Kiejstut Bunikiewicz, Koa Mahuna, Jerry Gehrung

Group 1

Professor Brahma

BSAN 6070: Intro to Machine Learning

05/05/21

Energy Data Analysis

**Project Objective and Domain Understanding**

For our project, we decided to focus on the topic of energy and concluded on the following predictive question: Can we predict which countries will produce the highest carbon emissions per capita? After initially pursuing a different question involving energy consumption’s effect on temperature change, we arrived at this inquiry because not only is it a topic of interest, but it would also provide useful answers and recommendations. Global carbon dioxide emissions have been rising for decades, totaling over 5 billion metric tons in 2016. Such a level is harmful because as carbon emissions invade our clean air, they form an invisible layer around the earth. This layer traps the heat inside the earth, leading to global warming. This process is what we call the Greenhouse effect. With melting polar caps and rising sea levels, it is time to make a change and reduce CO2 emission levels. With this predictive question, we then established the objective of our project to be determining: which countries produce the highest carbon emission, which variables have the greatest factor on carbon emission, and how to reduce carbon emission.

**Literature Survey**

To conduct our analysis, we drew vision from three literature surveys. *Random forest assessment of correlation between environmental factors and genetic differentiation of populations* showcases how random forest can be used to study the relationship between environmental factors such as nitrate and phosphate averages in water, geographic data, water salinity, and velocity averages in geographic areas, using the biological diversity of various marine mussels in the Baltic Sea as the independent variable. We found this survey helpful as it demonstrates how we can study environmental factors’ impact on another variable. It also exemplifies random forest as a viable model, beyond the classroom, for drawing conclusions from collected information. Our second literature survey is *The impact of long-term changes in air temperature on renewable energy in Poland*, which discusses how the renewable energy industry impacts climate change over time, substantiating the crisis that is our chosen topic. *Climate change impacts on renewable energy generation*, our final survey, addresses the most relevant studies that project quantitative estimates from climate change impacts on solar, wind, hydro, and other renewable generation technologies and calls for the development of public policies and private investment strategies. The suggested implementations in this work are valuable for us when brainstorming what recommendations to make.

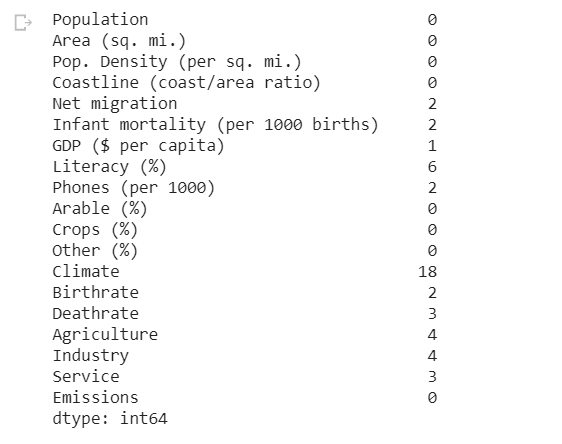
**Data Cleaning and Preparation**

In our project we analyzed two datasets, the first one *Countries of the World*, which we found on Kaggle. It is basically a world fact sheet, with information on population, region, area size, infant mortality, and such. This dataset consists of 227 rows, for the world’s 227 countries, and 20 columns for each country characteristic. Our second dataset is *CO2 Emissions per Capita*, which shows the CO2 emissions and CO2 emissions per capita in 2016. We retrieved this information from the Emission Database for Global Atmospheric Research (EDGAR) and the International Energy Agency (IEA). We read the *Countries* dataset into our Python notebook before concatenating it with *Emissions* via an inner join to take the intersection of the two, and this is the primary dataset that we will be using.

