Crop Plantation Recommendation using Feature Extraction and Machine Learning Techniques

Soumya Sri Attaluri
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
tp055361@mail.apu.edu.my

Nowshath K Batcha
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
nowshath.kb@apu.edu.my

Raheem Mafas
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
raheem@staffemail.apu.edu.my

Abstract*— India is an agriculture-based economy with 18% of its total Gross Domestic Product (GDP) coming from different agricultural products. Agriculture 4.0 with modern technologies and robots for precision farming is shaping the future of agriculture in many places. In this research latest technologies like data science and machine learning algorithms are applied to understand different factors contributing to a profitable crop in India. These methods are applied on historical data collected from different Indian government web sites and publicly available data sets. This research provides a crop recommendation system with a prime motive of creating economic welfare of farmers. Multiple factors such as cost of planting, cost of harvesting, rainfall, crop demand, cost of seed, cost of fertilizer and yield of crop are considered to generate a more accurate prediction of whether a crop will be profitable or not.

Keywords—Machine Learning, Recommender System & Feature Extraction

I. INTRODUCTION

India is ranked amongst the top five countries for its agricultural production in the world. Promoting the economic welfare of farmers is pivotal to progress India's agricultural produce further. Indian agriculture produces wide variety of crops which have different demands, production costs and climatic conditions. Indian agriculture is not only a source of livelihood but also related to diverse culture of India. Considering all the factors of costs, geographic diversity and socio-economic conditions of different states of India, profitability of crop varies. Hence, profitability of a crop depends on the demand, weather forecast and cost of cultivation which includes the costs of seeds, fertilizers, labour and machinery. Individual state governments provide varied subsidies on raw materials and interest on capital loans. Farmers are also supported by benchmarking minimum sale price for the harvested crop. Conventionally profitability of the crop is only realized at the end of the harvest. Applying data driven machine learning (ML) models on historical data could help predict if a crop is going to be profitable or not based on the prevailing cost conditions, weather forecast, yield of the crop and minimum forecasted sale price. Objective of the project is to predict if a crop is going to profitable or not based on different agricultural costs, yield and rainfall. Different features are collected from distributed datasets. Data is chosen from years 2009 to 2015. To compute if the crop is profitable or not, compute the per hectare production by deriving from number of hectares planted and yield for that year. Profitability is also influenced by price of the crop and costs incurred for agriculture and irrigation.

II. PROBLEM CONTEXT

Profitability of a crop majorly depends on weather conditions, yield of the crop and costs of cultivation and production. Economic welfare of farmer depends on not only yield of the crop but also demand for the crop. Agriculture being the primary livelihood of the work force in India, many factors need to be addressed while making a crop decision as it impacts the farmer's economic welfare. If a farmer can be recommended whether a choice of crop is going to be profitable or not based on key factors like yield, weather forecast, market demand and few others costs then it would promote economic welfare of farmers.

III. LITERATURE REVIEW

There are several agricultural crop recommendation systems available considering various parameters by using ML algorithms. Various ML techniques are applied in agriculture sector to study the historical data which can be helpful to farmers as well as nation's economy. This section reviews few studies done in agriculture for crop recommendation and profitability of crop to farmer.

Taj at el [1] applied both classification and regression techniques to build a crop recommendation system. Data used consisted of parameters related soil condition and environment conditions. Initial classification is done using K-Nearest Neighbors (KNN) to find the best parameters that have significant impact on the crop yield. Subsequently, a regression model using Artificial Neural Networks (ANN) is used to predict a crop for recommendation. This study was done primarily to address food security problem in Egypt. Banavlikar et al.

[2], came up with a comprehensive, accurate and powerful crop recommendation system built by using neural networks with a motive of helping farmers to choose a right crop for a particular area of land. Soil and temperature are considered and Components like soil moisture sensor, humidity sensor and a temperature sensor are deployed to measure the water content in soil, amount of vapor present in surrounding air.

