Changelog - July 19th, 2023

Dane:
- Weather data and soil moisture aggregation
  - Added weeks and months
  - Data exports to CSV - too many columns
  - Called from createSet modules as needed
- Datasets and modules to make easy modifications
- Decision tree visualization
- Feature reduction
- RandomForests, SVM, ANNs testing and evaluation
- Documentation

Daniel:
- Continue working on the machine learning models and implementing the autoencoder model (on hold)
- Working on create different datasets for different problem statements/predictions
- Experimenting the ML models on the newly created datasets
- Researched on dimensionality reduction

Dharmit:
- Worked on ML models(KNN, SVM)
- Experimented models on created dataset
- Testing and Hypertuning the models

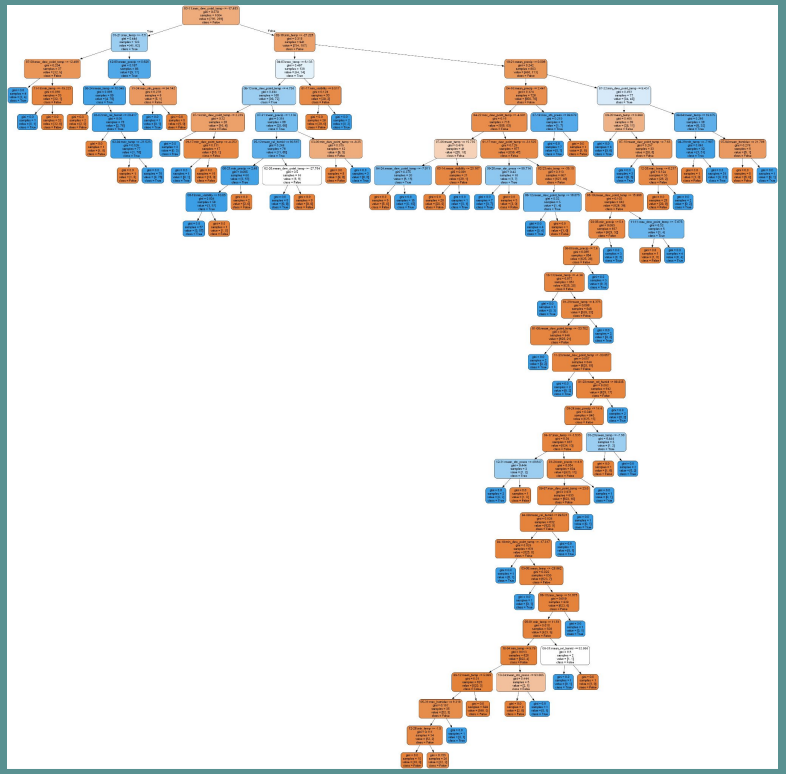
Joseff:
- Created new Ergot table with more attribute to predict: downgrade, quantile bin, arbitrary bin
- Combined daily weather station datasets into one table for easier subsetting
- Created new Ergot previous year data table using severity instead of incidences and using 0.04 threshold
- Researched ML techniques that deal with imbalance data sets
- Contemplated the implications of our data attributes, what we're predicting and what we're using to predict

