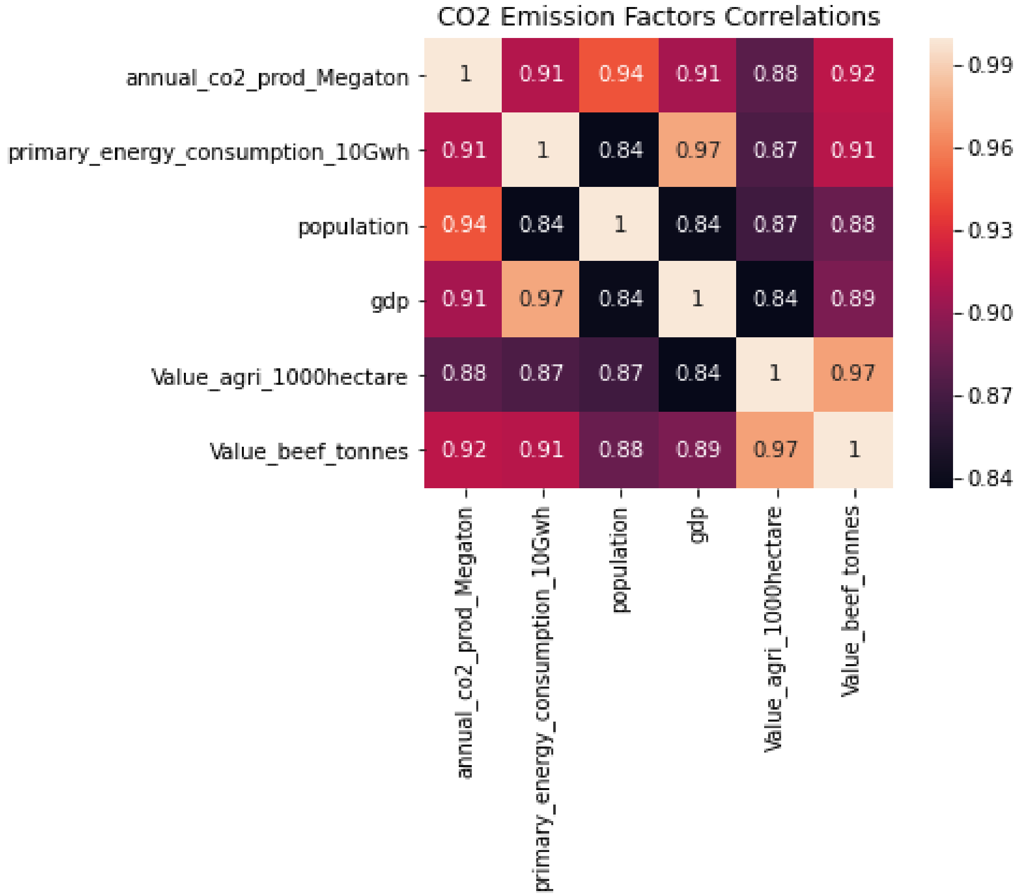
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14378 entries, 0 to 14377
Data columns (total 4 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Country or Area        14377 non-null  object
 1   Year                   14369 non-null  float64
 2   Unit                   14369 non-null  object
 3   Value_agri_1000hectare 14369 non-null  float64
dtypes: float64(2), object(2)
memory usage: 449.4+ KB
```

```
