IN-FIELD COTTON YIELD AND LOAN RATE PREDICTION USING MACHINE LEARNING REGRESSION ALGORITHMS

Jianing He
Wesley Porter
Luke Fuhrer
University of Georgia
Tifton, GA
Joe Thomas
USDA-ARS
Stoneville, MS
Christopher Delhom
USDA-ARS
New Orleans, LA

Abstract

The producers' revenue is highly dependent on cotton (*Gossypium hirsutum L.*) yield and loan rate associated with cotton fiber quality. It is vital to explore the determining parameters in a field for Site-Specific Management with the goal of obtaining maximum loan rate and maximum yield. The main goal of this study was to use available crop production parameters from a full-size cotton production field to develop relationships between production parameters, yield, and loan rate. This study was conducted in a field near Colquitt, Georgia. The cotton harvester was equipped with the John Deere's Harvest Identification System (HID), and each module harvested from the field has spatial fiber quality information associated with its developed location. Forty-two machine learning regression algorithms were compared to predict in-field cotton yield and loan rate with parameters measured infield. Due to the limitation of available production parameters measured infield, it is insufficient to analyze the underlying relationship between cotton fiber quality properties and the in-field parameters statistically. The initial results showed that the Light Gradient Boosting Machine (LGBM) Regression Method worked best to predict cotton loan rate and yield. The R-squared value was 0.45 with a root mean square error (RMSE) of 0.22 cents/lb. for cotton loan rate prediction and the R-squared value of 0.68 with an RMSE of 0.53 bale/ac for cotton yield prediction.

Introduction

In Precision Agriculture, by combining a global positioning system (GPS), geographic information system (GIS), and variable rate technology (VRT), the field can be divided into smaller management zones to optimize profit, which is known as Site-Specific Crop Management (SSCM). Also, with the development and application of cotton yield monitors and seed rate meters on cotton harvesters and planters respectively, cotton yield, heading position, elevation, speed, and seeding rate can be obtained during cotton harvest and seeding (Kachman and Smith, 1995; Wilkerson et al., 2001; Thomasson and Sui, 2003; Vellidis et al., 2003; Singh et al., 2005).

The cotton samples were previously collected manually from various locations in a field to determine the within-field variation of fiber quality (Ge et al., 2008). The new harvest identification (HID) system on modern John Deere (Deere & Company, Moline, Illinois, USA) on-board module building cotton harvesters utilizes Radio Frequency Identification (RFID) technology to label and track cotton modules from cotton harvest to gin. The cotton fiber quality properties are then measured in a unit of a module by the United States Department of Agriculture (USDA) Agricultural Marketing Service (AMS) classing measurements. The loan rate was calculated based on associated cotton fiber quality. Based on USDA Commodity Credit Corporation (CCC), the loan rate reflects the differences (loan rate premium and discounts) in market prices for color (reflectance (Rd) and yellowness (+b)), staple length, leaf, extraneous matter, micronaire, length uniformity, and strength (USDA 2017, 2021).

As cotton yield and fiber quality are two primary concerns in cotton production for farmers, growers, and researchers, previous studies have been conducted to explore the predominant parameters in cotton fields that impact cotton yield and fiber quality in applying SSCM. For environmental effects, cotton fiber strength reduced 3% with low light (70% of incident sunlight) than incident light and cotton lint yield was 10% on average lower in the warm regime which was 1℃ was warmer than ambient temperature (26.4℃) during growing seasons (Pettigrew, 2001, 2008). The nitrogen (N), and potassium (K) stress were investigated separately at the flowering stage on lint yield and fiber quality. With a relatively high boll load but low-quality fiber, N deficiency had an indirect influence on cotton fiber quality while

K deficiency led to significant reductions in lint yield and micronaire (Read et al., 2006). For Phosphorus (P), cotton yields were not significantly different between blanket-rate and variable-rate P treatments. Variable-rate P treatment applied thirty-eight percent less P than blanket-rate (Bronson et al., 2003). Also, plant densities, soil electrical conductivity (EC), landscape positions, and irrigation had some impact on cotton yield and/or cotton fiber quality (Bronson et al. 2003, 2006; Bednarz et al. 2000, 2005; Terra et al., 2006; Guo et al., 2012).

