MTP-II End-Term Evaluation Report

**Time Series Modelling for Tomato Price Prediction: A Comprehensive Analysis and Forecasting Approach**

Thesis/Project Report

Submitted By

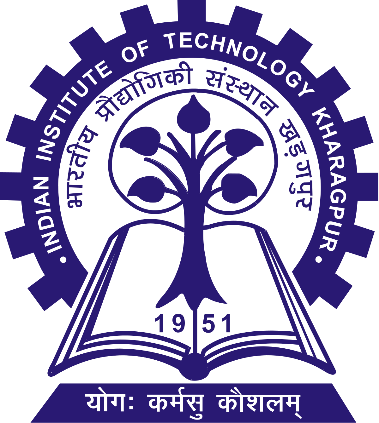
**Param Chauhan - 19IM30007**

Supervised By

**Dr. Sayak Roychowdhury**

**Department of Industrial and Systems Engineering**

**IIT Kharagpur**



Department of Industrial and Systems Engineering

IIT Kharagpur

Spring Semester 2023 – 2024

**Introduction**

The agricultural sector serves as a fundamental pillar of the global economy, where the prices of agricultural goods have significant effect over food security, economic equilibrium, and the livelihoods of millions worldwide. Given the inherent price volatility in agricultural markets in India, especially the tomato market, the need for forecasting becomes paramount to enable informed decision-making among farmers, traders, and policymakers in India. This project uses a comprehensive methodology, implementing various time series models to analyze historical data on tomato prices.

This research primarily focuses on using different time series and machine learning models to analyze and forecast the tomato prices of Chennai City. An extensive dataset of different features, spanning across 7 years (i.e. from 2013 to 2019), consisting of modal prices of tomato, consumer price index (CPI), diesel prices, rainfall data and temperature data were collected for the same.

The research methodology involves comprehensive data analysis conducted on the collected features, along with Seasonal-Trend decomposition utilizing LOESS, which provides insights into the underlying patterns and trends in the data. The research firstly utilizes the Autoregressive Integrated Moving Average (ARIMA) model to identify trends and seasonality in the tomato prices. Then, to identify the effects of the other exogenous variables and to combat the limitations of ARIMA, the Long Short-Term Memory (LSTM) machine-learning model is used.

Lastly, more advanced analytical technique like the Structural Equation Modeling (SEM) is used to further know the effect of different factors on the tomato prices and find out the extent to which, a factor affects the tomato prices.

This research has great implications on the agricultural sector in India, especially concerning tomatoes, which stands as a pivotal element of the nation's economy. For farmers, it enables well-informed decisions regarding planting timelines and marketing approaches, thereby optimizing resource management and increasing profitability. Similarly, traders and distributors stand to gain from optimized supply chain processes, leading to minimized wastage and heightened operational efficiency.

**Literature Review**

Smith et al (2019) utilized Structural Equation Modeling (SEM) to explore the interconnections among self-esteem, social support, and depression in adolescents. Their findings revealed that social support played a partial mediating role in the association between self-esteem and depression.

Muthén and Asparouhov (2018) proposed Bayesian Structural Equation Modeling (SEM) as a viable alternative to conventional maximum likelihood estimation methods. This approach presents several benefits, including increased resilience to non-normal data distributions and enhanced flexibility in model specification.

Lama, A., et al (2015) investigated the volatility of prices for edible oil and cotton, employing the Autoregressive Integrated Moving Average (ARIMA) model, GARCH model, and EGARCH model. Their analysis highlights that the price data of numerous agricultural commodities inherently exhibit noise and volatility. This volatility stems from the swift reactions of agricultural commodity prices to both real and perceived alterations in supply and demand conditions. Additionally, fluctuations induced by weather conditions in farm production further exacerbate the inherent volatility in these prices.

Ramirez, O. et al (2003) examined the price volatility of soybeans and wheat in the futures markets of the United States using the GARCH model.

Sekhar, C., et al (2017) applied GARCH and EGARCH to examine volatility in agricultural prices in India. According to it, in lower-income countries, food inflation is not only more volatile but is also, on average, higher than nonfood inflation. The study shows that food inflation is more persistent than nonfood inflation.

Onour, I., et al. (2011), applied the ARIMA-GARCH model for the analysis and prediction of price volatility in selected agricultural products on a global scale.

Bhardwaj, S.P., et al (2014) utilized the ARIMA-GARCH model for the examination and prediction of price volatility in the Delhi agricultural market for grams. Their findings revealed that the ARIMA model struggled to encapsulate the inherent volatility within the dataset, contrasting with the GARCH model, which effectively captured the observed volatility.

Kiran M. Sabu et al (2019) utilized methods like SARIMA, HoltWinter’s Seasonal method, and LSTM neural network, and their performance was evaluated based on the RMSE value on the arecanut dataset with prices from 2007 to 2017.

Zhiyuan Chen et al (2021) experimented with 5 different methods, namely ARIMA, SVR, Prophet, XGBoost, and LSTM on large historical datasets and found the LSTM method to be the best with an average MSE of 0.304.

**Data Collection and Pre-processing**

In order to perform an analysis on the trend of tomato prices in Chennai, a variety of data was collected from different sources to account for different factors contributing in the analysis, which are described below:

1. **Tomato Dataset**: Sourced from the website of National Horticultural Board, which provides data on daily minimum, maximum, and modal prices of tomatoes along with arrival quantity in quintals.
2. **Diesel Price**: Diesel prices can affect the cost of transportation and logistics, that is why it is taken into account. Sourced from mypetrolprice.com
3. **CPI**: Data of CPI is needed due to its critical role as a key economic indicator in providing valuable insights into inflation trends of different commodities like fruits, vegetables, fuels, etc. It is sourced from the website of the Central Statistical Office.
4. **Rainfall**: The data of rainfall is crucial factor in forecasting tomato prices as it affects the production of tomato. The rainfall data of the two districts, Chittoor and Kolar which mainly supply tomatoes to Chennai, is sourced from website of Indian Meteorogical Department.
5. **Temperature**: The data of temperature is also another crucial factor in forecasting tomato prices. The temperature data of Bengaluru is collected, as it is the closest to both the aforementioned districts and it is safe to assume that the average temperature of those district is same as average temperature of Bengaluru. The data is sourced from opencity.in

**Exploratory Data Analysis and Data Pre-Processing**

The tomato price data, along with all other data, spanning from 2013 to 2019, gives valuable insights into market dynamics of tomato prices. However, the collected data is raw and must be cleaned and polished for analysis. Steps on data preprocessing are described below.

