Evaluating the performance of crop yield forecasting models coupled with feature selection in regression framework

Manoj Varma[1,2], Achal Lama[1,*], K. N. Singh[1] and Bishal Gurung[1,3]

[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India
[2]Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi 110 012, India
[3]Department of Statistics, North-Eastern Hill University, Shillong 793 002, India

As crop yield is determined by numerous input parameters, it is important to identify the most important variables/parameters and eliminate those that may reduce the accuracy of the prediction models. The feature selection algorithms assist in selecting only those relevant features for the prediction algorithms. Instead of a complete set of features, feature subsets give better results for the same algorithm with less computational time. Feature selection has the potential to play an important role in the agriculture domain, with the crop yield depending on multiple factors such as land use, water management, fertilizer application, other management practices and weather parameters. In the present study, feature selection algorithms such as forward selection, backward selection, random forest (RF) and least absolute shrinkage and selection operator (LASSO) have been applied to three different datasets. Regression forecasting models have been developed with selected features for all the algorithms. The forecasting performance of the proposed models was compared using statistical measures such as root mean square error, mean absolute prediction error and mean absolute deviation. A comparison was made among all the feature selection algorithms. The regression models developed with LASSO, RF and backward selection algorithms were the best for different datasets.

Keywords: Crop yield, feature selection, prediction models, regression framework, statistical measures, weather indices.

THE effect of weather on crop growth varies with the stage of crop development. The influence of weather on crop yield is dependent on the magnitude of weather variables and how the weather is distributed across various growth stages of the crop. This is because different growth stages of the crop have different sensitivities to weather parameters; some are sensitive to weather fluctuations, while others are not[1]. Hence, we must divide the entire crop growth phase into narrow intervals for accurate forecasts. The interactions between weather parameters are crucial and should be included in crop–weather models[2]. Discriminant function analysis has been used to develop wheat yield forecast models. It has been observed that rainfall and temperature during key periods of wheat growth substantially impact wheat yield[3]. As a result, when the number of variables in the model increases, more parameters must be evaluated from the data. In such conditions, a series with large number of observations are required for accurate parameter estimation, which is difficult to obtain. The solution to this problem is finding a model based on a small number of parameters that can be easily evaluated. Also, it should consider the pattern or manner in which the weather is distributed throughout the crop growth period. The linear regression model aims to give an accurate forecast and maintain the complexity of the model to a minimum. The complexity of a model depends on the set of predictors, and determining this subset is known as the variable selection or feature selection.

Feature selection algorithms are important in data mining for finding useless attributes that should be removed from the dataset. In predictive analytics, feature selection refers to the process of finding the few most important attributes or features that are required to develop a model that can make an accurate forecast. Feature selection enhances the performance of the data mining algorithms and makes it easier for the analyst to interpret the modelling results. Including feature selection in the analytical process has numerous benefits; for example, it simplifies and narrows the scope of the features that are important in building a predictive model, and it helps save computation time. As the crop yield is determined by numerous input parameters, it is vital to find important variables and omit the redundant ones, which may decrease the accuracy of predictive models[4]. The feature selection algorithms assist in selecting only those relevant features in the predictive algorithms[5]. Instead of a complete set of features, feature subsets give better results for the same algorithm with less computational time[6]. Feature selection algorithms improve the performance of software defect prediction (SDP) models. There is no single best feature selection method because the performance of different methods varies according to the datasets and models used for prediction[7]. Sequential forward feature selection (SFFS),

*For correspondence. (e-mail: chllm6@gmail.com) and recursive feature elimination (RFE) feature selection strategies outperform other methods, with the bagging classifier working best with ten-fold and 70–30% data splitting range. Also, RFE with the bagging method outperforms other methods[8]. Feature selection has the potential to play an important role in the agricultural domain, with the crop yield depending on multiple factors such as land use, water management, fertilizer application, other management practices and weather parameters. Several variable selection techniques are available, like forward selection, backward selection, stepwise regression, ridge regression, etc.