By applying the information gathered from sensors, the model predicts the best suitable crop for that particular land/area. Since neural networks are best fit for huge volumes of data, the same can be applied for irrigation system and in many other fields. Unlike other studies where the crop yield is predicted by using many machine learning algorithms, Priya, Muthaian and Balamurugan [3] in their study describes ability of an algorithm to predict crop yield. Random Forest is chosen

by considering parameters like rainfall, perception, production and temperature for different seasons of kharif and rabi for rice production in Tamil Nadu throughout the period of 1997 to 2013. Conclusions were drawn by summarizing the results that an accurate prediction of crop yield is predicted by random forest which makes random forest an able machine learning model for crop prediction by which farmers can be helped by choosing a right crop for cultivation.

Devadhe et al [4] proposed an expert system for crop selection considering parameters like district wise monthly rainfall and productivity of crops during 2000 to 2014 in Maharashtra. Linear regression, Decision tree and Random forest algorithms are applied to predict the yield rate of crops which may help improve the selection of sequence of seasonal crops to be planted. However, due to consideration of only couple of parameters, the study is limited only for seasonal selection of crops. Random Forest algorithm gave better prediction results when compared to linear regression ad decision tree.

Jain et al [5] proposed a crop selection method based on various factors like environmental, economic and yield rate to maximize the crop production which can help farmers and economy of nation to overcome the food supply demand. To achieve the increase in yield, selection of crop based on certain factors plays an important role. The factor of price is added to other parameters like soil type, rainfall, temperature for crop selection. Weka classifiers and regression methods are used to predict the appropriate selection of crop and then a crop sequencing method is proposed by using crop sequencing algorithm based on yield rate and market price. Also suggested machine learning models can be used in different ways in agriculture sector like irrigation, disease detection, pattern findings which can further advance agriculture sector.

Rajak et al. [6] came up with a crop recommendation system to maximise crop yield by considering soil specific attributes collected from soil testing laboratories from pune, Maharashtra. Also combined general crop data. Soil plays a major role in productivity of crop. The study is more specific towards soil attributes like PH, colour of soil, texture etc. An Ensembling technique called majority voting technique is used to predict the crop.

Ensembling is nothing but combining two or more models for better and efficient results. Support vector machine, Naïve Bayes, Artificial neural network, Random forest are used as model learners. Class labels are predicted by these learners for training data. By applying the majority voting technique, the majority of class labels are decided by voting. The rules are generated by the ensembled models by giving an output of crop recommendation which can help farmers in crop decision.

Ahamed et al [7] discuss various factors like effects of environment, soil salinity and area of production to predict the productivity of crop in Bangladesh across districts. Clustering is applied in order to group the districts assuming that the districts containing similar attribute values should belong to same cluster. Then models were built by using logistic regression, KNN and neural net to predict the yield of the crops.

However due to use of small dataset, the results were not accurate and as expected.

A recommendation of crop system is proposed from results obtained. The proposed system can recommend three best possible crops across all agricultural districts which could help farmer to plant different crops in different districts.

Dumbre, Chikane and More [8] has considered additional attributes for providing recommendations and included climate conditions. Data mining techniques were used to find patterns in the data. Main input data considered is from weather data store, fertility of soil data, crop history data and customers data. Genetic Algorithm and rule mapping are used for generating higher accuracy of prediction.

Dr. Ganesh, Cindrella and Christy [9] included crop management along with weather and soil conditions. This parameter inclusion explains a better view of profit that a farmer could expect. Online Analytical Processing (OLAP) system is used to create multi-dimensional data set before applying J48 tree technique. J48 technique has given higher accuracy compared to Naïve Bayes (NB) and simple chart.

Lakshmi et al. [10] made a critical observation to include water utility and land mass along with weather conditions and soil condition. Big data analytics platform is used to generate recommendations based on the data collected. The data was taken from laboratory testing for soil condition, fertilizers and biomass. Data is collected from government sites using web scraping with flume. Hadoop distributed file system (HDFS) was used to store the scraped pages and extract data. Finally, K-NN model is used to find the related data points to make crop recommendations.

