Hackathon Group Name rEBooT rEBeLs

Group member Name **Anurag Sarva** and **Gulshan Shriram Hatzade**

Group member roll number **cs24mtech14003** and c**s24mtech14006** respectively

FoML Hackathon Report

Hackathon Challenge - Predicting Agricultural Field Efficiency

1. Introduction

The objective for this hackathon was to develop the predictive model to classify agricultural fields based on their efficiency levels. The target variable had 3 categories which are- high performing, moderately performing & low performing fields. Our aim was to identify the relative performance of fields considering factors like size, resource usage & the other field specific attributes rather than just yield output. This task of classification was evaluated using Macro F1-Score, which is measuring avg F1 Score across all 3 categories.

2. Feature Selection

For selecting features for this hackathon project, we focused on attributes that capture essential aspects of agricultural property & resource management, for improving the accuracy of prediction. Below is the breakdown of why each feature set was chosen-

1. **Crop & Field Characteristics**- Features like CropSpeciesVariety, CropFieldConfiguration & CultivatedAndWildArea provides specific information about types & configurations of crops on that land. These variables will help to differentiate between different types of crop setups & their varieties, which offers information of agricultural practices that can significantly influence yield, land value, & taxation.
2. **Structural Information**- Features such as FieldShadeCover, FieldConstructionType, FieldEstablishedYear & FieldZoneLevel explains geographical & structural aspects of that particular land. So, these can influence value of land, its usage efficiency. FieldSizeSqft adds essential measure of scale which is impacting land utility & potential of that agricultural.
3. **Environmental & Resource Management**- For capturing resource management practices & the environmental elements that are part of the land, features like HasPestControl, HasGreenHouse, NaturalLakePresence, MainIrrigationSystemCount, & NumberGreenHouses are important. These mentioned features not only affect agricultural productivity but also contribute for sustainable farming practices & indicates the potential for resource optimization.
4. **Water & Irrigation Resource-** The efficient water resources and irrigation systems directly impacts health of crop & yield, which makes these features highly relevant.

Water access are very important for agricultural productivity & features such as WaterAccessPoints, TypeOfIrrigationSystem, PartialIrrigationSystemCount, WaterReservoirCount, ReservoirWithFilter, & PerimeterGuardPlantsArea cover diverse aspects of management of water.

5. **Soil & Fertility**- SoilFertilityType indicates the quality of soil, which is an essential factor for crop success & also overall farm productivity. This feature is providing the insight into types of crops that can be grown & potential yields.
6. **Land Valuation & Taxation**- Valuation metrics for cultivated area, reservoir capacity and assessed taxes is covered by including features for training -TotalCultivatedAreaSqft, TotalReservoirSize & TotalTaxAssessed. These features reflects economic standing and valuation of the land, which are very important for financial assessment and allocation of resource.

We chose these features for their relevance to capturing agricultural infrastructure, crop management, environmental factors & financial assessment metrics. All these are integral to provide the robust basis for predictive analysis in this hackathon task.

3. Approach

3.1. Data Preprocessing

- **Feature Selection**- From original dataset, we chose the set of features that were deemed most relevant based on our knowledge.
- **Missing Value Handling**- We dropped feature with more than 70% missing values for ensuring the cleaner dataset. For the remaining missing values, we used median imputation for handling numerical columns.
- **Feature Engineering**- We created new features like Area_per_Equipment (calculated as PrimaryCropAreaSqft divided by FarmEquipmentArea) & Area_Sqrt (square root of PrimaryCropAreaSqft), where data was available. These features did non linear relationships & field efficiency ratios.

3.2. Data Transformation

- We used RobustScaler for normalizing data which is making the model less sensitive to outliers.
- We implemented the Pipeline that combined data imputation & scaling for the streamlined preprocessing flow.

3.3. Model Selection and Hyperparameter Tuning

- We selected Random Forest Classifier for its robustness & its ability of handling tabular data effectively, particularly with missing values and categorical features.
- Here we addressed class imbalance by setting class_weight='balanced'.
- We employed RandomizedSearchCV by hyperparameter tuning for optimizing parameters like- No. of trees, Max depth & Min samples.

3.4. Model Evaluation

- We evaluated model using Macro F1 Score on the validation set.
- Achieved our best Validation Macro F1 Score of 0.441, indicating the balanced performance across all 3 categories.

4.Observations

1. During preprocessing, several features had over 70% missing values. We dropped these features were dropped to ensure a cleaner dataset and improving model efficiency.
2. The features which we chose covered essential aspects of crop management, soil fertility, irrigation, field configuration & water access. These were expected to provide the comprehensive view of factors influencing crop outcomes, allowing the model to capture significant relationships in the data.
3. Feature Engineering Impact- Creating derived features like Area_per_Equipment & Area_Sqrt added new dimensions to the data. These features captured interactions between field area, productivity, equipment offering the model additional insights into resource allocation & management of field.
4. The target variable was imbalanced across categories low, medium, and high, so we applied the balanced class weighting in our RandomForestClassifier. This helped us to address the imbalance and improved the model's predictive accuracy across all the classes.
5. Randomized hyperparameter tuning led to an optimized Random Forest model. Adjusting parameters like n_estimators, max_depth & min_samples_split helped our model for achieving strong validation F1 macro score, ensuring it performs well across all target categories rather than favoring any single one.
6. Imputation with median values & scaling using RobustScaler helped to standardize the data, especially in the presence of outliers. This preprocessing improved robustness of model which made it less sensitive to extreme values and enhanced its ability to generalize across various data points.
7. Our final model showed comparatively better performance, as reflected in the F1 macro score. This suggests selected features & tuned parameters effectively captured complexities of this dataset compared to our earlier models.

The model has been tailored to predict crop outcomes in varying conditions, by focusing on features relevant to agricultural field & water resource management by optimizing hyperparameters. The final achieved validation F1 score demonstrates the effectiveness of model in handling categorical variations with imbalanced data.

7. Contribution

1. **CS24MTECH14003 Aurag Sarva-** Handled data preprocessing, including treatment of missing value treatment, feature engineering & pipeline creation. Contributed to feature selection process & development of new features like Area_per_Equipment & Area_Sqrt.
2. **CS24MTECH14006 Gulshan Hatzade-** Focused on model training, hyperparameter tuning & validation. Contributed feature selection process & configured a RandomizedSearchCV parameters & also explored tuning strategies to maximize F1 performance. Additionally, managed test set preparation, predictions & submission formatting.