Capstone Project - Plant Recommendation Engine

8/18/2021

Joseph Hicks

Table of Contents

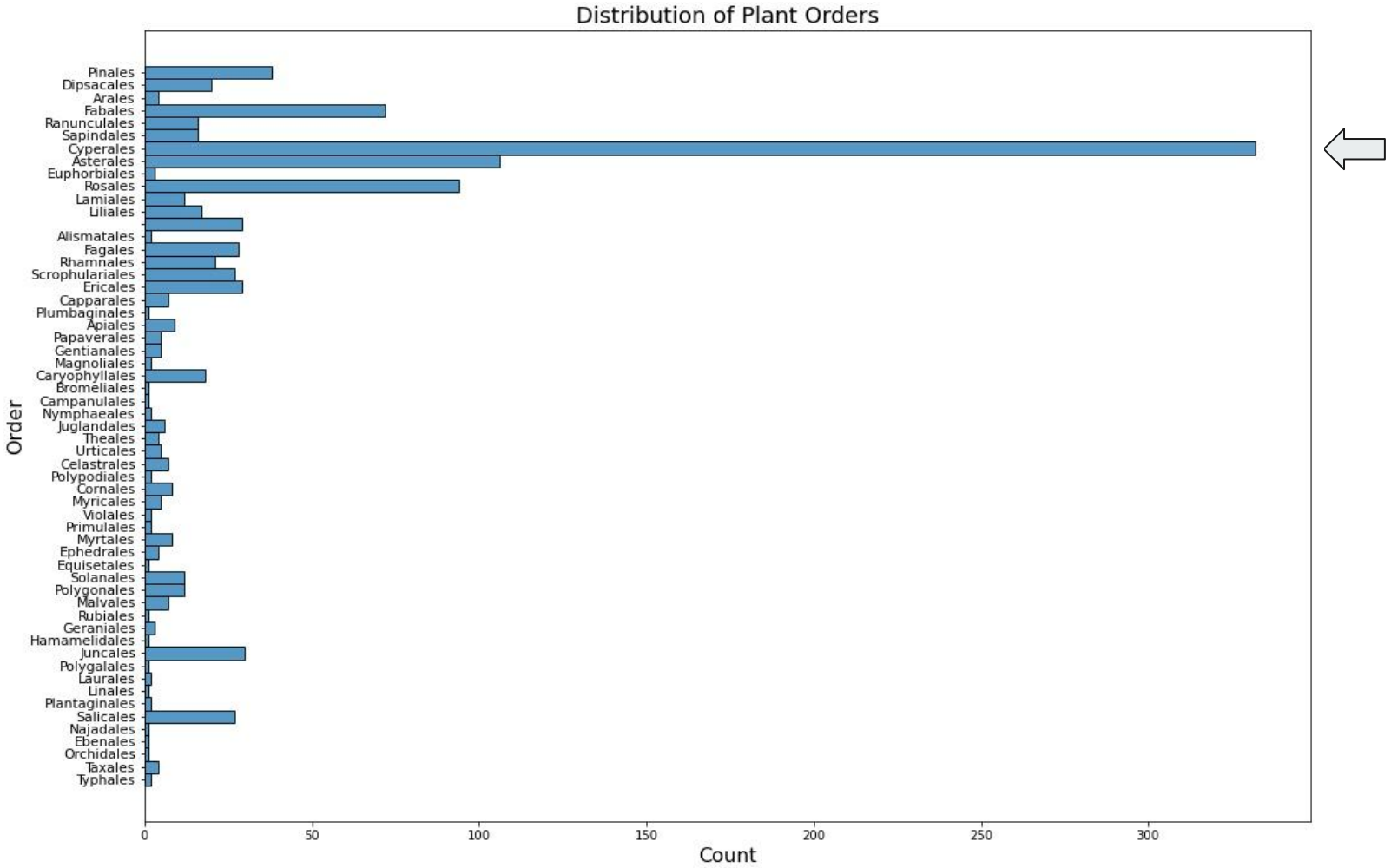
Background

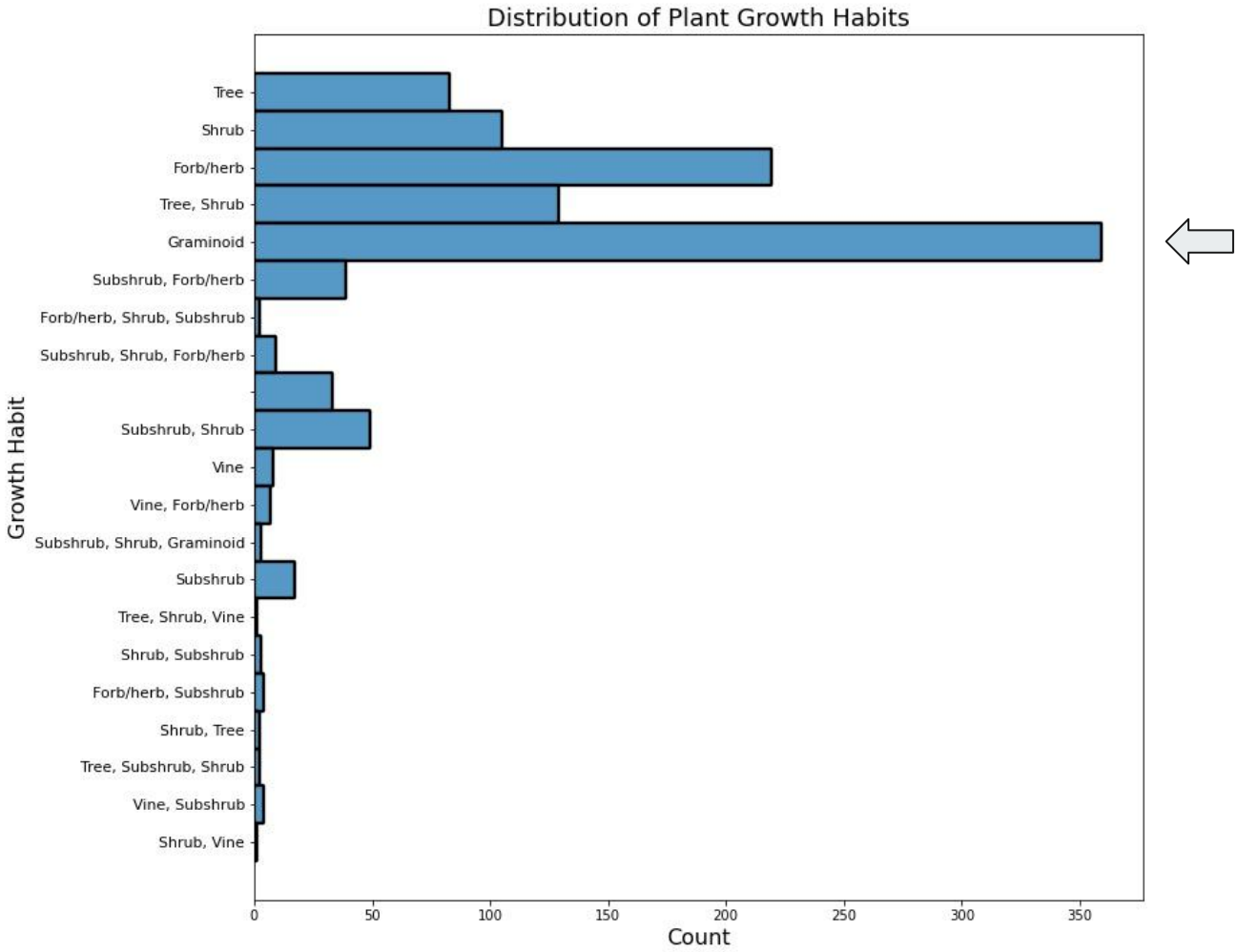
- People are increasingly turning to gardening as a hobby
  - Note record sales of seeds in 2020
- Starting a new hobby can be daunting
- Lots of questions about what plants to grow
  - More difficult to answer than it seems at first
- Lots of information online about growing plants

Problem Statement:
Create a model that can recommend plant species to grow based on growing conditions

Data Collection

- Surprisingly difficult to find a good database
  - Many paywalls and databases lacking detail
- Used the USDA's PLANTS database
- Information about ~90,000 species across North America
- Detailed information about **1078 species**
- Accessed this via a SQL query
  - Database came from a USDA API
- Data cleaning
  - Removed null values
  - One-Hot Encoding categorical features


Distribution of Plant Orders
Order
Pinales
Dipsacales
Arales
Fabales
Ranunculales
Sapindales
Cyperales
Asterales
Euphorbiales
Rosales
Lamiales
Liliales
Alismatales
Fagales
Rhamnales
Scrophulariales
Ericales
Capparales
Plumbaginales
Apiales
Papaverales
Gentianales
Magnoliales
Caryophyllales
Bromeliales
Campanulales
Nymphaeales
Juglandales
Theales
Urticales
Celastrales
Polypodiales
Cornales
Myricales
Violales
Primulales
Myrtales
Ephedrales
Equisetales
Solanales
Polygonales
Malvales
Rubiales
Geraniales
Hamamelidales
Juncales
Polygalales
Laurales
Linales
Plantaginales
Salicales
Najadales
Ebenales
Orchidales
Taxales
Typhales
0
50
100
150
200
250
300
Count

