International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

A Study of Location Data on Hadoop Spark Framework, Implementation & Identification of Parameters to Optimize Profit Estimation of Mandis (Market Places)

Miss Anju Shravankumar Yadav
Department of Computer Science and Engineering, SOCSE, SandipUniversity, Nashik, India

Mr. Vipin K. Wani
Assistant Professor, Department of Computer Science and Engineering, SOCSE, SandipUniversity, Nashik, India

Mr. Aditya Kumar Sinha
Associate Director(ACTS & HPC) HOD-ACTS CDAC, Pune, India

Abstract- Agriculture is the primary source of livelihood for about 58 per cent of India's population. Farmers face many problems after they have yielded the agricultural commodities: farmers sell at a very low rate, and consumers buy at non-reasonable rates due to middlemen. As most of the farmers sell their produce to mandis (market places), they find it difficult to estimate which of the nearby places will be more profitable to them. The aim of this research is to help the farmers find the most profitable marketplace on the basis of location, transportation cost and price of the commodity and many other parameters that affect the price. Pattern finding and analysis can be done to find out hidden insights from the data that can help to study the economic growth in agricultural sector. The data is downloaded from the agmarknet.gov.in website operated by the Indian govt, which has regular updates for about 30 commodities from each state and marketplace associated with it in India for several years.

Keywords— Agriculture Commodity, Apache Spark, Factors affecting price, Location data, Optimized prediction, Price Forecast

I. INTRODUCTION

Mandi in Hindi language implies commercial centre. Customarily, such commercial centres were for nourishment and agri-wares. In any case, after some time the inclusion of mandis got augmented to incorporate exchanging centre points for grains, vegetables, timber, jewels and precious stones; pretty much every tradable was incorporated. Mandis for creatures like steers, goats, ponies, donkeys, camels and wild oxen, and poultry are regularly sorted out as fairs. Accordingly, the word mandi accept the shapes of a catch-all commercial centre where anything is purchased and sold.

Regardless of the way that agribusiness represents as much as a fourth of the Indian economy and utilizes an expected 60 percent of the work constrain, it is considered profoundly wasteful, inefficient, and unequipped for taking care of the yearning and lack of healthy sustenance issues. Regardless of progress around there, these issues have kept on baffling India for quite a long time. It is assessed that as much as one fifth of the complete agrarian yield is lost because of wasteful aspects in collecting, transport, and capacity of government-financed crops.

India's horticulture is made out of numerous yields, with the principal nourishment staples being rice and wheat. Indian ranchers additionally, develop potatoes, sugarcane, oilseeds, and such non-nourishment things as cotton, tea and jute (a shiny fibre used to make burlap and twine). India is a fisheries mammoth too. An absolute catch of around 3 million metric tons yearly positions India among the world's main 10 angling countries.

Regardless of the mind-boggling size of the horticultural area, notwithstanding, yields per hectare of harvests in India are commonly
low contrasted with worldwide principles. Inappropriate water the board is another issue influencing India's horticulture. During a period of expanding water deficiencies and ecological emergencies, for instance, the rice crop in India is apportioned lopsidedly high measures of water. One consequence of the wasteful utilization of water is that water tables in districts of rice development, for example, Punjab, are on the ascent, while soil fruitfulness is on the decay. Disturbing the farming circumstance is a continuous Asian dry spell and severe climate. In spite of the fact that during 2000-01 a rainstorm with normal precipitation had been normal, possibilities of farming creation during that period were not viewed as splendid. This has somewhat been because of moderately negative appropriation of precipitation, prompting floods in specific pieces of the nation and dry seasons in some others.

This survey paper studies that which factors are important for the prediction of the agricultural commodity. Also, it studies which technology should be preferred to process such huge amount of data.

II. METHODS AND TECHNOLOGY

A. Big Data Analytics

Big data analytics is the often-complex process of examining large and varied data sets, or big data, to uncover information, such as hidden patterns, unknown correlations, market trends and customer preferences that can help organizations make informed business decisions.
In the research paper we will examine the price data of an agricultural commodity and try to forecast the prices for coming days by identifying different parameters to optimize profit.

