Capstone Proposal

Machine Learning Engineer
Nanodegree

Lucas Pontes

Apr 18,2021

Domain Background

Brazil is one of the world's largest producers of agricultural products, especially grains such as soybeans and corn and also meat, making up the country's main economic activity. At the same time, the country has one of the largest forested areas in the world and important biomes with endemic species of fauna and flora. Despite having a very complete and current environmental legislation that clearly defines the areas that can be used for agriculture and those that must be preserved, there are still many conflicts of interest between production and conservation. Such conflicts are gaining attention around the world as legal restrictions are adopted for agricultural products without certification.

Due to the extension of the agricultural area and the difficulty of access to forest areas and the monitoring of activities developed in these regions, the use of technologies, especially satellite images, has been increasingly adopted for monitoring and environmental regulation in Brazil. In this sense, several companies and startups have been created to use remote sensing data in agricultural production with surprising results that allow to increase productivity while guaranteeing the environmental conservation and the origin of the products.

One of the main difficulties is to monitor agricultural production areas and ensure that producers respect preservation areas. Therefore, the idea of using satellite images is very interesting to check whether the limits imposed by environmental legislation are being followed. However, in order to transform the information in the satellite images, it is necessary to classify the images automatically, which allows assessing changes in land use over time and whether the preservation areas are respected.

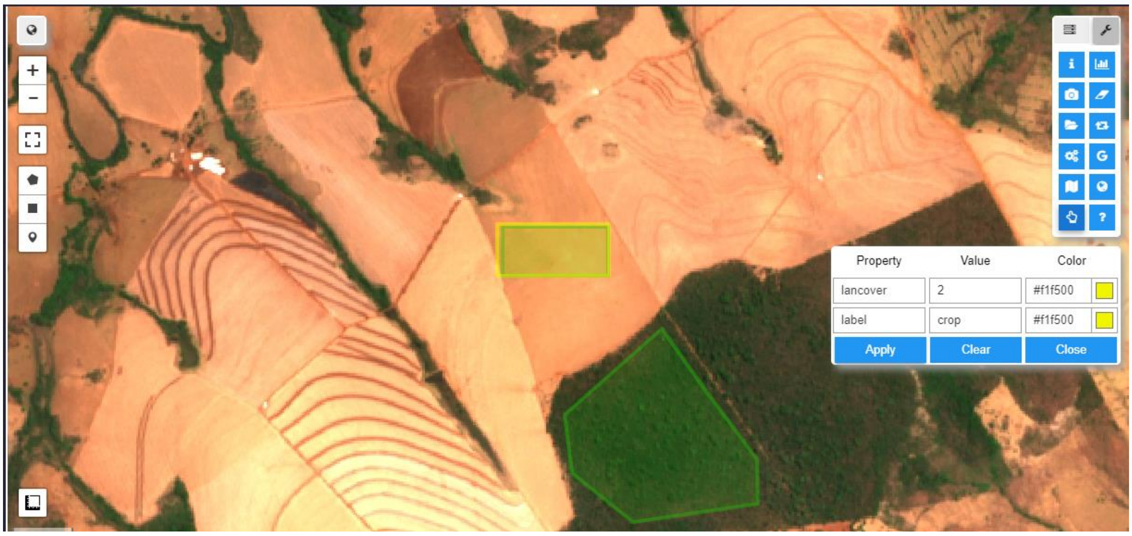
Having worked as a researcher in the environmental and agricultural fields, using geoprocessing techniques and data science for sustainable agriculture, I aim to work as a data scientist in a company related to agriculture.

Problem statement

The main problem I aim to solve with this project is to provide a easy to use application that allows to users to interactively choose the area and period for which they wish to know the LULC. That can be used to compare with the preservation areas stablish by law and declared by farmers.

