TIME SERIES AND MACHINE LEARNING MODEL FOR CORPORATION WASTAGE IN TIRUNELVELI.

Major Project Work Submitted to Manonmaniam Sundaranar University in partial Fulfillment of the requirements for the award of the degree of

Master of Science in Statistics

Submitted by

Velkumar. V

(Reg. No. 20214012529133)

In partial fulfillment for the award of the degree of

MASTER OF SCIENCE IN STATISTICS

DEPARTMENT OF STATISTICS

MANONMANIAM SUNDARANAR UNIVERSITY

TIRUNELVELI – 627 012

APRIL 2023

BONAFIDE CERTIFICATE

Certified that this Project titled "**Time Series and Machine Learning Model for Corporation Wastage in Tirunelveli**." is the Bonafide work of **Velkumar.V (Reg. No.20214012529133)** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

Signature of the HOD

Dr. A. RAJARATHINAM
Professor and Head
Department of Statistics
Manonmaniam Sundaranar University
Tirunelveli – 627 012

Signature of the Supervisor

Dr. P. ARUMUGAM,
Professor
Department of Statistics
Manonmaniam Sundaranar University
Tirunelveli – 627 012

Declaration

I hereby declare that the project entitled "TIME SERIES AND MACHINE LEARNING MODEL FOR CORPORATION WASTAGE IN TIRUNELVELI." submitted to Manonmaniam Sundaranar University I partial fulfillment of the requirements for the award of the degree of Master of Science in Statistics is my original work and that it has not previously formed the basis for the award of any degree, diploma or similar title of any University or Institution.

Tirunelveli (V. Velkumar)

April 2023

ABSTRACT

In recent years, there has been an increasing focus on sustainable waste management practices in India. Corporation wastage refers to the management and disposal of waste generated by municipal corporations or local governing bodies in Indian cities and towns. This study aims to explore the best models for predicting corporation wastage in Tirunelveli city. The data on generation of non-Biodegradable waste in Tirunelveli town has been collected from Tirunelveli corporation from July 2022 to February 2023. In this study accurate predictions of waste generation are made using the, SARIMAX and XG-Boost. The performance of each model has been assessed using root mean squared error (RMSE) value. Accurate wastage prediction models derived from this study can play a vital role in aiding the corporation in Tirunelveli to enhance waste management planning, optimize resource allocation, and implement effective strategies for waste reduction, thus contributing to the city's sustainable development objectives.

ACKNOWLEDGEMENT

I would thank God for being able to complete this project with success. Then I would like to special thanks of gratitude to my Guide **Dr. P. Arumugam,** Professor, Department of Statistics, for his guidance who gave me the golden opportunity to do this wonderful project on the topic '**TIME SERIES AND MACHINE LEARNING MODEL FOR CORPORATION WASTAGE IN TIRUNELVELI.**' which also helped me in doing a lot of analysis and I came to know about so many new things and as well as providing necessary information regarding project and also for his support to completing the project.

I wish to express my sincere thanks to **Dr. A. Rajarathinam,** Professor and head, Department of Statistics, **Dr. K. Senthamarai Kannan,** Senior Professor, Department of Statistics, **Dr.R.Sasikumar,** Assistant Professor, Department of Statistics, **Dr.V.Deneshkumar** Assistant Professor, Department of Statistics **Dr. K. Manoj,** Assistant Professor, Department of Statistics, for their help rendered during my study period.

V.Velkumar

TABLE OF CONTENT

CHAPTER -I

1.1 INTRODUCTION

Tirunelveli district is situated in the southern part of Tamil Nadu and is surrounded by Virudhunagar district to the north, Thoothukudi district to the east, the Gulf of Mannar and the Indian Ocean to the south, and Kanyakumari district and Kerala state to the west. The district boasts a diverse topography, ranging from the majestic Western Ghats in the north to the picturesque coastal plains in the south. Its tropical climate, with hot summers and moderate winters, shapes the natural environment.

According to the 2011 Census of India, Tirunelveli district has a population of approximately 3.7 million people, making it one of the most populous districts in Tamil Nadu. The district is home to various ethnic and linguistic groups, and the major religions practiced include Hinduism, Islam, and Christianity. With a commendable literacy rate of around 82%, the district demonstrates the importance of education and awareness.

Tirunelveli Corporation, the administrative body responsible for managing the district's urban areas, is divided into four zones. Each zone has its own unique characteristics and population distribution. The corporation generates a significant amount of solid waste, estimated to be around 375 tonnes per day. This waste comprises organic waste, plastic waste, paper waste, and electronic waste, reflecting the changing consumption patterns and urbanization in the district.

The rapid population growth, urbanization, and changes in consumption patterns are the primary factors contributing to the escalating waste generation in Tirunelveli district. As the population increases, so does the amount of waste generated. Urbanization further intensifies waste generation due to the growth of commercial and industrial establishments. Changes in consumption patterns, driven by evolving lifestyles and increased product consumption, also play a significant role.

One of the key challenges in waste management is the lack of awareness among the public about waste segregation and disposal. The importance of proper waste management practices, including recycling and composting, needs to be effectively communicated to ensure active participation. Additionally, Tirunelveli district faces a shortage of infrastructure and

resources for waste collection, transportation, and disposal. The lack of dedicated personnel trained in waste management further compounds the challenges.

The waste management system in Tirunelveli district is primarily based on door-to-door collection by sanitary workers. The collected waste is transported to transfer stations, where it undergoes sorting and compaction before being transported to landfill sites for disposal. Efforts to manage electronic waste and promote composting have been initiated but require further attention and implementation.

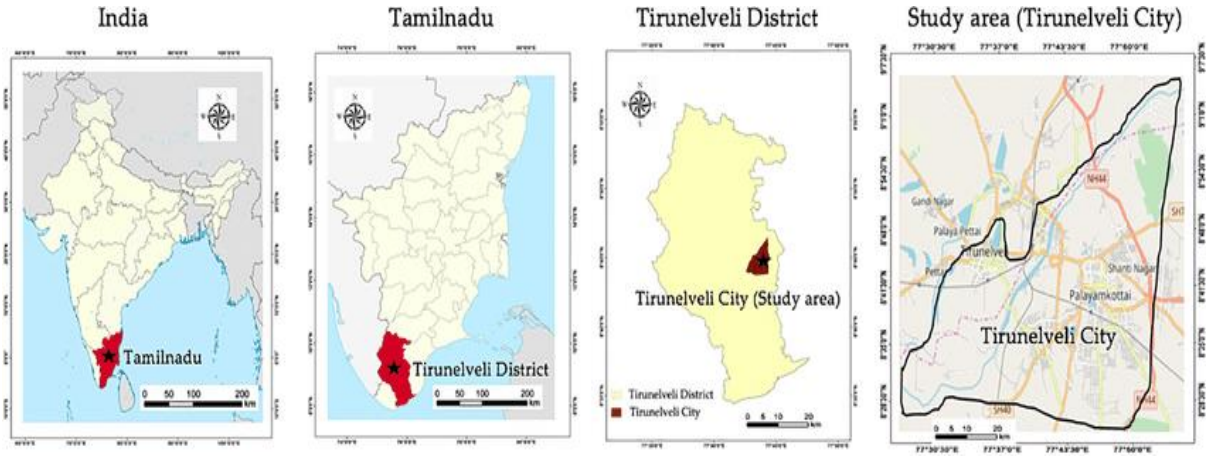

Figure 1.1

Tirunelveli district faces significant challenges in waste management, including waste generation, awareness, infrastructure, and human resources. This project aims to analyze these challenges and propose strategies to address them effectively. By fostering awareness, improving infrastructure, and implementing sustainable waste management practices, Tirunelveli can strive towards a cleaner and greener future for its residents and contribute to the larger goal of environmental sustainability.

Waste management is a crucial aspect of maintaining a clean and sustainable environment. In Tirunelveli district, located in the southern part of Tamil Nadu, India, effective waste management is essential to address the challenges posed by the growing population, urbanization, and changes in consumption patterns. This project aims to provide a detailed analysis of the waste management process and methods employed in Tirunelveli district, highlighting the challenges and exploring potential solutions.

Tirunelveli district generates a substantial amount of waste, comprising organic waste, plastic waste, paper waste, and electronic waste. The factors contributing to waste generation include population growth, urbanization, and changes in consumption patterns. As the population increases and urban areas expand, the amount of waste generated also rises. The evolving lifestyles and increased consumption of products further contribute to waste accumulation.

