Crop Variety Text Classification

PURDUE UNIVERSITY | The Data Mine

Lei Gong, Advait "Adi" Budaraju, Aneesh Vemuri, Saket Chamarti, Aviral "Avi" Rastogi, Thelma Yeboah, Arjun Chidambaram, Kyanna Fletcher

JOHN DEERE

Introduction

ABOUT

- John Deere is the world's leading manufacturer of agricultural machinery, turf equipment, forestry machinery, and construction machinery.

PROBLEM

- The Agricultural cycle is the annual cycle that involves the growth and harvest of a crop. During this cycle, the seed specifier is keyed in as a free-form text. This creates potential human error in the format and unique serial identifier.

MOTIVATION

- We want to find agronomic patterns to provide valuable insights and product automation systems for Customers (change)

GOAL

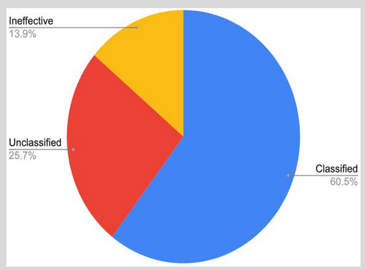
- The goal is to consolidate crop varieties. We want to reduce the amount of unique crop entries by identifying all unique entries of one crop variety and combining it into one entry

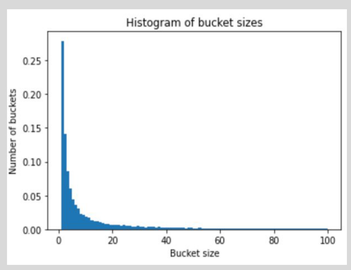
Unique crop Entry
P1108Q
1108Q
Pio 1108q
p1108q

Consolidated to → **P1108Q**

Categorization into Buckets by ID

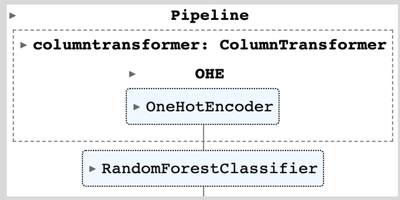
- Each crop variety is identified by an ID – for example, P1108Q can be represented by the ID 1108
- Our goal was to assign an ID to each entry in the dataset, grouping the data "buckets" by ID
- We used extensive string parsing to achieve this, with a lot of fine-tuning
- In this way, we managed to classify most of the usable data
- However, a lot of our buckets contained very few elements – so our next step was merging categories together if they contained the same information





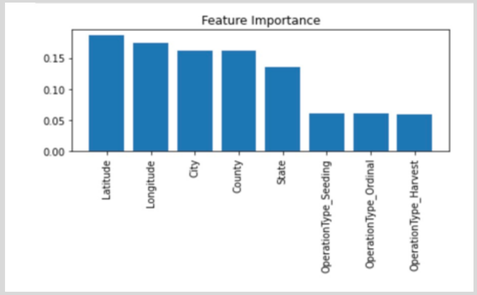
Bucket Merging

- "Merging" two buckets means consolidating all data elements from both the buckets into one larger, combined bucket
- For example, the IDs 1108 and 11-08 probably both correspond to Pioneer 1180Q, so their corresponding buckets should be merged
- Using various ways of string parsing, we were able to merge and thus trim down the amount of buckets
- Then, we used an ML model to predict bucket similarity to infer which other buckets to combine
- Overall, we were able to trim down the number of buckets from 40000 to 30000



Location-based classification & clustering

- Given the crop has one or several centroids by its property –Indiana is a really good location for crop to grow.
- The purpose is to using the current categorical data, including location, country, OperationType, ProductBrand, to predict the crop name.
- Geological information, including longitude, latitude(most applicable one), state, county, and city,
- Classification and cluster algorithms using geological information, which shows a strong pattern for different types of crop.
- 5 Encoding methods converting categorical data to numerical data.



Conclusions

- Classification by ID helped us efficiently clean up and organize the data
- Not all data has been classified and merged fully, but we've been able to apply this approach to a large majority of the dataset
- For the location-based approach, One-Hot Encoding does not work well dataset with too many categories, using Mean-Encoding and other encoding will increase the accuracy in machine learning model
- Using function in h3 package is a good way to find similarity between two rows based on geological information.
- Random Forest and KNN machine models could give good accuracy in predicting crop name using ProductBrand and location information.

Future Goals

- Further classification of data for easier analysis
- Recognition of performance patterns between different varieties using crop trials or other John Deere data.
- Constructing the function finding the similarity based on each row will be important to find the general patterns.
- Working with the John Deere machine team to make data entry more standardized

Acknowledgments

- We would sincerely like to thank our mentors from John Deere: Paul Readel, Dylan Roth, and Devin Becktell for their continuous support and encouragement.
- We also would like to extend our gratitude to the TAs Soumya, Praneeth and Victoria for coordinating the teams throughout the project.
- Lastly, we would like to thank the Data Mine for giving us the opportunity to work on this project.