Amazonian Deforestation Analysis

Gehan Velivitiya, Israel Booth, Tori Wright, Sri Harshini Duggirala

1 Project Statement and Motivation

The work presented in this paper aims to predict the area of deforestation in the Amazonian rainforest using different machine learning approaches. The Amazon rainforest – which stores around 76 billion tons of carbon and 20 billion tons of water – holds a vital role in regulating our climate and water cycle. As of 2023, about 20% of the Amazon rainforest has been deforested. Accurate deforestation forecasting allows for preventative measures on a governmental and corporate level. Our goal is to understand the trend of deforestation, create a model that can predict the area of deforestation using historical data, and analyze other factors that may contribute to deforestation.

2 Introduction and Description of Data

The historical data comes from TerraBrasilis, a web portal used to query spatial data from governmental environment organizations, detailing the amount of deforestation in kilometers for each of Brazil's states and municipalities from the year 2007 to 2022. There are 412 different municipalities belonging to 8 states in the dataset. For simplicity purposes in EDA, we started to examine the deforestation data for one state, using the average area of deforestation for all municipality per year. In our EDA and analysis, we used the state Acre. After the initial EDA we expanded the dataset by adding more features. Instead of the cumulative deforestation, we computed the deforestation increment as our target variable. This left us with observations from 2008 to 2022, since we had to leave out observations related to 2007. From Brazilian Institute of Geography and Statistics we collected gdp and gdp per capita information for the areas in our datasets from 2008 to 2020. We observed that the gdp and gdp per capita follows a monotone increasing trend. Taking that into account we imputed the values for 2021 and 2022 using a linear regression model trained for each municipality individually.

2.1 Exploratory Data Analysis

Our variable of interest is area in kilometers of deforestation. We began exploring the data by examining the distribution of our target variable. By plotting a histogram of the area variable (Figure 1), we can see that the area is skewed right. Area has a mean of 988.13 and a median of 1056.19, with a standard deviation of 656.94.

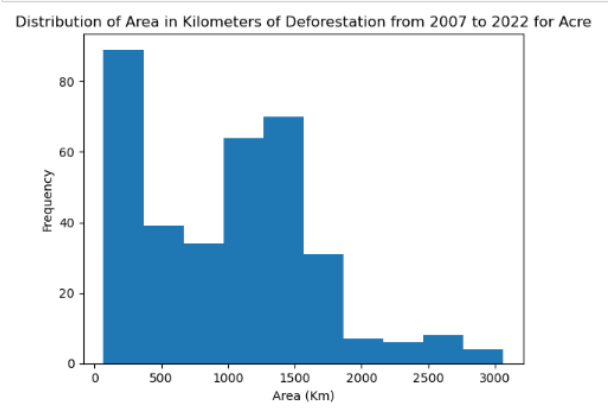

Figure 1

Since our data has a time component, we are interested in how the target variable behaves over time. To visualize this interaction, we plotted an area of deforestation against the variable "year".

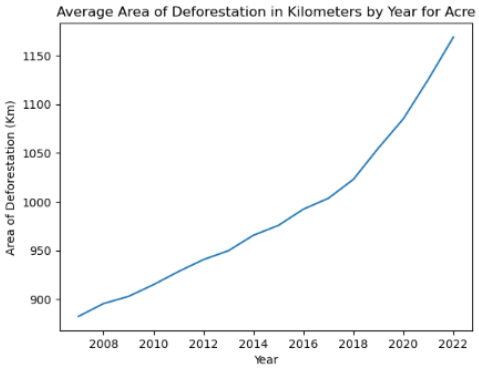

Figure 2

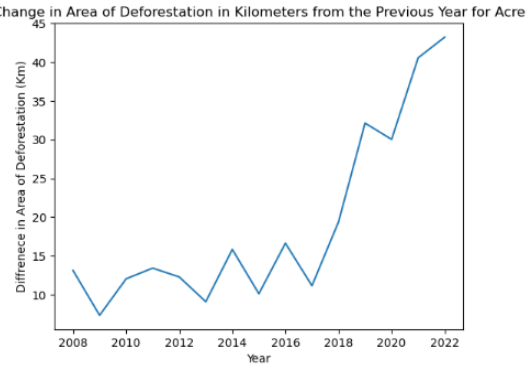

Figure 3

In figure 2, we can see that the area of deforestation is monotonically increasing over time. Interestingly, in recent years the area of deforestation has been increasing more rapidly than previous years. Around the year 2018, the area of deforestation noticeably accelerated. We further investigated this fact by computing the difference in area between the current year and the previous year.

