Farmer's Analytical Assistant

Aakash G Ratkal, Gangadhar Akalwadi, Vinay N Patil and Kavi Mahesh
KAnOE - Centre for Knowledge Analytics and Ontological Engineering,
PESIT, PES University, Bangalore, India
(aakashratkal1, gangadharakalwadi, vinaypatilloyola, drkavimahesh)@gmail.com

***Abstract*--About half of the population of India depends on agriculture for its livelihood, but its contribution towards the GDP of India is only 14 per cent. One possible reason for this is the lack of adequate crop planning by farmers. There is no system in place to advise farmers what crops to grow. In this paper we present an attempt to predict crop yield and price that a farmer can obtain from his land, by analysing patterns in past data. We make use of a sliding window non-linear regression technique to predict based on different factors affecting agricultural production such as rainfall, temperature, market prices, area of land and past yield of a crop. The analysis is done for several districts of the state of Karnataka, India. Our system intends to suggest the best crop choices for a farmer in order to address the prevailing socio-economic crisis facing many farmers today.**

Keywords-Regression Analysis; Prediction Algorithms; Agriculture; Market Research; Data Analytics.

I. INTRODUCTION

A possible reason for the poor contribution of the agricultural sector to the GDP of India may be the lack of adequate crop planning by farmers as well as by the government. Rapid fluctuations in crop prices are common in the market. In such a scenario, it is difficult for a farmer to make an educated choice of crop to grow in his land or to estimate the yield and price to expect from it. The objective of our work is to help the farmer by applying predictive analytics on data from previous years.

Data analytics is examining of the raw data with an aim to draw conclusions about that information. It is used by many companies and organization to make better business decisions and by scientists to validate models or theories. There are several processes in data analytics - data requirements, collection of data, processing, cleaning data and exploratory data analysis where scientists may apply different kinds of analysis techniques to understand the message in the data. Finally, algorithms are applied to the available data to discover relations between variables, often mere correlations on the basis of which causality is ascribed to properties.

In the process of analysis, the data being collected is rendered graphically, with the intent of finding some useful patterns. Regression modelling or "fitting the curve" is an important method of building a function which is the best possible fit to a series of available data points. Visualization of data is enabled by these curves, but their primary use is to infer values of the function where insufficient data points are available. They also serve to summarize the relationship between two or more variables. In this work, we apply regression modelling to build a Farmer's Analytical Assistant tool.

In order to find an algorithm for regression, there are several linear regression methods including multiple linear regression [1], online ridge regression [2] and stochastic gradient descent [3]. Key elements of each of them are incorporated into the design of the prediction algorithm in Farmer's Analytical Assistant (FAA). These algorithms work on the basis of reducing the least square difference between the curve and the actual data values. In our agriculture dataset, a linear regression technique would not suffice since the yield and price depend on several parameters. Inspired by multiple linear regression technique [1], the dimensions in predicting the yield has been reduced and the prediction carried out in multiple steps. A nonlinear regression algorithm is employed to find the best fit curve.

The FAA tool uses sliding window non-linear regression method in order to find the relation, given two variables. In this technique we take the window size (ideally the time for a crop to grow in this domain), plot the graph for this window period and find the pattern. In the case of agriculture, we find that the pattern of rainfall has a strong influence on the amount of production. We analyse the rainfall pattern for the period that a crop takes to grow. This pattern is extrapolated to the current year in order to find the expected yield. A Taylor series approximation algorithm is used in FAA since this algorithm guarantees a sufficiently smooth function that can fit the graph points well. Many graphs are plotted initially and the one with least variance is selected.

In previous work, the prediction of the production price has been done by the grey method [4] for a time span of one year. In FAA tool, the data is considered over a longer period of 10 years. Several factors

978-1-5090-4573-0/16 $31.00 © 2016 IEEE
DOI 10.1109/CCEM.2016.22

change in such a duration. Hence, prediction of the price is done not only by considering the previous year's data but also by considering factors like inflation.

The farmer's analytical assistant provides suggestions by looking at various parameters for a given farmer. Its key component is farm production analysis for traditional crops, with a special emphasis on crops which take 3 to 6 months to yield.