Spatial variability and spatial correlation were found in soil and fiber quality properties (Elms et al., 2001; Johnson et al., 1999, 2002; Ge et al., 2008). Thus, geostatistical analysis was applied to evaluate spatial autocorrelation, estimate target parameters at unknown locations, and generate high-resolution cotton yield and quality maps. However, it was insufficient to predict cotton yield and fiber quality properties based on soil parameters using conventional statistical and geostatistical methods (Wang et al. 2017). In this circumstance, machine learning (ML) regression algorithms were considered to approximate robust relationships between input-output data to make accurate predictions. Due to its advances in computer technology and associated techniques, ML has been applied in Precision Agriculture in many aspects to address complex problems with promising results (Lary et al., 2016; Noi and Kappas, 2018; Mao et al. 2019; Leo et al. 2020; Benos et al. 2021). However, little research has been done to predict cotton yield and loan rate using ML regression algorithms.

The research objectives of this study were (1) to compare the performance of different machine learning regressor models in cotton yield and loan rate predictions, and (2) to optimize the best model with hyperparameter tuning to predict cotton yield and loan rate.

Materials and Methods

Study Site

The field study was conducted at the Fire Tower field (Bowen Farms) (31.1713° N, 84.7333° W) near Colquitt, GA in 2020. The field is 41.02 acres in size. Based on USDA NRCS (Natural Resource Conservation Service) soil survey, the dominant soil type was Carnegie (Fine, kaolinitic, thermic Plinthic Kandiudults). Uniform litter was applied on 7th April 2020 with a rate of 4000.01 lb./ac. Three cotton varieties "DP 1646" (Delta and Pine Land Company, Scott, Mississippi, USA), "DG 3615", and "DG 3799" (Dyna-Gro Advanced Science Simplified, Richmond, California, USA) were planted on 4th June 2020 at four different Seeding rates of 20800, 24000, 27200, and 32000 seeds/ac across the field. The cotton was harvested by a John Deere on-board module building cotton harvester on 23rd Nov 2020 and wrapped up as twenty-five modules. These modules were then sent to USDA AMS for ginning to remove cottonseed, plant residue, and other foreign material and eventually be pressed into bales (one bale equals 480 pounds). Each module could be pressed into approximately four bales.

Data Collection, Statistical Analysis, and Spatial Maps

The input variables used in this study were Distance (D), Heading (H), Elevation (E), and Applied Seeding Rates (ASR). The first three were collected during cotton harvest by the cotton yield monitor mounted on the John Deere cotton harvester. The last one was collected by the seed rate meter mounted on the John Deere planter during seeding. The outputs were cotton Yield (Y) measured by the cotton yield monitor during harvest and Loan Rate (LR) associated with cotton fiber quality properties by averaging four bales.

In total, 12,695 data points were collected in this field with all input and output variables as well as geospatial information (Latitude and Longitude). Exploratory statistics of input and output variables are given in Table 1. Also, descriptive statistics of measured cotton yield (left) and loan rate (right) with three different cotton varieties are shown in Figure 1. Figures 2 and 3 depicted the spatial maps of output variables (modules and LR) and four input variables (D, H, E, and ASR), respectively.

Table 1. Exploratory statistics of input and output variables in the study field (n = 12695)

Variables	Type	Max	Min	Mean	SD	CV (%)
Distance (m)	Input	15.53	0.07	7.77	0.50	6.40
Heading (cm)	Input	360.00	0.00	142.52	121.58	85.31
Elevation (m)	Input	169.59	160.79	165.49	1.83	1.11
Applied Seeding Rate (seeds/ac)	Input	76490	0	26262	4487	17.09
Cotton Yield (bale/ac)	Output	5.98	0.03	2.39	0.92	38.54
Loan Rate (cents/lb.)	Output	56.55	55.33	55.12	0.30	0.54

