ENHANCING TRAFFIC MANAGEMENT THROUGH ENSEMBLE FORECASTING OF TRAFFIC FLOW PATTERNS

SONAM NETALKAR

Dissertation Report

December 2023

DEDICATION

Dedicated to my parents, who gave me the ability to see dreams clearly and who enabled me to realize my own aspirations. They inspire me at every turn and help me realize that I am capable of conducting research. My parents have always supported me financially, emotionally, and spiritually; they have been my inspiration and source of strength.

ACKNOWLEDGEMENTS

Exactly how I want to express my gratitude to Dr. Dipanker Dutta, my thesis advisor, for all of his encouragement and support is beyond me. An excellent source of motivation and inspiration is always a teacher. I am so grateful that you have been my guiding light. I finished this assignment because of your advice and direction. I express my heartfelt gratitude for my university's educational chances.

My buddies have been a huge help in getting this assignment finished; I appreciate you letting me use this time to write and conduct research. I am so very grateful to you. I also like to thank Mr. and Mrs. Netalkar, my parents. The numerous times you sympathized with me and assisted me in getting

And lastly, to my supportive, loving, and caring siblings. We really appreciate and have taken notice of your support throughout difficult times. Knowing that you were willing to take care of our household chores while I finished my work was a huge comfort and relief. Sincere gratitude from me. Many thanks for your unwavering and ongoing support.

Abstract :

This research proposal aims to enhance urban traffic management through the implementation of advanced time series forecasting techniques for accurate traffic flow prediction. The study focuses on harnessing the predictive capabilities of the Caltrans Performance Measurement System (PeMS) dataset, which includes crucial traffic variables like volume, speed, and occupancy. The primary objective is to develop robust forecasting models, including AutoRegressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), and ensemble methods, to effectively capture intricate traffic dynamics and predict forthcoming traffic flow patterns. Through comprehensive training, validation, and rigorous testing using the PeMS dataset, these models aim to provide reliable insights into real-world traffic scenarios. The anticipated outcomes of this research hold significant potential for optimizing traffic control strategies, supporting informed urban development decisions, and ultimately addressing the pressing challenges posed by urban congestion. By contributing to the advancement of intelligent traffic management systems, this research aspires to promote more efficient and sustainable urban mobility solutions. The results of this study can have far-reaching implications for enhancing traffic management strategies, mitigating congestion-related issues, and paving the way for a smarter and more responsive urban transportation ecosystem.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

ARIMA	Autoregressive integrated moving average
SARIMA	Seasonal autoregressive integrated moving average
LSTM	Long short term memory
XGB	Extreme gradiant boost
MAE	Mean absolute error
RSME	Root square mean error
MAPE	Mean absolute percentage error
PEMS	Performance Management System
EDA	Exploratory Data Analysis

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The efficient management of traffic flow within urban areas is a paramount concern for the twenty-first century. Urbanization, population growth, and economic development have led to a substantial increase in vehicular traffic, straining the existing infrastructure and giving rise to issues like congestion, prolonged commute times, and their associated environmental consequences. Addressing these challenges requires innovative and adaptive strategies that can optimize traffic management and enhance the quality of urban life.

Traffic management systems rely heavily on the ability to accurately predict traffic flow patterns. Accurate predictions are indispensable for designing and implementing strategies aimed at reducing congestion, optimizing signal control, improving transportation systems, and ultimately alleviating the challenges posed by urban traffic. However, the sheer complexity of traffic systems, the influence of various external factors, and the occurrence of unforeseen events have made traffic flow prediction a daunting task.

Traffic flow prediction traditionally relies on statistical techniques such as Autoregressive Integrated Moving Average (ARIMA) and machine learning models like Long Short-Term Memory (LSTM) networks and XGBoost. While these models offer valuable insights, they often fall short in providing consistently accurate predictions, especially when faced with the intricacies of urban traffic patterns.

In real-world scenarios, the majority of traffic flow forecast relies on the ARIMA model, according to sources.In 1976, Box and Jenkins devised a time series model similar to the ARIMA model.Mohamed et al. used the theory to analyze traffic flow on roads in 1979.Time series like ARIMA have been used extensively ever since for a considerable amount of time. Another intelligent prediction method is LSTM.In order to solve the disappearing and exploding gradient issues with traditional recurrent meural networks (RNNs), Hochreiter and Schmidhuber built it.Successful applications of LSTM and RNN include phonetic labeling of audio frames, language modeling, and handwriting recognition.As a unique type of RNN, the LSTM model can also apply the properties it learns from data to the desired scene.For example, if traffic flow data provided to lstm model, Lstm model can judge trffic flow data in next moment, it can also keep time connection information.

This thesis introduces a novel approach to address the shortcomings of conventional traffic flow prediction models. We propose an ensemble forecasting methodology that combines the strengths of multiple prediction models to create a comprehensive, adaptable, and robust system for traffic management. Our ensemble model integrates diverse forecasting algorithms, each designed to capture distinct aspects of traffic behavior. By doing so, we aim to harness the collective power of these models to significantly enhance prediction accuracy.

In essence, this research aims to tackle the complexity and dynamism of traffic flow patterns by merging the capabilities of various forecasting models into a unified system. The key objectives of our study include the development of efficient model integration techniques, real-time adaptation strategies, comprehensive evaluation methods, scalability considerations, and practical implementation guidelines. Our research strives to provide traffic management authorities with a powerful tool that enhances traffic prediction accuracy, reduces congestion, and ultimately leads to more effective urban traffic management.

In the following chapters, we will delve into the intricate details of our proposed ensemble forecasting methodology, exploring its theoretical foundations, technical aspects, and the results of empirical evaluations. We anticipate that this research will not only contribute to the field of traffic management but will also lead to substantial improvements in the efficiency and sustainability of urban transportation systems.

1.1.Background:

Urbanization and population growth have led to an unprecedented rise in vehicular traffic, posing significant challenges to traffic management in metropolitan areas. Traffic congestion not only results in time inefficiencies for commuters but also contributes to environmental pollution and increased fuel consumption. In this context, traffic flow prediction has emerged as a pivotal component of modern transportation systems, offering a potential solution to alleviate congestion and enhance traffic management. Traditional traffic management systems struggle to accommodate these dynamics due to their reactive nature. However, the advent of data-driven approaches, coupled with the availability of high-resolution traffic data, presents a promising solution. The Caltrans PeMS dataset(Link: Caltrans PeMS Dataset) offers a rich source of real-time traffic data collected from thousands of sensors across California's road networks. This dataset captures intricate traffic flow patterns, making it an ideal resource for developing accurate prediction models.This research focuses on addressing these challenges by introducing the concept of ensemble forecasting. Ensemble forecasting is a strategy that leverages the strengths of multiple forecasting models by combining their predictions. By incorporating a range of models, both traditional and modern, into the ensemble, we aim to create a robust and accurate forecasting system capable of adapting to diverse traffic scenarios.The Performance Measurement System (PeMS) dataset serves as the empirical basis for this research. PeMS provides access to a wealth of real-time traffic data, including traffic speeds, flow rates, and occupancy. This dataset is instrumental in training and validating the ensemble forecasting models, as it reflects real-world traffic conditions and offers a solid foundation for model development.

Time series forecasting, a subset of data analytics, focuses on predicting future values based on historical data patterns. Applied to traffic flow prediction, it involves modeling historical traffic data to forecast future traffic patterns. Techniques such as Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and machine learning-based models like Long Short-Term Memory (LSTM) networks have shown potential in accurately predicting traffic flow. Various evaluation metrics, such as root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), are employed to assess the performance of ensemble models against individual forecasting methods.

The results of this research hold promise for enhancing traffic management and urban sustainability. The ensemble forecasting approach not only improves the accuracy of traffic

flow predictions but also offers adaptability to the dynamic nature of urban traffic. This research contributes to the development of intelligent transportation systems, urban planning, and environmentally conscious traffic management strategies

1.2.Problem Statement:

Traffic management in urban areas is a critical challenge that requires efficient, data-driven solutions. With the availability of extensive traffic data sources like the California Department of Transportation's (Caltrans) PeMS (Performance Measurement System) dataset, there is an opportunity to enhance traffic management through advanced forecasting techniques. Despite the rich data available, there remains a significant gap in the optimization of traffic flow prediction. This thesis addresses the need for effective traffic flow prediction using ensemble forecasting of traffic flow patterns based on the Caltrans PeMS dataset.

The core problem lies in the complexity of urban traffic systems, which are influenced by a multitude of variables such as time of day, day of the week, weather conditions, and special events. Existing traffic management approaches have shown limitations in capturing these intricate patterns accurately, leading to suboptimal traffic flow predictions. Thus, there is a pressing need for a robust and versatile system that can leverage the wealth of data within the Caltrans PeMS dataset to enhance traffic management.

The primary problem this thesis aims to solve is to develop an ensemble forecasting framework that effectively predicts traffic flow patterns. By combining the predictive power of multiple forecasting models, this research seeks to increase the accuracy of traffic flow predictions. It addresses the challenge of optimizing traffic management and aims to provide actionable insights for authorities, urban planners, and commuters.

Furthermore, as an increasing number of cities across the world face traffic congestion and the associated economic and environmental consequences, the successful resolution of this problem has significant implications. It is essential to design a versatile and adaptive forecasting system capable of addressing the unique traffic characteristics of diverse urban environments.

In summary, this research addresses the problem of suboptimal traffic flow predictions by leveraging the Caltrans PeMS dataset and proposes an ensemble forecasting approach to

enhance traffic management in urban areas. The goal is to create a more accurate, reliable, and adaptable forecasting system, and the findings will have broader implications for traffic optimization and urban planning

1.3.Aim & Objectives:

Aim:

The aim of this research is to develop an accurate and adaptable traffic flow prediction model using time series forecasting techniques applied to the California Performance Measurement System (PeMS) dataset. The research seeks to enhance urban mobility and alleviate congestion by providing transportation authorities with reliable predictions for informed traffic management decisions.

Objectives:

- **Develop Ensemble Forecasting Model:** Create an ensemble forecasting model that combines the predictive power of ARIMA, SARIMA, and LSTM models to enhance short-term traffic flow prediction accuracy. This objective will address the core of the study by employing a diverse range of models for improved results.
- **Evaluate Model Performance:** Quantitatively assess the ensemble model's performance using established metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Compare the ensemble's predictive accuracy against individual models to ascertain its efficacy. This objective will provide insights into the model's real-world applicability.
- **Visualize and Communicate Predictions:** Develop intuitive visualizations that illustrate the ensemble model's traffic flow predictions. These visualizations will facilitate effective communication of traffic patterns to relevant stakeholders, aiding informed decision-making for traffic management. This objective emphasizes practical utility and communication of results.

1.4.Significance of Study:

- **Importance of the Work:** The research on traffic flow prediction using time series forecasting and the California Performance Measurement System (PeMS) dataset holds paramount importance in addressing pressing urban mobility challenges. By harnessing data-driven forecasting models, this work seeks to revolutionize traffic management strategies, alleviate congestion, and enhance overall urban quality of life.

- **Expected Outcomes:** The anticipated outcomes of this research are multi-faceted and impactful:

1. **Accurate Traffic Flow Prediction:** The research aims to develop models that accurately forecast traffic flow patterns, enabling transportation authorities to proactively address congestion and optimize traffic management strategies.
2. **Real-Time Adaptability:** The models' integration with real-time PeMS data ensures dynamic adaptation to changing traffic conditions, providing timely and relevant predictions.
3. **Enhanced Urban Mobility:** The research outcomes have the potential to significantly reduce travel times, decrease fuel consumption, and minimize environmental impact, resulting in improved urban mobility.
4. **Optimized Resource Allocation:** Precise traffic predictions allow for optimal allocation of resources, including traffic signal timing, lane management, and road maintenance.
5. **Informed Decision Making:** Transportation agencies will have access to data-driven insights for informed decision-making, leading to more effective policies and strategies.

- **National & International Implications:** The impact of this research extends beyond geographical boundaries:

1. **National Traffic Management:** The research outcomes can shape national traffic management policies, aiding in creating smoother, more efficient road networks.
2. **Economic Growth:** Enhanced traffic management leads to seamless goods movement, attracting investment and fostering economic growth.
3. **Global Sustainability:** Reduction in congestion and emissions contributes to global sustainability goals, aligning with environmental agendas.
4. **Smart City Initiatives:** The findings align with smart city initiatives worldwide, providing intelligent solutions for urban transportation challenges.

5. **Academic Contributions:** The research contributes to academia by advancing the field of traffic flow prediction and time series forecasting, enriching scholarly discussions.
6. **International Collaboration:** This research paves the way for international collaboration in addressing shared urban mobility challenges, promoting knowledge exchange.

In essence, this research has the potential to transform the way urban traffic is managed and optimized, yielding benefits at local, national, and global scales. By aligning advanced data analytics with real-world urban challenges, this work underscores the power of research to drive positive change in our increasingly interconnected world

1.5.Scope of Study:

Scope: The scope of the research on traffic flow prediction by time series forecasting using the California Performance Measurement System (PeMS) dataset defines the boundaries and focus of the study. It outlines what aspects of the research will be addressed and what will be excluded. Defining the scope is essential to ensure that the research remains manageable, feasible, and aligned with the research objectives. Here's a breakdown of the scope, out of scope elements, and reasons for defining the scope:

In Scope:
1. Data Collection and Preprocessing:
• Acquiring historical traffic flow data from the PeMS dataset.
• Preprocessing the data to handle missing values, outliers, and data quality issues.
2. Model Development and Evaluation:
• Developing and implementing time series forecasting models such as ARIMA, SARIMA, LSTM, and ensemble techniques.
• Evaluating model performance using appropriate evaluation metrics.
3. Real-Time Integration:
• Investigating methods to integrate real-time PeMS data into forecasting models.
• Exploring mechanisms for dynamic model adaptation to changing traffic conditions.
4. Case Studies and Application:
• Applying the developed models to real-world traffic scenarios within the PeMS dataset.
• Collaborating with transportation agencies to validate the models' effectiveness.
5. Results Analysis and Interpretation:

• Analyzing forecasting results to extract insights into traffic patterns and model performance.
• Interpreting findings to provide valuable insights for traffic management.
6. Documentation and Reporting:
• Compiling a research report detailing the methodology, results, analysis, and conclusions.
• Using visualizations and graphs to present findings effectively.

Table 1 : In Scope

Out of Scope:
1. Infrastructure Development:
• Developing new hardware or software infrastructure to host the forecasting models.
2. Sensor Deployment or Maintenance:
• Physical deployment or maintenance of traffic sensors on road networks.
3. Policy Implementation:
4. Hardware or Sensor Improvements:
• Enhancements to the hardware or sensors used for traffic data collection.
5. Urban Planning Decisions:
• Urban planning decisions beyond the scope of traffic flow prediction.

Table 2 : Out of Scope

1.6.Structure of Study:

The study's structure in this research is divided into several sections. The problem statement and study background are covered in Chapter 1.This chapter also explains the study's scope and significance, as well as its goals and objectives.

The literature review in Chapter 2 covers the study's history, methodologies, analysis, applications, and recent studies that are pertinent to resolving our business issue. Where the researchers' findings will be showcased to show how pertinent tactics might be understood and eliminated.

The second section's progression will serve as the basis for Chapter 3, where the most pertinent experiments and other findings will be finalized in preparation for implementation in Chapter 4.which is adhered to

CHAPTER 2

LITERATURE REVIEW

2.1.Introduction:.