Datasets and Inputs

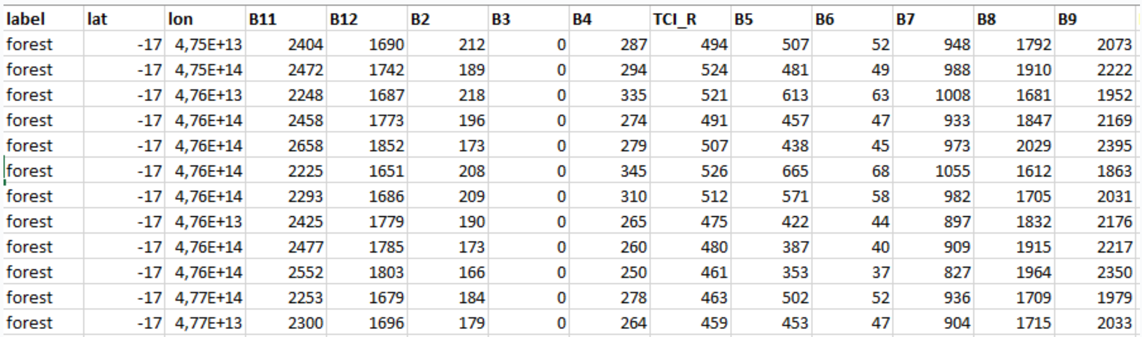
One of the main difficulties to use models to predict the LULC with satellite images is to tag the pixels of images given their band values, what is traditionally done manually. However, using geemap python package (Wu, 2020) I create an script that made the data acquisition and tagging much more easy once one can extract a large amount of samples simply by drawing polygons on a map plot (Figure 1).



Property	Value	Color	
lancover	2	#f1f500	🟨
label	crop	#f1f500	🟨

Figure 1. geemap used on jupyter notebook script to extract pixel band values with given classification.

The script extracts the band values and coordinates of each pixel from the polygons together with the classification given for each polygon (Figure 2, and file attached).



label	lat	lon	B11	B12	B2	B3	B4	TCI_R	B5	B6	B7	B8	B9
forest	-17	4,75E+13	2404	1690	212	0	287	494	507	52	948	1792	2073
forest	-17	4,75E+14	2472	1742	189	0	294	524	481	49	988	1910	2222
forest	-17	4,76E+13	2248	1687	218	0	335	521	613	63	1008	1681	1952
forest	-17	4,76E+14	2458	1773	196	0	274	491	457	47	933	1847	2169
forest	-17	4,76E+14	2658	1852	173	0	279	507	438	45	973	2029	2395
forest	-17	4,76E+14	2225	1651	208	0	345	526	665	68	1055	1612	1863
forest	-17	4,76E+14	2293	1686	209	0	310	512	571	58	982	1705	2031
forest	-17	4,76E+13	2425	1779	190	0	265	475	422	44	897	1832	2176
forest	-17	4,76E+14	2477	1785	173	0	260	480	387	40	909	1915	2217
forest	-17	4,76E+14	2552	1803	166	0	250	461	353	37	827	1964	2350
forest	-17	4,77E+14	2253	1679	184	0	278	463	502	52	936	1709	1979
forest	-17	4,77E+13	2300	1696	179	0	264	459	453	47	904	1715	2033

Figure 2. Head of the table obtained with geemap script that will be used as train and test data to the modelling process.

Columns named B1, B2, etc are the image bands, e.g. B4, B3 and B2 are the RGB bands respectively on the Sentinel image. These values are used to train machine learning supervised model. In the same way, the user will be able to select the area, by drawing polygons, and the app will retrieve the pixels and band values from the image inside the polygon, that will be used to model prediction.

The developer of geemap package achieved 95% accuracy with this same approach and using a random forest algorithm.

Solution statement

Given that the focus of ML engineer nanodegree is on deployment, I will not spend to much effort testing more complex and custom models such as CNN. As stated on benchmark there are good results using random forest, so I will probably use LightGBM or XGBoost.

The app will be hosted on Heroku. I will use Docker to encapsulate the deliverables (notebooks, html page and requirements). The fitted model will be saved as a pkl file that will be used to deploy the model and make predictions.

Evaluation metrics

Once this is not an imbalanced problem I will use Accuracy score as the performance metric.

References

Wu, Q., (2020). geemap: A Python package for interactive mapping with Google Earth Engine. The Journal of Open Source Software, 5(51), 2305. https://doi.org/10.21105/joss.02305