FOML-AI2000 - REPORT
Hackathon on Estimating Agrarian Land Efficiency
Team: Kanya Raasi

Report by: Challa vishweshwar Reddy -ES22BTECH11007
Team mare: BADAVATH SRIKANTH-ES22BTECH11004

Key points:

- Model : XGBoost Algorithm with optimized hyperparamters.
- Peak F1-score observed: 0.412

Preprocessing:

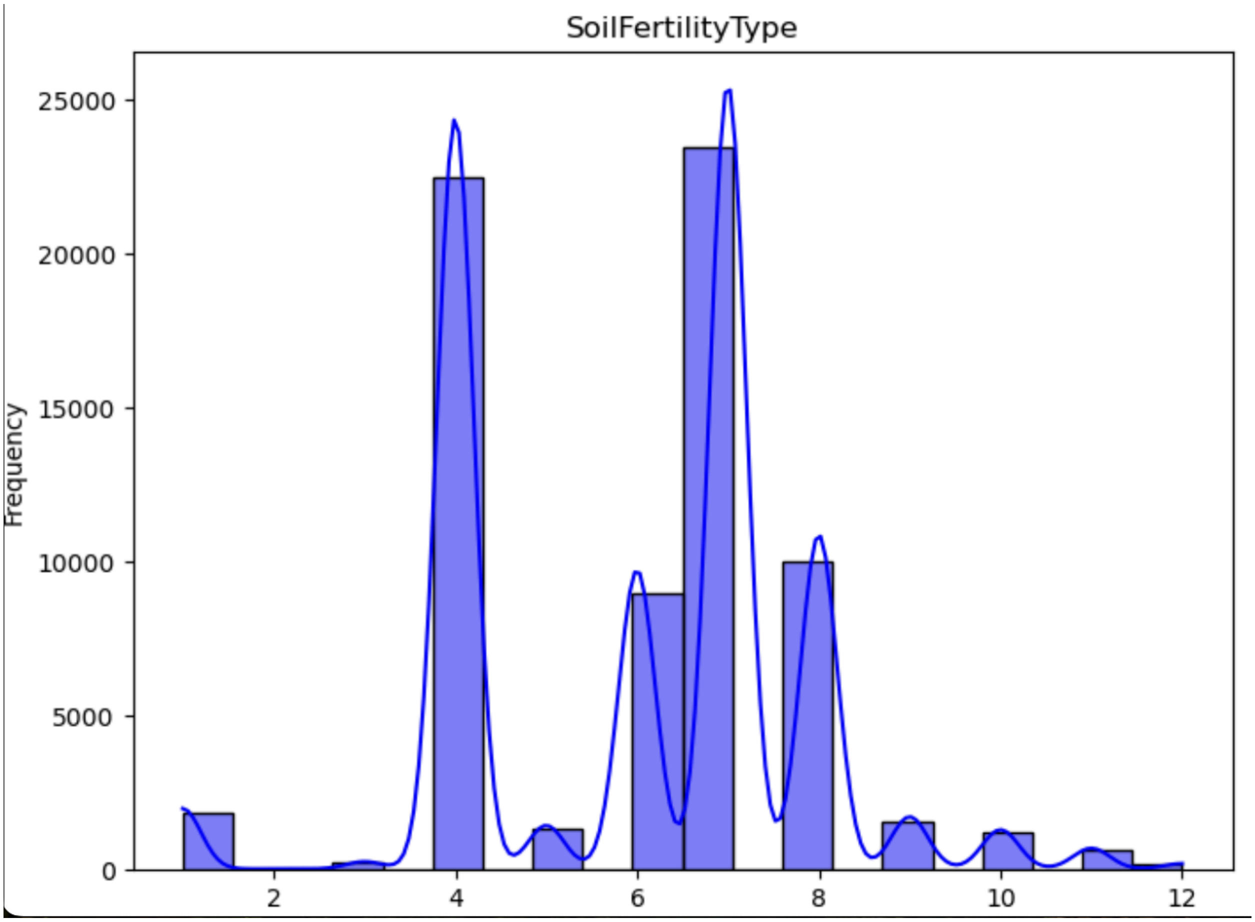
Deleting columns based on missing values:we first tried to delete the column with high missing values(>90).but it was much affecting after checking with cross validation.so we havent removed anything.