Figure:1.2

1.1.2 Waste Management Process in Tirunelveli District:

1. Waste Collection: The waste management process begins with door-to-door waste collection by sanitary workers. They collect waste from households, commercial establishments, and industries. Proper segregation of waste at the source is encouraged to facilitate efficient handling and disposal.
2. Transfer and Sorting: Collected waste is transported to transfer stations, where it undergoes sorting. The waste is segregated into different categories such as biodegradable waste, recyclables, and non-recyclables. This step is crucial for effective waste management and recycling initiatives.
3. Waste Treatment and Disposal: After sorting, the waste undergoes treatment based on its category. Biodegradable waste can be processed through composting, where it is

converted into nutrient-rich compost for agricultural use. Recyclable materials such as plastics, paper, and metals are sent to recycling facilities for processing. Non-recyclable and hazardous waste is disposed of in designated landfill sites, following proper safety and environmental regulations.

1.1.3 Segregation Of Wastes:

- **Biodegradable Waste:** Biodegradable waste, including food scraps, garden waste, and other organic materials, can be treated through composting. Composting is a natural process where microorganisms break down organic waste into nutrient-rich compost. This compost can be used as a soil conditioner, improving soil fertility and promoting sustainable agricultural practices. Composting facilities or compost pits are established to handle biodegradable waste effectively.
- **Recyclables**: Recyclable materials such as plastics, paper, glass, and metals are separated at the sorting facility and sent to recycling units. These materials are processed through various recycling techniques, including shredding, melting, and reprocessing. The recycled materials are then used as raw materials in the production of new products, reducing the need for virgin resources. Recycling initiatives help conserve natural resources, reduce energy consumption, and minimize environmental pollution.

4. **Non-Recyclables and Hazardous Waste:** Non-recyclable waste that cannot be composted or recycled, along with hazardous waste, requires special handling. These waste materials are disposed of in designated landfill sites that adhere to safety and environmental regulations. Proper landfill management practices, including the lining of the landfill area, leachate collection systems, and methane gas control measures, are implemented to minimize negative environmental impacts.

Promoting public awareness and education is an integral part of waste management in Tirunelveli district. Various initiatives are undertaken to educate residents, businesses, and schools about the importance of waste segregation, recycling, and responsible waste disposal practices. Awareness campaigns, workshops, and educational programs are organized to empower the community to actively participate in waste management efforts.

Continuous monitoring of the waste management process is essential to identify areas for improvement and ensure compliance with regulations. Regular assessments and audits are conducted to evaluate the efficiency and effectiveness of waste management practices in Tirunelveli district. Feedback from the public and stakeholders is considered to implement necessary changes and enhance the overall waste management system.

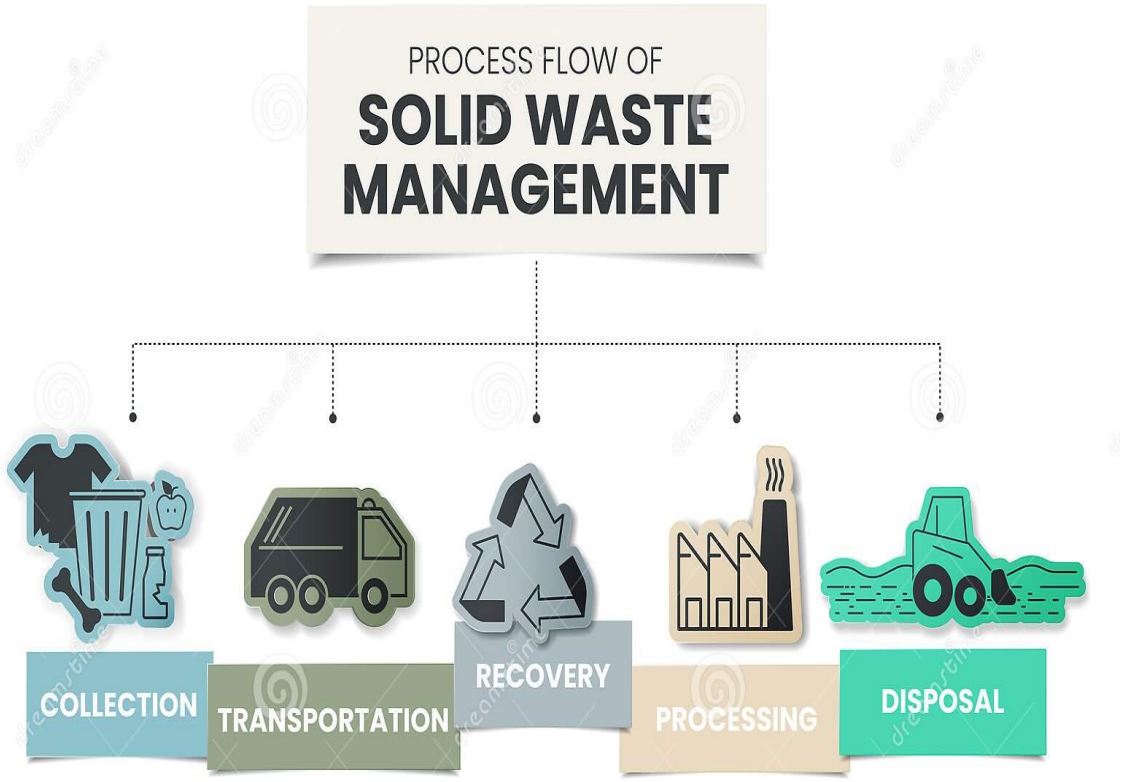

Figure:1.3

1.1.4 Challenges in Waste Management:

- Lack of Awareness and Education:

  One of the primary challenges in waste management is the lack of awareness and education among the public. Many residents are unaware of the environmental impact of improper waste disposal and the benefits of waste segregation and recycling. Without proper knowledge and understanding, individuals may not actively participate in waste management practices. Initiatives to educate the public about the importance

of waste management, including awareness campaigns, workshops, and educational programs, are crucial in addressing this challenge. By raising awareness and providing information on proper waste disposal methods, recycling, and composting, the community can be empowered to take responsibility for their waste and make informed choices.

- Insufficient Infrastructure:

  Tirunelveli district faces challenges related to insufficient infrastructure for waste management. This includes a lack of adequate waste collection vehicles, transfer stations, recycling facilities, and landfill sites. Inadequate infrastructure can hinder the efficient and effective handling of waste. For example, limited waste collection vehicles may lead to delays in waste collection, resulting in overflowing bins and unsightly litter. Insufficient transfer stations can lead to long transportation distances, increasing the cost and environmental impact of waste management. To address this challenge, there is a need for investment in infrastructure development, including the establishment of modern waste management facilities, expansion of transfer stations, and improvement of landfill sites. By enhancing the infrastructure, the waste management process can be streamlined, and the overall efficiency can be improved.

- Resource and Manpower Shortage:

  Waste management requires adequate resources and trained personnel to handle the various stages of the process effectively. However, Tirunelveli district faces challenges related to resource and manpower shortage. Insufficient financial resources may limit the implementation of waste management initiatives and hinder the procurement of necessary equipment and facilities. Additionally, there may be a shortage of skilled individuals with expertise in waste management techniques. The effective implementation of waste management practices requires trained personnel who can handle waste collection, segregation, treatment, and disposal in a safe and sustainable manner. To address this challenge, it is important to allocate sufficient resources for waste management programs and invest in capacity building initiatives. Training programs can be conducted to enhance the skills of the existing workforce and create a pool of trained professionals in waste management.

- Community Participation and Behavioral Change:

Figure:1.4

Encouraging community participation and promoting behavioral change are essential for the success of waste management initiatives. However, changing established habits and practices can be challenging. Some individuals may resist adopting waste segregation practices or may not feel motivated to reduce, reuse, and recycle their waste. Overcoming this challenge requires comprehensive engagement strategies that involve the community at various levels. It is important to involve residents, community groups, schools, and businesses in waste management programs and provide incentives and rewards for responsible waste disposal practices. Continuous communication, awareness campaigns, and community involvement can help foster a sense of ownership and responsibility among residents, leading to positive behavioral changes.

Figure1.5

1.1.5 Promoting Sustainable Waste Management:

- Public Participation and Awareness Programs:

  Creating awareness among the public about the importance of waste management is crucial. Public participation can be encouraged through campaigns, workshops, and community engagement programs. These initiatives can promote waste segregation, recycling, and responsible disposal practices.

- Strengthening Infrastructure:

  Investing in infrastructure development is essential for an efficient waste management system. This includes the establishment of waste collection centers, transfer stations, recycling facilities, and modern landfill sites. Adequate funding and collaboration between government bodies, private sectors, and NGOs can help improve infrastructure.