B. Hadoop Framework (Spark)

Apache Spark is an open-source distributed general-purpose cluster-computing framework. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. It will be used to store the data in a cluster environment. Further processing will be done using HIVE tool of Hadoop, to perform queries and data integration.

C. Octaparse

The supply data was an integrated data, that had to extracted from different URL's by clicking repeatedly. Hence a web scrapping tool was required to extract the daily data of supply in the market of brinjal.

D. Forecasting Profit Estimation

In virtually every decision they make, executives today consider some kind of forecast. Sound predictions of demands and trends are no longer luxury items, but a necessity, if farmers are to cope with seasonality, sudden changes in demand levels, price-cutting maneuvers of the competition, strikes, and large swings of the economy. Forecasting can help them deal with these troubles; but it can help them more, the more they know about the general principles of forecasting, what it can and cannot do for them currently, and which techniques are suited to their needs of the moment.

E. ARIMA Model & RNN

ARIMA model is used to forecast mostly univariate variables, when other factors are to be considered minimal. This model uses the historical prices of the variable to predict the prices for the coming week or month. The model in general has three parameters to consider, p (auto correlation), d (differencing), q (moving average). p gives us the value for number of lags, as to how many days previous values should be considered for the output. d is differencing, which needs to be done if the series is not stationary. And q is the moving average which denotes that the output variable is linearly depended on the inputs.

Neural networks can also be implied to such problem statements.

The factors responsible for the price prediction can be given as inputs to the networks. Large amount of data is required to train such a model, around 10 to 15 years of data. A single hidden layer can be used to have better accuracy. The model can learn linear and nonlinear relationships between the factors and the prices which ARIMA couldn't do. This is done in the training phase of the model. Even Neural networks tend to overfit.

F. Dataset Description

The datasets were a time series multi-variate data. All values were numerical values. The missing values were replaced by the previous days data. I have accounted for 3 years of data and took daily values for all the factors.

III. LITERATURE SURVEY

Haoyang Wu1, Huaili Wu, Minfeng Zhu, Weifeng Chen and Wei Chen [1] This paper forecasts weekly prices for some agricultural commodities in markets of China, by considering daily prices of data of one year. It implements a hybrid model. One part of the model considers Time and Space related factors, the other is concerned with predicting the price when there is a sudden rise in the prices of the commodities. The time factor model considers the historical prices of the commodity and forecasts the prices. The ARIMA model is used to implement the time factor. The space model considers the fact that other nearby agricultural markets also influence the price in a market. Here the Partial Least Squares method is used. The urgency model which deals with the sudden increase in prices, is implemented using Back propagation neural network. This considers the factors such as weather (temperature, humidity, wind speed, rainfall), international oil prices, and exchange rate (Sino-US). These factors have an accumulating effect on the change of prices, hence when these factors reach a certain degree, the prices rise all of a sudden. This paper defines the urgency model and tries to forecast when will there be a sudden rise in prices. This model was tested on many other commodities as well and it turned out that it had accurate results for a variety of agricultural products.

Myat Cho Mon Oo, Thandar Thein [2] Hadoop is been used to process huge amounts of data which may be structured, unstructured or semi-structured, also getting generated

constantly with high speed. Traditional tools are not able to process such data. This paper studies how Spark can be used to process such data by optimizing the hyperparameters for the scalable machine learning (Random Forest) Algorithm rather than using the default parameters. First Hadoop Spark is setup which has one master node and 3 slave nodes on Linux distributions. The dataset is processed and the algorithm of random forest is implemented on three different datasets (HPC, Google, DAS). The results which we get suggests that the number of trees that should be used in this ensemble learning was 128, beyond that no significant changes were seen and the maximal depth of the tree was 8 as per hyperparameter setting. More than this could result in overfitting. Hyperparameter optimization gives better results and reduced error rates. Features that affect the dataset are found out using dimensionality reduction technique, and only those are included for processing.

