Quest Journals
Journal of Software Engineering and Simulation
Volume 6 ~ Issue 1 (2020) pp: 01-13
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Crop Price Prediction System using Machine learning Algorithms

Pandit Samuel
(Assistant professor, Department of Information Technology , ANITS , Sangivalasa ,Visakhapatnam)

B.Sahithi , T.Saheli , D.Ramanika , N.Anil Kumar
(B.Tech , Department of Information Technology , ANITS , Sangivalasa ,Visakhapatnam)

ABSTRACT: *Price Prediction, nowadays, has become a very important agricultural problem which is to be solved only based on the available data. The aim of this paper is to predict the crop price for the next rotation. This work is based on finding suitable data models that helps in achieving high accuracy and generality for price prediction. For solving this problem, different Data Mining techniques were evaluated on different data sets. This work presents a system which uses data analytics techniques in order to predict the price of the crop. The proposed system will apply machine learning algorithms and predict the price of the crop based on multiple factors like Area harvested, Area planted etc. This provides a farmer with an insight of what the future price of the crop that he is going to harvest. Thus, the paper develops a system by integrating data from various sources, data analytics, prediction analysis which can help predict the target price of the crop and increase the profit margins of farmer helping them over a longer run. The complete research comes up to a conclusion that XGBoost is the suitable technique for our project.*

KEYWORDS: *Data Analytics, Prediction, Machine Learning, Linear Regression, Decision Trees, XGBoost, Neural Networks*

R*eceived 04 Apr., 2020;* ***A****ccepted 19 Apr., 2020 © The author(s) 2020.*
P*ublished with open access at* ***www.questjournals.org***

I. INTRODUCTION

Agriculture is the backbone of every economy. From ancient period, agriculture is considered as the main and the foremost culture practiced in any region. There are multiple ways to increase and improve the crop yield and the quality of the crops. Data mining is also useful for predicting the crop price. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Crop price prediction is an important agricultural problem. Each and every farmer always tries to know, how much price he will get from his expectation. In the past, price prediction was calculated by analyzing farmer's previous experience on a particular crop. Accurate information about history of crop yield is an important thing for making decisions related to price prediction of the crops. Therefore, this paper proposes an idea to predict the price of the crop. Data Analytics is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. Earlier, crop price prediction was performed by considering the farmer's experience on a field and crop. However, as the conditions change day by day very rapidly, farmers are forced to cultivate more and more crops. Being this as the current situation, many of them don't have enough knowledge about the losses that might incur and are not completely aware of the benefits they get while farming them. The proposed system applies machine learning and prediction algorithm like Logistic Regression, Decision Trees, XGBoost, Neural Nets, and Clustering to identify the pattern among data and then process it. This in turn will help predict the target price of the crop.

II. LITERATURE STUDY

A Survey on Crop Prediction using Machine Learning Approach:

The paper, A Survey on Crop Prediction using Machine Learning Approach[1] revolves around the idea of implementing techniques with the help of technical knowledge and improve the conditions of the farming sector by making it more reliable and inculcating it among the farmers to correctly predict the suitable crops according to the results obtained using certain machine learning techniques which takes into consideration of the factors like- soil, weather and the trends in the market. Certain conditions are also take into consideration as the PH, Nitrogen levels and the nutrients constitution in the soil. The machine learning algorithms are used

for the prediction which are Artificial Neural Networks, Information Fuzzy Network and Data Mining techniques. Finally, it is seen that Artificial Neural Network is the suitable technique for the project.

Predicting Yield of the Crop Using Machine LearningAlgorithm:

The paper, Predicting Yield of the Crop Using Machine Learning Algorithm [2] is based on the central theme that certain factors like weather conditions, soil parameters and the historic details of the crop has an effect on the yield of the crop in the present. So, it is important to take these factors or the parameters into consideration and predict the yield of the crop planted. Certain Machine Learning algorithms are taken into account for developing the models of the data obtained from the past historical yield of the crop and reflecting them in predicting the yield of the crop planted in the present.This paper focuses on predicting the yield of the crop by using Random Forest algorithm. Real data of Tamil Nadu were used for building the models and the models were tested with samples.The prediction will help to the farmer to predict the yield of the crop before cultivating onto the agriculture field. To predict the crop yield in future accurately Random Forest, a most powerful and popular supervised machine learning algorithm is used.

