AN ANALYTICS SOLUTION TO IDENTIFY ELITE SOYBEAN VARIETIES

Students
Surya Gundavarapu, Mayank Gupta, Eric Zheng
Purdue University Krannert School of Management
sgundava@purdue.edu; gupta363@purdue.edu; zheng279@purdue.edu;

Faculty Advisor
Matthew A. Lanham, PhD, CAP
Clinical Assistant Professor
Purdue University Krannert School of Management
Department of Management/Quantitative Methods
403 W. State St., Kran 466, West Lafayette, IN 47907
lanhamm@purdue.edu

Abstract

World Economic Forum put food security as the most pressing challenge of our generation. While agricultural sectors across the world have become more productive over the last half a century, there is still room for improvement when it comes to seed selection. Seed selection remains crucial since wrong choices of seed variety cannot be compensated with fertilization or mechanization. The purpose of this work is to design a strategy for selecting the elite soybean varieties that should be commercialized in the following year.

Keywords: Analytics, Predictive Analytics, Seed Selection

Introduction

The world's growing population and ever growing demand for food demand challenges the seed industry to develop and improve their seed varieties that maximize the yield and hence the profits for the farmers who use these high-quality seeds. However, it is a challenge to farmers and bio-tech firms alike since there are numerous parameters that influence a crop yield. Further, varying land management, disease resistance, weather conditions, and soil type add to the complexity of seed selection by adding to the secondary traits desired by the farmers according to Breene, K. (2016).

Syngenta is one such bio-tech firm looking to use the power of data to identify which soybean varieties to commercialize. This is a critical decision for Syngenta since its costly to perform a wide range of experiments. Moreover, Syngenta must find the right balance so as to provide seed varieties that meet or exceed customer's expectations.

To develop a solution that Syngenta can use to support their seed variety selection problem, we wanted to follow a structured analytical process. We found that INFORMS Certified Analytics Professional (CAP) framework of seven domain areas an ideal roadmap on how to understand the problem by eliciting intelligent questions from stakeholders and then develop a solution that is analytically valid and can be implemented in practice. According to INFORMS (2014), the analytical process can be decomposed into seven different yet co-dependent domain areas. This process is similar to the commonly known Cross-Industry Standard Process for Data Mining

(CRISP-DM)[1], both of which help align analytical solutions with business objectives to drive better business outcomes.

Figure 1 depicts INFORMS domain areas in sequential order, yet in practice one will likely revisit previous domains frequently. Essentially the process begins with defining the business problem, turning the problem into an analytical problem, and developing a mathematical solution that aligns with the original business problem. This is the strategy we followed to develop our solution to the Syngenta problem.

Figure 1: INFORMS seven analytics domain areas

In the remainder of this paper, we discuss the business problem in detail, transform the business problem into an analytics problem, discuss the data provided and relationships found, describe the methodologies we tested, and discuss the models developed. We provide R code that Syngenta could use to deploy our solution, which could be automated with just parameter recalibration for future experimental years.

Business Problem

Syngenta is a bio-tech company looking to use the power of analytics to identify the soybean varieties to commercialize for the year of 2015-16. To become a commercial variety each seed variety must pass through a series of "stage gates" performed annually. Each year the data from yield tests are analyzed, and a decision is made to either continue or discard. The final stage gate is the decision to commercialize which is also captured by graduation year. Syngenta continues testing on about 15 percent of the varieties tested in any given year as shown in **Figure 2.**

Figure 2: Syngenta stages [1]

Analytics Problem

[1] https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

In order to get an idea about the complexity of the given problem, we initially began by checking the correlation between YIELD and RM. While the description of the problem suggests that yield and RM share a significant positive correlation, we found the relationship to be weakly positive. The objective of the problem was to choose the top 15% elite soybean varieties that Syngenta has to commercialize to maximize their sales volume in the following year.

Data

The data set contained 258,254 entries, each representing one of 535 experiments done on 15,632 varieties. These experiments were done in one of 153 locations to control for variations in exogenous factors like soil quality and weather. **Table 1** provides a data dictionary of the variables provided as well as features generated in our analysis.

Variable	Definition
YEAR	the year that the experiment was conducted
EXPERIMENT	Experimental varieties of similar relative maturity are tested together in experiments. In the first year of yield testing, experiments often contain closely related experimental varieties, with the goal of selecting the best representatives of a family. In the second and third year of testing, varieties from different families are tested together to determine which varieties will be advanced to the next year or commercialized. In addition to the experimental varieties, designated "check" varieties are contained in the experiments for comparison.
LOCATION	Experiments are grown at many locations, depending on the stage of testing. Individual varieties may respond differently to different sets of environmental conditions. One of the reasons that varieties are tested over multiple years is to see how varieties will respond to a larger population of environments. For the purpose of this challenge, we are assuming that the yield trial locations are representative of the market that the varieties will be sold in. You may, however, find that some testing locations are more predictive than others as to the future performance of a variety.
VARIETY	the designation of the individual variety that is being evaluated in the experiment. From a botanical perspective, a variety is group of soybean plants that are genetically identical. They are selected for characteristics that are desirable to a grower (yield and agronomic traits). The seeds harvested from a soybean variety will be genetically the identical from one generation to the next.
FAMILY	identifies the "breeding population" from which a variety was derived. Members of a breeding population are highly related to each other since they are derived from the same parents. Many representatives from a breeding population are typically tested together every year with the goal of selecting the best representative of the population.
CHECK	commercial soybean varieties that are used as performance benchmarks in yield trials. Check varieties are typically elite commercial varieties that are used as benchmarks to measure experimental variety performance. Since the check varieties are already being sold, an experimental variety needs to outperform the check varieties to be considered to move to the next stage of testing. After an experimental variety graduates to commercial, it may become a check in the following years.
RM	Soybean Relative Maturity – Soybean varieties are affected by day length throughout the growing season. Day length triggers soybean plants to produce seed during the summer and to mature in the fall. Soybean varieties are assigned a relative maturity number (e.g. 2.5) which reflects differences in amount of time it takes individual varieties to reach physiological maturity. For example, a 2.5 RM variety matures relatively later than a 2.1 RM variety. Historical data show late maturing varieties have greater yields than early maturing varieties, so it is important to account for this effect.