- Encouraging Recycling and Composting:

  Promoting recycling initiatives and composting can significantly reduce the amount of waste that ends up in landfills. Implementing effective recycling programs and providing incentives for businesses and residents to participate can contribute to a more sustainable waste management process.

- Technological Interventions:

  Adopting advanced technologies such as waste-to-energy conversion, anaerobic digestion, and smart waste management systems can optimize waste treatment processes. These technologies can improve waste processing efficiency, reduce environmental impact, and generate renewable energy.

Effective waste management is crucial for the sustainable development of Tirunelveli district. By addressing the challenges and implementing comprehensive waste management strategies, the district can minimize the environmental impact of waste, promote recycling and composting, and enhance the overall cleanliness and well-being of its residents. Through public awareness, infrastructure development, and the adoption of innovative technologies, Tirunelveli can become a role model for efficient waste management practices in the region and contribute to a cleaner and greener future.

Tirunelveli Corporation, located in the southern part of Tamil Nadu, plays a vital role in ensuring the overall well-being and environmental sustainability of its residents. With the

city divided into four zones - Thachanallur, Palayamkottai, Town, and Melapalayam - effective waste management becomes crucial in maintaining cleanliness, hygiene, and a healthy living environment for the residents. In this chapter, we will delve into the statistical data of waste management in each zone of Tirunelveli district, including the amount and types of waste generated, as well as the current waste management practices.

1.2 Zones of Tirunelveli:

- Thachanallur Zone:

  The Thachanallur Zone, comprising two units, namely Unit 1 and Unit 2, covers several wards within its jurisdiction. Unit 1 includes ward numbers 5, 6, 7, 8, 9, 12, 13, 14, and 30, while Unit 2 comprises ward numbers 4, 10, 28, and 29. According to the data collected from the Thachanallur zone, an average of 20 metric tonnes of waste is generated daily. The majority of the waste in this zone is organic, followed by plastic and paper waste. To manage the waste effectively, the zone follows a door-to-door waste collection system. The collected waste is then transported to the designated dumping yard located on the outskirts of the city.

- Palayamkottai Zone:

  The Palayamkottai Zone, consisting of three units - Unit 5, Unit 6, and Unit 7 - covers a significant area of Tirunelveli. Unit 5 includes ward numbers 5, 6, 7, 8, and 37, while Unit 6 comprises ward numbers 9, 32, 33, 34, and 35. Unit 7 consists of ward numbers 36, 38, 39, and 55. The average daily waste generation in Palayamkottai Zone is approximately 40 metric tonnes. Similar to other zones, organic waste constitutes the majority of the waste generated, followed by plastic and paper waste. The waste collection system in this zone also relies on door-to-door collection, with the collected waste being transported to the designated dumping yard.

- Tirunelveli Town Zone:

  The Town Zone, divided into two units - Unit 3 and Unit 4 - encompasses various wards. Unit 3 includes ward numbers 15, 16, 23, 24, 25, 26, and 27, while Unit 4 comprises ward numbers 17, 18, 19, 20, 21, and 22. According to the statistical data collected from the Town zone, an average of 30 metric tonnes of waste is generated

daily. Organic waste constitutes the major portion of the waste generated in this zone, followed by plastic and paper waste. The waste collection system employed in this zone is door-to-door collection, ensuring efficient waste management. The collected waste is then transported to the designated dumping yard located on the outskirts of the city.

- Melapalayam Zone:

  The Melapalayam Zone, divided into three units - Unit 8, Unit 9, and Unit 10 - covers a substantial area within Tirunelveli. Unit 8 includes ward numbers 31, 44, 48, 49, and 50, while Unit 9 comprises ward numbers 43, 52, 45, 46, and 47. Unit 10 consists of ward numbers 40, 41, 42, 51, 53, and 54. The average daily waste generation in Melapalayam Zone is approximately 35 metric tonnes. Similar to other zones, organic waste is the primary component of the waste generated, followed by plastic and paper waste. The waste collection system implemented in this zone is also door-to-door collection, ensuring regular waste pickup. The collected waste is then transported to the designated dumping yard on the outskirts of the city.

Figure:1.6

Waste management plays a crucial role in maintaining a clean and healthy living environment in Tirunelveli district. Each zone, including Thachanallur, Palayamkottai, Town, and Melapalayam, follows a door-to-door waste collection system to effectively manage the waste

generated. The majority of the waste in all zones is organic, followed by plastic and paper waste. However, despite the existing waste management practices, Tirunelveli district faces challenges such as inadequate infrastructure, lack of public awareness, and insufficient funding. These challenges need to be addressed to enhance waste management and promote a sustainable and eco-friendly environment for the residents of Tirunelveli.

CHAPTER II

METHODOLOGY

2.1 METHOD OF DATA COLLECTION

The data were collected from the, 'Tirunelveli District Corporation Office'. The observation is based on a period of 243 days of day wise` data.

Data link: https://1drv.ms/u/s!ArQt3a3kkHmtgeBJeALAbJrxPrh8SQ?e=hXJhTH

2.2 Stationarity

A stationary time series is one whose statistical properties, such as mean, variance, and autocovariance, remain constant over time. In other words, it does not exhibit any systematic trends, cycles, or significant changes in statistical characteristics across different time periods

2.2.1 Importance of Stationarity:

Forecasting Accuracy: Stationary time series are easier to model and forecast compared to non-stationary series. Statistical models, such as ARIMA and SARIMA, are built on the assumption of stationarity, and violating this assumption can lead to inaccurate predictions.

Statistical Inference: Many statistical tests and techniques, such as hypothesis testing and confidence intervals, assume stationarity. Violating stationarity assumptions can lead to incorrect conclusions and unreliable results.

Time Series Decomposition: Stationary series can be decomposed into trend, seasonality, and residual components, facilitating a better understanding of the underlying patterns and dynamics in the data.

2.2.2 Testing of Stationarity

Several statistical tests can help determine whether a time series is stationary or not. The most commonly used test is the Augmented Dickey-Fuller (ADF) test, which tests the null hypothesis that a unit root is present in a time series (indicating non-stationarity).

If a time series is found to be non-stationary, it can be made stationary through a process called differencing. Differencing involves computing the differences between consecutive

observations to remove trends or seasonality. A stationary series obtained after differencing is suitable for modeling and analysis.

2.2.3 Formula

One commonly used statistical test for stationarity is the Augmented Dickey-Fuller (ADF) test. The ADF test examines whether a unit root (i.e., a root of the characteristic equation with modulus equal to 1) is present in the time series. The null hypothesis of the ADF test is that the time series is non-stationary, meaning it has a unit root. The alternative hypothesis is that the time series is stationary.

The general formula for the Augmented Dickey-Fuller (ADF) test is:

$$\Delta yt = \rho yt - 1 + \Sigma \varphi i \Delta yt - i + \varepsilon t$$

Where,

Δyt represents the differenced time series (first difference or higher order differences).

$yt - 1$ represents the lagged value of the time series.

$\Sigma \varphi i \Delta yt - i$ represents the sum of autoregressive terms on the differenced time series.

εt is the error term.

2.3 Exploratory Data Analytics (EDA)

Exploratory Data Analysis (EDA) is an approach to analysing and summarizing data sets to uncover patterns, relationships, and insights. It involves examining and visualizing data to understand its main characteristics, identify trends, detect outliers, and formulate hypotheses for further analysis. EDA is often the first step in the data analysis process and helps researchers and data scientists gain a better understanding of the data before applying more advanced statistical techniques.

There is no specific statistical formula or parameter for EDA as it encompasses a wide range of techniques and methods. However, some common statistical measures and techniques used in EDA include:

- Descriptive Statistics: Descriptive statistics summarize and describe the main characteristics of the data, including measures of central tendency (mean, median, mode) and measures of dispersion (range, standard deviation, variance).

- Data Visualization: Data visualization techniques, such as histograms, box plots, scatter plots, and bar charts, are used to visually explore the data and identify patterns, distributions, and outliers.
- 5. Data Transformation: Data transformation techniques, such as log transformations, scaling, and normalization, are applied to make the data more suitable for analysis and modelling.

2.3.1 Uses of Exploratory Data Analysis:

- Data Cleaning and Preprocessing
- Pattern Discovery
- Feature Selection
- Outlier Detection
- Hypothesis Generation

Overall, EDA provides a foundation for further analysis, model building, and decision-making by uncovering key insights and understanding the underlying structure of the data. It allows data analysts to explore data in a comprehensive and systematic manner, leading to more meaningful interpretations and informed decision-making.