Crop Price Forecasting System Using Supervised Machine Learning Algorithms:

The paper, Crop Price Forecasting System Using Supervised Machine Learning Algorithms [3] revolves around two main issues of agriculture- Profit and Price. This paper takes these two factors into consideration and develops a system which accurately predicts the price of the crop as well as the profit of the crop.The system comprises of two actors, the Administrator and the Agricultural Department. The Admin maintains the entire System. The Department performs two main roles, Price Prediction and the Profit Prediction. The Parameters considered for Price Prediction are- Rainfall, Maximum-trade, Minimum Support Price (MSP),Yield. The Parameters considered for Profit Prediction are- Crop Price, Yield, Cultivation Cost, Seed Cost. To predict the Price of the Crop we use Naïve Bayes Algorithm which is a Machine Learning Classification technique. To predict the Profit of the Crop we use K Nearest Neighbor(KNN) which is a Supervised Machine Learning Classification Algorithm. The Department Head can select the Crop and the Prediction methodology and provide the required parameter values. The algorithms run in the background and give the Output to the Department Heads which is in turn conveyed to the Farmers' for better preknowledge about the outcome. Thus, the System gives a beforehand prediction to the Farmers' which increases the rate of Profit to them and in turn the country's economy.

The paper we proposed deals with the price prediction of a certain crop based on the data obtained from the World Supply and Demand Estimate(WSDE) where the data relevant to the paper is obtained. The data is then processed through several steps of cleaning and transformation and made ready for the analysis process. Machine Learning algorithms are implemented to appropriately study the data and accurately predict the price. Unlike traditional statistical methods, ML does not make assumptions about the correct structure of the data model, which describes the data. This characteristic is very useful to model complex non-linear behaviors, such as a function for crop yield prediction. ML techniques most successfully applied to Crop Yield Prediction (CYP). The Machine Learning algorithms used are Logistic Regression, Decision Trees, XGBoost, Neural Nets which processes the data and predicts the price accurately. Results computed by this system are accurate as well as reliable. The study feeds data into Machine Algorithms and trains the algorithms using the training data and evaluate the test data set. The results are accurate and helps better predict the target price. It performs better as compared to traditional methods.

III. METHODOLOGY

Tools Used: The whole system is implemented using python programming language in the jupyternotebook.
Workflow: The approach for analysis is shown in Figure 1.

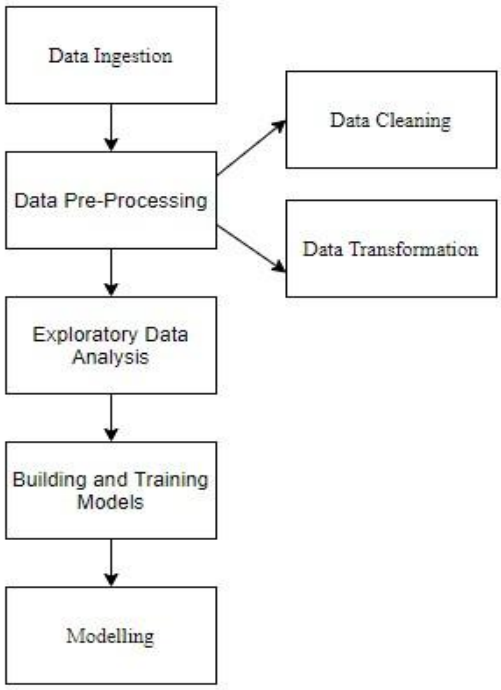


Figure 1

Data Ingestion:

Data ingestion is the transportation of data from assorted sources to a storage medium where it can be accessed, used, and analyzed by an organization. The destination is typically a data warehouse, data mart, database, or a document store. Sources may be almost anything – including SaaS data, in-house apps, databases, spreadsheets, or even information scraped from the internet. The data ingestion layer is the backbone of any analytics architecture. Downstream reporting and analytics systems rely on consistent and accessible data. There are different ways of ingesting data, and the design of a particular data ingestion layer can be based on various models or architectures.

