Predicting Crop Yield and Disease Trends in New Jersey Farms

Lauren Johnston, B.S.E Computer Science

Princeton University

ABSTRACT

Specialized farming technology is on the rise but efforts to accurately predict crop yield and identify crop disease are still in their infancy. In this project, I aimed to find a solution to both problems with a caveat: I wanted to ensure that yield prediction was cost-effective. I studied tomato yield at three New Jersey farms and attempted to predict yield based off of common crop measurements like normalized difference vegetation index (NDVI), evapotranspiration, and growing degree days (GDD). I found that elastic net was the best regression method due to both its feature selection and accuracy (R^2 = .39). In the second phase of my experiment, in which I sought to distinguish different types of crop disease from drone images, my results indicated that further data collection was needed to create a more even distribution of labels and allow for better model training.

METHODS

Part 1: Yield Prediction

For the tomato yield prediction phase of my project, I decided to collect NDVI, evapotranspiration, GDD measures, and other crop health measurements from local farms. These crop health measurements would act as descriptive features to predict yield, the target feature. To create a yield prediction model, I hoped to obtain cumulative tomato yield data from three different New Jersey farms collected over the same time interval to train the model. Training the model entails splitting the instances by day into a training and test set and then applying appropriate regression methods to the training set before testing on the test set.

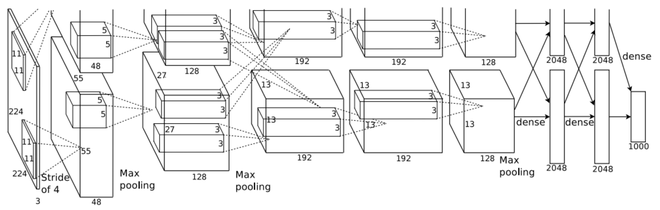
I decided to attempt canonical machine learning regression methods such as ridge, Lasso, and elastic net. Given that these methods are known for their feature selection abilities, I hoped to use them to generate model coefficient sets that eliminated less predictive factors. Through analysis of the efficacy of the combined NDVI, evapotranspiration, and GDD features I endeavored to determine an improved method for forecasting crop yield. I believed this approach would be successful with robust data collection over the growing season, the placement of sensors to collect crop health metrics directly in tomato fields at each farm in order to directly link the data to the yield numbers, and the use of proven machine learning methods in order to create a sparse prediction model. I used the following Sci-kit learn functions for yield prediction[14]:

- ***Ridge regression:*** Ridge(alpha=.5)
- ***Lasso regression:*** Lasso(alpha = 0.01)
- ***Lasso with 10 rounds of cross validation:*** LassoCV(cv=10, random_state=0)
- ***Lasso computed via LARS:*** LassoLars(alpha=0.01)
- ***Elastic net:*** ElasticNet(alpha=0.01, l1_ratio=0.7)

Part 2: Crop Disease Classification

For the crop disease classification phase of my project, the AlexNet CNN model was the most appropriate for the task due to its success in unsupervised image classification tasks [10]. AlexNet, invented by Krizhevsky et al. in 2012, was initially trained on the LSVRC-2010 ImageNet photo set and was both efficient and low in error.

AlexNet's architecture built upon previous CNN designs while adding certain unique modifications. The AlexNet network consists of eight layers as shown below, taken directly from Krizhevsky et al. [10].