II. PROCEDURE

A. Collection of data

The first challenge was to collect sufficient data to enable accurate analytics. Different datasets are required including (i) temperature data, (ii) rainfall data, (iii) production data and (iv) price data of each crop in different districts. There are different sources through which the datasets were obtained. Another issue comes from the different dimensions and units across datasets. For example, rainfall and temperature data are available for climatic regions (e.g., South Interior Karnataka) or for grid points at one-degree latitude and longitude increments on a daily or monthly basis. Crop wholesale price data of each crop in different markets of Karnataka are available on a daily basis for certain years. Crop yield data is available for different districts for each crop season. The datasets were cleaned and integrated to give us a consistent dataset to perform analytics.

B. Data Cleaning

One of the first and most important steps in any data processing task is to verify that the data values are correct or, at the very least, conform to some set of rules [5]. The dataset obtained were mainly in the form of XML/CSV files. Since the querying and searching in this format will not be convenient, we organized and indexed this data in the form of a relational database. There were several cleaning operations performed on the datasets in order to obtain accurate results:

1. Check for the missing values and replacing it with either default values or computing an interpolation to fit it in the curve.
2. Remove and replace the improper format entries (e.g., zero for the price typically indicates missing data).
3. Remove the outliers in the dataset or replace them with proper values.
4. Handle inconsistencies in the data, such as computing the season wise area and production in the cases where the values were given for the whole year.

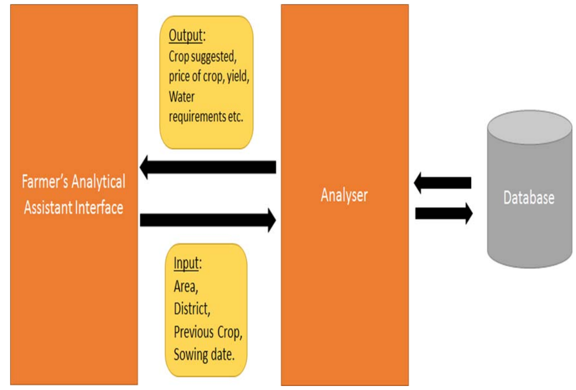

Figure 1. Architecture of Farmer's Analytical Assistant

After performing these steps, the data is ready to be used for analysis by the Farmer's Analytical Assistant tool.

C. Design

FAA tool consists of mainly 3 components: -

1. FAA User Interface: Users interact with the Farmer's Analytical Assistant through this interface. Currently we are providing a web interface to the tool. A farmer can login to our website and get various suggestions through it for which he provides the following inputs:
   i) Area of the land,
   ii) District in which the farm is located,
   iii) Previous crop which was grown on that piece of land and
   iv) The expected date on which he wants to sow the crop.

2. Analyser: The inputs provided by the farmer are passed to the analyser which in turn makes different computations by fetching the data from the database. The results of these calculations are given out to the User Interface in the form of a prioritized list of crops that a farmer can grow in his land to attain maximum benefit. For each of the crops we have different attributes such as
   i. Name of the crop
   ii. Expected price of that crop
   iii. Expected yield to be obtained from his land.
   iv. Water requirements (both rainfall and irrigation requirements)
   v. Soil requirements

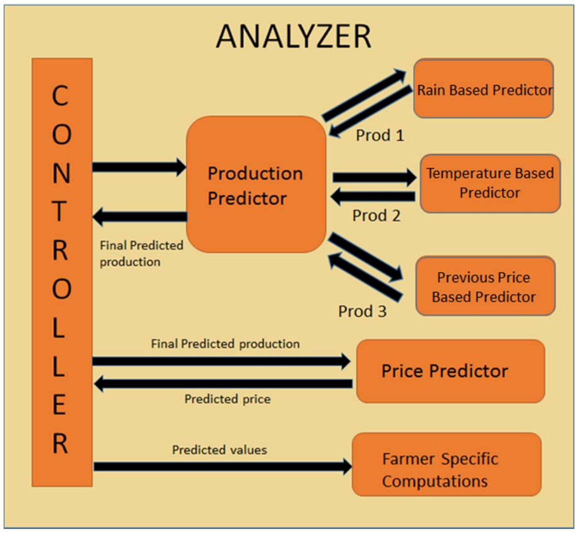

Figure 2. Different Modules of Prediction

vi. Other requirements such as ideal temperature, humidity, and pH value of the soil.