Xiangtuo Chen, Benoit Bayol, Paul-Henry Cournede [3] This paper predicts the wheat production in France using Weighted Regression method. Prediction of certain crops is important so that government can plan their policies beforehand. The Random Forest and Lasso Regression method gives good prediction and performance, where climate data is taken into account for predicting the production of wheat. The mean absolute error for the prediction is around 5.5% and achieves a good accuracy.

Ananthi Sheshasaayee, JVN Lakshmi [4] This paper does a comparison study between Apache Hadoop for MapReduce versus Apache Spark. Different algorithm categories like Classification, Regression, methods etc are used in various medical and science streams. This novel approach studies ensemble learning on Apache Spark platform to study that whether there is optimization for time and space factors. Jupyter environment is used to run the machine learning algorithm and Apache Spark automatically parallelizes the task. The value of temperature is predicted using a tree-based model. It is observed that Apache Spark performs much better than MapReduce Hadoop in terms of execution time and space utilization. Factors like humidity, moisture, fog, pollution will be considered in future times, to predict the temperature more efficiently.

Aakash G Ratkal, Gangadhar Akalwadi, Vinay N Patil and Kavi Mahesh [5] The agricultural yield is not well organized and planned by farmers in this changing era of global warming and environment. The contribution is just 14 % of the GDP, in spite of half of the population of India is dependent on it. This paper uses data analytics to help farmers in knowing which crop to produce, depending on the soil conditions and various other factors. The previous year's crop production is used to determine the next year's prices of the products. Factors like temperature, humidity, inflation, rainfall are extrapolated for the year we need to find the prices considering their values for the previous years. The dataset uses data between 2004-2013, for Karnataka state. The model used to predict the price of the agricultural commodity is nonlinear multiple regression technique. Also, the tool suggests crop rotation to the farmer so that the same crops are not grown again and again, which can have adverse effects. In one way it can affect the soil's nutrients, and can lead to deficiency of some nutrients in the soil, and secondly can lead to the price drop of the crop if many people produce the same crop again and again.

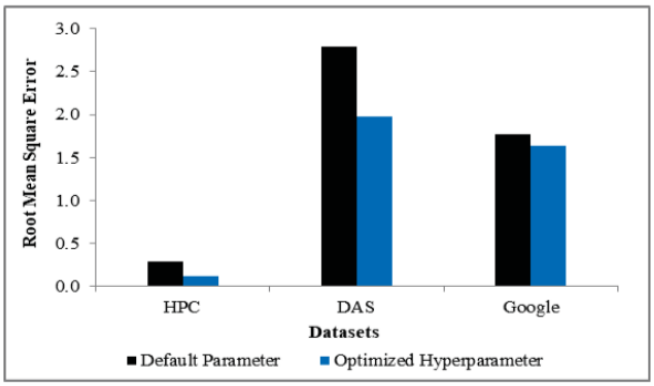

Fig-1 Graphical Representation of error rate [2]

Jorge Veiga, Roberto R. Exp´osito, Xo´an C. Pardo, Guillermo L. Taboada, Juan Touri˜no [6] MapReduce frameworks are now replaced with Spark or Flink which improves the programming API's that we use along with performance. This paper makes a comparison between the three by considering parameters like HDFS block size, input data size, interconnect network. The data used is static in nature, although there will be quite a difference if we use streaming data. The setup is done in standalone on the computers. The benchmarks used in the paper are Word Count, Grep, Tera Sort, connected components, Page Rank, K-means. There should be a fair comparison so that the three frameworks can be analysed as to which one is better. The parameters discussed above remain same for all of them. The different benchmarks are applied to test the scalability and performance. Also, the code of benchmarks is different for all the three frameworks, but they are written with proper optimization so that they can have the same output. Once the algorithm runs, we can compare which performed better by seeing the execution time, space complexity and other factors. The results show that the replacing Map Reduce with Spark or Flink can give

better results about 77% reduction in execution time on an average. In general terms we can conclude that Apache Spark is a good choice, as it is more mature and has a great shared community. The need to rewrite the code when converting from Map Reduce to Spark should also be taken into account.