beef_prod_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13197 entries, 0 to 13196
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Country or Area     13196 non-null  object
 1   Year                13194 non-null  float64
 2   Unit                13194 non-null  object
 3   Value_beef_tonnes   13194 non-null  float64
dtypes: float64(2), object(2)
memory usage: 412.5+ KB
```

Data Preprocessing

❖ Data Analysis

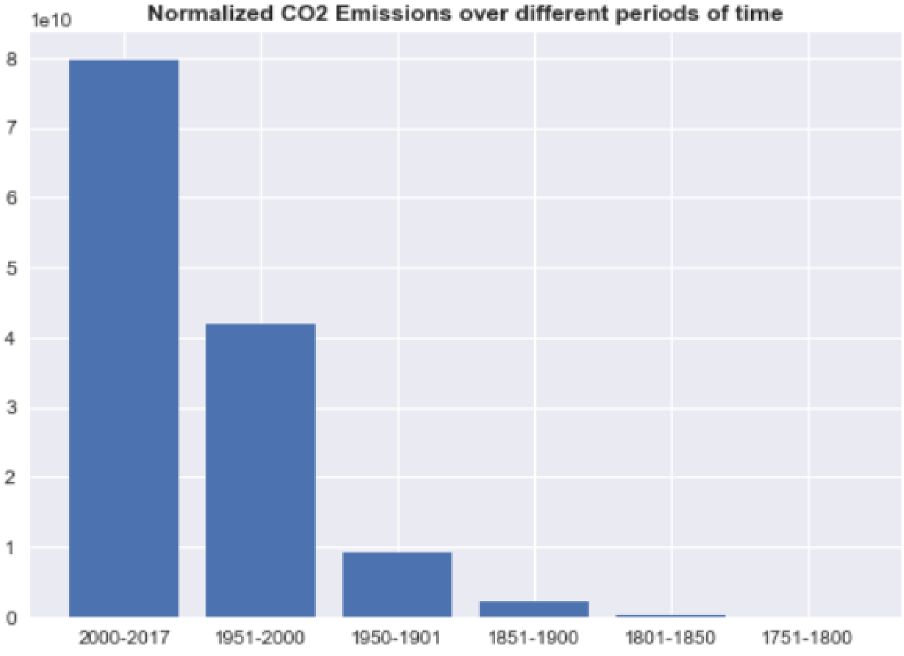