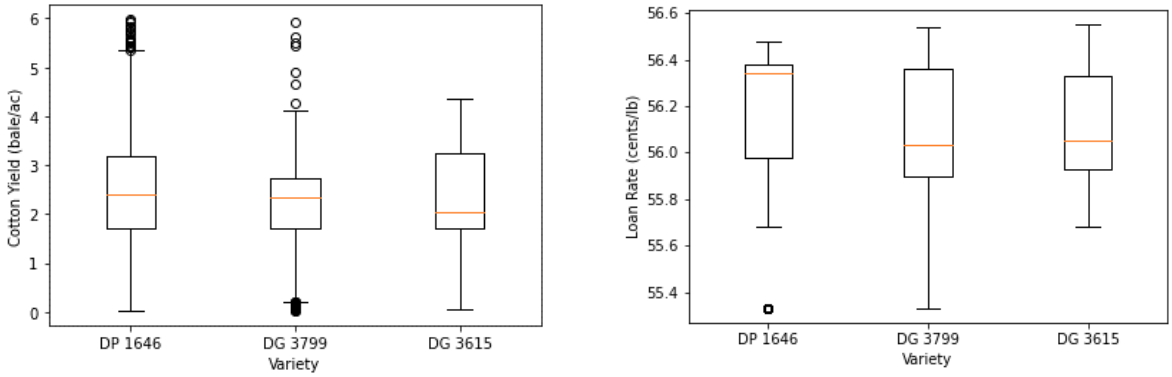

Figure 1. Descriptive statistics of Cotton Yield (left) and Loan Rate (right) on three cotton varieties are shown with boxplots

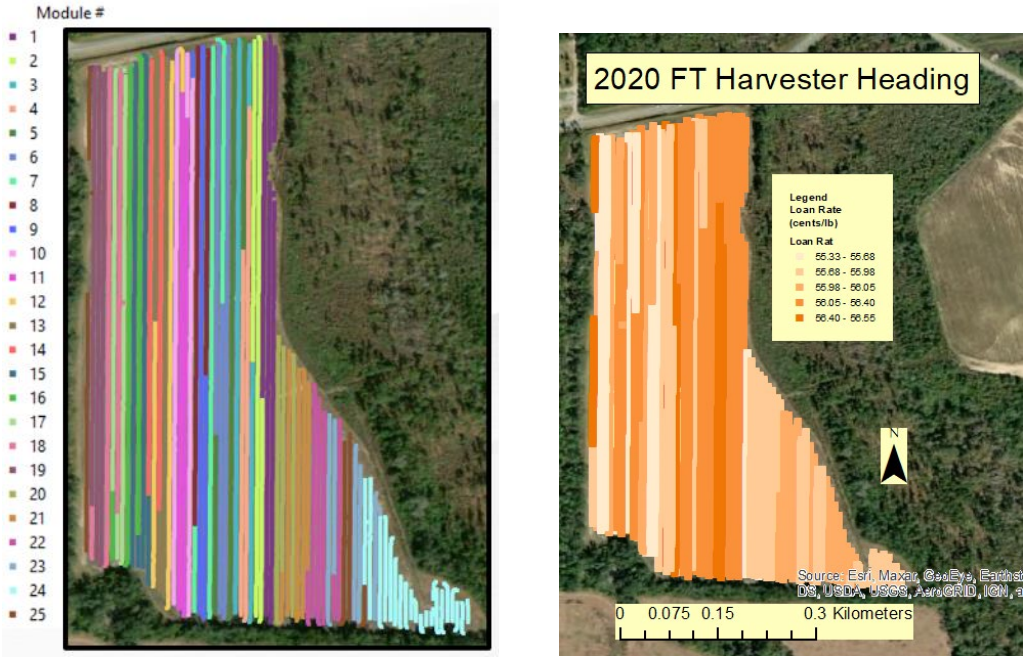

Figure 2. Spatial maps of output variables: modules (left) and Cotton Loan Rate (right)

