ARTIFICIAL INTELLIGENCE IN CROPLAND MAPPING

Julius Maina

julmngii@gmail.com

The author is a research student in data analytics at KCA University in Kenya

October 10th, 2023

Abstract *– This project addresses the pressing need for accurate global crop maps using advanced machine learning techniques. Current freely available land cover products have limitations, including infrequent updates and inconsistent definitions of "cropland." Leveraging new high-resolution satellite data and machine learning, this project aims to develop precise and cost-effective cropland classification models. The project offers an opportunity to contribute to global agricultural landscape mapping and enhance our understanding of crop extents worldwide.*

Keywords – Cropland mapping, machine learning models, remote sensing data, data preprocessing, feature engineering, hyperparameter tuning, accuracy metric, agricultural monitoring, optimization, and LightGBM classifier.

1. INTRODUCTION

In the realm of agriculture and environmental research, the accurate delineation of cropland extents plays a pivotal role. These maps serve as the foundation for various applications, including agricultural planning, resource management, environmental monitoring, and sustainability assessments. While several existing land cover products offer global cropland extent maps, they suffer from significant limitations.

Firstly, these products are not updated annually, hindering their utility for monitoring dynamic changes in cropland patterns over time. Secondly, each product employs its own definition of "cropland," which often deviates from the precise definitions used in agricultural research, such as FAOSATA's classification of "6620 cropland" or "6621 arable land." This disparity in definitions introduces discrepancies that affect the reliability of these maps. Consequently, there exists a critical need to address these challenges and advance the mission of global high-resolution cropland extent mapping using remote sensing data.

Fortunately, the field of machine learning and artificial intelligence offers promising solutions to improve the accuracy, consistency, and timeliness of cropland mapping using satellite imagery. With the advent of new earth observation plans and the increasing availability of high-resolution satellite data, the opportunity to harness these technologies for more precise and comprehensive land cover classification has never been more promising.

To tackle these challenges and further the cause of global cropland mapping, this challenge is centered on the development of an accurate and cost-effective classification model. This model, based on the powerful LightGBM framework, seeks to provide a single, robust solution for cropland extent mapping. By participating in this challenge, researchers and practitioners can contribute to advancing the state of the art in global cropland mapping. This, in turn, will enable a more precise and comprehensive understanding of agricultural landscapes on a global scale, benefiting agriculture, environmental science, and sustainable development.

2. DATASET

In the pursuit of global cropland mapping, this endeavor encompasses a diverse array of satellite imagery datasets, meticulously curated and harnessed to advance the boundaries of knowledge. Care was taken to use the timespans required by the competition host, that is Afghanistan in April 2022, and for both Iran and Sudan from July 2019 to June 2020 for both train and validation datasets.

In the test regions of Iran, Afghanistan and Sudan, participants were provided access to a structured 15-day composited Sentinel-2 time series dataset. For each of the countries,12 sentinel-bands were extracted from Google Earth Engine (GEE). This official dataset formed the foundation of the research, offering invaluable insights into the dynamic landscape of cropland.

The Afghanistan test region presented a unique challenge and opportunity for autonomous data

collection. This work ventured beyond provided data sources, seeking to augment the availed data and broaden the scope of analysis. As such, an additional 500 training samples, meticulously labeled from Afghanistan's Nangarhar province, were introduced to enrich the dataset. These samples, meticulously prepared, signify a commitment to precision and a contribution to the broader academic discourse.

The research journey was facilitated by the utilization GEE. Initiated with the registration of a Google account and subsequent GEE enrollment, this tool proved instrumental in data acquisition and analysis.