Deleting columns based on variance:we first tried to delete the column with low variance(0.01).but it was much affecting after checking with cross validation.so we havent removed anything.

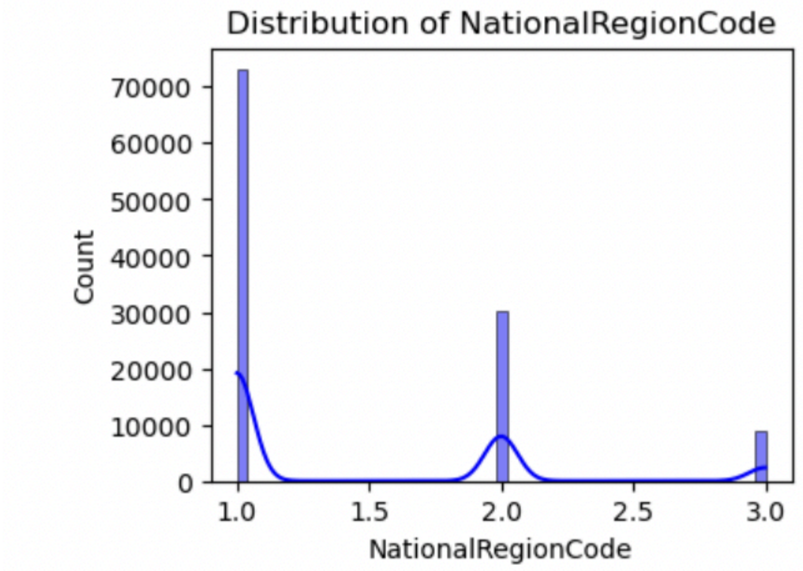
1.Identifying the catoforical variables:

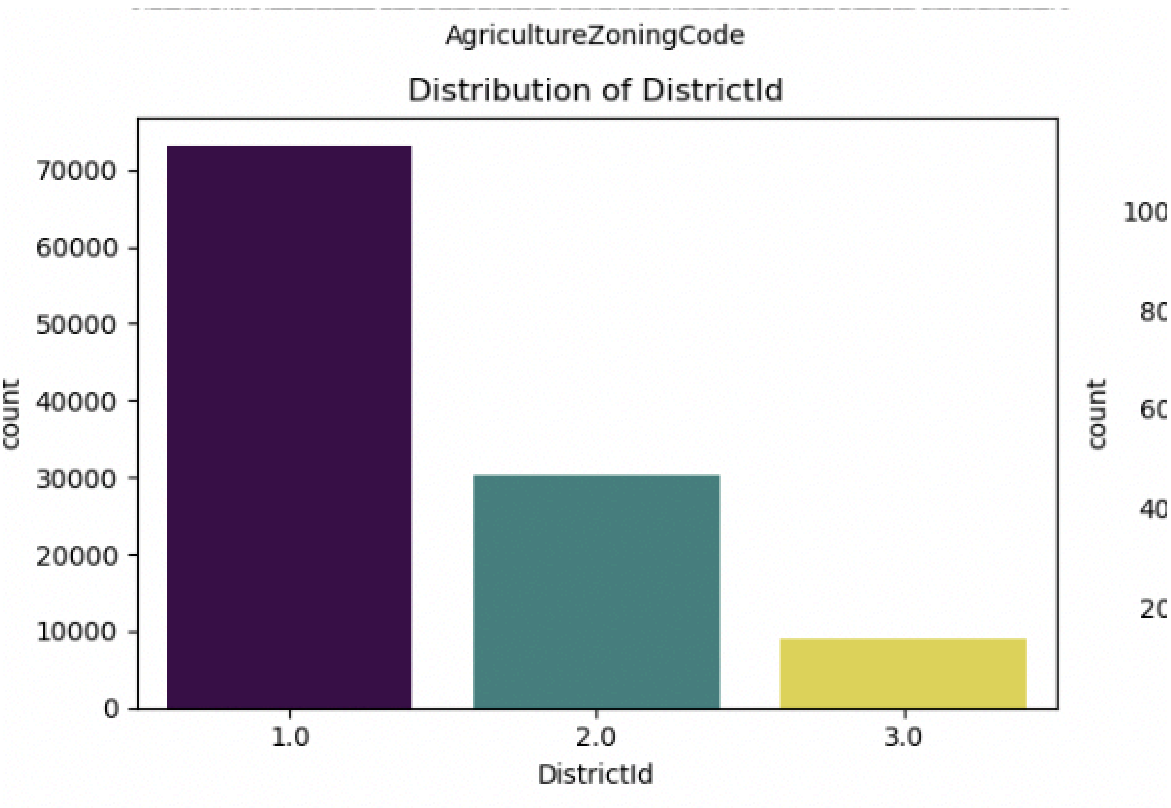
Eventhough the dats has all the data types as int or float there are many categorical variables.
We created graphs for some of the guessed columns.
Identifying categorical variables are important because we have to use mode for them while filling the missing values and use differnt encoding techniques.
We are concluding columns as categorical if we the column has only small number of categories.

