Global CO_2 Emissions Prediction

1. Introduction

1.1 Problem Statement

The growth of CO2 emissions is a major contributing factor to the speed of climate change. Global carbon emissions from fossil fuels have significantly increased since 1900. Since 1970, CO2 emissions have increased by about 90%, with emissions from fossil fuel combustion and industrial processes contributing about 78% of the total greenhouse gas emissions increase from 1970 to 2011. Agriculture, deforestation, and other land-use changes have been the second-largest contributors. Emissions of non-CO2 greenhouse gases have also increased significantly since 1900. In order to reduce CO2 emissions, governments and industries must play an active role to curb energy consumption activities most related to emissions growth. In the current energy consumption landscape, a lot of emphases is also placed on the growing consumption of renewable energy sources (e.g., wind, solar). In order to prioritize which initiative has the highest impact on reducing CO2 emissions, governments and industries must be equipped with high-performing, predictive tools for future emissions.

1.2 Key Stakeholders

Potential parties that could be interested in this project include:

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

2. Data Preprocessing

2.1 Data Overview

Source data obtained for this project contains information on different kinds of greenhouse gas emissions, energy consumption, agriculture, and food production. The CO_2 and Greenhouse Gas Emissions dataset is a collection of key metrics maintained by Our World in Data. It is updated regularly and 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.

The features and corresponding information contained in the raw CO_2 emission data set is shown in the following figures:

```
co2_raw_data.columns
```

```
Index(['iso_code', 'country', 'year', 'annual_co2_prod_Megaton',
       'co2_growth_prct', 'co2_growth_abs', 'consumption_co2', 'trade_co2',
       'trade_co2_share', 'co2_per_capita', 'consumption_co2_per_capita',
       'share_global_co2', 'cumulative_co2', 'share_global_cumulative_co2',
       'co2_per_gdp', 'consumption_co2_per_gdp', 'co2_per_unit_energy',
       'cement_co2', 'coal_co2', 'flaring_co2', 'gas_co2', 'oil_co2',
       'other_industry_co2', 'cement_co2_per_capita', 'coal_co2_per_capita',
       'flaring_co2_per_capita', 'gas_co2_per_capita', 'oil_co2_per_capita',
       'other_co2_per_capita', 'share_global_coal_co2', 'share_global_oil_co2',
       'share_global_gas_co2', 'share_global_flaring_co2',
       'share_global_cement_co2', 'cumulative_coal_co2', 'cumulative_oil_co2',
       'cumulative_gas_co2', 'cumulative_flaring_co2', 'cumulative_cement_co2',
       'share_global_cumulative_coal_co2', 'share_global_cumulative_oil_co2',
       'share_global_cumulative_gas_co2',
       'share_global_cumulative_flaring_co2',
       'share_global_cumulative_cement_co2', 'total_ghg', 'ghg_per_capita',
       'methane', 'methane_per_capita', 'nitrous_oxide',
       'nitrous_oxide_per_capita', 'primary_energy_consumption_10Gwh',
       'energy_per_capita', 'energy_per_gdp', 'population', 'gdp'],
      dtype='object')
```

```
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
```

The features and corresponding information contained in the raw agricultural land use data set is shown in the following figures:

```
agri_land_raw_data.columns
```

```
Index(['Country or Area', 'Element', 'Year', 'Unit', 'Value_agri_1000hectare',
       'Value Footnotes'],
      dtype='object')
```

```
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
```

The features and corresponding information contained in the raw beef production data set is shown in the following figures:

```
beef_prod_raw_data.columns
```

```
Index(['Country or Area', 'Element', 'Year', 'Unit', 'Value_beef_tonnes',
       'Value Footnotes'],
      dtype='object')
```

```
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
```

As shown in the above figures, two important considerations can be proposed and need to be handled using the data cleaning method before building machine learning models upon that:

1) The CO_2 data set contains excessive features (columns). Which ones are important key features? And which one is the target feature?
2) It seems many data are missing. How to deal with the missing data?

2.2 Data Processing

In the last section, two important considerations are proposed and need to be addressed.

Firstly, the CO_2 data set includes CO2 emissions by annual, per capita, cumulative, and consumption-based, and other greenhouse gases, energy mix, and other relevant metrics of different countries from the year 1750 - 2019. The objective of this project is to use machine learning methods to predict annual CO_2 production ("annual_co2_prod_Megaton"), which is the target feature. The features of primary energy consumption, population, GDP contained in this dataset are relevant and crucial for predicting CO_2 emissions. Accordingly, by joining the data sets of CO_2 emissions, agricultural land use, and beef production, the new CO_2 emission data set are shown in the following figures:

```
co2_data.columns
```

```
Index(['iso_code', 'country', 'year', 'annual_co2_prod_Megaton',
       'primary_energy_consumption_10Gwh', 'population', 'gdp', 'Unit_argi',
       'Value_agri_1000hectare', 'Unit_beef', 'Value_beef_tonnes',
       'energy_isnan', 'gdp_isnan', 'population_isnan', 'argi_isnan',
       'beef_isnan'],
      dtype='object')
```

