MARKET ANALYSIS FOR FARMERS USING MACHINE LEARNING

Sachin Desai[1], Kaushik Baug[2], Pragnesh Katkar[3]

[1-3]*Information Technology Engineering, Padmabhushan Vasantdada Patil Pratishthan's College of Engineering, Maharashtra India*

---***---

Abstract - *Use of Big Data Predictive Analysis and Data Mining have emerged to become useful for analyzing the agricultural crop price. This paper considers using techniques of data mining and Big Data using agricultural data. Farmers are more concern about obtaining the true value of the crop due to an old- chain system of selling the crops in their respective markets. The true capability of the market remains unexplored as farmers living nearby can only sell their crops indirectly affecting the consumers. Amount of data with the market keeps on increasing just like the other sectors. In this work, data from the various markets around Bangalore are taken into consideration and is used in our data mining decision tree models. The main goal is to provide a Farmers the best price for their crop using the Map-Reduce based decision tree model. The outcome of the suggested Map-Reduce model is precise to predict the best price from the available market data than the other models.*

Key Words: **Crop price, Big data, Data mining, Classification, Association rule, Decision tree**

1. INTRODUCTION

Agriculture has a significant contribution in the Indian economy,[1] contributes about 19.9% to the nation's GDP. India is leading in all kinds of crop production whether its food-grain or cash-crops. India's economy depends primarily on the growth of agricultural yields and their associated agro-industry. Most of the states in India use most of their geographical area for cultivation of crops.

Agricultural sector performance mainly depends on natural forces such as spatio-temporal distribution of rainfall, temperature, climate, etc., resulting in any deviation of the monsoon from the normal pattern resulting in huge fluctuations in area and production. Annually, a crop yield comes in to play on domestic and global economies, and the prediction yield [2] contributes much in the food and agriculture industry.

Agriculture sector entirely depends on 3 stakeholders Farmers, Market and Consumers. Due to increase in growth of population, demand of food will increase exponentially. Whenever the prices of food crops increases, marketmen keep an huge stock of crops and would not release from the stores until they get their required profit. As a result, consumers tend to buy less causing a threat to the agrarian economy.

Farmers are unable to make profits and as a result, losses are incurred more in the case of farming. The farmer has to rely on the local market to ensure that all of his crops should at least get sell at an MSP price.

[3] Data mining classification methods are used to create a pioneering model for anticipating the corresponding market price of commodities. Price estimation in agriculture is mandatory to forecast the market price for the chosen commodities, and it is supportive for farmers to schedule their crop cultivation operations and thus they gain more profit. Market price prediction is mandatory for both private and public sectors to plan and execute agricultural development programs to steady the commodities market price. Farmers become more aware of the price trends going on in the market and can trade with them accordingly, consumers can preplan their expenses based on their need.

This ground breaking idea comes in to play for the betterment of farmers, customers as well as markets solving the problem of utilization of agricultural goods as well as to increase farmer's capital.

[4] Data prediction modelling contains four phases, historical data analysis (descriptive), data pre-processing, data-modelling, and performance estimation. Classification technique in data mining provides the best solution for the process of prediction. Analysis of decision tree techniques observes an autonomous (predictor) and dependent (target) connection between the data set characteristics. This method enables the time series estimation of information and discovers the fundamental impact between these factors.

Data analytics (DA) is the method of examining data sets to draw conclusions about the information contained therein, progressively using specialized technologies and software. The forecast of sale of crop yields was carried out by considering the data of trade provided by various markets. However, as circumstances change very quickly every day, the trends in sale of food-crops gets changed every month.

Since this is the present scenario, many farmers have insufficient understanding of the market scenario and are not fully conscious of the advantages they gain from the market-driven agriculture. In addition, by knowing and predicting crop efficiency in a multitude of environmental circumstances, farm productivity can be improved.

As one who should understand how much he would expect for their plants, predicting prices as well as its associated market is a very significant problem for many farmers. Over the previous few years, price prediction has been produced by assessing farmers experience on a specific crop and field. Suppose they have prior year information accessible in which distinct respective price projections are recorded and these recorded price projections are used to classify future price projections [5].