From figure 3, there is an obvious spike in the difference in area of deforestation from the previous year. Between the years 2008 and 2017, the difference is somewhat stable between 10 and 15 kilometers. In 2018, the difference in area spikes and displays an upwards trend for the following years. In the year 2022, there was the greatest difference in deforestation with an area of about 45 kilometers.

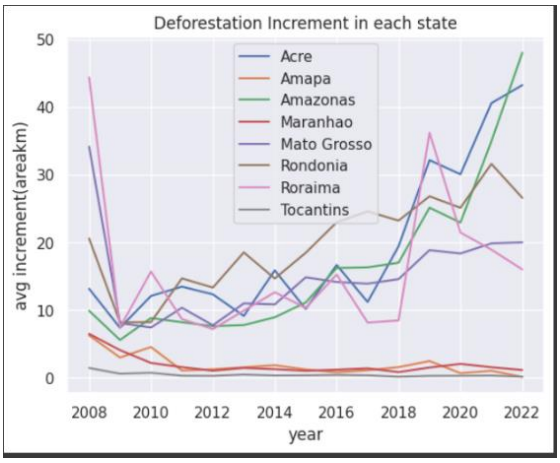

Figure 4

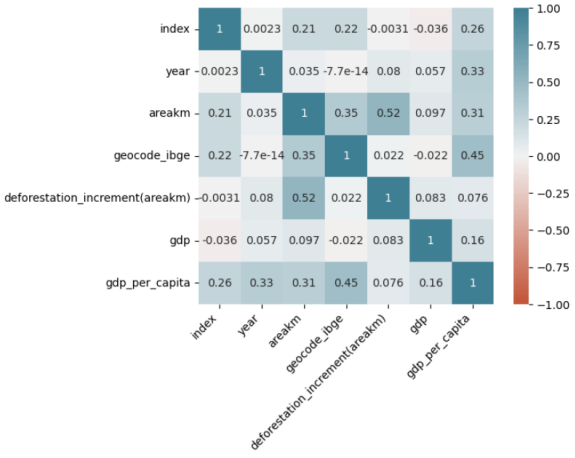

Figure 5

Figure 4 shows the trend of average deforestation increment over time for all 8 states. States Acre, Amazonas, Rondonia, Mato Grosso and Roraima follow a similar trend. Whereas the rest follows a significantly different trend, and the deforestation areas of those states are low. Observing the confusing matrix in Figure 5 we can see that most attributes are positively correlated to each other. The gdp has a very low correlation to the deforestation_increment(areakm), which was one of our goals to figure out from this project.

3 Modeling Approach

3.1 Time Series Forecasting

The time series forecasting task involves fitting models on historical data to predict future time steps. The accuracy of our predictions will entirely depend on how well our model fits the data. To accomplish this, we used the popular statistical model Autoregressive Integrated Moving Average (ARIMA). It is important to note that, though the target variable is skewed (Figure 1), we will not be scaling the data to fit a standard normal distribution as it made the model perform worse and ARIMA is not sensitive to scale. The forecasting process can be broken into two steps: build a model that fits the historical data then forecast future values. First, our data is plotted as a time series with the year along the x-axis and area of deforestation (km) along the y-axis. Since the data shows a clear upwards trend, we can determine that the series data is not stationary and requires some degree of differencing (Figure 2). By plotting an autocorrelation plot, we can see that there is a high degree of autocorrelation between consecutive observations. The data is not random and there exists some relationship between each data point (Figure 6).