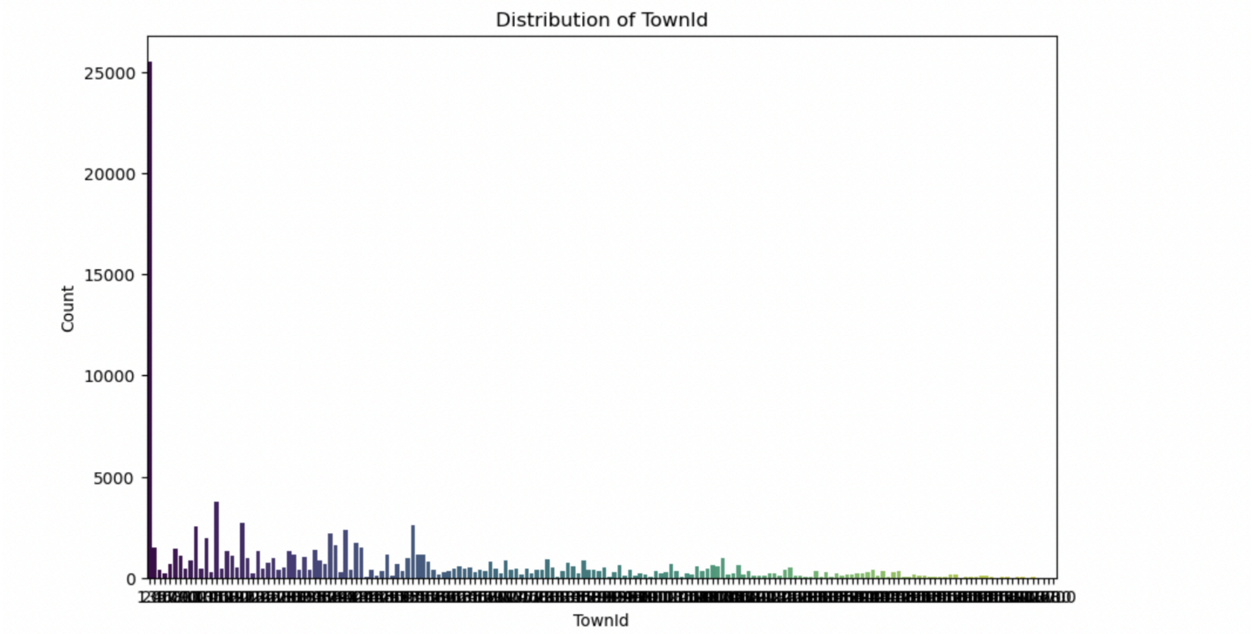
Here are some categorical variables i found.



We tried conclude area based columns are categorical or numerical.