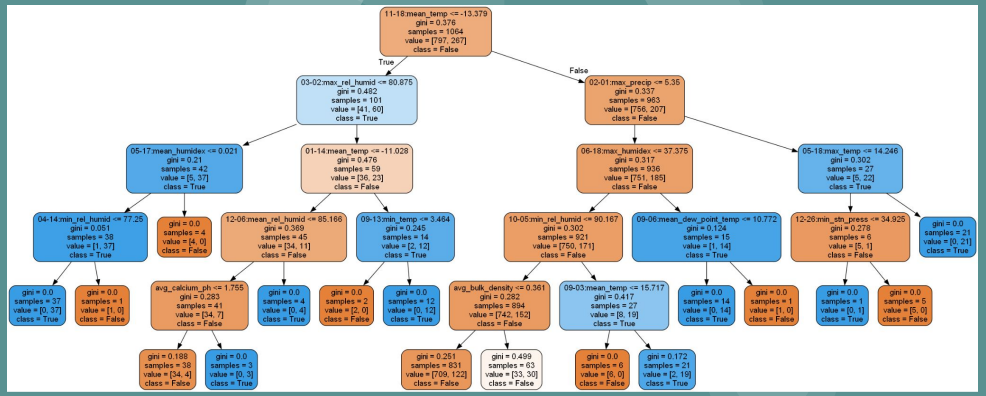
Understanding the Data

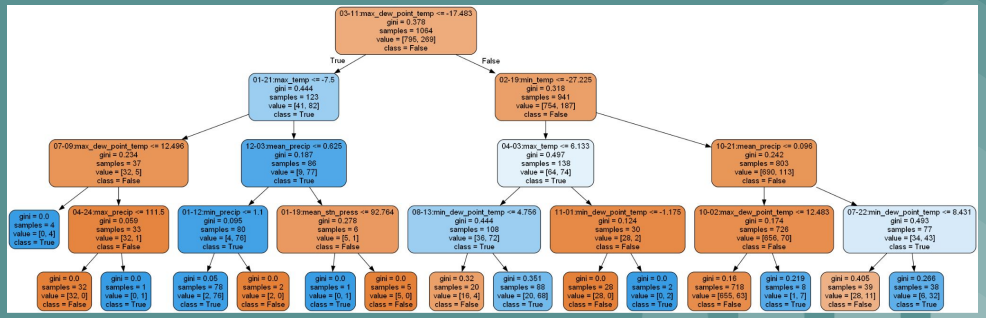
Which features are most useful?

- Inspired the extensive dataset coverage

[Module]







Curse of Dimensionality

PCA:

- Focuses on variance as opposed to importance
- Not necessarily better
- Creates entirely new attributes...

Gaussian Projection

- Random projection that focuses on distances between samples
- Also creates entirely new attributes...

Feature importance - as per this book:

- Similar to what I was already finding as per **model.feature_importances_**

```
[SUCCESS] reduced data in dataset: Data aggregated by
month [mean]|[minMax]|[straified on has_ergot]
(680, 24)
2:min temp 2:max temp 2:max dew point temp 3:max_precip
3:mean precip 4:max temp 4:mean temp 5:mean temp
7:mean precip 7:min rel humid 7:mean rel humid
9:max temp 9:mean temp 9:max humidex
9:mean humidex 12:mean dew point temp 12:max precip
12:mean precip 6:soil moisture max 7:soil moisture min
8:soil moisture min 8:soil moisture max
8:soil moisture mean 9:soil moisture max
```

```
Most relevant features = ['38:max temp', '03-10:max temp',
'9:max humidex', '41:mean temp', '9:mean humidex',
'9:max temp', '9:mean temp', '10-10:max temp', etc…
```

```
[SUCCESS] reduced data in dataset: First 15 years aggregated by
week [median]|[minMax]|[straified on has_ergot]
(364, 22)
8:mean temp 10:min temp 10:max temp 28:max temp 28:max humidex
28:mean humidex 38:max temp 38:mean temp 38:max humidex
38:mean humidex 41:min temp 41:max temp 41:mean temp
41:max dew point temp 41:mean dew point temp
43:min dew point temp 43:mean dew point temp 50:min temp
51:max temp 51:max dew point temp 33:soil moisture max
38:soil moisture max
```

Summer attributes appeared less than expected.
General trend seemed to favor fall/winter attributes

Most consistent were 38:max_temp, fall/winter soil moistures and dew point temperatures

Model Evaluation

Avg_accuracy: accuracy of stratified k fold cross validation using test data set (Perfect = 100)

- Creates balanced sets as per a selected attribute

R2: approximately how much of the observed variation can be explained by the model's inputs? (Perfect = 1)

Loss: summation of errors in our model (Perfect = 0)

Precision: the ability to classify positive samples in the model (Perfect = 1)

Recall: how many positive samples were correctly classified by the model (Perfect = 1)

F1: harmonic mean of precision and recall (Perfect = 1)

Auc: the ability to distinguish between all the Positive and the Negative class points (Perfect = 1)

neg_mean_squared_error: Mean squared logarithmic summation of errors in our model (Perfect = 0)

10 most relevant attributes

Datasets

Predictors:

Ergot_present_in_q3 (classification)
Ergot_present_in_q4 (classification)
Sum_severity_in_q3 (classification)
Sum_severity_in_q4 (classification)
Percnt_true (regression)
Sum_severity (regression)

[dataset1] - Exploratory set (seasons, worst years, first 15 years, all data, different data preprocessing)

[dataset2] - Best from dataset1 plus all data aggregated combinations (day, week, month)