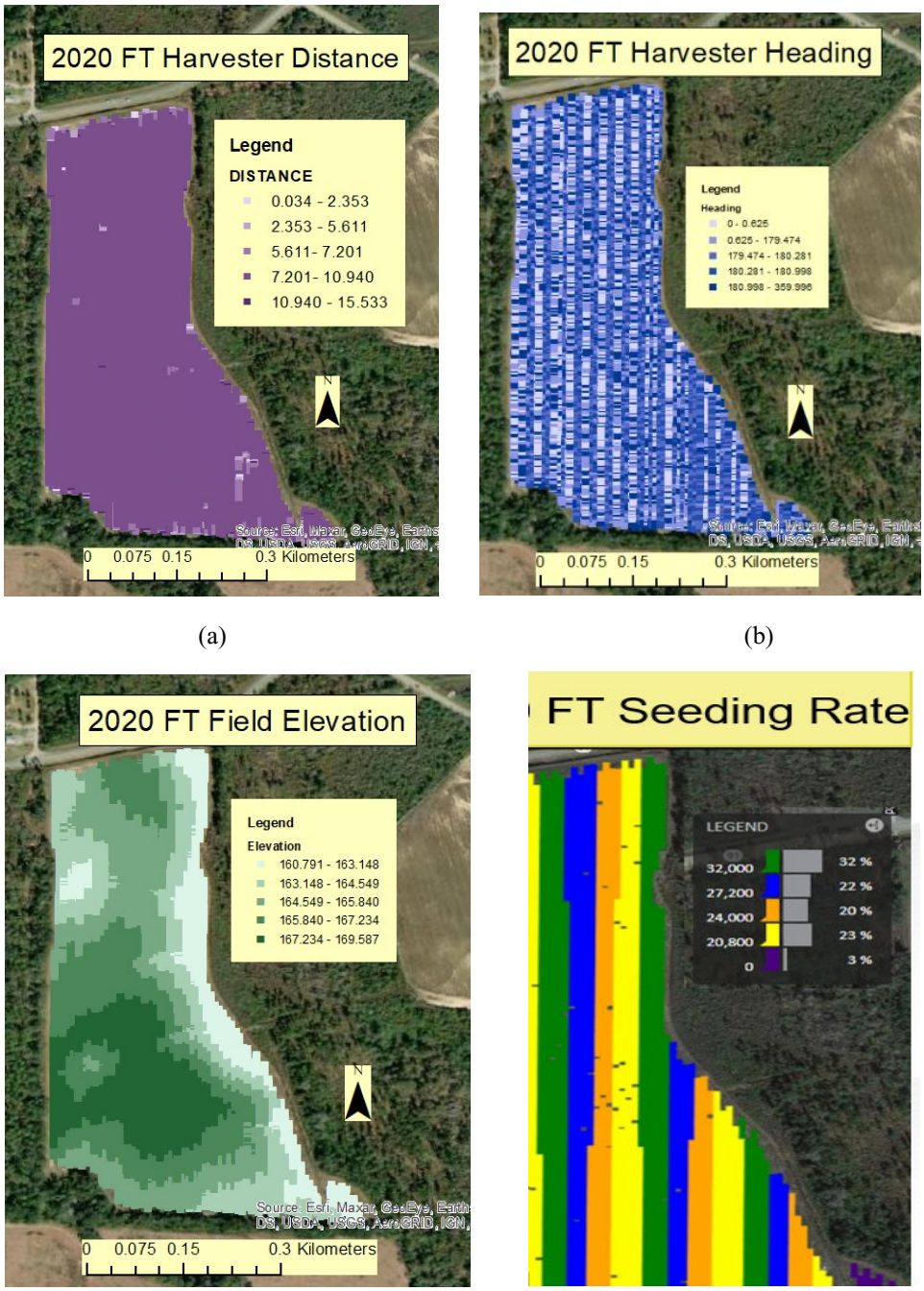

Figure 3. Spatial maps of input variables: Distance (a), Heading (b), Elevation (c), and Applied Seeding Rate (d)

Deployment of ML Regression Algorithms (MLRAs)

The relationship between input variables (D, H, E, and ASR) and output (Y and LR) can be learned automatically when implanting ML algorithms at the dataset. The scikit-learn package is an open-source Python module project integrating prevalent ML algorithms to perform classification, regression, and clustering (Van Rossum and Drake, 2009; Pedregosa et al., 2011).

To get an overview of different MLRAs' performance, the first step was to use another library named Lazy Predict in Python (Pandala, n.d.). The dataset was randomly split into 80% (n = 10156) for model training, and 20% (n = 2539)

for model testing. It listed and compared the performance (R-squared value, RMSE, and Time Taken) of forty-two different ML regression algorithms to help find the top five MLRAs to predict cotton yield and loan value as given in Table 2. The top five MLRAs were the same with even the same ranking to predict output variables, Yield, and Loan Rate, respectively. They are LGBM (Light Gradient Boosting Machine), XGB (Extreme Gradient Boosting), HistGradientBoosting (Histogram-Based Gradient Boosting), RF (Random Forest), and Bagging.

