Predicting Commercialization of Soybean Varieties

David Lee (dyl44), Clara Lishan Ong (lo88), Zilong Wang (zw243)
Cornell University

1. PROBLEM DESCRIPTION AND MOTIVATION

Unlike subsistence farms, commercial farms produce crops and other agricultural products for the sole purpose of generating a profit. In order for a crop variety to become commercialized, it often has to undergo several rounds of strict testing and experimentation to ensure that it offers good yield. Throughout these multiple phases of testing, crop varieties are often benchmarked against their "peers" and successfully commercialized varieties that came before, by their yield. Every year, varieties that fail to make the cut are discontinued, while those that survive compete again in the following year with newly introduced varieties. Figure 1 shows the testing process for varieties in the class of 2014, before commercialization in 2014.

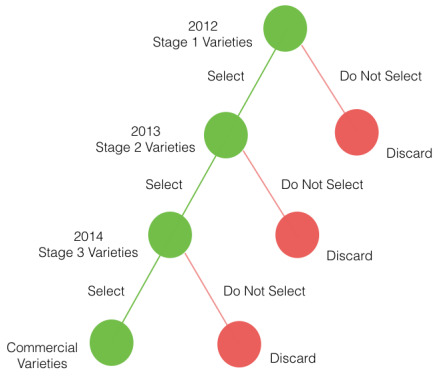

Fig. 1. Stage Gates.

Soybean varieties that have passed all tests and graduate (become commercialized) are expected to do well in sales. However, in reality, Type 1 errors occur, where some underperforming varieties slip past the checks and still get commericalized. It thus behooves us to find a learning model that can predict commercialization with a greater degree of accuracy and thus minimize our number of misclassified (mistakenly commercialized) varieties.

Therefore, the primary goal of the INFORMS competition is that given past experimental data, we should develop a predictive model that can be used to predict whether each soybean variety in the class of 2014 will graduate from the years of testing and be commercialized. This report describes our exploratory data analysis, predictive modeling methods and results.

2. THE DATASET

2.1 Variables

Before using any sophisticated procedures, we first performed exploratory data analysis to get acquainted with the dataset. Our data consists of 10 predictors and a response variable, and we briefly describe each variable in Table I. Our response variable is GRAD, which is a categorical variable. It takes on the value of -1 if the variety is not part of a class, 1 if the variety is part of a class and graduates, and 0 if it is part of a class and does not graduate.

2.2 Data Format

The type of data given to us is called panel data, which refers to multi-dimensional data involving multiple (repeated) measurements over multiple time periods.

(1) In an experiment, several varieties (the same varieties) are tested over multiple locations. Within each location, some varieties (the same varieties) are benchmarks, as indicated with CHECK = TRUE.
(2) Within the same experiment and location, the test for a variety is replicated. Replication is necessary because the yield can vary even within the same field due to, for instance, different soil conditions.
(3) Therefore, one row of the panel data represents a single replication of a specific experiment, location and variety.
(4) For each variety, the FAMILY, RM and GRAD are fixed, whereas YIELD varies.

2.3 Splitting the Data

We split our dataset into a training set, a test set and a prediction set. The prediction set consists of all unlabeled observations from the class of 2014. The training and test sets have a total of 253,663 rows, and the prediction set has 4,590 rows.

The training and test sets are split using the 5-fold cross validation for time series, based on the year of the experiments. Each training set comprises of past data, and the test set comprises of future data. This is slightly different from the usual cross validation for non-time series data, because, in our case, the sizes of the training and test sets differ for each fold.

Fold	Training Set	Test Set
Fold 1	2009	2010
Fold 2	2009, 2010	2011
Fold 3	2009, 2010, 2011	2012
Fold 4	2009, 2010, 2011, 2012	2013
Fold 5	2009, 2010, 2011, 2012, 2013	2014

Table I. Description of the Variables

Variable	Description
Year (Integer)	When the experiment was conducted.
Experiment (Categorical)	Consists of experimental varieties of relative similar maturity that are tested together.
Location (Categorical)	Where the experiment was conducted.
Variety (Categorical)	Groups of soybeans that are genetically identical.
Family (Categorical)	Sharing the same family means that the varieties have the same parents.
Check (Boolean)	Whether the commerical soybean varieties are used as performance benchmarks in yield trials.
RM (Float)	Soybean relative maturity. Every 0.1 stands for 1 day.
REPNO (Categorical)	Replication number. A variety under a specific experiment and location is tested more than once.
Yield (Float)	This refers to the amount of grain per unit of land that a soybean variety produces.
Class (Integer)	The batch the soybean variety belongs to (2011, 2012, 2013, 2014). Takes on a value of -1 if it is not part of a class.
Grad (Categorical)	Whether the soybean variety graduated from the last round of yield test and proceeded to be commercialized.

