SNP-based soybean variety selection

Damir Jajetic
IN2
Zagreb, Croatia
damir.jajetic@in2.hr

ABSTRACT

One of the main tools in soybean variety selection is analyzing data from yield testing. While most of the tested varieties never come to commercialization, some commercialized varieties are tested in the small number of testing fields. Current Marker-assisted selection (MAS) strategies are focused on major quantitative trait locus (QTL), whereas yield has the complex inheritance that is controlled by multiple QTL with minor effect on phenotype [14]. With the machine learning, yield potential can be efficiently predicted using high dimensional single nucleotide polymorphisms (SNP) markers of varieties. Therefore predicted yield potential could enable better allocation of varieties in testing fields, reducing the number of commercialization of unsuccessful varieties and reducing costs of production.

This paper will examine data from Syngenta AI Challenge [4, 9] and the effectiveness of Gradient Boosted Regression Trees for predicting yield potential from SNP markers.

CCS CONCEPTS

• **Applied computing** → **Computational biology**; • **Computing methodologies** → *Machine learning approaches*;

KEYWORDS

Gradient Boosted Regression Trees, variety selection, soybean, SNP

ACM Reference format:
Damir Jajetic. 2017. SNP-based soybean variety selection. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining - The Data Science for Intelligent Food, Energy and Water workshop, Halifax, Nova Scotia - Canada, August 2017 (KDD DSIFEr 2017),* 4 pages.
https://doi.org/TBD

1 INTRODUCTION

Sustainable usage of limited natural resources is one of the biggest challenges of today. It is expected that human population will reach 9.7 billion by 2050 [21, 22] , while usage of planet resources is reaching its capacity [2, 3, 7, 18]. The important part of the solution is creating more food per acre, and Syngenta created AI challenge to focus on this goal. Challenge and data are described in-depth on codalab [4] and idea connection [9] web pages, and here is a brief summary. From data given in 3 stages, participants should predict which soybean variety will perform best in 2015 and 2016. Participants got access to training data for varieties with known properties from past years and data for testing and benchmark varieties in the year 2012-2014. In stage 1 there are 1093 test varieties, and only 136 of them have corresponding SNP markers. Varieties from all other stages and all 32 varieties that belong to the class of 2014 have SNP markers available. For the examined model, that bias in data represents data leak, and all data is validated only on the subset of varieties with SNP data. In the training set there are 1169 different varieties and 46804 samples with SNP data out of 148129 samples total. For all experiments, corresponding data with geographic data is given with soil and weather conditions. Actual yield of variety is the result of their own genetic potential and weather and soil conditions.

There are several approaches to this problem, most notable being reinforcement learning [12, 15, 19, 20, 23] and supervised learning. Examined model use supervised learning approach for predicting yield potential directly from SNP markers, with distinguished benefit of the ability to use the model prior any testing and for better allocation of varieties in testing field. Model is based on machine learning gradient boosting regression algorithm [1, 5, 6] from scikit-learn package[17] while using only genetic data for training. Source code and detail description of model pipeline are available at corresponding Codalab worksheet page [10].

2 SOYBEAN VARIETY SELECTION

For predicting elite varieties and actual yield, machine learning regression model is used, trained only on training genetic data from varieties as features and yield as target variable. Genetic data in the provided dataset consists of 2163 SNP markers with 12 distinct base-pairs (AA, AC, AG, AT, CC, CG, DD, DI, GG, II, NN, TT). The meaning of base-pairs is available at the Nomenclature for incompletely specified bases in nucleic acid sequences [16]. Values are transformed to distinct numerical values, however still categorical. Rather than translating categorical values to binary, the model is based on a decision tree method that in practice can handle categorical variables without further encoding.

For the model, gradient boosting from scikit-learn package [17] is used. Detail explanation of tree boosting is out of the scope of this document and can be found in referenced materials [1, 5, 6, 17]. However, some characteristics important for variety selection should be recognized. For yield prediction of elite varieties downside of regression tree is that each leaf of tree predicts the value that has the least error on given metrics for all data in particular leaf from the training dataset. Consequently, predicted values are within range of target variable (yield) in training dataset. If we had a dataset where expected value for test data, fall within values from training data, this would not be an issue. In this particular business scenario, we are interested in the varieties that perform better than any varieties in training data. Hence this kind of method cannot