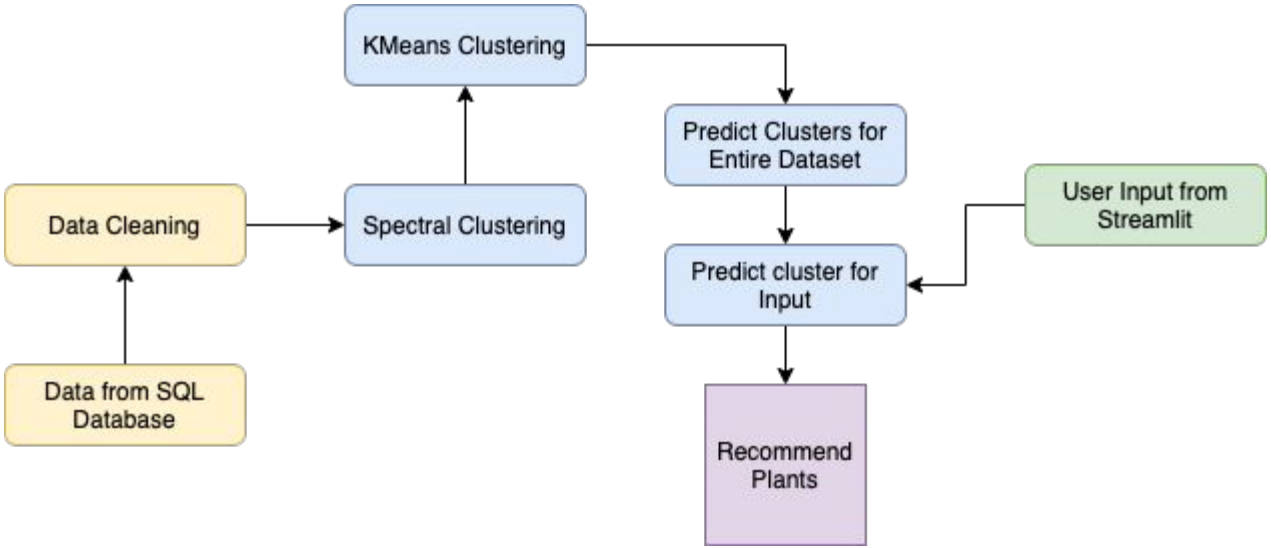

Distribution of Plant Growth Habits
Growth Habit
Tree
Shrub
Forb/herb
Tree, Shrub
Graminoid
Subshrub, Forb/herb
Forb/herb, Shrub, Subshrub
Subshrub, Shrub, Forb/herb
Subshrub, Shrub
Vine
Vine, Forb/herb
Subshrub, Shrub, Graminoid
Subshrub
Tree, Shrub, Vine
Shrub, Subshrub
Forb/herb, Subshrub
Shrub, Tree
Tree, Subshrub, Shrub
Vine, Subshrub
Shrub, Vine
0
50
100
150
200
250
300
350
Count

Features Included in the Model

- Growth Habit
- Family
- Native Status - whether or not a plant is native to a region
- Toxicity
- Shade Tolerance - how much sun does a plant need?
- Minimum Temperature (F) - The coldest temperature a species can survive
- Soil pH
- Drought Tolerance
- **22 features in total!**

Modeling

An Overview of the Modeling Process



Clustering

- Tried out several clustering models
- DBSCAN
  - Helped figure out the number of clusters needed (recommended 68 clusters)
  - Silhouette Score of 0.14
- KMeans
  - Silhouette Score of 0.049
- Went with Spectral Clustering
  - Creates a graphical representation of the dataset
  - Works with KMeans to predict clusters
  - Outperforms other clustering models regularly, but is more error prone
  - **Silhouette Score of 0.51!**

Building the Recommendation Engine

1. Make a dummy entry based on the user input
2. Use clustering model to predict cluster locations for the whole dataset
3. Predict the cluster of the dummy entry
4. Sample plants randomly from the same cluster as the dummy entry
5. Display the species of the plants suggested

- Item-based Recommender System
- Created a test case purely in Python first

Creating an App

- Lots of different options available here
- Decided to go with **Streamlit**
  - New library/ hosting service
  - Can be used to turn Python scripts into interactive apps
- My Streamlit app can be found on this project's repo
  - If you want to run the app, make sure to install Streamlit via Conda/pip and download the entire project repo

Conclusions

Results

- Successfully created a plant species recommendation engine
- Very bare bones recommender, however
- Few distinct clusters make the results sometimes inaccurate
- App runs slowly
- Recommender would be useful for:
    - Large commercial clients
    - Native plant restoration

Next Steps and Recommendations

- Try to get more data on more commercially available plants
  - Too many grasses in this dataset!
- Improve upon the Streamlit app
  - Add Wikipedia links
  - Add Images for each suggested plant
  - Add more features from the dataset
- Try out a hybrid approach with collaborative filtering
  - Try to include what plants sell in conjunction with each other
  - These models tend to outperform item-based recommenders
- Make the app available to the public

Thank You!