Problem Statement

Can a neural network or multiclassification model be trained to identify crops planted in a field with at least 90% accuracy, so that Prime Agri's management team can make a decision on the most competitive seed line to pursue in the new territory within the next year?

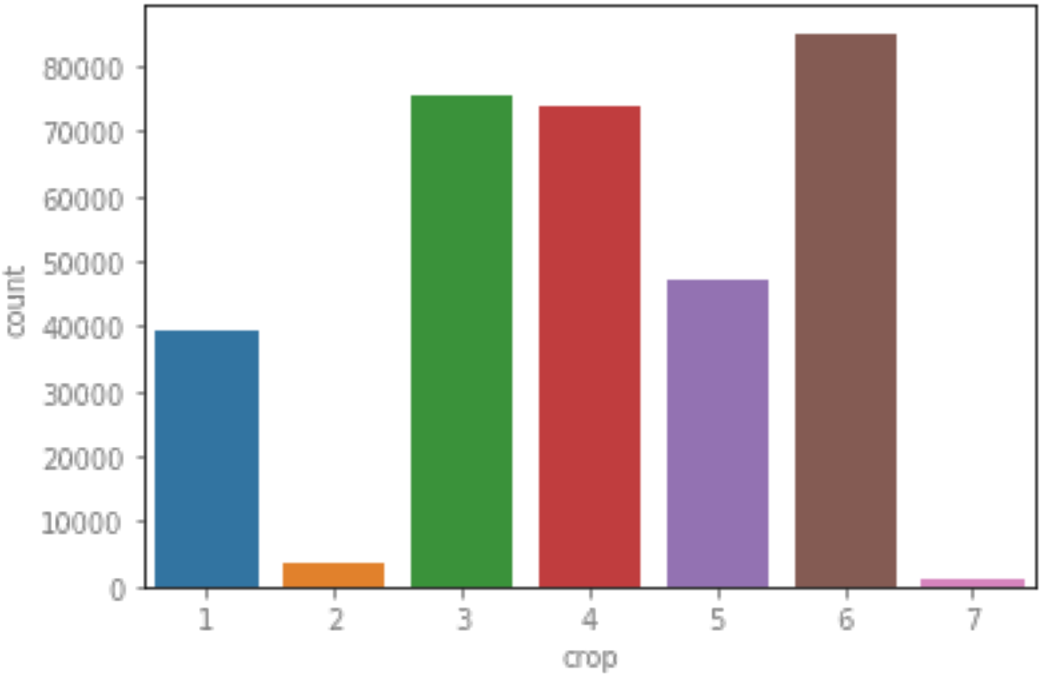
Data

The data for this model come from the UCI Machine Learning Library. It contains a dataset with 175 features and 325,834 observations. It contains optical and PolSAR remote sensor readings, taken across fields of 7 different crops including; corn, peas, canola, soybeans, oats, wheat, and broadleaf field crops.

Methods

The goal of this project is to train and deploy a multiclassification model that can identify the crop being grown in a field with at least 90% accuracy. The number of target classes and the high dimensionality of the dataset mean that proper feature selection and class balance are the most important issues to address prior to training ML models. Failure to address said issues will negatively affect model performance.

The crops grown in a given field depend on several factors, some of which are due human choice (personal/family experience or farming practices) and some which are beyond the control of the individual farmer (soil contents, weather, market demand). For these reasons there is great variability in determining which crops are being grown in any given field. This variability is seen in the dataset in the drastic imbalance between datasets. Two common and simplistic practices for achieving greater class balance are through oversampling or bootstrapping minority classes and under sampling majority classes. However, each of these techniques has rather significant short comings. Under sampling essentially decreases the size of the data available to train the ML model, and oversampling generates new data in the minority classes by duplicating existing observations. Given the number of observations and a great disparity between crop classes I choose to use a Synthetic Growth Technique to solve the class imbalance in the model, specifically SMOTE (Synthetic Minority Oversampling Technique). SMOTE can increases sample size without duplicating, it creates new data points by 'polling' the K nearest neighbors to each point in the minority class and then randomly plotting a new point linearly between the chosen point and each neighbor until the number of observations is balanced with that of the majority class.