**Missing Values and Outlier Detection**

The Inter-Quartile Range (IQR) approach was used to identify outliers or extreme values that deviated from other observations, especially in the "Arrival Quantity" and "Retail Price" columns. The missing values were filled in using linear interpolation. These outliers were carefully eliminated to preserve the integrity of the data and prevent skewing the results.

**Adjusting for Inflation**

Considering inflation is crucial when analyzing data over time, especially with price information covering a long period. This step helps us understand how the real value changes over time by removing the effects of general price increases due to inflation. Inflation adjustment ensures a more accurate and meaningful comparison of prices across different years.

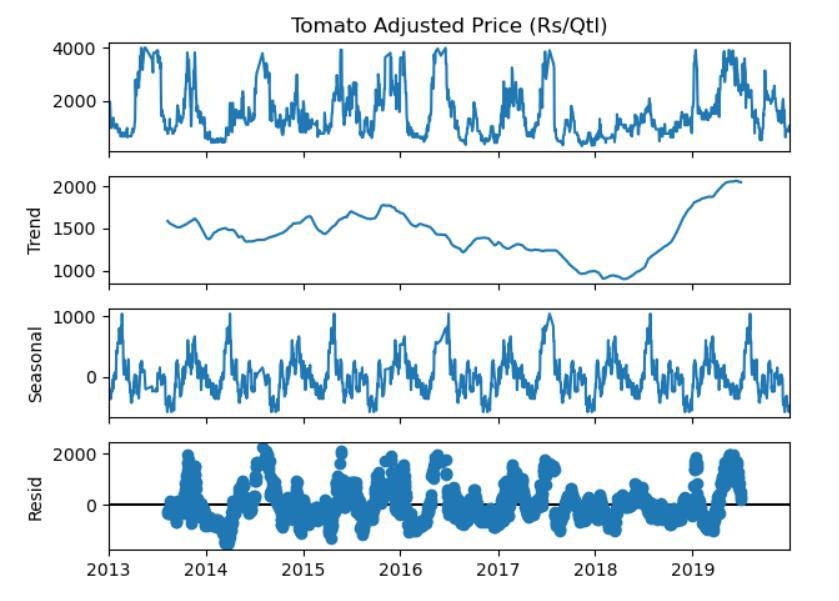
To adjust for inflation, select a base year, the Consumer Price Index (CPI) of that year is considered as reference CPI.

To adjust for inflation, the prices are multiplied by CPI of their respective year and divided by the reference CPI.

**Adjusted Price = Price x CPI of Base Year / CPI of Respective Year**

All parameters like tomato price, potato price and diesel price are adjusted to inflation.

**Seasonal Trend Decomposition**

The Seasonal trend Decomposition using LOESS (STL) is done on the time series data of modal price of tomato is done and shown here:

**Original**: This is the raw data we started with, representing the observed values of tomato prices over time. It includes all the patterns, trends, and fluctuations present in the original time series.

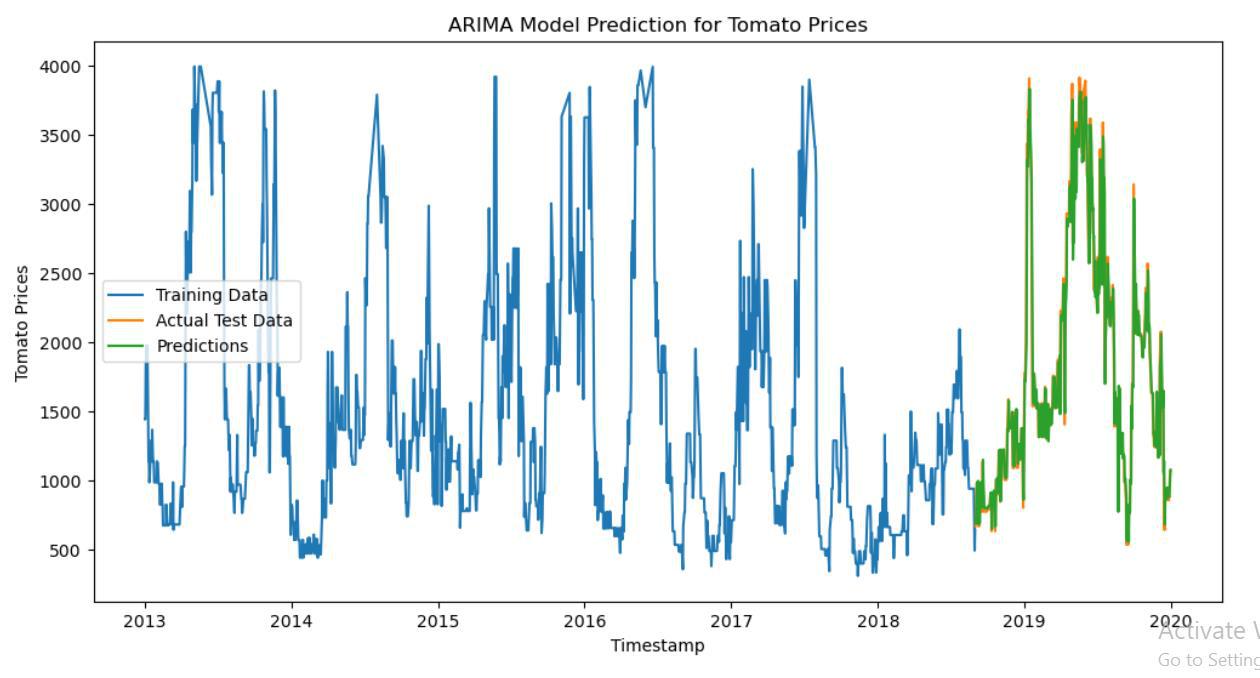
**Trend:** The trend component signifies the enduring fundamental pattern or motion within the tomato prices dataset. It aids in encapsulating the general trajectory in which the data is progressing.

**Seasonal**: The seasonal component embodies the recurring patterns or seasonality within the time series. Seasonal patterns tend to manifest at consistent intervals, such as daily, monthly, or yearly cycles.

**Residual**: The residual component characterizes the stochastic or unpredictable fluctuations, often referred to as noise, in the time series, which cannot be ascribed to the seasonal or trend components. Essentially, it captures what is left after eliminating the recognized patterns

**The ARIMA Model**

Developed in the 1970s by Box and Jenkins, the Autoregressive Integrated Moving Average (ARIMA) model serves as a mathematical technique to understand fluctuations in time series data. Over time, it has emerged as a powerful tool for analyzing and forecasting time series. With applications spanning various domains including finance, economics, and environmental science, the ARIMA model provides a flexible framework for capturing both linear trends and temporal relationships within a dataset. In the subsequent section, we examine the structure of the ARIMA model, investigating its constituent elements and practical uses.