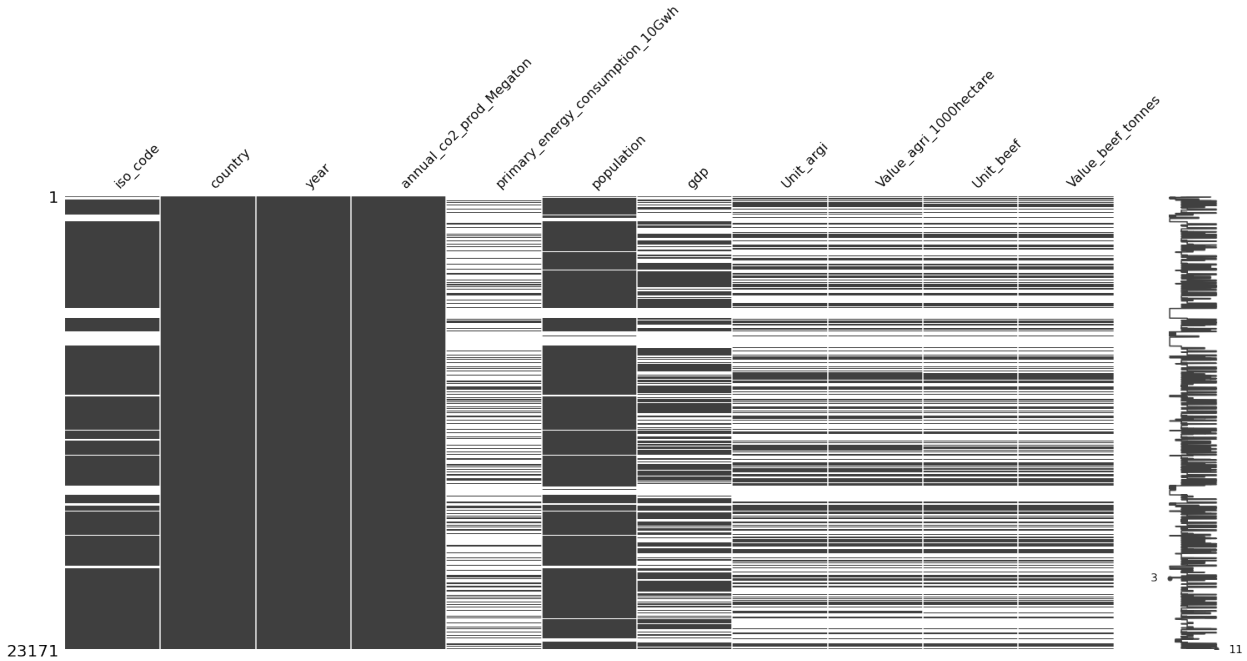
```
co2_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 23171 entries, 0 to 23708
Data columns (total 16 columns):
iso_code                            20440 non-null object
country                             23171 non-null object
year                                23171 non-null datetime64[ns]
annual_co2_prod_Megaton             23171 non-null float64
primary_energy_consumption_10Gwh    6045 non-null float64
population                          20583 non-null float64
gdp                                 12973 non-null float64
Unit_argi                           9818 non-null object
Value_agri_1000hectare              9818 non-null float64
Unit_beef                           9377 non-null object
Value_beef_tonnes                   9377 non-null float64
energy_isnan                        23171 non-null bool
gdp_isnan                           23171 non-null bool
population_isnan                    23171 non-null bool
argi_isnan                          23171 non-null bool
beef_isnan                          23171 non-null bool
dtypes: bool(5), datetime64[ns](1), float64(6), object(4)
memory usage: 2.2+ MB
```

```
co2_data.describe()
```

	annual_co2_prod_Megaton	primary_energy_consumption_10Gwh	population	gdp	Value_agri_1000hectare	Value_beef_tonnes
count	23171.000000	6045.000000	2.058300e+04	1.297300e+04	9.818000e+03	9.377000e+03
mean	270.234760	1638.034068	6.053309e+07	4.405589e+11	7.341125e+04	7.789382e+05
std	1509.880287	9665.709679	3.773372e+08	3.670729e+12	3.935006e+05	4.531116e+06
min	-1.165000	0.208000	1.000000e+03	6.378000e+07	3.000000e-01	0.000000e+00
25%	0.546000	46.326000	1.433000e+06	8.911988e+09	3.340000e+02	3.233000e+03
50%	5.170000	148.688000	5.004000e+06	2.946853e+10	3.495500e+03	4.080000e+04
75%	44.785000	518.789000	1.632450e+07	1.220000e+11	2.052425e+04	1.705510e+05
max	36441.388000	153848.433000	7.713468e+09	1.065610e+14	4.882180e+06	7.160131e+07