2.4 Sampling the Training Data

The original dataset was heavily imbalanced, where the vast majority of the varieties are not part of a class (GRAD = -1). The table below shows the distribution of classes for each year in the training sets.

Year	GRAD = -1	GRAD = 0	GRAD = 1
2009	98.9%	0.7%	0.4%
2010	97.2%	1.9%	0.9%
2011	93.8%	3.4%	2.8%
2012	91.4%	3.8%	4.8%
2013	78.4%	9.8%	11.8%
2014	78.3%	5.2%	16.5%

In order to reduce the imbalance, we oversampled the underrepresented classes (GRAD = 0 and GRAD = 1) and undersampled the overrepresented class (GRAD = -1). To obtain the number of samples required for each class, we summed the number of rows in a particular training set and divided it by three, thus ensuring that each class was equally represented. Sampling was done randomly with replacement. The test set remained the same, following the original data. The following table displays the number of rows per class with the second sampling method.

	GRAD			No. of Rows	
Training Set	-1	0	1	Total	Per Class
Set 1	11084	77	46	11207	3736
Set 2	30794	468	230	31447	10482
Set 3	128234	4006	3113	135353	45118
Set 4	203956	7112	7111	218179	72726
Set 5	221230	9269	9709	240208	80069

3. EXPLORATORY DATA ANALYSIS

Next, we decided to visualize some of our data. When plotting Yield against RM for the year 2009, we saw that there is a positive relationship between these two variables. We also saw no observable difference between the varieties that graduated and those that did not (Figure 2).

In Figure 3, we saw that varieties without a class, graduates and non-graduates have similar yields. From Figure 4, we noticed that

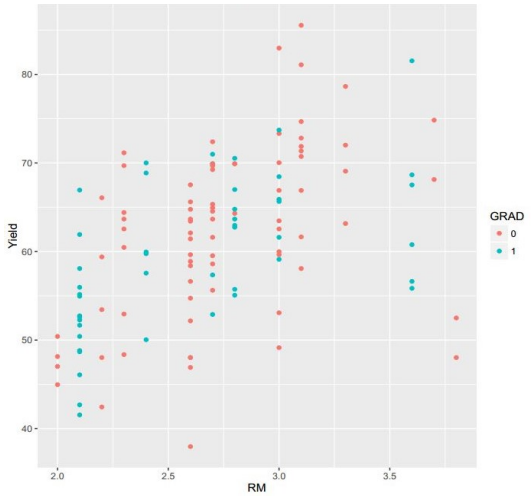

Fig. 2. Yield vs Relative Maturity, by Grad (2009).

benchmark varieties make up approximately 1/10th of those without a class and non-graduates, while they make up approximately one-half of the graduates. Lastly, from the bar chart, it can also be seen that GRAD = -1 is the overrepresented class, as mentioned in section 2.4.

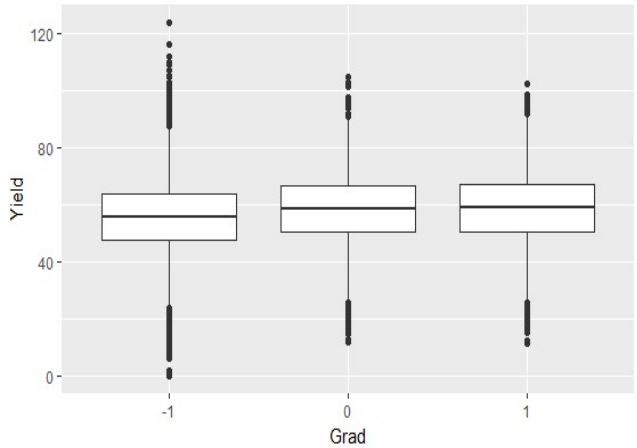

Fig. 3. Boxplot of Yield vs Grad.