REPNO	replication number. Soybean yield experiments are typically replicated. Data from the individual replicates are included in this dataset.
YIELD	the amount of grain per unit of land that a soybean variety produces. Grain yield in soybeans in the United States is measured in bushels per acre.
CLASS_OF	the final year that a soybean variety is tested prior to commercialization.
GRAD	varieties that graduate to commercialization following their final year of experimental evaluation.
BAGSOLD	the number of bags of seed sold in the second year after commercialization. High relative sales volume in the second year of sales is associated with the superiority of a variety relative to other choices in the marketplace.
YieldperRM (derived variable)	YIELD/RM, so that we have a better variable than just YIELD and RM

Table 1: Data dictionary

In the pre-processing phase we detected some outliers on RM < 1. We also removed the records where GRAD value was NULL. We only considered those records where GRAD was either YES or NO to train the proposed models.

Methodology Selection

Here we briefly discuss some of the methods we believed to be necessary in our data analysis.

First, Analysis of Variance (ANOVA), is a commonly used statistical method to detect difference between group means, which partitions the total variation in the response variable (SST) into either explained by different treatments (SSTR), or unexplained (SSE). Thus, if the variation triggered by treatment is relatively large as compared to that unexplained, we will be able to conclude that there is indeed a difference among groups greater than what would be realized by chance along (i.e. statistically significant). This is particularly useful in the factor screening process. We conducted ANOVA tests to prove statistical significance of our predictors.

Next, decision trees were used for predictive modeling purposes. A classification tree will be trained as the target variable CHECK is a categorical variable. Decision trees selects the most significant features on the basis of which we can divide the dataset into different groups. This is done so that the records with different behavior are contained in different data groups or clusters. We used decision trees to make the clusters of the data so that clusters have records which have similar behavior within each cluster.

Lastly, we implemented a target distribution heuristic. In the target distribution, we aggregated the data on the basis of variety and year of experiment. Then, we used this dataset to train and test our model. The distribution of BAGSOLD for years 2012 and 2013 can be seen in **Figure 3** and **Figure *4*** and later we show how our solution is acceptable in reference to these plots.

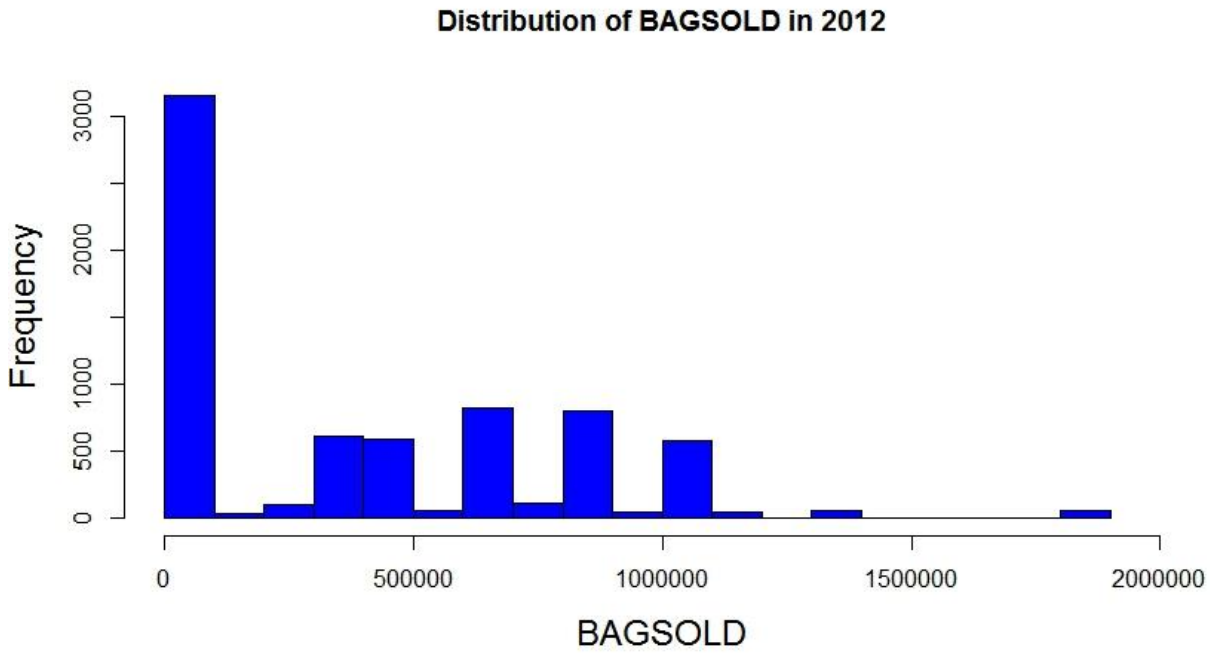

Figure 3: Histogram of BAGSOLD for those experimented in year 2012

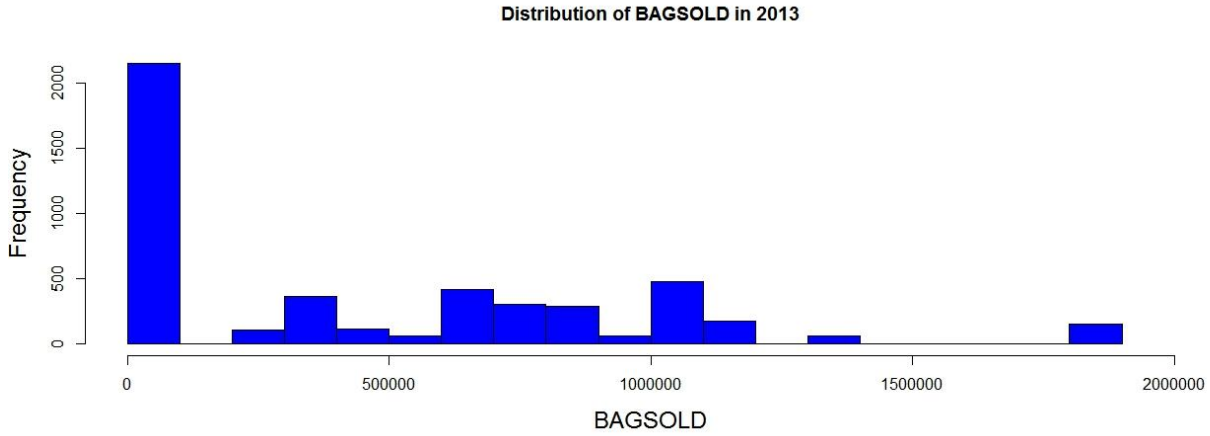

Figure 4: Histogram of BAGSOLD for those experimented in year 2013

Model Building

Variable Importance

To screen the factors, we run one-way ANOVA tests on the potential quantitative (and ordinal) predictors of which the result of Yield by Year stands out. As shown in the output (**Figure 5**), the F-score is 1207.895, yielding a p-value of 0.000 leading to a rejection of the null hypothesis that all means of yield are equal across different years. To prove our conclusion, the mean plot of yield demonstrates a clear upward trend after the year 2010, despite a subtle drop in 2012 (**Figure 6**).