**Descriptive Statistics/Data Quality Analysis**

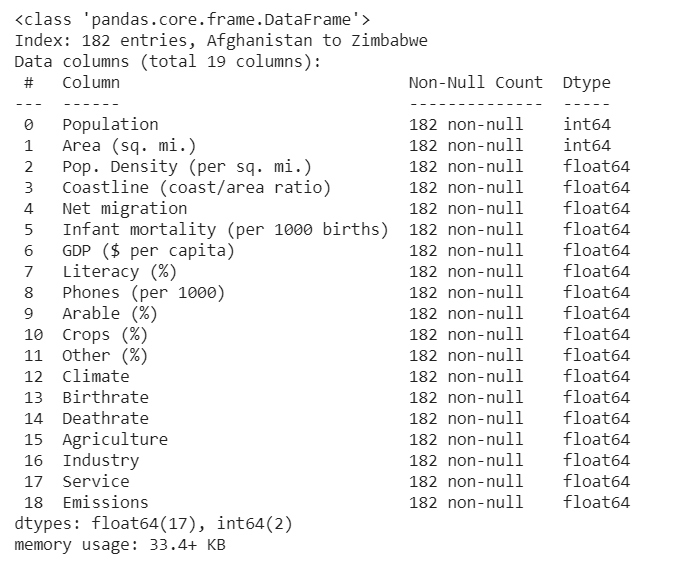
Our first step in our data quality analysis was to find the number of null values present in our dataset. None of the features in the dataset contained more than 10 percent of null values. This was considered an acceptable amount for our data to be considered high quality from the perspective of null values. Importantly, our dependent variable contained no nulls. We filled in the null values by imputing the median value for each column into the nulls.

Initial Count of Null Values:



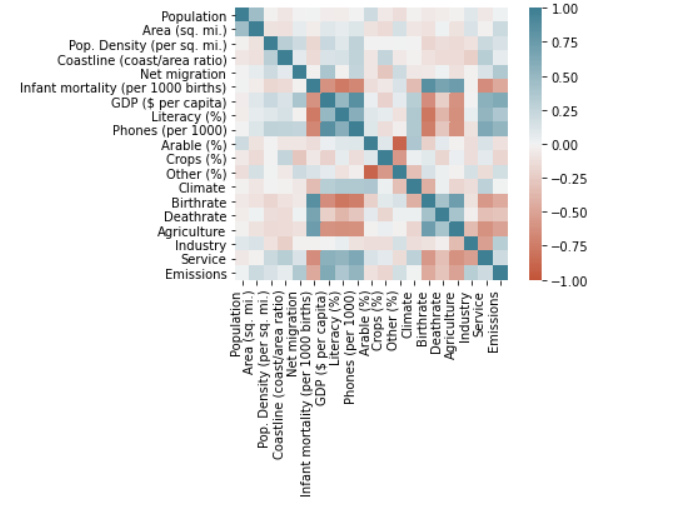
After imputing the null values, we inspected the data types and counts of values in the columns. All of the values were appropriately of the integer or float data type. The columns also contained the same number of values, indicating that there were no remaining null entries.

Data Frame Info:

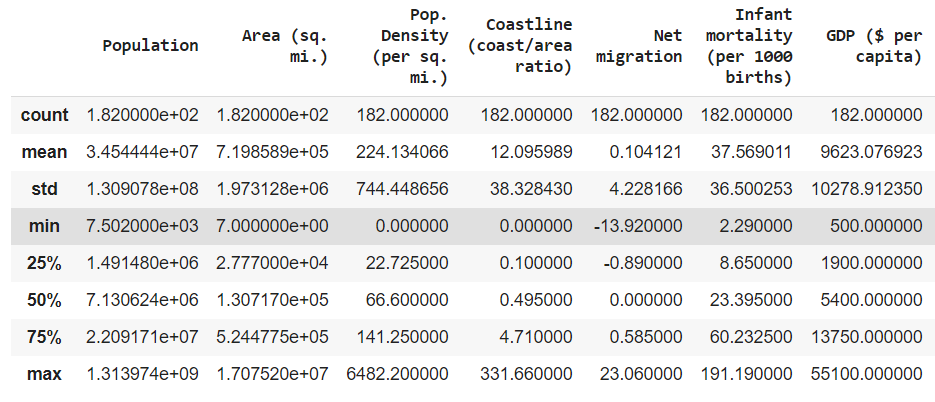


Upon completing our data cleaning and processing, we created exploratory visualizations and descriptive statistics to understand the data. The first data visualization we used was a correlation plot. Although the correlation plot shows that some of the independent variables are correlated with each other, none of them are above a threshold that we would consider removing them.