However, before constructing an ARIMA model, it is imperative to ensure the stationarity of the data. Stationarity implies that the mean and variance of the same time series dataset remain constant, and the covariance between the ith term and the (i+m)th term is not a function of time.

After performing ADF test for stationarity and applying ARIMA(4,0,0) we find the RMSE score to be **208.197**

The ARIMA model is a powerful and widely used time series forecasting method, but it does have certain limitations:

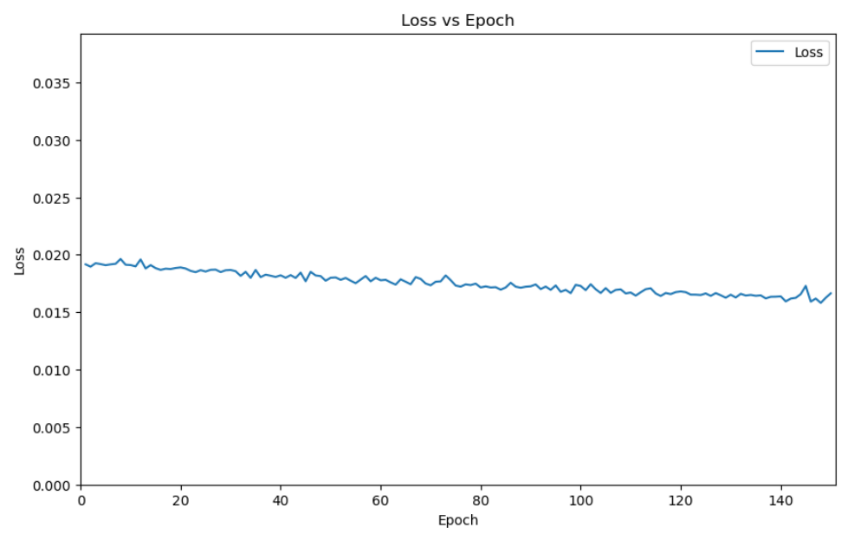
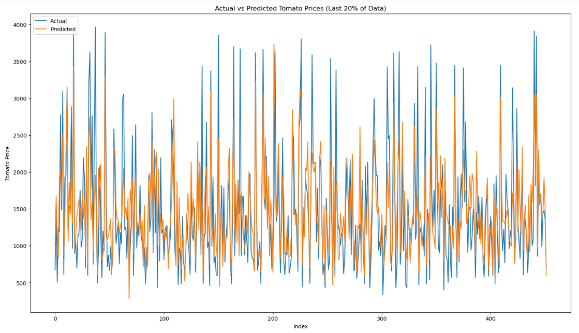
1. **Linear Assumption**: ARIMA assumes that the relationships between variables are linear. If the underlying patterns in the time series data are highly nonlinear, ARIMA may not capture them effectively.
2. **Stationarity Requirement**: ARIMA necessitates that the input time series data be stationary, implying that its statistical properties, such as mean and variance, remain constant over time. Although differencing can be employed to transform non-stationary data into a stationary form, this procedure may not always be straightforward or appropriate for every dataset.
3. **Sensitivity to Outliers**: ARIMA models can be sensitive to outliers in the data, and outliers can have a significant impact on the parameter estimation and forecasting performance.
4. **Limited Seasonal and Trend Patterns**: ARIMA is designed to capture autoregressive and moving average patterns, but it may struggle with complex seasonal and trend patterns, especially if they are not strictly periodic.
5. **Effect of Exogenous Factors**: ARIMA models do not inherently consider external factors or exogenous variables that may influence tomato prices
6. **Manual Selection of Parameters**: Selecting appropriate values for the order of autoregressive (p), integrated (d), and moving average (q) components in ARIMA models requires manual tuning and, in some cases, domain expertise. This process can be subjective and time-consuming.

**The LSTM Model**

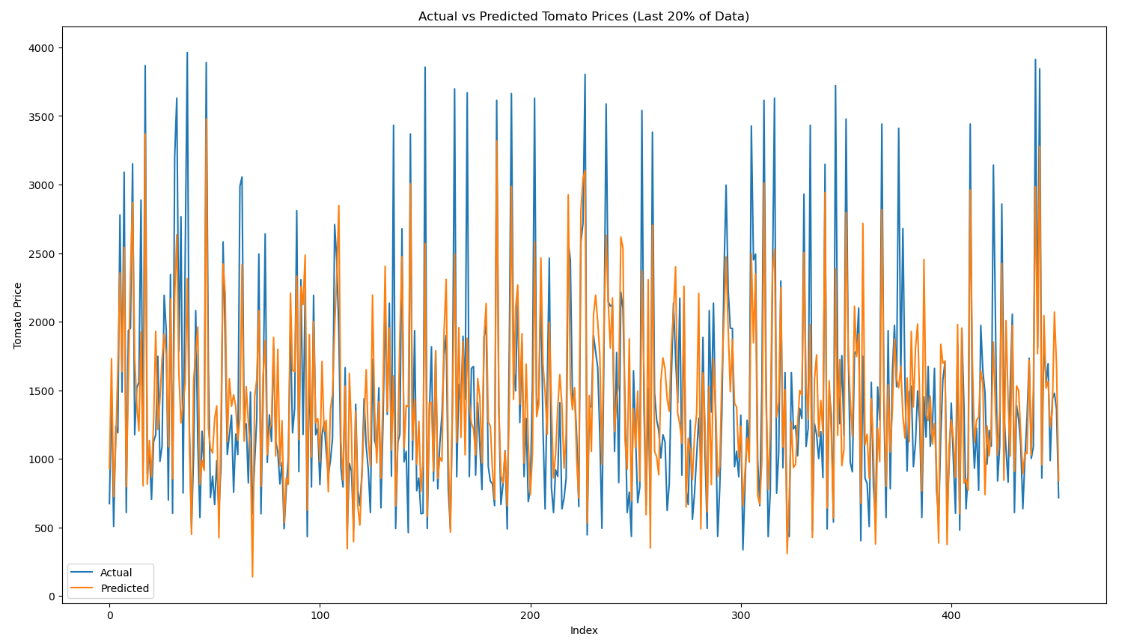
Using LSTM (Long Short-Term Memory) networks instead of ARIMA (AutoRegressive Integrated Moving Average) models for tomato price forecasting is often motivated by the inherent characteristics of time series and the strengths of LSTM in handling such data. Here's an explanation of why LSTM might be preferred over ARIMA in this context:

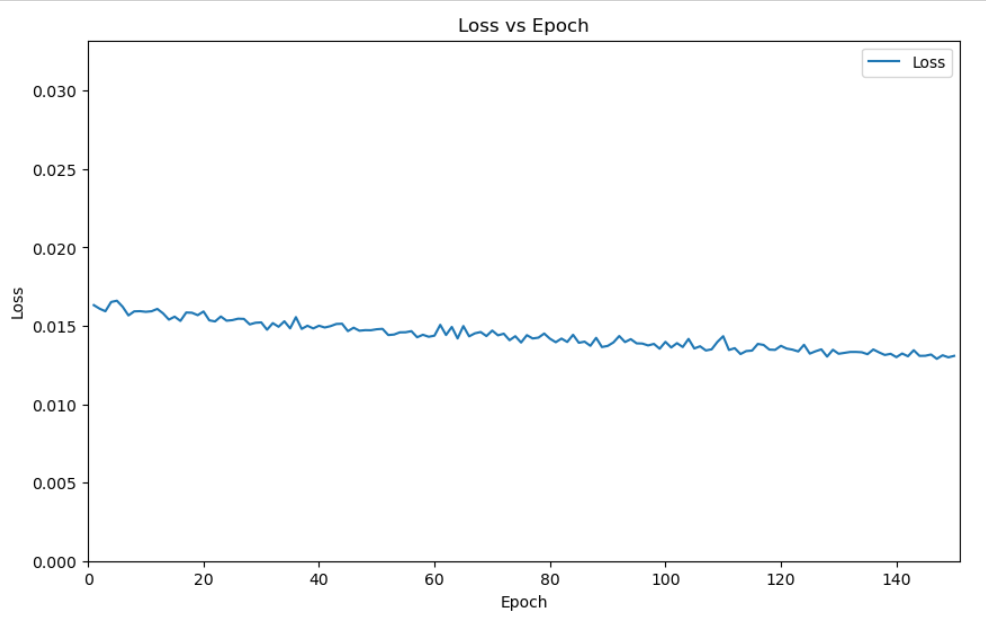
1. **Non-linear Relationship Capture**: LSTM models effectively capture complex, non-linear relationships in data, making them suitable for forecasting tomato prices influenced by various factors.
2. **Long-Term Dependency Handling**: Unlike ARIMA, LSTM models handle longterm dependencies and non-stationarity, which are common in economic and agricultural data like tomato prices.
3. **Flexible Input Handling**: LSTM models accommodate multiple input variables and their temporal relationships, allowing for comprehensive modeling of factors influencing tomato prices.
4. **Irregular Time Interval Handling**: LSTM models gracefully handle irregularly spaced time series data, which is common in agricultural markets where events like harvests occur at irregular intervals.
5. **Handling Non-Stationary Data**: These networks can handle non-stationary data directly. tomato market data, which are often non-stationary (mean and variance change over time), can be challenging for traditional time series models but are less problematic for LSTMs.

For testing the LSTM Model, the daily temperature and rainfall statistics are not quite useful in forecasting, as tomatoes take many days to grow. So, another solution to this problem is to take mean of the past 30, 60, 90, 120 and 150 observations consecutively for each data point, and then use them one by one as explanatory variables. We plot actual data vs predicted data for each case.

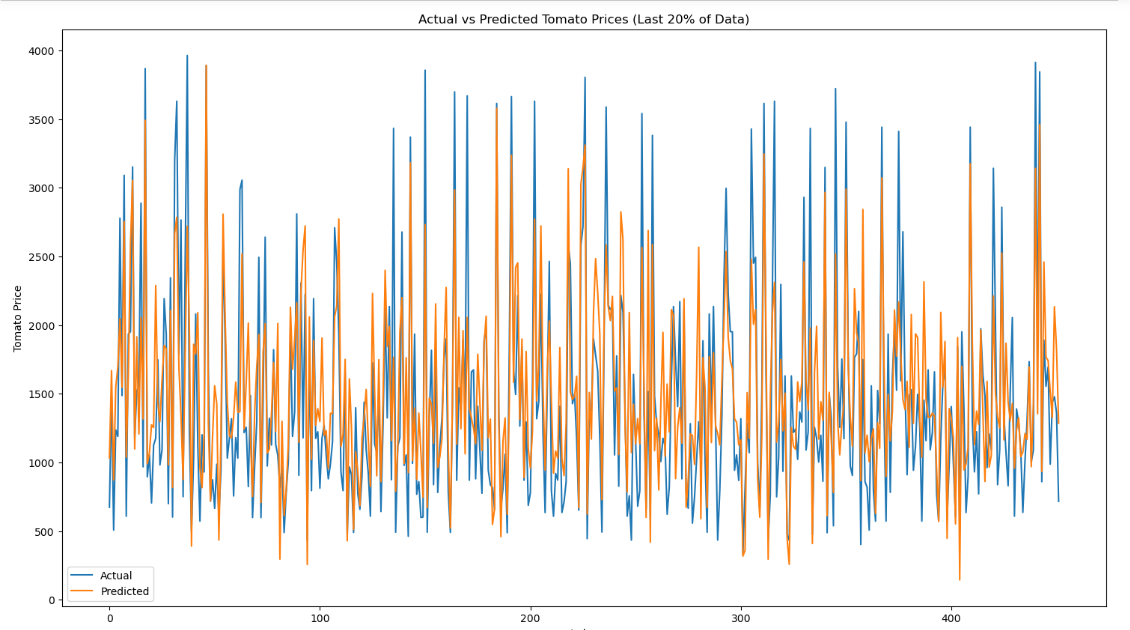
****

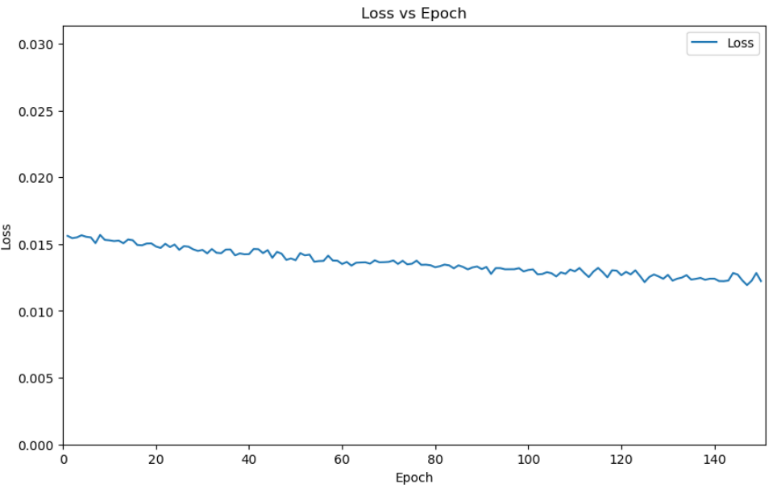
The RMSE Value for 30 Days Average is found to be 501.176