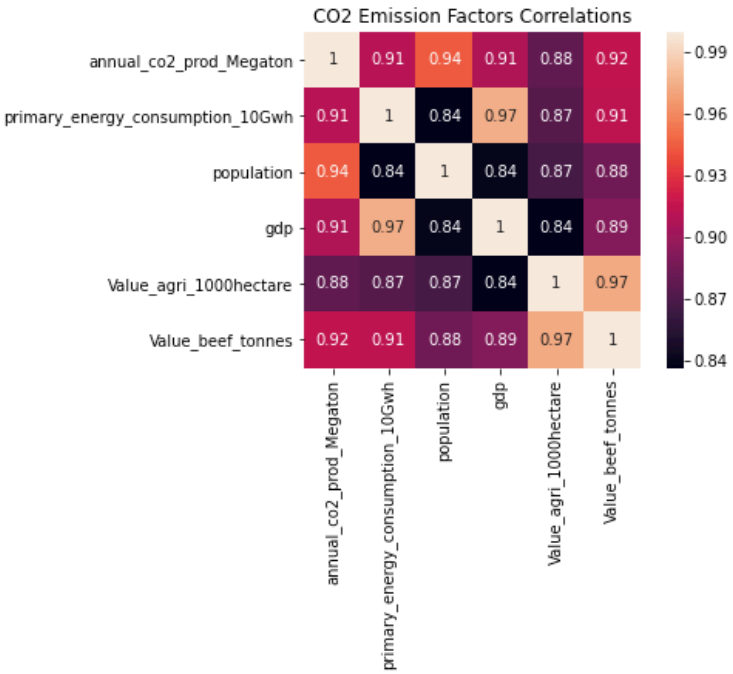
Secondarily, it seems there are a lot of missing values. To visualize the missing data, the package of "missingno" is imported and utilized. The results are shown in the following figure:



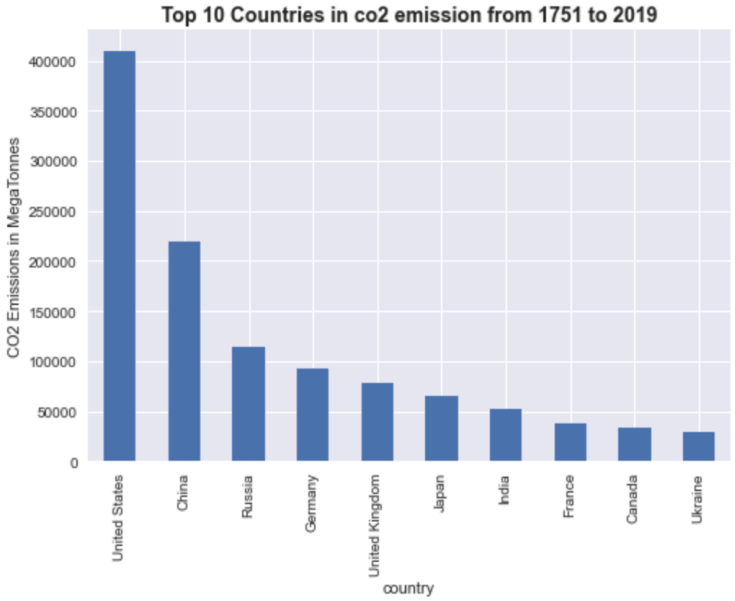
The results show that, compared with the columns of country, year, annual_co2_prod_Megaton, and population, there is a significant amount of data is missing in the rest columns. The missing data mostly belongs to the early time data of different countries due to the lack of recording intentions and techniques. The method to deal with NaN is elaborated in the following section.

2.3 Data Analysis

With the data preprocessing finished, exploratory data analysis (EDA) can be utilized for us to better understand the data. The pairwise correlation of the main 6 features are calculated and investigated. As shown in the heatmap below, the highest Pearson's correlation coefficient appears in the pairs of GDP vs. primary energy consumption and agriculture land use vs. beef production. In addition, annual CO2 production has the strong correlation with population (0.94) and beef production (0.92). The correlations between annual CO2 production with primary energy consumption and GDP are slightly weaker (0.91).

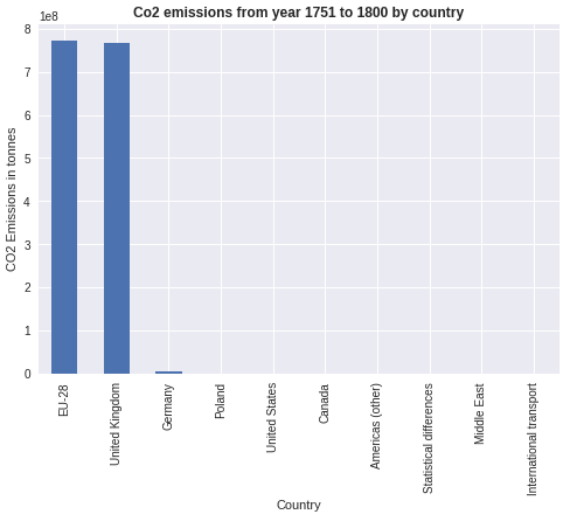


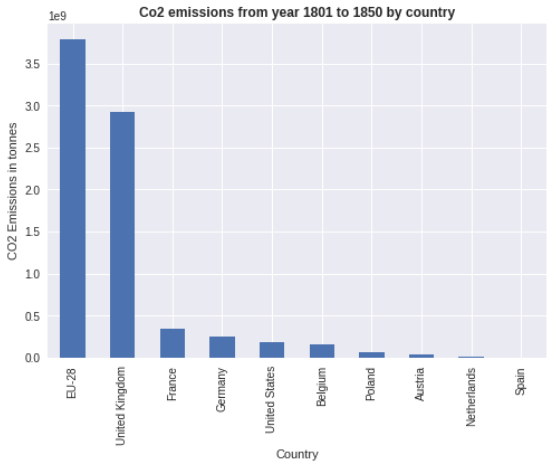
The figure below shows the top 5 countries in cumulative CO_2 emissions from 1751 to 2017. The results show the US has the highest cumulative CO_2 emission among all the countries and regions.