- First 15 years aggregated by week [median]|[minMax]|[straified on has_ergot]

[dataset3] - Results from Feature Reduction on dataset2

[dataset4] - Best results from dataset3

- First 15 years aggregated by week [reduced]|[median]|[minMax]|[straified on has_ergot]
  - Years not included used for testing
  - Scores decreased significantly

[dataset5] - Second best results from dataset3

- Moisture data from years with bad ergot aggregated by month [mean]|[minMax]|[straified on has_ergot]
  - Years not included used for testing
  - Scores decreased but overall still consistent scoring

Notable Results

Predictors for ergot presence did not apply well to sum_severity

Using engineered Ergot Features tended to increase performance
- Reduces bias?

Random forests best hyperparameters:
- n_estimators = 200/500, max_depth = 10/15

SVM best hyperparameters:
- gamma = 0.5, c = 0.75/1, degree = 3

Best results:

[dataset1] All data aggregated by month [mean]|[minMax]|[straified on has_ergot][Ergot_present_in_q4][RandomForests] [n_estimators=200][max_depth=15]
avg_accuracy: 0.8782, r2: 0.9249, loss: 2.7075, precision: 1.0, recall: 1.0, f1: 0.84, auc: 0.8763

[dataset1] Moisture data from years with bad ergot aggregated by month [mean]|[minMax]|[straified on has_ergot] [Ergot_present_in_q4][RandomForests][n_estimators=100][max_depth=5]
avg_accuracy: 0.8720, r2: 0.92, loss: 2.8835, precision: 0.9884, recall: 0.8673, f1: 0.7500, auc: 0.8263

[dataset5] Moisture data from years with bad ergot aggregated by month [mean]|[minMax]|[straified on has_ergot] [Ergot_present_in_q4][RandomForests][n_estimators=100][max_depth=5]
avg_accuracy = 0.7637, r2 = 0.8938, loss = 3.8288, precision = 0.9636, recall = 0.7681, f1 = 0.7327, auc = 0.8011

Dataset creation

Dataset V1: (1064 x 8)

- It contains only has_ergot as an output from ergot table and all weather attributes as inputs from weather table
- Allows to test on multiple months/seasons
- **Problem statement**: Given a district and its weather attributes -> predict if the district is gonna have ergot or not.

Dataset V2: (1026 x 26)

- It contains only has_ergot as an output from ergot table and all weather attributes, soil moistures, soil data as inputs from weather, soil moisture, soil data table
- Allows to test on multiple months/seasons
- **Problem statement**: Given a district and its weather, soil moisture, soil data attributes -> predict if the district is gonna have ergot or not.

Dataset V3: (154048 x 27)

- It contains only incidence as an output from ergot table and all weather attributes as inputs from weather table
- Allows to test on multiple months/seasons
- **Problem statement**: Given an ergot sample and its attributes (ergot, weather, soil moisture, soil data) -> predict if the given sample is gonna have ergot or not.

Dataset V4: (154048 x 56)

- It created a sellable (severity > 0.4) column as an output all weather, ergot, soil moisture, soil data attributes as inputs from ergot, weather, soil moisture, soil data table
- Allows to test on multiple months/seasons
- **Problem statement**: Given an ergot sample and its attributes (ergot, weather, soil moisture, soil data) -> predict if the given sample can be sold or not based on the severity (in particular, the created column sellable).

Recorded results - MLP

Dataset v1: 500 epoches

- [48, 32, 24, 16, 8, 1]
  - Accuracy: 0.6917
  - Precision: 0.8009
  - Recall: 0.8232
  - F1-Score: 0.8119
  - AUC Score: 0.5830
- [24, 24, 24, 24, 24, 1]
  - Accuracy: 0.6954
  - Precision: 0.8130
  - Recall: 0.8093
  - F1-Score: 0.8111
  - AUC Score: 0.5988
- [32, 16, 32, 16, 8, 1]
  - Accuracy: 0.7443
  - Precision: 0.8075
  - Recall: 0.8976
  - F1-Score: 0.8502
  - AUC Score: 0.5932

Dataset v2: 200 epoches

- [48, 32, 24, 16, 8, 1]
  - Accuracy: 0.7568
  - Precision: 0.8457
  - Recall: 0.8490
  - F1 Score: 0.8473
  - AUC Score: 0.6692
- [24, 24, 24, 24, 24, 1]
  - Accuracy: 0.7743
  - Precision: 0.8287
  - Recall: 0.8950
  - F1 Score: 0.8605
  - AUC Score: 0.6216
- [32, 16, 32, 16, 8, 1]
  - Accuracy: 0.8054
  - Precision: 0.8472
  - Recall: 0.9150
  - F1 Score: 0.8798
  - AUC Score: 0.6567