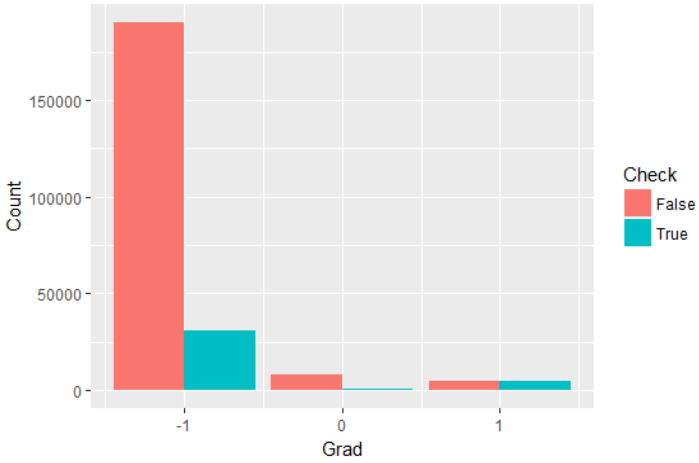

Fig. 4. Count of Check by Grad.

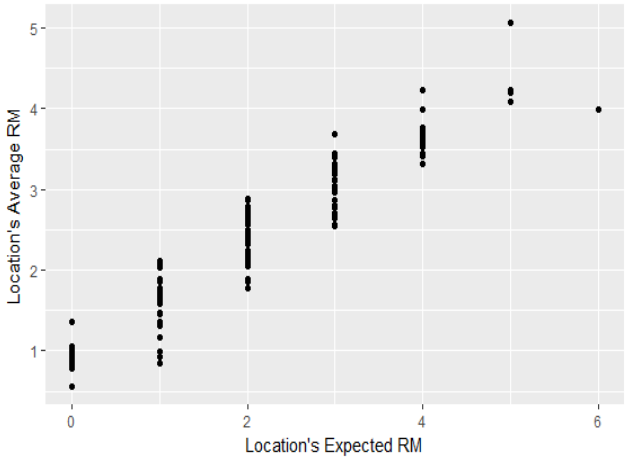

Fig. 5. Location's Average RM vs Expected RM.

4. FEATURE ENGINEERING

A major challenge that we encountered was the large number of categories for the variables EXPERIMENT, LOCATION, VARIETY and FAMILY.

Variable	Number of Categories
Experiment	534
Location	152
Variety	15632
Family	1938

The variable EXPERIMENT was discarded because the experiments were different for each year. Therefore, it was not possible to test our model on experiments that were not present in the training sets.

The variable LOCATION was discarded. The first digit of the location number indicates the expected relative maturity of the site. For example, LOCATION 3130 has an expected relative maturity of 3. From our analysis, the expected relative maturity is close to the actual relative maturity (Figure 5), hence using RM was sufficient.

We did not use VARIETY as a predictor because there are varieties that appear in the test sets but do not in the training sets. Also, different varieties belong to different graduating classes (or non-class where CLASS = -1), and each variety is only present in one possible class. Therefore, all of the varieties that appear in the prediction set (class of 2014) do not appear in any of the training and test sets. As a result, we could not predict the outcomes using varieties that were not seen before.

Because there are 1,938 families, our models took a long time to run. From doing set intersections, we discovered that there are families that appear in the test sets but do not in the training sets. Also, there are 3 out of 61 families that appear in the prediction set but do not in both the training and test sets. Therefore, we performed feature engineering on the variable FAMILY. For each year, we generated a list of unique families and calculated the average yield by family. We did this for each year, as the average yield rankings differ by year. We then sorted the families by descending order of average yields, and allocated the families into 5 equal segments according to the average yields. In other words, the families with the highest average yields for that year will be in segment 1 and those with the lowest will be in segment 5. This resulted in a new feature called "FAMILY SEGMENT," and one-hot encoding was subsequently done across the 5 segments.

Eventually, we shortlisted these variables as predictors:

(1) YEAR
(2) CHECK
(3) RM
(4) YIELD
(5) FAMILY SEGMENT [new; five variables generated from one-hot encoding]

5. METHODS AND RESULTS

This section describes the evaluation metrics, classification methods used and results obtained.

5.1 Precision, Recall, and F-score

In these three-class classification problems, we are concerned about both the precision and recall. Precision refers to the fraction of responses predicted to be in class i that were actually class i, and recall refers to the fraction of responses actually from class i that were predicted to be from class i. Precision and recall take values between 0 to 1.

In a binary classification problem, precision and recall are calculated as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{2}$$

For our problem, precision and recall are calculated for each predicted class i in each cross-validation fold.