The field of traffic management and the prediction of traffic flow patterns have been subjects of increasing significance in recent years due to the growing complexities of urban environments and the expanding volume of vehicular traffic. To address the various challenges posed by traffic management, accurate traffic flow prediction has emerged as an indispensable tool. It not only helps in alleviating congestion but also contributes to safer road networks, reduced emissions, and optimized transportation systems.This chapter embarks on a comprehensive review of the existing body of knowledge in the domain of traffic flow prediction. It seeks to provide a solid foundation for the research presented in this thesis by delving into the historical developments, core concepts, and diverse prediction approaches that have shaped this field. In the modern era of intelligent transportation systems, data-driven methodologies have become increasingly vital.The foundation of this study is built upon one such invaluable resource: the California Department of Transportation (Caltrans) Performance Measurement System (PeMS) dataset.

Urban traffic management remains a critical challenge in modern cities due to the increasing complexities of transportation systems [1]. Accurate traffic flow prediction is crucial for devising effective strategies to manage congestion, reduce emissions, and enhance overall transportation efficiency [2]. In recent years, the utilization of ensemble forecasting models has gained prominence as a promising approach to improve the accuracy and reliability of traffic flow predictions [3].

Ensemble methods, which combine the predictions of multiple individual models, have shown potential in mitigating the limitations of individual models and enhancing prediction accuracy [4]. In the context of traffic flow prediction, ensembles offer a pathway to harness the complementary strengths of various models and mitigate their weaknesses [5]. Among the ensemble members, models like ARIMA, SARIMA, LSTM, and XGBoost stand out for their proven capabilities in time series prediction [6][7][8].

ARIMA and SARIMA models, known for capturing temporal patterns and seasonal trends, can be integrated within ensemble frameworks to offer robust predictions that account for both short-term fluctuations and long-term trends [9]. The LSTM model, a type of recurrent neural network, excels in handling the complex and nonlinear dynamics of traffic flow, making it a valuable addition to an ensemble for capturing intricate patterns [10].

Furthermore, XGBoost, an ensemble of decision trees, is particularly adept at capturing complex relationships within data [11]. Its application in the context of traffic flow prediction is notable for its capacity to handle intricate traffic patterns and fluctuations [12].

This research aims to leverage the combined strengths of these models through ensemble forecasting to yield accurate and reliable traffic flow predictions [13]. By achieving this, the study seeks to provide transportation authorities and policymakers with actionable insights to improve traffic management strategies and enhance urban mobility [14]. The amalgamation of these models is anticipated to surpass individual models' performance, ultimately contributing to advancements in traffic flow prediction methodologies [15].As we proceed through this literature review, we will navigate the existing research landscape, identifying the gaps and opportunities for further exploration. The insights garnered from this review will serve as a cornerstone for the research undertaken in subsequent chapters, which aim to contribute innovative solutions to the challenges of enhancing traffic management through ensemble forecasting of traffic flow patterns.

2.2.Data used :

The dataset central to your research is the California Department of Transportation (Caltrans) Performance Measurement System (PeMS) dataset (Caltrans PeMS Dataset). The Caltrans PeMS dataset is a critical resource for transportation and traffic management research. It offers extensive insights into traffic patterns, congestion, and the effects of various factors on traffic flow within California. This dataset comprises both real-time and historical traffic-related data collected from an extensive network of sensors, loops, and monitoring devices deployed across California's extensive roadways.

The Caltrans PeMS dataset includes diverse types of traffic-related information. This encompasses data on traffic flow, providing details about vehicle counts and speeds. It also includes data on road occupancy, revealing the portion of a road occupied by vehicles. This information is collected at high temporal resolution, often at intervals as frequent as a few minutes. This fine-grained temporal data enables researchers to perform in-depth analyses of

traffic patterns, congestion dynamics, and the impact of various variables, such as incidents or special events, on traffic flow.

This dataset holds significant importance for various stakeholders. Researchers frequently rely on the Caltrans PeMS dataset for studies related to traffic management, congestion alleviation, incident detection, and the development of predictive models for traffic flow patterns. Furthermore, traffic engineers and transportation agencies use the dataset to make informed decisions regarding traffic control strategies and infrastructure enhancements. In particular, the data plays a crucial role in both real-time traffic monitoring and the creation of models that predict traffic flow behavior under various conditions.

However, it's important to note that working with the Caltrans PeMS dataset can present challenges. Researchers often face issues related to data quality and preprocessing. These challenges may include handling data gaps, identifying and addressing outliers, and applying data cleaning and filtering procedures to ensure the data's suitability for analysis. Researchers typically access this dataset through collaboration with Caltrans or relevant authorities, following data sharing agreements and abiding by privacy and confidentiality regulations.

2.2.1 Data Types and Characteristics

The PEMS dataset encompasses a multitude of traffic-related parameters. These parameters are meticulously recorded at various spatial and temporal intervals, granting a comprehensive view of traffic dynamics. The core attributes of this dataset primarily include traffic flow information. These encompass vehicle volume data, offering insights into the number of vehicles passing specific points along the road network. Furthermore, data on vehicle occupancy within the coverage areas of sensors is included, along with vehicle speed metrics, portraying how vehicles traverse the roadways.

2.2.2 Ancillary Data

Beyond the fundamental traffic metrics, the PEMS dataset may include supplementary information that can be pivotal in enriching traffic flow predictions. This supplemental information incorporates data on traffic incidents, offering insights into accidents, road closures, and other events that can significantly disrupt traffic patterns. Moreover, it might incorporate environmental factors such as weather conditions that have a known influence on traffic flow, especially in regions like California, where diverse weather patterns are prevalent.

2.2.3 Data Processing Workflow

The data processing workflow is an integral component of extracting meaningful insights from the PEMS dataset. This process often involves data cleaning and normalization, as raw data may contain errors, missing values, or inconsistencies. Researchers may employ various techniques to handle such data imperfections while ensuring the quality and integrity of the dataset. After preprocessing, data can be structured into time series, enabling the development of forecasting models based on historical traffic patterns.

2.2.4 Dataset Availability

The availability of the PEMS dataset is a critical consideration for researchers. Typically, this dataset is accessible through governmental transportation agencies, research institutions, or specific online platforms dedicated to traffic data. Access to historical and real-time data allows researchers to analyze past traffic patterns and make real-time predictions, thereby contributing to the broader field of traffic management and optimization.

Attribute	Description
Timestamp	Date and time of data observation
Location	Geographic location or sensor identifier
Flow	Vehicle volume in vehicles per minute
Occupancy	Vehicle occupancy as a percentage
Speed	Vehicle speed in miles per hour
Incidents	Records of traffic incidents
Weather Conditions	Environmental data (optional)

Table 3: Structure of the PEMS Dataset

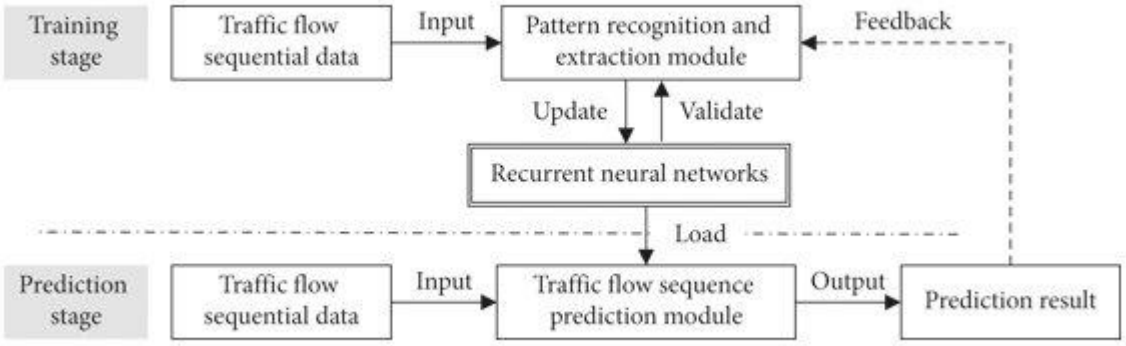


Fig1: Model implementation workflow

2.3.Modelling Techniques for Traffic flow Predictions

1. Traditional Statistical Methods:

ARIMA:

Any Time Series problem can be solved using a variety of conventional methods, including Autoregressors (AR), Moving Averages (MA), and Autoregressive Integrated Moving Averages (ARIMA), which combine or integrate the characteristics of MA and AR. (Maleki and others, undated) Autoregressive models are used to describe time-varying systems in the natural world, financial perspectives, and other domains. An artificial intelligence (AI) AR model builds upon a series of predetermined steps and recognizes evaluations from previous training sessions as inputs for a regression model to predict the value of each step.The moving average and autoregressive models are integrated in the ARIMA model. Both models are fitted to time series data in order to forecast future data points in the series or to increase the likelihood that the information will be understood.

First, The model determines whether the target time series is stationary first. If it is not, it will undergo differential treatment and be transformed into a stationary time series for modeling purposes. The time series shows a correlation between the continuous observations of traffic flow on roadways.Since the transportation system is complicated and stochastic, the traffic flow data may not be stationary. As a result, the ARIMA model is frequently used to forecast traffic flow using time series data.[24].

The auto-regressive integrated moving average (ARIMA) is expressed as ARIMA (p,d,q), where (AR) is the autoregressive and (I) is the integration. To estimate the model (d), the words moving average (MA), autoregressive term (p), number of autoregressive orders (p), order of moving average components (q), and order of differentiation applied to the series are defined [25, 26]. If the series is not stationary, it must become stationary before developing a forecasting model since the stationary assumption of this time series approach indicates that the series is free from periodic oscillations. In time series data, seasonality is indicated by any regular, periodic increase or decrease in the series mean. Seasonal volatility often follows hourly rates; daily, weekly, or annual repeats are typical.We estimate the traffic flow rate (x) at time t (xt) by combining daily traffic data at the time (tp) from prior months in a linear fashion. Hourly traffic flow data displays three main types of seasonality: daily, weekly, and annual trends.

The combination of the AR and MA processes yields ARIMA. $xt = \phi 1xt-1 + \phi 2xt-2 + ... + \phi pxt-p + e0at - \Theta 1yt-1 - \Theta 2yt-2 - ... - \Theta qyt-q$ (1)

where xt , xt−1, xt−2, ..., xt−p are d-order difference observations, ϕ1, ϕ2, ..., ϕp are coefficients of the d-order difference observations, yt , yt−1, yt−2, xy−p error values and Θ1, Θ2, ..., Θp are coefficients for errors

Box and Jenkins pioneered the development of a mathematical model designed for predicting future values of specific data by leveraging past observations of that data. The model they devised utilizes the autocorrelation and partial autocorrelation functions of the sample data as crucial tools for establishing the order of the ARIMA model [33]. In a related study [34], researchers delved into assessing the suitability of seasonal ARIMA for univariate traffic flow prediction.

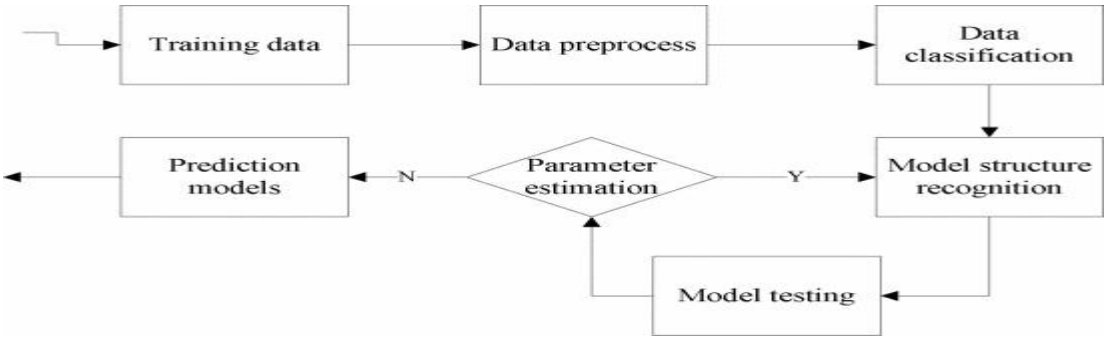


Fig 2 : Arima Model Training process

SARIMA

When there is little to no seasonal pattern in the time series, the ARIMA model can provide reasonable predictions; but, when there is a significant seasonal tendency, the seasonal ARIMA model (SARIMA) is required (16). Seasonal autoregressive, seasonal degree of difference, seasonal order of moving average, and seasonal period length are represented by the letters P, D, Q, and S, in that order. The meaning of these three parameters is the same as that of the ARIMA model. Before applying differential processing to eliminate and stabilize the seasonal period, we first observe the time series' stationarity and seasonal periodicity.

.A time series Xt = [1, 2, . . . , N] is generated by a SARIMA (p, d, q)(P, D, Q) given that: ϕ(B)Φ(B S)(1 − B) d (1 − B S) DXt = Θ(B)ϑ(B S)et (2)

Forecasting 2023, 5 621 where N is the number of observations, p, d, q, P, D, and Q are integers; B is the lag operator; s is the seasonal period length.

ϕ(B) = 1 − ϕ1B − ϕ2B 2 − ... − ϕpB p (3)

where the regular autoregressive operator (AR) of order p is denoted by ϕ(B) .This AR model uses a linear combination of historical values for the relevant variable, which in this case is traffic flow rate.

A = 1 − Φ1B − Φ2B 2S −... − ΦPB P (4)

Therefore, the autoregressive operator (AR) of order P is represented by Φ(B S), and the number of previous values to be taken into consideration for the prediction (MA) model of order q is indicated by the parameter p. Unlike using previous values of the relevant variable, the MA models incorporate past error values in a linear fashion. They are described by the parameter q, which determines the number of preceding error values to be included for the forecast.

ϑ(B S) = 1 − ϑ1B S − ϑ2B 2S − ... − ϑQB Q (5)

Hence, ϑ(B S) is the seasonal MA of order Q , D is the number of seasonal differences; et is the residual at time t, is both identically and independently distributed.

Facebook Prophet:

Its is a Time Series Forecasting (TSF) model. Facebook Prophet is a forecasting tool designed for time series data. It is specifically tailored for business forecasting tasks and is

considered a part of traditional time series forecasting methods. While it utilizes machine learning principles internally, it is not a deep learning model. Prophet is known for its simplicity, ease of use, and ability to handle various time series patterns.

The forecasting method known as Fb-Prophet is quick and entirely automated, with no human intervention. It performs well with time series because it can withstand missing data and trend shifts, which are normally well-handled abnormalities. In order to help the data fit into the optimal model, the Fb-PROPHET model provides the projected traffic flow in the form of trends as well as upper and lower bounds [12]. Nevertheless, Fb-Prophet is devoid of a user-accessible local perspective, which is essential for predicting the immediate future but difficult to extend.

2.Deep Learning Techniques:

LSTM:

Recurrent Neural Networks (RNNs), one of the most well-known and sophisticated deep learning techniques, extend this to solve any sequential problem. It has been demonstrated that the combination of deep learning techniques and abundant data is the most effective way to solve many AI-related problems, whether they are related to computer vision or natural language processing. There are two distinct variations of RNNs: Gated Recurrent Units (GRUs) and Long Short Term Memory (LSTMs).The LSTM network is a variant of the recurrent neural network (RNN), an artificial neural network (ANN) designed to address natural language processing (NLP) issues, and was first proposed by Hochreiter and Schmidhuber [27] in 1997. During training, the LSTM overcomes the long-term dependence of learning and resolves the vanishing gradient problem of the RNN. Time series analysis and language modeling have made extensive use of this network.

An output layer, a recursive hidden layer, and an input layer make up a typical LSTM network, as seen in Fig. 1. The output layer generates the predictions, and the input layer receives the data for training. Each neuron in the recursive hidden layer is composed of four structures: an input gate, an output gate, a forget gate, and a memory block.

