Machine Learning for Data Integrity: Validating Rural Property Identities in the Brazilian Amazon

Nelson Institute for Environmental Studies
UNIVERSITY OF WISCONSIN–MADISON

Cainã M Couto-Silva[1]; Matthew Christie[1], Lisa Rausch[1], Holly K Gibbs[1]

[1]**Nelson Institute for Environmental Studies, UW-Madison;** contact: coutodasilva@wisc.edu

Background

Brazil continues to have the world's highest rates of deforestation and most efforts aiming to address this target individual properties and producers. Over the last 15 years, we have focused on addressing deforestation in cattle supply chains by collecting and integrating land use, cattle sales, and environmental data about properties from multiple sources. This has proved to be challenging due to the variations in names and attributes the same rural property can have across different data sources, which led us to implement a matching algorithm that labels observations referring to the same property, assigning each group a unique property group identifier (PGI). However, given the large amount of data, there has been no easy way to validate the data integrity of PGIs to date.

To address this, we aimed to train a classifier using machine learning to predict whether a PGI is ill-formed or not (i.e., whether or not all properties labeled to the same PGI refer to the exact property unit), enabling us to validate the integrity of PGIs quickly and reliably.

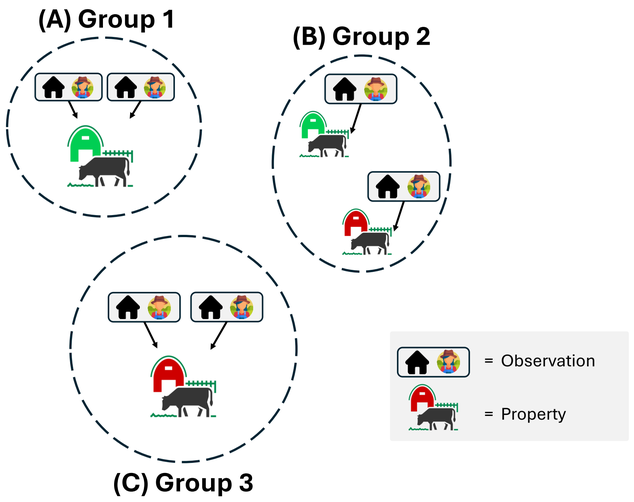

Figure 1. Illustration of PGIs generated by a matching algorithm (black dashed lines). Each PGI may contain one or more observations that ideally refer to the same property, even with slight name variations or different owners. (A) A well-formed PGI where both observations have the same owner and refer to the same property. (B) An ill-formed PGI with two owners and two properties. (C) Well-formed PGI with multiple owners.

Challenges

- **Data complexity:** Vast variability in rural property names and attributes across data sources.
- **Scale:** Managing and processing 15M+ records with over 3 million unique property identifiers.
- **Annotation:** Absence of pre-existing labeled data for model training.

Goal

To develop a scalable machine learning pipeline that:

- Maximizes detection of ill-formed PGIs (high recall)
- Maintains acceptable precision

Methods

Dataset

Data acquisition: In-premises data from multiple data sources with >15M records.

Data annotation: Based on domain knowledge and manual inspection, we annotated 130 PGIs using a random sampling approach stratified by group size.

Data Cleaning: Handled missing values and invalid records.

Feature engineering

Raw features: Variables based on property and owner attributes, like property names and geolocation data.

Derived features: Applied DBSCAN to provide cluster labels for property observations within each PGI. Observations with close property boundaries were clustered together under the same label.

Aggregated features: For each feature, we computed aggregate descriptive and custom statistics for all PGIs, including missing ratio, distinct category count, ratio of the most common category, gini impurity, and more.

Modeling

Baseline models: Evaluated three baseline models: a rule-based classifier based on exploratory data analysis, logistic regression, and decision tree.

Machine Learning models: Tested ensemble complex models including bagging (Random Forest) and boosting models (Catboost and LightGBM).

Validation: Model validation using a holdout (train-test-split) approach and cross-validation with 5 K-folds.

Results – Exploratory Data Analysis

We employed data visualization, Mann-Whitney U tests, Spearman correlation, and mutual information scores to assess distributions relative to target (ill-formed vs. well-formed PGIs).

Feature	Statistic	p-value
cluster_gini_impurity[1,2]	470	6.4821×10^{-18}
property_distinct_count[1,2]	237	9.3034×10^{-14}
shape_gini_impurity[1,2]	413	4.0281×10^{-13}
property_gini_impurity[1,2]	317	3.1172×10^{-12}
cluster_distinct_count[1,2]	378	3.2175×10^{-12}
distinct_count_SICAR[1,2]	345	4.9749×10^{-12}
property_most_common_ratio[1,2]	2793	8.7998×10^{-12}
shape_distinct_count[1,2]	411	9.8630×10^{-11}
total_count_SICAR[1]	435	3.0879×10^{-10}
same_property[1]	2384	1.8733×10^{-7}

[1] indicates features that are also in the top 10 Spearman correlation.
[2] indicates features that are in the top 10 mutual information scores.