	Actual		
Predicted	-1	0	1
-1	M_{11}	M_{12}	M_{13}
0	M_{21}	M_{22}	M_{23}
1	M_{31}	M_{32}	M_{33}

$$\text{Precision}_i = \frac{M_{ii}}{\sum_i M_{ij}}. \tag{3}$$

$$\text{Recall}_i = \frac{M_{ii}}{\sum_j M_{ij}}. \tag{4}$$

Precision and recall are much better metrics than accuracy, as it will lead to misleading results for an imbalanced dataset like ours. For instance, a high accuracy could be achieved by predicting all responses belonging to the majority class.

For each fold, we calculated the precision and recall, and then computed the average of these metrics over all folds and all classes. Since there is a tradeoff between precision and recall, it is important to have a single metric to evaluate the models. So, for each model, we computed the F-score, which is the harmonic mean of the precision and recall. Ideally, we would want this value to be close to 1.

$$\text{F-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

5.2 Naive Bayes Classifier

Naive Bayes classifiers are a family of simple probabilistic classifiers based on Bayes' Theorem. The model is "naive" because we assume that the features are independent, conditional on the class label [1]. The predicted class label $\hat{y} = C_k$ for some $k \in 1, \ldots, K$ is assigned to maximize the *a posteriori* probability. Because the features are conditionally independent given the class label, the class conditional density can be written as a product of one-dimensional densities [1].

$$\hat{y} = \underset{k \in 1,\ldots,K}{\arg\max}\ p(C_k) \prod_{i=1}^{n} p(x_i|C_k). \tag{6}$$

Our model performed poorly, with the average precision and recall being slightly better than random chance (one-third). Although the average precision was higher in the test set than in the training set, the test set had a lower average recall. The test set had only a slightly lower F-score of 0.390 than the training set, which suggested only a small amount of overfitting. The performance for Naive Bayes was marginally better than random chance (0.333, with 3 possible classes). There was a small amount of overfitting, which is to be expected since the classifier is fitting a joint distribution over many features [1].

When the model was run using the second sampling method with equally represented classes (see section 2.4), similar results were obtained.

Set	Avg Precision	Avg Recall	Avg F-score
Training	0.385	0.464	0.421
Test	0.431	0.356	0.390

5.3 K-Means Clustering

K-means clustering is an approach for partitioning a dataset into k distinct, non-overlapping clusters, with the aim of finding homogeneous subgroups among observations [2]. A good clustering is one where the sum of squared Euclidean distances between each point and its cluster centroid is minimized.

Here, we used k-means clustering in a semi-supervised context. On the training sets, the algorithm discovered clusters without being given the labels. The model performance on both the training and test sets were evaluated using the actual predefined labels. Here, we ran the algorithm to search for 3 clusters. Because a different solution could be obtained depending on the initial cluster centroids, we let our algorithm use 20 initial configurations and report the best one.

K-means clustering performed worse than naive Bayes, with the average precision, average recall and F-score being very close to random chance (0.333, with 3 possible classes).

Set	Avg Precision	Avg Recall	Avg F-Score
Training	0.333	0.326	0.330
Test	0.335	0.331	0.333

From the confusion matrices, we saw that the clusters found by the algorithm did not match the predefined clusters. As an example, the confusion matrix for fifth test set is shown below.

	Actual		
Predicted	-1	0	1
1	4437	285	981
2	5047	355	1033
3	1055	53	209

Subsequently, we attempted another approach that viewed varieties without a class (GRAD = -1) as unlabeled data, where the varieties could either eventually graduate or not. Hence, we ran the k-means algorithm for 2 clusters and calculated the average F-score for only the graduates and non-graduates.

Again, the k-means clustering performed just as poorly, with the average precision, average recall and F-score being close to random chance (0.5, with 2 possible classes).

Set	Avg Precision	Avg Recall	Avg F-Score
Training	0.0.514	0.514	0.514
Test	0.493	0.493	0.493

As an example, the confusion matrix for the fifth test set is shown below. The clusters found by the algorithm did not match the predefined clusters.

	Actual		
Predicted	-1	0	1
1	7305	476	1556
2	3234	217	667

When the 2-cluster and 3-cluster models were run using the second sampling method with equally represented classes, similar results were obtained.

5.4 Adaptive Boosting

Adaptive Boosting (AdaBoost) is an ensemble learning method where the output of various decision trees (weak learners) are combined into a weighted sum to develop a stronger learning algorithm. In a nutshell, the way this technique works is that it regularly updates the weights of each coefficient according to the performance at each stage of the decision tree [3]. As long as the weak learners have individual error rates that are lower than random chance, AdaBoost has been shown to converge to a better solution.

