Assessing Rice Yield

Sudhanshu Kumar Singh

Uddeshya Barnwal

Team- dhaasu

PROBLEM STATEMENT :

Given the data of rice yield collected from different states of India. Different factors

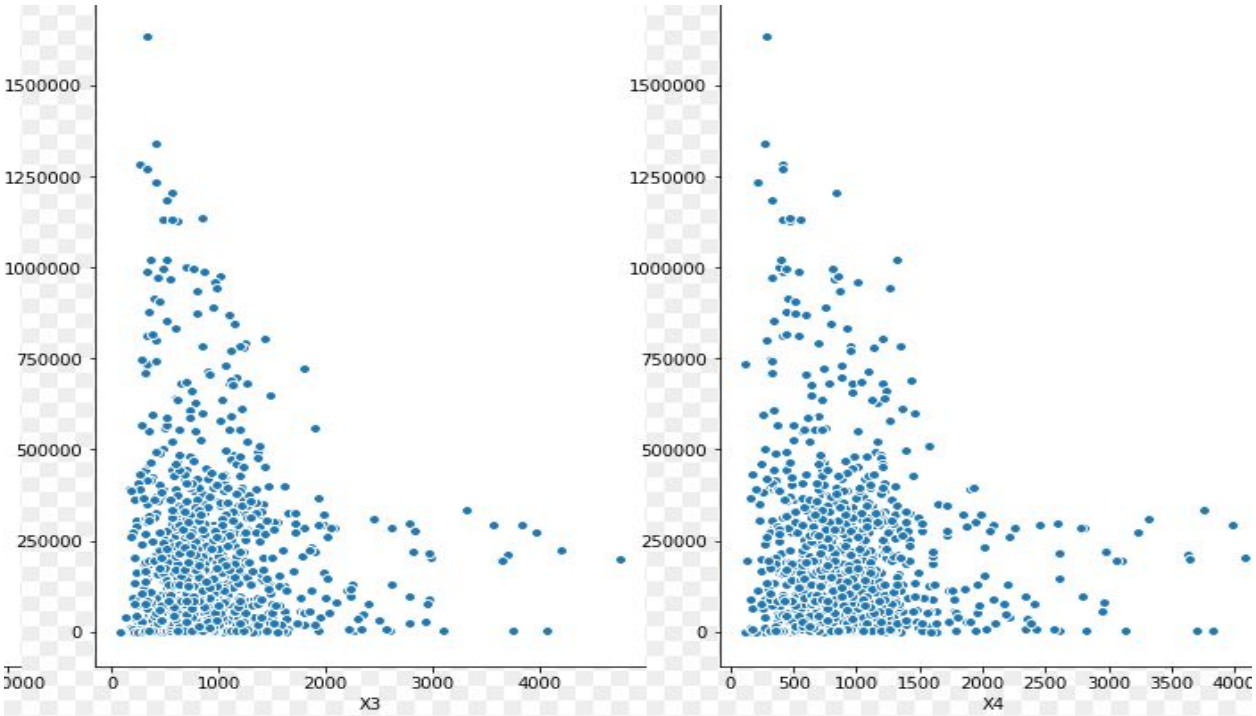
involving yield of rice is anonymous in given data named as X1, X2, X3, X4 along with some known factors such as crop area and state in which it is grown. Predict the yield of rice given the parameters.

DATA :

The training dataset contains the yield of 1451 different location. I am required to train the model in this dataset and predict the yield on test dataset which contains 480 locations.

DATA PREPROCESSING :

