Aira Domingo

Data Mining Individual Report

Note: I worked alone so the following is the same as the Final Group Project Report.

I worked on all the codes in datamining.py and app.py.

Percentage of code found on internet 90/471 = 0.19 = 19%

Ecological Footprint

Introduction

We are currently facing an environmental crisis. The global temperature is rising, the ocean is getting warmer, ice sheets are shrinking, sea levels are rising and so many other extreme events. Many scientists agree that the main reason for current global warming is the human contribution to the greenhouse effect. The greenhouse effect is the warming that results from gases radiating from Earth being trapped in the atmosphere. These gases include water vapor, methane, nitrous oxide, and carbon dioxide. Since the Industrial Revolution began, humans have increased the atmospheric carbon dioxide concentration by a third and it has become the main contributing factor to climate change ("NASA").

In the United States, the primary sources of greenhouse gases are transportation, electricity production, industry, commercial and residential, agriculture, and land use forestry. Land use has especially become a major factor in global warming since there is now less land for plants that absorb the carbon dioxide from the atmosphere ("EPA"). It's important to understand the big players in the rising levels of carbon dioxide and the Earth's difficulty in absorbing carbon dioxide, which is leading us to global warming.

This is because we are now at a point where we are using more of the Earth's resources than it can provide for us. The Earth is in an ecological deficit. Currently, humanity is using the equivalent of 1.7 Earths to provide our resources and absorb our waste. We know this by measuring a given population's ecological footprint and comparing it to their biocapacity.

Ecological footprint is a metric that measures the natural resources that a given population consumes, including, plant-based, livestock, fish, timber and other forest products, space for urban infrastructure, and resources to absorb its waste, especially carbon emissions. Biocapacity, on the other hand, represents the ecological productivity of crop land, grazing land, forest land, fishing grounds and built-up land for urban infrastructure. If most of these areas are lest unharvested, they can also absorb much of the waste that a population generates, especially carbon emissions.

Ecological footprint and biocapacity are both measured in global hectares which is standardized with world average productivity to be globally comparable. If a given population's ecological footprint is greater than their biocapacity, this population is in an ecological deficit. If a given population's ecological footprint is less than their biocapacity, this population is in an ecological reserve ("Global Footprint Network").

Since it is important to understand what exactly is contributing to the rise in carbon emissions, this project will focus on answering the question: How is the amount of carbon emission affected by a country's land use?

Description of Data Set

I found a Global Footprint Network dataset called "Ecological Footprint Accounts 2018" on Kaggle. The data set contains 1,305,300 observations. There is a total of 15 features: country, ISO alpha-3 code, UN region, UN subregion, year, record, crop land, grazing land, forest land, fishing ground, built up land, carbon, total, Percapita GDP (2010 USD), and population. The ISO alpha-3 code is a 3-letter code that identifies each country. UN region is the continent the country is located, while the UN subregion is the part of the continent the country is in. The column year ranges from 1961-2014, although, not all countries have data from all these years.

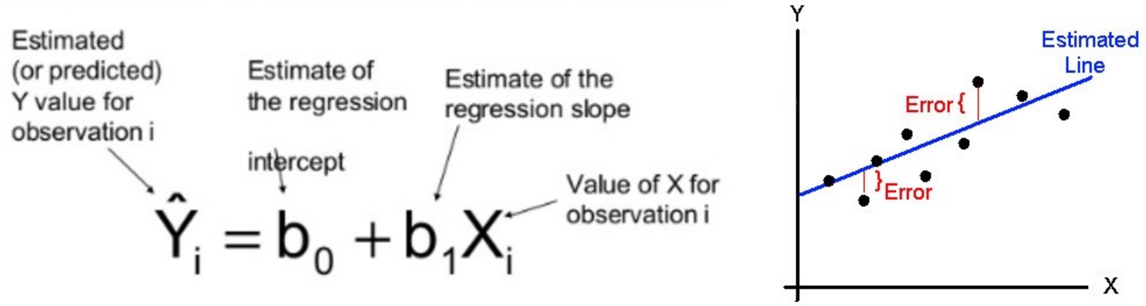
There are 10 types of ‘record’: BiocapPerCap, BiocapTotGHA, EFConsPerCap, EFConsTotGHA, EFExportsPerCap, EFExportsTotGHA, EFImportsPerCap, EFImportsTotGHA, EFProdPerCap, EFProdTotGHA, and each record has 130,530 observations which totals to the 1,305,300 observations, altogether. ‘PerCap’ after each record indicates the values measured per capita while ‘TotGHA’ indicates the values measured in global hectares. ‘Biocap’ is the country’s biocapacity for that year, ‘EFCons’ is the total Ecological Footprint Consumption for the year, ‘EFExports’ is the country’s amount of resources exported from the country that year, ‘EFImports’ is the country’s amount of resources imported to the country that year, and lastly, ‘EFProd’ is the consumption of resources resulting from production processes. Ecological Footprint Consumption is the sum of Ecological Footprint Production (EFProd) and resources used within international trade (EFImports – EFExports).