Table 2. Top five MLRAs to predict Yield and Loan Rate

Output variables	MLRAs' Performance Ranking				
	1	2	3	4	5
Yield (bale/ac)	LGBM	XGB	HistGradientBoosting	RF	Bagging
Loan Rate (cents/lb.)	LGBM	XGB	HistGradientBoosting	RF	Bagging

LGBM and Hyperparameter Tuning

LGBM is short for Light Gradient Boosting Machine, which is a kind of ensemble algorithm developed by Microsoft in 2007 to use a special type of decision trees, also called weak learners, to capture complex and non-linear patterns. The most significant difference between LGBM and other boosting algorithms is that LGBM grows trees vertically by using a leaf-wise algorithm while other boosting algorithms grow trees horizontally with a level-wise algorithm as shown in Figure 4 (Ke et al., 2017). LGBM is famous for its high speed, relatively low memory requirements, and great performance on large-size datasets. However, LGBM is sensitive to overfitting and cannot work well on small datasets. Also, the parameters of LGBM exceed one hundred, which makes it difficult and time-consuming to tune the hyperparameters.

Figure 4. LGBM's leaf-wise tree growth (left) and other bossing algorithms' level-wise tree growth (right)

In this study, LGBM regression method was chosen to predict cotton yield and loan rate with default parameter settings (base model). Also, a randomized search on hyperparameters was tuned to find out their optimal combination with the minimum loss function to achieve better results (tuned model). The selected hyperparameters were the number of leaves, minimum child samples, learning rates, and the alpha in regression.

Results and Discussion

Table 3 was truncated to the top five MLRAs with R-squared values, RMSE, and Time Taken for cotton yield prediction. The R-squared values ranged from 0.63 to 0.68, RMSE ranged from 0.51 to 0.56, and the time taken to run the MLRAs ranged from 0.18 to 3.79 seconds.

Table 3. Summary of Top five MLRAs' Performance to predict Yield

MLRA	R-squared value	RMSE	Time Taken
LGBM	0.68	0.51	0.18
XGB	0.68	0.52	0.66
HistGradientBoosting	0.67	0.53	0.60
RF	0.66	0.53	3.79
Bagging	0.63	0.56	0.40

Again, table 4 was truncated to the top five MLRAs with R-squared values, RMSE, and Time Taken for cotton loan rate prediction. The R-squared values ranged from 0.40 to 0.45, RMSE ranged from 0.22 to 0.24, and the time taken to run the MLRAs ranged from 0.18 to 3.43 seconds.

In these five MLRAs, RF took the longest (3.79 and 3.43, respectively) time while LGBM was the quickest (both

0.18) with the highest R-square values (0.68 and 0.45, respectively). LGBM was approximate two times faster than Bagging and HistGradientBoosting, three times faster than XGB, and more than six times faster than RF while maintaining promising R-squared value and RMSE. Before hyperparameter tuning, these results indicated LGBM worked best for both cotton yield and loan value predictions.

Table 4. Summary of Top five MLRAs' Performance to predict Loan Rate

MLRA	R-squared value	RMSE	Time Taken
LGBM	0.45	0.22	0.18
XGB	0.44	0.23	0.71
HistGradientBoosting	0.43	0.23	0.49
RF	0.43	0.23	3.43
Bagging	0.40	0.24	0.37

To predict cotton yield, the training and testing R-squared values of the LGBM base model were 0.763 and 0.653, respectively. The mean absolute error (MAE), accuracy, and root mean squared error (RMSE) were also calculated for this base model. The values were 0.38 bale/ac, 68%, and 0.53 bale/ac. After hyperparameter tuning, the training R-squared value improved to 0.828 (8.5%) and the accuracy improved to 71.9% (5.4%). There were no significant changes between the base and tuned models' testing R-squared value, MAE, and RMSE as shown in Table 5.

Table 5. The performance of the base LGBM and tuned models to predict cotton yield

Models	Training R-squared	Testing R-squared	MAE	Accuracy	RMSE
Base	0.763	0.653	0.38	68.0%	0.53
Tuned	0.828	0.659	0.37	71.9%	0.53

To predict cotton loan rate, the training and testing R-squared values of the LGBM base model were 0.565 and 0.441, respectively. The MAE, accuracy, and RMSE were also calculated for this base model. The values were 0.16 cents/lb., 99.7%, and 0.22 cents/lb. After hyperparameter tuning, the training R-squared value improved to 0.678 (20%). There were no significant changes between the base and tuned models' testing R-squared value, MAE, accuracy, and RMSE as shown in Table 6.

