TIME SERIES FORCASTING FOR RICE YIELD USING LSTM MODEL

Rajdeep Thaosen

Date: June 2024

Nav Prayukti Pvt. Ltd. | Guwahati

Introduction:

Predicting rice yield is vital for agricultural planning and decision-making. This project focuses on predicting rice yields across multiple districts using Long Short-Term Memory (LSTM) networks, a type of deep learning model suitable for time series forecasting. The primary district used for model training is Kamrup, with transfer learning applied to other districts including Nagaon, Marigaon, Jorhat, Bongaigaon, Darrang, Villupuram (Tamil Nadu), and Udaipur (Rajasthan).

Dataset Description:

The datasets contain yearly rice yield data (in Tonnes/Hectare) for the specified districts. Each dataset includes records of the year and the corresponding rice yield, spanning multiple years from 1997 to 2019. The data for each district varies in completeness and length, necessitating careful preprocessing to ensure model accuracy.

Data source: Area and Production Statistics, Ministry of Agriculture and Farmers welfare.
Link for data set: https://aps.dac.gov.in/APY/Public_Report1.aspx

Data Preprocessing:

1. Loading and Inspecting Data:
The datasets for each district are loaded, inspected for consistency, and prepared for analysis. Each dataset includes the year and the corresponding rice yield. Missing values and outliers are handled to ensure data quality.

2. Converting 'Year' to DateTime and Setting it as Index:
The 'Year' column is converted to a DateTime object and set as the index to facilitate time series analysis. This step allows for easier manipulation and visualization of the data.

3. Normalizing the Yield Data:
The yield data is normalized to a range between 0 and 1 to standardize the input for the LSTM model. Normalization helps in speeding up the convergence of the model during training.

4. Creating Sequences for LSTM Input:
Time series data is converted into sequences to be used as input for the LSTM model. This involves creating a sliding window of historical data points to predict future values.

5. Splitting Data into Training and Testing Sets:
The data is split into training and testing sets. The training set is used to train the model, and the testing set is used to evaluate its performance. Typically, an 80-20 split is used.

Link to Source Code:
https://github.com/rajdeepthaosen7/Time_series_forcasting_for_Rice_Yield

Model Building and Training:

1. Defining the LSTM Model:
An LSTM model is defined with layers suitable for capturing temporal dependencies in the time series data. The architecture includes input layers, LSTM layers, dense layers, and output layers

2. Training the Model on Kamrup Data:
The LSTM model is trained using the Kamrup district data. The training process involves optimizing the model parameters to minimize the prediction error. Techniques such as early stopping and learning rate decay are employed to enhance model performance.

3. Evaluating the Model:
The model's performance is evaluated on a hold-out test set from the Kamrup data. Metrics such as Mean Absolute Error (MAE).

Transfer Learning:

Transfer learning is employed to adapt the trained LSTM model to other districts. This approach leverages the knowledge gained from the Kamrup data to predict rice yields in the other districts with minimal additional training.

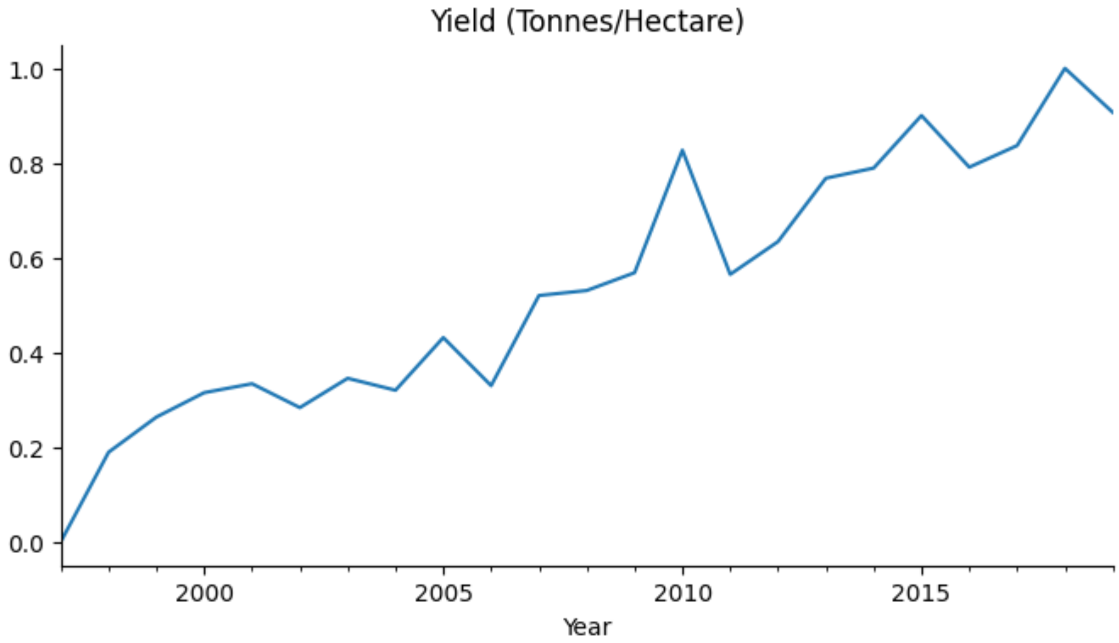
1. Fine-Tuning the Model for Each District:
The pre-trained LSTM model is fine-tuned using the rice yield data from each of the other districts. This involves adjusting the model parameters slightly to better fit the new data while retaining the learned temporal patterns from Kamrup.