By using these predictive models, the government can do agricultural development planning to stabilize the respective product prices. The suggested scheme applies machine learning and forecast algorithms such as decision tree algorithms to define information patterns and then process them according to input circumstances.

This primary objective is to suggest the best market with its associated price to the farmers based on the circumstances of the environment in the market as well as in the farmer's field.

Based on the factors of farmers and several markets this scheme recommends the price and suitable market using the available past data sets. This paper is structured as follows: the literature review required for the study work is described in Section II. Section III describes the Methods and Materials required for the study. Dataset description discussed on Section V. Section VI was discussed with Results and analysis. Section VII concludes the work with possible potential enhancement

2. RELATED WORKS

The food is on the top chart of primary needs for every human being, so the farmer's role is inevitable. Agricultural product price prediction supports the farmers to take beneficial decisions. This section discuss about various related works already done in data mining techniques using agriculture dataset. Most of the researchers focused on the similar problem like agricultural product commodity price prediction are as follows.

[6] Rajeswari, K. Suthendran developed an agricultural product price prediction model using HADT algorithm. The applications and techniques of data mining as well as Big Data using agriculture data is considered in this paper. Its proposed methodology consisted of two parts – the first part was known as Rule Creation. It first searches the training data set and discovers all the frequent item set. Then selects the values of the attributes in support descending order, F-list, and scans the training data to build a frequent item set again.

Next it sorted the rules in a compact tree structure. It investigates the possibility of exchanging rules and thus saves room as the rules that have common frequencies share the route portion. Lastly arranging and putting the rules in order by using general and high trust rules to prune more specific and lower trust rules. Select only rules that have been corrected positively. Select a subset of rules based on the coverage of the database. The second part consisted of the multiple rules generated using classification algorithm. For prediction existing C5.0: ADT classifier and association rule mining algorithms were used. These two algorithms were suitable to process the big amount data with better performance. Here, the fundamentals of C5.0: ADT decision tree algorithms were included in the MapReduce concept and the decision tree rules are generated. Decision tree rules were taken into the proposed Association rule mining algorithm for design the new MapReduce framework functions for generate the candidate itemset derived from the support, confidence count. The derived frequent itemset counts for selected crops prediction patterns were generated .

[7] Gauravjeet Dagar proposed a study of agricultural marketing information systems model and their implications. The information required to prepare the material principally came from various reports, documents and web site resources. The information was analysed to illustrate the use of market information, described the underlying concepts and generate lessons, ideas, and insights useful for developing and strengthening agriculture marketing information system in our country. it revisited some of the themes identified from the general literature and relates the case study experiences to them.

Availability of market information would also encourage spatial arbitrage between two markets, especially in cases where information and transport costs are relatively low. If no trade exists between two markets, both will clear supply and demand at their respective equilibrium prices. When price differences between the two are larger than the transaction costs, trade relations will be developed if there are no controls to inhibit exchange. A new equilibrium price will be determined for the combined market for the two regions. The availability of correct price information would lower the traders' cost of information gathering, as well as the risk of sudden unfavorable price changes.

[8] The Demand-Prediction Model for Forecasting AGRI-Needs of the Society proposed that in order to reduce the mismatch in demand and supply of food crops effectively, primarily the expected demand for various food commodities needs to be forecasted and guide the farmers accordingly. So there needed a system that could guide the farmers in selecting and growing the crops to satisfy the actual demand of the society. An effective forecasting model is proposed and has been implemented in this paper that **(i) Determine the gap between the demand and supply** of the crops that have to be reduced. **(ii) Forecasts the demand of various food commoditie**s that helps the system to guide the farmers in selecting and growing the appropriate crops to satisfy the demand and hence reducing the gap or mismatch between the demand and supply of the crops. **Map-Reduce** algorithm was used where large volume of crops have been collected into 'n' appropriated partitions and are stored in Hadoop Distributed File Systems

[9] The GSP Algorithm in Dynamic Cost Prediction of Enterprise. To provide decisional information, it was important to do the reasonable cost analysis and cost control. The GSP algorithm first decided the support degree of the database that was to decide the number of data sequence in the transactional database. GSP algorithm scans the sequence database, and get a sequence mode with length 1, then regard it the initial seed set. It produces the candidate sequence mode with length +1 by connection and cutting operation. Then scanning the sequence database, counting the support number of each candidate sequence mode. Then the sequence mode with length +1 will be produced and regard it the new seed set. The above process will be repeated until no new sequence mode or new candidate sequence mode is produced

Prediction of the most favourable market with its associated trading price around the markets of Bangalore.

