Analysis of Long Short Term Memory (LSTM) Network for Rice Crop Yield Prediction

Manasvin A

Information Science Department, R.V. College of Engineering, Bengaluru, India
manasvina.is17[at]rvce.edu.in

Abstract: ***Agriculture is a crucial aspect of India, both in terms of economy and culture. It employs more than 50% of the Indian workforce. There is a need to provide a helpful guidance to the farmers with regards to what their yield is going to be. In the modern economic situation, farmers are under a lot of stress to better plan their schedules and harvests. Having a tool that can provide a estimation or prediction on the crop yield can help the farmers make predictions and also observe the trends. Then there is also the aspect that the climate is experiencing striking and impactful changes due to rising pollution, use of harmful energy sources, etc. Traditional yield predictors tend to only consider discrete current climate parameters, and not using climate data gathered up to that point in time. Past studies have shown that the adverse negative effects of climate change are going to be experienced by developing countries. Hence, there is a need to take into account the past climate trends when predicting crop yields. This paper explores the applicability of Long Short Term Memory (LSTM) networks, a form of Recurrent Neural Network (RNN), using time series crop yield data in prediction of rice yields in the states of Uttar Pradesh, Bihar and Karnataka. Possible improvements and extensions that can be applied to the model to improve its accuracy are also explored.***

Keywords: Rice yield, time series, Long Short-Term Memory.

1. Introduction

Rice plays a vital role in India. India is the second biggest rice producer with 22% of global rice production. The hot and humid climate found in parts of India is ideal for rice cultivation.

Over 50% of the India workforce is employed in agriculture. Most of these are small and rural based farmers, who place a lot of stake on each and every harvest. In recent times, stories of farmers committing suicide has become common. They are usually under a lot of economic burden and hence leaving anything up to chance is not acceptable. There needs to a smart farming system which can help the farmers make better decisions.

The Food and Agriculture Organization (FAO) of the United Nations argues that a smart crop production forecast system can lead to reductions in risks associated with local/national food systems. It also states that implementation of smart crop yield prediction systems can and should lead to better results in terms of the environment, socio-economic aspect (better income, better employment, etc.), and health & nutrition (reduced diseases, morbidity,etc.). **[9]**

Crop yield prediction is one of the most vital tools that can help the farmers. Crop yield is influenced by a myriad of factors**[3]** including climatic conditions and farm management practices, including but not limited to rainfall, environment, crop genotype and management practices.

Most of the past predictive tools and studies often predict yield using discrete climate parameters. Not a lot has been to look into the prospect of prediction using the data collected up to that point.

This can help in considering the impact of trends in climate change.

Studies have indicated that climate change will affect the agricultural sector hard, due to its direct dependence on weather.**[2] [6]**

There have been numerous studies which show that certain crops experience reduction in yields due to the impact of rainfall and or temperature trends.**[8]**

Therefore, there is a need to consider the effects of recent trends of weather on plants in addition to the effects of current weather conditions on the plants and their yields.

This paper tries to explore all this by utilizing the Long Short-Term Memory (LSTM). LSTM is a type of Recurrent Neural Network**[5]**. The main advantage provided by LSTM in this case is its efficiency in predicting time series and in capturing long term dependencies which is ideal as we want to consider the long term climate impact of crop yields.

The ability of the LSTM to allow for information to persist enables us to develop a model that has the ability to consider a certain part of the time series ordered data in order to estimate the yield in the current situation.

Since different regions of India are located in different climate zones**[7]**, the effect of certain trends in a climate parameter might not be the same in different parts of the country and another point is that farmers in areas more susceptible to harmful environment condition for certain crops are less likely to cultivate those crops and may even use alternate procedures for those crops**[1]** and using nation level data might overlook these points and overstate the yield losses for these situations. So, the LSTM model is applied on data from a particular region only.

Hence, this paper tries to use time-series data in localized regions (states) to take into account the effects of climate changes on yields.

For the sake of this study, the regions considered are the rice producing states of Uttar Pradesh, Bihar and Karnataka which have been chosen after considering both the relevance of their respective rice cultivation sector and also the availability of adequate data for that state.