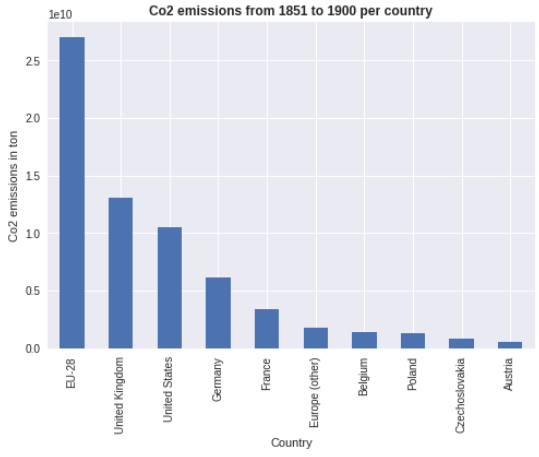


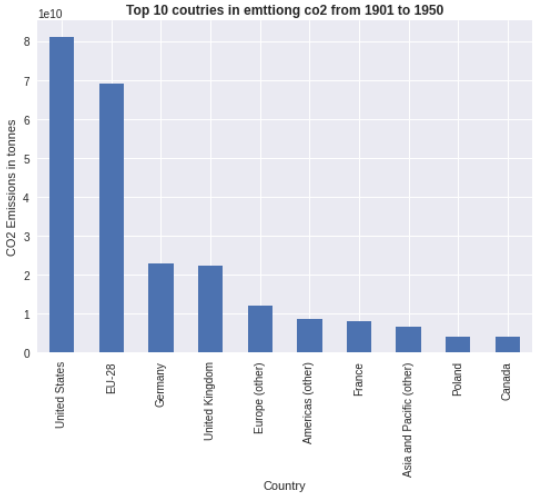
The figures below show the top 10 countries in cumulative CO_2 emissions within a 50-year period from 1751 to 2017. The result shows during the pre-industrial stage (1750 - 1850), only the UK had significant CO_2 emissions. With the start of the industrial revolution (the 1850s), the CO_2

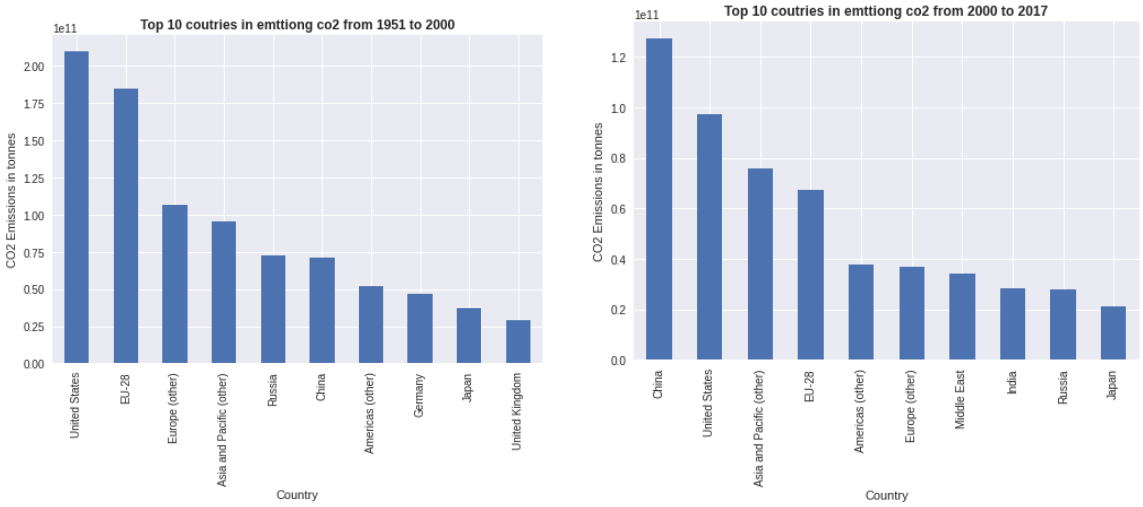
emission of the US and Germany increased rapidly and exceeded the UK during 1901 – 1950. Starting from late 20th, China and India began their first industrial revolutions and appeared on the list of top 10 countries after 1951 and 2000, respectively, while others, such as the United States and western Europe, began undergoing "second" industrial revolutions by the late 19th century. In 21st, China exceeded the US and became the No. 1 CO_2 emission countries.