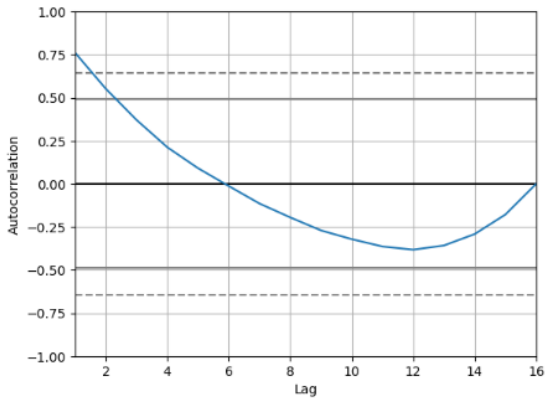

Figure 6

Identifying and understanding the trend in our data allows for faster modeling, making the model selection and evaluation process simpler. From Figure 2, we can identify a deterministic trend – meaning the trend is consistently increasing. To model this trend, we fit a linear regression model. By modeling the trend, we can remove the trend from the series data and inspect how the detrended dataset behaves. To detrend the model, we used the model fitting method. This method involves fitting a linear model to the data that represents the general trend and subtracting the observed series data from the predicted data (1).

$$value(t) = observed(t) - predicted(t) \qquad (1)$$

Figure 7 shows the linear regression model fitted to the series data against the observed data.

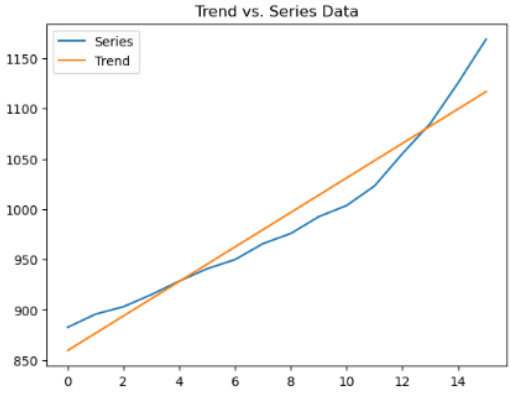

Figure 7

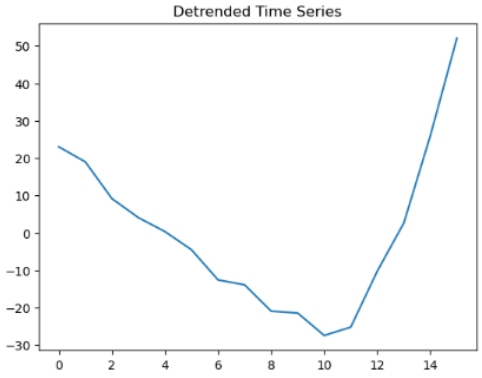

Figure 8

After subtracting the linear regression's predictions from the actual series data, we can plot the detrended time series data. From the detrended plot (Figure 8), we can infer that a nonlinear trend may fit the data better.

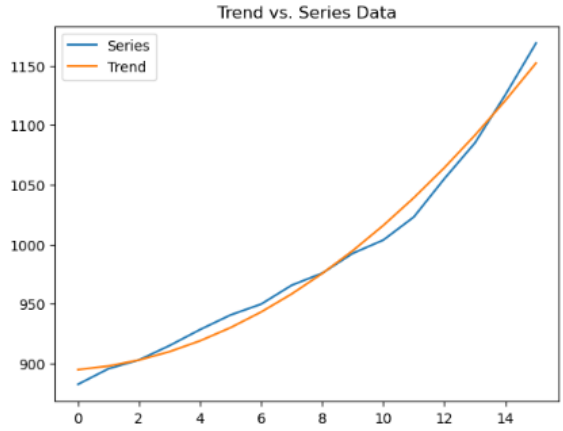

Figure 9

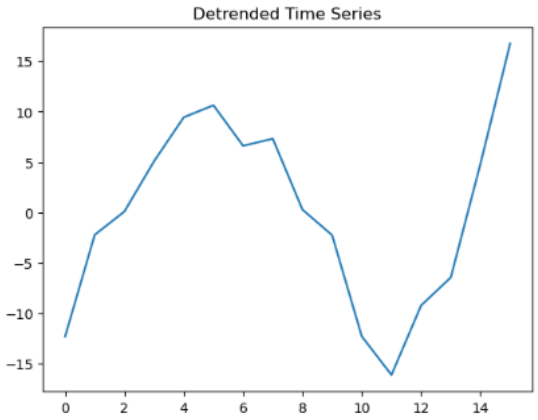

Figure 10

Figure 9 and Figure 10 show a polynomial model of degree 2 against the observed values, and the corresponding detrended time series. Based on these results, we can conclude that a polynomial model of degree 2 fits the trend of deforestation best.