The Analyser consists of four main components:

1. Controller: Controller is the module that acts as the interface to the UI. It takes the input provided by the front end and directs it to different modules for further computations. Controller also takes the output of the computation and forwards it to the UI which in turn will present it in user readable format.
2. Yield or Production Predictor: From the controller, the control is first passed to the Production Predictor. The main function of this module is to make a correlation between production and the different factors on which it depends. In current implementation we have modules for rain, temperature and previous price dependencies. Each of these modules return a production value, where each of these values will have a different influence factor on the final production calculation, which will later be passed as a result to the controller.
3. Price Predictor: This module takes the predicted production as input, applies nonlinear regression to obtain a pattern in which the price in market depends on the amount of yield produced. This relation is used to predict the price of the crop after its harvest. This computed price is returned to the controller.
4. Farmer Specific values: Once we get the predicted produce and price for a crop, the control is passed to this module. Here farmer specific computations are made, such as, the yield he will get out of his land, expected price of the crop at the nearest market, crop rotation suitability, and the soil, irrigation, humidity, temperature conditions the crop requires for its optimum growth.

3. Database: It stores all available datasets in a suitable schema, to facilitate quick access.

D. Algorithm and Implementation

For the prediction of the future values based on the classified data, non-linear regression is used which enables us to use sophisticated hard models for extrapolation [7]. The calculation of various predictions begins with the formulation of the nonlinear least-squares fit line. The nonlinear least-square fit makes use of a nonlinear model to fit a set of observations.
Consider a nonlinear model, it can be written in the form of following equation:

$$Y = f(X,\theta) + \varepsilon$$

Where Y represents the data to be modelled, $X = (x_1, x_2, \ldots, x_k)'$ are independent variables, $\theta = (\theta_1, \theta_2, \ldots, \theta_p)'$ are the model parameters, and ε are the residuals or errors.

In nonlinear model we try to estimate the parameter values in such a way that it forms most appropriate mathematical equation to describe the data points. This is generally done by finding the best fit curve, which chooses the parameter values to minimize the deviation from the curve. This is popularly known as chi-square minimization, with

$$\chi^2 = \sum_{i=1}^{n} [\frac{Y_i - f(x_i'; \hat{\theta})}{\sigma_i}]^2$$

Now to estimate $\hat{\theta}$,

$$\frac{\partial \chi^2}{\partial \hat{\theta}_p} = -2 \sum_{i=1}^{n} \frac{1}{\sigma_i^2} [Y_i - f(x_i'; \hat{\theta})][\frac{\partial f(x_i'; \hat{\theta})}{\partial \hat{\theta}_p}] = 0$$

Since it is very difficult to obtain solutions to the equations, an iterative strategy is employed to estimate the parameter values.

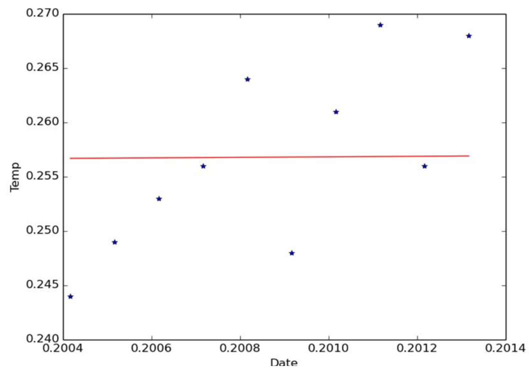

Figure 3. Temperature Variation

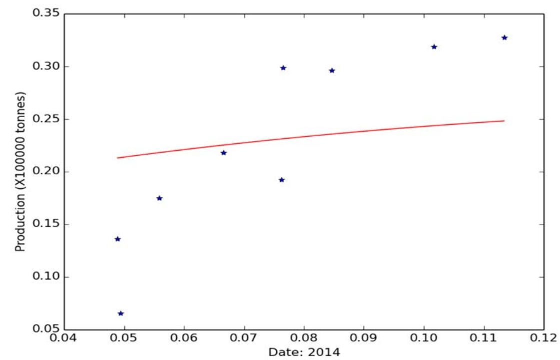

Figure 4. Production Variation

The nonlinear model which we employed for the predictions is:

$$F(p_1, p_2, x) = p_1 \cos(p_2 x) + p_2 \sin(p_1 x)$$

The iteration is started with some initial values, p_1, p_2 with each iteration, a χ^2 value is calculated and with an aim to reduce the χ^2, the parameter values are adjusted. When the χ^2 value is comparable between two successive iterations, we can say that the fitting procedure is converged.

1. Rainfall Prediction:
From the rainfall dataset we find a trend, if any, in the monthly rainfall pattern. Using this we extrapolate the pattern so as to obtain the expected rainfall for the growing period of crop.

