DDU – GKY BHOPAL

Joseph Sriharsha and Mary Indraja Educational Society

Project On

"MSP Rate"

Minimum Support Price

"Importance of Data Science in Farming for Farmers, MSP Rate & Crop Yield Information"

Project Team Members

Sejal Mishra, Kapil Singh, Jaishree Malode, Vikash Vishvkarma, Seetu Prajapati, Shailendra Lodhi, Sachin Rangote.

GUIDED BY

PROF. Salem Shaikh

Introduction

Minimum Support Price (MSP) is the value of crops that is decided by the government of India. It is considered as an extremely essential mainstay of the pricing policy of Indian Agriculture which took off with the objective of giving value security to cultivators. It is a powerful tool in stabilizing the economy, resulting in the sustained development of the agricultural sector in the country. It is announced by the commission of agricultural cost and price to ensure for the farmers and the grain merchants the safe selling and plantation of the crops. Due to different components, cultivators wind up selling their harvests at costs underneath MSP, which either means misfortunes or exceptionally low gaining for the farmers. Hence, this paper has employed different machine learning algorithms as well as numerical techniques to predict the precise value of MSP so that cultivators could get the optimum price of their crops. Experimental results show that Lagrange's interpolation produces best results over other presented techniques for MSP prediction. This work will help farmers as well as the government to strengthen the agricultural sector of our country.

Project Problems

- MSP Price of next upcoming year.
- Decrease suicide rate.
- Recommend solve farmer's problem plan.
- Govt. decrease scarcity of capital for Farmer's.

Machine Learning

The machine learning models are built to predict crop yields by taking into consideration different factors that affect it such as MSP rate and weather data (temperature, rainfall), crop, soil moisture sensors, astronomy images etc., thus predicting accurate yield values for an agriculture field before harvest time. These techniques can be used by farmers on daily basis with high accuracy which enable them to make decisions on when to harvest crops, how much pesticide needs to be applied and what fertilizers need to be used. Deep learning models can be used to predict agriculture production in large scale with an accurate estimation of the yield. This will help farmers make important decisions related to cropping patterns and crop management leading to better yields and MSP during harvest season. Algorithms such as **multi-linear regression, Lasso regression, LightGBM, random forest, XGBoost** and **deep neural networks (CNN, LSTM etc)** have been used for crop yield predictions.

Data Set Pre-Processing

The performance of different ML algorithms strongly depend on the size and structure of the input data. Thus, the correct choice of algorithm often remains unclear unless we test out our algorithms through plain old trial and error.

Our trials consisted of different combinations of the following three parameters:

1**. Class of algorithms**: Time series forecasting, decision trees and advanced regression algorithms

2. **Number of price determinants included while training**: Ranging from using 1 to 14 determinants

3. **Data from number of mandis for training**: Training the algorithm on data ranging from 1 to 30 mandis

The choice for the classes of algorithms was based on a literature review for price forecasting methods. For time series forecasting, we selected ARIMA (Auto Regressive Integrated Moving Average), for decision trees we applied Random forest and LASSO (least absolute shrinkage and selection operator), SVM (Support vector machine) and GLM (generalized linear model) for Regression. The explanation of the various algorithms is provided in appendix A.

Data quality and pre-processing

As mentioned in the Table 1, the data collection process involves a wide variety of sources. This resulted in a time-consuming data pre-processing step as all sources have their own format and frequency of reporting. These datasets needed to be converted into a uniform 'machine readable' format.

Some of the challenges faced during this phase were:

Many datasets such as MSP, area, yield and production (APY) of other states and countries are provided in a pdf format hence data

entry cannot be automated. Additionally, the unit of measurement differs in different sources, APY can vary from district, state to a country level. Finally, the reporting frequency varies from yearly (MSP and APY) to monthly (imports and exports) and daily (currency exchange rates).