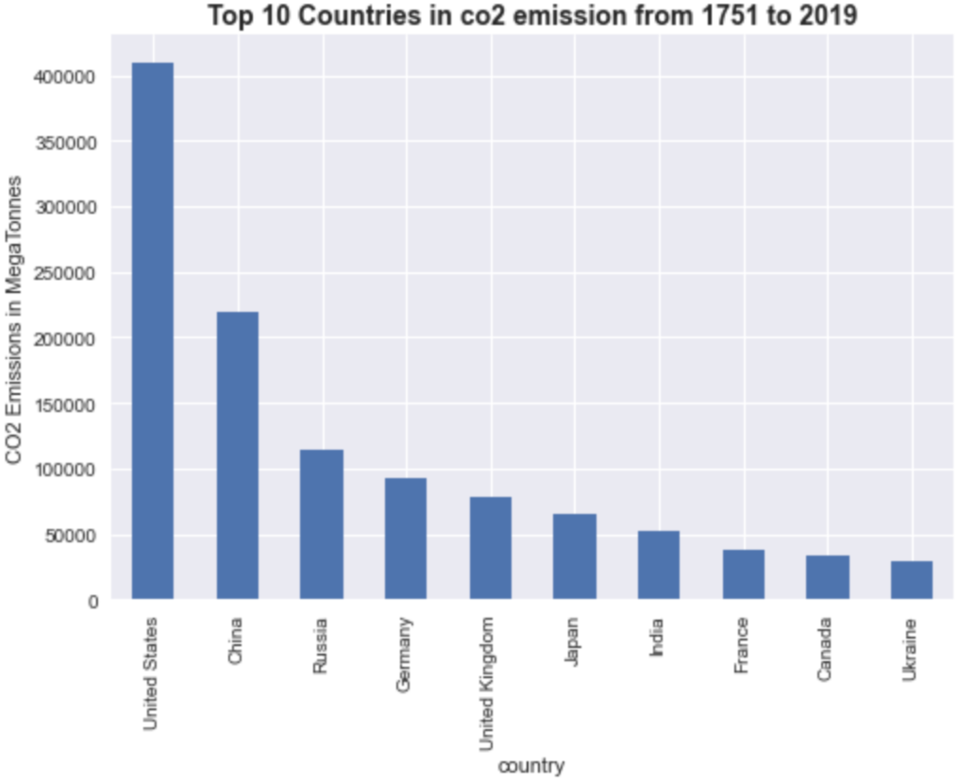
Annual CO2 production has the strong correlation with:

- population (0.94)
- beef production (0.92)
- primary energy consumption and GDP (0.91)

Data Preprocessing

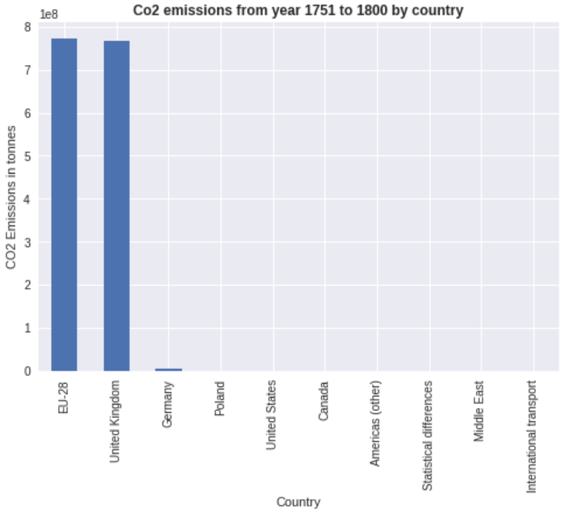
❖ Data Analysis

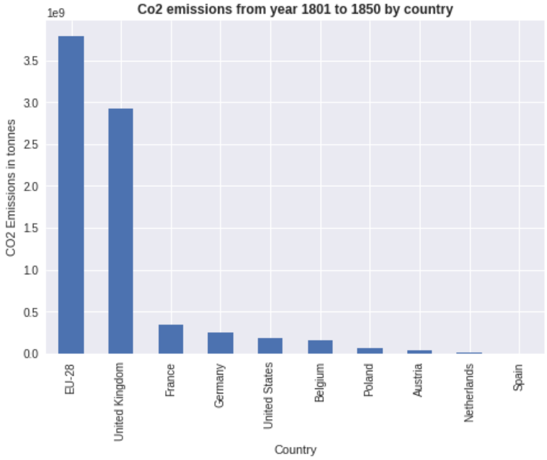


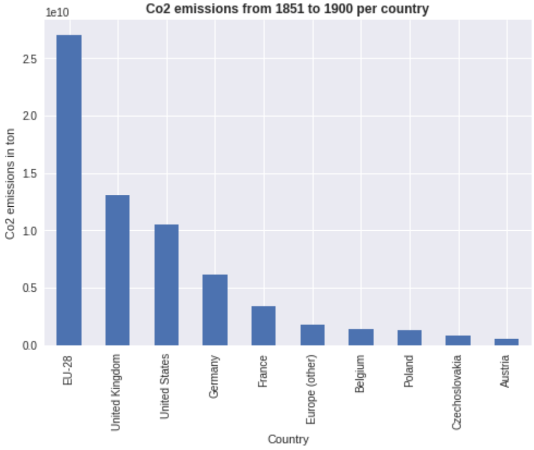


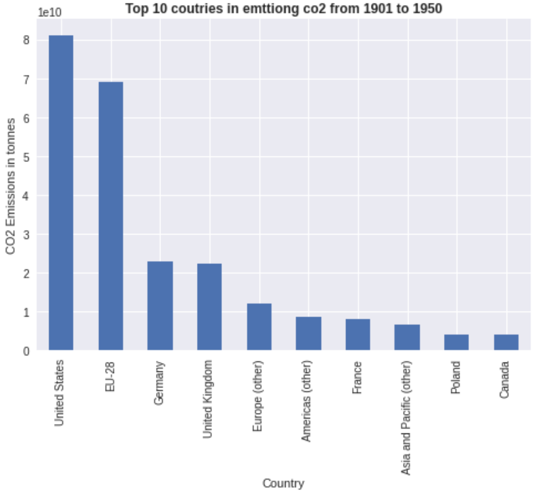
Data Preprocessing

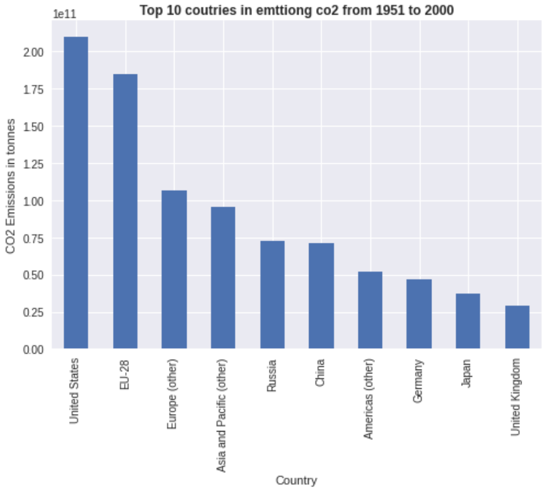
❖ Data Analysis