2.4 SARIMAX

2.4.1 History of Sarimax

SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) is an extension of the SARIMA (Seasonal AutoRegressive Integrated Moving Average) model. SARIMA itself is an extension of the ARIMA (AutoRegressive Integrated Moving Average) model. The ARIMA model was first introduced by George Box and Gwilym Jenkins in the 1970s, providing a powerful framework for time series analysis and forecasting. SARIMAX was developed to incorporate exogenous variables, allowing for better modeling of time series data.

2.4.2 Advantages of Sarimax

- ❖ SARIMAX is effective for modeling and forecasting time series data with seasonality and other time-dependent patterns.
- ❖ It can handle both univariate (single variable) and multivariate (multiple variables) time series.

- The inclusion of exogenous variables enables the model to capture additional factors that may influence the time series.

2.4.3 Applications of Sarimax

- Economic forecasting: SARIMAX can be used to forecast economic indicators such as GDP, inflation rates, and stock market prices.
- Demand forecasting: It is widely applied in retail and supply chain industries to predict demand for products or services.
- Energy consumption forecasting: SARIMAX is used to forecast electricity, gas, or water consumption for efficient resource planning.
- Environmental modeling: It can be used to analyze and predict environmental variables like temperature, air quality, or precipitation.

2.4.4 Formula and Parameters of Sarimax

SARIMAX stands for Seasonal AutoRegressive Integrated Moving Average with eXogenous variables. It is an extension of the SARIMA model that incorporates additional exogenous variables to improve forecasting accuracy. SARIMAX is useful when the time series data exhibit seasonality and require additional explanatory variables.

The general formula for SARIMAX is as follows:

$$Y(t) = c + \phi 1Y(t-1) + \phi 2Y(t-2) + \ldots + \phi pY(t-p) + \theta 1\varepsilon(t-1) + \theta 2\varepsilon(t-2) + \ldots + \theta q\varepsilon(t-q) + \beta 1X1(t) + \beta 2X2(t) + \ldots + \beta nXn(t) + \varepsilon(t)$$

In this formula:

Where,

$Y(t)$ -represents the dependent variable at time t.

c - is a constant term.

$\phi 1, \phi 2, \ldots, \phi p$ - are the autoregressive coefficients.

$\theta 1, \theta 2, \ldots, \theta q$ - are the moving average coefficients.

$\varepsilon(t)$ - represents the error term at time t.

$X1(t), X2(t), \ldots, Xn(t)$ - are the exogenous variables at time t.

$\beta 1, \beta 2, \ldots, \beta n$ - are the corresponding coefficients for the exogenous variables.

The parameters to be determined when fitting a SARIMAX model include:

p: The order of the autoregressive component.

d: The order of differencing (to make the time series stationary).

q: The order of the moving average component.

P: The seasonal order of the autoregressive component.

D: The seasonal order of differencing.

Q: The seasonal order of the moving average component.

s: The length of the seasonal period.

β1, β2, ..., βn: Coefficients for the exogenous variables.

2.4.5 Seasonality

SARIMAX accounts for seasonal patterns in the data by including seasonal differencing. Seasonal differencing removes the seasonal component by subtracting the series value from the same time in the previous season. The seasonal differencing order is denoted by the "D" parameter in SARIMAX.

2.4.6 Model Identification

The next step is to identify the appropriate order of autoregressive (AR), integrated (I), and moving average (MA) terms for the non-seasonal part of the model. This is done by examining the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the differenced data. The ACF helps determine the MA order (q), while the PACF helps determine the AR order (p). The integrated order (d) is already determined based on the differencing performed in the first step.

2.4.7 Seasonal Model Identification

Similarly, the appropriate order of seasonal autoregressive (SAR), seasonal integrated (SI), and seasonal moving average (SMA) terms needs to be identified. This is done by analyzing the ACF and PACF plots of the seasonally differenced data. The seasonal MA order (Q) is determined by the ACF, while the seasonal AR order (P) is determined by the PACF. The seasonal integrated order (D) is already determined based on the seasonal differencing.

2.4.8 Exogenous Variables

SARIMAX allows for the inclusion of exogenous variables, which are external factors that can influence the time series. These variables can be incorporated into the model as additional regressors. The presence of exogenous variables is denoted by the "X" parameter in SARIMAX.

2.4.9 Model Estimation

Once the model order is determined, the parameters of the SARIMAX model are estimated using maximum likelihood estimation or a similar optimization technique. This involves fitting the model to the available historical data.

2.4.10 Model Validation and Forecasting

After estimating the model parameters, it is crucial to validate the model's performance using appropriate evaluation metrics and diagnostics. Once the model is deemed satisfactory, it can be used to forecast future values of the time series, including the impact of exogenous variables.

2.5 XGBoost

XGBoost (eXtreme Gradient Boosting) is a machine learning algorithm introduced by Tianqi Chen in 2014. It is a powerful and scalable implementation of the gradient boosting framework, which combines the predictions of multiple weak learners (decision trees) to create a strong predictive model.

2.5.1 Advantages of XGBoost

- XGBoost is highly flexible and can handle a wide range of regression and classification tasks, including time series forecasting.

- It offers excellent predictive performance and is known for its ability to handle large and complex datasets.
- XGBoost provides built-in regularization techniques to prevent overfitting and enhance model generalization.
- It supports parallel processing, making it efficient for training and prediction on modern computing architectures.

2.5.2 Applications of XGBoost

- ❖ Time series forecasting: XGBoost can be used to forecast time series data by considering historical patterns and relevant features.
- ❖ Financial modeling: It is widely used in quantitative finance for tasks such as stock market prediction, credit risk modeling, and algorithmic trading.
- ❖ Healthcare: XGBoost is applied in medical research for disease diagnosis, patient outcome prediction, and personalized medicine.
- ❖ Natural language processing (NLP): It can be used for sentiment analysis, text classification, and document similarity tasks.

2.5.3 Gradient Boosting

XGBoost is based on the gradient boosting framework. Gradient boosting involves building an ensemble of weak learners (typically decision trees) sequentially, with each learner trying to correct the errors of the previous learners. This results in a strong predictive model.

2.5.4 Objective Function

XGBoost defines an objective function that needs to be optimized during model training. The objective function consists of a loss function to measure the model's performance and a regularization term to control the complexity of the model. The choice of the loss function depends on the task at hand, such as regression or classification.

2.5.5 Tree Boosting

XGBoost builds decision trees as weak learners. Decision trees are constructed in a greedy manner by recursively partitioning the data based on features and thresholds that minimize the

objective function. The trees are added to the ensemble one at a time, with each subsequent tree attempting to reduce the errors of the previous trees.

2.5.6 Feature Splitting

XGBoost incorporates a technique called "feature splitting" to handle missing values. Instead of treating missing values as a separate category, XGBoost automatically learns how to use the missing values during tree construction, which reduces the need for data preprocessing.

2.5.7 Regularization

XGBoost includes regularization techniques to control the complexity of the model and prevent overfitting. Regularization parameters include the learning rate, which controls the contribution of each tree, and the tree-specific parameters such as maximum depth, minimum child weight, and subsampling rate.

2.5.8 Parallel Processing and Optimization

XGBoost is designed to be highly efficient and scalable. It supports parallel processing, allowing multiple cores or machines to work together during training. Additionally, XGBoost includes various optimization techniques, such as approximate tree learning and column block compression, to further enhance its speed and memory efficiency.

2.5.9 Model Evaluation and Tuning

During training, XGBoost continuously monitors the model's performance on a separate validation set. This allows for early stopping, where training can be halted if the model's performance starts deteriorating, preventing overfitting. XGBoost also provides tools for hyperparameter tuning to find the best combination of parameters for optimal performance.

2.5.10 Prediction and Feature Importance

Once the model is trained, it can be used to make predictions on new unseen data. XGBoost provides methods to interpret the model's predictions, such as calculating feature

importance scores, which indicate the relative importance of each feature in the prediction process.

CHAPTER- III

Analysis and Interpretation

3.1 Descriptive Statistics

Table:3.1 Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error	Statistic	Std. Error
TH	243	0	19165	9817.68	182.860	2850.505	-.355	.156	1.630	.311
Valid N (listwise)	243									

From the above table3.1 given dataset consists of 243 observations. The minimum v alue is 0, and the maximum value is 19,165. The mean (average) value of the dataset is 9,817. 68, with a standard deviation of 182.860, indicating some variability in the data. The skewnes s value of -0.355 suggests a slight negative skew, indicating that the data is slightly skewed to the left. The kurtosis value of 1.630 indicates positive excess kurtosis, suggesting heavier tails and potential outliers in the distribution. Overall, the dataset exhibits some departure from a n ormal distribution. There are no missing values in the dataset, as indicated by the valid N (list wise) count of 243.