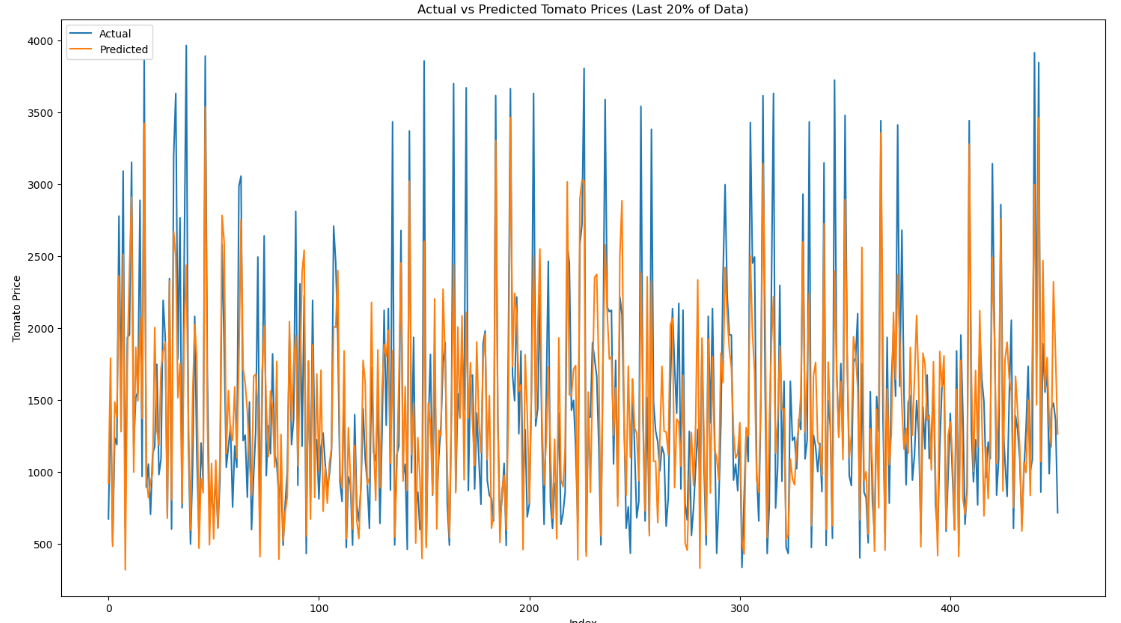
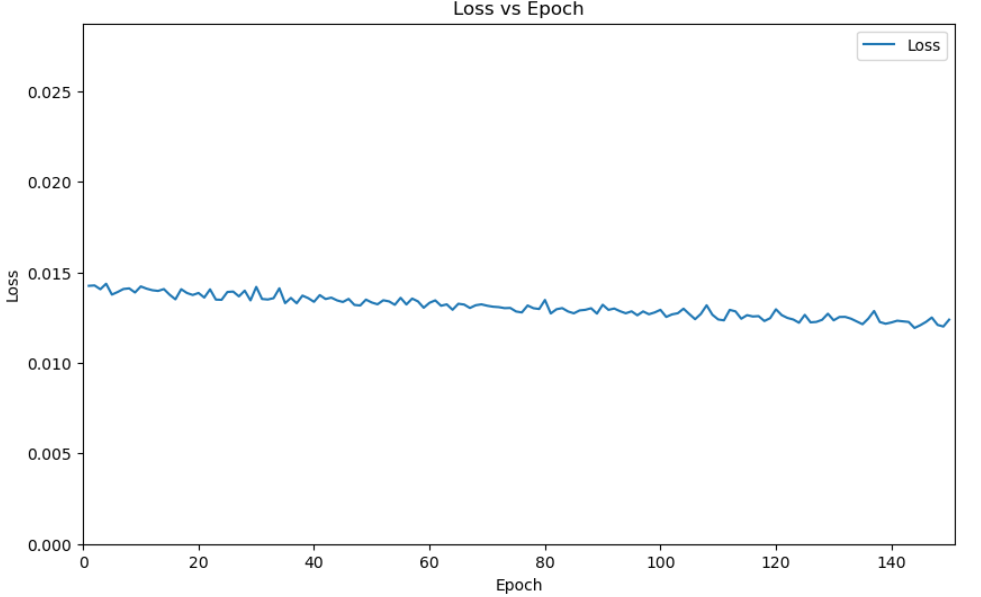