Corn - 1	Peas - 2	Canola - 3	Soybeans - 4	Oats - 5	Wheat - 6	Broadleaf - 7
39,162	3,598	75,673	74,067	47,117	85,074	1,143

The high level of dimensionality in the dataset can lead to what is known as the "curse of dimentionality". This "curse" causes the volume of the learning space to increase so rapidly that it actually makes the data increasingly more sparse. Meaning that a ML model trained on high dimensional data is less generalizable to unseen data. To solve this problem feature selection must be performed to reduce the dimensionality in the data. This can be done by eliminating less important or highly correlated features; however, this eliminates any information that these features might add to model. In order to maintain the predictive information of all the features while still reducing the dimensionality of the data overall PCA was chosen as the reduction method.

Data Cleaning and EDA

Data_import – Github page
EDA_Cleaning – Github page

There were relatively few cleaning issues with this dataset. The data downloaded from the UCI library contained no missing values, and all of the features were continuous measurements so there was no need to transform categorical features prior to model training. However, the feature names had to be scraped from the library website and appended to the data using beautiful soup and regex.

Exploratory data analysis identified a few major issues within the dataset that had to be addressed prior to model training:

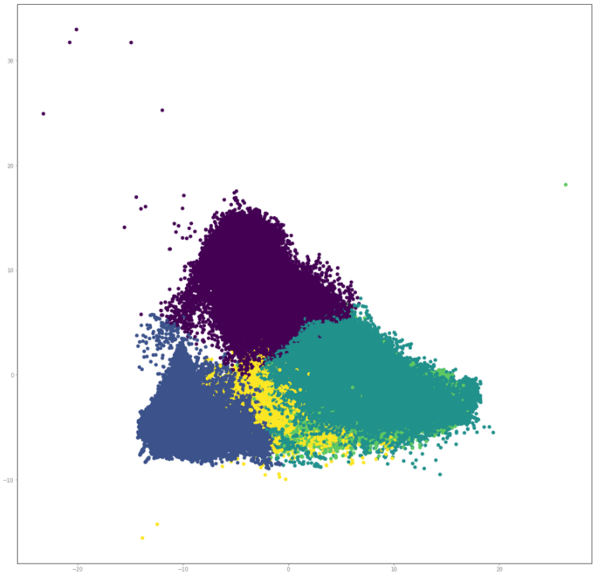
Problem 1: Class Imbalance

The classes within the dataset are significantly imbalance. Particularly within the Broadleaf class, which contained only 1,143 total observations out of a total of 325,834 observations. This significant lack of representation paired with the knowledge that Broadleaf is a term that can be applied to numerous crop species, including both corn and soybeans, lead to the decision to drop the Broadleaf class entirely. With the dropping of the Broadleaf class complete, the dataset was split between training and test sets and the SMOTE was applied to balance the remaining classes within the training set. This step is extremely important, because of how SMOTE generates new samples failure to separate the test data from the training data prior to the algorithm usage will lead to information from the test data leaking into the training data.

Problem 2: Curse of Dimensionality

As mentioned above, the dataset began with 174 features. Each feature was a different sensor measurement taken by a satellite. Upon exploration it was discovered that of the 174 original features 134 features had an absolute correlation greater than .9 with at least one other feature in the dataset. Preliminary unsupervised clustering methods like KMeans were unable to accurately separate wheat from oats.