3.2 Time plot

Here, “TH” variables in analysis refer to Tirunelveli Town Zone.

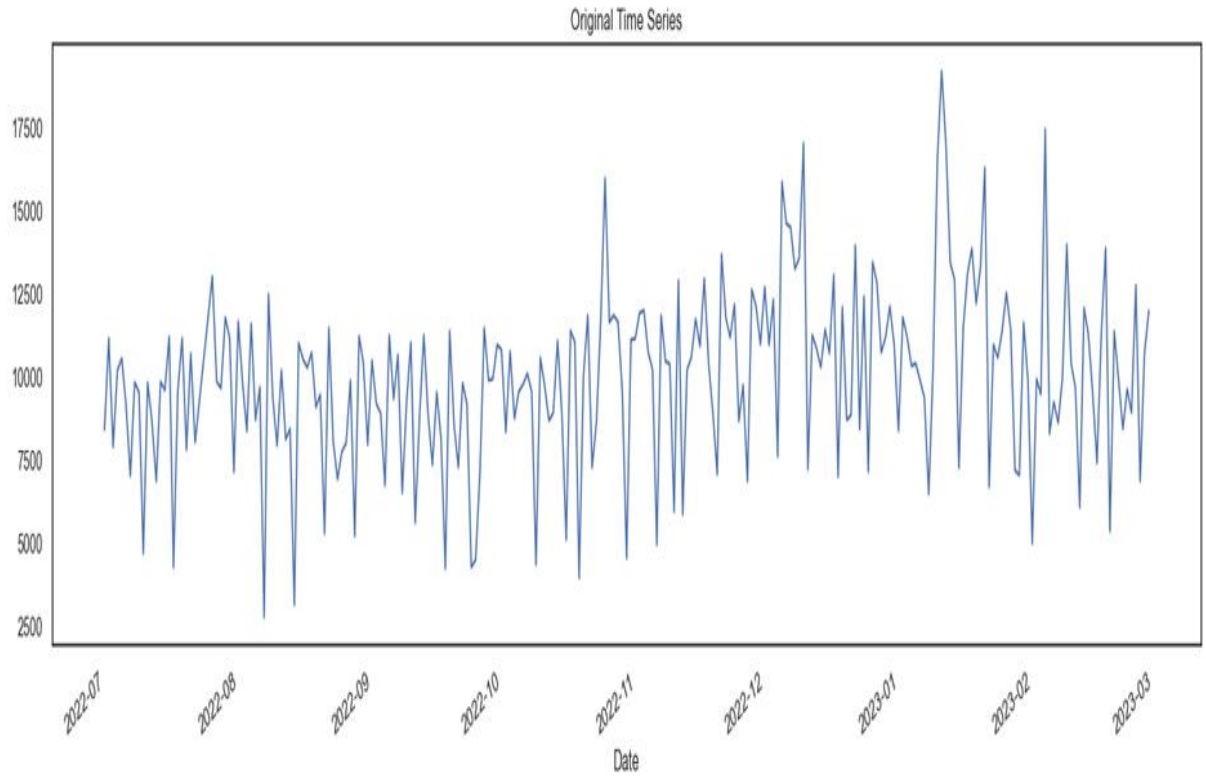

Figure:3.1-Time Plot

From the above (figure 3.1) time plot graph, the x-axis represents the dates or time periods, while the y-axis represents the values of the "TH" variable. Each data point represent s the value of "TH" recorded at a specific date.

The time plot allows you to visualize the changes and patterns in the "TH" values o ver time. By examining the plot, you can identify trends, seasonal patterns, outliers, or other i nteresting features in the data.

3.3 Stationarity

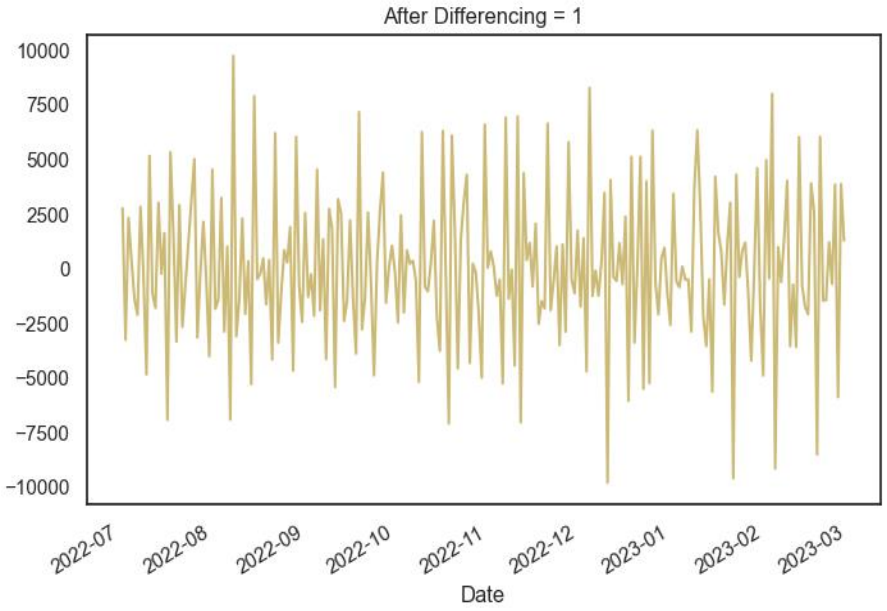

Figure :3.2

From above the (figure:3.2) “TH” variable has been differenced once (first-order dif ferencing), it means that the original values have been subtracted by their lag-1 values.

3.4 Histogram

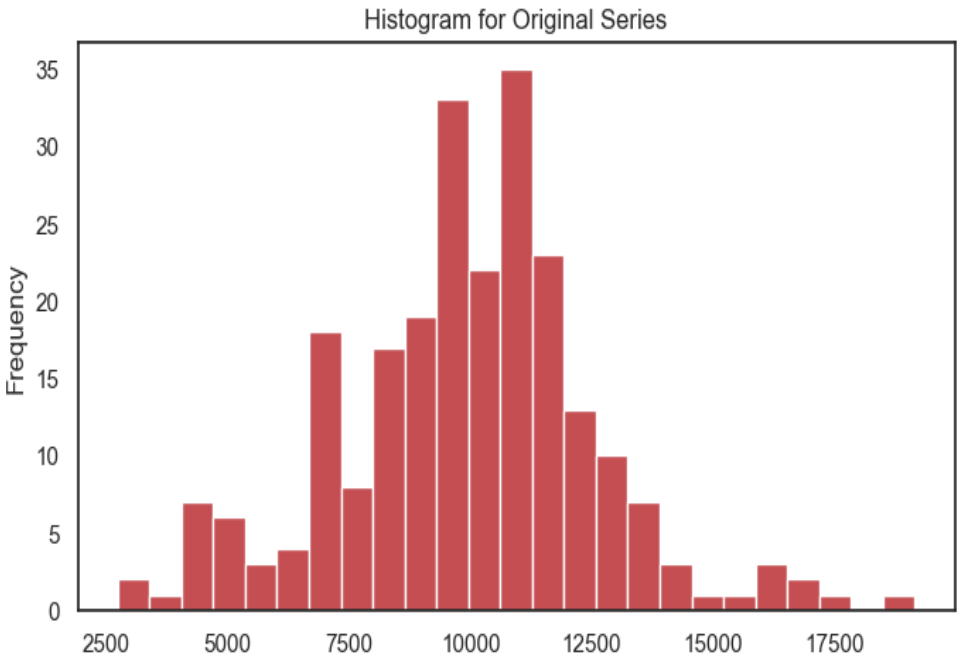

Figure:3.3-Histogram

For above (figure 3.3) "TH" variable the histogram displayed above, the frequency o r count of different values or value ranges of "TH" on the x-axis, and the corresponding frequ ency or count on the y-axis.

```
Running Augmented Dickey-Fuller test with paramters:
    maxlag: 31 regression: c autolag: BIC
Results of Augmented Dickey-Fuller Test:
+------------------------------+------------------------------+
|                              | Dickey-Fuller Augmented Test |
+------------------------------+------------------------------+
|        Test Statistic        |     -2.9707497269184038      |
|           p-value            |     0.037717776216177876     |
|          #Lags Used          |             6.0              |
| Number of Observations Used  |            233.0             |
|     Critical Value (1%)      |      -3.458731141928624      |
|     Critical Value (5%)      |     -2.8740258764297293      |
|     Critical Value (10%)     |     -2.5734243167124093      |
+------------------------------+------------------------------+
```

Based on the results of the ADF test, the test statistic (-2.9707497269184038) is more negative than the critical values at all commonly used significance levels. Additionally, the p-value (0.037717776216177876) is less than 0.05. Therefore, there is sufficient evidence to reject the null hypothesis of non-stationarity.