The ARIMA model was built with parameters p, d, and q set to 1. The predictions were made with rolling forecasting. Since the observations at each time step are dependent to the previous one, we will recreate the ARIMA model after receiving each new observation. As seen in Figure 9, the plots for the time series data and rolling forecast are close. The test RMSE of the ARIMA model is 10.697.

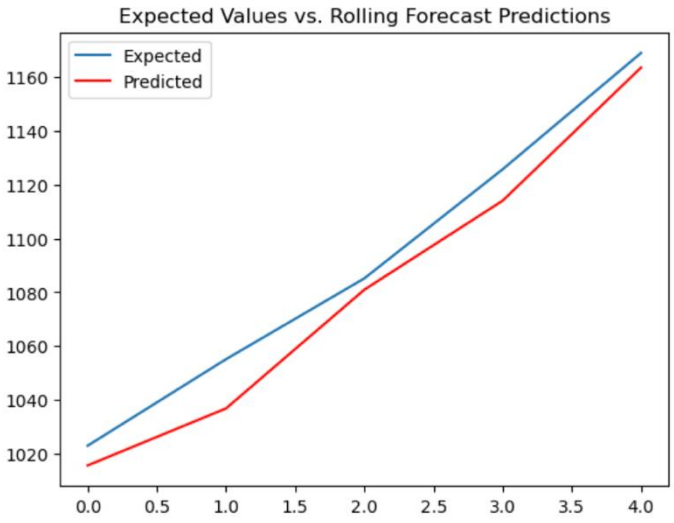

Figure 11

Our next approach uses machine learning algorithms using time series data by restructuring the data to look like a supervised learning problem. To accomplish this, we use the previous time steps as the input variables and the next time step as the output variable, making it a supervised learning problem (one step forecasting). For prediction, we used both linear regression and stochastic gradient descent. Prior to modeling, we normalized the data. To measure the performance of the linear regression, we calculated the RMSE and R-squared statistic, which were 0.02871 and 0.91999 respectively. Next, we found the coefficients of the regression function using a gradient descent algorithm and made predictions. Our RMSE and R-squared for the stochastic

gradient descent algorithm was 0.03023 and 0.9113 respectively. Based on the MSE and R-squared statistics for these two models, the linear regression model had superior performance.

3.2 Ensemble Stacking

In the ensemble stacking modeling approach we treated deforestation as a regression problem instead of time series. To achieve this, we trained multiple regression models as our base layer and stacked an OLS on top of it to predict the final deforestation increment(areakm).

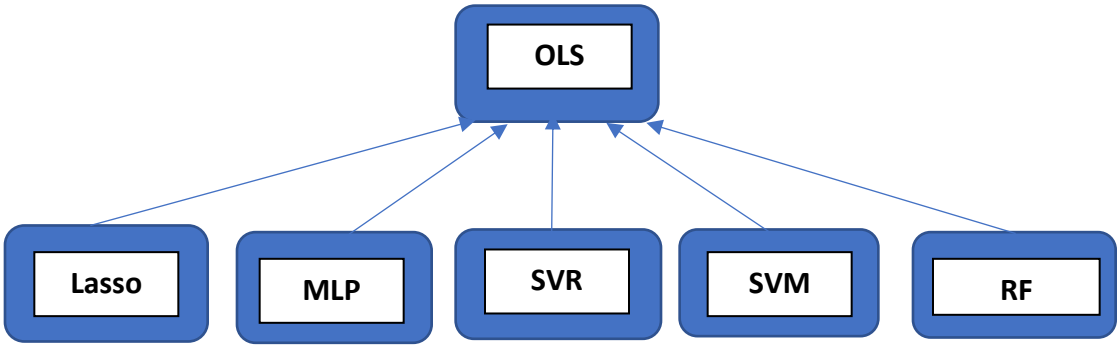

Figure 12

Figure 12 shows the architecture of the stacking approach.