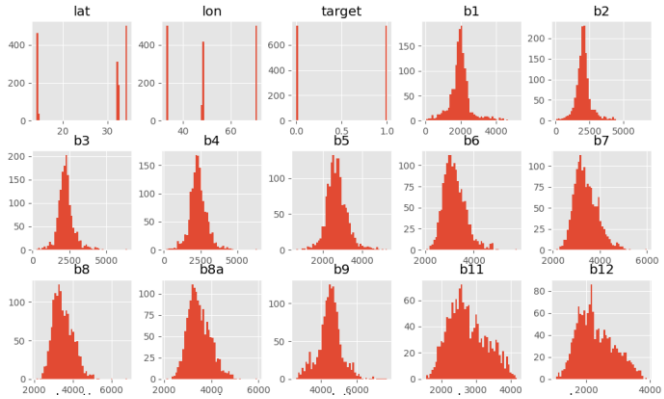
3. DATA PREPROCESSING

The foundation of any machine learning endeavor rests upon the quality and suitability of the data at hand. In this section, we delve into the intricacies of the data preprocessing pipeline employed to prepare the raw dataset for subsequent modeling tasks. Each step in this meticulous process is elucidated to shed light on its rationale and impact on the overall model performance.

Categorical variables, an intrinsic component of the dataset, necessitated encoding to facilitate their integration into the model. It is worth noting that specific columns ('id' and 'target') were excluded from this encoding process. The LabelEncoder from the scikit-learn library was deployed for this purpose, ensuring that categorical attributes were translated into a numerical format understandable by machine learning algorithms.

Robustness against outliers is paramount in any predictive modeling endeavor. To mitigate the potential influence of outliers on the model's performance, a robust capping strategy was employed. Specifically, for a set of pre-defined columns, the values falling beyond a specified threshold (3 standard deviations from the median) were replaced with their respective threshold values. This strategic capping ensured that the dataset remained resilient to extreme values.

The issue of skewed features was tackled through a log transformation strategy. Specifically, for a predefined set of columns, a natural logarithm transformation was applied to the data. This approach, implemented using `np.log1p`, effectively mitigates the skewness while avoiding challenges associated with zero or negative values.



4. FEATURE ENGINEERING

Feature engineering is the cornerstone of predictive modeling, as it has the potential to unlock hidden patterns and relationships within the data. In this section, we explore the intricate steps taken to transform raw data into a feature-rich dataset, poised to empower our machine learning models.

4.1 Geographical Enrichment

Geospatial data often holds valuable insights. To harness this potential, we employed a geocoder to enrich our dataset. By inferring location-based information from latitude and longitude coordinates, we expanded our feature set to include place names, states, counties, and country codes. This augmentation allows our models to account for geographic variations and patterns.

4.2 Location-Based Features

In addition to geospatial enrichment, we introduced a location-based feature that combines rounded latitude and longitude values. This feature not only facilitates sorting but also captures the spatial context of data points, enabling our models to consider the proximity of observations.

4.3 Country Classification

To address the unique characteristics of different countries within our dataset, we categorized data points into three distinct regions: Afghanistan, Iran, and Sudan. This classification allows our models to adapt to country-specific dynamics, potentially improving predictive accuracy.

4.4 Quartile Binning

A fundamental aspect of feature engineering is creating informative bins for continuous features. In our case, we employed quartile binning for select features, dividing them into four distinct quartiles.

This approach transforms continuous variables into categorical ones, providing models with a more granular view of the data.

4.5 Soil-Related Features

Soil characteristics can significantly impact land use patterns. To capture these nuances, we computed several soil-related features, including 'BI,' 'BIXS,' 'BaI,' and more. These features provide insights into soil conditions and their potential influence on cropland mapping.

4.6 Urban-Related Features

Urban areas exhibit distinct patterns in satellite imagery. To leverage this information, we derived urban-related features such as 'BRBA,' 'DBI,' 'NDBI,' 'IBI,' 'NBAI,' and 'NBUI.' These features shed light on urbanization trends and their implications for land use.

4.7 Water-Related Features

Water bodies have a significant impact on land classification. To account for this, we engineered features like 'ANDWI,' 'LSWI,' 'MBWI,' and more. These features capture the presence and characteristics of water bodies, facilitating water-related land use analysis.

4.8 Vegetation-Related Features

Vegetation plays a vital role in land cover classification. To account for vegetation dynamics, we introduced features such as 'AFRI1600,' 'AFRI2100,' 'ARI,' 'ARI2,' 'ARVI,' 'BNDVI,' and 'BWDRVI.' These features illuminate vegetation patterns and dynamics within the dataset.