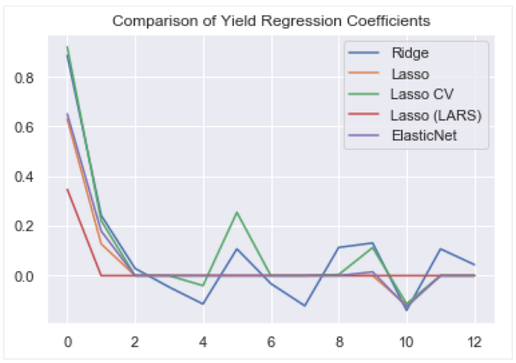
RESULTS

Feature	Ridge	Lasso	Lasso CV	Lasso LARS	Elastic net
GDD (cumulative)	0.885973	0.629643	0.920113	0.347498	0.650474
Chlorophyll index	0.242645	0.127821	0.221249	0	0.179403
Evapotranspiration	0.029076	0	0	0	0
GDD for day	-0.045835	0	0	0	0
LfAirDelta	-0.113918	0	-0.040104	0	0
NDVI	0.107615	0	0.254656	0	0
shortwave DW radiation	-0.03192	0	0	0	0
maxT	-0.121239	0	0	0	0
meanT	0.113613	0	0.004334	0	0
minT	0.131308	0	0.112843	0	0.015313
sea level pressure	-0.139502	-0.117789	-0.113361	0	-0.125171
crop coefficient Kc	0.107623	0	0	0	0
crop evapotranspiration	0.044176	0	0	0	0

Table 2: Regression coefficients for Arable data features by model.

Model	R^2 Score	Time (train)	Time (predict)
Ridge	0.403	0.0016	0.0002
Lasso	0.386	0.0013	0.0002
Lasso CV=10	0.417	0.1156	0.0001
LARS Lasso	0.218	0.0038	0.0002
Elastic Net	0.390	0.0012	0.0007

Table 3: Accuracy and speed tests for yield regression models (with normalized yield).



GDD Coefficient	Ridge	Lasso	Lasso CV	Lasso LARS	Elastic Net
GDD Max/Min Clamp	-0.021	0	0	0	0
GDD Mean Clamp	-0.021	0	0	0	0
GDD Max/Min Clamp (Cumulative)	0.224	0.266	0.253	0	0.148
GDD Mean Clamp (Cumulative)	0.224	0.000	0.000	0.048	0.146
GDD (Cumulative)	0.765	0.618	0.847	0.400	0.607
GDD	-0.021	0	0	0	0

Table 4: GDD coefficients by method of GDD calculation.

APPROACH

Part 1: Yield Prediction

- **Install Arable Mark sensors** in New Jersey farms during June - October 2018
  - Capture NDVI, temperature, GDD, upwelling/downwelling radiation, and other metrics (measured every 5 minutes)
- **Collect yield numbers** from farmers, weigh crates of produce, test soil composition, count/analyze insects on crops
- **Compare tomato yield** at **three farms** that grow standard-size tomatoes

Part 2: Crop Disease Classification

- **Equipment:** DJI Phantom 4 Pro, modified with dual RGB and Near Infrared (NIR) camera and Sentera Field Agent software
- **Approach:** Use Sentera Field Agent App to take series of images along paths of farm fields
- Drone images collected at 400ft altitude using the Digital Surface Model (DSM)/Ortho setting with 80% overlap
- **Crops of interest:** corn and soybean

SUMMARY

From the tomato yield prediction phase of my experiment, I determined that Lasso regression with cross validation, ridge regression, and elastic net regression generate the most accurate yield models. Of these three, Lasso with cross validation and elastic net are the most promising due to the sparsity of their solutions. However, given that the highest R^2 score was 0.417, there likely needs to be further data collection and exploration of new crop health metrics to create a more accurate yield prediction method. These results, despite their limitations, have still generated interesting feature subsets that may be used to reduce farming technology in future. In particular, Table 2 indicates that cumulative GDD and Chlorophyll index are important while NDVI eliminated in three out of five models. Furthermore, by testing two alternative clamping methods of GDD I was able to show that with the current data it is unclear if there is any benefit to clamping GDD methods over traditional GDD calculations.

With regards to the crop disease classification portion of my experiment, I found that AlexNet had a prediction accuracy of 56.1% across all classes. I concluded that this was likely due to the uneven distribution of class labels in the training data. I also found that there was no significant decrease in BCE loss over epochs during model training.

Given the opportunity to further develop this project, I would consider performing image segmentation before training my model with AlexNet. Furthermore, I would make use of the NIR images collected and add them as an additional input dimension to the model. This could potentially act as well as boundary segmentation due to the high contrast of NIR that showcase areas of healthy vs. unhealthy crops. Additionally, I would reconsider the class labels that I chose for the model. It was difficult to differentiate between crop disease and crop decay, so often I simply ended up classifying images as both. Part of the issue in differentiating classes was due to the fact that images were taken from the high altitude of 400ft. If I were to collect more data, I could consider flying at a much lower altitude (around 100-200ft.) so that images contained only 1 or 2 classes rather than most of them.