This series is stationary

There is some differencing needed in this datasets for stat models

Model Cross Validation Results:

```
Model Cross Validation Results:
----------------------------------------
    MAE (Mean Absolute Error = 2281.18
    MSE (Mean Squared Error = 8780477.52
    MAPE (Mean Absolute Percent Error) = 25%
    RMSE (Root Mean Squared Error) = 2963.1871
    Normalized RMSE (MinMax) = 21%
    Normalized RMSE (as Std Dev of Actuals)= 101%

    Finding the best parameters using AutoArima:
Performing stepwise search to minimize aic
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=4656.904, Time=0.02 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=4601.674, Time=0.05 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=4523.874, Time=0.28 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=4654.910, Time=0.03 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=4524.318, Time=0.42 sec
 ARIMA(0,1,2)(0,0,0)[0] intercept   : AIC=4524.350, Time=0.50 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=inf, Time=0.50 sec
 ARIMA(0,1,1)(0,0,0)[0]             : AIC=4522.907, Time=0.14 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=4523.635, Time=0.31 sec
 ARIMA(0,1,2)(0,0,0)[0]             : AIC=4523.638, Time=0.25 sec
 ARIMA(1,1,0)(0,0,0)[0]             : AIC=4599.685, Time=0.04 sec
 ARIMA(1,1,2)(0,0,0)[0]             : AIC=inf, Time=0.37 sec

Best model:  ARIMA(0,1,1)(0,0,0)[0]
Total fit time: 2.914 seconds

Best model is a Seasonal SARIMAX(0,1,1)*(0,0,0,12), aic = 4522.907
Refitting data with previously found best parameters
    Best aic metric = 4489.4
```

Interpret:

The cross-validation results you provided indicate the performance of a model on a dataset. Here are the metrics obtained:

Mean Absolute Error (MAE): 2281.18

Mean Squared Error (MSE): 8780477.52

Mean Absolute Percent Error (MAPE): 25%

Root Mean Squared Error (RMSE): 2963.1871

Normalized RMSE (MinMax): 21%

Normalized RMSE (as Std Dev of Actuals): 101%

These metrics are commonly used to evaluate the accuracy and performance of regression models.

The next part of the information suggests that an AutoArima algorithm was used to find the best parameters for the model. The algorithm performs a stepwise search to minimize the Akaike Information Criterion (AIC), which is a measure of the model's goodness of fit. Based on the search, the best model is identified as ARIMA(0,1,1)(0,0,0)[0].

The best model is a Seasonal SARIMAX(0,1,1)*(0,0,0,12), where (0,1,1) refers to the non-seasonal ARIMA parameters and (0,0,0,12) represents the seasonal ARIMA parameters with a seasonal period of 12 (assuming monthly data). The AIC value for this model is 4522.907.

Lastly, the model is refitted with the previously found best parameters, and the best AIC metric achieved is reported as 4489.4.

```
   SARIMAX Results
==========================================================================
=======
Dep. Variable:                        TH   No. Observations:
240
Model:                SARIMAX(0, 1, 1)   Log Likelihood                -2
196.035
Date:                 Mon, 15 May 2023   AIC                            4
400.069
Time:                         10:19:31   BIC                            4
413.942
Sample:                              0   HQIC                           4
405.661
                                 - 240
Covariance Type:                   opg
==========================================================================
=======
                 coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------
-------
intercept      6.0694     27.329      0.222      0.824     -47.495
59.634
drift         -0.0219      0.191     -0.115      0.909      -0.396
0.352
ma.L1         -0.9463      0.024    -38.722      0.000      -0.994
-0.898
sigma2      7.613e+06      0.000    3.1e+10      0.000    7.61e+06
==========================================================================
============
```

Ljung-Box (L1) (Q):	0.08	Jarque-Bera (JB):	3.26
Prob(Q):	0.77	Prob(JB):	0.20
Heteroskedasticity (H):	1.64	Skew:	-0.21
Prob(H) (two-sided):	0.03	Kurtosis:	3.40

The output of a SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exog enous Variables) model is displayed in above table. This type of model is commonly used for time series analysis and forecasting, incorporating both autoregressive and moving average components along with seasonal adjustments and optional exogenous variables.

Dependent Variable: This indicates the dependent variable in your model, labeled as "TH."

Number of Observations: The number of observations used in the model is 240.

Model: The specified model is SARIMAX(0, 1, 1), indicating the order of the autoregressive (AR), differencing (I), and moving average (MA) components.

Log Likelihood, AIC, BIC, HQIC: These are statistical measures used for model selection bas ed on their goodness-of-fit and complexity. Lower values of Log Likelihood, AIC, BIC, and H QIC indicate a better-fitting model. In this case, the Log Likelihood is -2196.035, the AIC is 4 400.069, the BIC is 4413.942, and the HQIC is 4405.661.

Intercept: The estimated coefficient for the intercept term is 6.0694. However, its p-value (0. 824) suggests that it is not statistically significant at conventional significance levels (typicall y 0.05). The confidence interval [0.025, 0.975] indicates the range of plausible values for the c oefficient.

Drift: The drift coefficient measures a linear trend in the model. In your case, the estimated co efficient is -0.0219, but it is also not statistically significant (p-value: 0.909). Again, the confi dence interval [0.025, 0.975] provides a range of plausible values.

ma.L1: This refers to the coefficient of the moving average lag 1 term in the model. The estim ated coefficient is -0.9463, and it is statistically significant (p-value: 0.000). It suggests a stron g negative relationship between the dependent variable and its lag 1 moving average term.

sigma2: This represents the estimated variance of the error term in the model. In your case, it i s 7.613e+06 (7.613 million).

Ljung-Box (Q): This test assesses the residuals' autocorrelation up to a specified lag. In this ca se, the Ljung-Box statistic (L1) is 0.08, and the associated p-value is 0.77, indicating no signif icant autocorrelation in the residuals.
Jarque-Bera (JB): This test evaluates the residuals' skewness and kurtosis to check if they foll ow a normal distribution. The Jarque-Bera statistic is 3.26, and its associated p-value is 0.20. A higher p-value (>0.05) suggests that the residuals are approximately normally distributed.

Heteroskedasticity (H): This test examines whether the residuals have constant variance. The Heteroskedasticity statistic is 1.64, and its associated p-value is 0.03, indicating some evidenc e of heteroskedasticity.

Thus, SARIMAX model with the specified parameters (0, 1, 1) was applied to the time series data. The intercept and drift coefficients were not statistically significant. The moving averag e lag 1 coefficient was significant, indicating a negative relationship. The model diagnostics s

howed no significant autocorrelation, approximate normality of residuals, and some evidence of heteroskedasticity.

Best Model is: SARIMAX

Best Model (Mean CV) Score: 2937.17

The model's performance is measured by a mean cross-validation (CV) score of 2937.17. The lower the CV score, the better the model's predictive performance.