2. Literature Survey

While earlier works such as **Kaylen and Koroma (1991) [11]** do note the relation between weather and crop yield, especially with regards to their non uniformity, they argued that the modelling of crop yields needs to be limited to temperature and rainfall.

Monisha and Robert (2004) [14] found that the machine learning models that produced the best results for the task of crop yield prediction were those that were built upon neural networks. They used rainfall data with the intention of predicting yields of soybean and corn.

Deschenes and Greenstone (2007) [12] through their findings concluded that yield lowers on increasing temperature but shows positive response to increasing rainfall. They still argued for a linear relationship between the yield and climate parameters.

Guiteras (2009) [15] studies the impact of temperature and rainfall on crop yield on the district level and tries to estimate the sensitivity that yields tend to have to climate changes. They focus on using gathered information and using it to get future predictions on the impacts of climate change and not get an assessment of its historical impact.

The non-linear effect on yield that most climatic conditions do show was shown by **Schlenker and Roberts (2009) [13]**. Using national level data, they were able to show a definite decline in yield when certain climate parameters passed a threshold value. Similar studies by **Tian Yu (2011)** also showed the nonlinear relation between yield and climate using Bayesian approach.

Droesch (2018) [2] assess the impact that climate change has on agriculture and looks at machine learning methods for crop yield prediction. The paper tries to project the impact of climate change on agriculture sector. They show how the negative climate change impacts also depend upon the climate zone of the region.

The use of time-series based networks for crop yield prediction was proposed and explored by **Jiang,et al(2018) [5]**. They argued for the importance of recognizing the impact of capturing long term dependencies. They also tried to argue for the usage of LSTM in fields other than natural language processing.

3. Data

1) Production (Yield) data

The dataset for crop yields was taken from ‘data.gov.in’ titled ‘District and Season wise Crop Production Statistics from 1997-2015’. This provides us with crop production along with area of land, state and district name, year of cultivation, season of cultivation and crop type.

2) Rainfall data

The dataset used for rainfall is taken from Kaggle, and is titled, ‘rainfall in India 1901-2015’. This provides us with monthly rainfall measurements in different regions of India, from 1901 to 2015.

3) Pre-processing

Any records with missing production entries were dropped. For missing entries in rainfall data, the method used to fill them was by using a weighted mean by considering the previous months recorded rainfall at that district and also the recorded rainfall during that month in the previous year in that district. This accounts for both historical trends in that district and also this year’s observed rainfall.

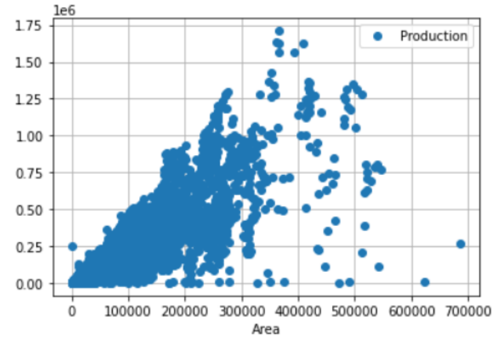
The scope of this paper is limited to rice in UP and Bihar and Karnataka.Hence, the first task is to filter the production dataset by considering only the relevant records.
Now, we map the rainfall data to the yield data records. For this we consider the state and district names,and year and season of the records in the yield dataset and use that to identify the relevant rainfall readings and take into account the rainfall readings of all the months that fall under that records mentioned season in that year at that district/state.

4) Data Visualization

The following plots were generated to get a better understanding of the feature importance and to see the statistical data visually,

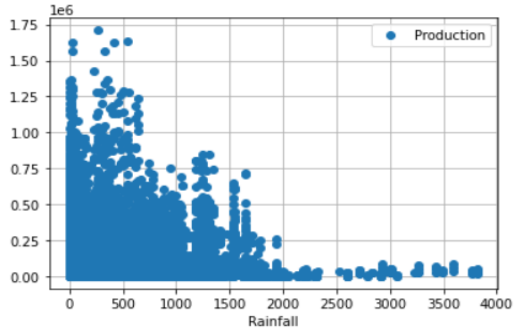
Area(m2) vs Production(kg)