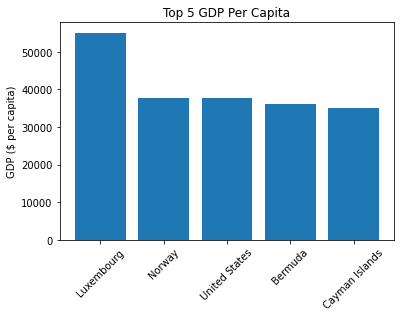
Correlation Plot:

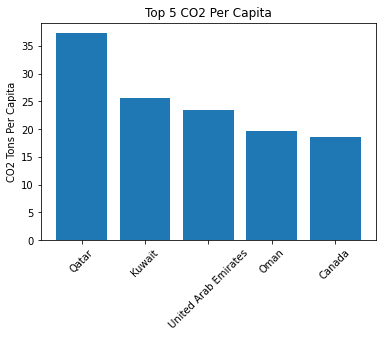


Descriptive Stats for First 5 Columns:



The descriptive statistics showed that the data was acceptable for continued analysis because there were few outliers in each feature. We then moved on to creating some exploratory visualizations to see the distributions of countries for certain variables and which countries were the top 5 CO2 emitters per capita and top 5 GDP per capita.





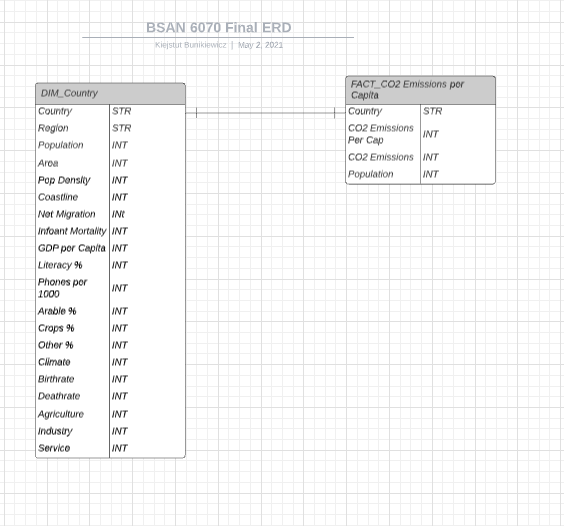
Interestingly, we found that the top CO2 emitters per capita were not necessarily those with the highest GDP per capita. This was surprising because one of our initial assumptions was that these two groups would have significant overlap. However, there is no overlap among the top 5 countries for either of these two metrics.

**Feature Selection Process**

Our initial data consisted of two tables. The first table was a Dim\_Country table that held descriptive information about the countries being used for analysis. The initial features included Country and Region name and the country’s Population, Area in Square Miles, Population Density, Coastline (coast/area ratio), and Climate category. Other features included demographic information such as Net Migration, Infant Mortality per 1000 Births, Birth Rate, and Death Rate. Finally, this table included a country’s economic and land use information such as GDP ($ per Capita), Phones per 1000 People, Arable Land %, Crop Land %, Other Land %, and the ratios of economy in the Agricultural, Industrial, and Service sectors.

The second table we used was a Fact\_CO2 Emissions per Capita country which contained information about the CO2 emission rates for the citizens of the countries. The columns in this table included the country’s name, CO2 emissions in tons per capita, CO2 emissions in tons in 2016, and population.

ERD Depicting Initial Features:



The feature we eliminated from the Country information table was the Region column. This feature was unnecessary because the analysis we are running is at a country level, not at a regional level. From the Emissions table, we eliminated the CO2 Emissions totals and the Population information. The Population column was redundant because our country information table already had a column on country population. The overall CO2 emissions was also removed because the question we wanted to answer related to the per-capita emissions. Therefore, this feature was unnecessary.