YIELD

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	849126.607	5	169825.321	1207.895	.000
Within Groups	36308518.980	258247	140.596		
Total	37157645.590	258252			

Figure 5: F-test YEAR vs YIELD

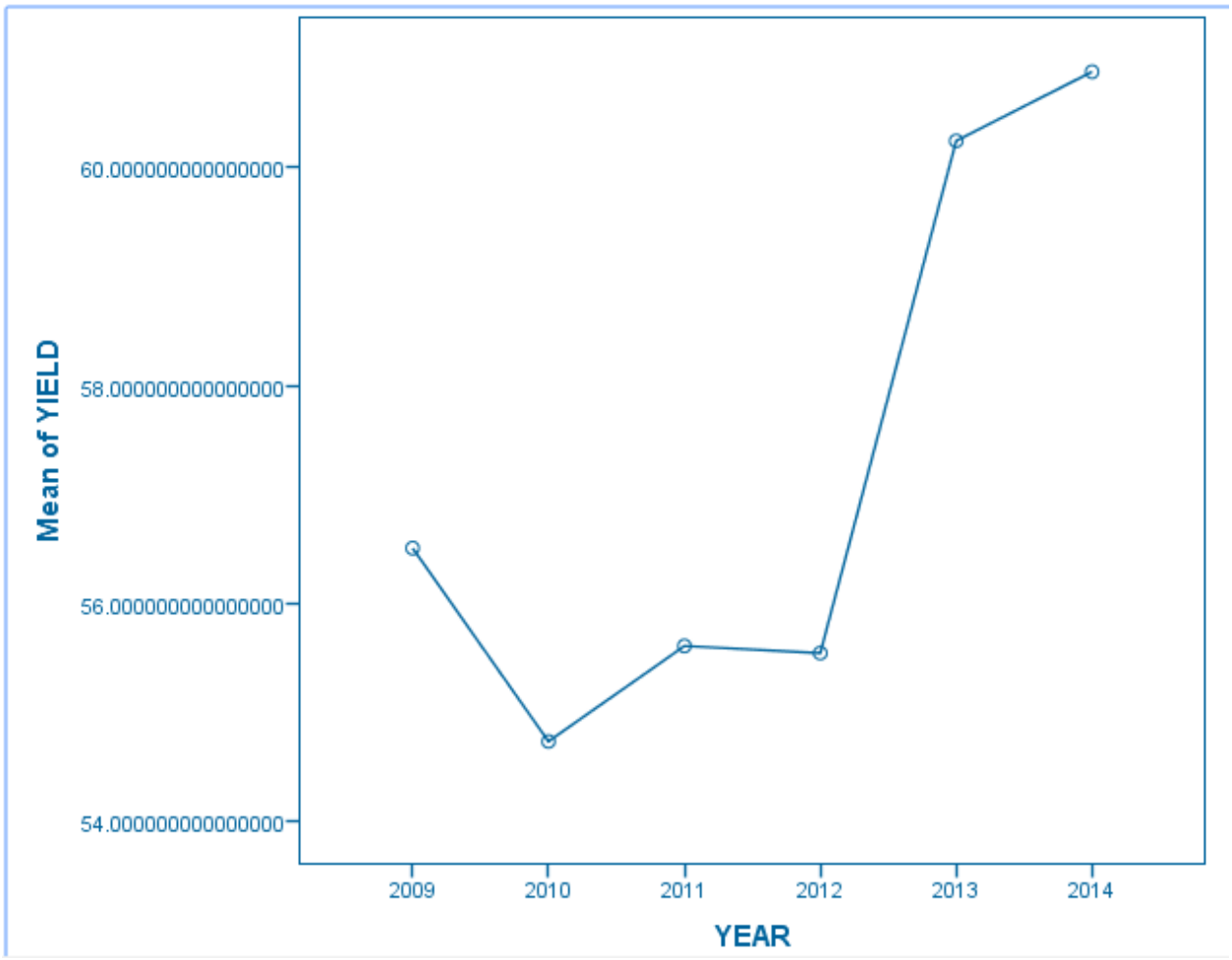

Figure 6: Plot of mean YIELD vs YEAR

We also conducted multiple comparisons using Tukey's Honest Significance Difference Test[2] to ensure valid results (**Figure 7**). We conclude that year contributes to the variation in yield, and will be included in the model.