2. Computing the production taking rainfall dependency:
The production of a particular crop in a district is plotted against the rainfall in that district and the pattern in it is observed. This pattern is extrapolated and the yield for the crop is found out for the rainfall expected in the next season. This yield acts as the Production 1 in our future references.

3. Temperature Prediction:
From the temperature data of a district, we find the pattern of changes in temperature in a particular agricultural season. This pattern is extrapolated to next year and expected temperature for next season is obtained. The graph for the temperature data and fitting curve for this data are shown in Figure. 3.

4. Computing production taking temperature dependency:
The yield of the crop for a season is plotted against the temperatures for that season and then predictions for the yield that will be obtained in the next season is made. This yield acts as Production 2.

5. Previous year crop price influence on the production this year:
We know that the crop prices in the market influences the farmers to grow that particular crop in their land. We take into account this factor and predict the production of the crop using the previous year crop prices. This is done by looking at their relation in the past years and then extrapolating it to the next year. This production is taken as the Production 3.

6. Computing the final production:
Using the production values obtained in the previous modules, we take a weighted sum of all the three depending upon their influence on the production.

Sl.no	*Dependency Factor*	*Weight*
1	Rainfall	50%
2	Previous Price	35%
3	Temperature	15%

Thus, we obtain a production value which acts as a final production value predicted by our tool.

7. Finding the price of crop:
Using the final production value, we plot dependency between production and price, and find out the optimum price of the crop.

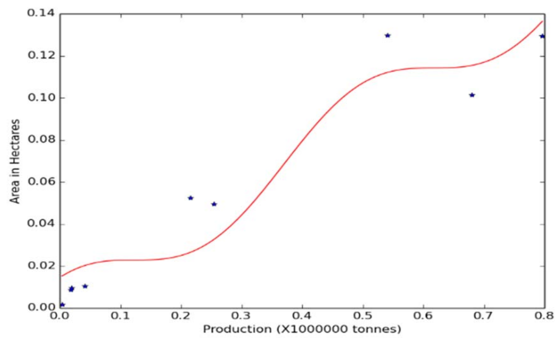

Figure 5. Production against Area of Land

8. Finding farmer specific details:

The price and production of the crop are computed specific to the farmer's area of land using various ratios.

9. Suggesting the crop:

The crops suitable to that region are selected and they are sorted on the basis of maximum income a farmer can make out of his land. The suggestion is completely in favour of the farmer.

10. Crop Rotation:

Crop rotation is one of the fundamental practice in modern agriculture. This is done so as to keep the soil nutrient level high. Crops are always selected in such a way that no two consecutive crop exhaust the same nutrients from the soil. In this way, soil remains fertile for longer time.

The list of crops from previous module are sorted again based on their feasibility towards crop rotation. We need to make sure that the crop we are suggesting doesn't barren the farmer's land.

11. Taking some other considerations:

Some of the other factors, such as humidity, ideal temperature for the crop, water requirements, pH level of the soil etc. are taken care before the farmer selects what crop he wants to grow in his piece of land.

12. Price regulation:

With our tool we also have to make sure that we direct the farmers to different variety of crops so that we regulate the prices based on the demand in the market. This is done by taking into account the amount of area that has been decided by the farmers for a particular crop, and how much more of the supply of that crop is needed in the market. For this we ask the farmer to select any one of the crop which he would like to grow, the selected crop and the farmer's area is stored in the database. For the subsequent farmers we show the availability of each crop in a particular