The main aim is not to prove that above mentioned methods are best, but it is to show that they are applicable to daily life problems. Using our model which is primarily based on the Map-Reduce model enables the farmers to chose their favourable markets on basis of their accountability. It enables the farming community make decisions and thus making profit is achievable.

3. METHODOLOGY AND DATASET

For their economic growth, an agro-based nation relies on agriculture. As the country's population increases dependence on agriculture, it also improves and affects the country's subsequent economic growth. Prediction is a forward-looking statement [10]. Agricultural commodity pricing has become the hour's need for farmers. Although future occurrences are unsure, it is not feasible to predict precision. This paper includes a model of decision-making support that can be helpful in predicting prices for farmers.

Our system proposes the analysis of market demands and prediction of demand and cost of goods for farmer. Market analysis will be done of the basis of

- Demand of crop
- Population
- Transportation facilities
- Market location

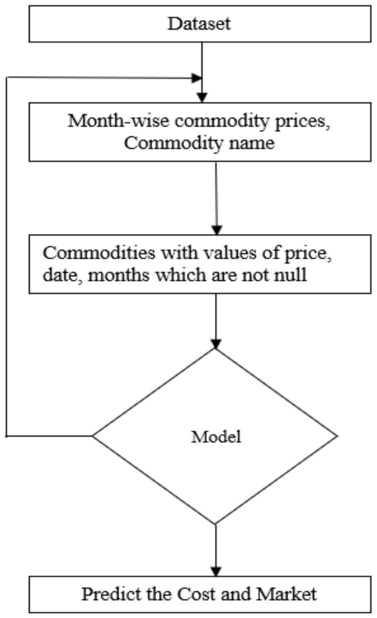
Dataset
Month-wise commodity prices, Commodity name
Commodities with values of price, date, months which are not null
Model
Predict the Cost and Market

Fig. 1.Proposed Method Work flow

Figure1. shows that the markets must enter the name of the commodity and the prior crop selling price. Based on past prices, our model will be able to provide average rates and market for a specific crop to help farmers make better choices and sell to the market accordingly.

3.1. Dataset and Study Area:

Many variables such as environment, supply and demand, etc., affect market prices of agricultural goods. It is more complex to predict than business products. It is hard to correctly and timely obtain the information of the impact factor [11]. Thus, in this research, information on agricultural product prices gathered as experimental materials for the city of Bangalore, India. The daily cost for the year 2011-2012 was gathered from the wholesale market of Bangalore City. The monthly price is gain to prepare for forecasting after it has been weighted. Dataset has 4 characteristics and 551948 samples in total. The explanations for the input characteristics are shown in

Table 1.

Attributes	Explanation
Item_Name	Commodity Name
Price	Commodity Price
Datesk	String in date, month, year format
Date	Date of the sale

3.2. Proposed Methodology:

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables and the single output variable. More specifically, the output variable can be calculated from a linear combination of the input variables. When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression [12]. Random Forest Regressor is a tree based algorithm that uses the features of multiple decision trees for making decisions. It merges the output of multiple decision tress to generate a final decision output [13]. The dataset will be tested with the two models mentioned above and the model with the best accuracy will be used for the prediction of crop cost and market.

The data has to go through the process of extraction and cleaning for removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

First part: Data Cleaning

1) The data should be first cleaned and all the null, repeated, unformatted values should be removed.
2) After removing the unwanted data, we are left with 395650 samples in total. Earlier there were 551948 samples
3) The datesk and date columns in the datset are terminated and converted into 2 new columns named Month and Date. This is done to improve data classification.
4) This data will be then used to perform One-hot-encoding

Second Part: One Hot Encoding

The parameters of the models indicates that it requires commodity names as input variables and its importance will be the output variable. The commodity names having samples less than 2048 will be discarded so as to maintain data consistency for the model. The commodity names will be converted to their respective columns in the dataset. The resultant dataset will consist of commodity name columns, day and month. The commodity which was sold on that day will be valued as 1 where as if its not sold then it will be valued as 0. This process is known as One-Hot Encoding