Notable results - ML

Dataset v1: StratifiedKFold

- Logistic Regression
  - Accuracy: 0.789
  - F1-Score: 0.88
  - AUC Score: 0.49
- Random Forest
  - Accuracy: 0.82
  - F1-Score: 0.9
  - AUC Score: 0.52
- Decision Tree
  - Accuracy: 0.67
  - F1-Score: 0.78
  - AUC Score: 0.60
- Gradient Boosting
  - Accuracy: 0.79
  - F1-Score: 0.87
  - AUC Score: 0.68

Dataset v2: StratifiedKFold

- Logistic Regression
  - Accuracy: 0.82
  - F1-Score: 0.9
  - AUC Score: 0.52
- Random Forest
  - Accuracy: 0.83
  - F1-Score: 0.91
  - AUC Score: 0.56
- Decision Tree
  - Accuracy: 0.77
  - F1-Score: 0.86
  - AUC Score: 0.56
- Gradient Boosting
  - Accuracy: 0.81
  - F1-Score: 0.89
  - AUC Score: 0.59

Dataset v3: StratifiedKFold

- Logistic Regression
  - Accuracy: 0.76
  - F1-Score: ~
  - AUC Score: 0.5
- Random Forest
  - Accuracy: 0.81
  - F1-Score: 0.32
  - AUC Score: 0.59
- Decision Tree
  - Accuracy: 0.83
  - F1-Score: 0.72
  - AUC Score: 0.86
- Gradient Boosting
  - Accuracy: 0.94
  - F1-Score: 0.89
  - AUC Score: 0.94

Notable results - ML

Dataset v1:

(KNN)

K value = 27

- Accuracy: 0.6525
- Precision: 0.9186
- Recall: 0.6384
- F1 Score: 0.7533
- AUC Score: 0.6803

(SVM)

Kernal = poly, C = 1

- Accuracy: 0.7887
- Precision: 0.8882
- Recall: 0.8531
- F1 Score: 0.8703
- AUC Score: 0.6626

Dataset v2:

(KNN)

K value = 23

- Accuracy: 0.771
- Precision: 0.9160
- Recall: 0.7692
- F1 Score: 0.8362
- AUC Score: 0.7746

(SVM)

Kernal = rbf, C = 10

- Accuracy: 0.8203
- Precision: 0.8888
- Recall: 0.8717
- F1 Score: 0.8802
- AUC Score: 0.7658

Dataset v3:

(KNN)

K value = 54

- Accuracy: 0.6290
- Precision: 0.8640
- Recall: 0.6319
- F1 Score: 0.7299
- AUC Score: 0.6249

(SVM)

Kernal = linear, C = 10

- Accuracy: 0.7176
- Precision: 1.0
- Recall: 0.6441
- F1 Score: 0.7835
- AUC Score: 0.8220

Caveats for current models

- Some use same year ergot data as predictors e.g. neighbor has ergot
- Some are using weather after the wheat is harvested as predictors
- Interpolation models
- Has_ergot

Next model iterations

Ergot sample feat eng
- Downgrade 0.04
- Quantile Bins
- Arbitrary 0.02 0.04 0.08
- Severity
- Incidence

Agg Ergot sample v2
- Severity based calculations for previous year data

Temporal data
- Ensure that the temporal data columns we are using for prediction are from before the wheat harvest date

Goals for next 2 weeks

Dane:
- Improve documentation
- Improve data pipeline
- Further model experimentation
    - Focus data test/train split on years instead of stratification (as per Josef)
    - Ensemble split (as per Josef)
    - Aggregate on growing seasons
    - Remove same year ergot features/Copernicus
    - Finish up ANN

Daniel:
- Start documenting and code cleanup
- Clean up model experimentation to focus on the problem statement
- Fix/Added more ergot visualization

Dharmit:
- Some Further Research on Models
- Improve Models
- Start Documentation for the Project

Jay:
- Improve model
- Documentation on model
- More dataset preparation for testing model

Joseff:
- Any requests from stakeholders?
- Multi model ml model
- Ensemble strategy
- XGBoost