Shinde, Andrei and Oke [11] have made a crop recommendation using the factors such as year of cultivation, market demand for the crop, crop yield per unit land, season and geographical region. By applying random forest algorithm, a 90% accuracy is achieved, however, the features considered are minimal and no relation to weather conditions and seed availability are made. Below table articulates all the work reviewed in this paper. Table compares different input parameters used, model applied and data manipulation methods.

IV. DATASET

Dataset for this research is sourced from Indian government websites published by different ministries related to agriculture. The data is stitched based on states, year and crop as keys.

The dataset for predicting if a crop is going to be profitable or not, depending on various factors like crop type, rainfall, different costs involved in agriculture, yield and market price are collected. The dataset consists of 23 variables and 2049 observations across seven years from 2009 to 2015. Python programming is used to collect, filter and integrate data from different sources. The sources include Agmarknet [12], Eands. dacnet.nic.in.[13]. and Mospi.nic.in. [14]. Figure 1 shows the summarized dataset after filtering and integration.

Atribute	Description
Principal Crops	Name of the crop that farmer chooses to grow. 11 crops ['Barley','Jowar','Bajra','Maize','Wheat','Cotton','Sesamum','Tur','Rice','Ground nut','Ragi']
type of crop	The crop cycle either ['Rabi', 'Kharif']
States	Geographical states in India
year	year of production 2009 to 2015
Agricultural area	Area under agrigulture for a specific crop in 1000 hectares
Production in thousand Tonnes	Quantity of produced crop in 1000 tonnes
Average Yield	Yeild in kilograms per hectare for a crop in a given state and year
Rainfall Actual	Summed average of rainfall in cm across the year
Average Price_quintal	Price in Rs. per quintal. Conversion is 1 Quintal = 100 Kg
Seed Quantity in KG per hectare	Quantity of seed in Kg per hectare
Fertilizer_Kg	Quantity of Fertilizer in Kg per hectare
Manure_Qtl	Quantity of Manure in Quintal per hectare. Convertion 1 Quintal = 100 Kg
Human_Labour_Man Hrs	Number of human labour hour required per hectare for a given crop
Animal Labour_Pair_Hrs	Number of animal labour hour required per hectare for a given crop
cost of Seed_Kg_hector	Cost of seed in Rs per Kg
cost of Fertilizer_RS_Kg	Cost of fertilizer per kg
cost of Manure_RS_Qtl	Cost of Manure in Rs per Qtl
Human Labour_RS_Man_Hrs	Human labour in Rs. per hour
Animal Labour_RS_Pair Hrs	Animal labour in Rs. per hour
Derived Yield_Qtl_Hectare	Yeild in Qtl per hectare
Insecticides	cost of Insecticides in Rs. per Hectare
Irrigation Charges	Cost of irrigation in Rs. per Hectare
Interest on Working Capital	Interest on working capital in Rs. per Hectare

Fig. 1. Dataset description

V. EXPERIMENTATION

Machine learning models work very effectively on features that are least correlated. It is important to extract features that are relevant and yet having least information overlap. Relevant and important features are generally known by gaining domain knowledge of the problem that is being solved. Information overlap can be found by either deriving covariance or by calculating Variance inflation factor (VIF). Machine learning models in general can only handle limited number of dimensions. Increase of number of dimensions increases the complexity of the model. Each feature in the dataset is considered as one dimension and this effectively means that a greater number of features increases the complexity and error of the model. These factors need to be considered while making feature selection. There are methods like principle component analysis (PCA) to reduce the dimensionality of high dimensional dataset.