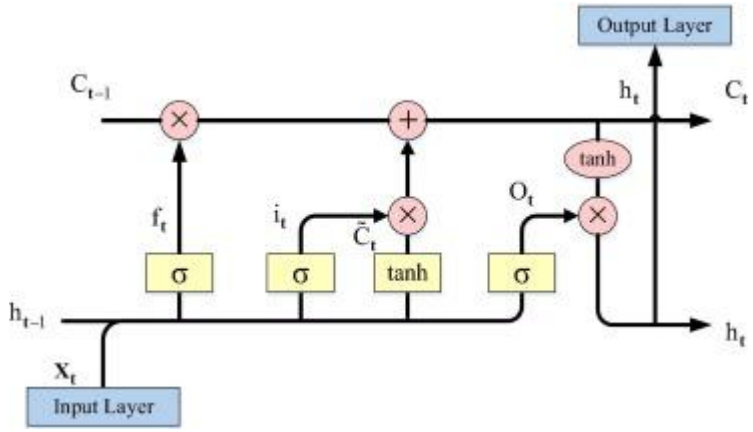


Fig. 2: Structure of a LSTM network.

The input layer in Fig. 1 receives the training data, while the output layer produces the LSTM network's prediction results. The forget gate, input gate, output gate, and memory cell are the four primary neuronal structures that make up the recursive hidden layer. The three gates can read, update, and remove features that are stored in the memory cell, while the cell's state reflects the features of the input.

X = (x1,x2,⋯ ,xt) represents the traffic flow time series; H = (h1,h2,⋯ ,ht) represents the hidden layer state of the original RNN; C = (c1,c2,⋯ ,ct) represents the memory cell state of the LSTM network; and y = (y1,y2,⋯ ,yt) represents the LSTM network target output series, with t representing the prediction period. Next, the state of the hidden layer and the target output of the LSTM network at time t can be calculated by:

ht=H(Wh)·[ht−1,xt]+bh(3)yt=Wyht+bh

where W are weight matrices (e.g. Wh is the weight matrix of the hidden layer); · is the dot product operation; b are vectors of bias; H is the hidden layer function. Equation(3) can be decomposed as follows:

it=σ(Wi·[ht−1,xt]+bi)

ft=σ(Wf· [ht−1,xt]+bf)

ot=σ(Wo· [ht−1,xt]+bo)

ct=ft* ct−1+it* tanh(WC· [ht−1,xt]+bC)(4)

ht=ot* tanh(ct)

where * is the Hadamard product operation between matrices; it, ft, and ot are the outputs of the input gate, forget gate, and output gate, respectively; tanh is the memory cell's activation function;Standard sigmoid function, denoted by σ:(5)σ(x)=11+e−x

In this work, the activation function of the memory cell is the sigmoid function projected to the interval [−1, 1]. By capturing both short- and long-term aspects of the time series, the memory cell can resolve the RNN's vanishing gradient issue.

2.4 Summary

The literature review unveiled a spectrum of models for traffic flow prediction. Traditional methods like ARIMA and SARIMA were explored for their historical significance. Machine learning models, notably ensemble methods, displayed improved predictive accuracy by blending diverse models. Deep learning, particularly LSTM networks, demonstrated prowess in capturing temporal dependencies for accurate forecasting. Facebook Prophet emerged as an accessible tool, with an additive modeling approach and automatic changepoint detection.

A notable trend in the literature is the inclination toward ensemble forecasting, combining models like Gradient Boosting, and Prophet. This reflects a concerted effort to enhance prediction robustness. As this research progresses, we aim to leverage and comprehend the unique contributions of these models in predicting traffic flow. The subsequent sections will delve into the research methodology, data collection, preprocessing, and the application of these models to the Caltrans PEMS dataset, contributing to the evolution of traffic management through ensemble forecasting.

CHAPTER 3

RESEARCH METHODOLOGY

3.1.Introduction:

Traffic flow prediction, using time series forecasting techniques, has emerged as a critical area of study to facilitate data-driven decision-making and enhance urban mobility. This research seeks to delve into the intricate dynamics of traffic flow, utilizing various time series forecasting models, and harnessing the wealth of data provided by the Caltrans PEMS dataset. The primary aim of this study is to develop a comprehensive understanding of traffic flow prediction by exploring a multitude of time series forecasting models. From classical statistical methods like ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal ARIMA), to cutting-edge deep learning approaches like Long Short-Term Memory (LSTM) networks, we aim to evaluate the performance and applicability of each model. Additionally, the study investigates the potential of ensemble forecasting, which combines the strengths of multiple models to improve predictive accuracy.

The research also involves a detailed exploration of data preprocessing techniques, transformation methods, and data augmentation strategies to ensure the quality and readiness of the dataset for predictive modeling. Furthermore, it emphasizes the significance of comprehensive exploratory data analysis (EDA) to gain insights into the underlying patterns and relationships within the traffic flow data.

The primary objective of this research is to contribute to the development of robust traffic flow prediction models, enhancing the accuracy of forecasting and enabling advanced traffic management strategies. By investigating a range of models and methodologies, this study aims to provide urban planners and transportation authorities with valuable tools for more efficient and data-driven traffic management, ultimately contributing to safer and more accessible urban environments.

3.2.Workflow:

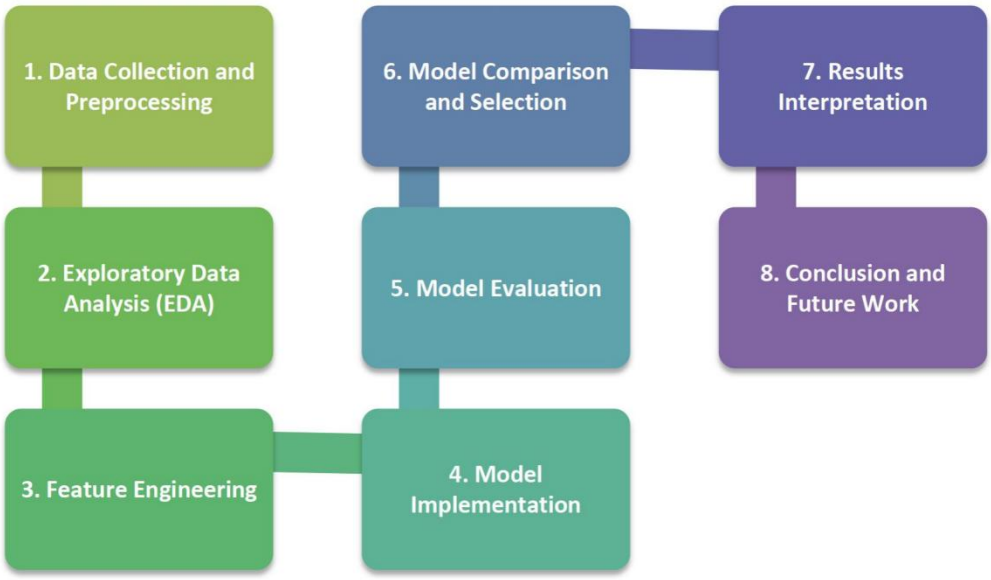


Fig 4 :Workflow

3.3..Data Collection and Understanding:

This data set is a high-speed data set in California, USA. The data set is collected in real time by more than 39,000 independent sensors at 5min intervals. The PeMS data set is huge, and several commonly used data subsets have been derived at present, including PeMS-03, PeMS-04, PeMS-07, PeMS-08, PeMS-BAY, etc.

Data is gathered from the PeMS (Performance Measurement System) of Caltrans. All of California's main metropolitan cities' motorway systems are home to roughly 40,000 separate detectors that collect data in real time.A useful tool for traffic monitoring and analysis is the California Department of Transportation (Caltrans) PeMS dataset. PeMS stands for Performance Measurement System, and the dataset provides a comprehensive collection of real-time traffic flow, speed, and occupancy data collected from various sensors deployed on roadways across California. The dataset covers a wide range of highways and road segments, making it a rich source of information for understanding traffic patterns, congestion, and overall transportation system performance.

Data Granularity: The dataset captures data at fine time intervals, often ranging from a few seconds to a minute. It collects data from each detector every 30 seconds and aggregates them into five-minute interval values by lane.This high temporal resolution allows researchers to

analyze traffic patterns in detail and observe variations over short time spans..PeMS is also provides over ten years of data for historical analysis.

Data Attributes: The dataset contains a variety of data attributes that provide insights into traffic conditions. Common attributes include:
Timestamp: Precise time at which the data point was collected.
Vehicle Count: The number of vehicles passing a specific point in a given time interval.
Speed: The average speed of vehicles during the time interval.
Occupancy: The proportion of time that a sensor is occupied by a vehicle.
Flow: The rate of vehicle passage over a specific point.
Lane Information: Data from individual lanes to analyze lane-specific traffic patterns.
Location Information: Geographical coordinates or mile markers to identify the sensor's placement.
Spatial Coverage: With sensors distributed across a wide geographic area, the dataset offers broad spatial coverage. This aspect is crucial for understanding traffic dynamics in different regions and for developing models that can generalize across diverse traffic scenarios.
Data Challenges: However, it's important to note that real-world datasets often come with challenges such as missing values, outliers, and noise. Preprocessing steps are crucial to address these issues and ensure the quality of data used for modeling.

Data Collection:

- **Sensor Network:** The Caltrans Performance Measurement System (PEMS) dataset draws its data from a vast sensor network deployed on California's roadways. This network consists of diverse sensor types, including inductive loop detectors, cameras, and GPS devices. These sensors continuously monitor and collect various traffic-related metrics.

- **Government Oversight:** The data within the Caltrans PEMS dataset is under the purview of governmental transportation agencies. This dataset is a product of California's commitment to efficient transportation management.

- **Multimodal Data:** The dataset encompasses data from multiple modes of transportation. It encompasses information related to highways, urban roads, tunnels, and toll booths. This

multimodal aspect makes the dataset highly versatile and valuable for a wide array of traffic management applications.

- **Real-time Monitoring:** The Caltrans PEMS dataset is characterized by real-time data collection. It features precise timestamping, allowing for temporal analysis and prediction.

3.4. Data Preprocessing:

Data preprocessing is a crucial step in preparing the PeMS dataset for traffic flow prediction using time series forecasting. Here's a comprehensive guide on data preprocessing steps you can follow:

1.Data Cleaning: Raw data from the Caltrans PEMS dataset is subject to data cleaning to ensure data quality. This process includes the identification and handling of missing data points, dealing with outliers, and rectifying any discrepancies. It's vital to ensure that the dataset is free from data imperfections.

2. Data Extraction:

- Extract the relevant data from the PeMS dataset, focusing on attributes like timestamp, traffic flow, and potentially speed and occupancy.

3. Handling Missing Data:

- Check for missing data points in the dataset.
- Decide how to handle missing values: interpolate, forward-fill, backward-fill, or drop the missing entries.

4. Outlier Detection and Treatment:

- Identify outliers that could impact the accuracy of your predictions.
- Consider using statistical methods or visualization tools to detect outliers.
- Decide whether to remove or adjust outlier values.

5. Resampling and Aggregation:

- Depending on the desired forecasting granularity (e.g., hourly, daily), resample or aggregate the data.
- Use methods like averaging or summing to consolidate data points within the chosen time intervals.

6. Time Series Alignment:

- Ensure that the data is aligned chronologically.
- Sort the dataset based on the timestamp.

7. Feature Engineering:

- Create additional features that could improve prediction accuracy, such as day of the week, time of day, holidays, and weather conditions (if available).

8. Normalization/Scaling:

- If you're using neural network-based models like LSTM or BLSTM, consider normalizing or scaling the traffic flow values to help the model converge faster.

9. Splitting into Training and Testing Sets:

- Divide the preprocessed data into training and testing sets. A common split might be 80% training and 20% testing.
- Ensure that the training data comes before the testing data in terms of time.

10. Data Visualization and Exploration:

- Create visualizations to understand the distribution of traffic flow, patterns, and potential seasonality in the data.
- This step helps in identifying any anomalies or trends.

11. Data Integrity Check:

- Ensure that the preprocessed data is consistent and accurate.
- Perform cross-checks to validate the correctness of your preprocessing steps.

3.5.Exploratory Data Analysis:

In this phase of the research, we embarked on an in-depth exploration of the Caltrans Performance Measurement System (PEMS) dataset to glean insights essential for our traffic flow prediction model. The dataset, comprising information about traffic flow, occupancy, and speed across various locations and time intervals, was subjected to a comprehensive EDA. Understanding the Dataset Structure: We initiated the EDA by scrutinizing the basic structure of the dataset. This involved checking for the presence of key features such as 'flow,' 'occupancy,' 'speed,' and 'location.' Descriptive statistics were computed to get an overview of the central tendencies and dispersions of these variables.

Temporal Patterns and Trends: Temporal analysis was conducted to uncover patterns and trends in traffic flow over time. Utilizing line plots and time series decomposition techniques, we identified recurring patterns, seasonality, and any long-term trends in the traffic data. This temporal understanding is pivotal for capturing the dynamics of urban traffic systems.

Spatial Analysis: Spatial aspects were explored to comprehend the geographical variations in traffic flow. Heatmaps and geospatial visualizations were employed to reveal high-density traffic zones, aiding in the identification of critical locations with consistent traffic congestion.
Correlation Analysis: To discern relationships among variables, correlation matrices were constructed. This facilitated the identification of factors influencing traffic flow, such as the impact of occupancy and speed on traffic patterns.
Handling Missing Values: During the EDA, attention was given to addressing missing values, ensuring that the dataset was cleansed for subsequent modeling. Imputation techniques were applied judiciously, considering the nature of the missing data.
Outlier Detection: Outliers, if any, were identified and addressed to prevent them from unduly influencing the predictive modeling. Box plots and statistical tests were employed to pinpoint potential anomalies in the traffic data.
The insights gleaned from this thorough EDA not only enriched our understanding of the dataset but also guided decisions regarding data preprocessing steps. It laid a robust foundation for the subsequent phases of model development, ensuring that our traffic flow prediction models are built on a sound and well-understood data structure.

3.6.Data Selection：

Dataset Requirements

The choice of an appropriate dataset is crucial for successful traffic flow prediction using diverse modelling techniques.

- **Time and Traffic Flow Columns:** A fundamental requirement is a time column representing temporal information, and a traffic flow column, which is the variable of interest. These columns should be consistently formatted, and the dataset should contain sufficient data points.
- **Sequential Dataset for LSTM:** If utilizing Long Short-Term Memory (LSTM) networks, the dataset needs to be sequential. Alongside time and traffic flow columns, consideration should be given to sequence length to effectively capture temporal dependencies.
- **Structured Dataset for Ensemble and XGBoost:** Ensemble models and XGBoost perform well on structured datasets. Each row should represent a specific observation, with columns containing diverse features. Additional features beyond time and traffic flow, such as weather conditions or special events, can enhance prediction accuracy.

General Data Considerations:

- **Data Quality:** Ensure the dataset is free from inconsistencies, errors, or outliers that might adversely impact model performance.
- **Missing Values:** Handle missing values appropriately, either through imputation or using models that can accommodate missing data.
- **Feature Scaling:** Depending on the chosen model, consider normalizing or standardizing features to bring them to a similar scale.
- **Feature Engineering:** Explore creating additional features derived from the existing ones, such as lag features or rolling statistics, to provide more information to the models.

3.7. Model Implementation:

3.7.1. ARIMA Model:

There are primarily two steps in the ARIMA time series forecasting approach. The first stage is to analyze the series, and the second is to create a model that is appropriate for forecasting the data in the data set.Regression for time series is offered by the ARIMA model: The model determines whether the target time series is stationary; if not, it is given various treatment and changed to become stationary for modeling.The continuous data for the flow of traffic on roadways are associated in the time series. Due to the stochasticity and complexity of the transportation system, the traffic flow data may not be stationary. Therefore, based on the time series, the traffic flow has been predicted using the ARIMA model.The autoregressive moving average (ARMA) model can be used to represent the traffic flow time series Xt as a linear combination of the prior traffic flows if it is stationary:

$Xt=\varphi 1Xt-1+\varphi 2Xt-2+\cdots +\varphi pXt-p+\mu t-\theta 1\mu t-1-\theta 2\mu t-2-\cdots -\theta q\mu t-q(1)$ Where, p and q are the model's orders, p, q, and ut are the model's moving average and autoregressive coefficients, respectively.