[2] https://en.wikipedia.org/wiki/Tukey%27s_range_test

Multiple Comparisons

YIELD
Tukey HSD

(I) YEAR	(J) YEAR	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2009	2010	1.774971092655299*	.139613140055625	.000	1.377111777188278	2.172830408122321
	2011	.897940780026389*	.117891857831813	.000	.561981185983789	1.233900374068988
	2012	.964063463977603*	.119310242265509	.000	.624061861714039	1.304065066241166
	2013	-3.738808412200520*	.135489996783491	.000	-4.124917895530177	-3.352698928870863
	2014	-4.372664423266627*	.147034092094609	.000	-4.791671425098754	-3.953657421434501
2010	2009	-1.774971092655299*	.139613140055625	.000	-2.172830408122321	-1.377111777188278
	2011	-.877030312628911*	.091101897553728	.000	-1.136645837242366	-.617414788015456
	2012	-.810907628677697*	.092930078576896	.000	-1.075732969860979	-.546082287494415
	2013	-5.513779504855819*	.112953994277510	.000	-5.835667536488165	-5.191891473223473
	2014	-6.147635515921927*	.126570493532022	.000	-6.508326851242584	-5.786944180601269
2011	2009	-.897940780026389*	.117891857831813	.000	-1.233900374068988	-.561981185983789
	2010	.877030312628911*	.091101897553728	.000	.617414788015456	1.136645837242366
	2012	.066122683951214	.055160318809382	.838	-.091069158261487	.223314526163915
	2013	-4.636749192226908*	.084647894776471	.000	-4.877972571760334	-4.395525812693482
	2014	-5.270605203293016*	.102111464095783	.000	-5.561594985417372	-4.979615421168660
2012	2009	-.964063463977603*	.119310242265509	.000	-1.304065066241166	-.624061861714039
	2010	.810907628677697*	.092930078576896	.000	.546082287494415	1.075732969860979
	2011	-.066122683951214	.055160318809382	.838	-.223314526163915	.091069158261487
	2013	-4.702871876178122*	.086612411677003	.000	-4.949693592293317	-4.456050160062928
	2014	-5.336727887244230*	.103745818548038	.000	-5.632375133152766	-5.041080641335694
2013	2009	3.738808412200520*	.135489996783491	.000	3.352698928870863	4.124917895530177
	2010	5.513779504855819*	.112953994277510	.000	5.191891473223473	5.835667536488165
	2011	4.636749192226908*	.084647894776471	.000	4.395525812693482	4.877972571760334
	2012	4.702871876178122*	.086612411677003	.000	4.456050160062928	4.949693592293317
	2014	-.633856011066108*	.122007377584875	.000	-.981543712298991	-.286168309833224
2014	2009	4.372664423266627*	.147034092094609	.000	3.953657421434501	4.791671425098754
	2010	6.147635515921927*	.126570493532022	.000	5.786944180601269	6.508326851242584
	2011	5.270605203293016*	.102111464095783	.000	4.979615421168660	5.561594985417372
	2012	5.336727887244230*	.103745818548038	.000	5.041080641335694	5.632375133152766
	2013	.633856011066108*	.122007377584875	.000	.286168309833224	.981543712298991

*. The mean difference is significant at the 0.05 level.

Figure 7: Tukey's HSD result

A similar procedure is implemented for relative maturity. Although RM is a quantitative variable, SPSS software artificially categorized it by assigning numerous bins. As seen in the ANOVA output of Yield by RM (**Figure 8**), the F-score is 178.333, considerably less than that of Yield by Year yet remains statistically significant. Consequently, we decide that as least one mean of all RM groups is not equal to the rest. As seen further in the mean plot (**Figure 9**), suggests an upward trend before RM reaches roughly 3.7, which starts a fluctuating downfall in yield, in spite of a spike when RM is between 4.7 and 5.8.

YIELD

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1240356.257	50	24807.125	178.333	.000
Within Groups	35917289.330	258202	139.105		
Total	37157645.590	258252			

Figure 8: F-test for YIELD vs RM

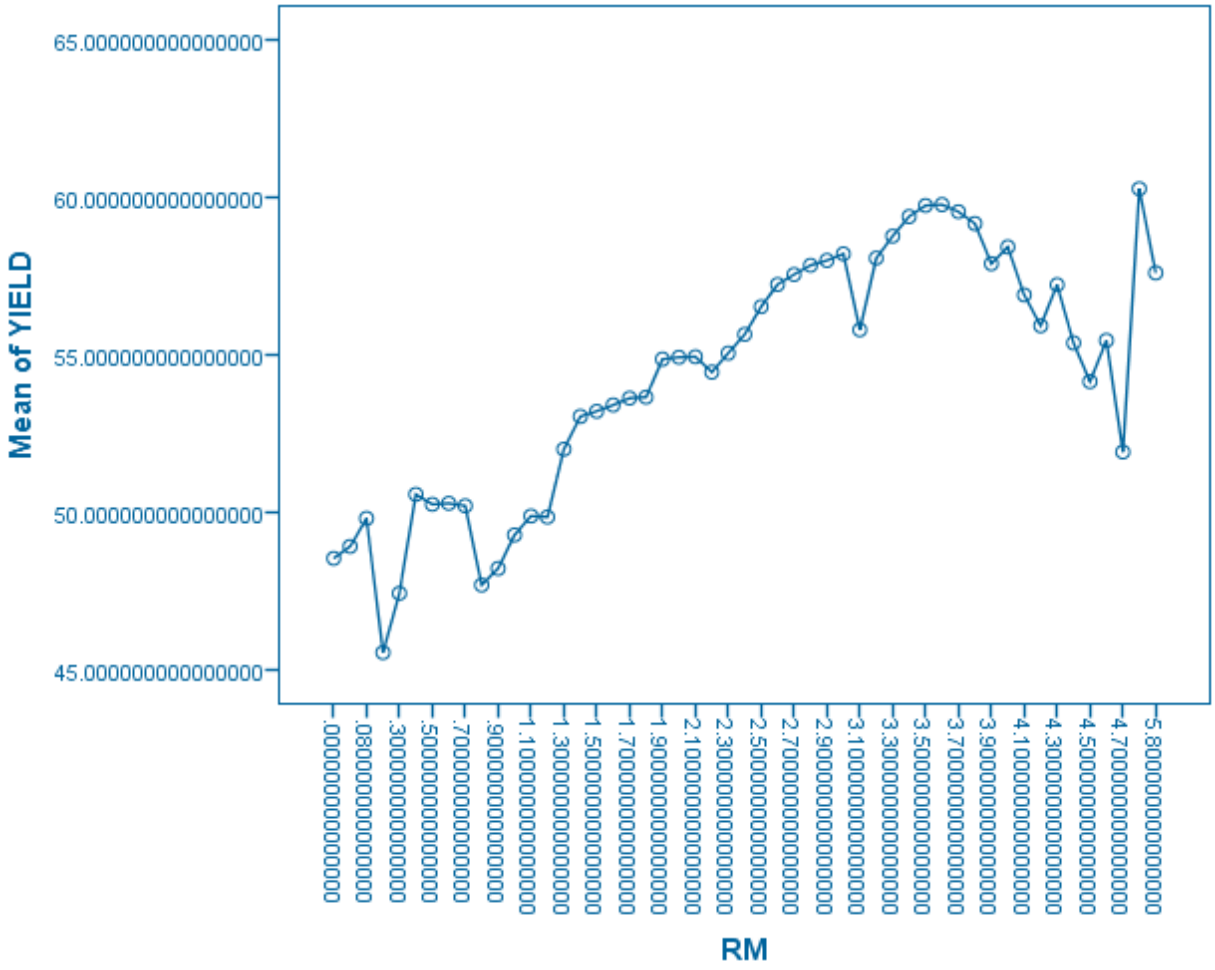

Figure 9: Plot of mean YIELD vs RM

Predictive Modeling

Model 1: (The model we tried but could not generate convincing results)

The unknown variable that we need to predict to get the Top 15% varieties is to predict the BAGSOLD for those who are from the class of 2014. As we have a target variable, this problem can be solved using supervised learning algorithms. We tried different nonlinear models using backward stepwise selection on the entire dataset of 258,254 records. We could not obtain quality fits for the entire dataset. This could have been due to the fact that there might be different

clusters that were behaving differently. Clustering the data into different clusters and then fitting appropriate models on those clusters separately seemed to be a good pathway.