There were a few instances concerning data discrepancies in agmarknet. The types of anomalies are stated at length in the previous Kharif and Rabi pilot reports. The main challenges were:

```
In [30]: # test_pred = knn.predict(test_df)
         test_df.head()
```

Out[30]:

	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21	2021-22
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	1000.0	1080.0	1250.0	1310.0	1360	1410	1470	1550	1750	1815	1868	1940
2	1030.0	1110.0	1280.0	1345.0	1400	1450	1510	1590	1770	1835	1888	1960
3	880.0	980.0	1500.0	1500.0	1530	1570	1625	1700	2430	2550	2620	2738
4	900.0	1000.0	1520.0	1520.0	1550	1590	1650	1725	2450	2570	2640	2758

Sudden and significant shocks in prices reported

```
In [30]: # test_pred = knn.predict(test_df)
         test_df.head()
```

Out[30]:

	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21	2021-22
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	1000.0	1080.0	1250.0	1310.0	1360	1410	1470	1550	1750	1815	1868	1940
2	1030.0	1110.0	1280.0	1345.0	1400	1450	1510	1590	1770	1835	1888	1960
3	880.0	980.0	1500.0	1500.0	1530	1570	1625	1700	2430	2550	2620	2738
4	900.0	1000.0	1520.0	1520.0	1550	1590	1650	1725	2450	2570	2640	2758

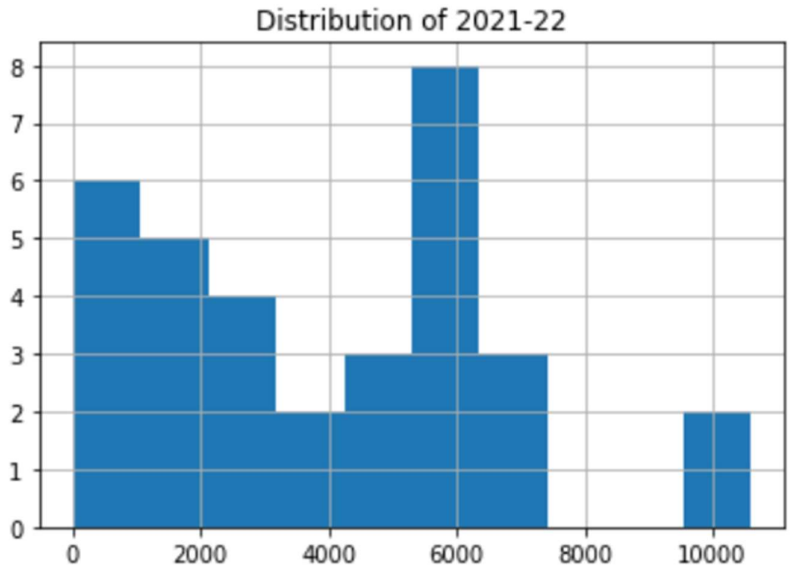
In the 2020-21 while the prices were normalized on 2021-22. This can be attributed to inferior quality of produce arrived in the market on a day. Adding a quality parameter could have explained this drop in prices at the market.

Machine Learning Model

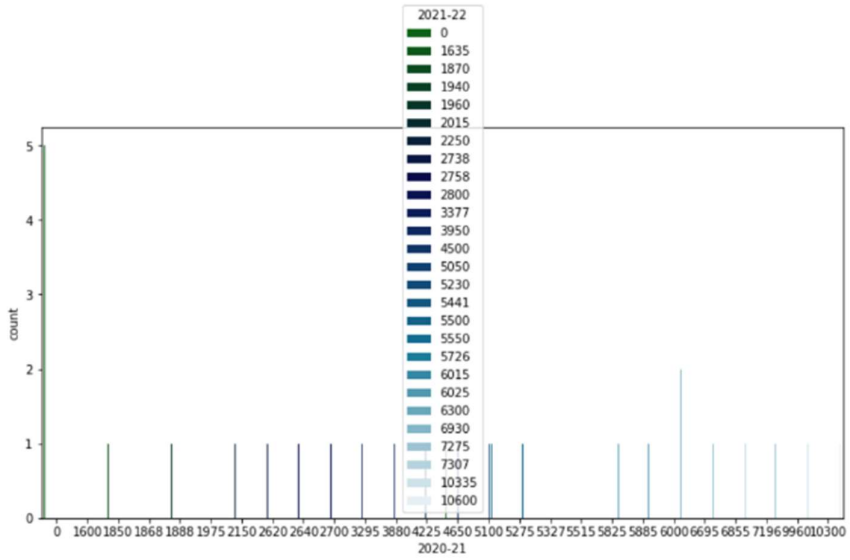
We used Support Vector Machine (SVM) for this prediction and SVM: Support Vector Machine is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, each data item is plotted as a point in n dimensional space (where n is number of features) with the value of each feature being the value of a coordinate. Then, classification is performed by finding the hyper-plane that differentiate the two classes very well.