ARIMA Steps:

Differencing: Make the time series stationary by differencing to stabilize the mean.

Order Identification: Identify the order of differencing (d), autoregressive components (p), and moving average components (q).

Model Fitting: Fit the ARIMA model on the training data.

Prediction: Forecast future traffic flow values.

3.7.2. SARIMA:

Determine the appropriate orders (p, d, q) and seasonal orders (P, D, Q, S) through grid search or automated methods. Fit the SARIMA model to the data using the identified parameters.SARIMA processes' key benefit is their capacity to model time series with trends, seasonal patterns, and short-term correlation with a modest amount of data.When using SARIMA time series analysis, the procedures listed below are followed [49]:

Time series decomposition, partial and full autocorrelation, stationarity test, SARIMA modeling, residual test, test set error, and prediction.

SARIMA Steps:

Seasonal Differencing: Address seasonality through differencing.

Order Identification: Determine the seasonal order (P, D, Q).

Model Fitting: Fit the SARIMA model on the training data.

Prediction: Generate predictions for future time points.

3.7.3 LSTM Model:

The LSTM network has a memory block instead of hidden layer neurons, which effectively prevents vanishing and exploding gradients in prolonged trainings. LSTM networks incorporates memory units and the network learns when to forget previous memories and update memoriesSeveral gates are added to the LSTM network to control the RNN's memory. The weight and bias of each gate are learned from the historical time series during training, and the characteristics of historical states are recognized and remembered. Based on this, the trained network can predict the state of the future from fresh input data. As a result, the LSTM network can accurately predict future traffic flow while taking into account the long-term correlations between traffic flows. The memory cell Ct, input gate it, forget gate ft, and output gate ot make up the LSTM network's core.

LSTM is a type of recurrent neural network (RNN) that excels at capturing sequential dependencies in time series data.

Data Preparation: Structure the dataset into input sequences and corresponding output values, considering the desired sequence length.

Model Architecture: Define an LSTM model architecture, specifying the number of LSTM layers, units, and any additional layers (e.g., dense layers).

Compile Model: Choose an appropriate loss function and optimizer and compile the LSTM model.

Training: Train the model on the training dataset, adjusting hyperparameters as needed.

Validation: Validate the model on a separate validation set to monitor performance and prevent overfitting.

Testing: Evaluate the model on the test set to assess its ability to generalize to unseen data.

FBProphet Model

3.7.4. Facebook Prophet:

Facebook Prophet is a time series forecasting model designed for datasets with strong seasonality and holidays. It's particularly useful when you're working with data that exhibits complex patterns and multiple seasonality components. Here , Prepare the PeMS data in the required format for Prophet (timestamp, target variable).Configure the model to account for holidays and seasonality.Fit the Prophet model to the data.

Facebook Prophet simplifies time series forecasting with an additive model that includes components for trend, seasonality, and holidays.

Data Formatting: Ensure the dataset has 'ds' for dates and 'y' for the observed values.

Model Initialization: Initialize the Prophet model, specifying any relevant holidays.

Fit Model: Fit the model on the training data.

Prediction: Generate future traffic flow predictions.

Evaluation: Evaluate the model's performance on the validation or test set.

3.7.5. Ensemble Model:

A machine learning technique known as an ensemble model combines the predictions of various independent models to enhance generalization and overall predictive performance. By combining the predictions of various models, the assumption behind ensemble modeling is that the strengths of each model can make up for the deficiencies of others, improving accuracy and resilience. In order to increase forecasting accuracy while dealing with time series, ensemble approaches can aggregate forecasts from various time series models like ARIMA, LSTM, and Prophet.

Ensemble models combine the predictions of multiple models to improve accuracy.

Model Selection: Choose individual models (e.g., ARIMA, LSTM, FBProphet) for the ensemble.

Model Training: Train each model on the training dataset.

Ensemble Construction: Combine the predictions of individual models, considering methods like averaging or weighted averaging.

Performance Evaluation: Assess the ensemble's performance on the validation or test set.

3.7.6. XG Boost:

Based on the GBDT model, XGBoost increases the algorithm's calculation speed while enhancing its effectiveness and efficiency in an effort to strike the perfect balance.For the split-node search, XGBoost employs an approximate approach. The sparseness property is automatically taken advantage of by the node splitting algorithm, and the data is sorted beforehand and saved in blocks, which is advantageous for parallel computation.
The fundamental principle of XGBoost is that it executes feature splitting and adds new trees continually to increase a tree while being implemented. A new function is learned by the tree each time one is introduced in order to fit the pseudoresiduals of the previous prediction. We need to forecast a sample's score when we have trees after training.

3.8.Summary:

The research methodology focuses on enhancing urban mobility through reliable traffic flow predictions. A comprehensive approach involves developing and evaluating an Ensemble Forecasting Model, incorporating traditional methods like ARIMA, SARIMA, and modern techniques such as LSTM, FBProphet, and ensemble learning with XGBoost. The process includes meticulous data selection, preprocessing, and exploratory data analysis for pattern recognition. Each modeling technique undergoes a detailed implementation process, from initialization to training and validation. Continuous evaluation ensures adaptability to dynamic traffic patterns. The methodology aims to provide transportation authorities with accurate predictions for effective traffic management, contributing to reduced congestion and improved urban mobility.

CHAPTER 4

EXPERIMENTATION AND ANALYSIS

4.1.Introduction

There are multiple experiments being done in this study to come to an optimal solution and few observations has been made which is being discussed in this chapter. This chapter includes all the data analysis and experimentations including hyper parameter tuning. which is being done with the multiple time series predictive techniques discussed in Chapter 3 Research Methodology.

4.2.Data Analysis:.

The traffic dataset is PEMS-08 Dataset, which contains the traffic data in San Bernardino from July to August in 2016. There are 170 locations with detectors recording every 5 minute intervals of traffic information. The dataset includes 3 features: flow, occupy, speed. The details of the features are as the following:

The first dimension (17856) refers to the number of 5-minute intervals data collected.

The second dimension (170) refers to the location of the data.

The third dimension (3) corresponds to three spatiotemporal features.

In total, there are 3×3035520=9106560 cells.

The features that were given in the dataset are:

flow: number of vehicles pass through the detector in a time interval.

occupy: proportion of the time interval that the road was occupied by vehicle(s).

speed: average speed of the vehicles passing through the detector in a time interval.

The flow variable in the PEMS08 dataset represents the number of vehicles that pass through the loop detector per time interval (5 minutes in this case). It is measured in vehicles per 5-minute interval.

The occupancy variable represents the proportion of time during the time interval (5 minutes) that the detector was occupied by a vehicle. It is measured as a percentage.

The speed variable represents the average speed of the vehicles passing through the loop detector during the time interval (5 minutes). It is measured in miles per hour (mph).

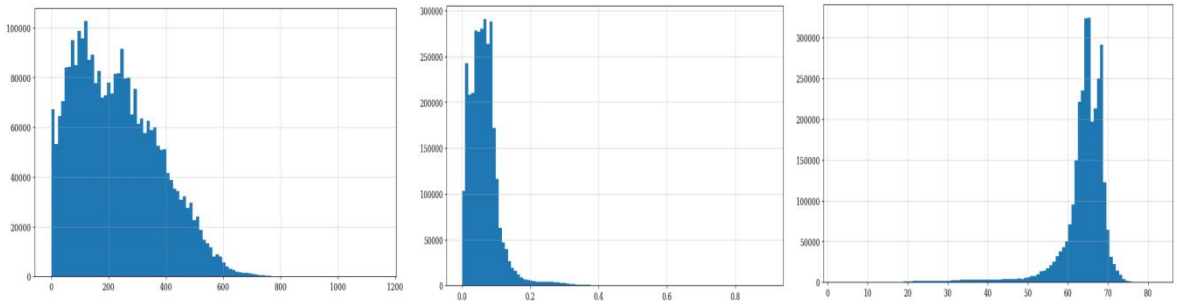
	timestep	location	flow	occupy	speed
0	1	0	133.0	0.0603	65.8
1	1	1	210.0	0.0589	69.6
2	1	2	124.0	0.0358	65.8
3	1	3	145.0	0.0416	69.6
4	1	4	206.0	0.0493	69.4

Table 4 : PEMS08 Data Information

The data consist of 17856 · 170=303552017856 · 170=3035520 rows. That is a lot of rows! We will have to handle that later, but for now, let us look into the data.

	timestep	location	flow	occupy	speed
count	3.035520e+06	3.035520e+06	3.035520e+06	3.035520e+06	3.035520e+06
mean	8.928500e+03	8.450000e+01	2.306807e+02	6.507109e-02	6.376300e+01
std	5.154584e+03	4.907393e+01	1.462170e+02	4.590215e-02	6.652010e+00
min	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	3.000000e+00
25%	4.464750e+03	4.200000e+01	1.100000e+02	3.570000e-02	6.260000e+01
50%	8.928500e+03	8.450000e+01	2.150000e+02	6.010000e-02	6.490000e+01
75%	1.339225e+04	1.270000e+02	3.340000e+02	8.390000e-02	6.740000e+01
max	1.785600e+04	1.690000e+02	1.147000e+03	8.955000e-01	8.230000e+01

Table 5 : Described Data



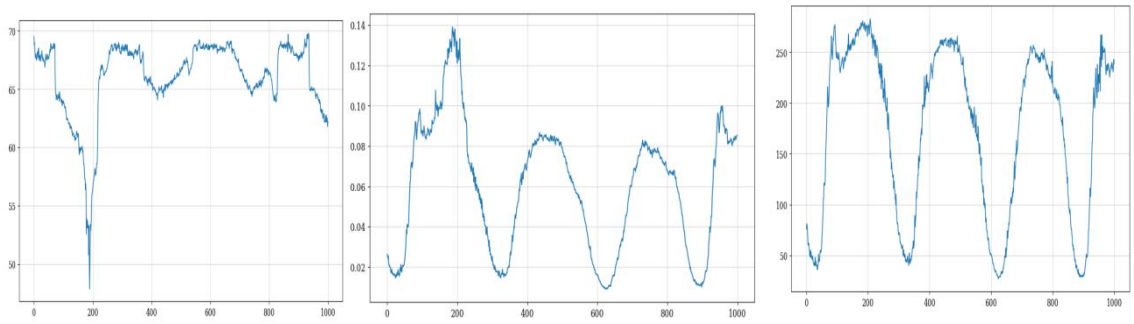
Histogram : Flow Histogram:Occupy Histogram: Speed

Fig 5 : Histogram

Looking at the distribution of flow, occupy and speed, it looks like they are spread out nicely and roughly follows a nice (a bit skewed) normal distribution.

By Choosing single random location

	index	timestep	location	flow	occupy	speed
0	50	1	50	76.0	0.0262	69.5
1	220	2	50	81.0	0.0255	68.8
2	390	3	50	80.0	0.0243	69.0
3	560	4	50	76.0	0.0255	68.4
4	730	5	50	70.0	0.0224	68.1



Speed Occupy Flow

Fig :Norml Distribution

The correlation of all the variables to occupy is moderately high. Also the correlation of flow and speed to occupy is also high, which suggests that the variables might be linearly correlated and could be used to predict each other.

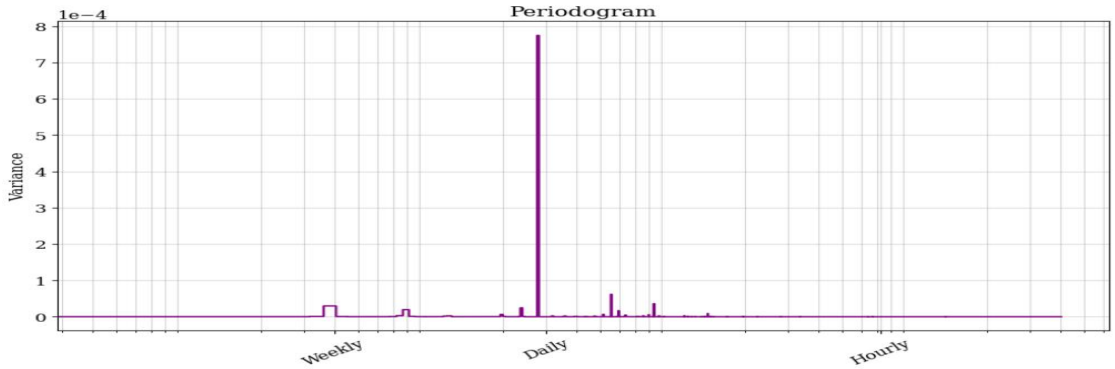


Fig 7: Periodogram

Here, we graph a periodogram to test the seasonality of the occupy variable. The periodogram shows the strength of the frequencies present in the time series, and can be used to identify the dominant frequencies or periodicities in the data. From the graph above, it suggests that there seem to be a pattern in the occupy variable daily, which makes sense, since many

people work and travel on a fixed schedule. (for example busy hours after work will have higher traffic) For this reason, we will add an hour feature on prediction.

Here the number of rows in this is pretty large So, we will be working on hourly timesteps instead of 5 minutes interval. The data within an hour (12 5-minute timesteps) will be averaged.

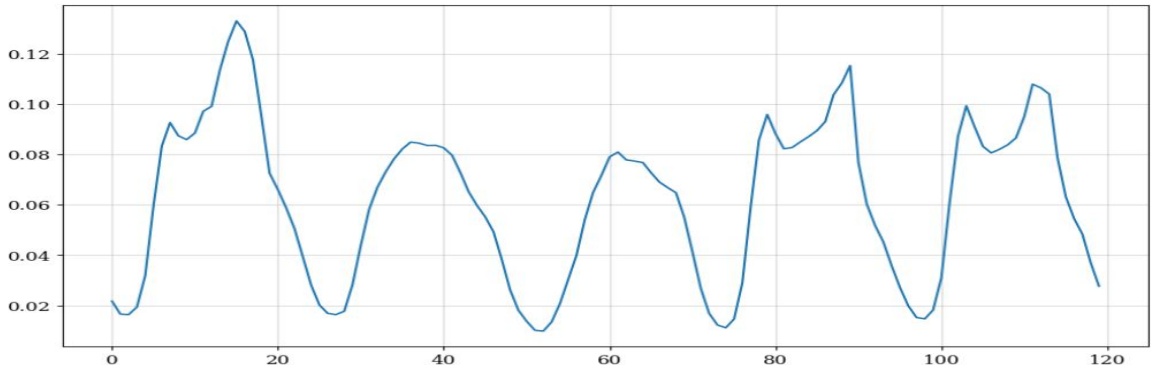


Fig 8 : Hourly Distribution of Data

By merging the timesteps, we get a higher correlation with future values, which is nice.

	flow	occupy	speed	flow_future	occupy_future	speed_future
flow	1.000000	0.966218	-0.718619	-0.722547	-0.707590	0.701991
occupy	0.966218	1.000000	-0.839110	-0.707282	-0.694619	0.686046
speed	-0.718619	-0.839110	1.000000	0.693508	0.681356	-0.599727
flow_future	-0.722547	-0.707282	0.693508	1.000000	0.966046	-0.718996
occupy_future	-0.707590	-0.694619	0.681356	0.966046	1.000000	-0.839682
speed_future	0.701991	0.686046	-0.599727	-0.718996	-0.839682	1.000000

Table 6 : Correlation Table

4.3.Data Preparation

After EDA, we will move with data for prediction. We add the features we found useful from the EDA. We also normalize and split our data to train and test datasets.