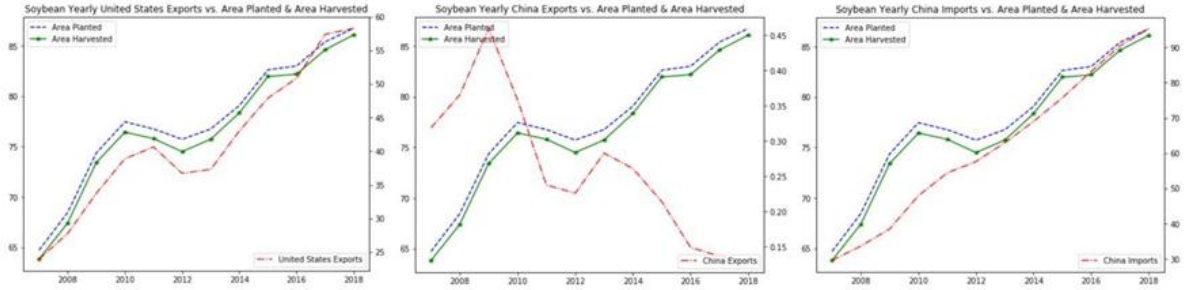
Data Preprocessing:

Data Preprocessing is a data mining technique used to transform the raw data into useful and efficient format.
The data here goes through 2 stages

1. Data Cleaning: It is very important for data to be error free and free of unwanted data. So, the data is cleansed before performing the next steps. Cleansing of data includes checking for missing values, duplicate records and invalid formatting and removing them.
2. Data Transformation: Data Transformation is transformation of the datasets mathematically; data is transformed into appropriate forms suitable for data mining process. This allows us to understand the data more keenly by arranging the 100's of records in an orderly way. Transformation includes Normalization, Standardization, Attribute Selection.

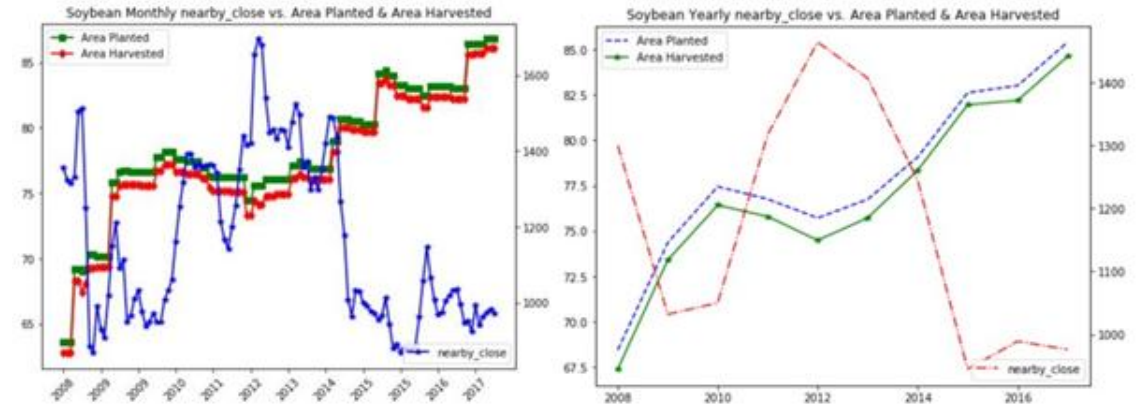
Exploratory data analysis:

Exploratory data analysis(EDA) is an approach to understand the datasets more keenly by the means of visual elements like scatter plots, bar plots, etc. This allows us to identify the trends in the data more accurately and to perform analysis accordingly.
From the yearly trends graphs it is observed that, US Exports depend on and follows the areas planted and harvested annually.
A sudden drop in China's Exports in the year 2009 is observed and in the mean time its imports kept increasing in the last 12 years regardless of the global yield, which implies China has a huge and lasting demand of soybean crop but now it relies on the global supply to meet the needs.