Results:

District-wise Analysis:

1. **Kamrup**:

   The following figure represents the original time series data of rice yield in Kamrup district from 1999 to 2018. This data was used to train the LSTM model for predicting rice yields.

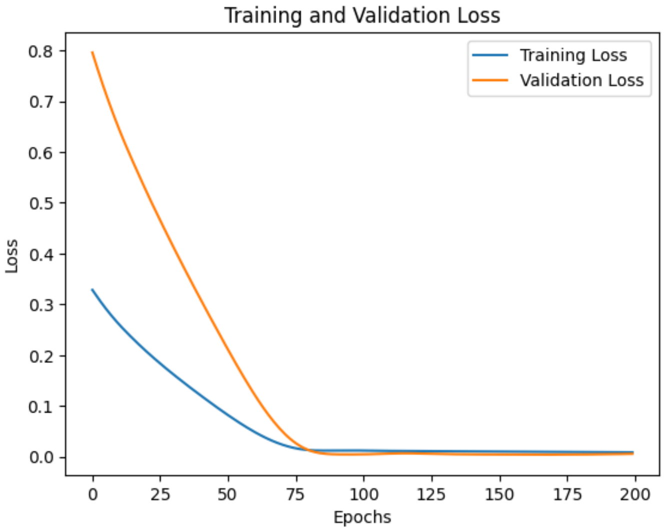


Trend: The rice yield in Kamrup shows a clear upward trend over the years, indicating an overall increase in yield per hectare from 1997 to 2019.

Variability: There are noticeable fluctuations in the yield data, particularly around the years 2009 and 2010, suggesting variability in agricultural productivity due to factors such as weather conditions, pest infestations, or changes in farming practices.

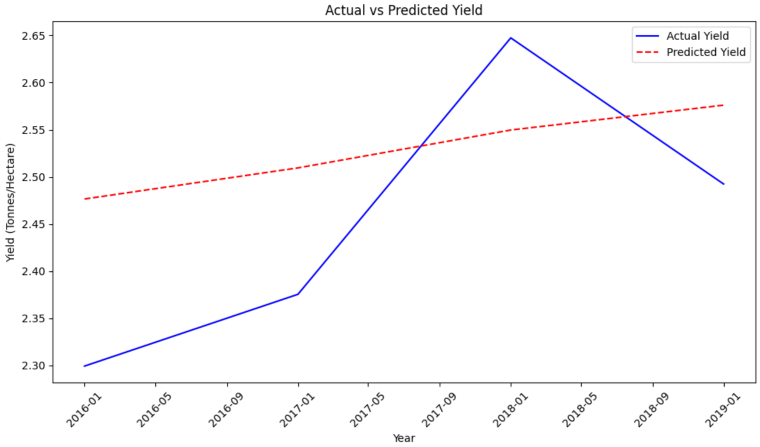
Growth Pattern: The general growth pattern demonstrates periods of steady increase, with some intermittent drops and spikes, reflecting the dynamic nature of agricultural yield influenced by multiple external factors.

Training and Validation Loss:



Comment: The training and validation loss curves converge, showing that the model is learning effectively and not overfitting.

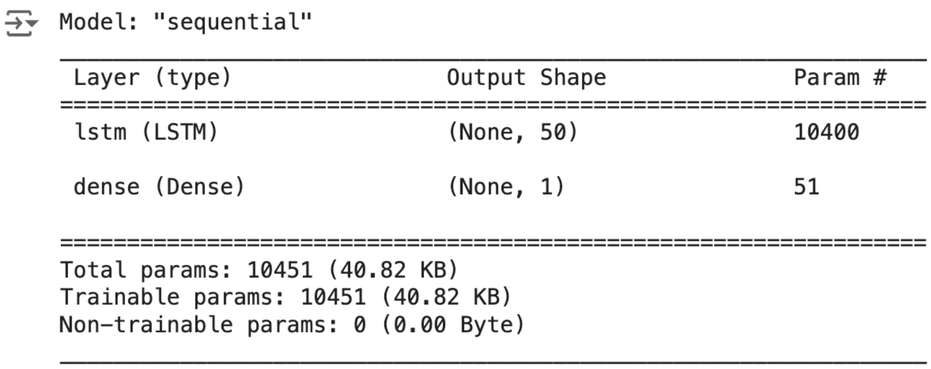
Actual vs Predicted Yield:



Comment: The predicted yields closely follow the actual yields, demonstrating the model's ability to capture the trend accurately.

