Satellite Data Monitoring Global Deforestation

By James Hoang

Overview of The Subject Area

Problem Statement: Using machine learning, how might we predict areas at risk for deforestation such that we can determine areas ideal for logging by least environmental impact?

Solution: Using carbon and forest cover data to help illustrate areas which have the highest contribution to deforestation where its suitable for logging

Potential Impact Estimate: environmental longevity allows for companies to assess risk management on areas to log, along with adhering to regulations to avoid potential fines or penalties

Overview of dataset & preprocessing

Datasets:

- 2 main datasets, carbon data and tree cover data
- Cover loss and gross carbon emissions from 2001-2023
- Forest thresholds to classify forested areas

Quality Concerns & Preprocessing:

- Environmental data is small
- High collinearity and dimensionality when binarizing country data (265), including subnational data (3000+)
- Lasso was ineffective despite a high alpha and would not be feasible for modelling

Important EDA Findings

Preliminary EDA Findings:

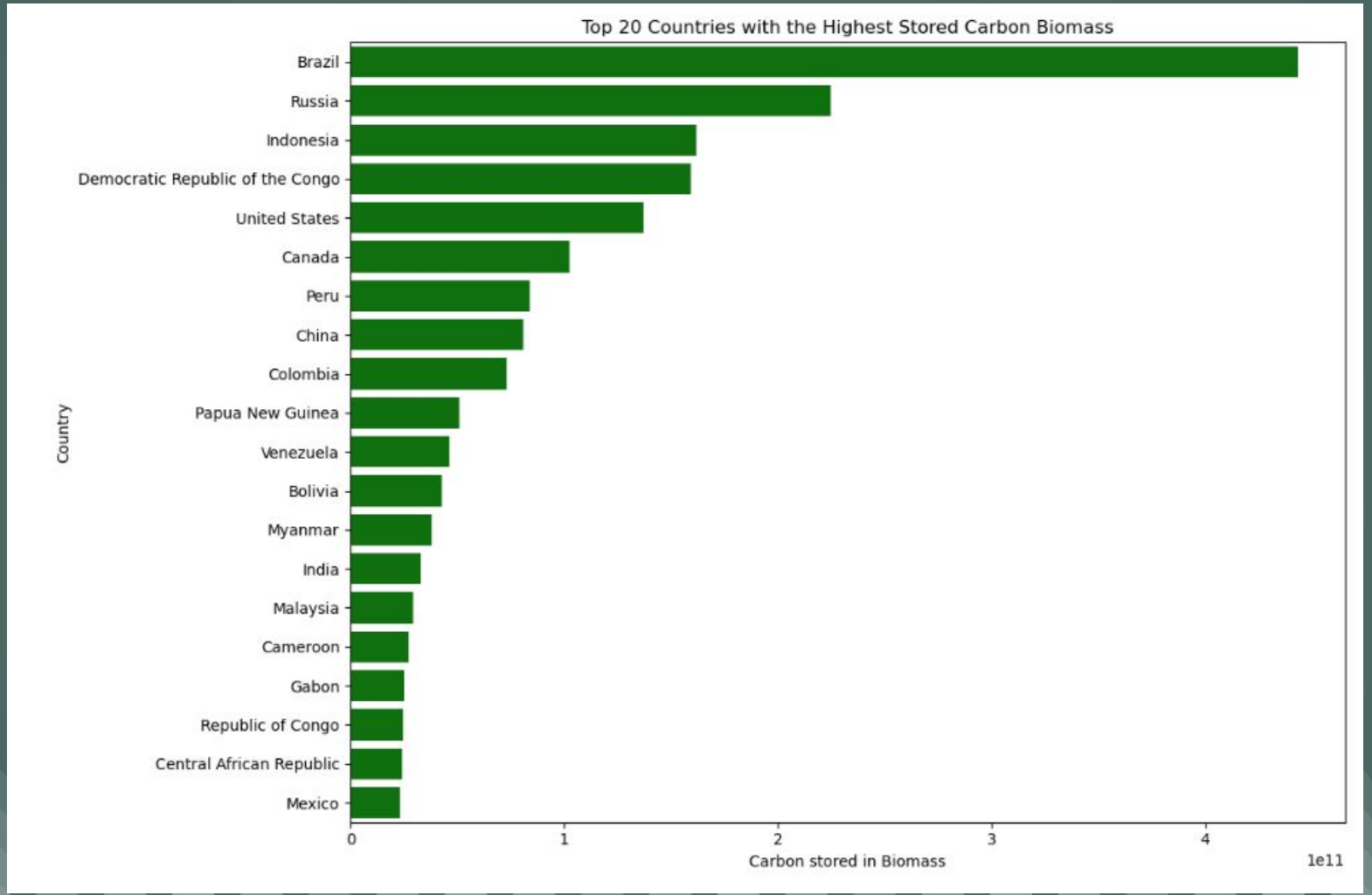
- Gradual canopy density loss over the years (spike in 2015-2016)
- Brazil followed by Canada are huge carbon emitters

