CS 6240 Final Project

Darshan Patel

Sriramprabhu Sankaraguru

Content

1. Data Engineering
2. Model Training
3. Prediction
4. Sample run results
5. What's Next
6. Questions?

1. Data Engineering

How to clean the data and give it as input for Machine Learning library

- Remove less important columns
- One hot encoding – for logistic regression
- Convert the values of certain columns
- Handle missing values in different columns (Birds, Distance etc.)
- Convert each column to float and target column to binary

Removing Columns

- Columns that are not going to help in prediction

SAMPLING_EVENT_ID - <Reason>

LOC_ID, DAY

COUNTRY

STATE_PROVINCE

COUNTY

OBSERVER_ID

GROUP_ID

BAILEY_ECOREGION

OMERNIK_L3_ECOREGION

SUBNATIONAL2_CODE

- Filter rows by PRIMARY_CHECKLIST_FLAG

One-Hot Encoding

Converting columns with categorical feature to work better with Logistic Regression

- COUNT_TYPE: Categorical column with 20 different values encoded as 20 different columns.
- TIME: Split into four columns, each represents one 6 hour slot.

Convert Values

Convert the values and merge certain columns together so Machine Learning library learns better.

- YEAR: Converted into odd/even
- LONGITUDE, LATITUDE: Converted into xyz plane
- ELEV_GT, ELEV_NED: Dropped and replaced with average of the two
- CAUS_*, CAUS_*_MM: Dropped around 60 columns by replacing CAUS_* values with the value of particular month (MONTH) from CAUS_*_MM
- NLCD_* - Replace them with corresponding year rather than having all the columns

Handle Missing Values

Birds

- ? And other values are replaced with 0
- x replaced with rand(2, 10)

Others

- ? Replaced with 0

Normalize the values

Correlation/Sampling

2. Model Training

3. Prediction

4. Sample results

5. What’s Next

6. Questions?

Thank You!