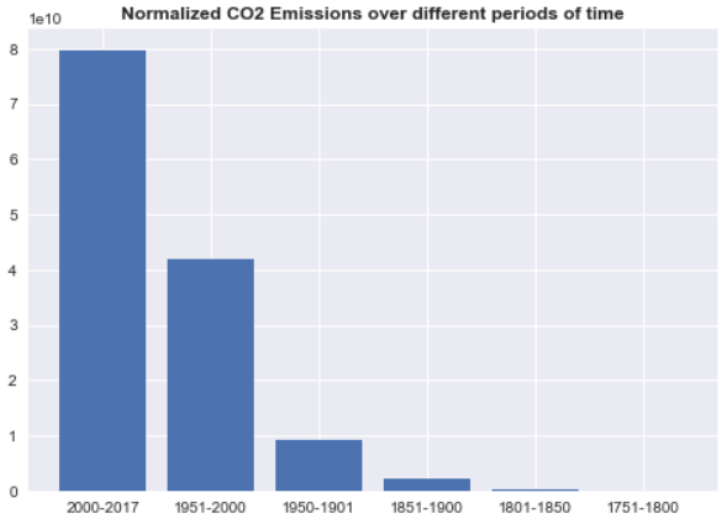








In addition, the figure below compares the total CO2 emissions within different periods from 1751 to 2017 normalized by the period length, which indicates the average CO_2 emissions per year for different period. We can observe the exponential increase of CO_2 emission with time. It is noted that the normalized CO_2 emission of 2000 – 2017 is two times higher than that of of 1951-2000.



2.4 Feature Engineering

In order to prepare the datasets and solve the problem of missing data, the featuring engineering should be performed on the raw data sets. The cleaned dataset of CO2 emission contains the following features which will be used for the train-test dataset:

- iso_code: categorical feature. The ISO country codes are internationally recognized codes that designate every country and most of the dependent areas a two-letter combination or a three-letter combination
- country: categorical feature.
- Year: date/time feature
- annual_co2_prod_Megaton: numerical feature. annual CO2 emissions of each country in a megaton.
- primary_energy_consumption_10Gwh: numerical feature. Annual primary energy consumption of each country in 10Gwh.
- Population: numerical feature.
- Gdp: numerical feature.
- Value_agri_1000hectare: numerical feature. Annual agricultural land use of each country in 1000 hectares.
- Value_beef_tonnes: numerical feature. Annual beef production of each country in tonnes.

However, due to a large number of missing data, especially in annual_co2_prod_Megaton, NaN data need to be treated and some new features need to be created to fully use the dataset.

First, NaN data is treated by using the function of "fillna" to be replaced by the value of 0.

Second, several columns of Booleans variables are created corresponding to the numerical features which representing if the numerical features contain NaN data.

Third, two columns of Booleans variables are created especially for primary energy consumption. The two columns are the threshold for primary energy consumption corresponding to the CO2 emission higher than 70k and 110k respectively.

Last but not least, a column of "region_name" is created for each example which contains categorical features of 6 regions/countries that this work will focus on, which are Africa, Europe, USA, China, India, and Russian. The region names are assigned to each example corresponding to the ISO code, and the countries not in these 6 regions are assigned with "Other". In addition, the columns of dummy variables are created based on the "region_name".

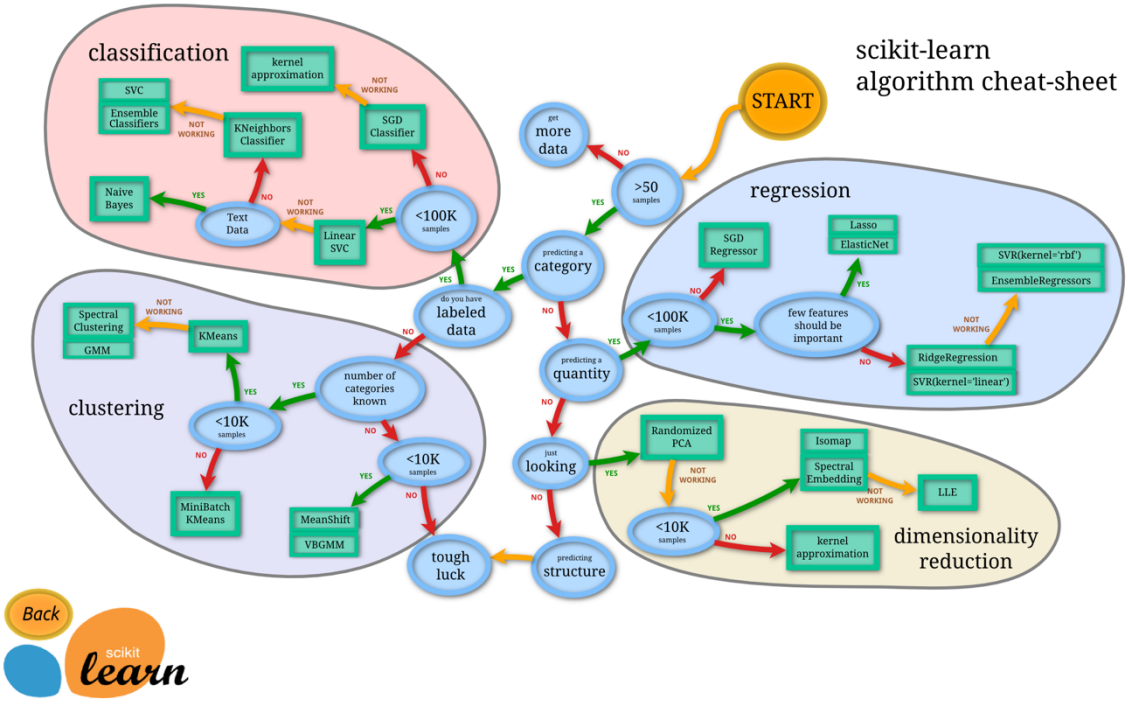
In conclusion, the CO2 dataset after applying feature engineering contains features:

```
Index(['index', 'iso_code', 'region_name', 'country', 'year', 'annual_co2_
prod_Megaton', 'primary_energy_consumption_10Gwh', 'population', 'gdp', 'V
alue_agri_1000hectare', 'Value_beef_tonnes', 'energy_isnan', 'gdp_isnan',
'population_isnan', 'argi_isnan', 'beef_isnan', 'emissions_gt_70k', 'emiss
```

```
ions_gt_110k', 'dm_China', 'dm_Europe', 'dm_India', 'dm_Russia', 'dm_USA
'].
```