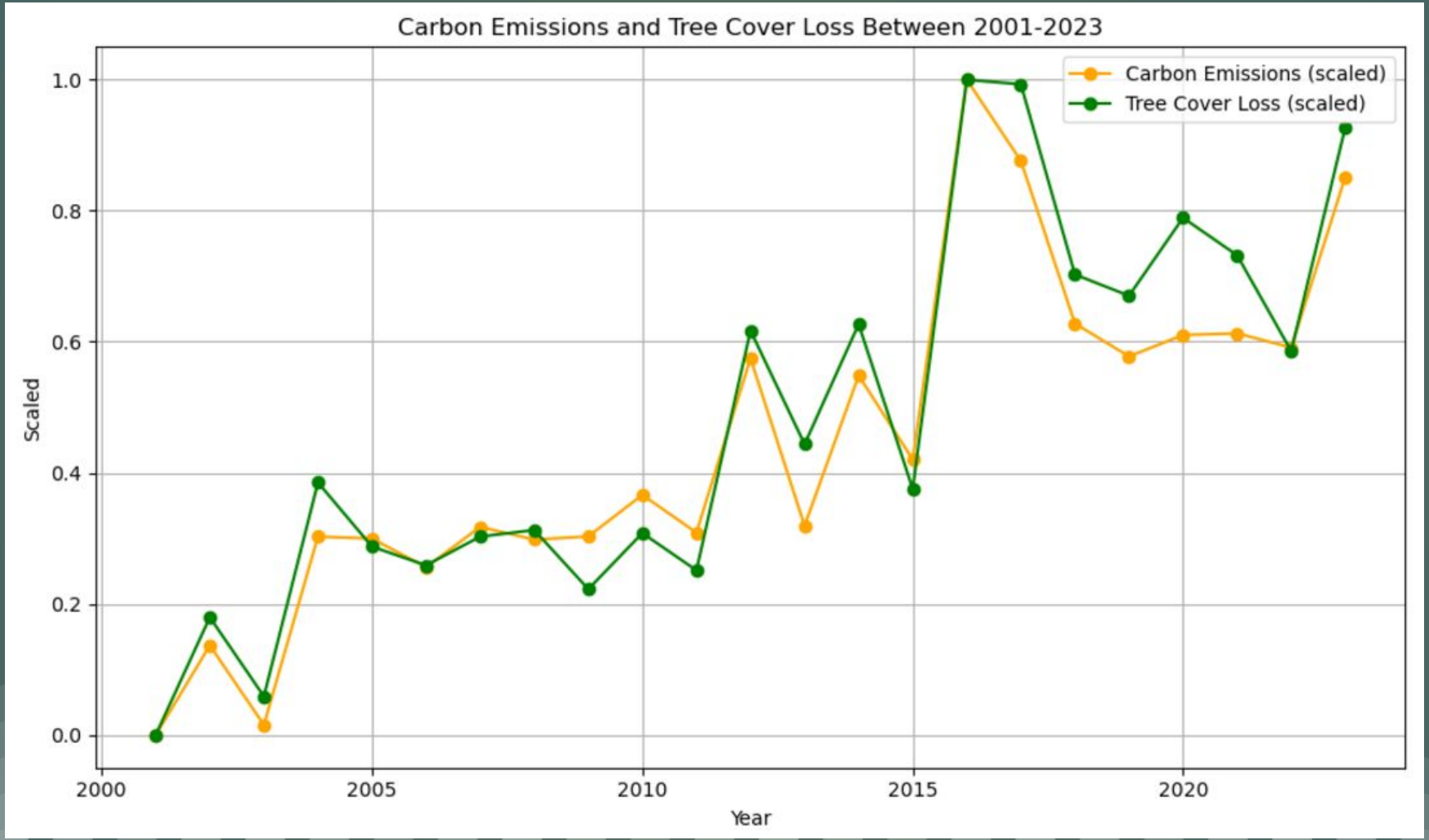
3 Main Findings:

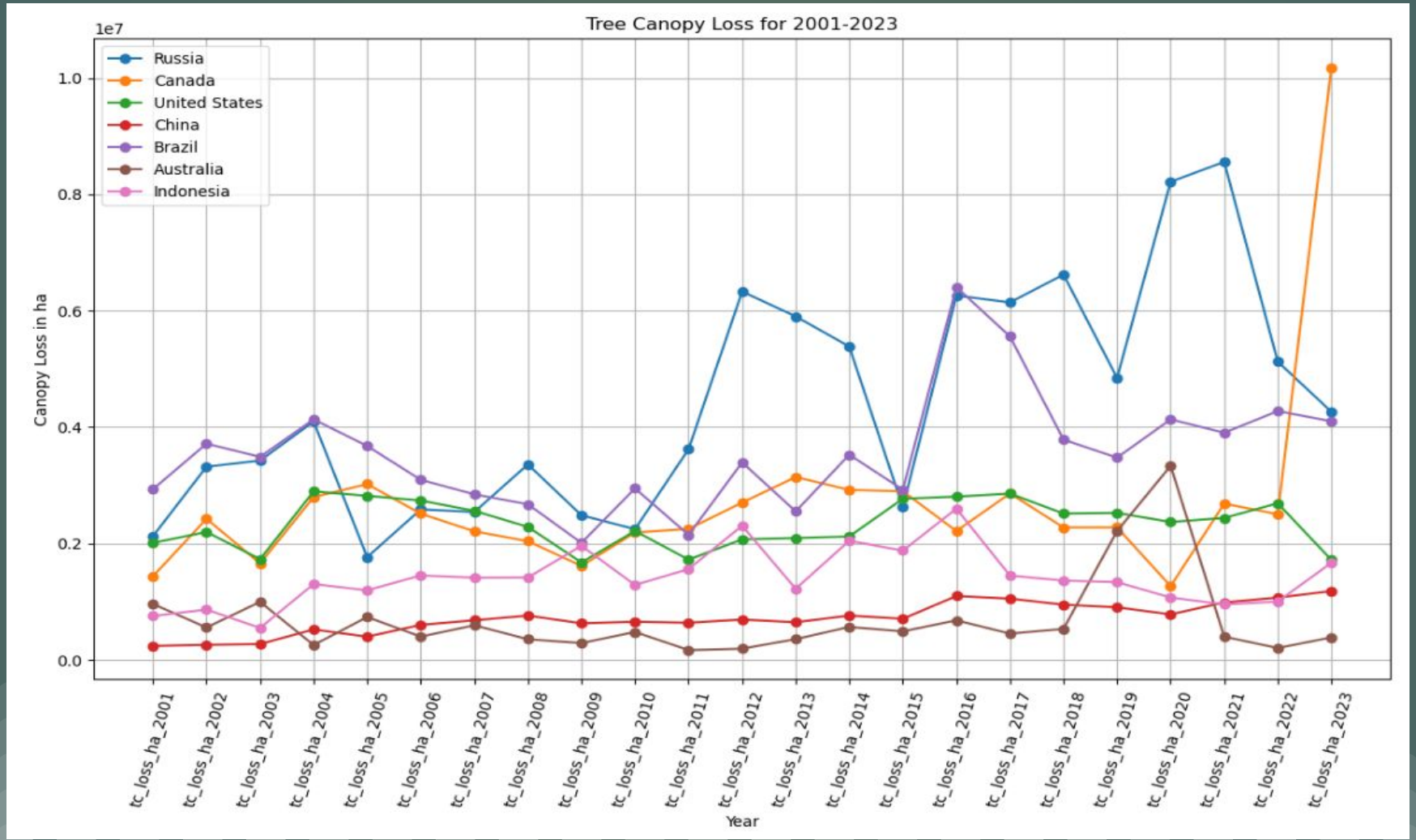
- Most carbon is stored in primary forests which are areas that are most of interest for logging