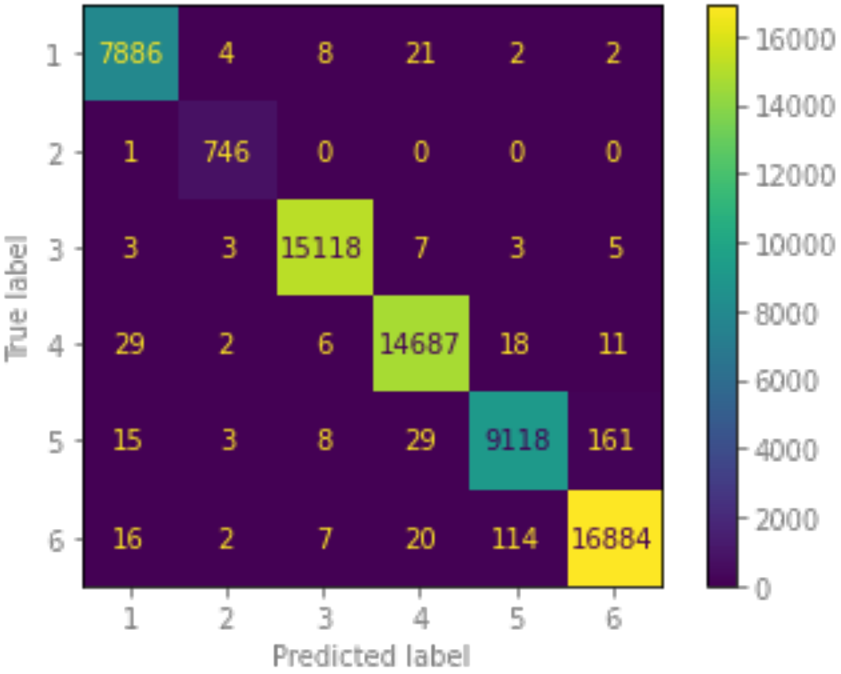
KMeans Clustering (2D)