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fdac6acea!
```



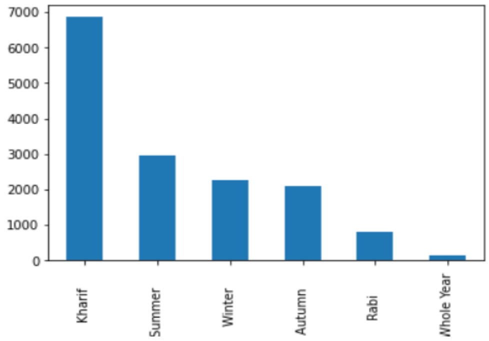
Rainfall(mm) vs Production(kg)

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fdac6564b:
```



Season Frequency

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fdac64f8128>
```

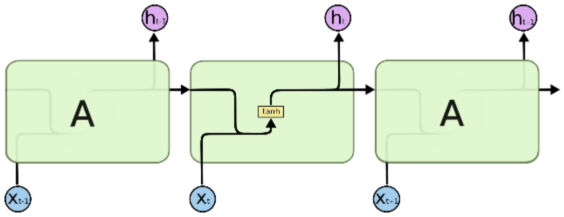


4. Technology/ Concept

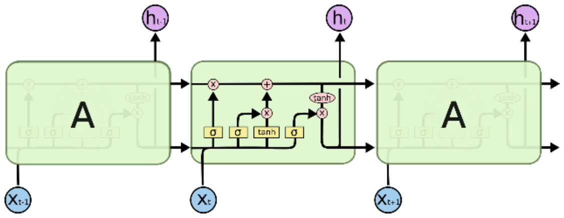
The paper makes use of Long Short-Term Memory network. LSTM is a special type of RNN, and hence they allow information to persist. They were introduced by Hochreiter & Schmidhuber (1997) are also capable of learning long-term dependencies.

They differ from standard RNN in that their repeating modules have much more complex and bigger layers.

The repeating module in a standard RNN contains a single layer.



The repeating module in an LSTM contains four interacting layers.



Information flow through each unit is controlled by structures called gates (input/output/forget).

- Input gate: regulates flow of new value into cell
- Output gate: controls extent to which value in cell is used to compute o/p activation of unit.
- Forget gate: regulates extent of value staying in cell.

The activation function is the logistic sigmoid function.

The cells have the ability to remember information over arbitrary time intervals.

Another advantage that LSTM have over standard RNN is that they are less vulnerable to the vanishing gradient problem which can occur during back-propagation of gradients (gradients can vanish to zero or explode to infinity).

5. Implementation

Firstly, the dataset is split into datasets for each state separately.

1) Feature Selection and Data Modification

We have already seen how area and rainfall play some role at least in the yield prediction. These attributes are normalized so that they fall to the same scale. The normalization method chosen is the z_score normalization, based on its ability to handle outliers and based on the distribution shown by the features.
z_score normalization is given by,
new_value = (value- mean of feature) / std.dev of feature

Then, the categorical features like district name and season are one hot encoded.

The dataset is sorted on time as LSTM requires a time series to work as required.

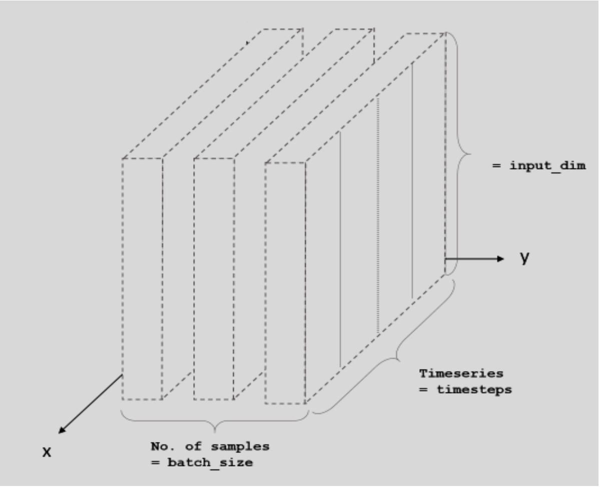
Now, we convert the dataset into a supervised learning dataset fit for our LSMT model. This is accomplished by shifting the records so that the information for n previous records in the time series is seen as the features and the current records output is the target.

2) Test/Train Split

The prepared dataset is split into train and test sets (75:25 ratio).

The train and test sets are further split into their respective features and targets.