$$\text{EFCons} = \text{EFProd} + (\text{EFImports} - \text{EFExports})$$

Crop land is the global hectares of land used for crops and crop derived products available or demanded. Grazing land is the global hectares of land used for meat, dairy, leather, etc. It includes global hectares for the animal’s grazing but not the crop land used to produce feed for the animals. Forest land is the global hectares of land available or demanded for sequestration, timber, pulp or timber products. Fishing ground is the global hectares of marine and inland fishing grounds used for fish or fish products that are available or demanded. Built-up land is the global hectares of built up land or land used for human infrastructure that is available or demanded. Carbon is the global hectares of world-average forest required to sequester carbon emissions. The total column is the sum of all the land types for each country in that year and record. Percapita GDP is the per capita GDP in constant 2010 USD. Lastly, population is the country's population of that year rounded to thousands ("Global Footprint Network").

Data Mining Technique

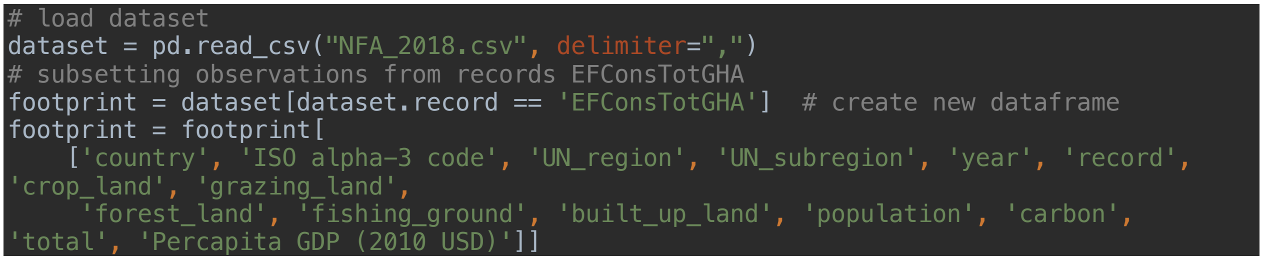
Linear regression is a common and simple statistical data analysis technique used to understand the real-world phenomena of the relationship between two or more variables. This allows us to gain information on one variable (y) by knowing the values of the others (x). Regression means fitting a straight line to data and can be used for prediction and modeling causal relationships between a dependent (y) and independent (x) variables. Linear regression is only valid for a data set with a continuous dependent/response/target variable. It is important to minimize the error, which is the difference between the observations and the predictions made by the linear model ("Statistically Significant Consulting").


Estimated (or predicted) Y value for observation i
Estimate of the regression intercept
Estimate of the regression slope
Value of X for observation i
$\hat{Y}_i = b_0 + b_1 X_i$
Y
Estimated Line
Error {
}Error
X

In this case, carbon will be the dependent or response variable and crop land, grazing land, forest land, fishing ground, and built-up land will be the predictors.

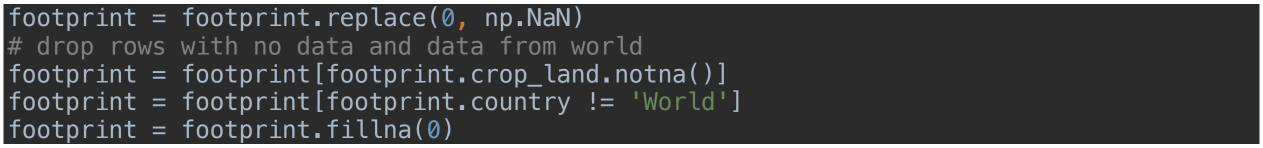
Experimental Setup

Using Pycharm, I imported the packages numpy, pandas, and sklearn. Since the record 'EFCons' is the total of the other records and I am interested in predicting 'carbon' using 'crop land', 'grazing land', 'forest land', 'fishing ground', and 'built up land' which are all measured in global hectares, I will only be using observations with the record 'EFConsTotGHA'.