KDD DSIFEr 2017, August 2017, Halifax, Nova Scotia - Canada

ACM ISBN TBD…$TBD
https://doi.org/TBD

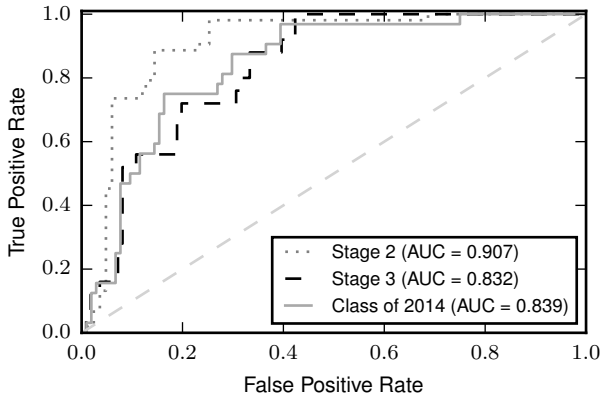

Figure 1: ROC curves for model predictions of varieties with genetic data

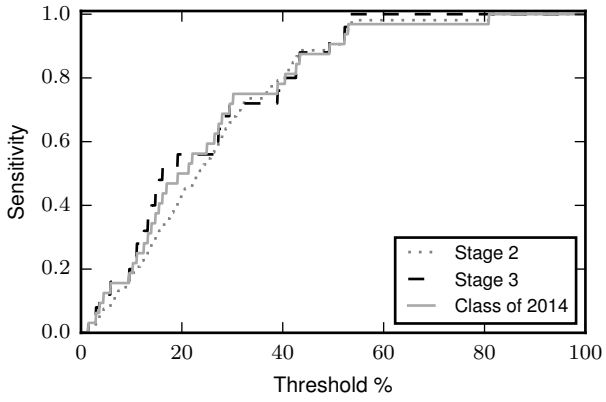

Figure 2: Sensitivity diagram for model predictions for varieties with provided genetic data

provide suitable estimation for exact yield potential. However, the model can be suitable in predicting ranking of yield potentials, from best to worst.

All 2163 SNP markers are used in the model and this is top 10 markers X220, X1106, X6008, X4353, X5192, X5193, X11, X4939, X218, X737. The model did not show any marker that significantly stands out in feature importances. Model is tested on predicting best varieties, where varieties with higher predicted yield potential are considered better than varieties with predicted lower yield potential.

With predicted yield as model result, area under the receiver operating characteristic curve is calculated. ROC AUC is binary classification metrics for relative probabilities, where 1 is the perfect score, 0 is the perfect score for inverse sorted probabilities, and 0.5 is a random guess. In this case, it measures if the model predicts higher yield potential for selected varieties than varieties that are not selected in next stage. In more detail, RAC and AUC are explained in scikit-learn [17] article on classification metrics and paper from Hanley and McNeil [8]. As we expect that predicted yield is not accurate on absolute values, an advantage for this metrics is that ROC AUC is not sensitive to actual values. As the aim of the model is to estimate best varieties for testing field allocation regardless of actual values, actual values will be measured on testing fields. With that metrics, we can measure how accurately model predicts if variety will be selected for stage 2, stage 3 or in the class of 2014.

Model predictions are validated on intermediate stages as predictions of the relative probability of selecting variety in next stage for stage 2, stage 3, and for being part of the class of 2014. Predicted yield is used as the probability that variety is elite, meaning variety with higher yield potential predicted has more probability to be elite variety than the variety with lower yield potential predicted.

In challenge data, out of 1093 total test varieties, only 136 have genetic data, and none of the varieties without genetic data proceeded to later stages. Besides qualifying that as data leak, since the essence of the model is SNP data, we cannot rely on model validation for data subset without SNP data itself. On the other

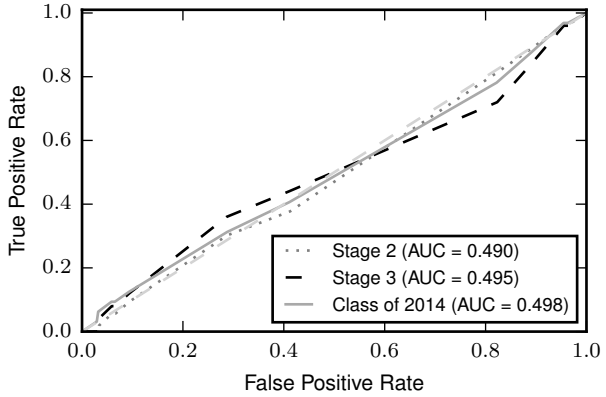

Figure 3: ROC curves for number of trials of varieties in Stage 1 vs. selected Stage 2, Stage 3, and the class of 2014 varieties

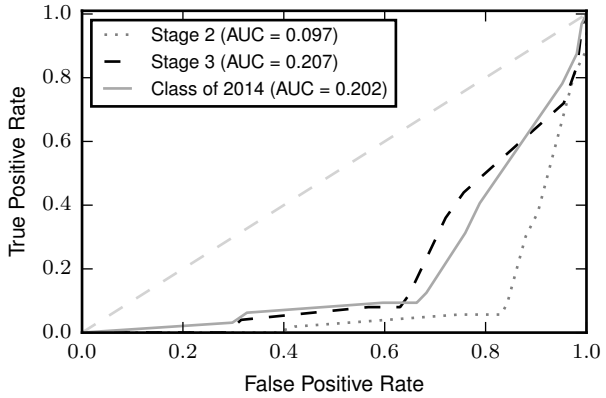

Figure 4: ROC curves for number of trials of varieties in Stage 1 with available genetic data vs. selected Stage 2, Stage 3, and the class of 2014 varieties

hand, the subset of varieties with SNP data is the cluster of good varieties in the set of all varieties.