Agricultural dataset chosen has 24 columns and is a good candidate for dimensionality reduction. However, on a closer look some features like states, crop and year can be dropped since these are categorical in nature. In PCA analysis, it is not easy to explain the outcome or the importance of different features. Hence the models like random forest and neural networks are chosen which don't suffer the curse of dimensionality. Target variable needs to be derived based on the price of crop, yield of crop and total expenses incurred. This derivation is not straightforward since profit estimation on the crop varies and cut-off point for each crop is different. The ML algorithms are used to predict a binary classification output, if a crop is profitable to the farmer or not was done using a) Random Forest b) Logistic Regression & c) Artificial Neural Network (ANN). The performance of a model can be evaluated using metrics like Accuracy, AUC, Sensitivity, Specificity, Confusion matrix etc. In this study, Confusion matrix, AUC ROC (Receiver Operating Characteristic curve), F1 score are used to measure the performance of as model. All these measures are gathered across all experiments/simulations and compared to explain the best model. Figure 2 and Figure 3 shows confusion matrix and model performance measures.

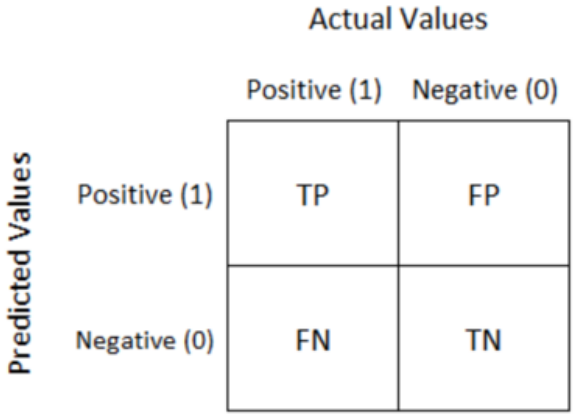

Fig. 2. Confusion Matrix

Accuracy	Sensitivity	Specificity	True Positive Rate (TPR)	False Positive Rate(FTR)
$\frac{TP+TN}{TP+FN+FP+TN}$	$\frac{TP}{TP+FN}$	$\frac{TN}{TN+FP}$	$\frac{TP}{TP+FN}$	$\frac{FP}{TN+FP}$

Fig. 3. Model Performance measures

VI. RESULTS

All experiments and simulations are done on the same split of training and testing data. The results after each simulation is captured in a data frame for comparison. True Negative(TN), False Negatives (FN), True Positives (TP), False Positives (FP) and the metrics are captured individually. Figure 4 shows the table of all the metric collected from individual models.

	index	TN	FN	TP	FP	accuracy	sensitivity	specificity	roc_auc	precision	f1_src	Experiment
0	0	448.00	48.00	114.00	5.00	0.91	0.70	0.99	0.85	0.20	0.81	Exp1-Sim1:RF_Default
1	0	437.00	17.00	145.00	16.00	0.95	0.90	0.96	0.93	0.25	0.90	Exp1-Sim2:RF_HyperParams
2	0	438.00	30.00	132.00	15.00	0.93	0.81	0.97	0.89	0.23	0.85	Exp2-Sim1:LogitRegression_base
3	0	402.00	17.00	145.00	51.00	0.89	0.90	0.89	0.89	0.27	0.81	Exp2-Sim2:LogitReg_updatedprods
4	0	433.00	8.00	154.00	20.00	0.95	0.95	0.96	0.95	0.26	0.92	Exp3-Sim1:NN_3H_Sigmoid

Fig. 4. Comparison of Results

It is important to note that there is a class imbalance in the data. Profitable crop entries (label: 1) are very less compared to non-profitable crops (label: 0). Hence maximizing true positive (TP) and minimizing false negative (FN) is very important compared to wrongly classifying a non-profitable crop. Neural Network model classified the data with highest TP and least FN. All the metrics of the experiments and simulations are compared using a bar graph as shown in figure 4. From the bar graph it is clear that ANN model provided the highest percentage of all the metrics which include f1 score, AUC ROC and accuracy.