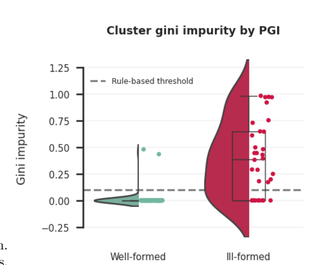

Figure 2. Statistical analysis. On the left, the table shows the Mann-Whitney (non-parametric version of the T-test for independent samples) statistic and p-values. Superscripted numbers indicate features with high correlation or mutual information to ill-formed groups. The right panel shows the distribution of cluster gini impurity (a derived feature). This feature was chosen based on the statistical results, and a threshold was manually selected to create a rule-based model.

Results – Modeling

We evaluated several machine learning models to predict ill-formed PGIs, from a rule-based approach (feature and threshold chosen after EDA) to more sophisticated algorithms. A Light GBM model was tuned to optimize recall (hyperparameters and probability threshold) with acceptable precision. The performance metrics for each model are present in the following table.

Model	Accuracy	Balanced Accuracy	Recall	Precision	F1-score	ROC-AUC
Rule-based	0.92 ± 0.06	0.84 ± 0.11	0.70 ± 0.22	0.93 ± 0.15	0.78 ± 0.16	0.84 ± 0.11
Logistic regression	0.83 ± 0.04	0.74 ± 0.09	0.58 ± 0.21	0.71 ± 0.18	0.62 ± 0.14	0.86 ± 0.12
Decision tree	0.86 ± 0.03	0.79 ± 0.08	0.65 ± 0.19	0.79 ± 0.15	0.68 ± 0.12	0.79 ± 0.08
Random Forest	0.88 ± 0.03	0.81 ± 0.07	0.68 ± 0.18	0.84 ± 0.17	0.72 ± 0.08	0.95 ± 0.03
Catboost	0.87 ± 0.03	0.82 ± 0.07	0.74 ± 0.23	0.80 ± 0.19	0.73 ± 0.07	0.96 ± 0.04
Light GBM	0.90 ± 0.03	0.86 ± 0.03	0.78 ± 0.10	0.85 ± 0.15	0.80 ± 0.03	0.90 ± 0.10
Light GBM (tuned)	0.83 ± 0.11	0.85 ± 0.09	0.90 ± 0.15	0.65 ± 0.15	0.74 ± 0.12	0.91 ± 0.08

Table 1. Model results for the cross-validation approach. The table shows the mean ± standard deviation for each model-metric. The highest recall is highlighted in red (corresponding precision = 0.65).

With the holdout approach, we achieved 1 and 0.67 as recall and precision, respectively. The following plot shows the probability estimates for the test set and the selected threshold for light GBM.

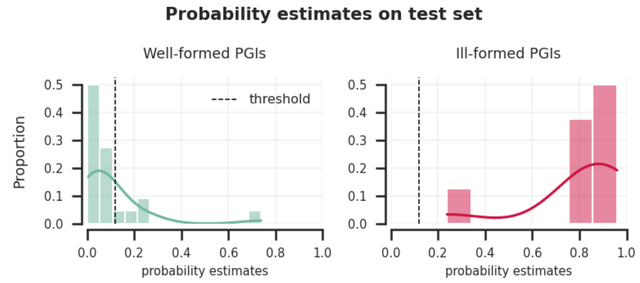

Figure 3. Probability estimates for the selected model. The black dashed line shows a conservative chosen threshold, aimed at avoiding false negatives.

We computed SHAP values to assess how each feature impacts the model. The following figure shows that features like multiple clusters and properties play an important role to the decision function.

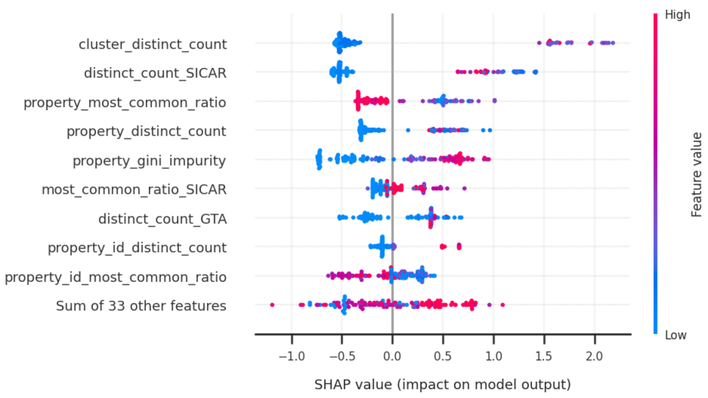

Figure 4. Feature importances. SHAP values on the x-axis show the impact on the model output (higher values convey a higher impact to the ill-formed PGI outcome). Every point is an instance of the data, and red points denote a higher value compared to the average value.

Conclusion

We successfully developed a model with high recall and acceptable precision to assess ill-formed groups at scale. Additionally, we built a web app to validate the results. This pipeline helps promote data integrity in efforts to end deforestation in Brazilian cattle supply chains.

Supported by

Bezos Earth Fund and the Gordon and Betty Moore Foundation.