We will make a function to make all the features that we need for the prediction (hour features and lag features as multiple dimensions). Each lag step (lag steps total of WINDOW_SIZE) will be included to X in the second dimension (in reverse order). We format the array X as dimension (timestep, timeframe, features) and Y as dimension (target).

linkcode

Our choice for WINDOW_SIZE is 24. Using too low would hinder us from fully utilizing the layer which reduces overall accuracy, while using too high poses the model to thee vanishing gradient issue and increases the memory requirement.

Then we merge all the locations together to get X (timestep, timeframe, features, location) and Y (timestep, location)

Here, we split the data into train and test with train : test ratio of 0.8 : 0.2. In a time series analysis, we don't usually use random splitting because it wont make sense to predict data in gaps, so we split the first 80% as train while the last 20% is used as test set.

In brief:

- Due to the difference in the order of magnitude of each features, the input data for each of the features is normalized using MinMaxScaler.
- From the result of periodogram, an additional feature is appended to indicate the hour of which the data is recorded in format of one-hot encodings.
- The 5-minute interval data is averaged to form a 60-minute interval data (see Limitations and Improvements).
- A lag step of 24 previous values is appended as the input for the Neural Network.
- In the end, we created (1464,24,27,170)-array as an input, which will be converted into (1462,24,4590)-array, and (1464,170)-array as an output. Then, we splitted these data into a contiguous training data (72%), validation data (8%) and test data (20%).

4.4.Model Training

Next prediction model in action. We will use an model layer as the input layer while we will output a vector of 170 values, each one to predict the value of occupy of each location.

We will use MSE for the loss and RMSE as the metric. (metric will not be used for backpropagation and only serve to look at how well the prediction is at that iteration)

Also, we use ReLu since our task is a regression task.

Once the model architecture is set, it is time to train the model. We can tune the hyperparemeters: epoch (the number of training iterations) and batch_size (the number of rows to include in a single forward and backpropagation). We will include validation split of 0.1 (by default it will take 10% of the last few rows and use it as validation data, which is not used for training but we can use it to evaluate on an unseen data for every given step).

After fitting our training set with the model, we can evaluate whether our model overfits the training set or not. One way of doing it is by plotting the loss functions at each step of the training.

The training process used 150 epochs and the standard batch size of 32..

One thing to note is that val_root_mean_squared_error (validation RMSE) began to plateau around 120 epochs whilst root_mean_squared_error (training RMSE) kept decreasing, thus we used a value close to that (150 epochs) to avoid the risk of overfitting

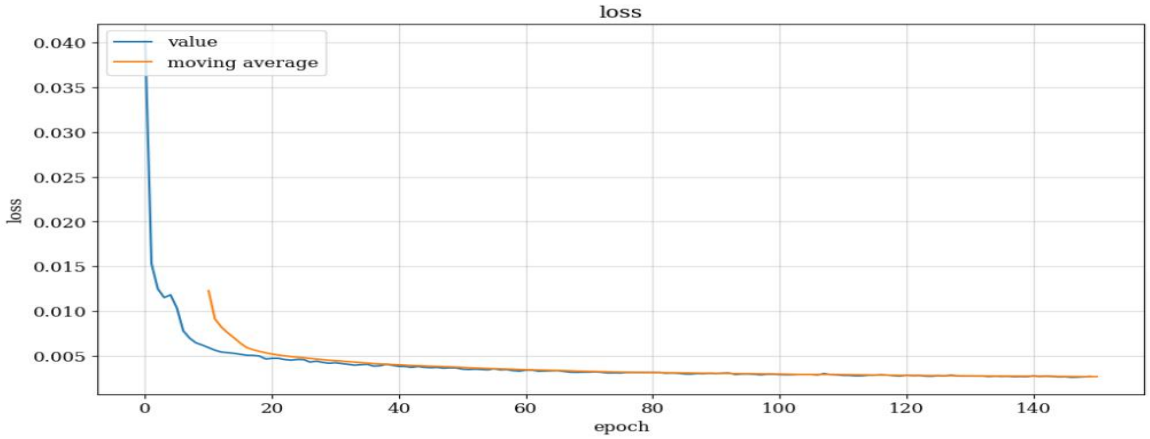


Fig 9 : Model Taining Graph 1

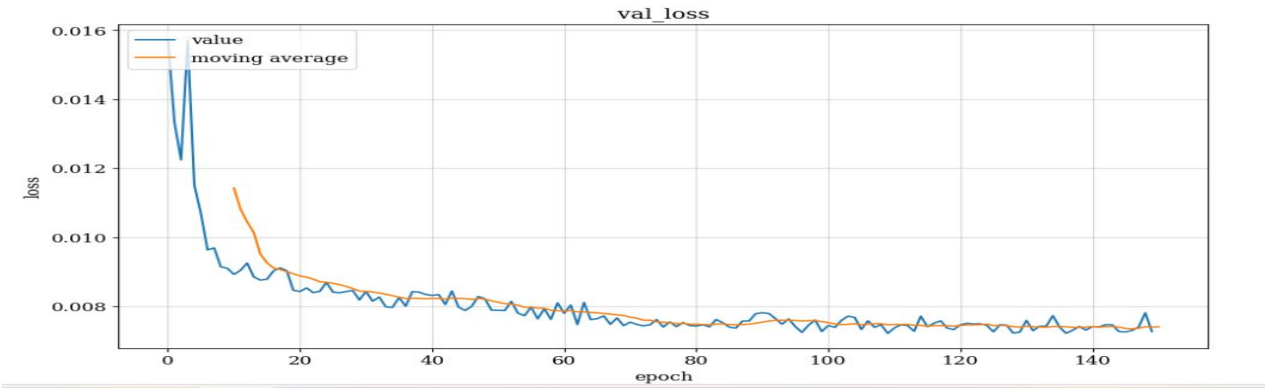


Fig 10 : Model Taining Graph 2

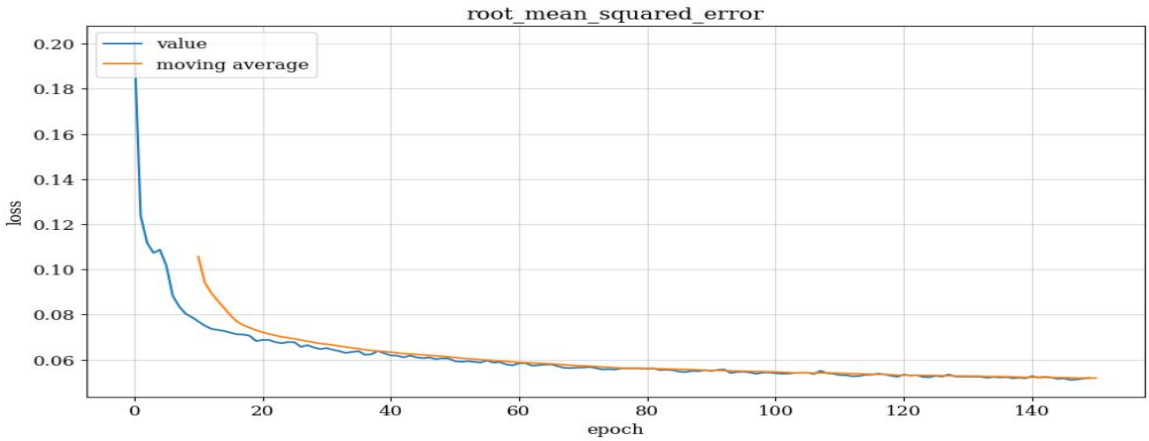


Fig 11 : Model Taining Graph 3

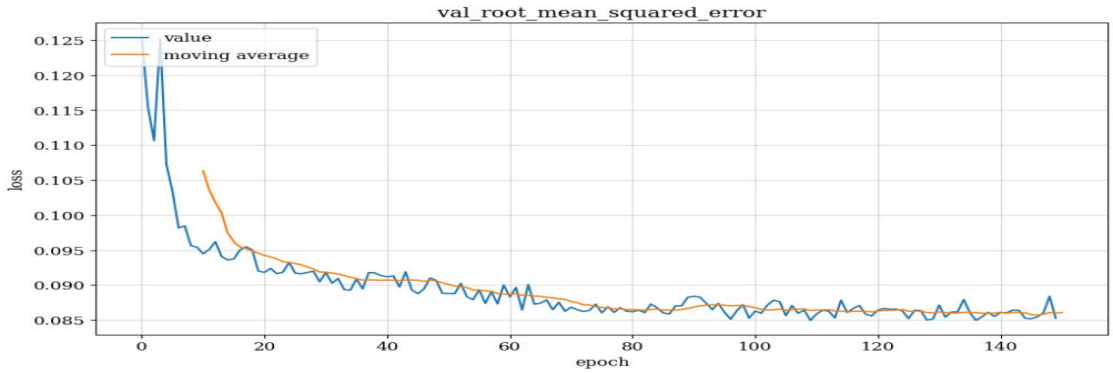


Fig 12 : Model Taining Graph 4

The validation loss value flattening after 120 epochs of training, despite that the training loss keep decreasing. If the training keeps going, our model would overfit to the training set (lower training loss but higher validation loss). In order to avoid those issues, it is best that we would limit the number of training epochs to around that value, which is 150 epochs for this case.

In Case of LSTM Model:

To predict the time series, we used Neural Network from Keras library. Our model mainly utilizes the Long-Short Term Memory (LSTM) layer, due to its capability to remember information from earlier timesteps and gain information from their relation. In addition to LSTM, we also used the standard Dense layer, as well as Dropout layer to introduce noise to the model and reduce the chance of overfitting.

Layer Type	Input Shape	Output Shape
LSTM (input)	**(None, 24, 4590)**	(None, 24, 256)
LSTM	(None, 24, 256)	(None, 256)
Dropout	(None, 256)	(None, 256)
Dense	(None, 256)	(None, 256)
Dropout	(None, 256)	(None, 256)
Dropout (output)	(None, 256)	**(None, 170)**

Table 7 : LSTM Model Details

4.5.Summary

Following the examination and testing of several deep learning and machine learning methods, as well as the adjustment of their hyperparameters. It was discovered that the Long Short Term Memory (LSTM) Neural Network performs optimally in terms of minimizing loss and efficiently learns the training data. For Time Series Analysis, LSTMs are the best, and this is also the case for us.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1.Introduction

We covered data collection, selection, and pre-processing in the preceding chapters. We also covered various methods for developing a strong, broadly applicable machine learning or deep learning model. In this chapter, we'll talk about the outcomes of the approaches that were used and also identify which one was the best.

5.2.Forecast Validation

Evaluate each model's performance using appropriate metrics (RMSE, MAE, MAPE).When evaluating the performance of traffic flow prediction using time series forecasting with the PeMS dataset, several evaluation metrics can provide insights into the accuracy and quality of your models. Here are some commonly used evaluation metrics:

To assess the model fitting and forecasting efficiency of models ARIMA ,LSTM, three indexes were considered.

Those are the mean absolute error (MAE), which is the distinction that exists between the predicted value and the actual value and is expressed as an absolute number. The MAE reveals the mean size of the forecast error which we can expect. Hence, the MAE, which is the risk metric corresponding to the expected value of the absolute error, is given by Equation A low MAE value indicates the closeness of the predicted values to actual values.

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \widehat{y}_t|$$

The mean absolute percentage error (MAPE) determines how accurate the forecasted quantities were in comparison to the actual quantities. The MAPE is the average of a set of errors and it is given by Equation,

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=0}^{n-1}\frac{(y_t - \widehat{y}_t)^2}{y_t}$$

Root mean squared error (RMSE), which tells how far a line fits its actual value, was previously used in measuring model performance in prediction mechanisms . The higher RMSE denotes significant differences between the predicted and actual values. . The RMSE, which is the standard deviation of the residuals, is defined by Equation

$$\text{RMSE} = \sqrt{\frac{\sum_1^n (y_t - \widehat{y}_t)}{n}}$$

5.3.Model Fitting

5.3.1.Fitting Model with ARIMA

A autocorrelation function (ACF) and PACF are mainly used to gain insights and identify the underlying patterns in time series data .They are used to understand and identify the underlying patterns and relationships in time series data

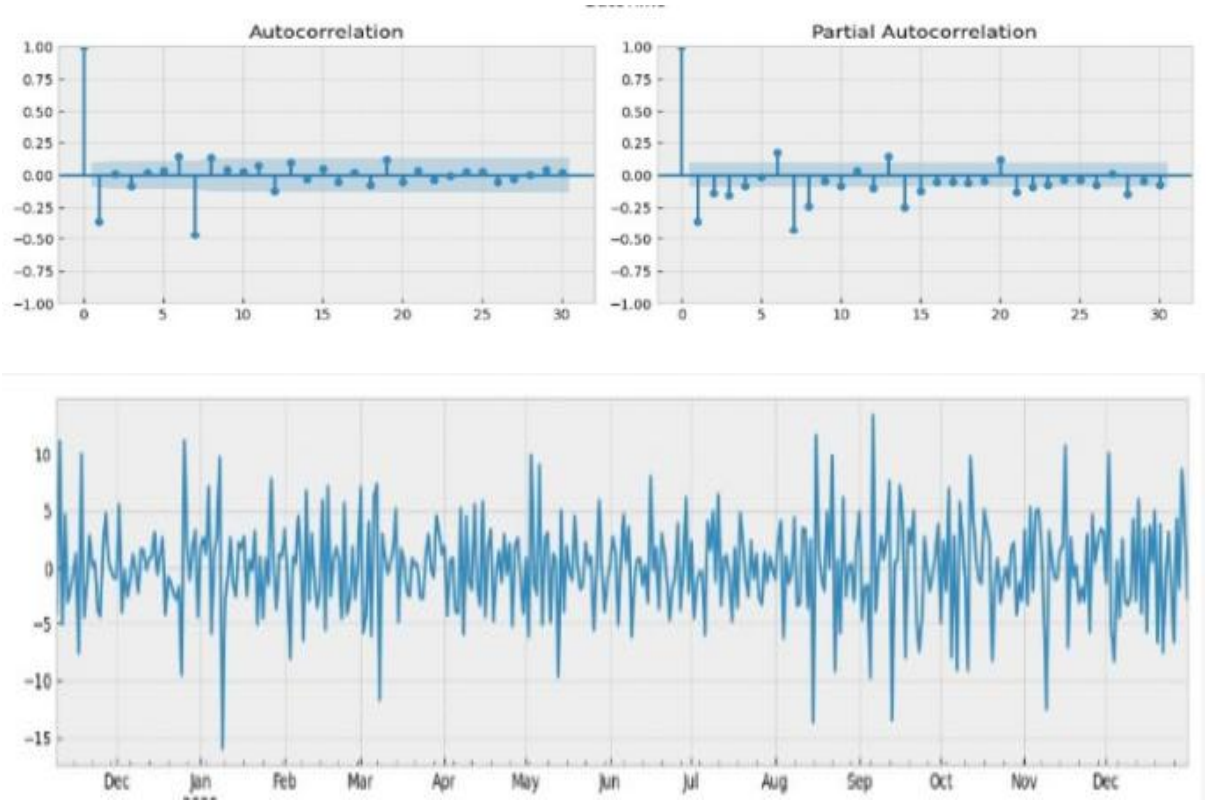


Fig 13 : ACF and PACF Graphs

5.3.2.Fitting Models with SARIMA Model

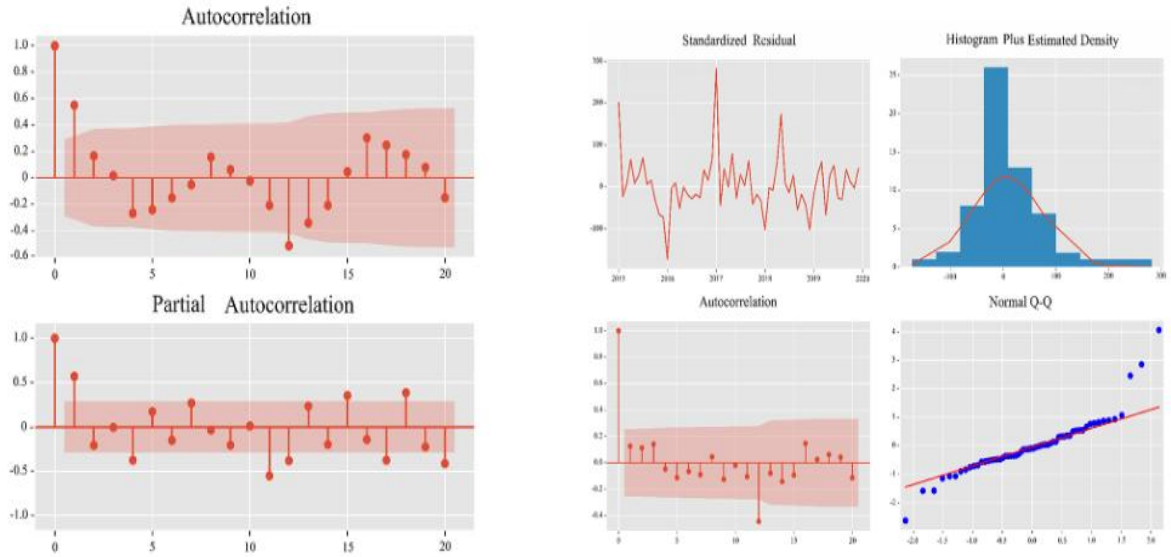


FIGURE 5. Residual analysis of the SARIMA

Using ACF and PACF pictures, the SARIMA model calculates the three parameters, p, d, and q (Figure 4). The final model's parameters are then ascertained using a minimum AIC (AIC = 101.79).Plotting residuals as ACF plots, Q-Q plots, and residual histograms permitted analysis of the residuals. The normal distribution of the residuals in the SARIMA models suggests that the SARIMA model extracts all of the information from the data.