3. Modeling

The primary goal of this work is to build and evaluate machine learning models to predict CO2 emissions based on selected features. Often, the hardest part of solving a machine learning problem can be finding the right estimator for the job. Different estimators are better suited for different types of data and different problems. The flowchart below is designed to give users a bit of a rough guide on how to approach problems with regard to which estimators to try on your data. The topic investigated in this work is a problem of regression. Therefore, 3 machine-learning algorithms, multivariate, Lasso, and Ridge Regression were selected and applied to train-test the model separately.



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

Multivariate Regression is a method used to measure the degree at which more than one independent variable (predictors) and more than one dependent variable (responses), are linearly related. The method is broadly used to predict the behavior of the response variables associated to changes in the predictor variables, once a desired degree of relation has been established. In this

work, multivariate regression is tested with hyper-parameters of fit_intercept=True, normalize=True.

```
# 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 is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. The tuning parameter, α controls the strength of the L1 penalty. α is basically the amount of shrinkage: When α = 0, no parameters are eliminated. The estimate is equal to the one found with linear regression. As α increases, more and more coefficients are set to zero and eliminated (theoretically, when α = infinity, all coefficients are eliminated). In addition, as α increases, bias increases; and as α decreases, variance increases. In this work, Lasso regression is applied with hyperparameters shown in the following figure. It is worth noting that 10 alpha values are tested between 1e-5 to 1 in a log space.

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

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where independent variables are highly correlated. Ridge regression considers L2 penalty by adding “squared magnitude” of coefficient as penalty term to the loss function. if α is zero then the ridge regression is equivalent to ordinary least squares. However, if α is very large then it will add too much weight and it will lead to under-fitting. Having said that it’s important how α is chosen. This technique works very well to avoid over-fitting issue. In this work, Ridge regression is applied with hyperparameters shown in the following figure. It is worth noting that 10 alpha values are tested between 1e-5 to 1 in a log space.

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

In the following section, these 3 regression models are implemented and their results are compared.

4. Results and Discussions

A hypertable is established to track the model performance, which contains the following hyperparameters, as shown in the table below.

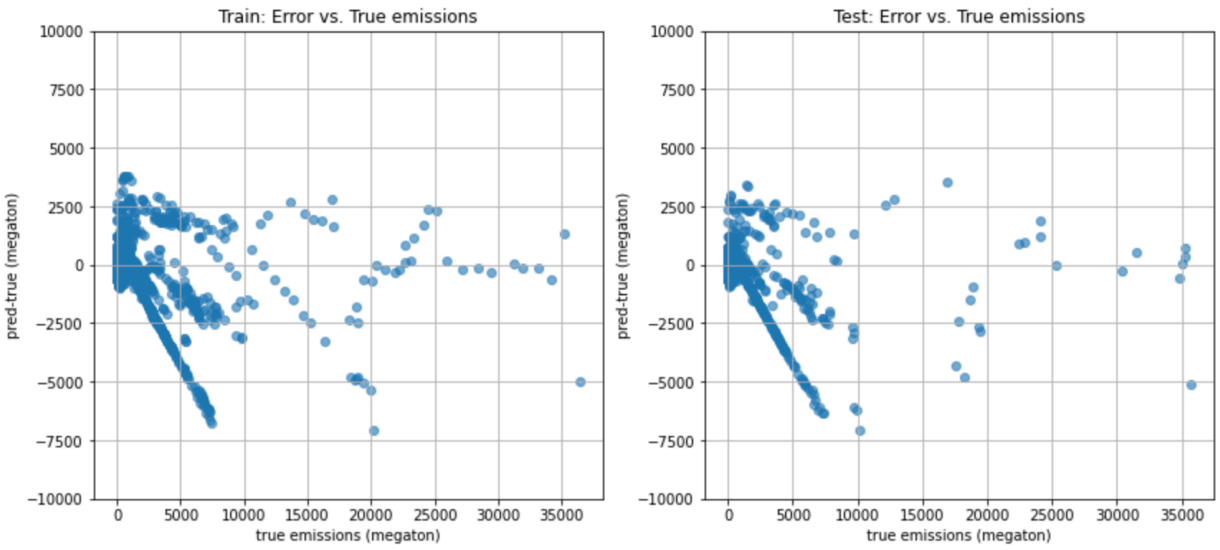
r2_train	rmse_train	r2_test	rmse_test	alpha	coef	incpt	model	features	description

4.1 Multivariate Regression

The R-squared and RMSE of using multivariate regression on train and test data set are calculated and shown in the following table. The results show that the R-squared and RMSE of train and test data set are very close to each other. This indicates there is no obvious bias or variance applying multivariate regression.