The question which was posed to us now was that on what basis to cluster them. We thought of applying a classification tree with CHECK as our target variable. The CHECK variable here indicates which of the varieties are elite. But, before applying decision trees on the entire data set we needed to be sure that CHECK levels (TRUE and FALSE) are uniformly distributed among all the experimental years. We started to explore the distribution. We found that both the levels (TRUE and FALSE) were only present in experimental years 2012 and 2013 as shown in **Table 2** below.

	CHECK levels	
YEAR	FALSE	TRUE
2009	123	
2010	575	
2011	6421	
2012	3507	3597
2013	2288	2467
2014		2916

Table 2: The distribution CHECK levels across years.

Therefore, we were required to use the data of only experimental years 2012 and 2013 in order to train the models and come up with an authentic model that predicts the variable BAGSOLD with greater accuracy. We tried to first cluster the data using the decision tree. The tree showed RM as the most significant variable (**Figure 10**). Once we found the clusters, we now tried to fit models separately on these clusters. For this we split each cluster into training and testing datasets and then fitted non-linear models (with BAGSOLD as target and RM, YIELD and YEAR (dummy variables) as predictors), which yielded a respectable adjusted R-squared (~70%) on the training data, as well as testing dataset. However, when we tested these trained models on the dataset of experimental year 2014, we got unrealistic (distribution was not similar at all to previous years) results for the values of BAGSOLD.

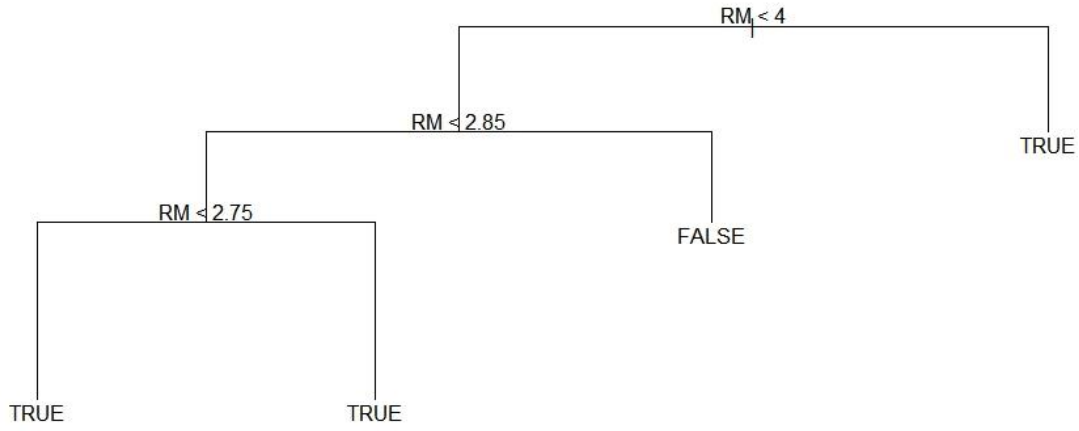

Figure 10: Decision tree with CHECK as the target variable

Model 2: (This model generated more conclusive results)
We started from scratch and we again explored the raw data with year on year trend of each variety. We aggregated basically two columns BAGSOLD and YIELD on the basis of VARIETY and YEAR to see the trend of only those varieties which graduated in the given time period of data. The YIELD and BAGSOLD showed pretty much similar trends over the years and the varieties as shown in **Figure 11**. We took sum aggregation of BAGSOLD and YIELD and average aggregation of RM on the basis of YEAR and VARIETY. We then applied a non-linear model using backward stepwise selection with aggregated value of BAGSOLD as target variable and other aggregated variables as predictors. This model overfit the data as we observed from the comparison of adjusted R-squared values among the training and testing data sets.

BAGSOLD							YIELD						
VARIETY	2009	2010	2011	2012	2013	2014	VARIETY	2009	2010	2011	2012	2013	2014
V103484			5,913,756	5,092,401			V103484			1,940	1,323		
V103620				3,096,533			V103620				492		
V103624			36,672,472				V103624			2,784			
V111237			62,296,311	58,211,307	175,655,172	59,232,558	V111237			3,512	3,295	9,945	3,538
V111336			14,766,552	124,695,328	93,521,496	51,682,932	V111336			1,166	8,685	7,441	4,085
V120038			222,390	163,086			V120038			780	524		
V120410				19,213,184			V120410				3,377		
V136868				52,873,594			V136868				8,204		
V137136			10,981,546				V137136			3,582			
V137240			12,593,171	12,355,564			V137240			3,246	2,820		
V148493	978,576	1,712,508	3,363,855	3,547,338			V148493	841	1,446	2,977	3,025		
V150834			27,213,310				V150834			2,776			
V150844			21,892,780	67,867,618	30,649,892		V150844			4,079	11,557	5,559	
V150847			18,357,732		30,492,504		V150847			3,635		5,635	
V150853			22,357,964	41,417,212	40,317,640	42,516,784	V150853			3,749	6,463	6,303	6,764
V151663			3,638,821				V151663			3,139			
V152053			11,282,208	48,889,568	56,411,040	176,754,592	V152053			767	2,896	3,769	11,367
V152079			10,861,960	73,318,230	80,106,955	243,036,355	V152079			500	3,410	3,883	11,271
V155820	8,491,368	37,149,735	441,551,136	496,745,028	195,301,464	66,869,523	V155820	502	2,115	24,277	27,956	11,100	3,848
V155842	3,805,026	20,927,643	198,495,523	454,700,607	149,664,356	113,516,609	V155842	366	1,975	18,017	41,777	13,783	10,555
V155843	5,228,046	28,754,253	371,545,215	519,319,236	104,560,920	110,660,307	V155843	394	2,057	63,737	34,233	7,123	7,467
V155853	3,045,568	25,125,936	46,444,912	40,353,776	140,857,520	91,367,040	V155853	251	1,896	3,575	3,198	11,036	7,503
V155918	2,451,942		23,293,449	241,516,287	47,812,869	26,154,048	V155918	391		3,507	34,052	7,764	4,025
V156247		14,920,496	104,443,472	108,173,596	289,084,610	110,038,658	V156247		494	3,318	3,491	8,805	3,527
V156305		5,316,256	41,200,984	70,440,392	121,609,356	78,414,776	V156305		526	3,738	6,398	10,963	7,230
V156314		4,530,936	46,064,516	43,799,048	91,373,876	91,373,876	V156314		367	3,796	3,690	7,672	7,490
V156368				61,180,652	209,426,078	158,834,385	V156368				2,972	11,496	8,530
V156553			7,141,428	33,326,664	35,707,140	113,072,610	V156553			783	3,548	3,817	11,778
V156642			3,572,700	15,183,975	17,268,050	40,192,875	V156642			696	2,955	3,667	7,632
V156774			8,029,980	40,952,898	46,573,884	51,391,872	V156774			668	2,943	4,012	4,203
V156786			3,167,461		14,973,452	16,413,207	V156786			651		3,015	3,412
V156797			3,139,930	17,897,601	18,839,580	21,351,524	V156797			645	3,783	3,726	4,213
V156806			12,464,400	56,089,800	124,644,000	123,605,300	V156806			705	3,301	7,712	7,652