4.9 Burn and Kernel Features

Understanding burned areas and employing kernel-based indices can provide valuable insights into land use. We introduced 'KEVI,' 'kIPVI,' 'kNDVI,' 'kRVI,' and 'kVARI' features, which capture various aspects of land dynamics, including burn patterns.

4.10 Snow-Related Features

In regions prone to snow, detecting snow cover is essential for accurate land use analysis. We engineered 'NBSIMS,' 'NDGlaI,' and 'NDSI' features to capture snow-related patterns, aiding snow-related land use classification.

4.11 Rotation Features

Rotating geographical coordinates can reveal different perspectives of the landscape. We applied rotation to latitude and longitude coordinates, introducing features 'rot_45_x,' 'rot_45_y,' 'rot_30_x,' and 'rot_30_y.' These features offer alternative representations of the data, potentially enhancing model performance.

4.12 Geographical Clustering

Geographical clustering can unveil spatial groupings within the data. Employing K-means clustering with 30 clusters, we created 'gspatial_' features, allowing our models to consider spatial relationships and dependencies among data points.

The final dataset now contained 260 features.

5. MODEL SELECTION

To In this section, we elaborate on our approach to model selection, parameter tuning, and ultimately utilizing the LightGBM classifier for the task at hand.

5.1 Model Comparison

To ensure the efficacy of our predictive model, we commenced with a comprehensive comparison of various machine learning algorithms. The following models were considered for evaluation:

- ☐ Logistic Regression
- ☐ Random Forest
- ☐ AdaBoost
- ☐ XGBoost
- ☐ HistGradientBoosting
- ☐ LightGBM
- ☐ CatBoost

These models were trained and tested to gauge their performance on our dataset. The evaluation was based on accuracy, a fundamental metric for classification tasks. The results of this initial comparison are summarized below:

Model	Accuracy
XGBoost	0.884
LightGBM	0.880
HistGradientBoosting	0.874
Random Forest	0.866
CatBoost	0.864
AdaBoost	0.828
Logistic Regression	0.682

It is evident from the comparison that both XGBoost and LightGBM outperformed other models in terms of accuracy.

5.2 Parameter Tuning with Optuna

Having identified LightGBM as the optimal choice because of its speed over XGBoost, we proceeded to fine-tune its hyperparameters to maximize its predictive potential. We employed Optuna, a powerful hyperparameter optimization library, to systematically search for the best combination of hyperparameters. The objective function for Optuna was defined to minimize the misclassification error, which is equivalent to maximizing classification accuracy.

The hyperparameters subjected to optimization included:

- ☐ Learning rate
- ☐ Number of boosting rounds
- ☐ Maximum depth of trees
- ☐ L2 regularization strength (lambda_l2)

Through a series of trials, Optuna systematically explored the hyperparameter space and converged to the best set of hyperparameters that enhanced model performance. The optimized hyperparameters were then utilized for the subsequent modeling phase.

5.3 Model Building with Optimized Parameters

With the best hyperparameters determined, we constructed the final LightGBM classifier model. This model was trained on the entire training dataset using the following hyperparameters:

- ☐ Objective: Binary classification
- ☐ Random state: Fixed for reproducibility
- ☐ Number of threads: 4 for parallel processing
- ☐ Number of boosting rounds: Optimized value
- ☐ Early stopping rounds: 20 for efficient training
- ☐ Learning rate: Optimized value
- ☐ Maximum depth of trees: Optimized value
- ☐ L2 regularization strength: Optimized value

We employed repeated k-fold cross-validation with 5 folds and 3 repeats to ensure robust model evaluation and selection of the best hyperparameters. This approach allows for thorough testing and validation of the model's performance.

5.4 Model Prediction

Upon training the LightGBM model with the optimized parameters, we utilized this model to predict the test dataset. The predictions were generated for the specific task at hand, which is a binary classification problem.

6. TESTING AND EVALUATION

Log loss learning curves for evaluating results were considered. Learning curves provide insight into the dependence of a learner's generalization performance on the training set size. This important tool can be used for model selection, to predict the effect of more training data, and to reduce the computational complexity of model training and hyperparameter tuning [2].