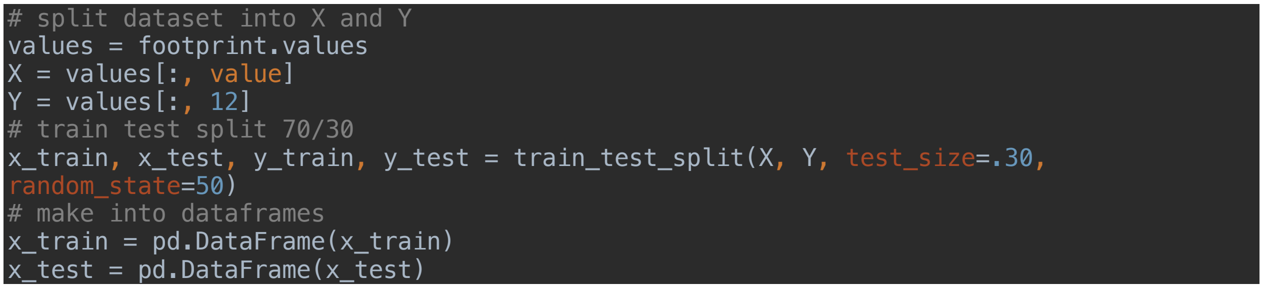
```
# load dataset
dataset = pd.read_csv("NFA_2018.csv", delimiter=",")
# subsetting observations from records EFConsTotGHA
footprint = dataset[dataset.record == 'EFConsTotGHA']  # create new dataframe
footprint = footprint[
    ['country', 'ISO alpha-3 code', 'UN_region', 'UN_subregion', 'year', 'record',
'crop_land', 'grazing_land',
     'forest_land', 'fishing_ground', 'built_up_land', 'population', 'carbon',
'total', 'Percapita GDP (2010 USD)']]
```

After looking at the data set, I found that many of the countries didn't have any data for any of the land use columns before 2014 so the observations where the crop land column was Nan were dropped. World was also included as a 'country' to represent the World's total. The values for this would have been far greater than any of the countries so the observations for world were also dropped. For imputation later, all Nan were replaced with 0. The remaining number of observations is 94,470.



```
footprint = footprint.replace(0, np.NaN)
# drop rows with no data and data from world
footprint = footprint[footprint.crop_land.notna()]
footprint = footprint[footprint.country != 'World']
footprint = footprint.fillna(0)
```

To prepare the data to fit the model, the dependent (Y) and independent (X) variables were assigned. The data was then 70/30 split into the training and testing data, meaning 70% of the data is for training and 30% for testing the model. The split in to X and Y created numpy arrays and need to be converted back to dataframes.



```
# split dataset into X and Y
values = footprint.values
X = values[:, value]
Y = values[:, 12]
# train test split 70/30
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=.30,
random_state=50)
# make into dataframes
x_train = pd.DataFrame(x_train)
x_test = pd.DataFrame(x_test)
```

```
y_train = pd.DataFrame(y_train)
y_test = pd.DataFrame(y_test)
```

Since there are still empty cells for some observations, I had to impute the missing data with the mean for the column.

```
# impute missing value with mean
imputer = SimpleImputer(missing_values=0, strategy='mean')
# impute training set
imp_x_train = imputer.fit_transform(x_train)
imp_x_test = imputer.transform(x_test)
imp_y_train = imputer.fit_transform(y_train)
imp_y_test = imputer.transform(y_test)
```

Next, I scaled the data using scaler = MinMaxScaler() with a maximum of 100 since the features have different ranges. In scaling, to ensure that the test data remains "unseen", I scaled the training set with scaler.fit_transform() and the test set with scaler.transform() only ("Sebastian Raschka"). Figures are in the appendix showing total ecological footprint and carbon emissions before and after data cleaning.