**Performance Criteria**

Due to the use of both Regressions and Classifiers for our initial models, we used two sets of criteria for model performance. The criteria we used for regression based models was R^2. This criteria was effective for comparing models because the R^2 estimates the model’s performance based on the amount of variation in the dependent variable that can be predicted by the use of the model. The R^2 is also a useful criteria because it establishes a useful baseline for effectiveness of 0.5. If a model is below this score, it explains less than half of the variation between predictions and therefore does not contain much useful information.

The second set of criteria we used was the set of metrics most used for classification models: precision, recall, F1 score, and accuracy. Accuracy is a good criteria for baseline estimation because it calculates the number of correct decisions made over the number of total decisions made. This would be useful because it would give the group members a general idea of their model’s performance. However, the accuracy metric may be too simplistic when evaluating a model’s performance in depth. In this case, we chose to consider the precision, which calculated the true positives over the sum of true positives and false positives to give an idea of the model’s positive predictive value. We also used recall to estimate what percentage of the target class in the data was correctly identified. This metric would be useful for our group because it is important to predict and strategize for countries that have a positive value (CO2 per capita emissions above average) rather than to penalize countries that are doing their part to mitigate emissions per capita. The final and most important metric we wanted to use was the F1 Score. This score provides a harmonic mean of precision and recall, which indicates how well our model balances the precision and recall. The F1 score will be useful because it will allow us to evaluate our model’s ability to not only classify positives (or when the citizens of a country emit more emissions than usual), but also to minimize the number of misclassifications of negative outcomes. In short, the F1 score will allow us to see which countries emit more CO2 per capita than average while minimizing the number of countries that will be falsely classified in that group.

**Possible Model Choices and Justification**

Classification and Regression Tree:

The first model we decided to run was the Classification and Regression Tree (CART). Our reasoning is that this model works well for categorical as well as continuous target variables. CART requires neither normalization nor the scaling of data. Furthermore, missing values in the data do not affect the process of building a decision tree to a considerable extent. Although this does not matter greatly since we replaced all our nulls with median values, we believe it makes CART a strong model impervious to underlying flaws in the data. On the other hand, the drawback of using CART is that there is a high probability of overfitting the data. Additionally, it gives low prediction accuracy for a dataset compared to other machine learning algorithms and calculations can become complex when there are many class labels. As far as assumptions, typically we do not need to make any at all. However, this lack of assumptions is a doubtful advantage because, while those assumptions are not guaranteed, it gives this method a relative advantage. Conversely, when those assumptions hold, more information can be derived from the data. Overall, we believed a CART analysis would be extensive and definitive. Classification and regression trees are simple to comprehend as they follow a similar approach to humans when making decisions.

Gradient Boosted Techniques:

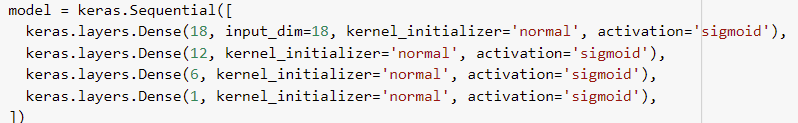
Two other models that could be useful for answering our research question are the Gradient Boosted Regression and the Gradient Boosted Classifier. We thought that these two techniques could be useful for our question because they come with very few statistical assumptions. The Gradient Boosted models use decision trees as a basis, but combine the trees with gradient descent, which allows the model to minimize its errors as it iterates through the predictions. The models then decide a predictive value or make a classification by using an aggregate decision, like an average or voting. The Gradient Boosted models are also robust because they do not require data normalization and can accept both categorical and continuous variables. The Gradient Boosted Regressor requires a continuous target value and the Classifier requires a categorical target value. This categorical target value was computed as 1 for above global average per capita CO2 emissions and 0 for below average. The Gradient Boosted models come with the benefit of being a strong predictor because they have the ability to iterate through multiple tree predictions to reduce error. The Gradient Boosted techniques’ process of arriving at a decision by aggregate also means that this model is not prone to overfitting. This occurs because each model iteration will fit differently and use an aggregate to arrive at a prediction. The Gradient Boosted techniques do come with two limitations. These models are time and resource intensive and are a black box. However, these drawbacks are not too concerning because the dataset is small enough to avoid concerns about computational time and resources. The models will also allow us to see feature importances, which is an additional benefit when answering our predictive question. The feature importances will allow us to identify which factors have the highest effect on the CO2 per capita emissions and inform us on which policy recommendations we can make to reduce CO2 emissions per capita.