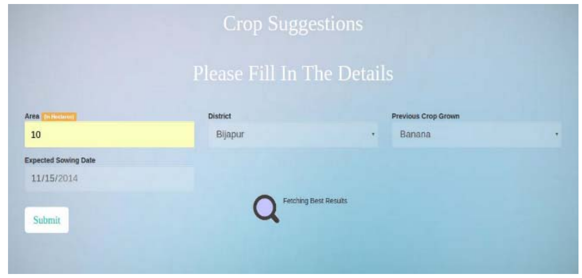

Figure 6. Crop Suggestion Request Form

Commodity	Production 1 Tonne	Production 2 Tonne	Production 3 Tonne	Final Production Tonne	Predicted Price Rs/Tonne	Total Area Required Hectares	Expected Farmer's Production Tonne	Expected Farmer's Income Rs	Min Rain MM	Max Rain MM	Actual Price Rs/Tonne	Percentage Error %	Available	Crop Rotation Compatible
Maize	64741.84	64451.8	100055.12	77067.98	13503.56	19444.26	39.83	535146.08	500	800	12282.13	9.94	✔	✔
Groundnut	17353.19	17288.7	15660.32	16750.94	34744.86	16983.61	9.86	337654.37	500	700	31600	8.37	✔	✔
Wheat	35980.65	35962.56	39344.32	37155.22	17934.36	24783.68	11.99	268836.06	450	650	22442.78	10.09	✔	✔
Gram	15602.58	15092.19	16682.0	15898.82	32212.65	20222.94	7.86	252869.3	600	1000	34673.6	7.1	✔	✔
Jowar	123363.13	123280.93	124731.86	123829.86	12714.32	124332.36	9.96	126634.63	450	650	14100	9.83	✔	✔
Sunflower	30733.28	30539.62	24128.91	28392.7	35095.43	5697.64	4.98	174775.24	350	500	31167.19	12.6	✔	✗

Figure 7. Prioritized List of Suitable crops

district in that season based on the previous selections made by the farmers, if a crop indicates no availability means most of the farmers are already growing that crop and he can go for some other crop which is available.

III. RESULTS

After running Farmer's Analytical Assistant tool, the dataset from the year 2004 to the year 2013, the prices predicted by the tool were compared with the actual market prices of 2014. The analysis shows an average accuracy of less than 10% difference between the predicted and actual market price.

IV. CONCLUSIONS

This tool intends to help farmers to make educated choices about the crop which he plans to grow next. We have implemented features like production prediction and price prediction which will help the farmer make a reasonable estimate of the price and yield he may get. The proposed tool also helps the farmer in estimating crop requirements such as water (both irrigation and rainfall), soil, humidity, pH value of soil etc. It also makes sure that the crops suggested follow crop rotation patterns so as to make sure that the land remains fertile for long. Also, we have made sure that the farmers are suggested different crops to make sure that all the farmers are not growing the same crop leading to drop in its price. With these features implemented we hope that the farmers will benefit from the tool and it will reduce the problem of crop insecurity and rapid fluctuations in market prices.

V. FURTHER ENHANCEMENTS

Some of the future enhancements in Farmer's Analytical Assistant are:

1. We have currently covered districts in Karnataka. We hope to extend FAA to whole of India.
2. Data is available only for a few of the prominent crops. As such, many regional crops are yet to be added to the tool.
3. Currently we are predicting only the income from the crops. We can add the prediction for investment needed for different crops. Factors like fertilizer, pesticide, preparation of farm for sowing, farm equipment, and bore wells play an important role in deciding which crop to grow.
4. We can also suggest fertilizer supplement needs of the soil if the farmer gives the soil analysis results.
5. The suggestions for intercropping and cash crops can also be added to increase productivity.
6. Several vegetation indices like Normalized Difference Vegetation Index (NDVI), Vegetation Condition Index (VCI) and Temperature Condition Index (TCI) can be used to detect draught conditions and several other weather impacts on the yield of the crop [6].

ACKNOWLEDGEMENTS

This work is supported in part by the World Bank/Government of India research grant under the TEQIP programme (subcomponent 1.2.1) to the Centre for Knowledge Analytics and Ontological Engineering (KAnOE) at PES University, Bangalore, India.

REFERENCES

[1] Yan Xiaozhen, Xie Hong, Wang Tong, "A Multiple Linear Regression Data Predicting Method Using Correlation Analysis for Wireless Sensor Networks", 2011 Cross Strait Quad-Regional Radio Science and Wireless Technology Conference

[2] Poala Arce, Luis Salinas, "Online Ridge Regression method using sliding windows" ,2012 31st Internation Conference of the Chilean Computer Science Society.

[3] Tng Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms", Proceedings of the twenty-first international conference on Machine Learning

[4] Jiajun Zong and Quanyin Zhu, "Apply Grey Prediction in the Agriculture Production Price", 2012 Fourth International Conference on Multimedia Information Networking and Security

[5] Ronald Cody, Ed.D., Robert Wood Johnson Medical School, Piscataway, NJ, "Data Cleaning 101"

[6] Aakunuri Manjula and Dr.G .Narsimha, "XCYPF: A Flexible and Extensible Framework for Agricultural Crop Yield Prediction", IEEE Sponsored 9th ISCO 2015.

[7] E.V. Bystritskaya, A. L. Pomerantsev and O.Ye. Rodionova, "Non-linear regression analysis: new approach to traditional implementations",2000 Journal Of Chemometrics