At this point, we need to reshape the test and train feature sets into a 3Darray with the dimensions representing samples, time steps, features so as to work with the LSTM model.

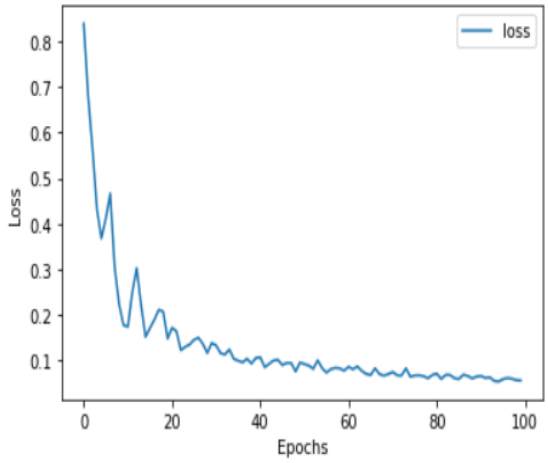


3) Fitting and Training the model

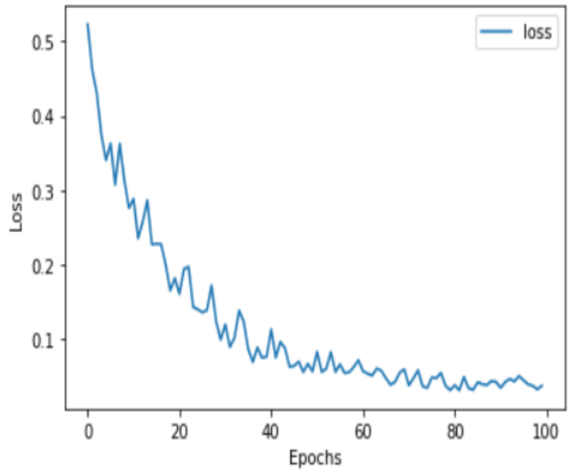
The model developed used the adam optimization technique, as this showed better results than when using other techniques such as stochastic gradient descent. The adam is a specially designed adaptive learning rate optimizer used for training deep neural networks.**[10]**

Then the model was fit with the training set and the following results were seen after conducting 100 epochs.

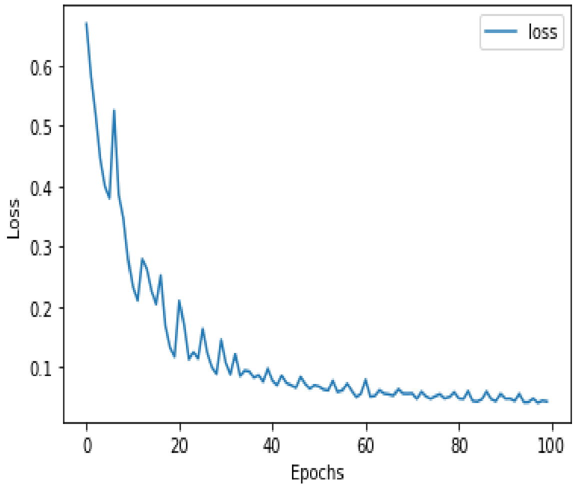
Uttar Pradesh



Bihar



Karnataka



6. Results & Conclusion and The Road Ahead

The following were the results observed when the fitted model was run on the UP, Bihar and Karnataka data.

The error measurement metrics chosen are the RMSE and MAPE (Mean Absolute Percentage Error).

STATE	RMSE	MAPE
UP	0.57	35%
Bihar	0.69	40%
Karnataka	0.56	40%

The RMSE scores indicate the numeric error value in the predicted yields, whereas the MAPE expresses the error as a percent of the actual value.

The state of UP provided better results mainly due to the presence of a larger dataset for it.

The degree of localization of the rainfall data being state level whereas yield data having district level data demanded a level of assumption, and our model had to act like all districts of a state experienced similar rainfall patterns.

The number of previous records in the time series to be considered was taken to be 50.
A major factor that impeded the model was the fact that the yield data time is not highly specific such as with the exact date of harvest. Thus, when sorting the data on time, there is a degree of inaccuracy in that records of similar times may get sorted inaccurately if a more accurate timestamp was present.

In this paper, there was a demonstration of a possible LSTM model for rice crop yield prediction. The results, while not quite as good as results obtained by certain other methods, perform somewhat well and show the potential of LSTM in the field of crop yield prediction.
The paper also tried to bring to light the importance of considering climate trends and past climate data in building a model to predict crop yield information.

LSTM are mostly used only in language processing; this paper hopes to show the potential that they hold in being used in certain other fields.

The Road Ahead

The first way to improve this model is by use of a better dataset. More localized data would help. This model used climate data based on the district of the farm, it would be more beneficial if climate conditions of the farms themselves were measured and available.

A larger dataset could also be useful in providing us with more records for each state.