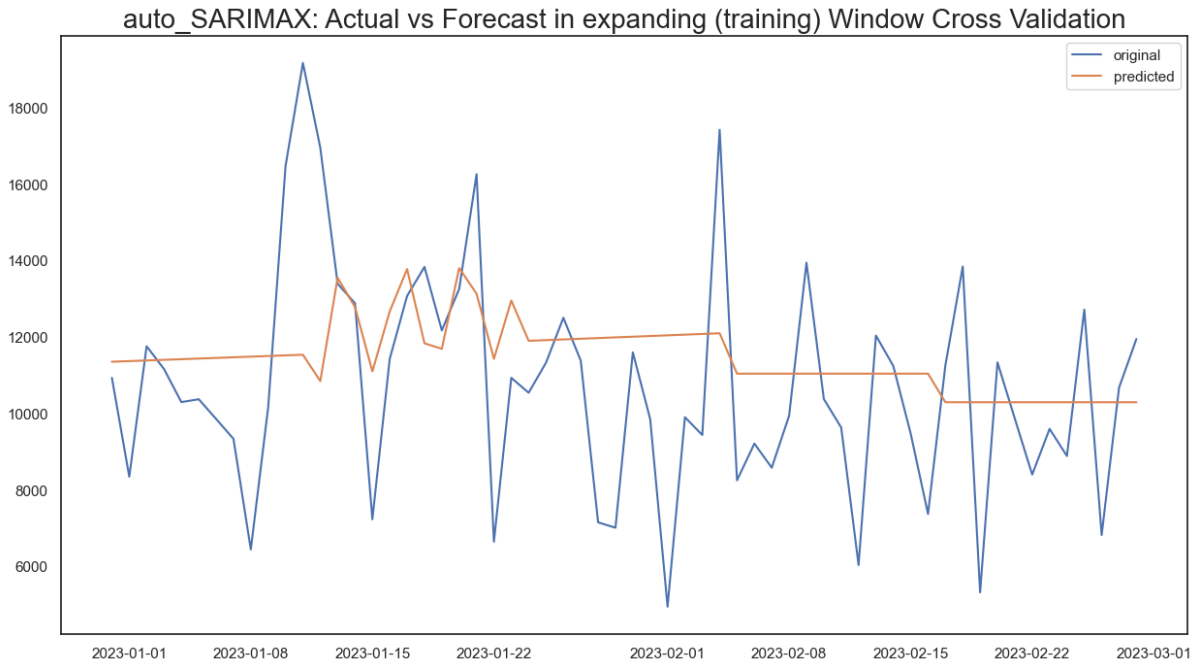

Figure:3.4 Sarimax

The graph showing the TH predicted and actual values using the SARIMAX model, the x-axis represents the data time periods, while the y-axis represents the values of the TH variable (the target variable) for each data point.

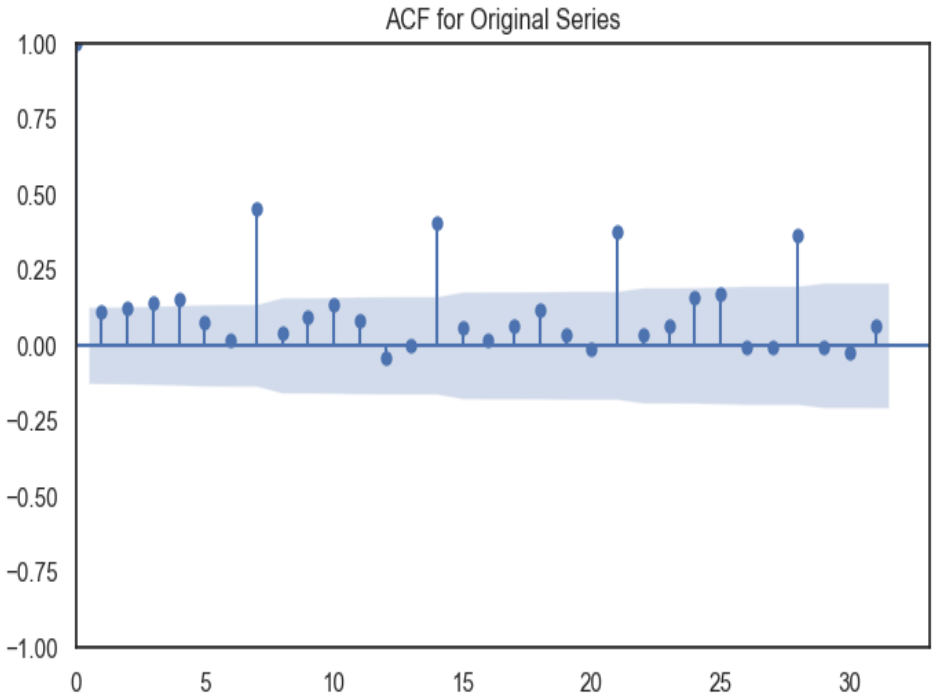

Figure:3.5 – ACF Plot

An ACF plot shows the correlation between a time series and its lagged values. The height of the bars indicates the strength of the correlation, while the significance bounds determine statistical significance. Positive values indicate positive correlation, negative values indicate negative correlation. Lags represent time intervals between observations. Patterns in the ACF plot can reveal recurring patterns, seasonality, or decay of correlation over time.

In the ACF plot, there are four bars that exceed the threshold of 0.25. This suggests a relatively strong positive correlation between the time series and its lagged values at those specific lag intervals.

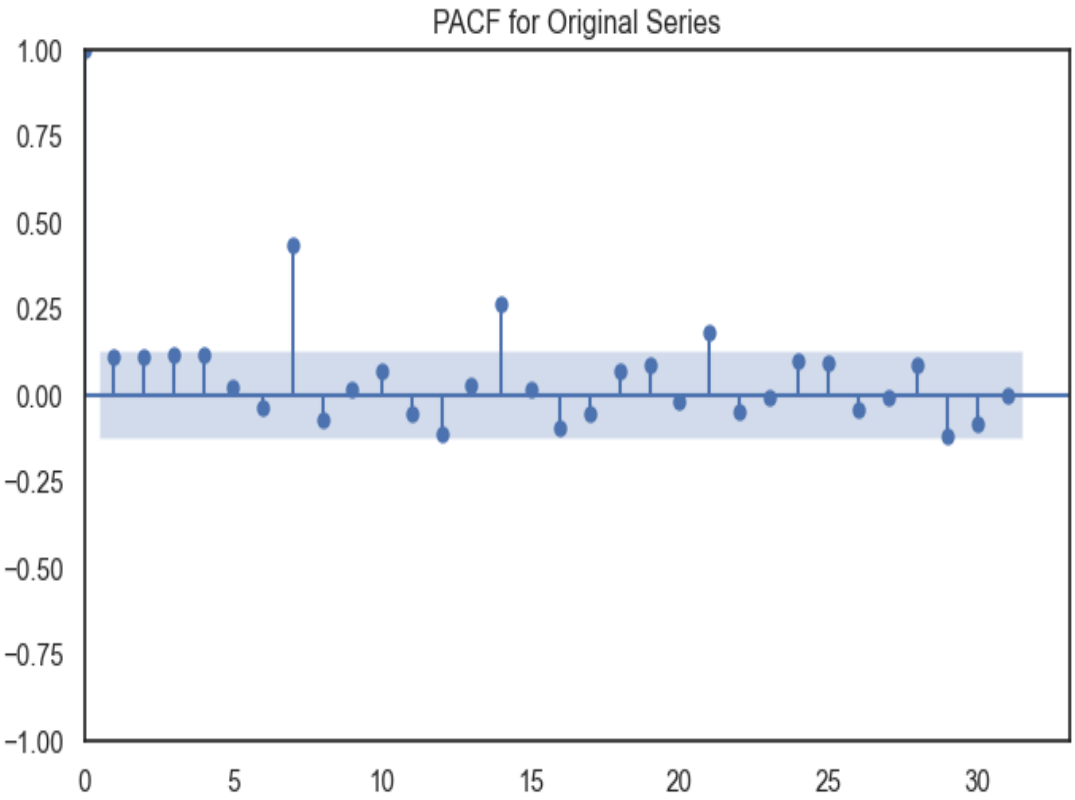

Figure:3.6 – PACF Plot

The Partial Autocorrelation Function (PACF) plot is used to analyze the correlation between a time series and its lagged values while controlling for the effects of intermediate lags. Interpretation of a PACF plot involves examining the correlation coefficients displayed on the y-axis for different lag values on the x-axis.

In the case where two bars in the PACF plot are above the threshold of 0.25, it suggests a relatively strong partial correlation at those specific lag intervals. The PACF measures the direct relationship between the current observation and a particular lag, while removing the influence of shorter lags.