IV. PROPOSED WORK

We have data for prices of agricultural commodities of all states. The data is more than 10 years old. The different factors that affect commodities are Weather, Pests/Diseases, Transportation Costs (Profit optimization), Labour Costs, Govt policies and programs, Supply-demand relation (inflation), Exchange rates, Petroleum price change, Festival. Using some of these factors we can predict the weekly price of an agricultural commodity.

We need to categorize crops, according to their similarities: then apply same algorithm for each category. Example we can take categories like Fruits, Spices, Cereals, Pulses, Nuts, Meat, Oil, Livestock, eggs, Animal Products, Aquatic products. And then apply the same concept on all states (big data pre-processing), and process them on different machines. There are 2 candidate Algorithms that can be used for predicting the prices of the commodities. Two agricultural commodities have quite different price change trends [1]

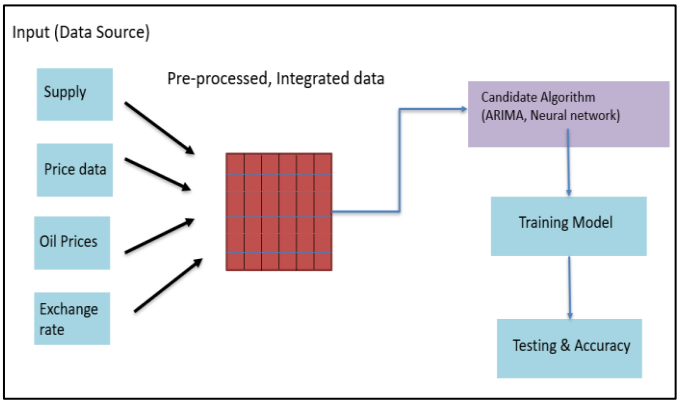

Fig-2 Architecture of the system

V. EXPERIMENT

We have to setup a computer node with Spark/ Hadoop on the systems, on the basis of data available. Run the Algorithm on the dataset and do analytics. The strategy for the evaluating the performance was first running the ARIMA model to use its result as a benchmark and then run and analyze the output of RNN model. First, we have taken data for 3 years for brinjal production in Maharashtra. The price data had to be scrapped via Octaparse, which contained three years of data. Commodities can be forecasted for a week, based on markets in a city. And then can be progressively applied to other cities.

VI. RESULT ANALYSIS

The ARIMA Model was used as a benchmark to compare the result for an RNN Model.

As due to constraints like, not able to setup a fast processing cluster in the lab, the setup of all the software and frameworks were done on a standalone mode on a single PC.

The output and the comparison of both the graphs can be done. And it is seen that as the data constraint was there, where we could not use the neural network model with large amount of data, the results are similar to the ARIMA, which is itself with high error.

In the ARIMA model the black line shows the prediction while the pink line is the actual values.

In the RNN model the blue dots represent the true values and the red dots are the predicted values.