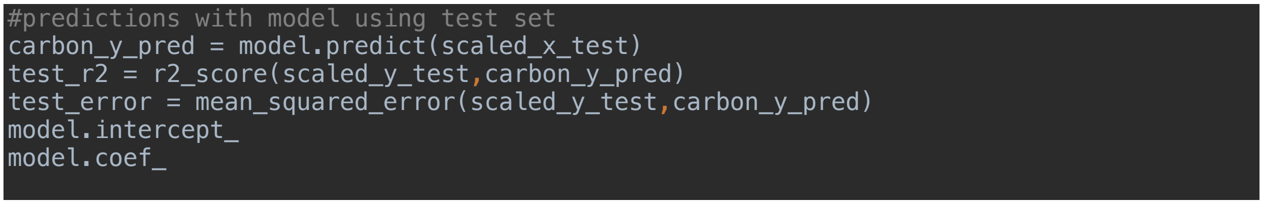
```
# scale data
scaler = MinMaxScaler()
scaled_x_train = scaler.fit_transform(imp_x_train)
scaled_x_test = scaler.transform(imp_x_test)
scaled_y_train = scaler.fit_transform(imp_y_train)
scaled_y_test = scaler.transform(imp_y_test)
```

Then I fit the data to the model using scaled x and y training sets and predicted carbon only using scaled x testing set.

```
#train with linear regression model using training sets
model = LinearRegression(normalize=True)
model.fit(scaled_x_train, scaled_y_train)
```

In order to judge the performance of the model, I am using r-squared (R2) and mean squared error (MSE). R-squared measures how close the data fits the regression line built by the model. R-squared = Explained variation/Total Variation. This means that 0 (0%) indicates that the model explains none of the variability of the response data around its mean. An r-squared of

1 (100%) means that the model explains all the variability of the response data around its mean. Overall, the higher r-squared, the better the model fits the data ("Minitab") measures the average of the squares of the errors. Errors are the differences between the predicted y values and target y values. The objective is to minimize the mean squared error and have an r-squared value close to 1. The model's coefficient and intercept were also calculated.



```
#predictions with model using test set
carbon_y_pred = model.predict(scaled_x_test)
test_r2 = r2_score(scaled_y_test,carbon_y_pred)
test_error = mean_squared_error(scaled_y_test,carbon_y_pred)
model.intercept_
model.coef_
```

Results

In total, 5 models were created with crop land, grazing land, forest land, fishing grounds, and built up land acting as predictors of carbon in their own models. The crop land model had an r-squared of 0.59 and MSE of 11.722 and resulted in a linear regression of y= 0.71x + (-0.002) + error. The grazing land model had an r-squared of 0.469 and MSE of 15.308 and resulted in a linear regression of y= 0.375x + (-0.037) + error. The forest land model had an r-squared of 0.81 and MSE of 5.48 and resulted in a linear regression of y=0.578x + (-0.44) + error. The fishing ground model had an r-squared of 0.438 and MSE of 16.191 and resulted in a linear regression y = 0.58x + (-0.214) + error. The built-up land model had an r-squared of 0.354 and MSE of 18.620 and resulted in a linear regression of y = 0.7786x +0.45 + error.

Linear Regression of Carbon Emissions vs Forest Land Use

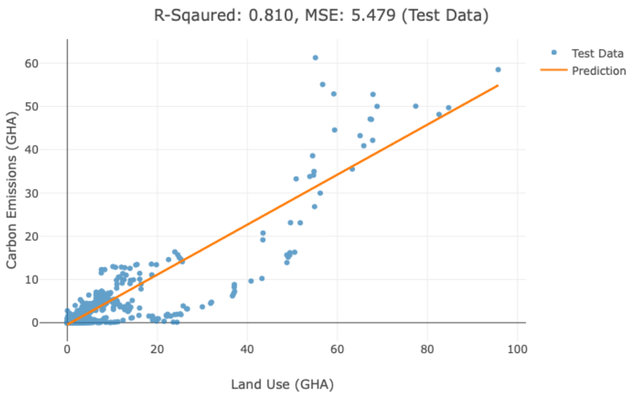


Figure 1 r2=0.8099, MSE = 0.0000548, y = 0.578x+(-0.44)

The figure above illustrates the plot of the test data for forest land and the prediction. The forest land model is the best model as a predictor for carbon. It has the highest r-squared value which we wanted to be as close to 1 as possible. It also has the least mean squared error, which we wanted as close to 0 as possible. This meant that the test data had the least amount of variability from the predicted model. It can be observed that the majority of the data points are near (0,0). There is an unequal distribution of the data points. There is a positive, linear relationship. As forest land increases, carbon emission also increases.

