**Dataset and Exploration**

Our primary model is to predict the deforestation in Brazil’s Amazon rainforest. For this we collected two datasets from the ( [Terrabrasilis – Geographic Data Platform](http://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/legal_amazon/rates))).

**Dataset 01**: This contains state wise deforestation increment data for five ecoregions in Brazil from 2001 to 2021. The ecoregions are Pantanal, Cerrado, Caatinga, Mata Atlantica and Pampa. The dataset has 811 rows and 4 features. Figure 1 shows how the dataset looks like.

Text

Description automatically generated Chart

Description automatically generated

*Figure 1: First 5 rows in Dataset 01 Figure 2: Box plot for dataset 01*

Chart, histogram

Description automatically generated Square

Description automatically generated

*Figure 3: Distribution of ‘areakm’ in dataset 01 Figure 4: Correlation matrix of dataset 01*

According to Figure 2, feature ‘areakm’ has a considerable number of outliers over the upper bound. This is because some forest areas are larger than the rest. Looking at Figure 3 we can see that the distribution of area is skewed towards the right. Observing heatmap in Figure 4 we can see that ‘year’ and ‘areakm’ have a negative correlation. This negative correlation can be observed in Figure 5 as the deforestation increment eventually decreases.

Chart, line chart

Description automatically generated

*Figure 5: Variation of deforestation region wise*

**Dataset 02**: This contains municipality wise accumilated deforestation data related to 8 states in Brazil from 2007 to 2022. The states are Acre, Amapa, Amazonas, Maranhao, Mato Grosso, Rondonia, Roraima and Tocantins. The dataset has 6640 rows and 5 features. Figure 6 shows how the dataset looks like.

Table

Description automatically generated Chart, box and whisker chart

Description automatically generated

*Figure 6: First 5 rows in Dataset 01 Figure 7: Box plot for dataset 02*

According to Figure 7, feature ‘areakm’ has a considerable number of outliers over the upper bound. This is because some forest areas are larger than the rest. Looking at Figure 8 we can see that the distribution of area is skewed towards the right. Observing heatmap in Figure 4 we can see that ‘year’ and ‘areakm’ have a positive correlation. As the accumulated deforestation area increases each year. Figure 10 shows the autocorrelation plot for the average deforestation of state Acre. There is a high degree of autocorrelation between consecutive observations. The data is not random and has some relationship.

**Analysis:**

Chart, histogram

Description automatically generated Chart, treemap chart

Description automatically generated

*Figure 8: Distribution of ‘areakm’ in dataset 02 Figure 9: Correlation matrix of dataset 02*

Chart, histogram

Description automatically generated Chart, line chart

Description automatically generated

*Figure 10: Variation of deforestation for state Acre Figure 10: Autocorrelation plot for the series of state Acre*

**Baseline Model**

We fit different models for the two datasets separately. For dataset 01 we fitted a Linear Regression model which gave us a mean accuracy of 0.483. The model we obtained at first is almost meaningless.

For convenience in analysis, we took the mean deforestation over the years in state Acre and trained an ARIMA model as the baseline. We got a RMSE of 11.524 and according Figure 12 we can see that the expected value and rolling forecast are quite close.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

*Figure 11: Plot of trend in the dataset 02 Figure 12: Expected value vs Rolling forecast comparison.*

**Conclusions**