We split the dataset into three different parts. The observations of the last three years of 2020,2021 and 2022 were separated as the holdout set, and this hold out set will be used to validate the final output of the stack model. From the remaining data, the observations of the last three years 2017,2018 and 2019 we separated as the meta set to validate the first layer of models, and the remaining data will be used as the base set to run the grid search.

Following are the steps:
Step 01: Using the base set run GridSearchCV for all the ML algorithms to find the best hyperparameters.

Algorithm	**Hyperparameters from GridSearchCV**
Lasso	{'alpha': 0.1, 'fit_intercept': 1}
MLP Regressor	{'activation':'relu','alpha':1e-06, 'hidden_layer_sizes': (24, 12)}
SVR	{'C': 33, 'epsilon': 14, 'fit_intercept': 0}
SVM	{'C': 0.2, 'gamma': 0.2, 'kernel': 'sigmoid'}
Random Forest	{'max_features': 'log2', 'min_samples_split': 3, 'n_estimators': 300}

Step 02: Predict using the meta set for all five regression models and create a data frame with results. Fit linear regression model on this result data.
Step 03: Predict the hold out set using the base models and create a new data frame with those results as features and deforestation area increment of the hold out set as the output variable.
Step 04: Using Linear Regression model which is our Stack Model, predict the hold outset deforestation area increment using the results of base layer models.

3.3 Multistate Model

Another minor approach taken was using an ensemble method, both with raw and normalized data. The year and GDP of the states were the data columns used to predict the area increment of deforestation. This ensemble method initially started as a multistate method. Using a gradient boosted model, an xgradient boosted model, and a random forest regression, the average prediction was taken after being fit on the raw/normalized data from each state. This created a dictionary of models that would be tailored to each states individual data.

4 Project Trajectory, Results, and Interpretation

4.1 Time Series Forecasting

Algorithm	RMSE
ARIMA (Rolling Forecast)	10.697
Linear Regression	0.02871
Stochastic Gradient Descent	0.03023

The model that most accurately modeled the historical data was the linear regression model. This model had the lowest RMSE of all the models we trained, with an RMSE of 0.0287. Since this model most accurately modeled our data, it should be used for forecasting future areas of deforestation.

4.2 Ensemble stacking approach.

Looking at the confusion matrix(Figure 13), the correlation between lasso and mlp is the highest and they perform about the same. Whereas Other models yields different results. Its evident looking at the table below, because MSE of Lasso and MLP are about the same and RMSE is quite similar. Out of the models in the base layer Random Forest has the lowest MSE and SVR has the lowest RMSE.

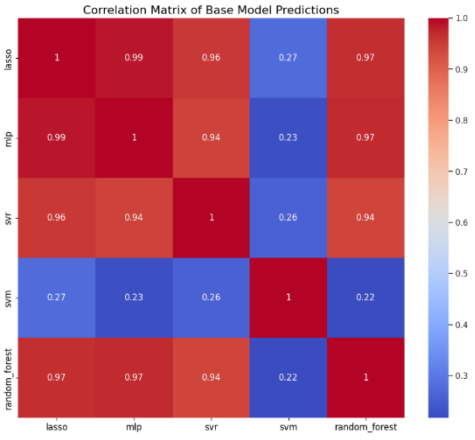

Figure 9

Model	MSE	RMSE
Lasso	11.0522	40.4272
MLP	11.0825	39.2084
SVR	22.2282	29.7057
SVM	18.6302	55.5488
Random Forest	10.3567	37.4177
Stack Model	9.2552	28.3874