Linear Regression of Carbon Emissions vs Built Up Land

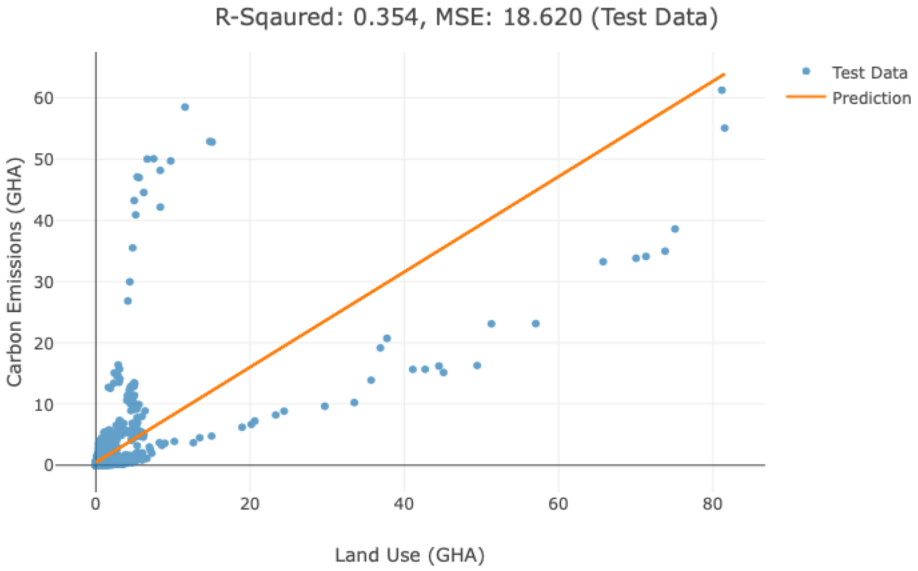


Figure 2 r2=0.354, MSE = 0.00186, y = 0.7786x+0.45

It is surprising that built up land was the least reliable model as a predictor for carbon. Although not as significant as a model, it had the highest coefficient resulting in the steepest slope or fastest rate of change between forest land use and carbon emissions. With a low r-squared, the test data had the greatest variability to the predicted model.

The second-best model was crop land, followed by grazing land and fishing grounds, in descending order. The performance of the crop land model was about 10% better than the other two. Grazing land and fishing grounds had about the same performance.

Conclusion

Forest land use turned out to be the best predictor for carbon emissions. This makes sense as the increasing forest land use means that there is less land available for plants to absorb the carbon dioxide in the atmosphere, leading to an increase in carbon dioxide concentration. It is noteworthy that built up land wasn't a great model for predicting carbon emissions due to the variability. I'm interested in looking at other datasets to see if this finding is specific to this particular dataset used in the report. In the future, I would also like to explore the rate of carbon emission increase through the years and pin together events in history that has caused the great increase in recent years.

References

(“EPA”)https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions

(“Global Footprint Network”) https://www.footprintnetwork.org

(“NASA”) https://climate.nasa.gov/causes/

(“Minitab”) https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

(“Sebastian Raschka”)https://sebastianraschka.com/faq/docs/scale-training-test.html

(“Statistically significant Consulting”)

https://www.statisticallysignificantconsulting.com/RegressionAnalysis.htm

Appendix

Figure 3 Ecological Footprint over the Years (Raw)

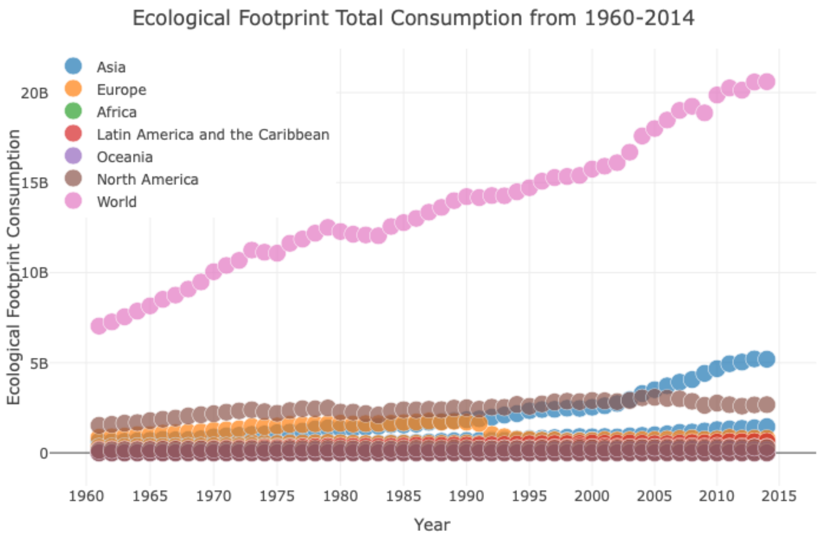


Figure 4 Ecological Footprint over the Years (Clean)