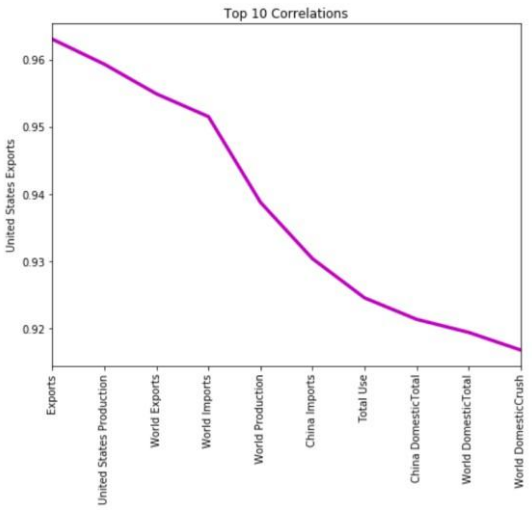
Also, from the monthly trends graphs it is observed that China's imports are historically positively correlated to the US exports trend, but in the recent years, Cina's import growth rate is becoming higher than US export increase rate, which in trun implies China also possibly sources it's supply from other parts of the world. Similar to the yearly trends, US exports roughly follows areas planted and harvested whereas China imports is in stable growth irrespective of the yield and production.

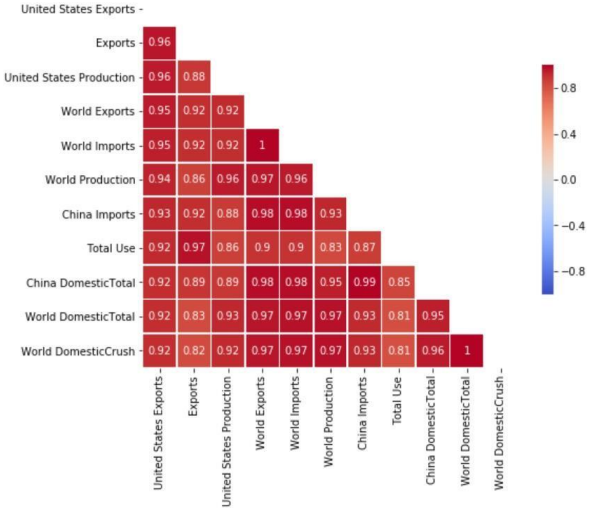
The monthly and yearly time trends allowed us to identify the relationship of price with area planted and harvested. Price is inversely proportional to the area planted and harvested



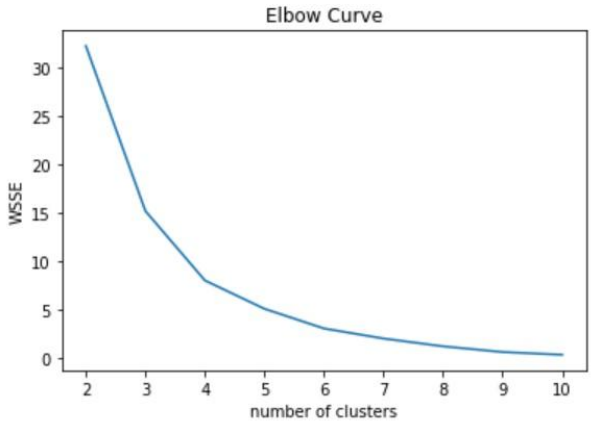
Building and Training models:

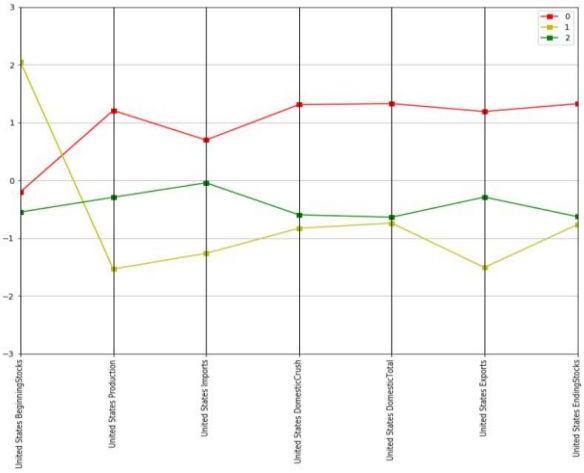
Correlation plots help explain the relationship of the factors with the target variable. Magnitude of the correlation coefficient of 0 indicates that there is no correlation between the selected attributes and a 1 indicates that the selected variables are at best correlated. It helps answer the question 'What are the main factors impacting US Soybean export?'