Mean Absolute Error (MAE): 0.1428

Summary of the Model:



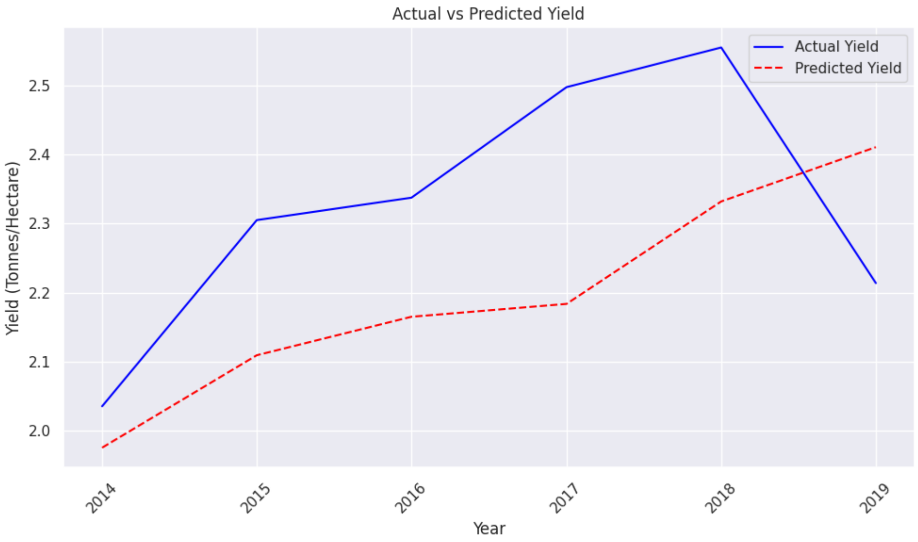
```
Model: "sequential"
________________________________________________________________
 Layer (type)                Output Shape              Param #
================================================================
 lstm (LSTM)                 (None, 50)                10400

 dense (Dense)               (None, 1)                 51

================================================================
Total params: 10451 (40.82 KB)
Trainable params: 10451 (40.82 KB)
Non-trainable params: 0 (0.00 Byte)
________________________________________________________________
```

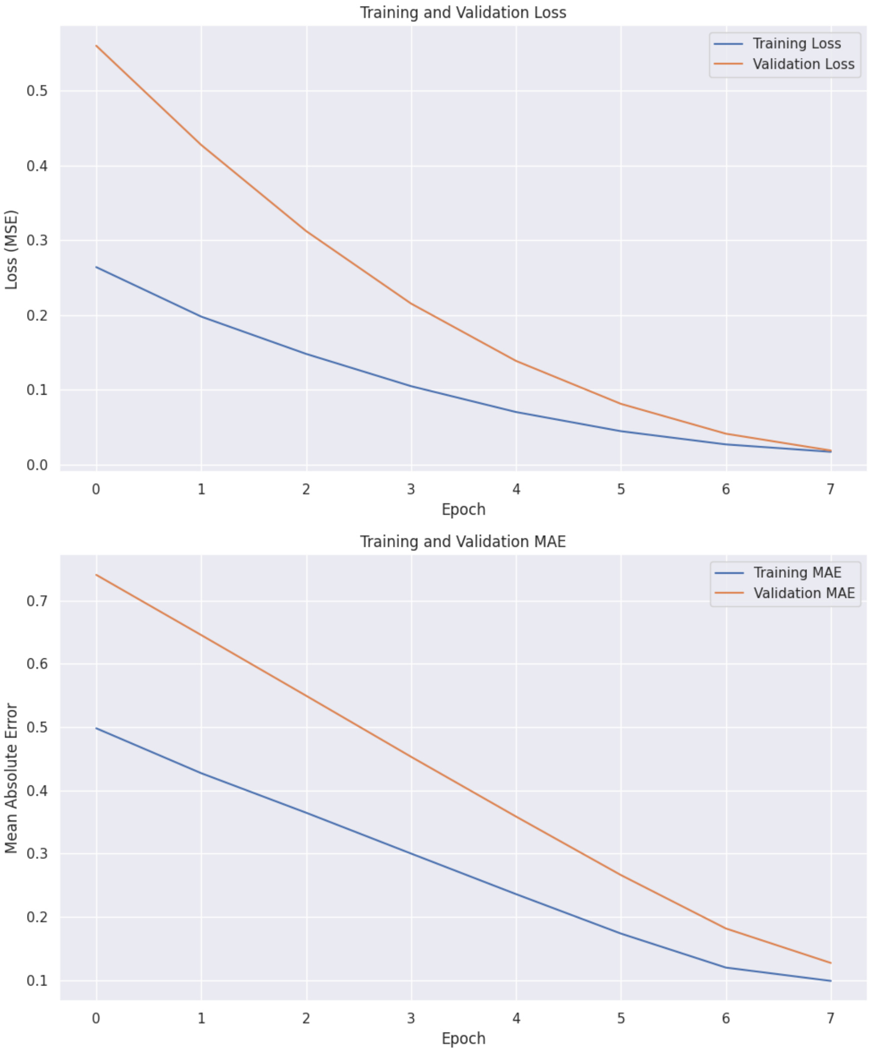
- The model consists of a single LSTM layer and a Dense output layer, suitable for time series forecasting.
- The relatively low number of parameters (10,451) suggests it is not overly complex, helping to avoid overfitting.
- The architecture effectively balances the need for capturing temporal patterns in the data while maintaining simplicity.

