Submitted to the Syngenta AI Challenge

FOR INTERNAL USE ONLY (leave blank)
Submission Number:
Team Number:

INSTRUCTIONS TO AUTHORS
• Use this template in either the .doc or text with style file format.
• Do not remove the INTERNAL USE ONLY line above
• Please do not change the font size margins or headers
• Unless otherwise requested, do not include personal names or information in your write-up
• Failing to do any of the above may result in a delay in evaluating your submission

Portfolio Selection and Yield Prediction Report

Shuo Chen, Qianqian Pan
5474 Delmar Blvd, St. Louis, MO 63112 U.S.A., schen833@wustl.edu;
6127 Pershing Ave, St. Louis, MO 63112 U.S.A., qianqian.pan@wustl.edu

The following seed varieties "V067897, " "V068083", "V083408", "V068101", "V114564", "V152334", "V081953", "V140352" are "elite" and will perform the best in farmers' fields in 2016. Four layers of model were used – a classification model was used to predict whether or not a variety will outperform the bench mark varieties in a certain location and accordingly a positive model (boosted tree) and a negative model (bagging tree) were used respectively for observations outperforming the bench marks and those underperforming to predict the yield difference between test varieties and benck mark. Besides, a risk management model was used to select the "elite" varieties with high yield but low variance. We also trained a Random Forest model to predict the yield both for "elite" variety group and unselected variety group.

Key words: Random Forest, Boosted Tree, Yield Difference, Portfolio Optimization, Risk Management

---

1. Introduction

The research goal is to help Syngenta select the true "elite" varieties and predict their yields in 2016. In order to select the "elite" varieties and predict their yields in 2016, we used 4 layers of model. The first step is to select the "elite" varieties. Our method is to set a certain risk level and selected varieties with highest yield but low variance. In this step, 3 layers of model were used. A Random Forest model was used to split all the varieties into two groups – in each location, if a variety outperforms the bench marks, it will be considered as a positive observation and negative observation otherwise, and accordingly a positive model and a negative model would be used to predict the yield difference between the yield of a variety in a particular location and that of the bench mark. By aggregating the yield difference of a variety in difference locations(both negative and positive), we can calculate the mean yield difference and variance and accordingly use a optimization model to select varieties with high yield but low variance. Next, we trained

another model for predicting the yield of "elite" varieties and result of 8 "elite" varieties yield prediction is listed below:

VARIETY_ID	PREDICTION
V067897	60.7341116
V068083	62.1997395
V068101	60.5991583
V081953	59.3652578
V083408	59.9730656
V114564	63.0822912
V152334	64.3856119
V140352	60.65392

* The result listed above is not the one we submitted on CodaLab. The list we submitted before is listed below. It was produced by using the same modeling and prediction method; however, because we didn't know how many "elite" varieties we were supposed to come up with, we just submitted a list with 20 varieties, which includes alomost all the truly elite comercial varieties selected by Syngenta in the real world but has a low F-score due to the size of the list (20). Therefore, above we come up with a new list which contains only 8 varieties and we believe the new list includes most truly elite varieties and will produce a higher F-score.

VARIETY_ID	PREDICTION
V052885	60.9681142
V053050	59.8115847
V067859	59.5479225
V067897	60.7341116
V067898	59.8287257
V067901	59.2570993
V068065	60.4185464
V068083	62.1997395
V068084	62.0025315
V068101	60.5991583
V068109	59.7146032
V068111	59.6946088
V081944	60.2225523
V081953	59.3652578
V083408	59.9730656
V114467	63.2260673
V114564	63.0822912
V114565	62.9751901
V152334	64.3856119
V152413	64.1383038

2. Criteria used to select the seed varieties

We select the seed varieties based on mean and variance of yield difference from benchmark varieties in all appeared locations. In general, an experiment variety which outperforms commercial varieties in different experiment has the potential to be elite. An elite variety is expected to have high yield difference from benchmark varieties across different locations on average. However, an experiment variety that has high mean of yield difference might have high variance in its advantage over commercial variety and we regard it as the "risk" of high yield advantage. Therefore, taking the trade of mean yield difference and risk into consideration, we expect an elite variety to have high mean of yield difference in all locations on average and low variance of yield difference among all experiment locations.

To get the criteria for each experiment variety, our solution is to develop a model to predict yield difference of 1093 experiment varieties from that of commercial varieties in 79 experiment locations. Then, We sorted and filtered top 50 varieties that have the highest mean yield difference (because of the variables number limitation on LINGO). Based on the predicted yield difference on 79 listed locations, we also calculated the variance and covariance matrix for the top 50 experiment varieties. Finally, the "truly elite" seed varieties come in as a portfolio that has the maximized mean of yield difference and minimized risk/variance.