Figure 11: Behavior of BAGSOLD and YIELD aggregated over years and Variety

To remedy model over-fitting, we tried some regularization techniques like Ridge regression and Lasso regression. Ridge regression did improve the result but it could not tackle the problem of over-fitting. On the other hand, Lasso improved the result as well as it solved the problem of over-fitting by driving the parameter coefficients all the way to zero, thus making the model less complex.

We then applied this Lasso model to the evaluation set which consists of only varieties from the class of 2014. We used their aggregated YIELD and RM (aggregated on the basis of VARIETY and for each YEAR) to predict the BAGSOLD for the following year of these 38 varieties. We took the predicted value for the latest year of experiment of every variety and extrapolated that value for the following year as a measure of potential sales volume of these varieties.

Results

The predicted potential sales volume for each variety in the evaluation set can be seen in **Table 3**.

CLASS_OF	VARIETY	FAMILY	RM	BAGSOLD
2014	V114655	FAM11247	2.5	1,885,261
2014	V114553	FAM06574	2.5	1,796,116
2014	V114649	FAM11251	2.5	1,722,934
2014	V114556	FAM13833	2.3	1,693,617
2014	V152322	FAM13828	3.1	1,012,118
2014	V114589	FAM11189	3.1	1,010,079
2014	V152306	FAM13804	3.2	995,892
2014	V114545	FAM01215	2.6	988,908
2014	V152300	FAM01223	3.4	986,674
2014	V152320	FAM11190	3.1	986,061

2014	V152440	FAM06776	3.9	983,577
2014	V114688	FAM06758	3.2	978,922
2014	V152415	FAM06784	3.5	972,922
2014	V113396	FAM05408	3.9	971,877
2014	V152325	FAM01215	3.4	971,124
2014	V152324	FAM01215	3.5	969,706
2014	V114685	FAM06758	3.0	956,324
2014	V114676	FAM11223	3.2	952,145
2014	V152334	FAM01223	3.6	941,047
2014	V114569	FAM06563	2.6	923,468
2014	V114565	FAM01215	3.0	922,083
2014	V140408	FAM01209	2.7	919,701
2014	V114530	FAM06560	2.7	916,572
2014	V114564	FAM01215	2.8	911,928
2014	V152312	FAM13852	2.8	906,345
2014	V114585	FAM06560	2.8	880,487
2014	V140364	FAM08217	2.0	855,586
2014	V114541	FAM06538	2.9	847,075
2014	V114538	FAM13851	2.8	839,303
2014	V151236	FAM13840	2.2	827,869
2014	V140393	FAM07868	2.1	804,164
2014	V140432	FAM13872	2.1	788,017
2014	V151161	FAM10485	1.9	778,135
2014	V151273	FAM10404	1.9	692,719
2014	V151284	FAM07855	1.9	626,773
2014	V104000	FAM10353	1.5	623,634
2014	V151283	FAM10354	1.8	606,035
2014	V152253	FAM01407	3.7	297,078

Table 3: Final Result: Predicted potential sales volume for each variety of class 2014.

The top 15 % varieties on the basis of potential sales volume are depicted in **Table 4.**

Top 15% Varieties
V114655
V114553
V114649
V114556
V152322
V114589

Table 4: List of top 15% varieties

We justify the performance of this model by examining the distribution of the BAGSOLD for the given data and compare it with distribution of the predicted data. **Figure 12** and **Figure 13** shows the distribution of BAGSOLD for years 2012 and 2013 respectively. **Figure 14** shows the distribution of predicted values of BAGSOLD which shows quite similar pattern as seen in **Figure 12** and **Figure 13**. As stated previously in the other models we tried we were not able to obtain a target distribution that resembled previous year's output. Since 2012 and 2013 were so similar we made the assumption that 2014 should be more or less in alignment with those years.