The RMSE Value for 60 Days Average is found to be 462.469

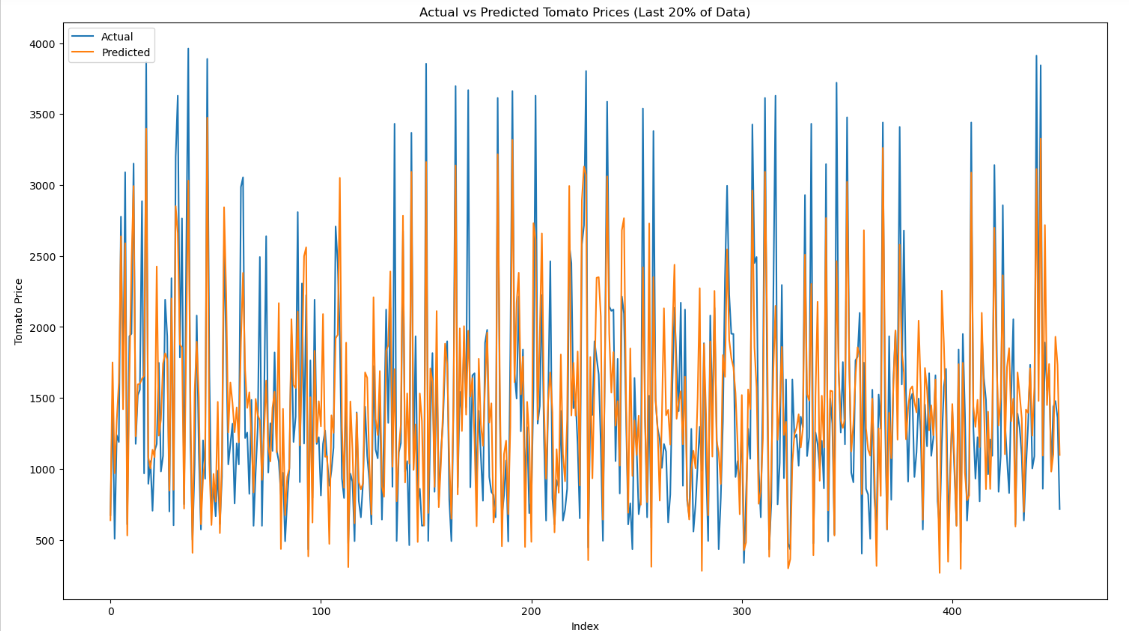


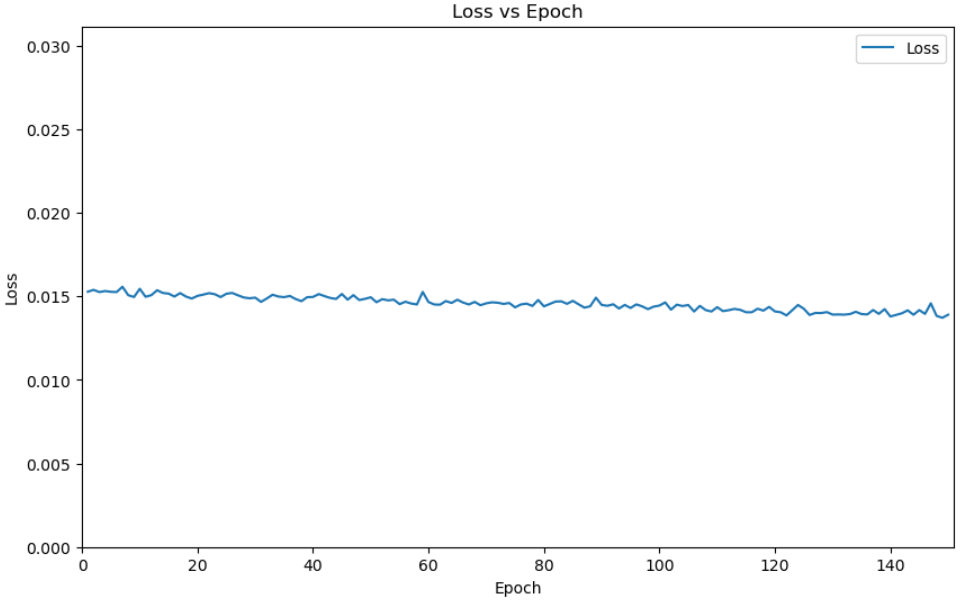


The RMSE Value for 90 Days Average is found to be 466.618

****

The RMSE Value for 120 Days Average is found to be 440.654





The RMSE Value for 150 Days Average is found to be 452.200

It is observed that the RMSE value is least for taking 120 Days Average of rainfall and temperature data, which should be quite intuitive, as tomatoes generally take 120 to 140 days to fully grow from plantation to harvest.

However, the least RMSE value for all the cases is still more than what we got through the ARIMA(4,0,0) model. Also, one should notice that the LSTM can predict general trends quite well, but struggles to predict the local minima and maxima accurately.

Hence, the LSTM model has its limitations

**Structural Equation Modeling**

We have seen how previous models are not accurate to a great extent, and fail to capture the effect of other factors. So, to get a better picture of the effect of other factors on the tomato prices, we use Structural Equation Modeling.

Structural Equation Modeling (SEM) stands out as a highly adaptable tool for comprehensively understanding the interplay among multiple variables in a sequential manner within a structural framework. It offers a robust approach for capturing the complexity of multivariate problems, closely mirroring real-world dynamics.

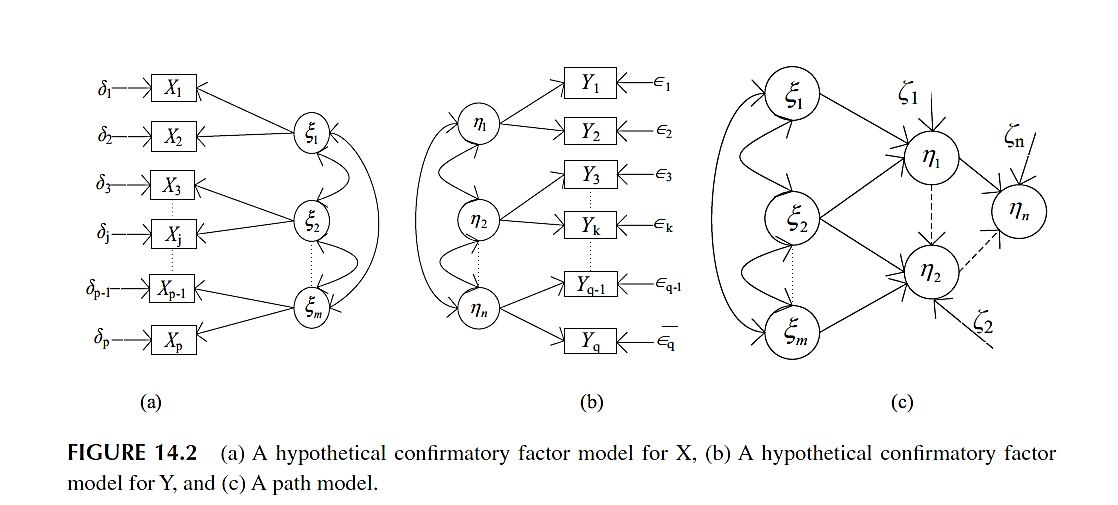
SEM involves the construction of multiple equations, reflecting its intricate statistical nature. It consists of two main components: the measurement model and the structural model. The measurement model serves to quantify latent variables (factors), while the structural model assesses the relationships between these latent variables.

Confirmatory factor analysis provides the mathematical framework for assessing the relationships between observed variables and latent constructs, while path analysis forms the foundation for examining the direct and indirect effects between variables in structural equation modeling.

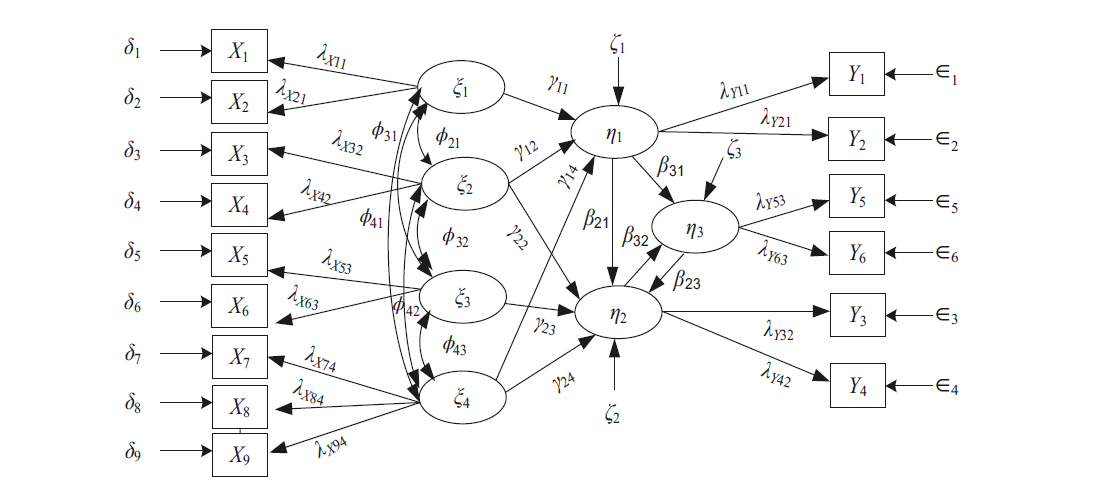
By utilizing these methodologies, SEM enables the exploration of the intricate interactions among latent constructs, facilitating a more profound comprehension of the research inquiry.