The optimization functions are listed below:

$$\text{MAX} \quad \text{E}\left(\sum I_i C_i YD_i\right)$$

$\text{YD}_i - \ i's\ mean\ of\ predicted\ yield\ difference$

$I_i = 1 - Variety\ is\ selected\ ;\ I_i = 0 - Variety\ is\ not\ selected\ ;$

$C_i - Variety's\ propotion\ in\ portfolio$

$$\text{SUBJECT TO} \quad \sum I_i \leq 10, \sum C_i = 1, C_i \leq I_i$$

$$\sqrt{\sum I_i {C_i}^2 var(YD_i) + 2\sum I_i I_j {C_i}^2 {C_j}^2 cov(YD_i, YD_j)} \leq tolerable\ risk(e.g = 2)$$

Note: This process will be realized by using LINGO

3. Estimates of Type I Errors

In terms of the methodology introduced in challenge background, we calculated and got the elite list selected from stage three. Assuming the 8 varieties selected by us are true elite, the type I error rate is 94.44%, which means only 5.56% of varieties selected using traditional method are truly successful after they become commercial.

We recommend that the type I error rate should be reduced in two ways. First of all, a good selection method is expected to avoid high yield underperformer. To address this problem, our solution is to calculate and compare experiment varieties' yield difference from benchmark varieties to cancel out the yield scale difference among locations. Further, another way to reduce type I error is to consider the trade-off between mean yield advantage and "risk" (variance of yield difference). To make that happen, we expect to estimate experiment varieties' potential performance in all locations, even though they might have not been tested in some of locations. A model needs to be developed in this step to estimate experiment varieties' yield difference in terms of all appeared experiment locations (details are illustrated in methodology part). Among all locations, it's also valuable to look at the variance and covariance of top 50 varieties. The elite varieties would compose as a portfolio that generates high yield advantage while the advantage variance across locations is minimized. By doing so, we will have a comprehensive evaluation on all experiment varieties and exclude those that are not sustainable across sites.

4. Methodology

We divided the research problem into two aspects: "elite" variety selection and yield prediction. The methodology we used are described as below.

4.1. "Elite" Varieties Selection Phase

As we mentioned above, the criteria we use is to consider both yield difference (i.e. difference between yield of a test variety and that of bench mark) and variance, ensuring the stability and reliability of the prediction result.

Accordingly, the first step is to fit a mode predicting the yield difference for all the potential varieties in all the potential locations – because we don't know the location information of farmer's fields, we plan to predict the yield/yield difference in all the locations which have been used as test location in 3-stage testing period; therefore, each varietiey will have totally 79 yield/yield difference prediction and the mean yield/yield difference will be used for elite variety selection.

We use the data of class 2014(all three stages) as the whole dataset. In each location, a variety was grown twice, so we treated them as two independent observations. Also, we use the average yield of bench marks in the same location as the yield level of bench marks and calculated the yield difference – that is, the yield of test variety minus that of bench mark level. In order to improve the prediction accuracy, we trained a classification model to predict whether or not a variety will outperform the bench mark in a certain location. If the prediction result shows the variety will outperform in a particular location, then this observation will be thrown into a "positive" yield difference prediction model to predict the yield difference between the yield of this test variety and that of the benck mark; otherwise, it will be thrown into a "negative" yield difference prediction model. Finally, the average of the "positive" and "negative" prediction results of a variety will be considered as the prediction yield difference result and be used in the elite selection model. The benefit of doing so is that it can definitely reduce the variance of training data, and improve the accuracy of yield difference prediction.

For the classification model, we considered "LATITUDE", "LONGITUDE", "AREA", "IRRIGATION", "TEMP", "PREC", "RAD", "CEC", "PH", "ORGANIC.MATTER", "CLAY", "SILT_TOP",

"SAND_TOP", "AWC_100CM" from location dataset and "RM" as independent variables. In order the capture the difference among varieties, we created three factor columns to represent totally 1093 varieties – the first two columns vary from 1 to 24, and the third one varies from 1 to 2, and therefore, there are totally 24*24*2 = 1152 unique factor numbers, which can definitely take care of 1093 varieties. Then, we randomly select 80% of observations as training data and 20% as test data, and tried logistic regression model, Linear Discriminant Analysis model(LDA), Quadratic Discriminant Analysis model(QDA), K Nearest Neighbors(KNN) model, Decision Tree model, Bagging Tree model, Random Forest Model, Boosted Tree model and Support Vector Machine model. The model with lowest test error rate(i.e. the Random Forest model with the test error rate at around 32%) was selected.

Then, as we have discussed above, we trained a positive model and a negative model respectively based on the observations outperforming the bench marks and those didn't. By grouping the observations, we can expect a lower variance among the observations in each group and a higher accuracy of the prediction model. In this step, we used the same independent variables as the classfication model, while we treated yield difference between test varieties and bench mark average level as the dependent variable. We tried Linear Regression model, Ridge regression model Lasso regression model, KNN model, Decision Tree model, Bagging Tree model, Random Forest Model and Boosted Tree model. For the positive model, the Bagging Tree model was selected; for the negative model, the Boosted Tree model was selected. With these two models, we can predict the yield difference of a variety in different locations.

After we got the yield difference prediction results, we calculated the mean of yield difference and variance-covariance for each variety across 79 locations. Our objective is to maximize the mean yield difference of variety portfolio while controlling the selection amount and portfolio risk within a tolerant level. By doing so, we believe our selection of varieties would perform sustainably and steadily in all farm sites.