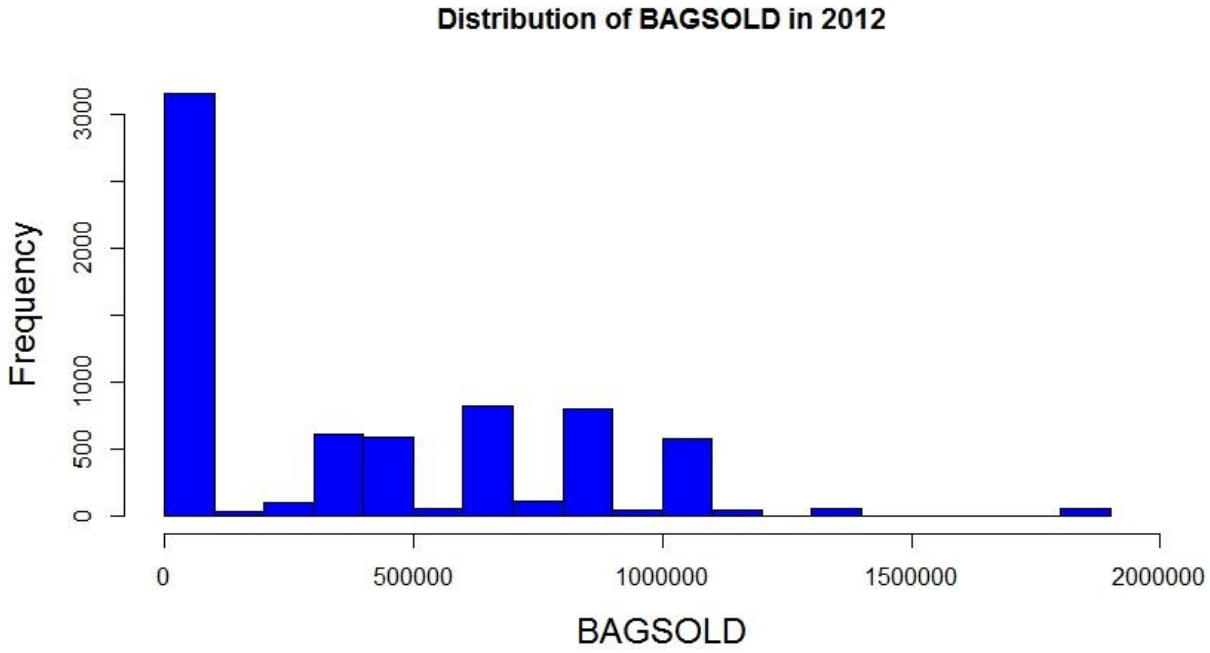

Figure 12: Histogram of BAGSOLD for those experimented in year 2012

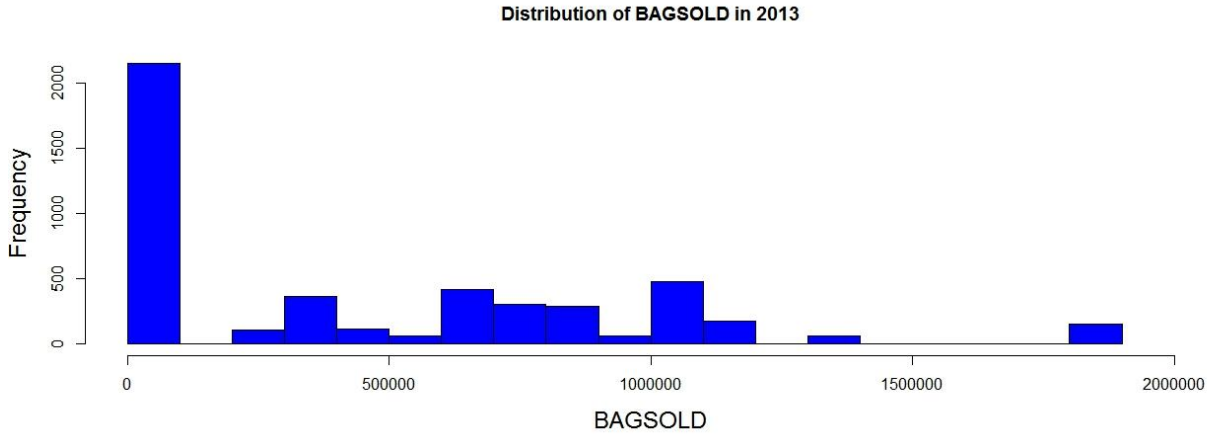

Figure 13: Histogram of BAGSOLD for those experimented in year 2013

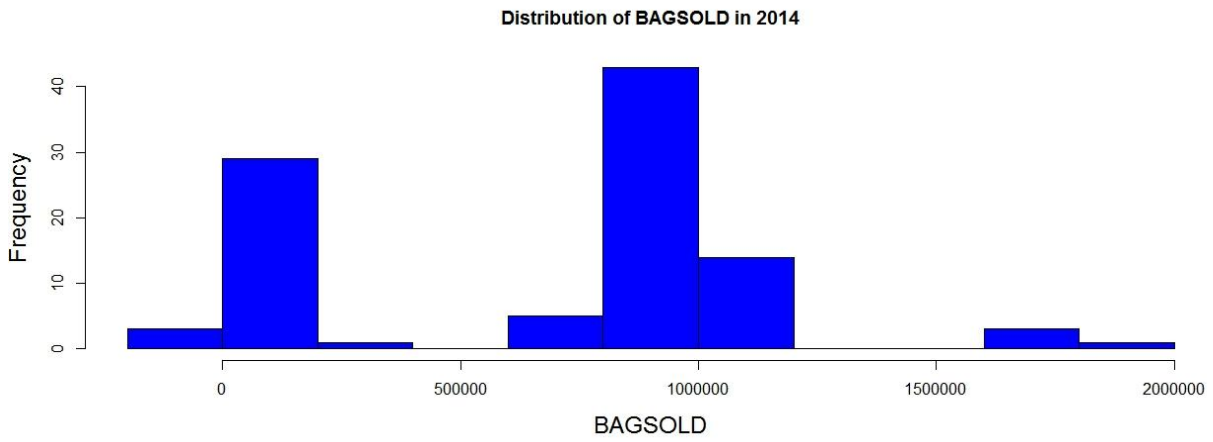

Figure 14: Histogram of BAGSOLD for those experimented in year 2014

Deployment

In practice, we would like to discuss our findings and recommendations with stakeholders to see if our assumptions of BAGSOLD for 2014 are indeed true with regard with previous years. We believe this domain expert feedback would help justify if our solution could indeed be deployed to support the business problem of identifying elite seed varieties going forward. If this solution is believed to be valid, the model can be deployed using the R-code we provided in the Appendix of this paper. Then just output results in a csv file containing the latest variety for each year.

Conclusions & Model Life Cycle Management

Identifying elite commercial varieties is a challenging analytics problem, but must be performed as best as possible by Syngenta to meet the expectations of their customers. We tested several models and identified one that we were most comfortable recommending to stakeholders for further discussion.

Our model generates results on the basis of year on year data of each variety. Model 1 gave very unrealistic results which may have been due to the repetition of each variety in multiple records. This is because each variety was experimented multiple times at different locations for different years. Moreover, the distribution of the response (BAGSOLD) did not seem realistic based upon the consistency of previous years

Model 2 on the other hand generated acceptable results as the training was performed on the aggregated dataset and the distribution of BAGSOLD did resemble a distribution that is in alignment with previous year. Thus, the final model we propose is Model 2.