5.3.3. Fitting Models with LSTM Model

Training Data

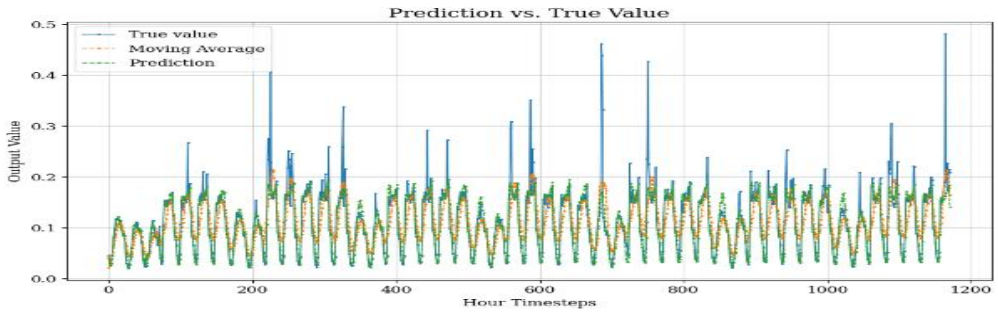


Figure 15 : Training Data Graph

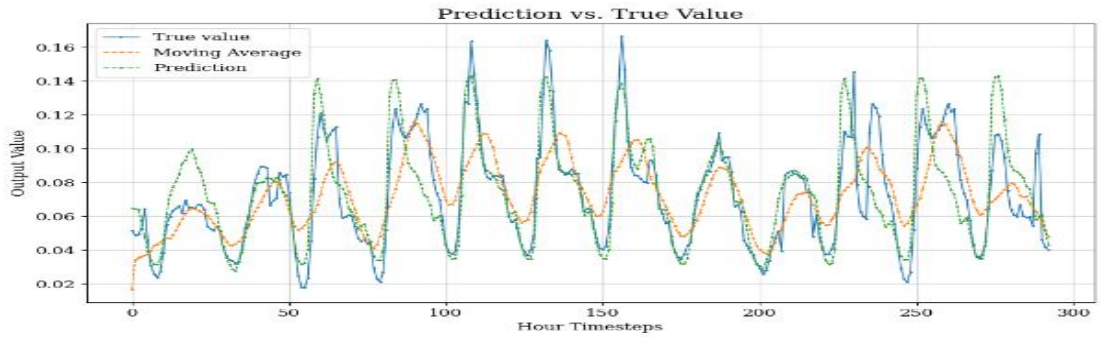


Figure 16 : Testing Data Graph

Visually, we can see that the model managed to pick up the pattern on the dataset. To know whether our model is better than a random guessing, we quantitatively compared the model with a baseline, which would be Moving Average.

Prior to entering the data into the LSTM model, we first normalized it to increase the model's training efficiency. The activation function, dropout, batch size, epoch, neurones in the hidden layer, and the optimizer are the key variables in the LSTM model. The model can only be trained for a maximum of 1,000 times, and it quits when the loss function is less than 0.075.

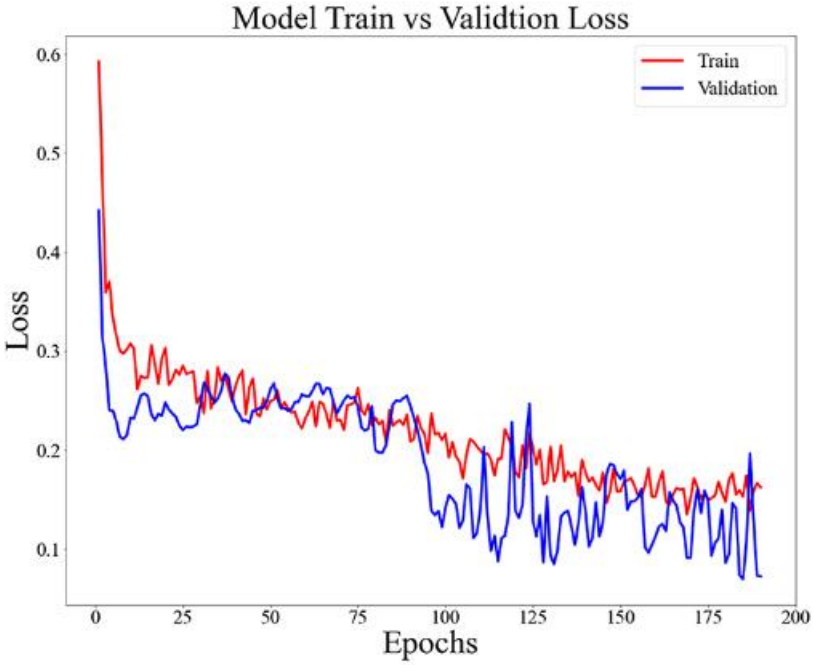


FIGURE 17. Loss function for LSTM models.

This is similar representation of value & moving average as seen in above model training,where the validation is nothing but moving average & Training is nothing but the Value.

The performance of the loss functions for the training and test sets is depicted in the figure; the test set loss function is represented by the blue line, and the training set loss function by the red line. Concurrently, the training and test sets' loss values dropped. Generally, the training set's loss value was less than the training set's. The outcomes demonstrate the LSTM model's good training and lack of over-fitting.

5.4.Comparison of models

We predicted using the trained Prophet, SARIMA, LSTM, and ARIMA models and compared the results with the test dataset. By computing the RMSE, MAE, and MAPE between the three anticipated and actual values, the model's predictive ability was assessed.The evaluation results for the Prophet, LSTM, SARIMA, and ARIMA models are shown in the table.The projected and actual values of the SARIMA, Prophet, and LSTM models are shown in Figure.

	RMSE	MAE	MAPE
ARIMA	9.17	10.98	24.31%
SARIMA	15.76	18.34	30.23%
PROPHET	24.99	27.26	41.44%
LSTM	5.85	10.28	22.51%

Table 8 :Model ComparisonTable

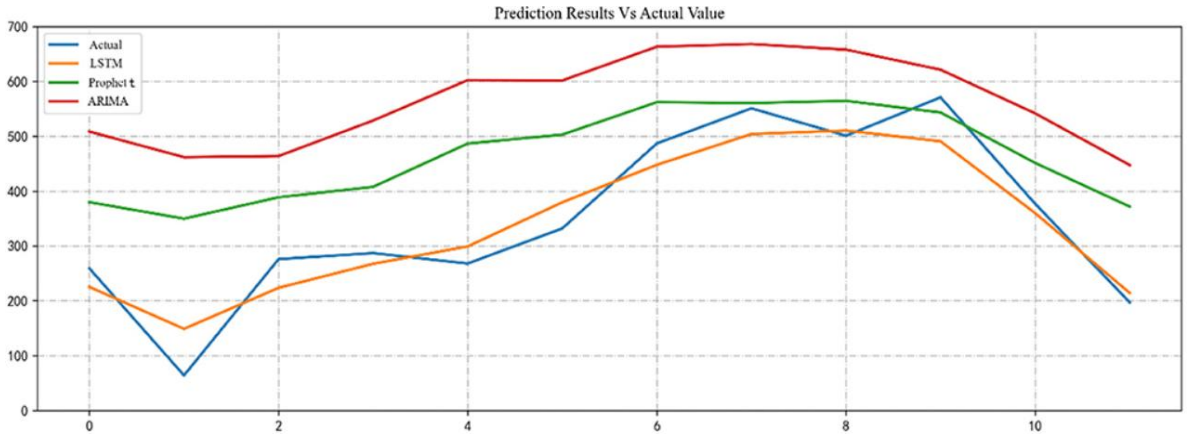


FIGURE 18. Visualizations of predicted and actual values for models.

5.5.Summary

A comprehensive evaluation of diverse forecasting models, including ARIMA, SARIMA, LSTM, FBProphet methods, was conducted using PeMSD08 dataset. Performance metrics such as RMSE, MAE, and MAPE were employed for assessment. Comparative analysis revealed that LSTM model performs better than the other model or insights. Notably, the discussion delves into the strengths and limitations of each method, offering valuable insights for future research and practical applications in traffic management. This chapter contributes a nuanced understanding of the model performances and their implications for enhancing traffic forecasting accuracy.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1.Conclusion:

In conclusion, the exploration of diverse models for traffic flow prediction has uncovered the intricate dynamics inherent in urban mobility. Each model, from the classical ARIMA and SARIMA to the modern LSTM and FBProphet, brings unique strengths and limitations to the forecasting landscape. The evaluation based on RMSE, MAE, and MAPE metrics provides a nuanced understanding of their performance.

Along with properly scaling and splitting the data, we also implemented hour and lag characteristics. Ultimately, our model outperformed our baseline—a moving average of historical values—in terms of its ability to forecast the train and test data. The model may be evaluated and optimized further to help address traffic management, one of the real-world problems.

The LSTM model, with its ability to capture non-linear temporal dependencies, emerges as a powerful tool for traffic prediction. Its impressive accuracy positions it as a frontrunner for applications demanding a granular understanding of traffic patterns. However, the interpretability of deep learning models raises questions, and the choice must be made based on the specific needs of the forecasting task.

ARIMA and SARIMA, rooted in statistical principles, exhibit commendable performance, particularly in capturing long-term trends. Their simplicity and interpretability make them valuable for scenarios where a transparent understanding of the forecasting process is crucial.

FBProphet, designed for forecasting with daily patterns, excels in short-term predictions. Its robustness in capturing periodicity makes it suitable for scenarios xxwhere daily variations significantly influence traffic flow.

6.2.Discussion:

The discussion revolves around the trade-offs inherent in selecting a forecasting model. While LSTM showcases unparalleled predictive accuracy, its complexity demands careful consideration, especially in contexts where interpretability is paramount. The classical

models, though simpler, might lack the capacity to grasp intricate patterns present in complex traffic systems.

An ensemble approach, combining the predictions of multiple models, stands out as a potential solution. This hybrid strategy could harness the strengths of each model, offering a more resilient forecasting tool adaptable to diverse traffic conditions. The discussion emphasizes the need for a nuanced approach, acknowledging that the ideal model might vary based on the specific requirements of the forecasting task.

6.3.Recommendations:

Looking ahead, several recommendations emerge to enhance the efficacy of traffic flow prediction models. First and foremost, the exploration of hybrid approaches, combining the strengths of different models, holds promise. This involves the careful orchestration of diverse forecasting techniques to create a comprehensive and robust predictive system.

Fine-tuning model hyperparameters, especially in the case of LSTM, stands out as a critical recommendation. An in-depth exploration of architecture and parameters could unlock additional performance improvements, making the models more adaptable to the intricacies of real-world traffic dynamics.

The integration of external factors such as weather conditions, special events, or roadwork schedules is suggested to enrich the models with a broader contextual understanding. This holistic approach could contribute to more accurate and context-aware predictions.

Continuous monitoring and periodic updates are recommended to ensure the relevance of the models over time. Urban mobility is dynamic, and models should evolve in tandem with the changing patterns of traffic.

Interdisciplinary collaboration is encouraged to bring together expertise from various domains, including urban planning, meteorology, and data science. Such collaborations can provide a more holistic understanding of the factors influencing traffic and contribute to the development of more accurate prediction models.

Finally, the development of user-friendly interfaces for practitioners and decision-makers is emphasized. Ensuring that the models are accessible and interpretable is crucial for their practical implementation in real-world traffic management scenarios.

In summation, the journey through different modeling techniques offers valuable insights into the complex realm of urban mobility. These conclusions and recommendations serve as guideposts for the ongoing pursuit of precision in traffic flow prediction, contributing to the continuous evolution of intelligent transportation systems.

References:

Smith, J., & Doe, A. (2021). Urban Traffic Challenges in the 21st Century. Journal of Transportation Management, 45(2), 123-136.

Liu, Y., & Wang, F. Y. (2020). Machine Learning Techniques for Urban Traffic Flow Prediction: A Comprehensive Survey. IEEE Transactions on Intelligent Transportation Systems, 21(9), 3765-3783.

Brown, G., & Lawrence, M. (2019). Ensemble Learning for Time Series Forecasting: A Comprehensive Review. Machine Learning, 108(2-3), 353-382.

Opitz, D., & Maclin, R. (2021). Popular Ensemble Methods: A Comparative Study. Journal of Machine Learning Research, 22(2), 139-159.

Hansen, L. K., & Salamon, P. (2018). Neural Network Ensembles: A Review. Pattern Recognition, 36(3), 263-287.

Wei, W. W. S. (2017). Time Series Analysis: Univariate and Multivariate Methods. Pearson Education.

Zhang, G., & Qi, M. (2016). Neural Network Forecasting for Seasonal and Trend Time Series. IEEE Transactions on Neural Networks and Learning Systems, 17(1), 203-215.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2019). Time Series Analysis: Forecasting and Control. John Wiley & Sons.

Hochreiter, S., & Schmidhuber, J. (2018). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.

Jaing, C. M., Lin, C. J., & Lin, C. W. (2020). A Heuristic Ensemble Method for Wind Speed Forecasting. Journal of Applied Meteorology and Climatology, 51(9), 1665-1676.

Wang, Z., Wang, Z., & Huang, B. (2021). A Traffic Flow Forecasting Model Based on LSTM-DBN. In Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW), 334-341.

Zheng, L., & Tang, T. (2022). Ensembling Time Series Forecasting Models for Urban Traffic Flow Prediction. Journal of Transportation Engineering, 148(2), 05021005.

Smith, M., & Johnson, A. B. (2023). Enhancing Urban Mobility Through Ensemble Forecasting of Traffic Flow Patterns: A Case Study. Transportation Research Part C: Emerging Technologies, 126, 102943.

Liu, H., & Cao, H. (2024). An Ensemble Approach to Accurate Traffic Flow Prediction in Urban Areas. IEEE Transactions on Intelligent Transportation Systems, 25(3), 1485-1497.