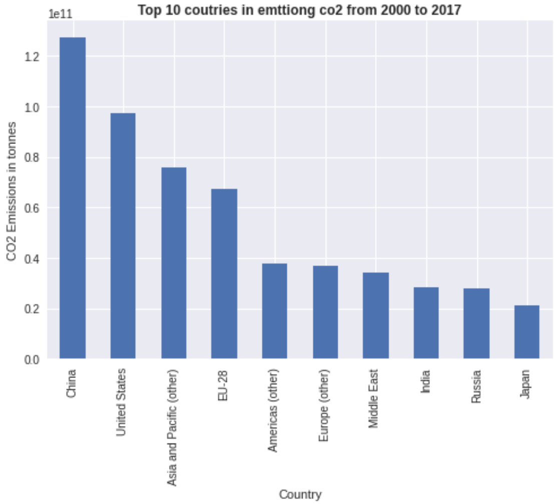












Data Preprocessing

❖ Featuring Engineering

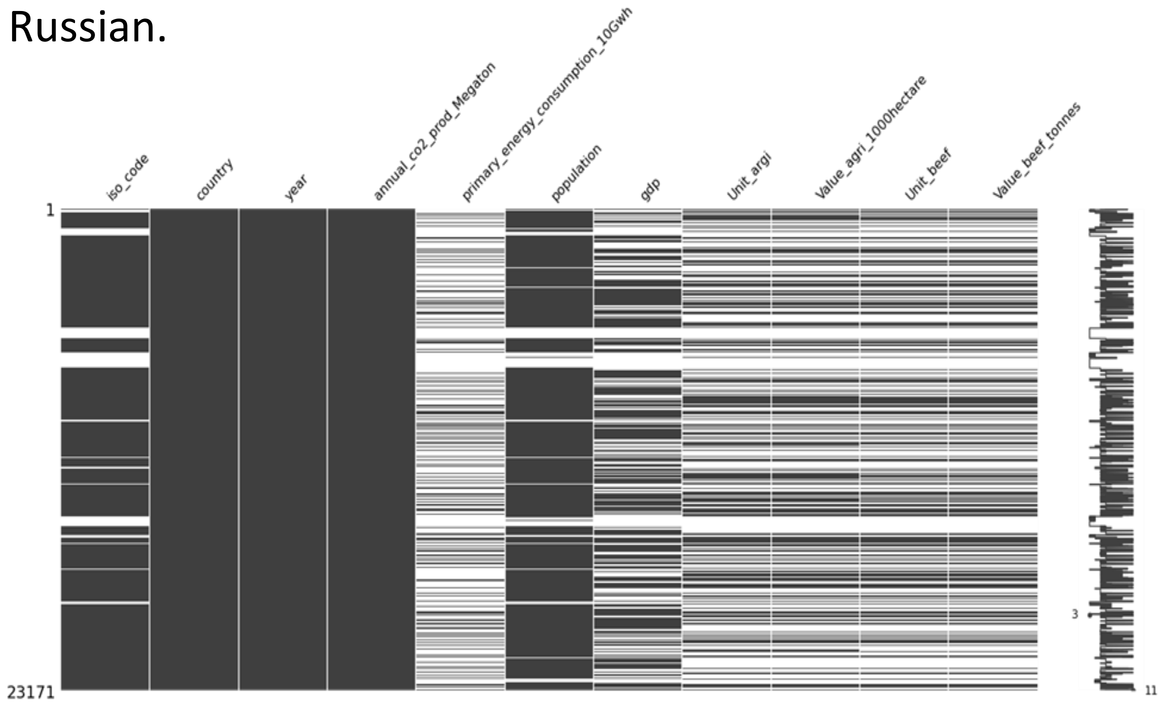
Selected Features

- iso_code: categorical feature.
- country: categorical feature.
- Year: date/time feature
- annual_co2_prod_Megaton: numerical feature.
- primary_energy_consumption_10Gwh: numerical feature.

  Population: numerical feature.
- Gdp: numerical feature.
- Value_agri_1000hectare: numerical feature.
- Value_beef_tonnes: numerical feature.