Third Part: Deciding Models

1. Random-Forest Regressor model:

   Import the train-test split module from the skilearn.model_selection library. It splits arrays or matrices into random train and test subsets[14]. The input parameters includes two arrays, test_size, train_size, random_state. It returns list containing test-train split of inputs. The resultant dataset in the earlier part will splitted into train-data and test-data. The train-data will consist of 80% samples whereas test-data will consist the remaining 20% data. The Random-Forest Regressor Model will be applied on both train and test-data and their accuracy of the prediction will be calculated. We'll import the RandomForestRegressor module from sklearn.ensemble library. The sklearn.ensemble module includes ensemble-based methods for

Model. The output is stored in another file having extension '.pkl'(indicates that it's a pickle file). The advantage of the Pickle library is that it serializes the object first before writing it to the file. Pickling is the way to convert python object into a character stream. The character stream contains all the information necessary to reconstruct the object in another python script[16]. Another advantage of using Pickle that keeps track of the objects it has already serialized, so later references to the same object won't be serialized again classification, regression and anomaly detection[15]. In the above code we can see the train data is used as a parameter for the RandomForestRegressor module

reg_rf = RandomForestRegressor()

reg_rf.fit(X_train, Y_train)

where X_train and Y_train are the train-data inputs.

'.fit' is used to insert the inputs into the model.

reg_rf = RandomForestRegressor()

reg_rf.fit(X_test, Y_test)

where X_test and Y_test are the test-data inputs.

The calculation of accuracy can be done using the '.score' method.

reg_rf.score(X_train, Y_train)

reg_rf.score(X_test, Y_test)

By using the above code, the accuracy of this model is calculated. For our dataset the accuracy comes out to be 88 percent for trained data and 80 percent for test data.

1) Linear Regression Model:
Import the train-test split module from the skilearn.model_selection library. The Liner Regression Model will be applied on both train and test-data and their accuracy of the prediction will be calculated. We'll import the LinearRegression Module module from sklearn.ensemble library. In the above code we can see the train data is used as a parameter for the LinearRegression module

reg_rf = LinearRegression()

reg_rf.fit(X_train, Y_train)

where X_train and Y_train are the train-data inputs.

reg_rf = LinearRegression()

reg_rf.fit(X_test, Y_test)

where X_test and Y_test are the test-data inputs.

The calculation of accuracy can be done using the '.score' method.

reg_rf.score(X_train, Y_train)

reg_rf.score(X_test, Y_test)

By using the above code, the accuracy of this model is calculated. For our dataset the accuracy comes out to be 87 percent for trained data and 85 percent for test data.

From the above accuracy its noticed that Linear Regression Model has the better accuracy and thus can be used as a firm model for the cost and market prediction. Linear Regression model came out to be accurate in both train and test data as noticed above. As the model is

decided we can easily use its output and can use for accurate cost and market prediction.

decided we can easily use its output and can use for accurate cost and market prediction. Pickle library is used to store the output generated by the Linear Regression

Third Part: Using Flask:

Flask is a web-application framework designed to make getting started quickly and easy. It has ability to scale up complex applications. It enables to write applications without worrying about low-level details such as protocol, thread management, and so on[17].

Flask will be used to portray our model in the form of a web-application. Its mainly used to make complex data look easier by using various tools. First we'll write our logic in the 'app.py' file provided by the Flask. In 'app,py' we'll open the 'pickle' file as mentioned in the earlier part. Python will be able to read the '.pkl' file and will then reconstruct them into objects. The template folder will contain all the static html pages used for making a web-application. The web-application will ask terms such as date, location, commodity-name

(the commodity that the farmer wants to sell) using form-fields.

The input of this form will be used in the 'app.py' file for predicting the crop cost and favourable market for that crop.

In 'app.py' file using the data provided by the pickle file, we'll predict the cost of the crop. The predicted cost is then displayed on the web-page after clicking the submit button. On clicking the submit button the 'app.py' file will run and will perform prediction based on the input data.