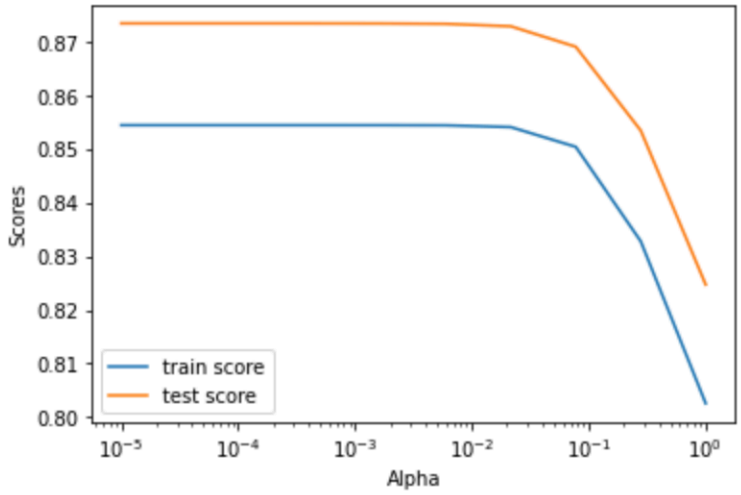
r2_train	rmse_train	r2_test	rmse_test
0.854439	554.310706	0.873422	581.648314

The figures below show the error (predicted value – true value in megaton) vs. true emission. The results also approve that multivariate regression predicts the same trent when it is applied to train and test set. As shown in the figures, the error is significant at low emission region. This is because the lack of data during early time when CO2 emission is low.



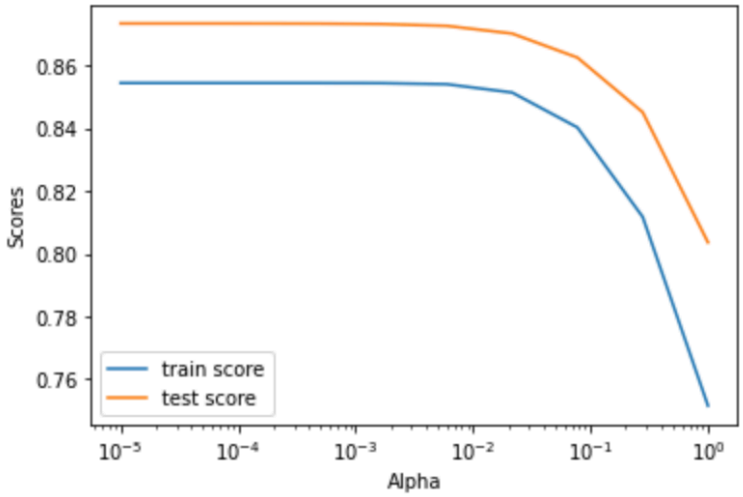
4.2 Lasso Regression

Lasso regression is applied on the training set with L1 regularization parameter alpha = np.logspace(-5, 0, num=10, base=10). The R-squared scores for all the models are obtained for both training and test sets. The obtained R-squared scores for training and test sets are plotted as a function of alpha, as shown in the following figure. The figure shows that the R-squared scores for both data sets keep stable within the alpha range of 1E-5 to 1E-2, but after this, decreases with the increasing alpha. For this case, lasso regression performs better with alpha approaching 0, which indicates that no features should be eliminated, and all of them are important to the prediction of CO2 emission.



4.3 Ridge Regression

Following the similar procedure as Lasso regression, Ridge regression is also applied on the training set with L2 regularization parameter alpha = np.logspace(-5, 0, num=10, base=10). The R-squared scores for all the models are obtained for both training and test sets. The obtained R-squared scores for training and test sets are plotted as a function of alpha, as shown in the following figure. The figure shows that for ridge regression, the R-squared scores for both data sets keep stable within the alpha range of 1E-5 to 5E-3, but after this, decreases quickly with the increasing alpha. α (alpha) in ridge regression is the parameter which balances the amount of emphasis given to minimizing RSS vs minimizing sum of square of coefficients. For this case, ridge regression performs better when alpha approaches 0, which indicates that to predict CO2 emission data set, most amount of emphasis should be given to minimizing RSS, which is the same objective as multivariate linear regression. In addition, it is worth noting that the three machine learning models yield slightly higher R-squared scores in the test data set than in the training data set. This suggest the three machine learning models are slightly biased.



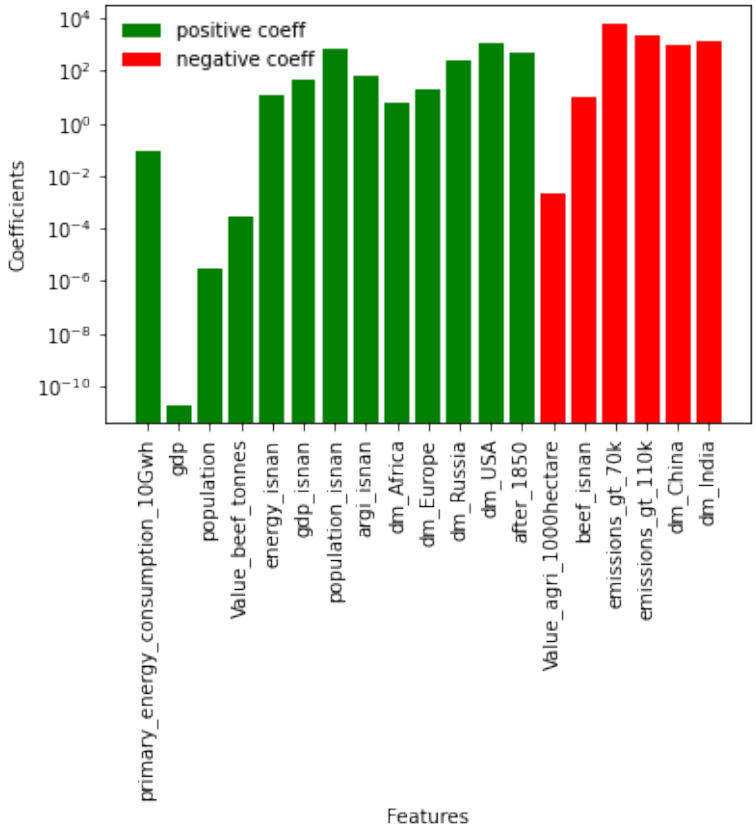
4.4 Best Model

With obtained R-squared and RMSE result, the performance of three models are compared and listed in the order of R-squared and RMSE for the test data set. As shown in the table below, for this CO2 emission case, the best model is multivariate regression (alpha = 0). The best lasso regression model is with alpha = 1E-5 and ranks No. 2nd. The best ridge regression model is also with alpha = 1E-5 but ranks No. 4th.

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	Multivariate

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

The coefficients from the best model, multivariate regression, are investigated and plotted in a bar chart in a semi-log coordinate, as shown in the following figure. The red colored group represents the feature with negative coefficients, and green for positive coefficients. Because the features of CO2 emission dataset were not standardized before regression, only the coefficients of the features with same value range can be compared with each other, such as features of "_isnan" group and dummy variables of "dm_" group with Boolean values. Some observations and conclusions can be obtained from the following figure.