Table 6. The performance of the base LGBM and tuned models to predict cotton loan rate

Models	Training R-squared	Testing R-squared	MAE	Accuracy	RMSE
Base	0.565	0.441	0.16	99.7%	0.22
Tuned	0.678	0.439	0.16	99.7%	0.22

In cotton yield and loan rate predictions, it was both found that the training R-squared values had a great improvement compared to other parameters. It should be noted that overfitting may occur if the training R-squared value is much larger than the testing R-squared. Also, the extremely high accuracy (99.7%) in loan rate prediction could be biased due to the relatively low testing R-squared value.

The variable importance indicates how much this model relies on each input variable to make accurate predictions. In cotton yield prediction, the importance of input variables from highest to lowest were D, ASR, E, and H as shown in Figure 5 (left). Figure 5 (right) shows the predictions versus residuals using LGBM to predict cotton yield. Most residuals ranged from -1 to 1 as expected. In loan rate prediction, the order of variable importance was the same as that of yield prediction though the values of importance were different for each input variable as shown in Figure 6 (left). However, linear patterns were found in the cotton loan rate predictions versus residuals plot, which is shown in Figure 6 (right). The reason behind these linear patterns was the imbalance between input variables and output loan rate. Each input variable had 12,695 different samples collected in the target field. However, just 25 different modules were collected in this field and each of them was associated with a unique loan value. There are two methods to solve this imbalance problem, ordinal regression and input downscale.

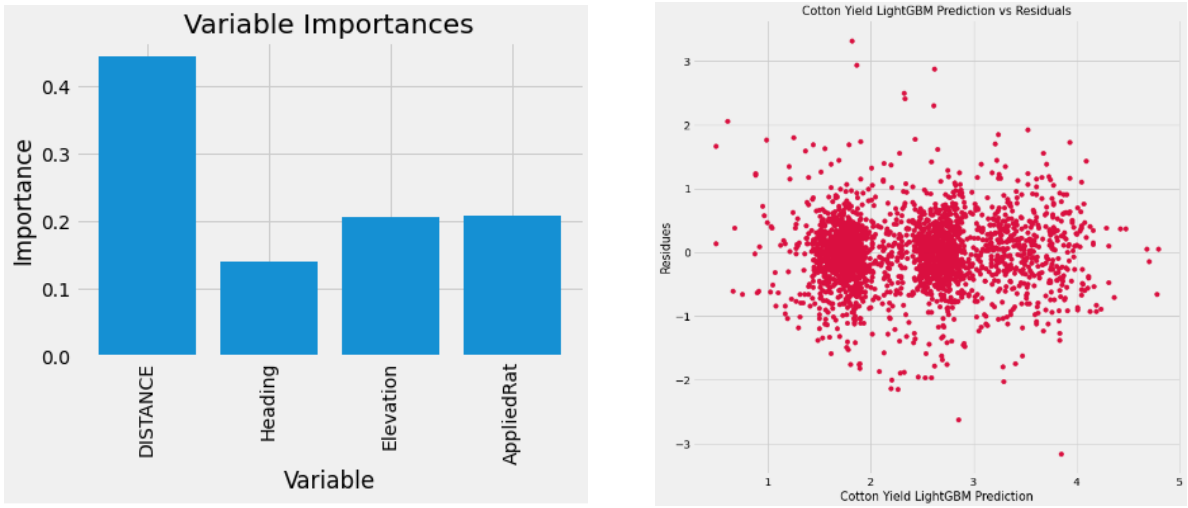

Figure 5. Input variable importance (left) and predictions vs. residuals (right) in cotton yield prediction