The advantages of this method are:

- Training the data is very capable apart from the volume of data set.
- Both the models gave a clear perspective about the accuracy generated in their prediction which enabled us to select the best model.
- Linear regression has a considerably lower time complexity when compared to some of the other machine learning algorithms. The mathematical equations of Linear regression are also fairly easy to understand and interpret.
- Linear Regression is used to find the nature of relationship between two variables.
- Storing data in a pickle file enables the 'app.py' file to read it from anywhere and do operations with it.
- Flask makes it easier to develop web-applications which use a complex data structure behind.
- Flask is lighter than any other framework as its codebase size is relatively smaller. This results in speed working of the application

4. RESULTS:

With an accuracy of 87 percent for the train-data and 85% with the test-data, Linear Regression model was able to predict the commodity price and market better than the Random-Forest-Regressor Model. the motivation of the proposed application is to recommend the right price and right market to the farmers based on factors of location, sale of the crops in the markets and storage facilities. The accuracy calculation should be a very essential component for classification algorithm. It will testify the past active classification algorithms are moreover "Right" or else "Wrong". The performance of the proposed model is compared and measured with other existing algorithms. The performance is measured by calculating accuracy, rate of error and time of execution.

The result facilitates that the farmers around Bangalore district are getting suggested of the markets of their convenience. The market which provides all favourable conditions right from cost, to transport and storage facilities are getting suggested to the farmer which makes them aware of the market trends and a perfect view of the crop-demand in the market. Also, its clear that the Linear Regression achieves better performance than any other algorithms. So the proposed approach is better for large datasets.

5. CONCLUSION AND FUTURE DIRECTIONS:

In accounting for better economic condition of farmers, commodity price and selection of the best market becomes important. Therefore analysing trends in the commodity prices in each and every market becomes important. Some markets may have a sale of lot of tonnes of foodgrains where as in some markets the foodgrains just remain lying for days and get spoiled. If farmer can chose in which market he has to sell by viewing its cost prediction then it will increase his knowledge of the sales made by the market and can do his farming accordingly.

In this paper the Linear Regression Model created and assessed the intelligent scheme for short-time forecast.

The proposed model has also some constraints in relation to the characteristics mentioned above. Implementing market forecasting based on cost predictions leads to step-by-step time consuming procedures.

The primary objective in the future will be to prevent any type of agricultural loss by utilizing each and every crop of the farmers by selling it to different kind of markets as per their predicted costs in our platform. Farmers should only focus on their farming and should not face difficulties in selling their crop and face loss economically and socially. The Linear Regression Model can be combined with any other better algorithm to increase its accuracy to a certain level. It will help to further enhance the precisions of predictions.

6. REFERENCES

1. Shagun Kapil, https://www.downtoearth.org.in/news/agriculture/agri-share-in-gdp-hit-20-after-17-years-economic-survey-75271.
2. Y. S. Son, R. Baldick, K. Lee, and S. Siddiqi, https://www.researchgate.net/publication/3267183_Short_Term_Electricity_Market_Auction_Game_Analysis_Uniform_and_Pay-as-Bid_Pricing.
3. https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/commodity-valuation/
4. https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/
5. https://www.researchgate.net/publication/338855496_Developing_an_Agricultural_Product_Price_Prediction_Model_using_HADT_Algorithm
6. Developing an Agricultural Product Price Prediction Model using HADT Algorithm
7. https://www.semanticscholar.org/paper/No-.-11-STUDY-OF-AGRICULTURE-MARKETING-INFORMATION-Dagar/d9c179a5595a0f8a3b732f7981cd3e245a6f9900
8. https://www.researchgate.net/publication/325420196_Demand-

prediction_model_for_forecasting_AGRI-needs_of_the_society

9. https://www.researchgate.net/publication/221163258_The_GSP_algorithm_in_dynamic_cost_prediction_of_enterprise
10. Population Dynamics in India and Implications for Economic Growth https://core.ac.uk/download/pdf/6494801.pdf
11. https://www.everycrsreport.com/reports/RL33204.html
12. https://courses.lumenlearning.com/boundless-statistics/chapter/multiple-regression/
13. https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/
14. http://scikitlearn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
15. http://scikitlearn.org/stable/modules/ensemble.html
16. https://github.com/pfnet/pkl
17. Flask_Documentation, https://pypi.org/project/Flask/