VII. CONCLUSION

In conclusion, aim and all objectives of this research are achieved. Prediction is performed on historical data using logistic regression, random forest and artificial neural network (ANN) models. Different simulations are performed using hyper parameters to improve the model using the criteria of accuracy, sensitivity and specificity. This study can be progressed further in future by considering more variety of crops. Current research is limited to eleven crops based on data availability. Granularity of geographical region is set to a state in this study. Future studies could consider a more granular geographical regions to accurately assess the soil fertility, rainfall and water availability. In summary, this research evaluated factors that affect the profitability of a crop. Market price of the crop and yield of the crop are major impacting factors followed by costs of fertilizer, interest on capital and average rainfall. Based on the study it is recommended that farmers not only require minimum support price from government but also guidance on demand and supply. Realtime sharing of agricultural data like area under cultivation among states would not only help farmers but also helps in diversification of crops.

Comparing the results from machine learning models, it is evident that reasonably accurate predictions can be made if a crop is profitable or not, with historical and forecasted data. Crop profitability prediction is a high dimensionality problem and neural network provided the best results of prediction.

REFERENCES

[1] Taj, M. B.N., Kavya, H.C., Nayana, R. R., Bindu, H.S., Meghana, D. P. , " A Crop Recommendation System for Precision Agriculture," International Journal of Engineering Research & Technology, 6(15), 2018.

[2] T. Banavlikar, A. Mahir, M. Budukh, S. Dhodapkar, "Crop recommendation system using Neural Networks," International Research Journal of Engineering and Technology (IRJET), vol. 5(5), pp.1475-1480, 2018.

[3] P. Priya, U.Muthaiah, M.Balamurugan, "Predicting yield of the crop using machine learning algorithm", IJESRT International Journal of Engineering Sciences & Research Technology, Volume 7(4), April 2018.

[4] S. Devadhe, A. Kausadikar, P. Daphal, A. Joshi, A. 2017, "Expert System for Crop selection. International Journal of Scientific Research in Science and Technology, Volume 3(3), pp. 436-438, 2017.

[5] N. Jain, A. Kumar, S. Garud, V. Pradhan, P. Kulkarni, "Crop Selection Method Based on Various Environmental Factors Using Machine Learning", International Research Journal of Engineering and Technology (IRJET),Volume (04), 2017.

[6] R. K. Rajak, A. Pawar, M. Pendke, P. Shinde, S. Rathod, and A. Devare, "Crop recommendation system to maximize crop yield using machine learning technique", Int Res J Eng Technol, vol. 4(12), pp. 950–953, 2017.

[7] A.T.M Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, Rashedur M Rahman (2015) , 'Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bang, vol. 4(12), pp. 950–953, 2017.

[8] N. Dumbre, O. Chikane, G. More, "System for Agriculture Recommendation Using Data Mining," International Education & Research Journal (IERJ), EISSN: 2454-9916, Vol: 1, Issue: 5, Dec 2015.

[9] S. Hari Ganesh, B D. Pritty Cindrella, "A Review on Classification Techniques of Agricultural data", International Journal of Computer Science and Mobile Computing, Volume 4(5), pp. 491-495, 2015.

[10] S. Pudumular, E. Ramanujam, R. Harine Rajashree, C. Kavya, T. Kiruthika, J. Nisha, "Crop Recommendation System for Precision Agriculture", presented at the 2016 IEEE Eight International Conference on Advanced Computing, Volume 6(5), pp. 1132 – 1136, 2017.

[11] Shinde, K., Andrei, J., Oke, A. , " Web Based Recommendation System for Farmers", International Journal on Recent and Innovation Trends in Computing and Communication, Volume 3(3), pp. 1444 -1448, 2015.

[12] Agmarknet.gov.in. (2016). *Statistical and Analytical Reports.* [Online]. Available from : http://agmarknet.gov.in/PriceTrends/ [Accessed 4 December 2019].

[13] Eands.dacnet.nic.in. (2016). *Directorate Of Economics And Statistics, Ministry Of Agriculture, Government Of India*. [online] Available at: http://eands.dacnet.nic.in [Accessed 6 December 2019].

Mospi.nic.in. 2016. *Ministry Of Statistics And Program Implementation, Government Of India*. [online] Available at: http://mospi.nic.in/ [Accessed 6 December 2019].