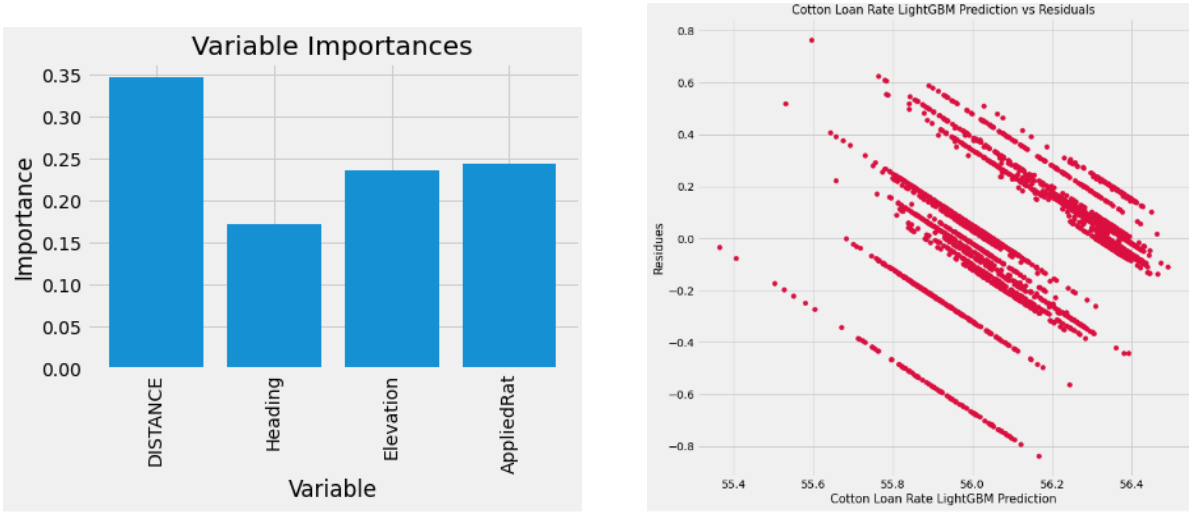

Figure 6. Input variable importance (left) and predictions vs. residuals (right) in cotton loan rate prediction

Summary

This study shows initial but promising results when using MLRAs to predict either cotton yield or loan rate. The R-squared values of the best MLRA were 0.68 and 0.45, respectively, which are much better than the traditional statistical models.

Compared to other MLRAs mentioned in this study, LGBM regression method had the best performance to predict cotton yield and loan rate with the highest R-squared values, lowest RMSE, and highest speed. The input variable importance of LGBM regression method from highest to lowest were Distance, Applied Seeding Rate, Elevation, and Heading. Also, the improvement of the LGBM regression methods was not significant after deploying hyperparameter tuning.

In the future, more input variables from external resources will be involved in this project, such as weather data from National Oceanic and Atmospheric Administration (NOAA) weather station, and soil data from Natural Resources Conservation Service (NRCS) soil surveys. Also, more fields will be included in this project to validate the accuracy and generality of the MLRAs as well as overcome the problems of overfitting and imbalance between the input and output variables.

Acknowledgments

We thank our colleagues and graduate students at the University of Georgia for their collaboration and for providing help on data collection and cleaning. Also, we thank John Deere and Cotton Incorporated for funding and supporting this project.

References

Bednarz, C.W., Bridges, D.C., Brown, S.M. (2000). Analysis of cotton yield stability across population densities. Agronomy Journal, 92, 128-135.

Bednarz, C.W., Shurley, W.D., Anthony, W.S., Nichols, R.L. (2005). Yield, quality, and profitability of cotton produced at varying plant densities. Agronomy Journal, 97, 235-240.

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., Bochtis, D. (2021). Machine Learning in Agriculture: A Comprehensive Updated Review. Sensors, 21(11), 3758. MDPI AG. Retrieved from http://dx.doi.org/10.3390/s21113758

Bronson, K.F., Booker, J.D., Bordovsky, J.P., Keeling, J.W., Wheeler, T.A., Boman, R.K., Parajulee, M.N., Segarra, E., Nichols, R.L. (2006). Site-specific irrigation and nitrogen management for cotton production in the Southern High Plains. Agronomy Journal, 98, 212-219.

Bronson, K. F., Keeling, J. W., Booker, J. D., Chua, T. T., Wheeler, T. A., Boman, R. K., et al. (2003). Influence of landscape position, soil series, and phosphorus fertilizer on cotton lint yield. Agronomy Journal, 95(4), 949–957.

Chen, T, Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 785–94.

Elms, M. K., Green, C. J., Johnson, P. N. (2001). Variability of cotton yield and quality. Communications in Soil Science and Plant Analysis, 32, 351–368.