In calculation metrics, to avoid data leak problems, only varieties that have genetic data are used for calculation. It should be recognized that we are at risk of underestimation of the model precision since in that case, we are trying to distinguish best from the set of good varieties. Actual model precision can be calculated only if genetic data was provided for all candidate varieties. The area under a receiver operating characteristic curve is calculated for model predictions, where for ground truth (values for "1") are used selected varieties in stage 2, stage 3 and varieties that belong to the class of 2014, for all varieties with SNP data, with results of 0.9068, 0.8324, 0.8392 respectively (Figure 1).

For the purpose of allocating varieties in testing fields, sensitivity diagram is created. This diagram is useful if we want to use model for testing only best varieties from the model, to estimate Type II error. In figure 2 we can examine if we choose to test only varieties with high yield from the model, how many varieties that are actually selected for later stages is covered with the model. This diagram is also created only for a subset of varieties with genetic data, with the same risk of underestimation. However, calculated results are still competitive and usable for proposed process of variety allocation in testing fields.

With the same method, we can compare an actual number of tests in stage I, with the fact if particular variety passes stage gates. If varieties that are more tested in Stage I, are part of later stages, ROC AUC value will be close to 1. Furthermore, if varieties that are not promoted to later stages are tested more, AUC value will be close to zero. If there is no connection between number of tested varieties in stage 1 and promotion to later stages, AUC value will be close to 0.5. If we create this diagram for all varieties, we can detect no pattern (Figure 3). However, for varieties with genetic data, we can recognize that varieties that are not promoted are tested more in testing fields in stage 1 (Figure 4). ROC AUC values are close to zero, that means high precision on negative probabilities. Train data from previous years were not cross-validated in the same way because of the smaller initial set of varieties.

As created model for yield estimation uses only genetic data of variety as features, it can be utilized before stage 1 to adjust the number of trials for varieties according to that estimation. Stage I testing fields can be allocated in a way that varieties with lower yield estimation are tested in fewer fields, yet enough trials to avoid Type II errors. Varieties with higher yield estimation should be tested with more trials. Repercussion could be more varieties selected for next stage and more Type I errors, however only for that stage of selection process. After the first stage, even with a small sample, most varieties could be discarded as varieties without yield potential. With cost saving of stage 1, in stage 2 more trials could be accomplished to reduce most Type I errors from the first stage. That way overall cost of testing would be lower, and varieties that will graduate will have more trials through all stages, that could lead to less of Type I errors. In dataset from Syngenta AI challenge, we can examine pattern that within varieties with known genetic data, varieties that are not selected to next stages are tested more in stage 1 than varieties that are selected to next stages (Figure 4). It is not disclosed to public if this is intentional by the business process.

For challenge purposes only, some additional estimations for varieties are made using predicted yield potential. Beside tested varieties, the model has also created the prediction for yield potential of benchmark (check) varieties. As we are expecting that elite varieties will have higher yield potential than benchmark varieties from previous years, estimations are created with following criterion. Tested varieties for which are predicted better yield potential than the mean yield potential of all benchmark varieties are inserted in the candidate list. Each variety in candidate list that is not in the class of 2014 is marked as Type II error estimate. Each variety that is in the class of 2014 and is not in the candidate list is marked as Type I estimate. Each variety that is both in candidate list and part of the class of 2014 is marked as "elite" variety.

Estimation of the model is that false positive (Type I errors) represents this 5 varieties V151236, V140432, V140364,V140393, V151283; False negative (Type II error) represents varieties with this top 5 V114663, V152414, V114687, V152313, V152321; and top 5 varieties that would perform best in farmers fields in 2016 are V152324, V114685, V152312, V152334, V152325

3 CONCLUSION

Analyzed data reveals the effectiveness of using Gradient Boosted Regression Trees for predicting ranking of yield potential of varieties from SNP data prior to yield testing. This information can be used very early in the breeding and selection process for cost reduction of variety selection process and better quality of selected varieties by adjusting the number of trials for each variety in testing fields. Although the examined model is focused only on yield potential, this is typically the first trait examined when selecting soybean varieties [11], while some other qualities do not need exhausting testing in many testing fields [13]. As data were analyzed only on the limited dataset, the model should be additionally tested and adjusted for production purposes.