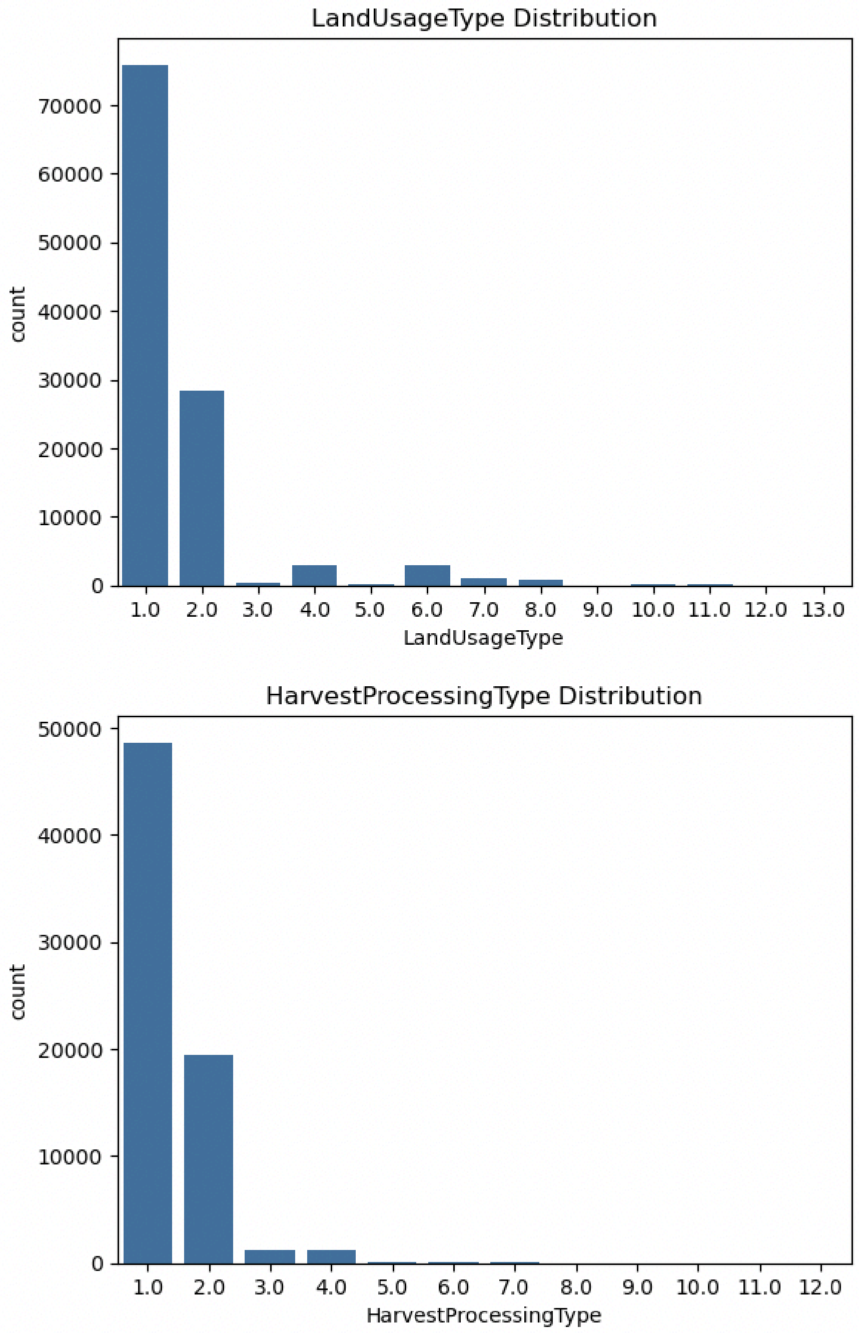






We concluded that district,national region code are categorical.and town id is numerical.(even though it is categorical it has high number of missing values)

We concluded following also as categorical.



2.We tried to create the following additional features we assumed may help increasing f1-score.what we observed is f1-score.the f1 score increased from 0.39 to 0.4

Water-Related Features

These features describe the field’s water availability relative to its area.