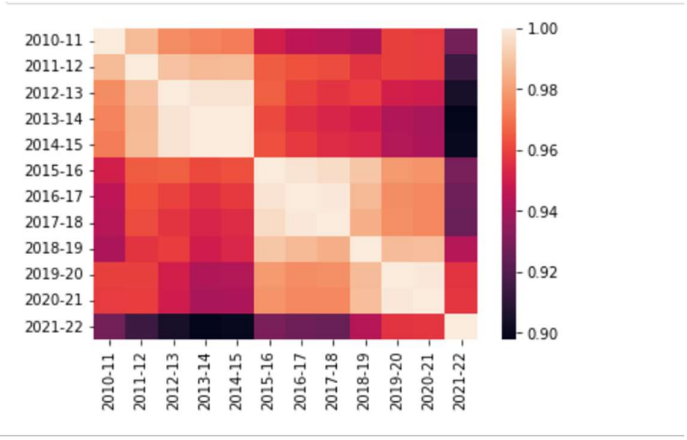
Testing Results of Various Algorithms

The evaluation metric which we chose to test our trained model was RMSE (Root Mean Squared Error). The lower the RMSE, the better the performance of the respective algorithm.

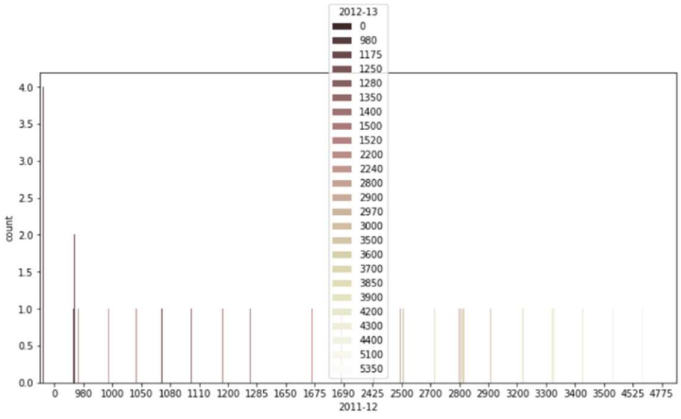


During the experiments, we discovered that Lasso is better for Maize predictions while Random Forest proved to be important for Arhar and Urad. For Urad, Maize and Soyabean, the period between 2020 and 2022 was used for testing various algorithms while for Arhar the algorithm had been tested between 2021 to 2022.





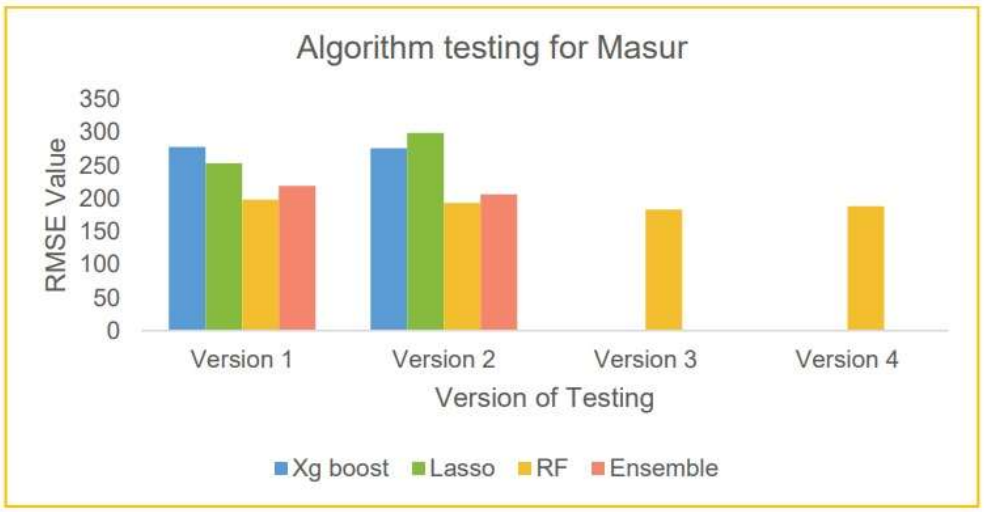
We used the Linear Regression Method to perform with data pre-processing.