Ge, Y., Thomasson, J.A., Sui, R., Morgan, C.L., Searcy, S.W., Parnell, C.B.J.P.A. (2008). Spatial variation of fiber quality and associated loan rate in a dryland cotton field. Precis. Agric, 9, 181–194.

Guo, W., Maas, S.J., Bronson K.F. (2012). Relationship between cotton yield and soil electrical conductivity, topography, and Landsat imagery. Precision Agric., 13 (6), 678-692.

Johnson, R.M., Bradow, J.M., Bauer, P.J., Sadler, E.J. (1999). Influence of soil spatial variability on cotton fiber quality. Proceedings of the beltwide cotton conference (p. 1319-1320). Memphis, Tennessee, USA: National Cotton Council.

Johnson, R.M., Downer, R., Bradow, J.M., Bauer, P.J., Sadler, E.J. (2002). Variability in cotton fiber yield, fiber quality and soil properties in a south eastern coastal plain. Agronomy Journal, 94, 1305-1316.

Kachman, S.D., Smith, J.A. (1995). Alternative measures of accuracy in plant spacing for planters using single seed metering. Trans. ASAE, 379-387.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W. et al. (2017). Lightgbm: A highly efficient gradient boosting

decision tree, Proc. Adv. Neural Inf. Process. Syst., pp. 3146-3154.

Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L. (2016). Machine learning in geosciences and remote sensing. Geosci. Front, 7, 3–10.

Leo, S., Migliorati, M.D.A., Grace, P.R. (2020). Predicting within-field cotton yields using publicly available datasets and machine learning. Agron. J. 2020, 1150–1163

Mao, H., Meng, J., Ji, F., Zhang, Q., Fang, H. (2019). Comparison of Machine Learning Regression Algorithms for Cotton Leaf Area Index Retrieval Using Sentinel-2 Spectral Bands. Applied Sciences, 9(7), 1459. MDPI AG. Retrieved from http://dx.doi.org/10.3390/app9071459.

Noi, P.T., Kappas, M. (2018). Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. Sensors, 18, 18.

Pandala, S. R. GitHub - shankarpandala/lazypredict: Lazy Predict, [online] Available: https://github.com/shankarpandala/lazypredict.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. (2011). Scikit-learn: Machine learning in Python. J. Machine Learning Res. 12, 2825–2830.

Pettigrew, W.T. (2001). Environmental effects on cotton fiber carbohydrate concentration and quality. Crop Sci, 41,1108–1113.

Pettigrew, W.T. (2008). The effect of higher temperature on cotton lint yield production and fiber quality. Crop Sci, 48, 278–285.

Read, J.J., Reddy, K.R., Jenkins J.N. (2006). Yield and fiber quality of Upland cotton as influenced by nitrogen and potassium nutrition. Eur. J. Agron., 24, 282-290.

Singh, R.C., Singh, G., Saraswat D.C. (2005). Optimization of design and operational parameters of a pneumatic seed metering device for planting cottonseeds. Biosyst. Eng., 92 (4), 429-438.

Terra, J. A., Shaw, J. N., Reeves, D. W., Raper, R. L., Santen, E. V., Schwab, E. B., et al. (2006). Soil management and landscape variability affects field-scale cotton productivity. Soil Science Society of America Journal, 70(1), 98–107.

Thomasson, J. A., Sui, R. (2003). Mississippi cotton yield monitor: Three years of field-test results. Applied Engineering in Agriculture, 19, 631–636.

USDA. Agricultural Marketing Service. (2017). Agricultural handbook 566: The classification of cotton. Washington, DC, USA: USDA.

USDA. Farm Service Agency. (2021). 2021 CCC loan schedule of premiums and discounts for upland and ELS cotton. Available at https://www.fsa.usda.gov/programs-and-services/price-support/commodity-loan-rates/index Accessed 01 Oct 2021.

Van Rossum, G., Drake, F.L. (2009). Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

Vellidis, G., Perry, C. D., Rains, G. C., Thomas, D. L., Wells, N., Kvien, C. K. (2003). Simultaneous assessment of cotton yield monitors. Applied Engineering in Agriculture, 19, 259–272.

Wilkerson, J. B., Moody, F. H., Hart, W. E., Funk, P. A. (2001). Design and evaluation of a cotton flow rate sensor. Transactions of the ASAE, 44, 1415–1420.