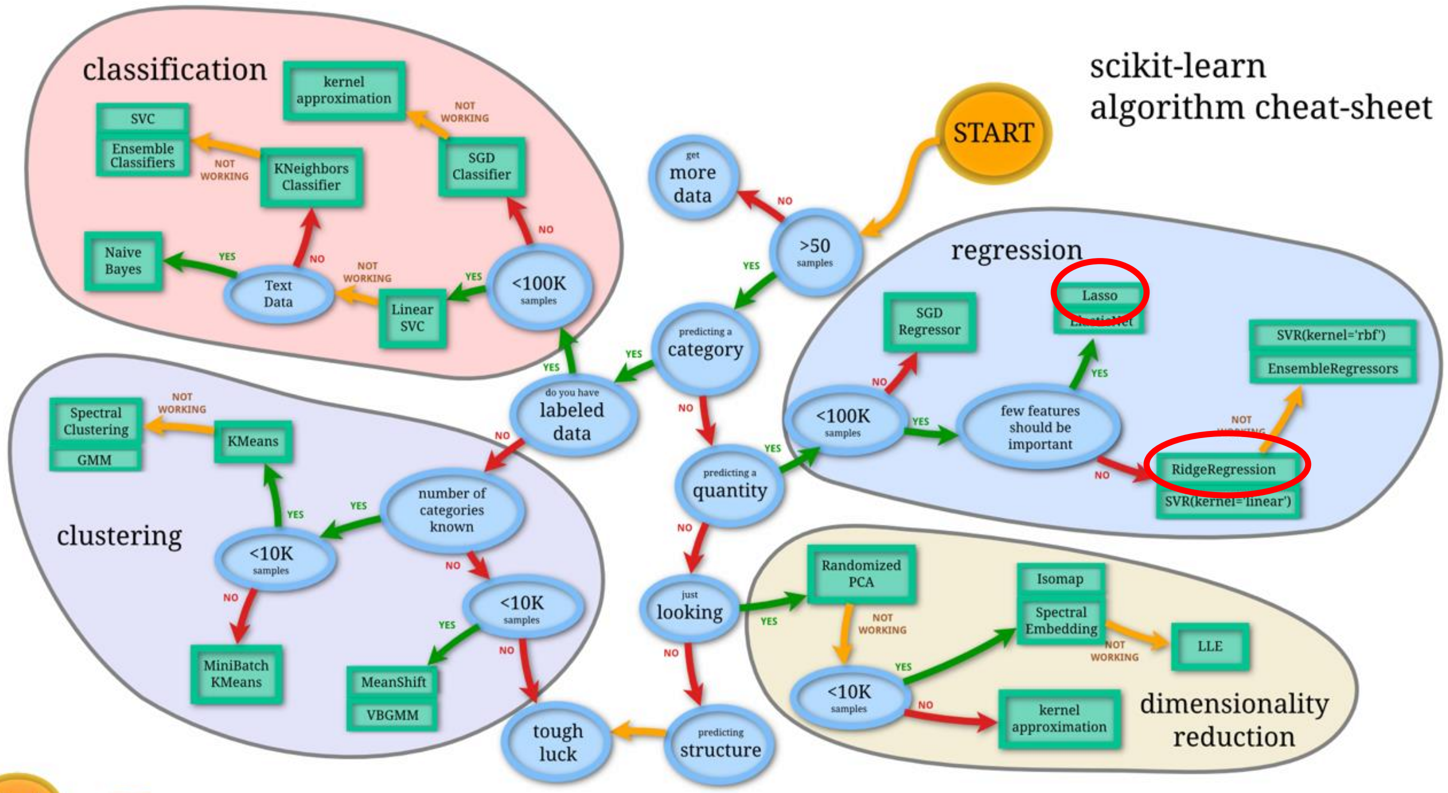
Processing

- NaN data: “fillna” to be replaced by the value of 0
- Thresholds: primary energy > 70k and 110k
- Region names: Africa, Europe, USA, China, India, and Russian.