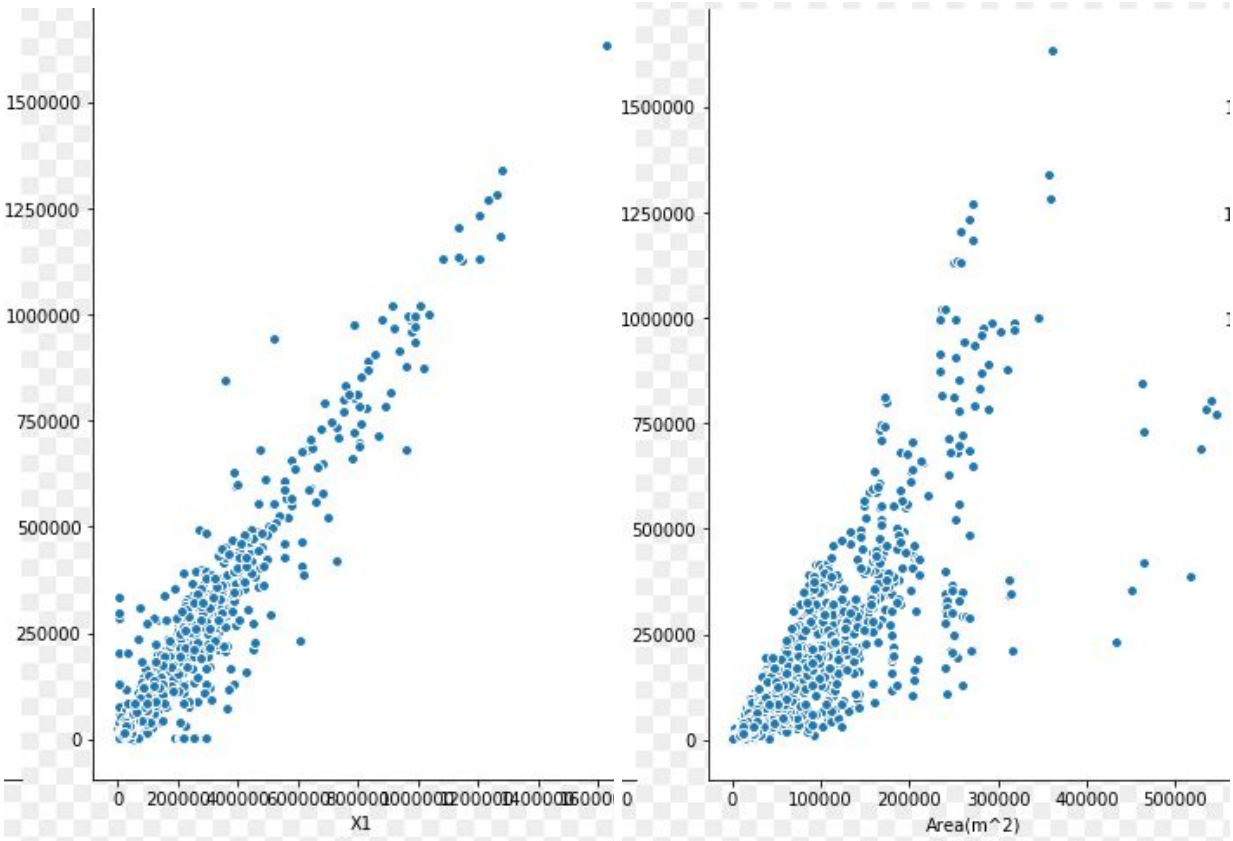
1. Removed the Sl.No column because it is not useful for analysis
2. There is high correlation between X1 and X2 and they are almost linearly dependent on Yield
3. There is high correlation between X3 and X4 but they act as noise in the prediction of Yield because their dependency on Yield is not linear.
4. Removed variables X3 and X4 for predicting the output that is Yield



5. Encoded the State_Name variable with integers for using as attribute in machine learning model.
6. Removed the outliers in X1, X2 and Area

DATA ANALYSIS (MODELLING) :

1. Divided The training data into training set and testing set using cross validation method into 70% to 30%
2. Used machine learning algorithms like PLS regression, Ridge , ElasticNet, Laaso, Linear Regression and Random Forest Regressor.
3. We have used these model because the relationship between variables and Yield is visually Linear.



4. Used all algorithms and found RMSE values corresponding to each one and found to be minimum for Random Forest Regressor then Ridge and so on.
5. I also used average of all algorithms prediction to minimise the RMSE but still

Random Forest Regression is having better result.

6. I also used Normalisation of data to optimise the model's RMSE value but the result is almost similar.
7. Tuned the n_estimators parameter of Random Forest Regressor to minimize the RMSE .
8. Used the Random Forest Regressor to predict the values on Test set.

CONCLUSION

Finally it is found that normalisation of data is not relevant here and Linear models of Machine Learning are showing good results with very less computation.

REFERENCES

1. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
2. http://scikit-learn.org/stable/modules/generated/sklearn.cross_decomposition.PLSRegression.html
3. http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
4. https://seaborn.pydata.org/