In the context of the training log loss curves, a distinct pattern was observed while using both models: the validation loss curve initially began at a value of 0.7 and experienced a decrease before it flattened around 0.3.

The training loss curve similarly commenced at the same value and underwent a decrease. However, the curve did not flatten and showed continued decrease of the log loss yet the validation log loss had plateaued. This was around iteration 200 showing the point at which overfitting started and therefore further training would be stopped.

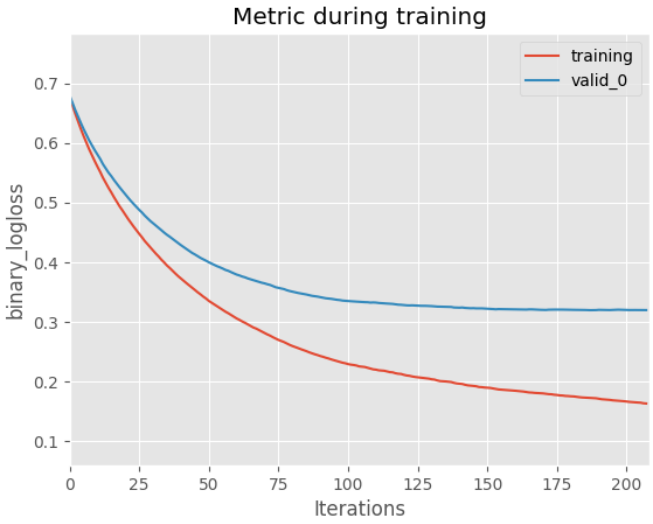


Fig. 5 – Training Loss curve with 10000 rounds and an early stopping of 20 rounds (80% Training, 20% Validation)

Accuracy, representing the ratio of correctly classified samples to the total samples, served as the primary evaluation metric for assessing the performance of the LightGBM model. The formula for accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

When assessing the performance of the LightGBM model on a validation set generated through an 80-20 split of the training dataset, it achieved an accuracy of 0.87 across different folds with low standard deviations of 0.0167. Subsequently, the same model

was utilized to train on the entire training dataset and make predictions on an unseen withheld dataset, resulting in an accuracy score of 0.913 on the public leaderboard and 0.918 on the private leaderboard, as detailed in Table 1.

Table. 1 – Performance of the model (80% Training, 20% Validation)

Metric	**LightGBM Accuracy**
Cross Validation	0.870
Public Leaderboard	0.913
Private Leaderboard	0.918

7. DISCUSSION AND FUTURE WORK

The successful development and deployment of the cropland mapping model mark significant progress in harnessing remote sensing and machine learning for agricultural monitoring. This section explores the achievements, potential areas for further improvement, and avenues for future research and development in the context of cropland mapping.

The cropland mapping model, primarily based on the LightGBM classifier, has demonstrated commendable performance in predicting and mapping cropland across diverse geographical regions. The project's key achievements include:

1. High Prediction Accuracy: The model's predictions have exhibited a high degree of accuracy, enabling the creation of detailed and reliable cropland maps. The accuracy metric, consistently high across validation and public/private leaderboard datasets, underscores the robustness of the model.

2. Efficiency and Scalability: LightGBM's efficiency in terms of memory and speed has ensured that the model is not only accurate but also scalable. This efficiency allows for the timely processing of large-scale remote sensing datasets.

3. Integration of Geospatial Data: The incorporation of geospatial data, such as latitude, longitude, and place information, has enriched the model's feature space, enhancing its ability to capture cropland patterns across different regions.

8. CONCLUSION

In conclusion, the cropland mapping solution developed in this project represents a significant step forward in the realm of agricultural monitoring and land-use classification. The culmination of remote sensing data, advanced machine learning techniques, and geospatial analysis has yielded a robust and accurate model for mapping cropland on a large scale.

REFERENCES

[1] Viering, Tom, and Marco Loog. "The shape of learning curves: a review." IEEE Transactions on Pattern Analysis and Machine Intelligence (2022). Available: https://shorturl.ac/7bny6

[2] International Telecommunication Union (ITU) official webpage: https://www.itu.int/