However, as AdaBoost is typically used for 2-class predictions (we have 3), we decided to keep it simple by excluding the data with GRAD = -1 labels. Here, we assumed that varieties without a class would eventually graduate (or not), and that we would not know whether our predictions of 0 or 1 were correct given the actual label of -1. After removing the data with GRAD = -1, there were 21,894 rows in the training and test sets in total. The number of rows in the prediction set remained unchanged at 4,590 rows.

As expected of AdaBoost, the training model fit the training sets extremely well, and when tested on the test sets, the model performed relatively well as compared to naive Bayes and k-means.

It is worth noting that for the training set, the precision, recall and F-score hover at around 0.98 to 0.99, and for the test set they are around 0.70, which is a noticeable but expected drop in performance due to overfitting.

Set	Avg Precision	Avg Recall	Avg F-score
Training	0.983	0.987	0.985
Test	0.705	0.700	0.702

Although AdaBoost had a much higher average F-score than naive Bayes and k-means, the average F-score fluctuated across the 5 test sets. AdaBoost performed the worst in the fifth test set, as shown by the confusion matrix below. In the fifth test set, there were more misclassified points than correctly classified ones.

Set	Avg F-Score
Test Set 1	0.673
Test Set 2	0.952
Test Set 3	0.662
Test Set 4	0.758
Test Set 5	0.467

	Actual	
Predicted	0	1
0	296	1117
1	397	1106

Again, when the model was run using the second sampling method with equally represented classes, similar results were obtained.

6. CONCLUSION

6.1 Model Comparison

Among all the classification methods we explored, AdaBoost achieved the highest average F-score of 0.702 across the five test sets. However, it fluctuated across the five folds, doing as well as 0.952 on the second training model but as poorly as 0.467 on fifth training model. In contrast, naive Bayes and k-means had average F-scores close to random chance, and the scores were similar across the five folds. Therefore, while AdaBoost outperformed the other methods, AdaBoost's reliability could still be strengthened.

All three models gave similar results regardless of how the training data was sampled from the original dataset.

6.2 Model Application

The goal of the model is to predict which varieties in the class of 2014 will graduate from the tests and become commercialized. Because the AdaBoost model trained on second training set gave the highest F-score (0.952), this model was used to make predictions for the class of 2014.

It is important to note that our models and predictions were done on panel data with multiple measurements for each variety. To address the goal of the model, we consolidated the predictions for each variety in the class of 2014 and checked whether GRAD = 0 or GRAD = 1 was the majority class. We discovered that 36 out of 38 varieties have predictions that were 100 percent in either GRAD = 0 or GRAD = 1, showing us that our model predicted a clear majority. This gives us confidence in our AdaBoost model. In contrast, naive Bayes predicted GRAD = -1 for all rows in our prediction set, which was suspicious. In k-means, for a particular variety, the predictions were more scattered across the three possible labels.

6.3 Limitations and Future Work

Our biggest obstacle was the ambiguity surrounding the varieties without a class (GRAD = -1), which made up 92% of our original data. We have explored various strategies such as removing these in AdaBoost, treating them as a "class" of their own in naive Bayes and 3-cluster k-means, and keeping them as unlabeled / missing data and predicting their graduation or non-graduation in the semi-supervised 2-cluster k-means clustering.

In order to remedy this issue, we plan on implementing the Generalized Low Rank Model (GLRM), which is a modern machine learning technique for imputing missing data entries and identifying important variables. Since many of our features consist of a large number of categories, the GLRM will aid us in finding similarities among these categories and thus reduce the dimensionality of the problem. Variables that were previously discarded, such as

LOCATION, can be incorporated into our model so that other abstract aspects of LOCATION apart from RM can be captured.

To prevent overfitting in AdaBoost, we would like to use various forms of regularization, such as early stopping (limiting the depth of the tree) and tree pruning (building a complex tree and simplifying it).

We believe that our method will help Syngenta, the host sponsor for the INFORMS competition, efficiently determine which soybean varieties to commericialize as well as ensure that their customers are satisfied with their products.

7. ACKNOWLEDGEMENTS

We would like to thank Syngenta and INFORMS for their data and assistance. We would like to thank our classmates for their valuable feedback in the peer reviews. Lastly, we are grateful to Professor Madeleine Udell for guiding us.

8. REFERENCES

[1] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT, pp. 82-84, 2012.

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. New York: Springer, 2013.

[3] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer, 2001.