Forward selection is a stepwise regression that starts with a blank model and gradually adds variables. Each time we take a step ahead, we add the one variable that improves the model the most. Unlike forward stepwise selection, it starts with a full least squares model with all p predictors and removes the least useful predictor one by one[9]. The least absolute shrinkage and selection operator (LASSO) is a regularization approach that minimizes the number of predictors in a regression model while identifying the most significant ones[10]. In the random forest (RF), high-importance variables greatly impact the outcome values. Low-importance variables, on the other hand, may be eliminated from a model, making it easier and faster to fit and predict[11]. The correlation-based method gave the best results for alfalfa yield prediction[12]. RF has been reported best for sugarcane (*Saccharum* spp.) yield modelling with data obtained from a sugarcane mill[13].

Feature selection algorithms

Forward selection

This is an iterative strategy in which no feature is included in the model at the start. We keep adding the feature that improves the model in each iteration. We keep iterating until adding a new variable has no effect on the performance of the model. In forward selection, the first variable chosen for inclusion in the model is the one with the highest correlation with the dependent variable. After selecting the variable, it is evaluated using a set of criteria. Two of the most common criterion are Mallows C_p and Akaike information criterion (AIC). If the first variable chosen meets the inclusion criteria, the forward selection process continues. When no more variables satisfy the entry criteria, the process is terminated.

Backward selection

This is also called backward elimination. It starts with all of the features and eliminates the least significant one at each step, which will improve the performance of the model. We continue the procedure until no improvement is observed. This feature selection technique is the reverse of the forward selection technique.

Strengths of forward and backward selection algorithms

- Stepwise selection is a computer-assisted strategy that can be used in almost every statistical package.
- Using stepwise regression to reduce the number of predictors in the model will increase out-of-sample accuracy (i.e. generalizability).
- Stepwise selection reduces the number of variables, resulting in a model that is simple and easy to interpret.
- When compared to manually choosing variables based on expert opinion, stepwise selection gives a reproducible and objective technique to reduce the number of predictors.

Weaknesses of forward and backward selection algorithms

- It is not certain that the best feasible combination of variables will be chosen.
- It generates biased regression coefficients, confidence ranges, P-values and R^2 values.
- Choosing variables using stepwise regression will be highly unstable, especially if the sample size is small in comparison to the number of variables to be studied.
- The causal relationship between variables is not considered.

Random forest

Here feature selection is carried out by determining the importance of each variable or feature. The default method to calculate variable importance is the mean decrease in impurity or the Gini importance. It is possible to compute how much each feature decreases the impurity. The more a feature reduces impurity, the more important it is. The impurity decrease from each feature can be averaged across trees in the RF algorithm to determine the final variable of importance.

Strengths of random forest

- RF does not require feature scaling because it uses a rule-based method rather than distance computation.
- It builds uncorrelated decision trees by implicitly performing feature selection. For this, it builds each decision tree using a random collection of features.
- Missing values can be handled automatically using RF.
- Both categorical and numerical data perform well with RF.
- Outliers are normally tolerated well by RF, which can handle them automatically.
- RF is less affected by noise compared to other methods.

Weaknesses of random forest

- It is difficult to interpret. RF gives a sense of how important a characteristic is but does not provide much interpretability of the coefficients as linear regression does.
- It is similar to the black box technique in that one has minimal control over the model output.

LASSO

This is a regularization technique for reducing the number of predictors in a regression model while selecting the most important ones. LASSO is a shrinkage estimator with a penalty factor that limits the size of the estimated coefficients and reduces predictive errors compared to the ordinary least squares (OLS) technique. Unlike ridge regression, LASSO sets more coefficients to zero as the penalty term increases.

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\cdot x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|,$$

where

$$\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\cdot x_{ij}\right)^2$$

is the residual sum of squares, λ the tuning parameter/regularization parameter and $|w_j|$ is the LASSO penalty.

Strengths of LASSO

- It prefers a collinearity-free subset of features.
- For better prediction and model interpretation, LASSO performs shrinkage and variable selection simultaneously.