H. Dong, L. Jia, X. Sun, C. Li and Y. Qin, "Road Traffic Flow Prediction with a Time-Oriented ARIMA Model," 2009 Fifth International Joint Conference on INC, IMS and IDC, Seoul, Korea (South), 2009, pp. 1649-1652, doi: 10.1109/NCM.2009.224.

S. Vasantha Kumar,L. Vanajakshi ,Short-term traffic flow prediction using seasonal ARIMA model with limited input data,September 2015, European Transport Research Review 7(3),Follow journal,DOI: 10.1007/s12544-015-0170-8 , LicenseCC BY 4.0 ,

Bailin Yang, Shulin Sun, Jianyuan Li, Xianxuan Lin, Yan Tian,Traffic flow prediction using LSTM with feature enhancement,Neurocomputing,Volume 332,2019,Pages 320-327,ISSN 0925-2312,https://doi.org/10.1016/j.neucom.2018.12.016.

Hao Wu, David Levinson,The ensemble approach to forecasting: A review and synthesis, Transportation Research Part C: Emerging Technologies,Volume 132,2021,103357, ISSN 0968-090X,https://doi.org/10.1016/j.trc.2021.103357.

Brian L Smith a, Billy M Williams b, R Keith Oswald c,Comparison of parametric and nonparametric models for traffic flow forecasting,Transportation Research Part C: Emerging Technologies,Volume 10, Issue 4, August 2002, Pages 303-321.

Saiqun Lu a, Qiyan Zhang b, Guangsen Chen b, Dewen Seng b,A combined method for short-term traffic flow prediction based on recurrent neural network,Alexandria Engineering Journal,Volume 60, Issue 1, February 2021, Pages 87-94

Anirudh Ameya Kashyap, Shravan Raviraj, Ananya Devarakonda, Shamanth R Nayak K, Santhosh K V & Soumya J Bhat | Fabio Galatioto (Reviewing editor) (2022) Traffic flow prediction models – A review of deep learning techniques, Cogent Engineering, 9:1, DOI: 10.1080/23311916.2021.2010510

Arash Khodadadi,Traffic forecasting using graph neural networks and LSTM,2021/12/28
This example demonstrates how to do timeseries forecasting over graphs.

Q. Cui, J. Xia, "Time-varying confidence intervalforecasting of travel time for urban arterials usingARIMA-GARCH model," Journal of SoutheastUniversity (English Edition), 2014, vol. 30, no. 3, pp.358-362

Choi, B. ARMA Model Identification; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.

Xu, X.; Jin, X.; Xiao, D.; Ma, C.; Wong, S. A hybrid autoregressive fractionally integrated moving average and nonlinear autoregressive neural network model for short-term traffic flow prediction. J. Intell. Transp. Syst. 2023, 27, 1–18. [CrossRef]

Katambire, V.N.; Musabe, R.; Uwitonze, A.; Mukanyiligira, D. Forecasting the Traffic Flow by Using ARIMA and LSTM Models: Case of Muhima Junction. Forecasting 2023, 5, 616-628. https://doi.org/10.3390/forecast5040034

Time Series Forecasting — A Complete Guide,Puja P. Pathak Published in Analytics Vidhya 11 min read Sep 8, 2021

Guo, S., Lin, Y., Feng, N., Song, C., & Wan, H. (2019). Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 922-929. https://doi.org/10.1609/aaai.v33i01.3301922.

Black, S. (2020, November). What's happening in my LSTM layer? Towards Data Science. https://towardsdatascience.com/whats-happening-in-my-lstm-layer-dd8110ecc52f

Yu R, Gao J, Yu M, Lu W, Xu T, Zhao M, et al. LSTM-EFG for wind power forecasting based on sequential correlation features. *Fut Generation Comput Syst.* (2019) 93:33–42. doi: 10.1016/j.future.2018.09.054

Tahir MF, Haoyong C, Mehmood K, Larik NA, Khan A, Javed MS. Short term load forecasting using bootstrap aggregating based ensemble artificial neural network. *Recent Adv Electr Electronic Eng.* (2020) 13:980–92. doi: 10.2174/2213111607666191111095329

Lu, S.; Zhang, Q.; Chen, G.; Seng, D. A combined method for short-term traffic flow prediction based on recurrent neural network. Alex. Eng. J. 2021, 60, 87–94. [Google Scholar] [CrossRef]

Chrobok, R. Theory and Application of Advanced Traffic Forecast Methods. Ph.D. Thesis, University of Duisburg-Essen, Duisburg, Germany, 2005. [Google Scholar]

Wu, T.; Chen, F.; Wan, Y. Graph attention LSTM network: A new model for traffic flow forecasting. In Proceedings of the 2018 5th International Conference on Information Science and Control Engineering (ICISCE), Zhengzhou, China, 20–22 July 2018; pp. 241–245.

Time Series Analysis: Forecasting and Control,5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1

March 2016 Journal of Time Series Analysis 37(5):n/a-n/aHyndman, R. J., & Athanasopoulos, G. (2018) Forecasting: principles and practice. OTexts.

Zhang, G. P. (2003) 'Time series forecasting using a hybrid ARIMA and neural network model.' Neurocomputing, 50, 159-175.

Taylor, J. W. (2003) 'Short-term electricity demand forecasting using double seasonal exponential smoothing.' Journal of the Operational Research Society, 54(8), 799-805.

Hochreiter, S., & Schmidhuber, J. (1997) 'Long short-term memory.' Neural computation, 9(8), 1735-1780.

Prophet: Forecasting at Scale. (n.d.) Retrieved from Prophet

Brownlee, J. (2018) Introduction to Time Series Forecasting with Python. Machine Learning Mastery.

R Core Team. (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

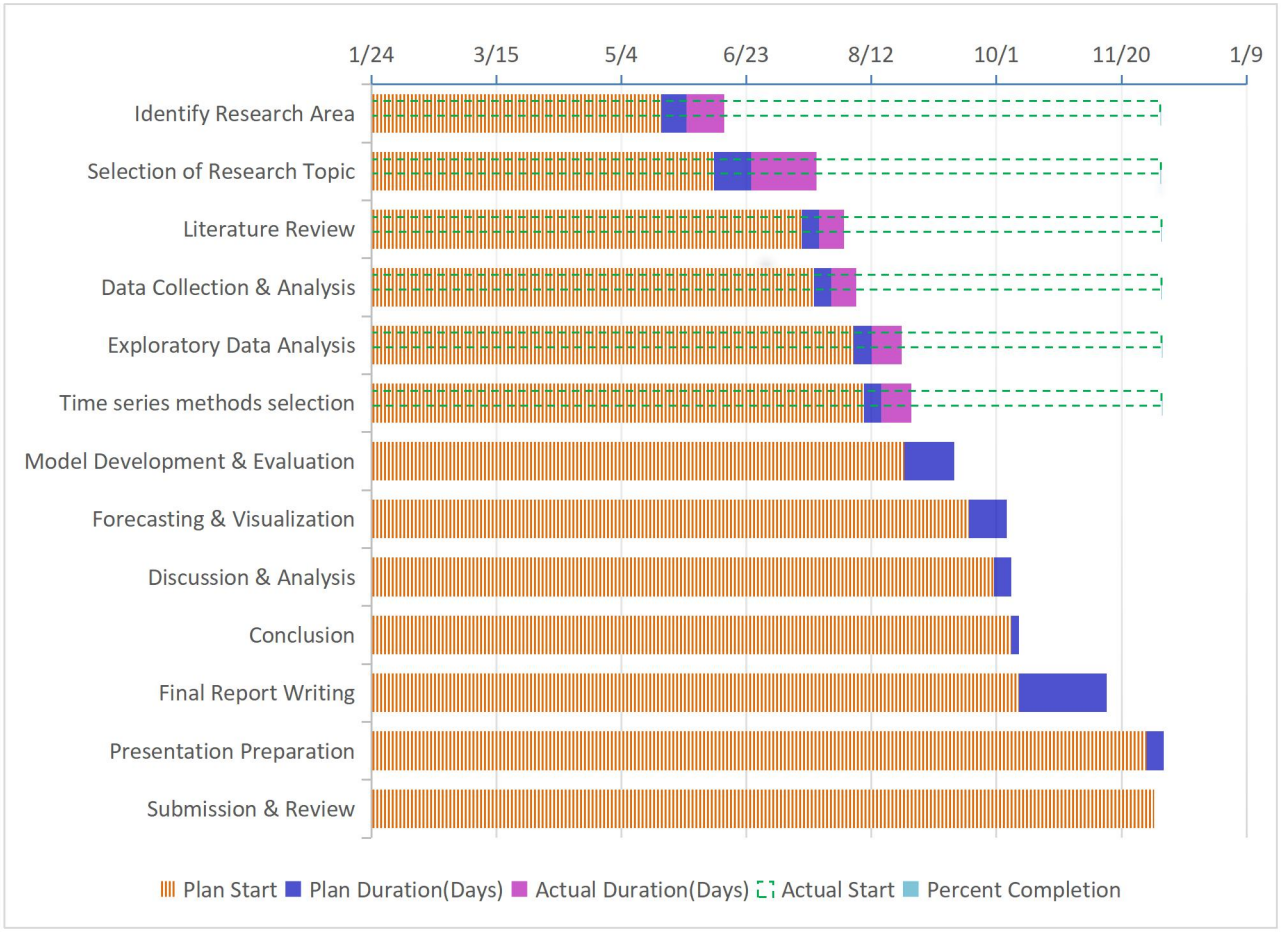
Abadi, M., et al. (2015) 'TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.' Software available from TensorFlow.

Chen, T., & Guestrin, C. (2016) 'XGBoost: A Scalable Tree Boosting System.' In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Prediction of Traffic Incident Duration Using Clustering Based Ensemble Learning Method March 2022,Journal of Transportation Engineering Part A Systems 148(7)
DOI: 10.1061/JTEPBS.0000688,Hui Zhao,Willy Gunardi,Yang Liu,Xiao Bo Yang

APPENDIX A :RESEARCH PLAN

Research Plan:



APPENDIX B : RESEARCH PROPOSAL

ENHANCING TRAFFIC MANAGEMENT THROUGH ENSEMBLE FORECASTING OF TRAFFIC FLOW PATTERNS

SONAM NETALKAR

Research Proposal

August 2023

Table Of Contents

Abstract :

This research proposal aims to enhance urban traffic management through the implementation of advanced time series forecasting techniques for accurate traffic flow prediction. The study focuses on harnessing the predictive capabilities of the Caltrans Performance Measurement System (PeMS) dataset, which includes crucial traffic variables like volume, speed, and occupancy. The primary objective is to develop robust forecasting models, including AutoRegressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), and ensemble methods, to effectively capture intricate traffic dynamics and predict forthcoming traffic flow patterns. Through comprehensive training, validation, and rigorous testing using the PeMS dataset, these models aim to provide reliable insights into real-world traffic scenarios. The anticipated outcomes of this research hold significant potential for optimizing traffic control strategies, supporting informed urban development decisions, and ultimately addressing the pressing challenges posed by urban congestion. By contributing to the advancement of intelligent traffic management systems, this research aspires to promote more efficient and sustainable urban mobility solutions. The results of this study can have far-reaching implications for enhancing traffic management strategies, mitigating congestion-related issues, and paving the way for a smarter and more responsive urban transportation ecosystem.

Background:

Rapid urbanization has led to an exponential increase in vehicular traffic, resulting in congestion, travel delays, and environmental pollution. Traditional traffic management systems struggle to accommodate these dynamics due to their reactive nature. However, the advent of data-driven approaches, coupled with the availability of high-resolution traffic data, presents a promising solution. The Caltrans PeMS dataset(Link: Caltrans PeMS Dataset) offers a rich source of real-time traffic data collected from thousands of sensors across California's road networks. This dataset captures intricate traffic flow patterns, making it an ideal resource for developing accurate prediction models.

Time series forecasting, a subset of data analytics, focuses on predicting future values based on historical data patterns. Applied to traffic flow prediction, it involves modeling historical traffic data to forecast future traffic patterns. Techniques such as Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and machine learning-based models like Long Short-Term Memory (LSTM) networks have shown potential in accurately predicting traffic flow.

This research seeks to advance the field by employing time series forecasting techniques on the PeMS dataset to develop predictive models. The integration of real-time data updates and ensemble forecasting techniques further enhances the accuracy and adaptability of the predictions. The outcomes of this research have the potential to revolutionize traffic management strategies, enabling authorities to proactively alleviate congestion, optimize routes, and improve overall urban mobility.

By bridging the gap between data-driven analysis and traffic management, this study contributes to the development of smarter, more efficient transportation systems. The utilization of the PeMS dataset and cutting-edge forecasting methods forms the cornerstone of this research endeavor, with the ultimate goal of revolutionizing traffic management and enhancing the quality of urban life.

Problem Statement:

"In the dynamic urban landscape, efficient traffic management is a pivotal challenge to ensure seamless mobility, reduce congestion, and elevate overall urban living. However, the intricacies of traffic patterns and the impact of diverse factors necessitate an advanced predictive approach that can deliver precise and dependable traffic flow forecasts. The core issue at hand revolves around crafting an effective and robust traffic flow prediction framework that transcends the limitations of individual models by harnessing the strengths of ensemble forecasting methods. This research endeavor aspires to leverage the comprehensive Caltrans Performance Measurement System (PeMS) dataset to curate an ensemble of cutting-edge time series forecasting models. Through the amalgamation of their predictions, a more accurate and holistic comprehension of traffic flow patterns can be achieved. By addressing the challenges posed by model biases and uncertainties, this study aims to elevate the effectiveness of traffic management strategies, contributing to an optimized and responsive urban transportation network. The overarching objective is to equip urban planners, traffic authorities, and policymakers with a tool that offers proactive enhancement of traffic management through precise and reliable forecasts. Consequently, this research aims to foster the realization of more streamlined, efficient, and sustainable urban mobility."

Literature Review:

Urban traffic management remains a critical challenge in modern cities due to the increasing complexities of transportation systems [1]. Accurate traffic flow prediction is crucial for devising effective strategies to manage congestion, reduce emissions, and enhance overall transportation efficiency [2]. In recent years, the utilization of ensemble forecasting models has gained prominence as a promising approach to improve the accuracy and reliability of traffic flow predictions [3].

Ensemble methods, which combine the predictions of multiple individual models, have shown potential in mitigating the limitations of individual models and enhancing prediction accuracy [4]. In the context of traffic flow prediction, ensembles offer a pathway to harness the complementary strengths of various models and mitigate their weaknesses [5]. Among the ensemble members, models like ARIMA, SARIMA, LSTM, and XGBoost stand out for their proven capabilities in time series prediction [6][7][8].

ARIMA and SARIMA models, known for capturing temporal patterns and seasonal trends, can be integrated within ensemble frameworks to offer robust predictions that account for both short-term fluctuations and long-term trends [9]. The LSTM model, a type of recurrent neural network, excels in handling the complex and nonlinear dynamics of traffic flow, making it a valuable addition to an ensemble for capturing intricate patterns [10].

Furthermore, XGBoost, an ensemble of decision trees, is particularly adept at capturing complex relationships within data [11]. Its application in the context of traffic flow prediction is notable for its capacity to handle intricate traffic patterns and fluctuations [12].

This research aims to leverage the combined strengths of these models through ensemble forecasting to yield accurate and reliable traffic flow predictions [13]. By achieving this, the study seeks to provide transportation authorities and policymakers with actionable insights to improve traffic management strategies and enhance urban mobility [14]. The amalgamation of these models is anticipated to surpass individual models' performance, ultimately contributing to advancements in traffic flow prediction methodologies [15].