XGBoost

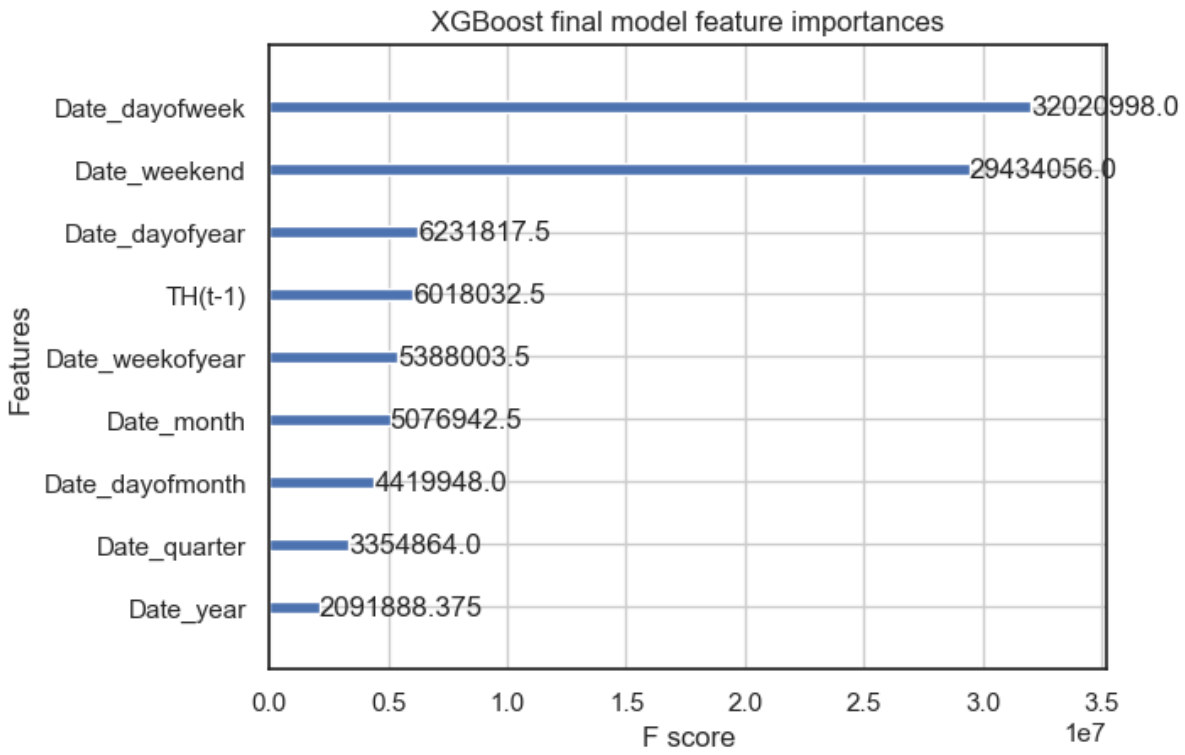

Figure:3.7 - XGBOOST

The above the (figure:3.7) graph of XGBoost model feature importance which provides insights into the relative importance of different features in predicting the target variable. The values shown on the y-axis represent the F scores, which are a measure of feature importance in XGBoost models.

Date_dayofweek: This feature has the highest importance with an F score of 32020998.0, indicating that it significantly contributes to the model's predictions.

Date_weekend: The feature of whether a date falls on a weekend is the second most important, with an F score of 29434056.0.

Date_dayofyear: This feature has an F score of 6231817.5, suggesting it is relatively important but not as influential as the first two features.

TH(1-1): This feature, labeled as "TH(1-1)," has an F score of 6018032.5, indicating its significance in predicting the target variable.

Date_weekofyear, Date_month, and Date_dayofmonth: These features have F scores of 5388003.5, 5076942.5, and 4419948.0, respectively, showing their importance but to a slightly lesser extent compared to the previous features.

Date quarter: This feature has an F score of 2091888.375, suggesting it has a relatively lower importance compared to other features.

The feature importance graph provides a ranking of the features based on their contribution to the XGBoost model's predictions. The higher the F score, the more influential the feature is in determining the target variable. This information can be useful in understanding the relative importance of different variables and potentially identifying key drivers or factors affecting the target variable

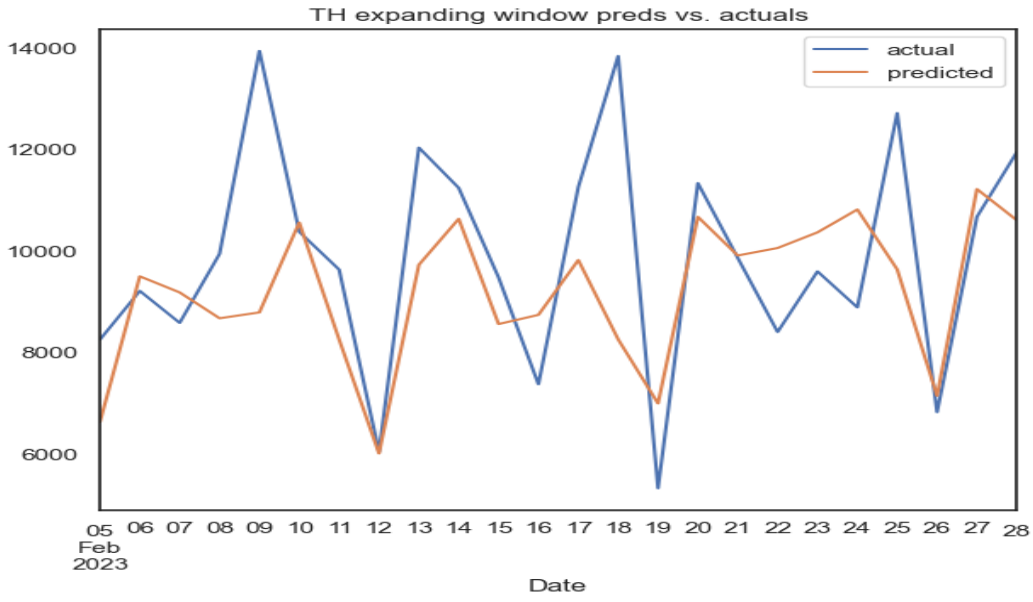

Figure:3.8

From the(figure:3.8) graph showing the TH predicted and actual values using the XGBoost model, the x-axis represents the data time periods, while the y-axis represents the values of the TH variable (the target variable) for each data point.

```
name        rmse
0   ML  2008.97981
```

A lower RMSE value indicates better model performance, as it represents a smaller average difference between the predicted and actual values. In this case, the XGBoost model has an RMSE of 2008.97981, which indicates the average prediction error is approximately 2008.97981 units.

CHAPTER- IV

Conclusion

In this work, the aim is to identify the best models for time series and machine learning tasks. After conducting the analysis, determined that SARIMAX performed the best for the time series, while XGBoost emerged as the top model for the machine learning task.

For the time series analysis, SARIMAX achieved a Mean CV Score of 2875.24. This indicates that SARIMAX demonstrated strong predictive capabilities and performed well in terms of cross-validation accuracy. SARIMAX is a widely used model for time series forecasting, and its superior performance in this study suggests its effectiveness in capturing and predicting the patterns and dynamics within the time series data.

On the other hand, for the machine learning, XGBoost emerged as the best model with an RMSE (Root Mean Squared Error) value of 2008.97981. The lower the RMSE, the better the model's predictive accuracy. XGBoost, an ensemble learning algorithm, has gained popularity for its ability to handle complex relationships and deliver high performance in various machine learning applications.

These findings suggest that SARIMAX is a reliable choice for time series analysis, while XGBoost is a strong contender for machine learning tasks.

Bibliography

1. Jason Brownlee (Year: 2019). XGBoost with Python: Discover Gradient Boosting with XGBoost.
2. Rob J. Hyndman and George Athanasopoulos (2018). Forecasting: Principles and Practice.
3. Wes McKinney (Year: 2017). Python for Data Analysis
4. Rami Krispin (Year: 2020). Hands-On Time Series Analysis with Python
5. Peter J. Brockwell and Richard A. Davis (2016). Introduction to Time Series and Forecasting.

APPENDIX

```
pip install auto_ts

from auto_ts import auto_timeseries

import pandas as pd

data = pd.read_csv('/users/velkumar/Desktop/project/II major  project/data.csv',
parse_dates=['Date'], index_col='Date')

data

target_col = 'TH'

time_col = 'Date'

model = auto_timeseries(

    score_type='rmse', forecast_period=12, time_interval='M', non_seasonal_pdq=None,

    seasonal_period=12, perform_feature_selection=True, model_with_lowest_error=True,
verbose=2

)

model.fit(

    traindata=data,

    ts_column=time_col,

    target=target_col,

    cv=5,

    sep=';'

)

forecast_df_folds = None

model = auto_timeseries(

    score_type='rmse',

    forecast_length=12,

    model_type=[ 'ml'],

    model_time_series=['XGBoost', 'Random Forest', 'LSTM'],

    verbose=2,
```

```
    lookback=3,
    time_interval='W',
    seasonal_period=4
)
model.fit(
    traindata=data,
    ts_column=time_col,
    target=target_col,
    cv=5,
    sep=';'
)
leaderboard = model.get_leaderboard()
print(leaderboard)
```