We have discussed the specific selection criteria before and Lindo was used to do the optimization analyisis. The optimization functions are listed below:

$$\text{MAX} \quad \text{E}\left(\sum I_i C_i YD_i\right)$$

$\text{YD}_i - \ i's\ mean\ of\ predicted\ yield\ difference$

$I_i = 1 - Variety\ is\ selected\ ;\ I_i = 0 - Variety\ is\ not\ selected\ ;$

$C_i - Variety's\ propotion\ in\ portfolio$

$$\text{SUBJECT TO} \quad \sum I_i \leq 10, \sum C_i = 1, C_i \leq I_i$$

$$\sqrt{\sum I_i {C_i}^2 var(YD_i) + 2\sum I_i I_j {C_i}^2 {C_j}^2 cov(YD_i, YD_j)} \leq tolerable\ risk(e.g = 2)$$

Note: This process will be realized by using LINGO

4.2. Yield Prediction Phase

In this phase, we use all the three stages observation data as the whole dataset. The same, we randomly select 80% of the observations as training data and 20% as test data. The same independent variables with previous models are used and yield of each observation is used as dependent variable. Accordingly, we tried Linear Regression Model, Ridge Regression Model, Lasso Regression Model, Decision Tree Model, Bagging Tree Model, Rondom Forest Model, and Boosted Tree Model. Finally, the Rondom Forest Model with the smallest MSE (46.64487) was selected.

The same, we predict the yield of each variety in 79 locations and then use the average of 79 prediction results as the estimated yield.

5. Quantitative results

5.1. Classification Model

Test Error Rate of the Best Classication Model – Random Forest Model: 32%

Test Accuracy of the Best Classication Model – Random Forest Model: 68%

5.2. Yield Difference Prediction Model

Mean Squared Error of the "Positive" Model – Bagging Tree Model: 8.400704

Mean Squared Error of the "Negative" Model – Boosted Tree Model: 10.55631

5.3. Yield Prediction Model

Mean Squared Error of Yield Prediction Model – Boosted Tree Model: 48.0296

5.4. Table of Top 50 Experiment Varieties (By Mean Yield Difference)

VARIETY_ID	mean_yield difference	variance_yield difference
V067897	8.532100412	11.66029708
V068083	7.105176541	12.58749233
V083408	6.96581998	10.67813447
V068101	6.944554925	8.550017406
V114564	6.893092059	10.52815956
V067898	6.719407878	14.4760688
V152334	6.71446663	3.73851493
V067901	6.612291237	9.188035571
V114565	6.548985776	11.99531093
V152413	6.521828861	6.400495443
V081944	6.477027339	14.44365344
V068084	6.451059953	16.72279838
V114465	6.427405635	10.22580436
V068231	6.41730841	8.043717967
V068165	6.416877542	3.469319874
V068111	6.415063017	5.773063613
V053050	6.386173763	5.957174132
V067859	6.358595023	7.091720927
V068109	6.357331258	9.182272273
V068065	6.3119563	8.692076647
V114467	6.280910231	12.02113834
V081953	6.268986352	4.062790089
V140352	6.263930512	8.93042636
V052885	6.258361705	5.279561916
V068112	6.208776237	6.837238095
V068064	6.189996788	10.42443086
V152286	6.158107023	14.18784402

V152440	6.120246078	10.0771353
V067872	6.101570457	10.18085901
V084855	6.079990604	9.884523925
V068063	5.998361777	9.511113107
V152280	5.992325799	9.620943727
V068075	5.979144362	12.98609168
V152262	5.974206373	9.303761351
V152281	5.968580119	11.7940791
V068113	5.961359143	6.266035076
V068166	5.949776012	5.717105746
V083390	5.93468835	17.1209571
V152253	5.928381036	9.966161692
V114541	5.922220283	33.77665262
V068102	5.905757865	6.995332632
V068068	5.900837714	20.86123724
V114501	5.898743204	10.76092929
V114663	5.87540877	10.06783595
V152283	5.873811242	10.52128712
V083409	5.854216591	33.53743835
V081494	5.848989889	4.046624032
V081495	5.840335766	3.977070458
V081954	5.840037162	5.6719454
V114552	5.833654582	16.29059661

5.5. Elite Varieties:

VARIETY_ID	Propotion(%)
V067897	27.33%
V068083	10.00%
V083408	10.00%
V068101	12.15%
V114564	10.52%
V152334	10.00%
V081953	10.00%
V140352	10.00%

5.6. Yield Prediction Result of Elite Varieties:

VARIETY_ID	PREDICTION
V067897	60.7341116
V068083	62.1997395
V068101	60.5991583
V081953	59.3652578
V083408	59.9730656
V114564	63.0822912
V152334	64.3856119
V140352	60.65392

6. Team members

- Shuo Chen, 5474 Delmar Blvd, St. Louis, MO 63112, U.S.A. schen833@wustl.edu
- Qianqian Pan, 6127 Pershing Ave, St. Louis, MO 63112 U.S.A., qianqian.pan@wustl.edu