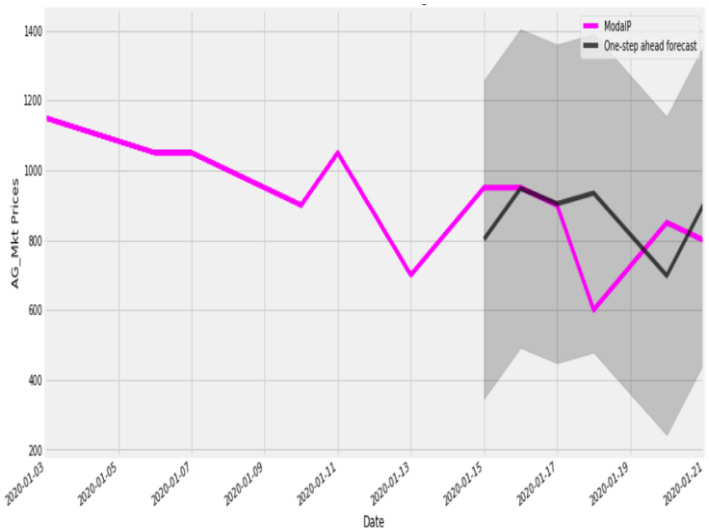

Fig-3 ARIMA Implementation

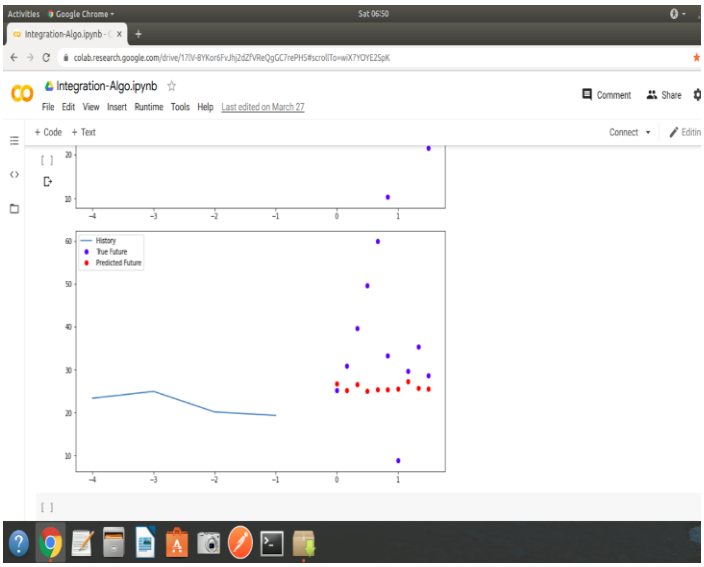

Fig-4 RNN Implementation

CONCLUSION

In this paper study has been done regarding different parameters affecting the price of various agricultural products for farmers. Hadoop framework is used to store data which is downloaded from the government website. Different tools will be used for the integration and processing of queries to find hidden pattern using setup done on the Hadoop Spark environment. Using the ARIMA and RNN model, forecasting can be done to estimate the profit optimization of their agricultural produce of mandis for farmers in India.

Most of these exogenous factors cannot be mathematically expressed like festivals, govt policies, pests etc. This is the reason, why accurate results cannot be made in certain cases. Algorithms like ARIMA, Neural Network models like RNN have been applied to different datasets of agriculture. Traditional methods cannot deal with such huge amount of data pre-processing hence Apache Spark should be used to analyse data at a large scale. Most of them have been able to achieve above average accuracy.

As the setup was installed on my computer, which could run only on a specific amount of data, hence taking 3 years of data lead to high error in the results, as neural networks require very high amount of data for better prediction.

FUTURE SCOPE

Based on the research done and the prototype implemented, it is possible to extend and enhance this proposed architecture in a way to include more factors so that can improve the overall accuracy of the prediction of the prices.

REFERENCES

1. Haoyang Wu1, Huaili Wu, Minfeng Zhu, Weifeng Chen and Wei Chen, "A new method of large‑scale short‑term forecasting of agricultural commodity prices: illustrated by the case of agricultural markets in Beijing", Wu et al. J Big Data, 2017.

2. Myat Cho Mon Oo, Thandar, Thein, "Hyperparameters Optimization in Scalable Random Forest for Big Data Analytics", IEEE 4th International Conference on Computer and Communication Systems, 2019.

3. Xiangtuo Chen, Benoit Bayol, Paul-Henry Cournede, "Application of Weighted Regression for the Prediction of Soft Wheat Production in France",6th International Symposium on Plant Growth Modeling, Simulation, Visualization and Applications (PMA), 2018.

4. Ananthi Sheshasaayee, JVN Lakshmi, "An insight into tree-based machine learning techniques for big data Analytics using Apache Spark", International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2017.

5. Aakash G Ratkal, Gangadhar Akalwadi, Vinay N Patil and Kavi Mahesh,"Farmer's Analytical Assistant", IEEE International Conference on Cloud Computing in Emerging Markets, 2016.

6. Jorge Veiga, Roberto R. Exp´osito, Xo´an C. Pardo, Guillermo L. Taboada, Juan Touri˜no, "Performance Evaluation of Big Data Frameworks for Large-Scale Data Analytics", IEEE International Conference on Big Data (Big Data), 2016.