Weaknesses of LASSO

- Regardless of whether all the predictors are relevant, LASSO regression will treat the majority of them as non-zero.
- If there are two or more highly correlated variables, it will choose one randomly, which is not a good data interpretation technique.

Materials and methods

For evaluation of the performance of crop yield forecasting models coupled with feature selection in a regression framework, different steps have been followed, starting with data collection and ending with a comparison between different regression models developed with selected features using various FS algorithms.

Collection of data

Collection of weekly weather data containing different weather parameters such as minimum temperature, maximum temperature, relative humidity, total precipitation, mean temperature and pressure was done. Also, crop yield data were collected for the particular districts.

Data preparation

From these weekly weather parameters, weather indices were developed. The number of weather indices obtained varies depending upon the number of weather parameters used.

Weather indices were developed using the following expression:

$$Y=A_0+\sum_{i=1}^{p}\sum_{j=0}^{1}a_{ij}Z_{ij}+\sum_{i'i=1}^{p}\sum_{j=0}^{1}a_{ii'j}Z_{ii'j}+cT+\varepsilon,$$

where

$$Z_{ij}=\sum_{w=1}^{m}r_{iw}^{j}X_{iw},\ Z_{ii'j}=\sum_{w=1}^{m}r_{ii'w}^{j}X_{iw}X_{i'w},$$

$r_{iw}/r_{ii'j}$ is the correlation coefficient of yield with the *i*th weather variable/product of the *i*th and *i*′th variables in *w*th week, *m* the week of forecast, *p* the number of weather variables used, *C* the constant and *T* the trend component included to account for the long-term upward or downward trend in the yield[14].

The number of indices formed from *n* weather variables = $2(n+\binom{n}{2})$.

So, for $n = 2$, the number of indices will be 6. For $n = 3$, it will be 12 and $n = 5$, the number of indices will be 30.

Feature selection

Weather indices have been used in place of weather variables. Different feature selection algorithms were applied to the weather indices to select important variables that can be used for further analysis.

Forecasting models

Crop yield forecast models were developed through multiple linear regression by taking the district-wise yearly yield data with the selected features using different feature selection algorithms.

Comparisons

Yield forecasts for the testing dataset were carried out with all the regression models developed. Finally, a comparison among all the regression models developed with the selected features using different FS algorithms was made. The comparison was made with the help of different measures such as mean absolute deviation (MAD), mean square error (MSE), root mean square error (RMSE) and mean absolute percentage error (MAPE).

Data description

The first dataset was collected for wheat crop yield and different weather parameters (maximum and minimum temperature, rainfall, relative humidity and precipitation) for Amritsar district, Punjab, India, provided by National Aeronautics and Space Administration (NASA), USA. The second dataset was collected for wheat crop yield and different weather parameters for Jalandhar district, Punjab, provided by NASA. The third dataset includes wheat crop yield data for the Patiala district, Punjab. The weather data were sourced from https://power.larc.nasa.gov/. Each of these datasets was partitioned into two, of which 80% data was used for fitting the prediction models and 20% for validation of the results. For all three datasets, 30 observations were used for training and fitting the models, and the remaining seven observations were for testing and validation purposes.

Results and discussion

Feature selection

The criterion of a model for forward selection is the significance level of the predictor. In the present study significance level is 0.05, which indicates that the predictor must have a *P*-value less than 0.05 to be included in the model. The Gini importance or mean decrease accuracy (MDA) is generally used as the criterion for choosing variables in the RF model for feature selection. Hence 2, 1 and 1 were the most important variables selected for the Amritsar, Ludhiana and Patiala datasets respectively. The size of the regression coefficient of the variable is used as the criterion in LASSO. This regression technique modifies the OLS objective function by including a penalty element that decreases the coefficients towards zero. A tuning parameter λ governs the penalty term. The value of λ for the Amritsar, Ludhiana and Patiala districts was 26.097, 48.946 and 58.803 respectively.

Table 1. Number of features selected with different methods

Algorithm	No. of features selected Amritsar data	Ludhiana data	Patiala data
Forward selection (FS)	3	5	2
Random forest (RF)	2	1	1
LASSO	6	4	7

Table 1 shows the number of features selected using different feature selection methods.