REFERENCES

[1] Arable, "Exploring evapotranspiration," Feb 2019. [Online]. Available: https://www.arable.com/2018/04/06/exploring-evapotranspiration/

[2] T. N. Carlson and D. A. Ripley, "On the relation between NDVI, fractional vegetation cover, and leaf area index," *Remote sensing of Environment*, vol. 62, no. 3, pp. 241–252, 1997.

[3] J. Dash and P. Curran, "The meris terrestrial chlorophyll index," 2004.

[4] I. de Castro, Ana, J. Torres-Sanchez, M. Peña, Jose, F. Jimenez-Brenes, O. Csillik, and F. López-Granados, "An automatic random forest-OBIA algorithm for early weed mapping between and within crop rows using uav imagery," *Remote Sensing*, vol. 10, no. 2, p. 285, 2018. [Online]. Available: https://search-proquest-com.ezproxy.princeton.edu/docview/2014750644?accountid=13314

[5] R. Fortes, M. H. Prieto, J. M. Terrón, J. Blanco, S. Millán, and C. Campillo, "Using apparent electric conductivity and NDVI measurements for yield estimation of processing tomato crop," *Transactions of the ASABE*, vol. 57, no. 3, pp. 827–835, 2014.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[7] H. Huang, J. Deng, Y. Lan, A. Yang, X. Deng, and L. Zhang, "A fully convolutional network for weed mapping of unmanned aerial vehicle (UAV) imagery," *PLoS One*, vol. 13, no. 4, 04 2018. [Online]. Available: https://search-proquest-com.ezproxy.princeton.edu/docview/2031413122?accountid=13314

[8] L. Johnston, "Predicting crop yield and disease trends in Central New Jersey farms," Feb 2019.

[9] M. Koller and S. Upadhyaya, "Prediction of processing tomato yield using a crop growth model and remotely sensed aerial images," *Transactions of the ASAE*, vol. 48, no. 6, pp. 2335–2341, 2005.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[11] G. S. McMaster and W. Wilhelm, "Growing degree-days: one equation, two interpretations," *Agricultural and forest meteorology*, vol. 87, no. 4, pp. 291–300, 1997.

[12] K. P. Murphy, *Machine Learning : A Probabilistic Perspective*. Cambridge: MIT Press, 2012. [Online]. Available: http://ebookcentral.proquest.com/lib/princeton/detail.action?docID=3339490

[13] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] H. L. Penman, "Natural evaporation from open water, bare soil and grass," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 193, no. 1032, pp. 120–145, 1948.

[16] I. C. Prentice, W. Cramer, S. P. Harrison, R. Leemans, R. A. Monserud, and A. M. Solomon, "Special paper: a global biome model based on plant physiology and dominance, soil properties and climate," *Journal of biogeography*, pp. 117–134, 1992.

[17] N. Quarmby, M. Milnes, T. Hindle, and N. Silleos, "The use of multi-temporal NDVI measurements from AVHRR data for crop yield estimation and prediction," *International Journal of Remote Sensing*, vol. 14, no. 2, pp. 199–210, 1993.

[18] J. Rouse Jr, R. Haas, J. Schell, and D. Deering, "Monitoring vegetation systems in the great plains with ERTS," pp. 1,29–31, 1974.

[19] Z. Wang, C. Liu, and A. Huete, "From AVHRR-NDVI to MODIS-EVI: Advances in vegetation index research," *Acta ecologica sinica*, vol. 23, no. 5, pp. 979–987, 2003.

[20] Y. Zhou and R. Chellappa, "Computation of optical flow using a neural network," 08 1988, pp. 71 – 78 vol.2.

[21] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

ACKNOWLEDGEMENTS

I would like to thank Professor Engelhardt for her guidance as well as Greg Gundersen for his help understanding CNNs and Pytorch. Furthermore, the research for this project was conducted as part of the Farm Project at the Rubenstein Lab in the Ecology and Evolutionary Biology (EEB) department. Many thanks to Professor Daniel Rubenstein, Gina Talt (Lab Fellow), Mina Arashloo (Graduate student), Axel Haenssen (OIT drone specialist), and the EEB students who worked on this project over the summer.