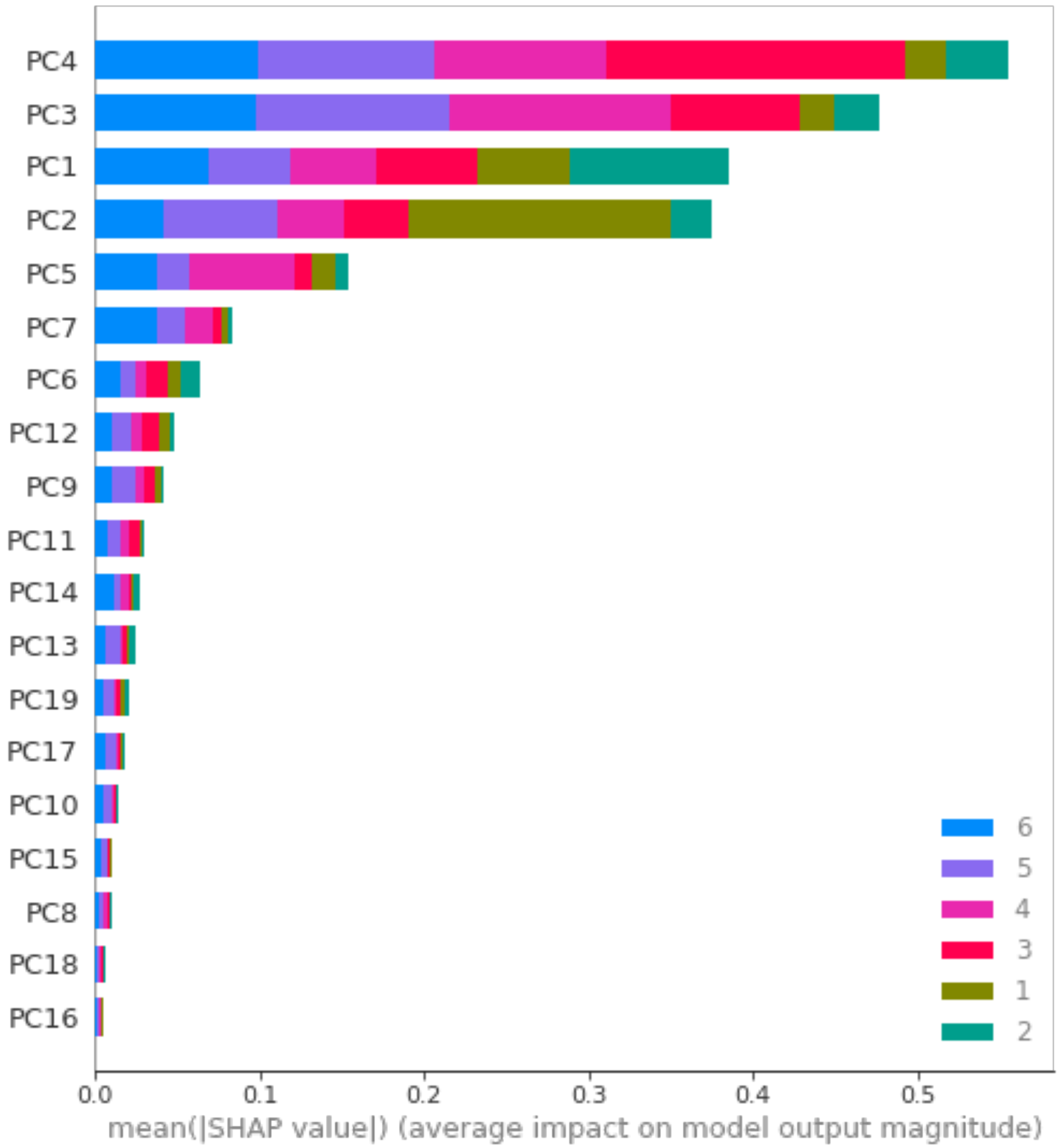
The high correlation among features, the knowledge that each feature is a different temporal, spectral, textural, and polarimetric sensor readings (not necessarily easy to understand outright), and the inability of unsupervised methods to accurately separate the 6 target classes, lead me to choose PCA as the preferred method of dimensionality reduction. Despite the loss of explicability of the models features inherent to PCA, it reduces dimensions of a model and preserves all of the information of those features by combining several features into Eigen values which then can be added together to explain a chosen threshold of variation within the model (this case 90% was chosen).

Model Selection

Model performance will be evaluated using 3 models, and if the goal threshold is not met then a neural network will be trained in an attempt to meet the threshold criteria. The initial models tested will be Logistic Regression, K Nearest Neighbors, and Random Forest Classifier. After cross validation the RF model significantly outperformed the other models in terms of score and computation time.



With an roc_auc score of .99986, the RF model was nearly perfect, and as such more than meets the predetermined goal of a 90% accurate classification score.



Using shap values we can see that the most important components in crop classification were, in order of importance globaly, PC4, PC3, and PC1. The table below depicts the order of importance on class (local) level.

Crop	Corn	Peas	Canola	Soybeans	Oats	Wheat
Shap 1	PC2	PC1	PC4	PC3	PC3	PC4
Shap 2	PC1	PC4	PC3	PC4	PC4	PC3
Shap 3	PC4	PC3	PC1	PC5	PC2	PC1

Please see Model Documentation for more detailed analysis.

Future Improvements

More sophisticated models can be fit to the data in an attempt to further improve roc_auc score. Perhaps Boosting algorithm or a neural network could achieve a score of 1.

More investigation into what has caused the model to incorrectly predict the falsely predicted observations would also be beneficial in improving the accuracy further.

More in depth analysis of Shap values may be beneficial, and perhaps finding a method other than PCA to reduce dimensionality of the data would further improve interpretability of results.

Credits

Thank you to UCI and Dr. Iman Khosravi for posting and allowing use of the dataset.

A huge thank you to my data science mento Wei Ang, without your time and guidance this project would not have been possible.