Modeling

❖ Model Selection



Back

(https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html)

❖ Linear Regression

```
# train the model
ml_fea_5 = LinearRegression(fit_intercept=True, normalize=True)
ml_fea_5.fit(x_fea_5_train,y_fea_5_train)
```

❖ Lasso Regression

```
lasso_rgns = Lasso(alpha=alp,
                   fit_intercept=True,
                   normalize=True,
                   selection='cyclic',
                   max_iter=10000,
                   tol=0.0001,
                   warm_start=False)
```

❖ Ridge Regression

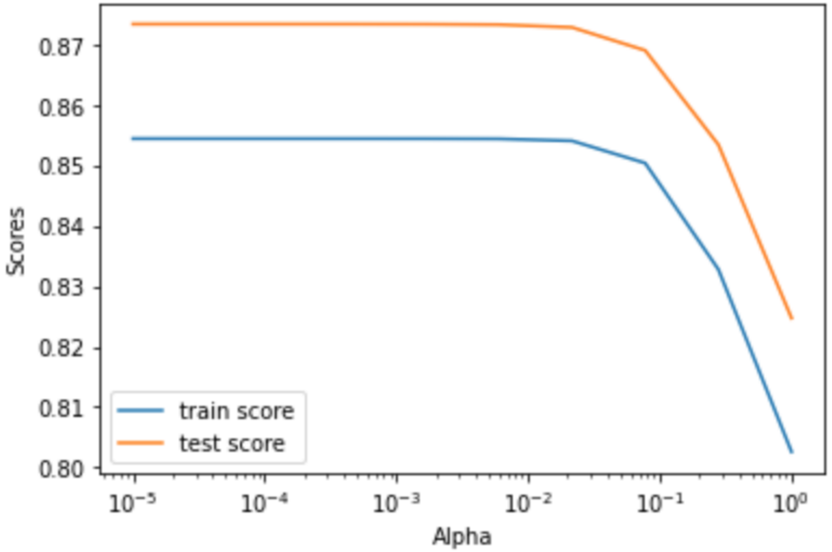
```
for alp in alpha_space:
    ridge_rgns = Ridge(alpha=alp,
                       fit_intercept=True,
                       normalize=True)
```

Results and Discussions

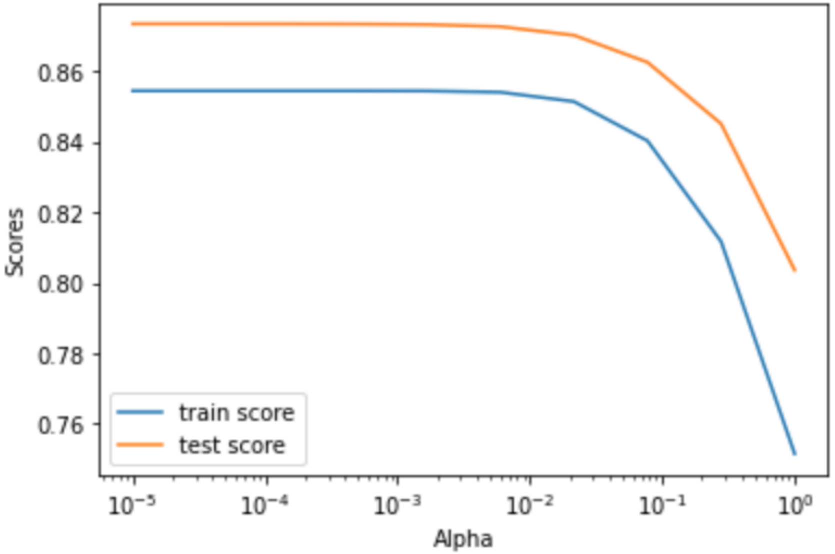
❖ Hypertable

Rank	r2_train	rmse_train	r2_test	rmse_test ↓	alpha	incpt	description
1	0.8544	554.3107	0.8734	581.6483	0.00E+00	-532.32	Linear
2	0.8544	554.3107	0.8734	581.6486	1.00E-05	-532.30	Lasso
3	0.8544	554.3107	0.8734	581.6493	3.60E-05	-532.26	Lasso
4	0.8544	554.3107	0.8734	581.6508	1.00E-05	-532.31	Ridge
5	0.8544	554.3107	0.8734	581.6520	1.29E-04	-532.11	Lasso

❖ Lasso Regression

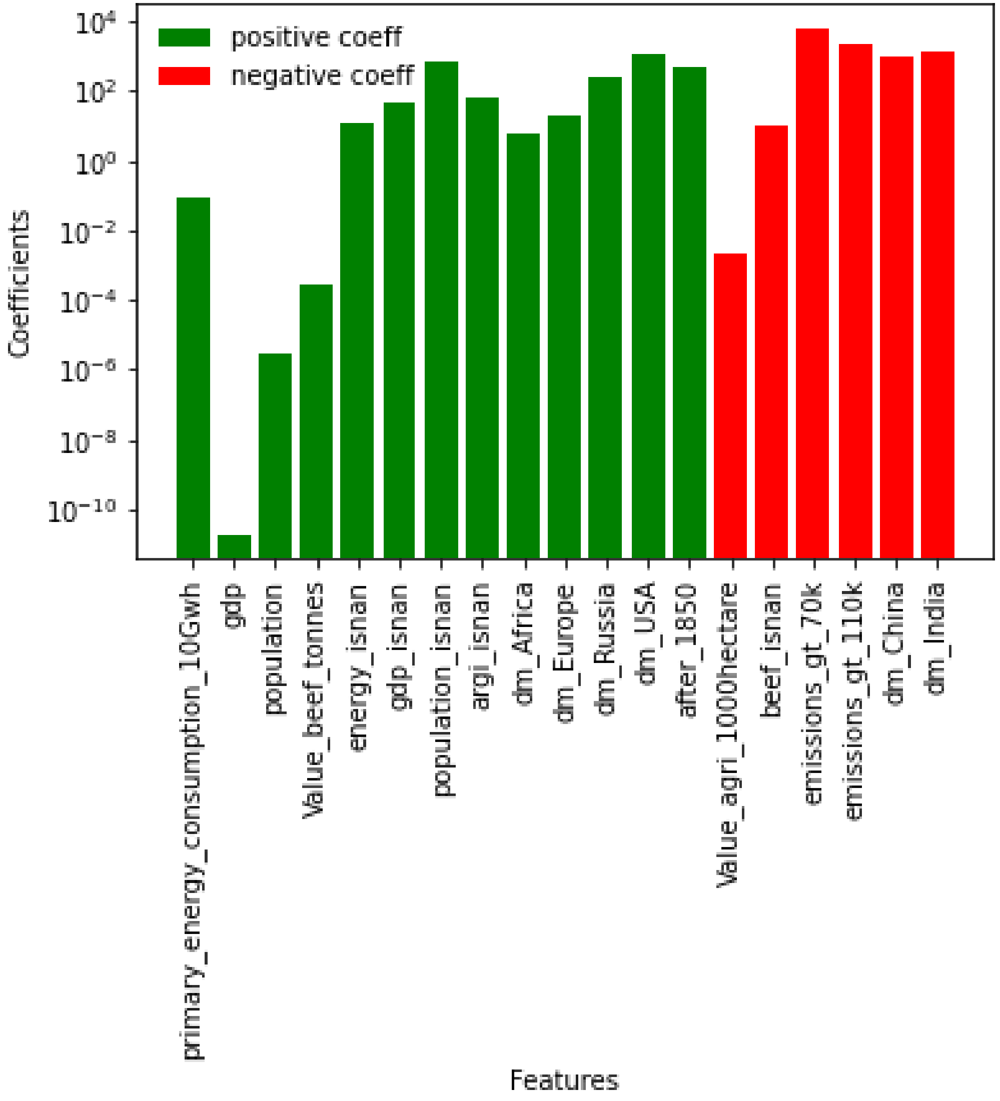


❖ Ridge Regression



Results and Discussions

❖ Coefficients



Observations

- CO2 emission is positively related to primary energy consumption, GDP, population, and beef production, but reversely related to agriculture land use.

- United States, Europe, Russian, and Africa contributions are positively related to CO2 emission, but China and India are negatively related to CO2 emission.

- The possible explanation can be:
  - Low contribution of CO2 emission per person of China and India due to huge population in these two countries causes low per capita CO2 emission.
  - China and India have larger agriculture land use and less beef production.

Conclusions

- This work investigated CO_2 emissions by considering five main factors, including primary energy consumption, GDP, population, agriculture land use, and beef production.
- Three machine learning regression models (linear, Lasso, and Ridge) are applied to the training and test data set in order to predict CO_2 emission.
- The results show that multivariate linear regression is the best performance model for this data set.
- The coefficients of each feature show that CO_2 emission is positively related to primary energy consumption, GDP, population, and beef production but agriculture land use has a negative effect on CO_2 emission.
- The dummy variable coefficients show that United States, Europe, Russian, and Africa contributions are positively related to CO_2 emission, but China and India are negatively related to CO_2 emission.

- **Future works:**
  - Add geological features based on the location of countries.
  - Add temporal features based on time period.
  - Standardize the features. It is important to standardize the features by removing the mean and scaling to unit variance.
  - Consider extra ML/AI models, including NLP and ANN.