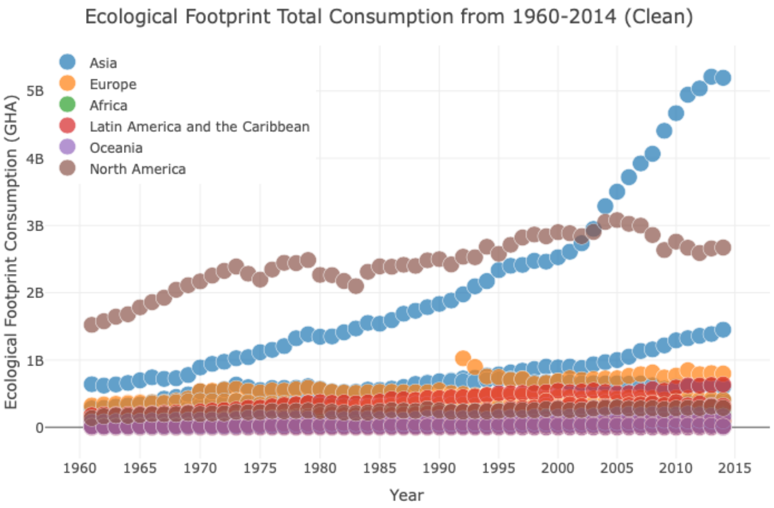
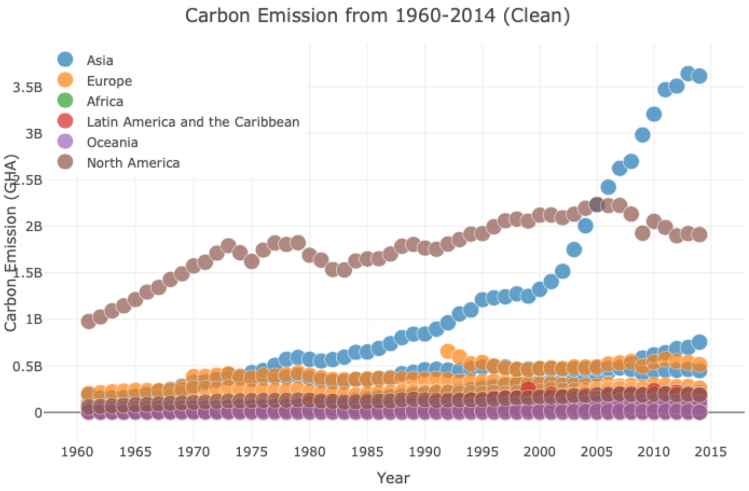
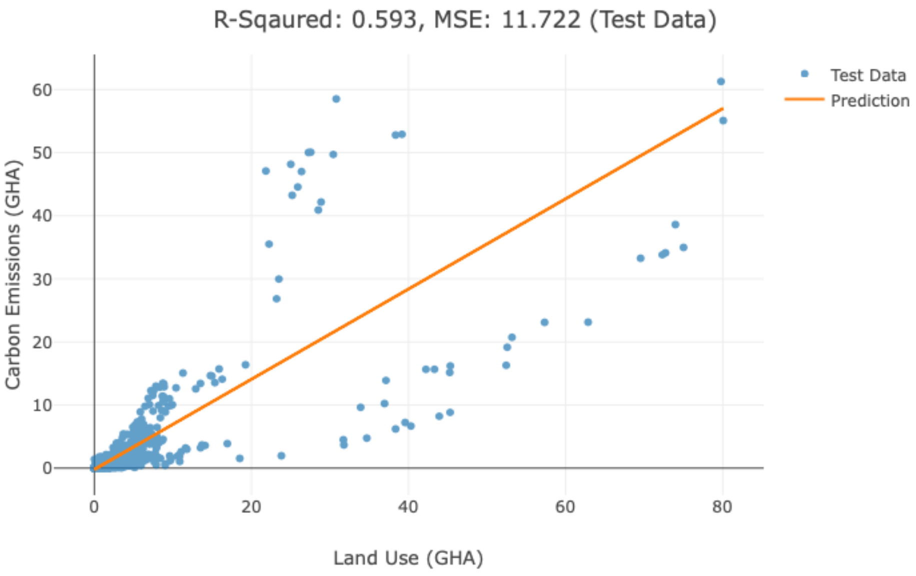