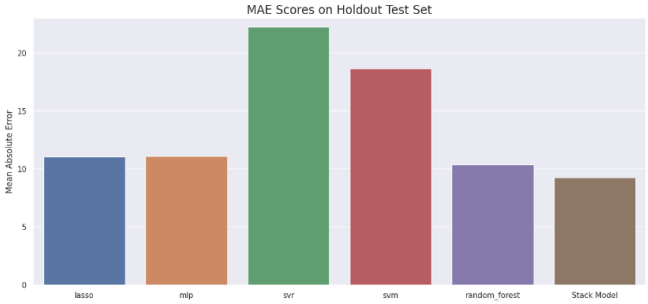

Figure 14

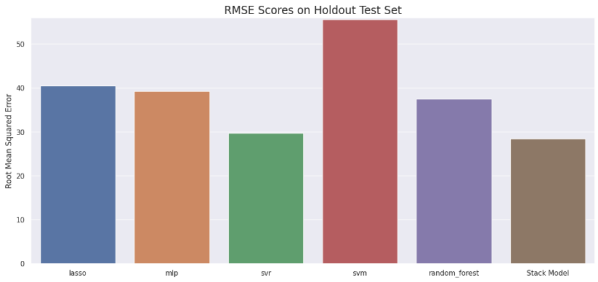

Figure 15

Overall the Stack Model(OLS) improves the accuracy by comibining the results of the base models.

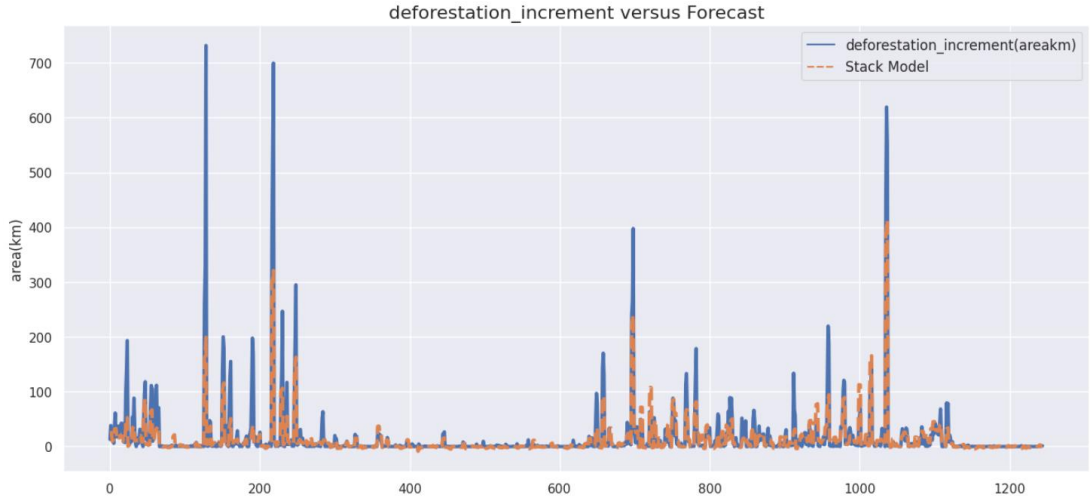

Figure 16

Figure 16 shows the difference between the true value and the stacked model prediction of the observations in the holdout set. We can observe that there are significant amounts of residuals.

If we manage to find or augment data withint a wider time range, time series forecasting will be the best approach.

4.3 Multistate Modeling

Gradient Boosted Regression	Raw Data MSE: 5977.14	Normalized Data MSE: 0.0017
XGradient Boosted Regression	Raw Data MSE: 786.05	Normalized Data MSE: 0.0008
Random Forest Regression	Raw Data MSE: 7319.39	Normalized Data MSE: 0.0019
Average	Raw Data MSE: 3365.66	Normalized Data MSE: 0.0012

5 Conclusion and Future Work

With the models we've provided, there is substantial evidence that shows we can use both the original forest area data to predict the future of the Amazon, and also use the

economic data to assist or entirely predict the rainforest's future condition. Some future ideas for these models include creating fake data to represent some of the past years where the rainforest was not being supervised as heavily, comparing models that use economic data to models that do not use it, and finally finding other economic factors that correspond to the Amazon's condition. The biggest weakness to our project so far has been the lack of data, since we are only using yearly data that translates to either a lot of disjointed sparse data, or a mass accumulation of spare data. we believe the best solution to this is to use augmented data created by trained models such as these to create highly accurate representations of what monthly data for the past half century should look like.