Predictions

When the developed linear regression models were validated using the testing dataset, different values of the predicted variables were observed for different feature selection

Table 2. Predicted values obtained from regression models of different FS algorithms for wheat yield (kg/ha) data of Amritsar district, Punjab, India

		Predicted values (kg/ha)		
Year	Actual values (kg/ha)	FS	RF	LASSO
2011	4283	3937.60	3886.70	4047.08
2012	4975	4429.98	3861.71	4057.76
2013	4654	3571.68	3920.52	3398.37
2014	4869	3879.33	3912.86	4094.43
2015	3914	4108.16	3706.00	2684.63
2016	4478	3983.11	4122.84	4238.80
2017	4948	4672.10	4009.24	4920.20
2018	4866	3849.94	4097.59	4259.02

Table 3. Predicted values obtained from regression models of different FS algorithms for wheat yield (kg/ha) data of Ludhiana district, Punjab, India

		Predicted values (kg/ha)		
Year	Actual values (kg/ha)	FS	RF	LASSO
2011	4964	4122.80	4293.71	4250.42
2012	5375	4709.60	4332.56	4497.55
2013	4853	3699.15	4302.89	3480.69
2014	5226	3892.81	4306.23	3966.86
2015	4462	3730.78	4261.76	2936.42
2016	4670	4495.27	4351.89	4479.80
2017	5093	4628.78	4362.79	4748.26
2018	5144	4312.36	4366.89	4289.50

Table 4. Predicted values obtained from regression models of different FS algorithms for wheat yield (Kg/ha) data of Patiala district, Punjab, India

		Predicted values (kg/ha)		
Year	Actual values (kg/ha)	FS	RF	LASSO
2011	4836	4095.57	4274.01	4054.29
2012	5472	3986.40	4226.55	4025.84
2013	4798	3387.32	4241.70	3353.21
2014	4968	3707.88	4256.43	3781.30
2015	4496	2371.33	4266.21	2642.29
2016	4585	3901.12	4174.89	3776.82
2017	5165	4188.49	4169.70	4343.37
2018	5272	4033.82	4188.08	3984.11

Table 5. Comparison of regression models for different FS algorithms for wheat yield (kg/ha) data of Amritsar district, Punjab

	Mean absolute deviation (MAD)		Root mean square error (RMSE)		Mean absolute prediction error (MAPE)	
Algorithm	Training	Testing	Training	Testing	Training	Testing
FS	260.64	617.93	300.69	703.42	6.79	13.13
RF	383.80	683.70	500.11	749.15	10.58	14.38
LASSO	304.84	660.84	366.52	791.46	8.17	14.58

Table 6. Comparison of regression models for different FS algorithms for wheat yield (kg/ha) data of Ludhiana district, Punjab

	MAD		RMSE		MAPE	
Algorithm	Training	Testing	Training	Testing	Training	Testing
FS	217.37	774.43	269.04	846.43	5.02	15.50
RF	349.52	651.04	447.83	703.83	8.36	12.82
LASSO	268.98	892.19	316.38	997.52	6.30	18.09

Table 7. Comparison of regression models for different FS algorithms for wheat yield (kg/ha) data of Patiala district, Punjab

	MAD		RMSE		MAPE	
Algorithm	Training	Testing	Training	Testing	Training	Testing
FS	283.57	1240.01	357.25	1313.48	7.16	25.22
RF	380.94	724.30	462.99	796.03	9.56	14.27
LASSO	280.94	1203.85	332.30	1256.05	7.00	24.47

techniques, as presented in Table 2 for Amritsar district wheat yield data. Table 3 for Ludhiana district wheat yield data and Table 4 for Patiala district wheat yield data.