- Water Density: Measures the concentration of water access points per square foot of the entire field.
  - **Formula**: Water Density = Number of Water Access Points / Total Field Size (sq ft)
- Water per Cultivated Area: Calculates the availability of water access points specifically in the cultivated area.
  - **Formula**: Water per Cultivated Area = Number of Water Access Points / Total Cultivated Area (sq ft)
- Reservoir Density: The density of water reservoirs relative to the entire field's area.
  - **Formula**: Reservoir Density = Number of Water Reservoirs / Total Field Size (sq ft)

Cultivation Efficiency Metrics

These features indicate how efficiently the field area is used for cultivation.

- **Cultivation Ratio**: Shows the proportion of the field size that is actively cultivated.
  - **Formula**: Cultivation Ratio = Total Cultivated Area (sq ft) / Total Field Size (sq ft)
- **Greenhouse Density**: The number of greenhouses per square foot of the field.
  - **Formula**: Greenhouse Density = Number of Greenhouses / Total Field Size (sq ft)
- **Farming Intensity**: Measures the concentration of farming units relative to the cultivated area.
  - **Formula**: Farming Intensity = Number of Farming Units / Total Cultivated Area (sq ft)

Infrastructure Utilization

These features measure how well infrastructure resources are utilized within the field.

- **Storage per Area**: Calculates the combined storage area (both underground and harvest storage) per square foot of field area.
  - **Formula**: Storage per Area = (Underground Storage Area + Harvest Storage Area) / Total Field Size (sq ft)

- **Equipment Ratio**: Represents the proportion of the field occupied by farming equipment.
  - **Formula**: Equipment Ratio = Farm Equipment Area / Total Field Size (sq ft)

Economic Indicators

These features give insights into the economic aspects of the field.

- **Value per Square Foot**: Measures the total value of the field per square foot, indicating its worth.
  - **Formula**: Value per Square Foot = Total Value of Field / Total Field Size (sq ft)
- **Tax Burden**: Represents the tax liability relative to the field's total value, reflecting the tax load.
  - **Formula**: Tax Burden = Total Tax Assessed / Total Value of Field

Operational Scale Indicators

These features indicate the extent of mechanization and irrigation in the field.

- **Vehicle per Area**: Shows the density of farm vehicles per square foot of cultivated area.
  - **Formula**: Vehicle per Area = Number of Farm Vehicles / Total Cultivated Area (sq ft)
- **Irrigation Coverage**: Measures the density of irrigation systems relative to the cultivated area.
  - **Formula**: Irrigation Coverage = Number of Irrigation Systems / Total Cultivated Area (sq ft)

Field Age and Development

This feature calculates the field's age, indicating its maturity.

- **Field Age**: Represents the age of the field by calculating the difference between the current year and the year it was established.
  - **Formula**: Field Age = Current Year - Year Field was Established

Location-Based Features

These features represent the geographic location of the field.

- **Location Cluster**: A unique identifier for each field based on its latitude and longitude coordinates, which can be used to distinguish different field locations.

- **Formula**: Location Cluster = Latitude + '_' + Longitude (concatenated as a string)

3.scaling

We used StandardScaler to standardize these numeric columns, transforming them to have a mean of 0 and a standard deviation of 1, and updates data with the scaled values.we found some increase in f1 score from .36 to 0.38

4. Outlier detection

We used outlier detection for every column .

Calculating IQR: For each column, the first quartile (Q1) and third quartile (Q3) are calculated, and the interquartile range is found (IQR = Q3 - Q1).

Setting Bounds: Defines a lower and upper bound as Q1 - 1.5 \times \text{IQR} and Q3 + 1.5 \times \text{IQR} , respectively.

Clipping Outliers: Uses .clip(lower_bound, upper_bound) to limit values outside this range to the calculated bounds, effectively reducing extreme outliers to the nearest boundary values.

Model selection:

We experimented with different models such as random forest,gradient boosting with different hyperparametrs.both of them given f1 score aroun 0.37

Then we experimented with XGboost.it gave best f1 score till now so we selected XG boost.

Then we best hyperparameters with grid search cv.finall we have our peak score as 0.42.

We even tried building a model with small 80 percent of data and testing on 20% and selecting the model with best accuracy.it hasnt improved much so we dropped that idea