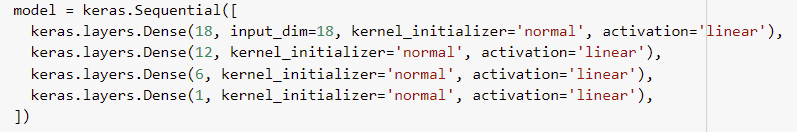
Neural Net:

For our final models we decided to try neural networks. We attempted to use a neural network regressor and a neural network classifier. We felt that a neural network might be best to identify the complex relationships between our features and our target column. Secondly, the back propagation methods appealed to us. We felt that our small dataset might be best leveraged by a model with rapid self-learning built in. Furthermore, we were able to add layers to add complexity to our model, so it seemed like a great model for us to utilize. Its adaptability was a primary selection reason.

The neural network operates by tuning weights over each trial according to the error. This means that, over time, a neural network would be able to identify the optimal weights and biases in order to accurately predict based on input nodes. We were optimistic that this model would be able to make these tunings quick enough to find optimal weights given our small dataset size. Additionally, for our neural nets, no assumptions were made about our data.

Architecture for Neural Networks:





**Final Model: Gradient BoostingClassifier**

We decided that the most optimal model was a Gradient Boosted Classifier. From our three individual models, we felt that it was the most appropriate. Its classification metrics were promising even prior to intense hyperparameter tuning. On top of its initial accuracy, we found it more intuitive than our other models. For example, we were able to produce feature importance plots which indicated far more information about our outcomes than neural networks. CART performed well but we preferred the Gradient Boosted model since we felt that its ability to account for strong and weak predictors would benefit our predictive power. The GBM could update weights as it iterated through trees which would create a more accurate model. We decided to use it as a classifier because it gave us better performance. Furthermore, although we sacrifice the specificity of a regressor’s prediction, we still have the feature importance plot. The feature importance plot can inform policy decisions, so we still have the ability to make meaningful suggestions without using a regression model. However, how can we be sure that our model is accurately performing prior to interpreting its results? In the next section, we will be discussing the performance criteria.

Hyperparameter Tuning:

To improve our model from its initial run session, we performed hyperparameter tuning. In order to quickly and efficiently tune these parameters, we used the GridSearch package. We constructed a parameter grid which included a handful of hyperparameters and a list of values for the package to try. In our initial attempts, we included far more values and hyperparameters but discovered that GridSearch was going to take over 10 hours to finish running. We felt that we should reduce the number of instances it had to run so that we could finish the tuning in a more reasonable amount of time. We eventually ended up with around 7,000 different combinations which took around 20 minutes to run. This resulted in acceptable scores for accuracy which we will detail in the next section. However, since we did not complete an exhaustive GridSearch, it is possible that we did not use the most optimal parameters.

Performance Criteria:

We decided to gauge the performance of our model by utilizing the standard classification metrics: precision, recall, F1 score, and accuracy. Obviously, while accuracy may be a good benchmark score to assess performance, it is important to ensure that our model is performing well with the classification of true positives, false positives, etc. This is where precision and recall are essential. Firstly, we can see that our accuracy is 95%. This is a good score and indicates that our model performed very well at identifying which countries would be producing the highest carbon emissions. However, could a class imbalance be impacting our results? We look at the F1 scores and see values of 0.96 and 0.91 for false and true respectively. These indicate that our model has done a good job at both identifying true instances and not misidentifying the class of a country. With these metrics, we can be more confident that our results have been produced by an accurate, reliable model.