**SEM Model Architecture**

The LSTM (Long Short-Term Memory) model, crafted for forecasting purposes, constitutes a sequential neural network featuring several layers of LSTM cells, which are subsequently followed by Dense layers. This framework is specifically engineered to grasp intricate temporal dependencies within the data, rendering it well-suited for forecasting endeavors, such as predicting tomato prices in time series datasets.



The figures (a), (b) and (c) represent hypothetical confirmatory factor model for X, hypothetical confirmatory factor model for Y and a path model respectively



(d)

Figure (d) represents a complete hypothetical SEM model

Each variable is explained as follows:

**X** : manifest variable vector of exogenous variable

**Y** : manifest variable vector of endogenous variable

**δ** : vector of measurement model error for X

**ϵ** : vector of measurement model error for Y

**ζ** : vector of measurement model error for η

**ξ** : exogenous latent construct

**η** : endogenous latent construct

**λx**: exogenous factor ξ loading on the manifest variable *X*

**λy**: endogenous factor η loading on the manifest variable *Y*

**β** : The path coefficient between endogenous factor η

As SEM involves many relationships with different types of variables (latent exogenous, latent endogenous, explanatory manifest X explained manifest Y, and error variables for X, Y and η). The notations are categorized into three groups: variables, effect parameters and variance-covariance matrices. For a general SEM model, the notations are as follows:

* **Latent Variables**

η(n x 1)  = [ η1, η2, . . . , ηn ] vector of latent endogenous variables.

ξ(m x 1)  = [ ξ 1, ξ 2, . . . , ξ m ] vector of latent endogenous variables.

* **Manifest Variables**

Y(q x 1)  = [ Y 1, Y 2, . . . , Y q ] vector of explained manifest variables.

X(p x 1)  = [ X 1, X 2, . . . , X p ] vector of explanatory manifest variables.

* **Error Variables**

ε(q x 1)  = [ ε 1, ε 2, . . . , ε q ] vector of error variables for ***Y***.

δ(p x 1)  = [ δ 1, δ 2, . . . , δ p ] vector of error variables for ***x***.

ζ(n x 1)  = [ ζ 1, ζ 2, . . . , ζ n ] vector of error variables for η.

* **Factor Loadings**

λx (i, k) = The kth exogenous factor loading on ith explanatory manifest variable.

λy (j, l) = The lth exogenous factor loading on jth explanatory manifest variable.

Λx (p x m) = matrix of factor loadings involving ξ and X.

Λy (q x n) = matrix of factor loadings involving η and Y.

* **Causal Parameters**

*ϒlk*  = The path coefficient between the *l*- th η and the *k*- th ξ

*βll’*  = The path coefficient between the *l*- th η and the *l’*- th η

Г(n x m)  = matrix of path coefficients for the effects of ξon η

β(n x n) = matrix of path coefficient showing the influence of variables ηon each other

* **Dispersion Metrices**

Φ(m x m)= variance- covariance matrix of the *m* latent exogenous variables

Ψ ζ (m x m)= variance- covariance matrix of ζ

ϴ ϵ (m x m)= variance- covariance matrix of ϵ

Φ δ (m x m)= variance- covariance matrix of δ

**Model Equations**

* + - **Measurement Model**

In structural equation modeling (SEM), latent variables, representing underlying concepts, are associated with observable measures (indicators) through a statistical method known as factor analysis. This mirrors the approach psychologists employ to assess personality traits: while traits like introversion cannot be directly observed, behavioral questions (indicators) can be used to infer them (latent variables).

Each latent variable functions as a "behind-the-scenes" factor influencing multiple related indicators. The magnitude of this influence is quantified by a parameter known as lambda (λ). Conceptually, lambda represents the strength of the connection between the latent concept and its corresponding indicator. A higher lambda value indicates a more robust association.

* + - **Structural Model**

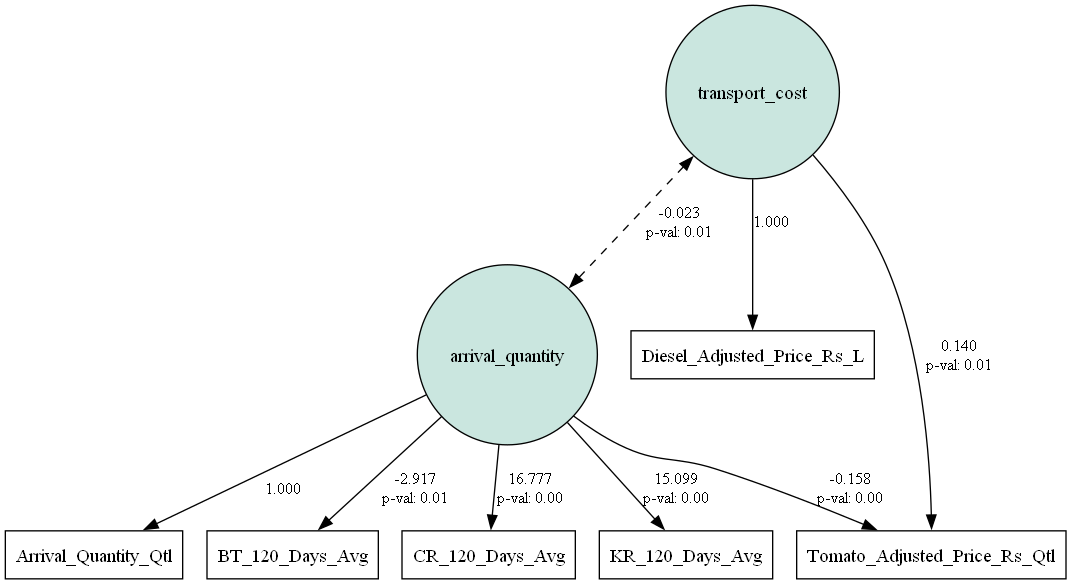
**The structural model is like a roadmap that shows how hidden concepts (latent variables) affect one another.** It focuses on the relationships between these underlying factors, not the individual measures used to assess them.