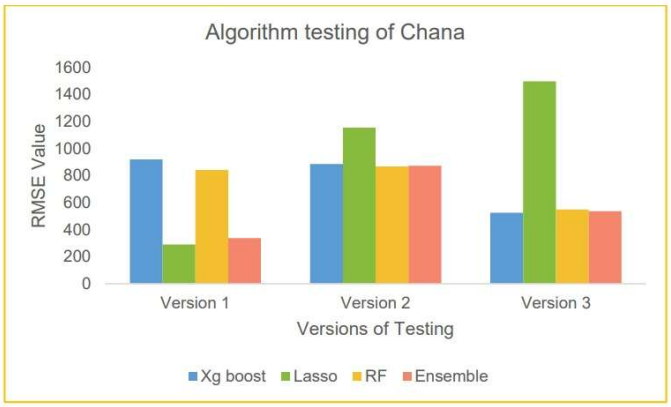
Model Evaluation

The steps followed in price forecasting includes, identifying price determinants, data collection, data cleaning and formatting, algorithm training, testing and evaluation followed by giving live predictions. On an average it takes around three months to complete the process and give prediction of mandi specific prices. The most time-consuming steps are that of data collection, cleaning and formatting (almost 30% of the total project time

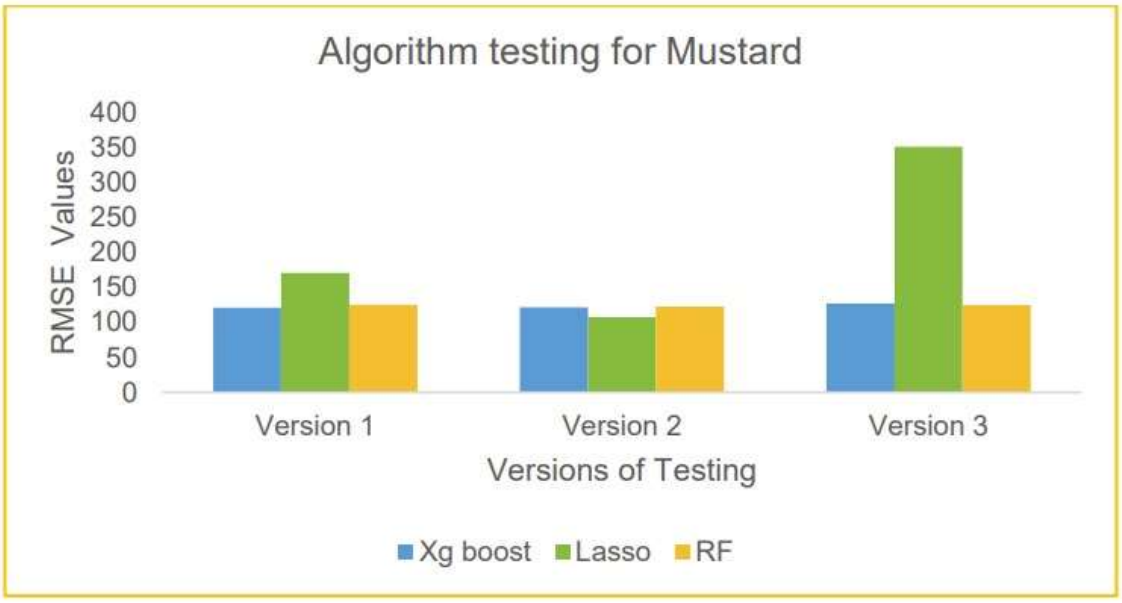
Analysis of Results



Testing Results for various algorithms Masur



Testing Results for various algorithms Chana



Testing results for various algorithms for Mustard

Notes:

1. Farm size refers to average land operated by the family for crop production in hectares. Minimum and maximum farm sizes for smallholder and otherfarms are also reported.

2. Value of crop production includes all crops produced on the farm. Value of food produced excludes cash crops.

3. Household income refers to gross annual earnings from all income generating activities (i.e. on-farm, agricultural wages, off-farm self-employment or wage earning, transfers and other).

4. Family labour days on-farm supplied over a day refer to the total number of person-days family members spend on- farm during one working day. Hired

labour days on-farm and family labour days in off-farm activities are computed similarly.

5. All animals are included in the calculation of livestock in Tropical Livestock Units. Depending on the country, these refer to horses, donkeys, oxen, cows,

sheep, goats, lambs, pigs, chicken, and ducks.

6. % of improved to total seeds is the ratio between quantity of improved seeds to the total quantity of seeds.

7. % of expenditure for inputs on value of production refers to the ratio between the total value of inputs to the value of agricultural production.

8. The average years of education at school is reported for the head of the household.

Thank you