It is obvious that CO2 emission is positively related to primary energy consumption, GDP, population, and beef production. Quantitively, per 10 Gwh primary energy consumption contribute 0.08 megatons of CO2 emission. Per billion GDP increase CO2 emission by 0.02 megatons. Per million population contributes 2.8 megatons of CO2 emission. In addition, per kilotons beef

production causes CO2 emission increases by 0.28 megatons. On the contrary, agriculture land use has a negative effect on CO2 emission. Per 1000 hectares agriculture land use causes CO2 drops by 0.002 megatons.

Another important observation is regarding the contributions of different regions/counties. From EDA in the previous section, in the order of the regions/countries producing most CO2 cumulatively from 1751 to 2019, the regions/countries under investigation can be ranked as United States, Europe, China, Russian, India, and Africa. However, the dummy variable coefficients show that United States, Europe, Russian, and Africa contributions are positively related to CO2 emission, but China and India are negatively related to CO2 emission. Combining the results in EDA section, a possible explanation is the industrial revolution started relatively late in China and India. The development of China and India initiated from late 20th century. Therefore, even though China and India contribute significant amount of CO2 emission cumulatively, they negatively affect the CO2 emission trend.

5 Future Research

The following tasks are considered and further investigated in my future research:

1) Add geological features based on the location of countries. Accordingly, a regional visualization and analysis can be performed
2) Add temporal features based on time period using Boolean values for the features like energy consumption. In this way, the negative effects of largely missing data on the model accuracy can be reduced.
3) Standardize the features. It is important to standardize the features by removing the mean and scaling to unit variance. The L1 (Lasso) and L2 (Ridge) regularizers of linear models assume that all features are centered around 0 and have variance in the same order. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected. In this way, the importance of features on CO2 emission can be comparted.

6 Conclusions

This work investigated CO2 emissions by considering five main factors, including primary energy consumption, GDP, population, agriculture land use, and beef production. Three machine learning regression models are applied to the training and test data set in order to predict CO2 emission. The results show that multivariate regression is the best performance model for this data set. The coefficients of each feature show that CO2 emission is positively related to primary energy consumption, GDP, population, and beef production but agriculture land use has a negative effect on CO2 emission. The dummy variable coefficients show that United States, Europe, Russian, and Africa contributions are positively related to CO2 emission, but China and India are negatively related to CO2 emission. The future work will include geological and temporal investigations and feature standardization.