Clustering is used to answer the question 'Which years have similar profiles of US Soybean production, export and import?'
Elbow curve is used to find the optimal number of clusters as 3 and the visualization of the centroids of the clusters is as follows:





Cluster0: Years 2015-2018, Cluster1: Years 2007-2008, Cluster2: Years 2009-2014
Following is the analysis from Clustering,

- Each cluster has similar profiles
- Cluster2 is quite similar to Cluster1, except in 'United States Beginning Stocks'
- Cluster2 is an interesting one, similar to Cluster1 in terms of US DomesticCrush, DomesticTotal, and Ending Stocks and similar to Cluster0 in terms of US Beginning Stocks, Production and Imports which in fact implies that 2014 might be a transitional year or adjustment period between two stable patterns
- Interestingly, that group patterns matches the soybean yearly yield and market price profiles:
  - 2007 to 2008: supply started to increase and price started to drop
  - 2009 to 14: supply and price were dramatically fluctuating in reverse directions
  - 2015 to 18: supply was being in constant increase and price keeps decreasing

Modeling: Modelling of data involves creating a data model for the data to be stored in the database. The process of modeling means training a Machine Learning Algorithm to predict the labels from the features, tuning it for business need, and validating it on the hold out data. The output from modeling is a trained model that can be used for inference, making predictions on new data points. Modeling is independent of the previous steps in the Machine Learning process and has standardized inputs which means we can alter the prediction problem without needing to rewrite all our code. If the business requirements change, we can generate new label times, build corresponding features, and input them into the model. Models are implemented and later evaluated for their accuracies using root mean square error

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

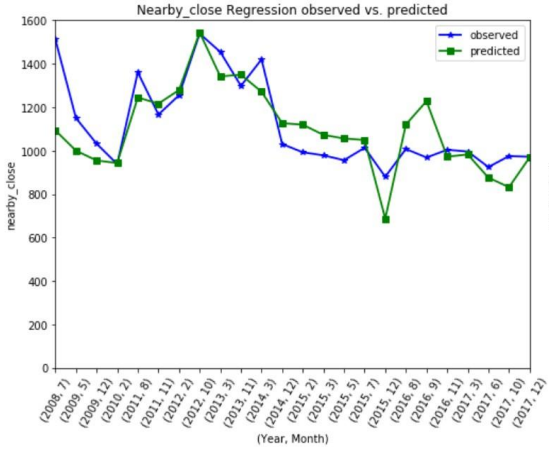
Root Mean Square Error:

The Root Mean Square Error is evaluated for every model and the accuracies are measured. It is observed that XGBoost has the best accuracy among all the models.

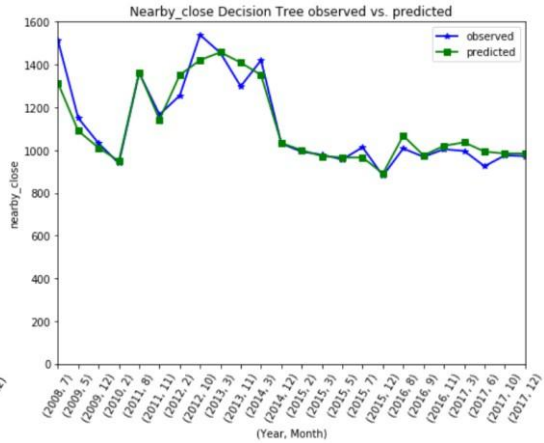
	Target mean	Target std.	Linear regression RMSE	Decision Tree RMSE	XGBoost RMSE	Neural Nets RMSE
Compare	1117.915874	205.366323	136.278178	63.778936	56.50975	551.159259

Following are the plots of actual vs predicted values for each model:

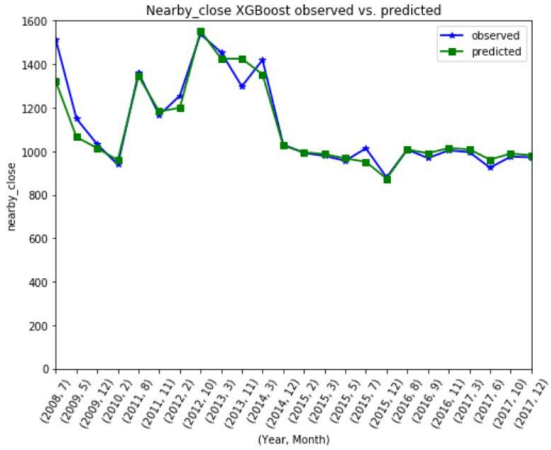
Linear Regression



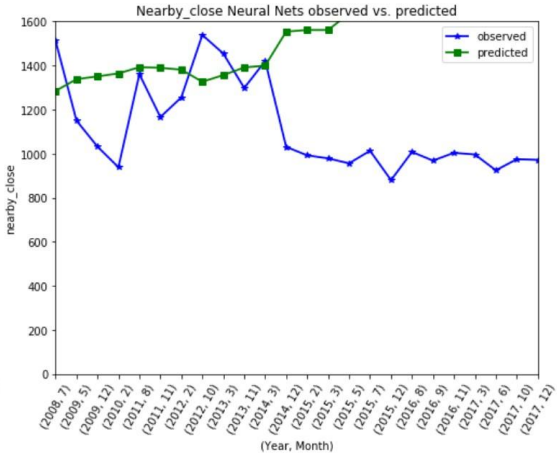
Decision Trees



XGBoost



Neural Nets



IV. CONCLUSION

In this paper certain Data Analytics techniques were adopted in order to estimate crop price analysis with existing data. Linear Regression is employed for discovering important information from the agricultural datasets. Neural Network is constructed for price prediction to increase the accuracy percentage. The root mean square error is calculated for every technique to accurately measure the accuracy of each system employed and the most accurate system is then selected. Strong dependencies between soybean yield/production and exports/imports is observed, similarly between US soybean export and other countries' import, for example, China. US soybean business has gone through 4 to 6 typical periods, from starting to grow before 2008, to big fluctuations of both yield and prices from 2009 to 2014 and to adjustment/transition in 2014 and then the stable increasing stage after 2015. It is reliable to use soybean growth areas, yield/productions and export/import profiles to predict the nearby_close market price, which is of practical values for financial purposes. XGBoost predicts the target better when compared to all the other algorithms.

REFERENCES

[1]. A Survey on Crop Prediction using Machine Learning Approach
[2]. Predicting Yield of the Crop Using Machine Learning Algorithm
[3]. Crop Price Forecasting System Using Supervised Machine Learning Algorithms
[4]. www.kaggle.com for datasets

[5]. https://en.wikipedia.org/wiki/Agriculture
[6]. https://en.wikipedia.org/wiki/Data_analysis
[7]. JeetendraShenoy, YogeshPingle, "IOT in agriculture", 2016 IEEE.
[8]. M.R. Bendre, R.C. Thool, V.R.Thool, "Big Data in Precision agriculture",Sept,2015 NGCT.
[9]. Monali Paul, Santosh K. Vishwakarma, Ashok Verma, "Analysis of Soil Behavior and Prediction of Crop Yield using Data Mining approach", 2015 International Conference on Computational Intelligence and Communication Networks.
[10]. Abdullah Na, William Isaac, ShashankVarshney, Ekram Khan, "An IoT Based System for Remote Monitoring of Soil Characteristics", 2016 International Conference of Information Technology.
[11]. Dr.N.Suma, Sandra Rhea Samson, S.Saranya, G.Shanmugapriya, R.Subhashri, "IOT Based Smart Agriculture Monitoring System", Feb 2017 IJRITCC.
[12]. N.Heemageetha, "A survey on Application of Data Mining Techniques to Analyze the soil for agricultural purpose", 2016IEEE.

Pandit Samuel,etal."Crop Price Prediction System using Machine learning Algorithms." *Quest Journals Journal of Software Engineering And Simulation, Vol. 06, No. 01, 2020, Pp. 14-20.*