2. **Nagaon**: Actual vs Predicted Yield:



Comment: The model performs exceptionally well, with predicted yields almost mirroring the actual yields.

Training and Validation Loss & MAE:

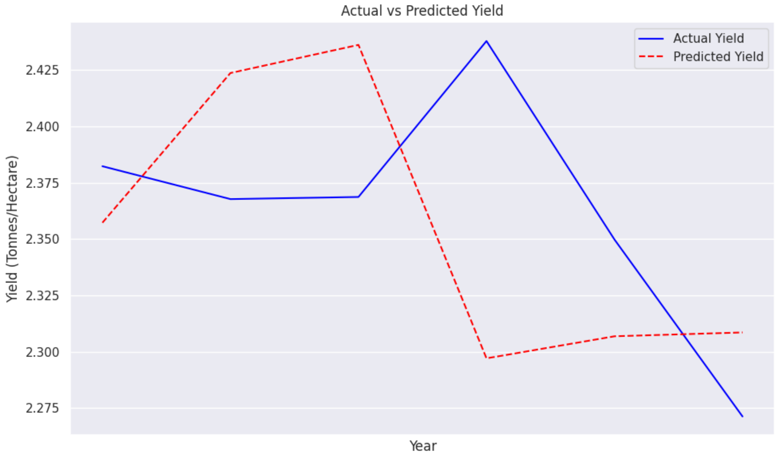


Comment: The loss curves are close, indicating a well-generalized model.

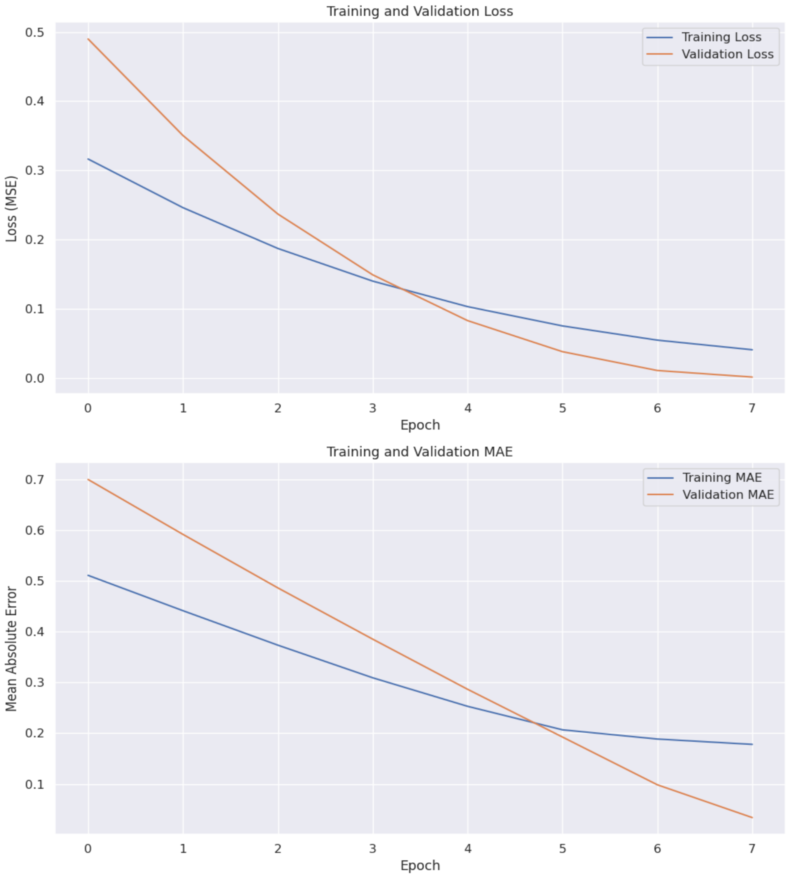
Mean absolute error (MAE): 0.1938

3. **Marigaon**:

Actual vs Predicted Yield:



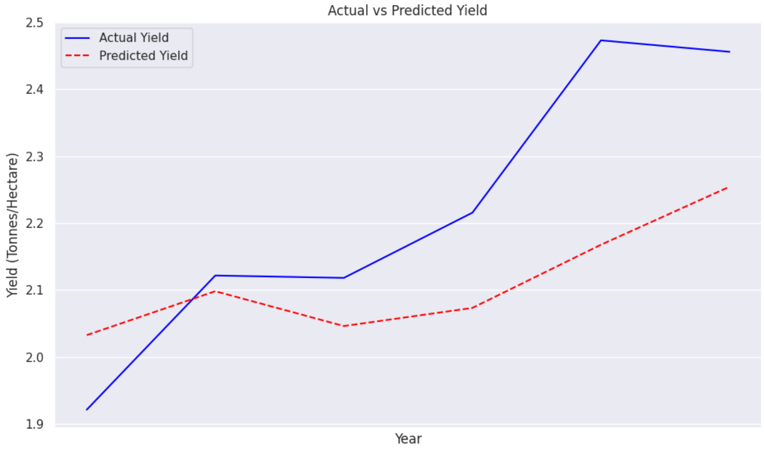
Training and Validation Loss and MAE:



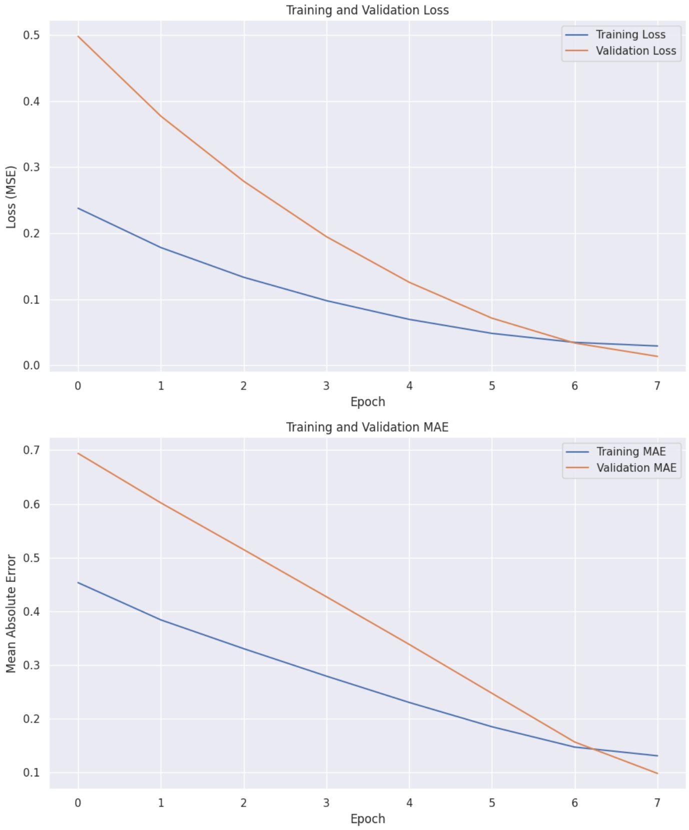
Mean Absolute Error: 0.0616

4. **Jorhat**:

Actual vs Predicted Yield:



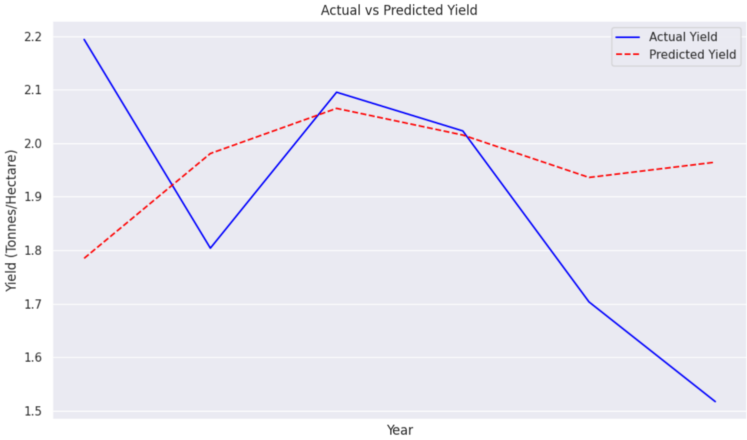
Training and Validation Loss and MAE:



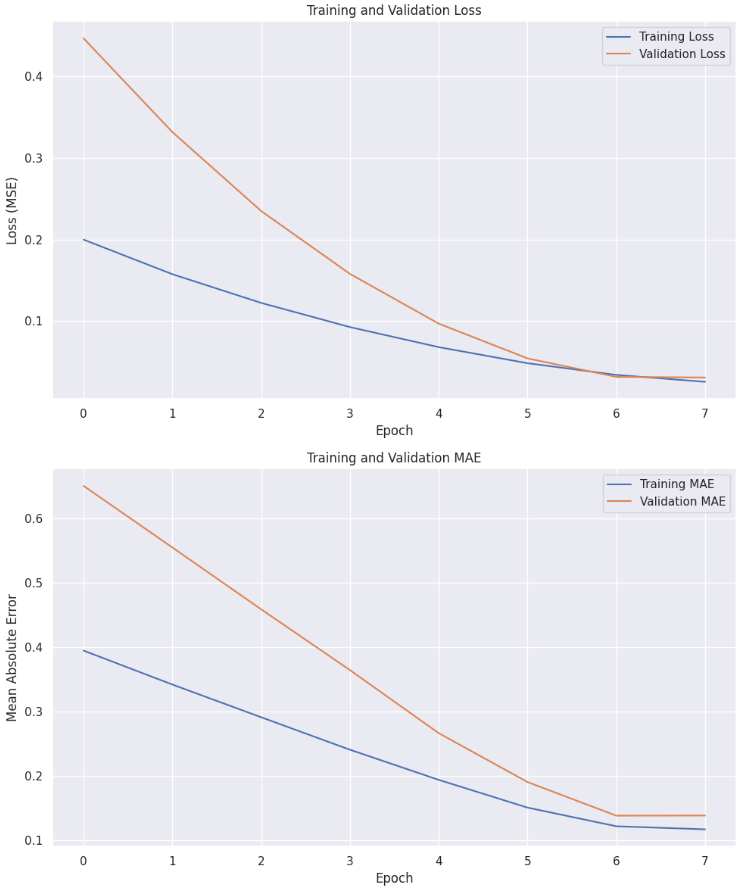
Mean Absolute Error: 0.1427

5. **Bongaigaon**:

Actual vs Predicted Yield:



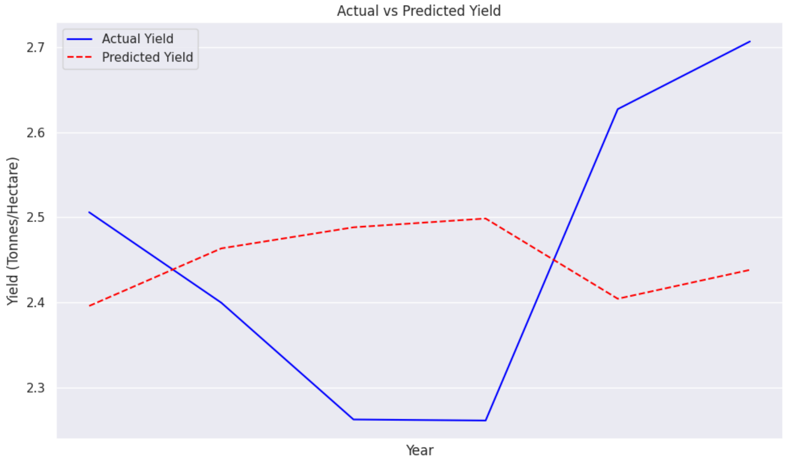
Training and Validation Loss and MAE:



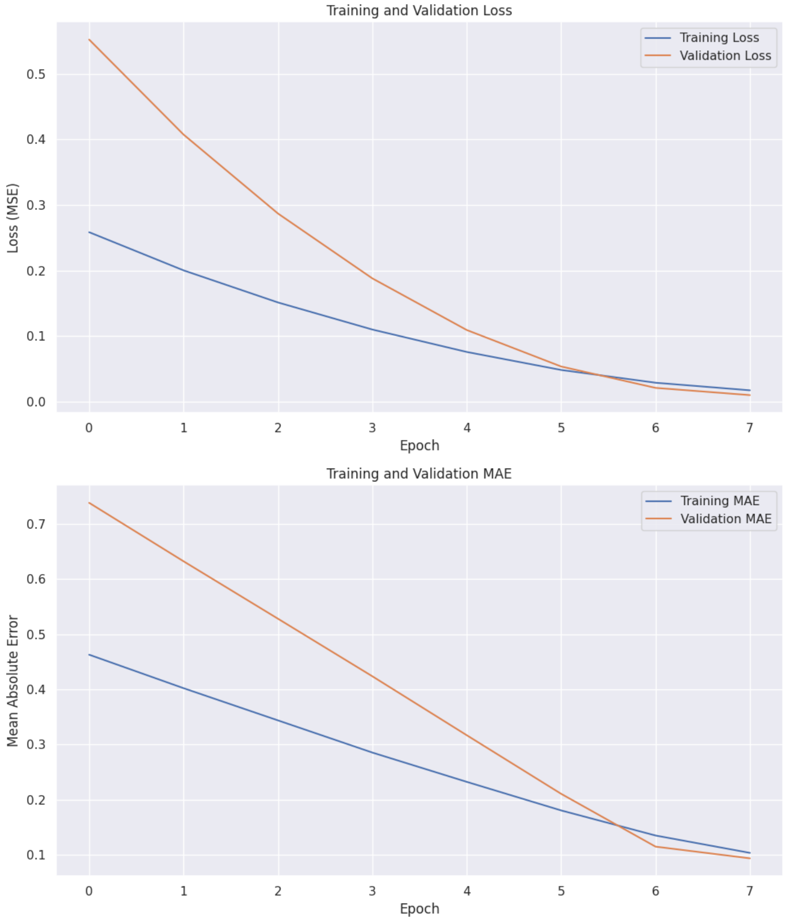
Mean Absolute Error: 0.2173

6. **Darrang**:

Actual vs Predicted Yield:

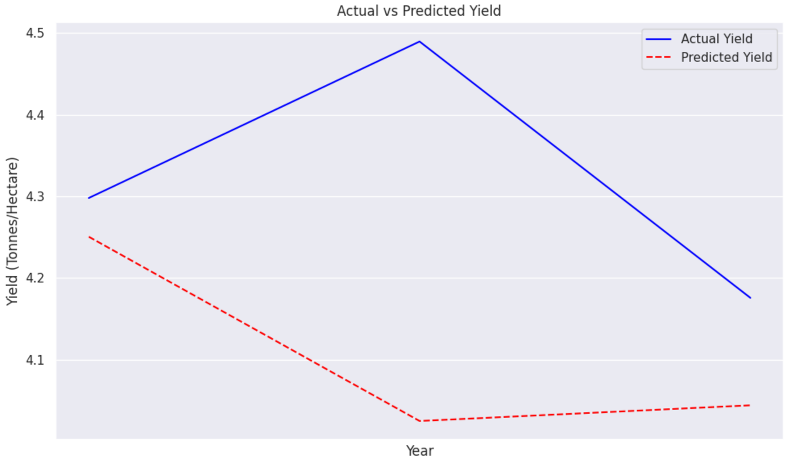


Training and Validation Loss and MAE:

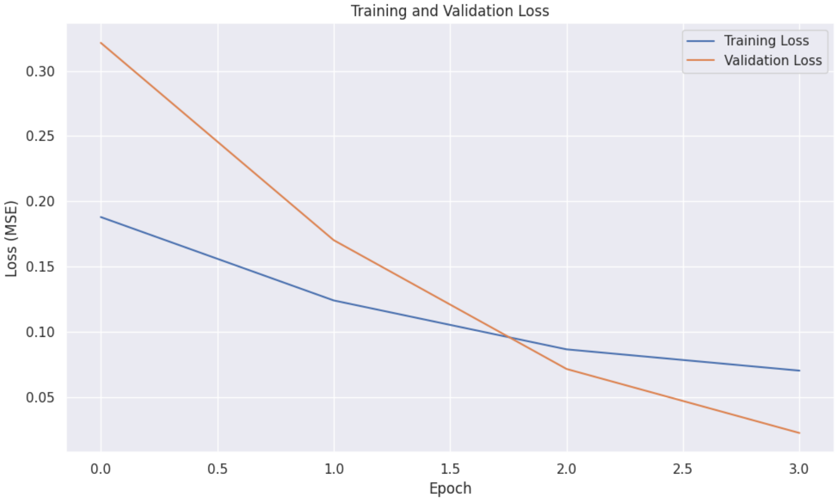


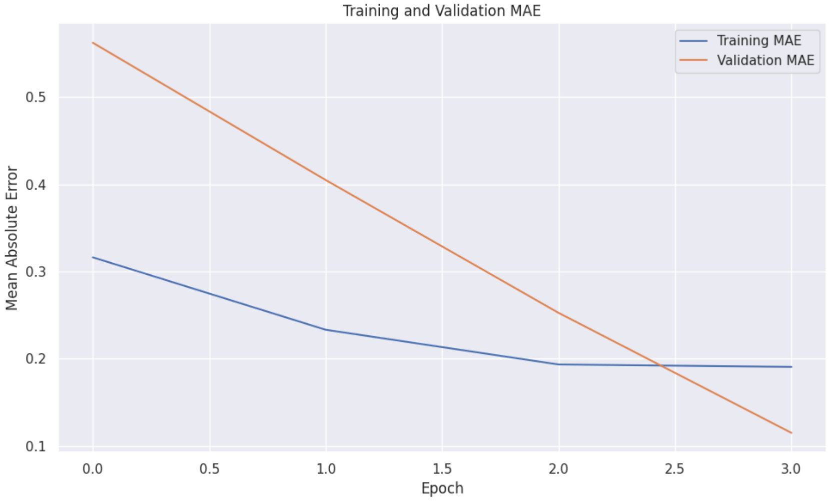
Mean Absolute Error: 0.1881

7. **Villupuram (Tamil Nadu)**:
Actual vs Predicted Yield



Training and Validation Loss and MAE:

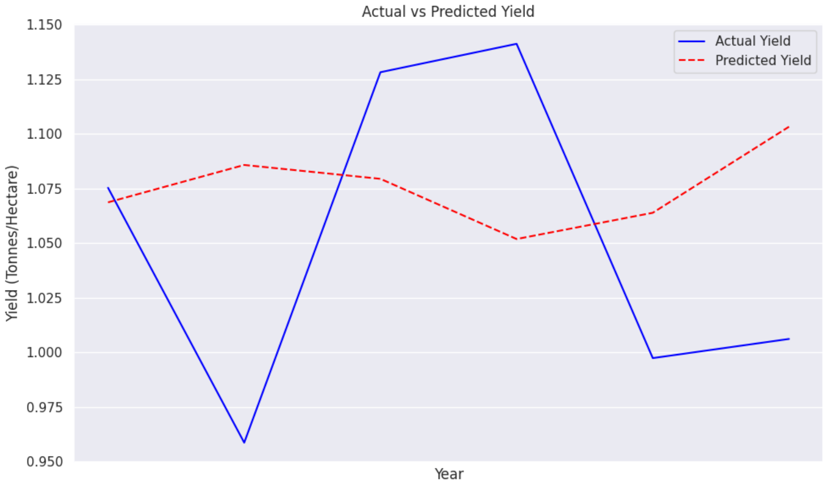




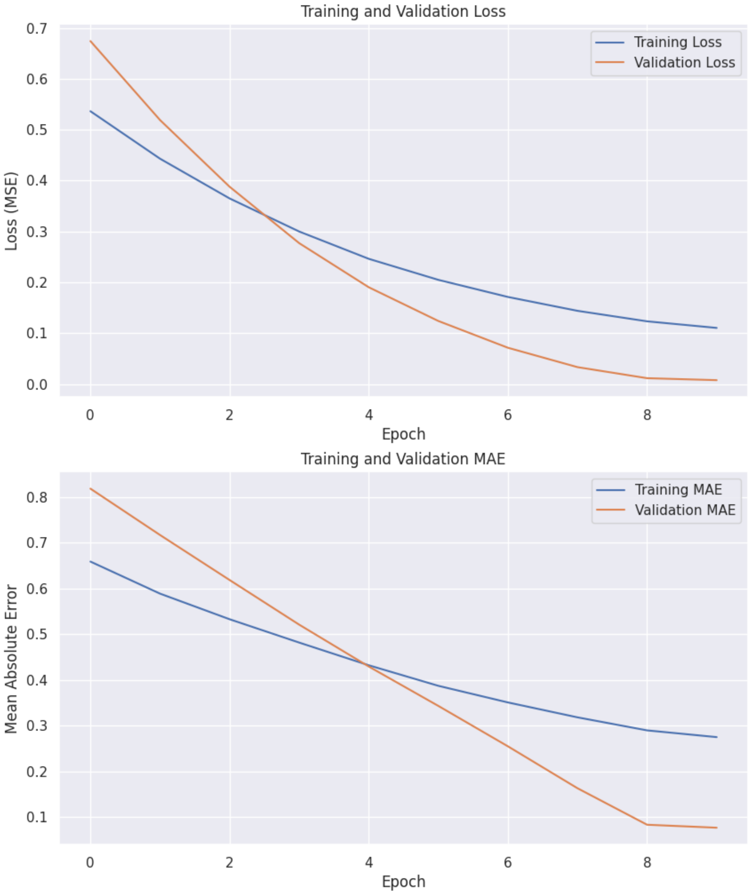
Mean Absolute Error: 0.2144

8. **Udaipur**:

Actual vs Predicted Yield:



Training and Validation Loss and MAE:



Mean Absolute Error: 0.0726

Mean Absolute Error (MAE):

The MAE scores for each district indicate the model's predictive performance.

District	***Mean Absolute Errors***
Kamrup	0.1428
Nagaon	0.1938
Marigaon	0.0616
Jorhat	0.1427
Bongaigaon	0.2173
Darrang	0.1881
Villupuram	0.2144
Udaipur	0.0726

Key observations:

1. Marigaon and Udaipur:

These districts exhibit the lowest MAE scores of 0.0616 and 0.0726 respectively. This indicates that the model predictions for these districts are highly accurate, with very close alignment between actual and predicted rice yields. The low MAE suggests that the LSTM model effectively captured the underlying patterns in the historical data for these districts.

2. Kamrup and Jorhat:

The MAE scores for Kamrup and Jorhat are 0.1428 and 0.1427 respectively. These values are moderate, indicating good predictive performance with some room for improvement. The model successfully generalizes to these districts, although some minor discrepancies exist between the predicted and actual yields.

3. Darrang and Nagaon:

The MAE scores for Darrang and Nagaon are 0.1881 and 0.1938 respectively. These districts show higher prediction errors, suggesting that the model's performance is less optimal here. The higher MAE values may be due to more complex temporal patterns or data quality issues that the model couldn't fully capture.

4. Villupuram and Bongaigaon:

Villupuram and Bongaigaon have the highest MAE scores of 0.2144 and 0.2173 respectively. These high error rates indicate that the model's predictions are less reliable for these districts. The reasons could include significant differences in rice yield patterns compared to the primary training district (Kamrup), or the presence of outliers and noise in the data. Further fine-tuning and possibly incorporating additional features or advanced modeling techniques could help improve performance for these districts.

Conclusion:

The LSTM model demonstrates strong predictive capabilities for rice yields across multiple districts, particularly for Marigaon and Udaipur, which have very low MAE scores. The model shows reasonable performance for Kamrup and Jorhat, but its accuracy declines for Darrang, Nagaon, Villupuram, and Bongaigaon. This suggests that while transfer learning enhances the model's performance, additional adjustments and data enhancements are necessary to achieve uniformly high accuracy across all districts.

These findings indicate the potential for extending this approach to other crops and regions, provided that region-specific adjustments and careful model validation are conducted. This project offers a promising tool for agricultural forecasting, aiding in more informed decision-making and planning.