- Carbon emissions and Tree cover are very positively correlated

- Tree Cover loss sporadicness has increased over time and can be explained through changes within this topic space like wildfires


Top 20 Countries with the Highest Stored Carbon Biomass
Brazil
Russia
Indonesia
Democratic Republic of the Congo
United States
Canada
Peru
China
Colombia
Papua New Guinea
Venezuela
Bolivia
Myanmar
India
Malaysia
Cameroon
Gabon
Republic of Congo
Central African Republic
Mexico
Country
0
1
2
3
4
Carbon stored in Biomass
1e11


Carbon Emissions and Tree Cover Loss Between 2001-2023
Carbon Emissions (scaled)
Tree Cover Loss (scaled)
Scaled
Year
0.0
0.2
0.4
0.6
0.8
1.0
2000
2005
2010
2015
2020


Tree Canopy Loss for 2001-2023
1e7
Russia
Canada
United States
China
Brazil
Australia
Indonesia
Canopy Loss in ha
1.0
0.8
0.6
0.4
0.2
0.0
tc_loss_ha_2001
tc_loss_ha_2002
tc_loss_ha_2003
tc_loss_ha_2004
tc_loss_ha_2005
tc_loss_ha_2006
tc_loss_ha_2007
tc_loss_ha_2008
tc_loss_ha_2009
tc_loss_ha_2010
tc_loss_ha_2011
tc_loss_ha_2012
tc_loss_ha_2013
tc_loss_ha_2014
tc_loss_ha_2015
tc_loss_ha_2016
tc_loss_ha_2017
tc_loss_ha_2018
tc_loss_ha_2019
tc_loss_ha_2020
tc_loss_ha_2021
tc_loss_ha_2022
tc_loss_ha_2023
Year

Baseline Model and Evaluation Metrics

Linear Regression:

- For carbon emissions and tree cover
- Target variable was net carbon emissions and canopy gain

Evaluation Metrics:

- R-squared and means squared error
- Overfitting issues

```
Training MSE: 0.0522535270901479, Training R squared: 1.0
Test mean squared error: 0.06347994685042575, Test R squared: 1.0
```

```
Training Mean squared error 38834315161.65923, Training R squared: 0.9951980990007283
Test Mean squared error: 37527616384.50387, Test R squared: 0.9959806816832032
```

Advanced modeling next steps

Within the problem space, our EDA illustrated key forests that contribute to deforestation

PCA: See how much dimensionality we can reduce from the country features with ideally 80-90% data explained

Hyperparameter tuning: Adjusting for most relevant countries

Contribution Metrics: Explore the possibilities of developing a more robust metric for contribution to deforestation

Thanks for Listening!

Additional Links for Field Overview

https://hub.arcgis.com/datasets/f0cc32be502b49bc87711249ff5dcdfb/explore?location=6.772634%2C7.042178%2C1.48

https://www.statista.com/statistics/238893/ten-countries-with-most-forest-area/

https://www.statista.com/statistics/1346900/largest-rainforests/

https://www.noaa.gov/climate

https://earthdata.nasa.gov/data/catalog?granule_data_format_h[]=CSV

https://research.wri.org/gfr/data-methods#data-sets

https://oec.world/en/profile/hs/wood-products?yearSelector1=2020