For the Amritsar district wheat yield dataset, forward selection was found to be the best feature selection algorithm. The regression model for LASSO is as follows

$$\hat{Y} = 1003.50 * X_7 + 43.47 * X_8 - 15.48 * X_{21} - 1342.51.$$

For the Ludhiana district wheat yield dataset RF was found be the most efficient algorithm. The regression model for RF is:

$$\hat{Y} = 0.521 * X_1 + 0.01.$$

For the Patiala district wheat yield dataset, the best feature selection algorithm was RF. The predictive model for RF is:

$$\hat{Y} = 0.41 * X_1 + 3780.52.$$

Prediction accuracy

Different measures of comparison of prediction, such as MAD, RMSE and MAPE, have been used for comparing the accuracy of predictions provided by different regression models developed on the selected variables using different FS algorithms. The lower values of MAD, RMSE and MAPE assure a comparatively more accurate model. For the Amritsar district wheat yield data, the minimum MAD was 617.93, minimum RMSE was 703.42 and minimum MAPE was 13.13 (Table 5). The minimum values of all three measures corresponded to forward selection. For the Ludhiana district wheat yield data, all three measures were found to be minimum for RF. The values corresponding to MAD, RMSE and MAPE were 651.04, 703.83 and 12.82 respectively (Table 6). In case of the Patiala district wheat yield data, minimum values of MAD, RMSE and MAPE were 724.30, 796.03 and 14.27 respectively, and all three measures were found to be minimum for the RF algorithm (Table 7). The results indicate that for the Amritsar district data, the regression model combined with forward selection yielded the best results. On the other hand, for both Ludhiana and Patiala districts' wheat yield data, the regression model coupled with the RF algorithm performed the best. These findings are further supported by the visualizations (Figure 1).

Conclusion

The present study was conducted on three separate datasets. For all three datasets, 30 weather indices were obtained. Different feature selection algorithms select features

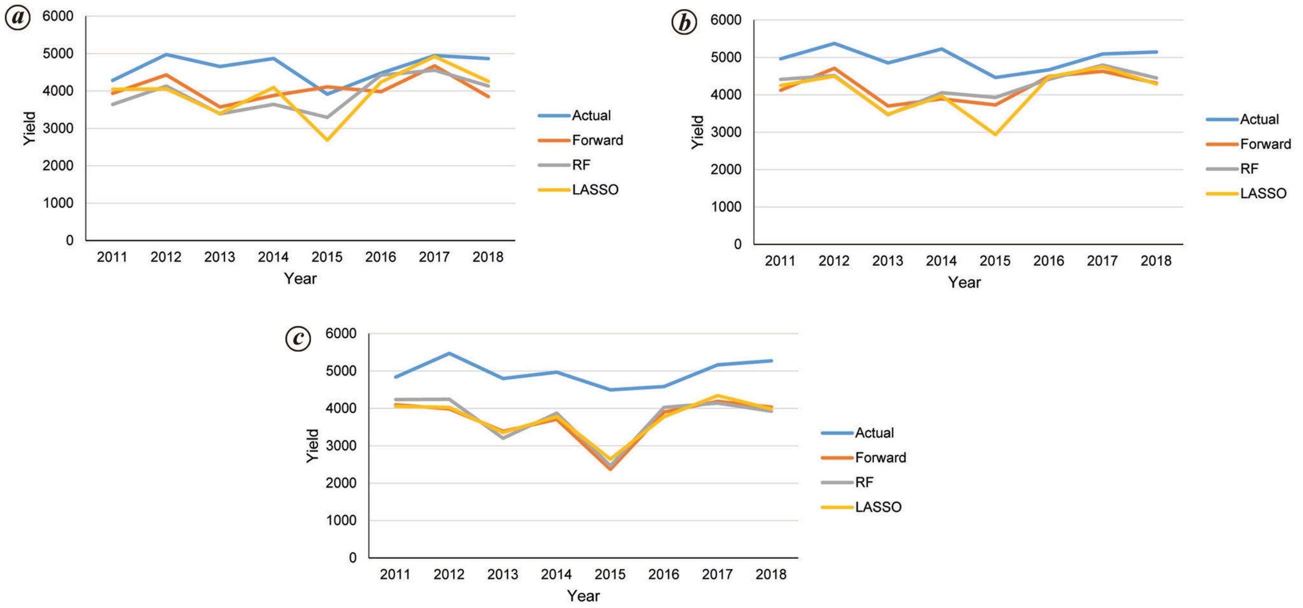

Figure 1. Fitting of regression models with different feature selection algorithms for wheat yield (kg/ha) data: (***a***) Amritsar district, (***b***) Ludhiana district, (***c***) Patiala district in Punjab, India.