(I −β)η = ξ + Γ ξ

Additionally, there are certain assumptions done by the model:

1. The error variables have zero mean and are not correlated to each other or any other variable.
2. The error variables are multivariate normal
3. The (I −β) term in the structural equation is non-singular
4. All casual relationships are considered only in path model and are linear.

* **Path Diagram**



We have taken 120 days average of rainfall and temperature as variables, as we previously found out that tomato takes 120 days to become ready to harvest.

From the path diagram, we can infer different types of variables

* **Latent Variables**

1. **arrival\_quantity**: inferred from Arrival Quantity, Temperature, Chittoor Rainfall and Kolar Rainfall
2. **transport\_cost**: inferred from Diesel Prices

* **Manifest Variables**

Tomato Prices, Arrival Quantity, Temperature, Diesel Prices, Chittoor Rainfall and Kolar Rainfall are manifest variables

* **Regressions**

1. **Tomato Prices**: regressed on arrival\_quantity and transport\_cost

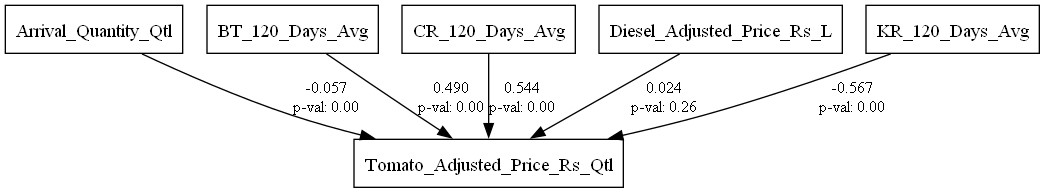
* **Fitness Measures of The Model**

For this model, the fitness parameters are as follows:

|  |  |
| --- | --- |
| **Fitness Index** | **Observed Value** |
| Chi-Squared | 1326.590166 |
| CFI | 0.7612 |
| GFI | 0.760583 |
| AGFI | 0.486964 |
| NFI | 0.760583 |
| TLI | 0.488285 |
| RMSEA | 0.289069 |
| AIC | 26.824466 |
| BIC | 106.929551 |
| LogLik | 0.587767 |

The χ2, RMSEA, AIC, BICvalues are supposed to be low, and the CFI, GFI, AGFI, NFI, TLI values should be close to 1, and LogLik should be very less if the model is supposed to be well fit, but the opposite is seen here.

Hence, we change the model, remove the latent variables, and regress the tomato prices on all the observed variables



Now, we calculate all the fitness parameters again

|  |  |
| --- | --- |
| **Fitness Index** | **Observed Value** |
| Chi-Squared | 0.000029 |
| CFI | 1.002717 |
| GFI | 1 |
| AGFI | 1 |
| NFI | 1 |
| TLI | 1.003623 |
| RMSEA | 0 |
| AIC | 12 |
| BIC | 46.330751 |
| LogLik | 1.30E-08 |

We can see that chi-squared is very low, CFI, GFI, AGFI, NFI, TLI are almost 1, along with low AIC, BIC and LogLik values. So, we conclude that this model is a near perfect fit. That means, no latent variables are required

* **Path Coefficients of The Model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Path** | **Estimate** | **Std. Err** | **z-value** | **p-value** |
| Tomato Price – Arrival Quantity | -0.056533 | 0.019289 | -2.9309 | 0.00338 |
| Tomato Price – Chittoor Rainfall | 0.54374 | 0.056188 | 9.6772 | 0 |
| Tomato Price – Kolar Rainfall | -0.567161 | 0.054135 | -10.476884 | 0 |
| Tomato Price – Bengaluru Temp | 0.489949 | 0.023844 | 20.548088 | 0 |
| Tomato Price – Diesel Prices | 0.023717 | 0.021082 | 1.124978 | 0.260598 |

The only observation is that the p-value of Tomato Price with Diesel Prices is greater than 0.05, so their relationship isn’t statistically significant

**Conclusion**

In conclusion, this research project delved into the intricate dynamics of tomato price prediction through a meticulous analysis employing various time series models and advanced forecasting techniques.

The research not only shed light on the volatility and complexities of agricultural markets, particularly the tomato market in Chennai, but also proposed effective forecasting methodologies to address the inherent challenges.

Through extensive data collection and pre-processing, including the incorporation of exogenous variables such as rainfall, temperature, and diesel prices, the study provided a robust foundation for analyzing tomato price trends. Utilizing models like ARIMA and LSTM, the research navigated through the nuances of time series data, capturing both linear trends and nonlinear relationships effectively.

Furthermore, the exploration of Structural Equation Modeling (SEM) offered a deeper understanding of the multifaceted factors influencing tomato prices. By constructing comprehensive measurement and structural models, the study unraveled the interplay among latent and manifest variables, providing valuable insights into the intricate relationships within the tomato market.

The findings of this research hold significant implications for stakeholders in the agricultural sector, including farmers, traders, and policymakers. By enabling informed decision-making through accurate price forecasting, the study empowers stakeholders to optimize resource management, enhance operational efficiency, and ultimately contribute to the sustainable growth of the tomato market in Chennai and beyond.

Overall, this thesis contributes to the growing body of knowledge in agricultural economics and time series forecasting, offering valuable methodologies and insights for future research and practical applications in agricultural markets.

**References**

● Smith et al (2019) "The Role of Social Support in the Relationship Between Self-Esteem and Depression: A Structural Equation Modeling Approach"

● Muthén and Asparouhov (2018) "Bayesian SEM: A More Flexible Representation of Substantive Theory"

● Jhareswar Maiti (2022) Multivariate Statistical Modeling in Engineering and Management

● Lama, A., Jha, G.K., Paul, R.K. and Gurung, B. (2015), “Modeling and forecasting of price volatility: an application of GARCH and EGARCH models”

● Ramirez, O. and Fadiga, M. (2003), “Forecasting agricultural commodity prices with asymmetric error GARCH models”

● Sekhar, C., Roy, D. and Bhatt, Y. (2017), “Food inflation and food price volatility in India trends and determinants”

● Onour, I. and Sergi, B.S. (2011), “Modeling and forecasting volatility in the global food commodity prices”

● Bhardwaj, S.P., Paul, R.K., Singh, D.R. and Singh, K.N. (2014), “An empirical investigation of arima and garch models in agricultural price forecasting”

● Zhiyuan Chen et al (2021), “Automated Agriculture Commodity Price Prediction System with Machine Learning Techniques”

● Kiran M. Sabu et al (2019), “Predictive analytics in Agriculture: Forecasting prices of Arecanuts in Kerala

● Feihu Sun et al (2023), “Agricultural Product Price Forecasting Methods: A Review”