REFERENCES

[1] L. Breiman, J.H. Friedman, A. Olshen, and Stone C.J. 1984. *Classification and regression trees.* Wadsworth.

[2] James H. Brown, William R. Burnside, Ana D. Davidson, John R Delong, William C. Dunn, Marcus J. Hamilton, Norman Mercado-Silva, Jeffrey C. Nekola, Jordan G. Okie, William H. Woodruff, and Wenyun Zuo. 2011. 61, 1 (2011), 19–26.

[3] Joseph R. Burger, Craig D. Allen, James H. Brown, William R. Burnside, Ana D. Davidson, Trevor S. Fristoe, Marcus J. Hamilton, Norman Mercado-Silva, Jeffrey C. Nekola, Jordan G. Okie, and Wenyun Zuo. 2012. The Macroecology of Sustainability. *PLOS Biology* 10, 6 (06 2012), 1–7. https://doi.org/10.1371/journal.pbio.1001345

[4] Codalab. 2017. Syngenta AI Challenge - Harness data to help feed our rising population. (2017). https://competitions.codalab.org/competitions/16194

[5] Jerome H. Friedman. 2000. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29 (2000), 1189–1232.

[6] Jerome H. Friedman. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38, 4 (28 Feb. 2002), 367–378. http://www.sciencedirect.com/science/article/B6V8V-451NMK5-2/1/25f688e042a2d32cfed9da4d20ebbd35

[7] Robert Goodland. 1995. The Concept of Environmental Sustainability. 26 (1995), 1–24.

[8] J A Hanley and B J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 1 (April 1982), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

[9] IdeaConnection. 2017. Harness data to help feed our rising population. (2017). https://www.ideaconnection.com/Syngenta-AI-Challenge/challenge.php

[10] Damir Jajetic. 2017. SNP-based soybean variety selection model. (2017). https://worksheets.codalab.org/worksheets/0x1aacb50a4e954e72b72bde6ae4acda5d/

[11] Chad Lee, Carrie Ann Knott, and Edwin L. Ritchey. 2014. Soybean Variety Selection. *Agriculture and Natural Resources Publications* 133 (2014).

[12] Gunar E. Liepins, Mike R. Hilliard, Mark R. Palmer, and Gita Rangarajan. 1991. Credit assignment and discovery in classifier systems. *Int. J. Intell. Syst.* 6, 1

(1991), 55–69. http://dblp.uni-trier.de/db/journals/ijis/ijis6.html#LiepinsHPR91

[13] Lei Ma, Bin Li, Fenxia Han, Shurong Yan, Lianzheng Wang, and Junming Sun. 2015. Evaluation of the chemical quality traits of soybean seeds, as related to sensory attributes of soymilk. *Food Chemistry* 173 (2015), 694 – 701. https://doi.org/10.1016/j.foodchem.2014.10.096

[14] J. Mammadov, R. Aggarwal, R. Buyyarapu, and S. Kumpatla. 2012. SNP markers and their impact on plant breeding. *Int J Plant Genomics* 2012 (2012), 728398. https://doi.org/10.1155/2012/728398

[15] Marvin Minsky. 1963. Steps Toward Artificial Intelligence. In *Computers and Thought*, Edward A. Feigenbaum and Jerome A. Feldman (Eds.). McGraw-Hill, New York, 406–450.

[16] Nomenclature Committee of the International Union of Biochemistry (NC-IUB). 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. 150 (1985), 1–5.

[17] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, Oct (2011), 2825–2830.

[18] J. Rockström, W. Steffen, K. Noone, Å. Persson, F.S. Chapin, E.F. Lambin, T.M. Lenton, M. Scheffer, C. Folke, H.J. Schellnhuber, et al. 2009. A safe operating space for humanity. *Nature* 461, 7263 (2009), 472–475. http://scholar.google.co.uk/scholar.bib?q=info:RmvzB-5al1UJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&ct=citation&cd=1

[19] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323, 6088 (Oct. 1986), 533–536. http://dx.doi.org/10.1038/323533a0

[20] Richard Sutton. 1984. *Temporal Credit Assignment in Reinforcement Learning*. Ph.D. Dissertation. University of Massachusetts, Amherst, MA.

[21] Department of Economic United Nations and Population Division Social Affairs. 2015. World Population Prospects: The 2015 Revision, Key Findings and Advance Tables. (2015).

[22] Department of Economic United Nations and Population Division Social Affairs. 2015. World Population Prospects: The 2015 Revision, Methodology of the United Nations Population Estimates and Projections. (2015).

[23] T. H. Westerdale. 1998. An approach to credit assignment in classifier systems. *Complexity* 4, 2 (1998), 49–52. http://dblp.uni-trier.de/db/journals/complexity/complexity4.html#Westerdale98