based on different criteria. When the feature selection algorithms were compared, the regression model with forward selection provided the highest accuracy of prediction for the Amritsar district wheat yield. RF was most efficient for both Ludhiana and Patiala district wheat yield datasets. Thus we can conclude that the weather indices-based regression model coupled with feature selection algorithms provides greater accuracy for crop yield forecasting. As future scope of work, one can explore the possibility of applying other advanced feature selection algorithms, such as AdaBoost and XGBoost, to name a few.

Conflict of interest: The authors declare that there is no conflict of interest.

1. Singh, K. N., Singh, K. K., Kumar, S., Panwar, S. and Gurung, B., Forecasting crop yield through weather indices through LASSO. *Indian J. Agric. Sci.*, 2019, **89**, 540–544.
2. Agrawal, R., Jain, R. C. and Jha, M. P., Joint effects of weather variables on wheat yields. *Mausam*, 1983, **34**, 189–194.
3. Agrawal, R., Has, C. and Aditya, K., Use of discriminant function analysis for forecasting crop yield. *Mausam*, 2012, **63**(3), 455–458.
4. Springenberg, J., Dosovitskiy, A., Brox, T. and Riedmiller, M., Striving for simplicity: the all convolutional net. In 2nd International Conference on Learning Representations, ICLR2014, Banff, AB, Canada, 14–16 April 2014, pp. 1–14; https://arxiv.org/abs/1412.6806.
5. Oreski, D., Oreskib, S. and Klicek, B., Effects of dataset characteristics on the performance of feature selection techniques. *Appl. Soft Comput.*, 2017, **52**, 109–119.
6. Gopal, P. M. and Bhargavi, R., Optimum feature subset for optimizing crop yield prediction using filter and wrapper approaches. *Appl. Eng. Agric.*, 2019, **35**, 9–14.
7. Balogun, A. O., Basri, S., Abdulkadir, S. J. and Hashim, A. S., Performance analysis of feature selection methods in software defect prediction: a search method approach. *Appl. Sci.*, 2019, **9**(13), 2764.
8. Suruliandi, A., Mariammal, G. and Raja, S. P., Crop prediction based on soil and environmental characteristics using feature selection techniques. *Math. Comput. Modell. Dyn. Syst.*, 2021, **27**(1), 117–140.
9. Huang, J. Z., *An Introduction to Statistical Learning: With Applications in R*, Springer, New York, 2014, pp. 225–282.
10. Tibshirani, R., Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc., Ser. B*, 1996, **58**(1), 267–288.
11. Breiman, L., Random forests. *Mach. Learn.*, 2001, **45**, 5–32.
12. Whitmire, C. D., Vance, J. M., Rasheed, H. K., Missaoui, A., Rasheed, K. M. and Maier, F. W., Using machine learning and feature selection for alfalfa yield prediction. *AI*, 2021, **2**, 71–88.
13. Bocca, F. F. and Rodrigues, L. H. A., The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. *Comput. Electron. Agric.*, 2016, **128**, 67–76.
14. Agrawal, R. and Mehta, S., Weather based forecasting of crop yields, pests and diseases – IASRI models. *J. Indian Soc. Agric. Stat.*, 2007, **61**, 255–263.

ACKNOWLEDGEMENTS. We thank the anonymous reviewers for their constructive comments that helped improve the manuscript. We also thank India Meteorological Department, New Delhi for providing the dataset and the Director (ICAR-Indian Agricultural Statistics Research Institute, New Delhi) for providing the necessary facilities to carry out this study. Manoj Varma acknowledges the financial support provided by Indian Council of Agriculture Research, New Delhi in the form of JRF fellowship.

Received 27 January 2023; revised accepted 2 May 2023

doi: 10.18520/cs/v125/i6/649-654