Figure 5 Carbon Emissions over the Years (Clean)



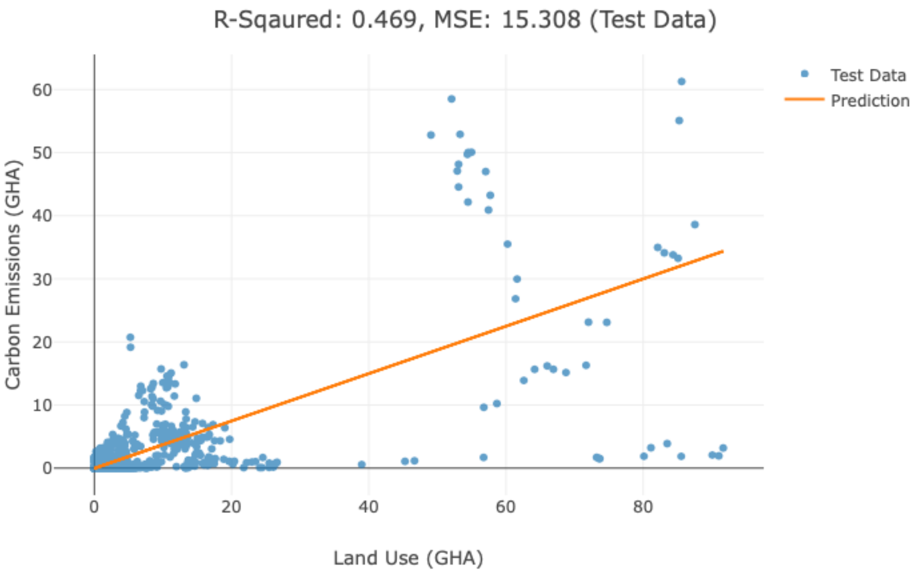
Linear Regression of Carbon Emissions vs Crop Land

Figure 6



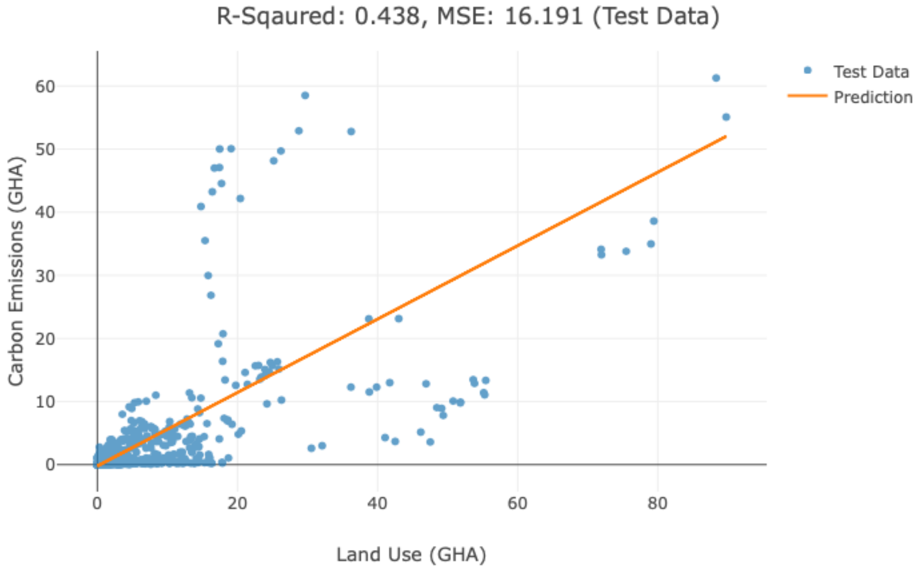
Linear Regression of Carbon Emissions vs Grazing Land

Figure 7



Linear Regression of Carbon Emissions vs Fishing Grounds

Figure 8


R-Sqaured: 0.438, MSE: 16.191 (Test Data)
Test Data
Prediction
Carbon Emissions (GHA)
Land Use (GHA)