The normalization done can also be done in a windowed fashion wherein the data within a certain time period are normalized with respect to each other, and not with respect to the entire dataset readings.

Further climatic and physical conditions such as soil/plant genotype/wind/temp can also be included. This will help the model gain additional information as these are all factors known to influence crop yield.

References

[1] Aravind Moorthy, et al., “The Impact ofClimate Change on Crop Yields in Indiafrom 1961 to 2010”, 2012

[2] Andrew Crane-Droesch 2018 Environ. Res. Lett. 13 114003

[3] Khaki S, Wang L and Archontoulis SV(2020) A CNN-RNN Framework forCrop YieldPrediction.Front. Plant Sci. 10:1750. doi: 10.3389/fpls.2019.01750

[4] Bejo, Siti, et al., “Application of Artificial Neural Network in Predicting Crop Yield: A Review”, Journal of Food Science and Engineering, vol.4, 2014

[5] Zehui Jiang, et al., “Predicting County Level Corn YieldsUsing Deep Long Short-Term Memory Models”, 2018

[6] Porter J R et al 2013 Food security and food production systemsClimateChange 2014: Impacts,Adaptation, and Vulnerability (Part A: Global and Sectoral Aspects) ed C B Field et al(NewYork: Cambridge University Press)

[7] Bansal, N.K., Minke, G., 1988. Climate Zones and Rural Housing in India. Kernforschungsanlage Jülich, Zentralbibliothek.

[8] D.B. Lobell, W. Schlenker, and J. Costa-Roberts. Climate trends and global crop production since 1980. Science, 333(6042):616, 2011.

[9] Yakob M.Seid, Crop forecasting: Its importance, current approaches, ongoing evolution and organizational aspects, FAO Statistics Division.

[10] Diederik P.Kingma, et al., “Adam: A method for stochastic optimization”, ICLR 2015.

[11] Michael S. Kaylen, Suffyanu S. Koroma, Trend, Weather Variables, and the Distribution of U.S. Corn Yields, *Applied Economic Perspectives and Policy*, Volume 13, Issue 2, July 1991

[12] Deschênes, Olivier, and Michael Greenstone. 2007. "The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather." *American Economic Review*, 97 (1): 354-385.

[13] Schlenker, Wolfram & Roberts, Michael. (2009). Nonlinear Temperature Effects Indicate Severe Damages to U.S. Crop Yields under Climate Change. Proceedings of the National Academy of Sciences of the United States of America. 106. 15594-8. 10.1073/pnas.0906865106.

[14] Monisha, K., Robert, L. & Charles, W. Artificial neural networks for corn and soybean yield prediction. Agricultural Systems 85: 1-18 (2005).

[15] R. Guiteras. The impact of climate change on indian agriculture. Manuscript, Department of Economics, University of Maryland, College Park, Maryland, 2009.

Dataset References

[16] Yield dataset https://data.gov.in/resources/district-wise-season-wise-crop-production-statistics-1997

[17] Rainfall Dataset https://www.kaggle.com/rajanand/rainfall-in-india