Insights from Model:

Now that we have trained our model and confirmed that its results are worth interpreting, we can begin to derive insights from it. Our primary source of information for these insights will stem from the feature importance plot. It is immediately clear that GDP per capita is the highest predictor of carbon emissions per capita. Intuitively, we can assume that richer countries have citizens with more vehicles, devices, appliances, etc. All of these additional devices generate more carbon when they are used and also cost a lot more carbon to create. This hypothesis is supported by the fact that the number of phones per 1000 people is the second-highest predictor. We can assume that phones are indicative of how many other devices or appliances an individual family might have. After GDP and phones, our predictors become more cryptic. For example the third- and fourth-highest predictors are infant mortality and net migration. Infant mortality rate is typically inversely correlated with GDP, which raises some questions about these results. Perhaps this relationship is a fluke and not as meaningful, but this would require additional research. Net migration is a ratio that becomes positive as more people immigrate into a country. This means that countries with a high value for net migration have a lot of people moving in and not a lot moving out. This is indicative of a country where quality of life might be higher, which may nod to a higher GDP and more amenities. These may be responsible for the higher emissions per capita. We can see that a number of variables have relatively insignificant impacts on emissions, such as: economic industry proportions, literacy, population, and especially geographic features.

This information allows us to identify what might need to happen in order to lessen our carbon emissions over time. As the rest of the world continues to modernize, more people around the world will have access to appliances and vehicles that first world countries are using today. This means that there will be a dramatic increase in the number of emissions globally. In order to avoid this catastrophic amount of emissions in the future, we alter what these items use for power. Renewable energy cars and energy-efficient appliances are the key to maintaining a more viable level of emissions. If other countries modernize and we continue to use the same energy sources that we have today, then we will see an unsustainable rise in emissions. Thanks to our model it has become clear that the technology utilized by richer countries is a powerful indicator of their carbon emissions. In order to curb these rises, we would need to reduce the use of these appliances or change how they are powered. Obviously, it is unrealistic to expect people to stop using modern appliances or vehicles, so the best option is to target renewable energy sources which will steadily decrease the emission per device as other countries begin adopting them. There may be asubstail emission during the initial development and production of these renewable technologies, but hopefully there is a net negative result over the course of their lifetimes.

Model Limitations:

Unfortunately, our model is not perfect. We do have some issues that could be addressed for future iterations. Firstly, the primary issue is a lack of data. We only had around 200 rows total which led to our training set being around 130 to 140 records. That means our model was only trained on a very small handful of data. We cannot be confident that these results are entirely generalizable as a result. A similar issue is the fact that our data itself embodies a number of assumptions. Since our records each represent a country, that means we are generalizing the data of each country to its inhabitants. For instance, we are assuming that people from New York and Kentucky operate in a similar function or enjoy a similar quality of life when the truth may be different. Another limitation may be that our model was not optimally tuned. As discussed earlier, we were unable to perform a comprehensive model tuning search due to time constraints. As a result, we are uncertain if this set of hyperparameters is the best performing set.

Future Improvement:

To address the issues, we have a couple suggestions. Firstly, we believe that if data could be collected at a more granular level, then we would improve the model’s predictive capability – for example, if we could have information on states within countries or territories within countries. This would allow us to have data separated into smaller and more individualized segments. At this point, our model would predict based on individual characteristics of regions as opposed to completing analysis on an average of an area. Furthermore, it would create far more records for our model to train itself on. This approach would drastically improve the model’s reliability. Secondly, we suggest performing a comprehensive GridSearch over the course of a day or two to determine the truly optimal hyperparameters. This, in conjunction with something like a K-Fold Cross-Validation method, would ensure that our results are meaningful and accurate.