Lastly, our model does not take into account the effect of family and location where the seed was grown. We cannot recognize which locations were good with yield and which were bad. Also, as new data is gathered year on year, the model will need to be re-estimated/re-calibrated. This is because we have trained our model on the recent past 2 years' data.

References

[1] Breene, K. (2016, January 18). Food security and why it matters. Retrieved January 29, 2017, from https://www.weforum.org/agenda/2016/01/food-security-and-why-it-matters/

[2] INFORMS. (2014). *INFORMS Certified Analytics Professional (CAP) Examination Study Guide*: www.informs.org.

[3] Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*: " O'Reilly Media, Inc.".

[4] INFORMS. *INFORMS O.R. & Analytics Student Team Competition: 2017 Problem*. 2016 [cited 2016 12/1/2016]; Available from: http://connect.informs.org/oratc/2017problem.

Appendix

R-code

```
############## Calling the libraries required #############
library(caret)
library(ISLR)

################# Input the Raw Data file #################
Experiment_data=read.csv("B:\\MS BAIM\\INFORMS OR case competition\\Experiment Data.csv"
                         ,head=TRUE,sep=","
                         ,colClasses = c("factor","factor","factor","factor"
                                         ,"factor","factor","numeric","numeric"
                                         ,"numeric","factor","factor","numeric"))

dim(Experiment_data)
################## Experimental data ######################
data_with_bagsold = Experiment_data[Experiment_data$BAGSOLD>0 & !is.na(Experiment_data$BAGSOLD),]
summary(data_with_bagsold)

############## Aggregating the Experimental data ################
YIELD = aggregate(YIELD ~ VARIETY + YEAR, data = data_with_bagsold, sum)
BAGSOLD = aggregate(BAGSOLD ~ VARIETY + YEAR, data = data_with_bagsold, sum)
RM = aggregate(RM ~ VARIETY + YEAR, data = data_with_bagsold, mean)
#CHECK = aggregate(CHECK ~ VARIETY + YEAR, data = data_with_bagsold, mean)

agg_data = merge(YIELD, BAGSOLD,by = c('VARIETY','YEAR'))
agg_data = merge(agg_data, RM,by = c('VARIETY','YEAR'))
head(agg_data)
summary(agg_data)

################# making dummy variables for Year ##################

dummies_YEAR = model.matrix(~agg_data$YEAR-1)
colnames(dummies_YEAR) = c('YEAR_2009','YEAR_2010','YEAR_2011','YEAR_2012'
                           ,'YEAR_2013','YEAR_2014')

agg_data=data.frame(agg_data,dummies_YEAR)
```

```
########### Splitting into the training and the test set ##########
inTrain=sample(nrow(agg_data),112*0.7, replace = FALSE)
train_with_bagsold=agg_data[inTrain,]
test_with_bagsold=agg_data[-inTrain,]

################ training the models ###################
linear.model=step(lm(BAGSOLD~YIELD+I(YIELD^2)+I(YIELD^3)+RM+I(RM^2)+I(RM^3)
                +YEAR_2012+YEAR_2013+YEAR_2014+YEAR_2009+YEAR_2010+YEAR_2011
                ,data = train_with_bagsold),direction='backward')
summary(linear.model)


validate_lm = defaultSummary(data=data.frame(obs=test_with_bagsold$BAGSOLD
                                ,pred=predict(linear.model,newdata=test_with_bagsold)))
validate_lm

###################### Ridgefit ########################

ctrl=trainControl(classProbs = FALSE,summaryFunction = defaultSummary)
Ridgefit=train(BAGSOLD~YIELD+I(YIELD^2)+I(YIELD^3)+RM+I(RM^2)+I(RM^3)
          +YEAR_2012+YEAR_2013+YEAR_2014+YEAR_2009+YEAR_2010+YEAR_2011
          ,data = train_with_bagsold
          ,method = 'foba'
          ,trControl = ctrl
          ,preProcess=c("center","scale")
          ,tuneLength = 16
          ,metric = 'RMSE')

Ridgefit

###################### Lassofit ########################

lassofit=train(BAGSOLD~YIELD+I(YIELD^2)+I(YIELD^3)+RM+I(RM^2)+I(RM^3)
          +YEAR_2012+YEAR_2013+YEAR_2014+YEAR_2009+YEAR_2010+YEAR_2011
          ,data = train_with_bagsold
          ,method = 'lars'
          ,trControl = ctrl
          ,preProcess=c("center","scale")
          ,tuneLength = 16
          ,metric = 'RMSE')
lassofit

validate_lm = defaultSummary(data=data.frame(obs=test_with_bagsold$BAGSOLD
                                ,pred=predict(lassofit,newdata=test_with_bagsold)))
validate_lm

pred = predict(Ridgefit,newdata = test_with_bagsold)
summary(pred)
hist(pred)

################ Preparing Evaluation set ###################
Evaluation_set = Experiment_data[Experiment_data$CLASS_OF=='2014',]
dim(Evaluation_set)
############## aggregating the Evaluation data ################
```

```
YIELD = aggregate(YIELD ~ VARIETY + YEAR, data = Evaluation_set, sum)
RM = aggregate(RM ~ VARIETY + YEAR, data = Evaluation_set, mean)

agg_eval_data = merge(YIELD, RM,by = c('VARIETY','YEAR'))
dim(agg_eval_data)
summary(agg_eval_data)
################ making dummy for Year ######################

dummies_YEAR = model.matrix(~agg_eval_data$YEAR-1)
colnames(dummies_YEAR) = c('YEAR_2009','YEAR_2010','YEAR_2011','YEAR_2012'
                           ,'YEAR_2013','YEAR_2014')

agg_eval_data=data.frame(agg_eval_data,dummies_YEAR)

############# Predicting using the appropriate model #############
############# and adding that to the column in the Evaluation set #########
pred = predict(lassofit,newdata = agg_eval_data)
agg_eval_data$BAGSOLD = predict(lassofit,newdata = agg_eval_data)

########## Output the final file with Predicted values on the Evaluation set ######
write.csv(agg_eval_data,file = 'B:\\MS BAIM\\INFORMS OR case competition\\Evaluation set with Bagsold.csv'
          ,row.names = FALSE)
```