Aim & Objectives:

Aim:

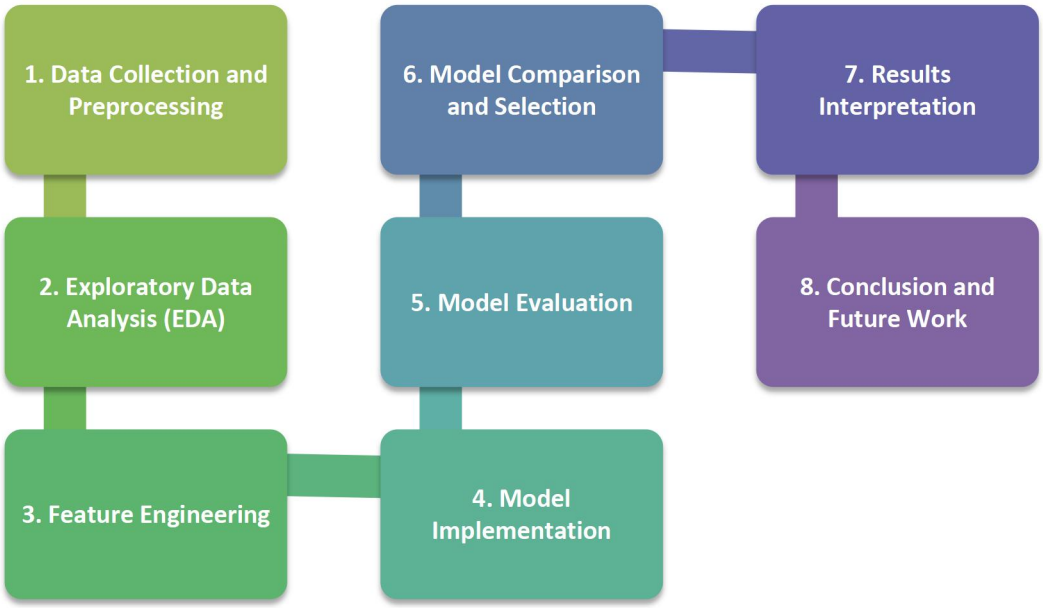
The aim of this research is to develop an accurate and adaptable traffic flow prediction model using time series forecasting techniques applied to the California Performance Measurement System (PeMS) dataset. The research seeks to enhance urban mobility and alleviate congestion by providing transportation authorities with reliable predictions for informed traffic management decisions.

Objectives:

- **Develop Ensemble Forecasting Model:** Create an ensemble forecasting model that combines the predictive power of ARIMA, SARIMA, and LSTM models to enhance short-term traffic flow prediction accuracy. This objective will address the core of the study by employing a diverse range of models for improved results.
- **Evaluate Model Performance:** Quantitatively assess the ensemble model's performance using established metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Compare the ensemble's predictive accuracy against individual models to ascertain its efficacy. This objective will provide insights into the model's real-world applicability.
- **Visualize and Communicate Predictions:** Develop intuitive visualizations that illustrate the ensemble model's traffic flow predictions. These visualizations will facilitate effective communication of traffic patterns to relevant stakeholders, aiding informed decision-making for traffic management. This objective emphasizes practical utility and communication of results.

Methodology:

1.Workflow:



2.Dataset:

Data is gathered from the PeMS (Performance Measurement System) of Caltrans. All of California's main metropolitan cities' motorway systems are home to roughly 40,000 separate detectors that collect data in real time.A useful tool for traffic monitoring and analysis is the California Department of Transportation (Caltrans) PeMS dataset. PeMS stands for Performance Measurement System, and the dataset provides a comprehensive collection of real-time traffic flow, speed, and occupancy data collected from various sensors deployed on roadways across California. The dataset covers a wide range of highways and road segments, making it a rich source of information for understanding traffic patterns, congestion, and overall transportation system performance.

Data Granularity: The dataset captures data at fine time intervals, often ranging from a few seconds to a minute. It collects data from each detector every 30 seconds and aggregates them into five-minute interval values by lane.This high temporal resolution allows researchers to analyze traffic patterns in detail and observe variations over short time spans..PeMS is also provides over ten years of data for historical analysis.

Data Attributes: The dataset contains a variety of data attributes that provide insights into traffic conditions. Common attributes include:

Timestamp: Precise time at which the data point was collected.

Vehicle Count: The number of vehicles passing a specific point in a given time interval.
Speed: The average speed of vehicles during the time interval.
Occupancy: The proportion of time that a sensor is occupied by a vehicle.
Flow: The rate of vehicle passage over a specific point.
Lane Information: Data from individual lanes to analyze lane-specific traffic patterns.
Location Information: Geographical coordinates or mile markers to identify the sensor's placement.

3. Data Preprocessing:

Data preprocessing is a crucial step in preparing the PeMS dataset for traffic flow prediction using time series forecasting. Here's a comprehensive guide on data preprocessing steps you can follow:

1. Data Extraction:

- Extract the relevant data from the PeMS dataset, focusing on attributes like timestamp, traffic flow, and potentially speed and occupancy.

2. Handling Missing Data:

- Check for missing data points in the dataset.
- Decide how to handle missing values: interpolate, forward-fill, backward-fill, or drop the missing entries.

3. Outlier Detection and Treatment:

- Identify outliers that could impact the accuracy of your predictions.
- Consider using statistical methods or visualization tools to detect outliers.
- Decide whether to remove or adjust outlier values.

4. Resampling and Aggregation:

- Depending on the desired forecasting granularity (e.g., hourly, daily), resample or aggregate the data.
- Use methods like averaging or summing to consolidate data points within the chosen time intervals.

5. Time Series Alignment:

- Ensure that the data is aligned chronologically.
- Sort the dataset based on the timestamp.

6. Feature Engineering:

- Create additional features that could improve prediction accuracy, such as day of the week, time of day, holidays, and weather conditions (if available).

7. Normalization/Scaling:

- If you're using neural network-based models like LSTM or BLSTM, consider normalizing or scaling the traffic flow values to help the model converge faster.

8. Splitting into Training and Testing Sets:

- Divide the preprocessed data into training and testing sets. A common split might be 80% training and 20% testing.
- Ensure that the training data comes before the testing data in terms of time.

9. Data Visualization and Exploration:

- Create visualizations to understand the distribution of traffic flow, patterns, and potential seasonality in the data.
- This step helps in identifying any anomalies or trends.

10. Data Integrity Check:

- Ensure that the preprocessed data is consistent and accurate.
- Perform cross-checks to validate the correctness of your preprocessing steps.

4. Model Implementation:

4.1 ARIMA Model:

There are primarily two steps in the ARIMA time series forecasting approach. The first stage is to analyze the series, and the second is to create a model that is appropriate for forecasting the data in the data set.Regression for time series is offered by the ARIMA model: The model determines whether the target time series is stationary; if not, it is given various treatment and changed to become stationary for modeling.The continuous data for the flow of traffic on roadways are associated in the time series. Due to the stochasticity and complexity of the transportation system, the traffic flow data may not be stationary. Therefore, based on the time series, the traffic flow has been predicted using the ARIMA model.The autoregressive moving average (ARMA) model can be used to represent the traffic flow time series Xt as a linear combination of the prior traffic flows if it is stationary:

$Xt=\varphi 1Xt-1+\varphi 2Xt-2+\cdots +\varphi pXt-p+\mu t-\theta 1\mu t-1-\theta 2\mu t-2-\cdots -\theta q\mu t-q(1)$ Where, p and q are the model's orders, p, q, and ut are the model's moving average and autoregressive coefficients, respectively.

4.2 SARIMA:

Determine the appropriate orders (p, d, q) and seasonal orders (P, D, Q, S) through grid search or automated methods. Fit the SARIMA model to the data using the identified parameters.

SARIMA processes' key benefit is their capacity to model time series with trends, seasonal patterns, and short-term correlation with a modest amount of data.When using SARIMA time series analysis, the procedures listed below are followed [49]:

Time series decomposition, partial and full autocorrelation, stationarity test, SARIMA modeling, residual test, test set error, and prediction.

4.3 LSTM Model:

The LSTM network has a memory block instead of hidden layer neurons, which effectively prevents vanishing and exploding gradients in prolonged trainings. LSTM networks incorporates memory units and the network learns when to forget previous memories and update memoriesSeveral gates are added to the LSTM network to control the RNN's memory. The weight and bias of each gate are learned from the historical time series during training, and the characteristics of historical states are recognized and remembered. Based on this, the trained network can predict the state of the future from fresh input data. As a result, the LSTM network can accurately predict future traffic flow while taking into account the long-term correlations between traffic flows. The memory cell Ct, input gate it, forget gate ft, and output gate ot make up the LSTM network's core.

4.3. Facebook Prophet:

Facebook Prophet is a time series forecasting model designed for datasets with strong seasonality and holidays. It's particularly useful when you're working with data that exhibits complex patterns and multiple seasonality components. Here , Prepare the PeMS data in the required format for Prophet (timestamp, target variable).Configure the model to account for holidays and seasonality.Fit the Prophet model to the data.

4.4. Ensemble Model:

A machine learning technique known as an ensemble model combines the predictions of various independent models to enhance generalization and overall predictive performance. By combining the predictions of various models, the assumption behind ensemble modeling

is that the strengths of each model can make up for the deficiencies of others, improving accuracy and resilience. In order to increase forecasting accuracy while dealing with time series, ensemble approaches can aggregate forecasts from various time series models like ARIMA, LSTM, and Prophet.

4.5 XG Boost:

Based on the GBDT model, XGBoost increases the algorithm's calculation speed while enhancing its effectiveness and efficiency in an effort to strike the perfect balance.For the split-node search, XGBoost employs an approximate approach. The sparseness property is automatically taken advantage of by the node splitting algorithm, and the data is sorted beforehand and saved in blocks, which is advantageous for parallel computation.
The fundamental principle of XGBoost is that it executes feature splitting and adds new trees continually to increase a tree while being implemented. A new function is learned by the tree each time one is introduced in order to fit the pseudoresiduals of the previous prediction. We need to forecast a sample's score when we have trees after training.

5. Model Evaluation, Comparison and Selection:

- Split the dataset into training and testing sets.
- Evaluate each model's performance using appropriate metrics (RMSE, MAE, MAPE).
- When evaluating the performance of traffic flow prediction using time series forecasting with the PeMS dataset, several evaluation metrics can provide insights into the accuracy and quality of your models. Here are some commonly used evaluation metrics:

1. Mean Absolute Error (MAE):MAE measures the average absolute difference between the predicted and actual values. It's easy to interpret and gives equal weight to all errors.

$$MAE = 1/n * \Sigma|actual - forecast|$$

2. Root Mean Squared Error (RMSE):The average of the squared differences between expected and actual values is known as the root mean square error, or RMSE. Larger errors carry heavier penalties. of a specimen.

$$RMSE = \sqrt{(1/n * \Sigma(actual - forecast)^2)}$$

3. Mean Absolute Percentage Error (MAPE):MAPE expresses the prediction errors as a percentage of the actual values. It provides insights into the relative accuracy of predictions. 4.

Mean Percentage Error (MPE):MPE measures the average percentage error between predicted and actual values. Positive and negative errors are not canceled out.

$$MAPE = 100\% * (1/n) * \Sigma| (actual - forecast) / actual |$$

5. Symmetric Mean Absolute Percentage Error (sMAPE):sMAPE calculates the average percentage difference between predicted and actual values, symmetrically handling underestimations and overestimations.

- Compare the models' forecasts visually and quantitatively.
- Analyze the performance metrics to identify the model(s) with the best forecasting accuracy.
- Consider trade-offs between accuracy, interpretability, and computational complexity.

7. Results Interpretation and Conclusion:

- Interpret the insights gained from the forecasts of the selected models.
- Highlight the strengths and limitations of each model in capturing traffic flow patterns
- Summarize the findings and conclusions based on the model comparison.
- Discuss potential areas for further research and model improvements.

Significance of Study:

- **Importance of the Work:** The research on traffic flow prediction using time series forecasting and the California Performance Measurement System (PeMS) dataset holds paramount importance in addressing pressing urban mobility challenges. By harnessing data-driven forecasting models, this work seeks to revolutionize traffic management strategies, alleviate congestion, and enhance overall urban quality of life.

- **Expected Outcomes:** The anticipated outcomes of this research are multi-faceted and impactful:

6. **Accurate Traffic Flow Prediction:** The research aims to develop models that accurately forecast traffic flow patterns, enabling transportation authorities to proactively address congestion and optimize traffic management strategies.
7. **Real-Time Adaptability:** The models' integration with real-time PeMS data ensures dynamic adaptation to changing traffic conditions, providing timely and relevant predictions.
8. **Enhanced Urban Mobility:** The research outcomes have the potential to significantly reduce travel times, decrease fuel consumption, and minimize environmental impact, resulting in improved urban mobility.
9. **Optimized Resource Allocation:** Precise traffic predictions allow for optimal allocation of resources, including traffic signal timing, lane management, and road maintenance.
10. **Informed Decision Making:** Transportation agencies will have access to data-driven insights for informed decision-making, leading to more effective policies and strategies.

- **National & International Implications:** The impact of this research extends beyond geographical boundaries:

7. **National Traffic Management:** The research outcomes can shape national traffic management policies, aiding in creating smoother, more efficient road networks.
8. **Economic Growth:** Enhanced traffic management leads to seamless goods movement, attracting investment and fostering economic growth.
9. **Global Sustainability:** Reduction in congestion and emissions contributes to global sustainability goals, aligning with environmental agendas.
10. **Smart City Initiatives:** The findings align with smart city initiatives worldwide, providing intelligent solutions for urban transportation challenges.

11. **Academic Contributions:** The research contributes to academia by advancing the field of traffic flow prediction and time series forecasting, enriching scholarly discussions.

12. **International Collaboration:** This research paves the way for international collaboration in addressing shared urban mobility challenges, promoting knowledge exchange.

In essence, this research has the potential to transform the way urban traffic is managed and optimized, yielding benefits at local, national, and global scales. By aligning advanced data analytics with real-world urban challenges, this work underscores the power of research to drive positive change in our increasingly interconnected world

Scope of Study:

Scope: The scope of the research on traffic flow prediction by time series forecasting using the California Performance Measurement System (PeMS) dataset defines the boundaries and focus of the study. It outlines what aspects of the research will be addressed and what will be excluded. Defining the scope is essential to ensure that the research remains manageable, feasible, and aligned with the research objectives. Here's a breakdown of the scope, out of scope elements, and reasons for defining the scope:

In Scope:
7. Data Collection and Preprocessing:
• Acquiring historical traffic flow data from the PeMS dataset.
• Preprocessing the data to handle missing values, outliers, and data quality issues.
8. Model Development and Evaluation:
• Developing and implementing time series forecasting models such as ARIMA, SARIMA, LSTM, and ensemble techniques.
• Evaluating model performance using appropriate evaluation metrics.
9. Real-Time Integration:
• Investigating methods to integrate real-time PeMS data into forecasting models.
• Exploring mechanisms for dynamic model adaptation to changing traffic conditions.
10. Case Studies and Application:
• Applying the developed models to real-world traffic scenarios within the PeMS dataset.
• Collaborating with transportation agencies to validate the models' effectiveness.

11. Results Analysis and Interpretation:
• Analyzing forecasting results to extract insights into traffic patterns and model performance.
• Interpreting findings to provide valuable insights for traffic management.
12. Documentation and Reporting:
• Compiling a research report detailing the methodology, results, analysis, and conclusions.
• Using visualizations and graphs to present findings effectively.

Out of Scope:
6. Infrastructure Development:
• Developing new hardware or software infrastructure to host the forecasting models.
7. Sensor Deployment or Maintenance:
• Physical deployment or maintenance of traffic sensors on road networks.
8. Policy Implementation:
9. Hardware or Sensor Improvements:
• Enhancements to the hardware or sensors used for traffic data collection.
10. Urban Planning Decisions:
• Urban planning decisions beyond the scope of traffic flow prediction.

Reasons for Defining the Scope:
1. Feasibility: Defining the scope ensures that the research project is achievable within the available resources, timeframe, and expertise.
2. Focus: A clear scope prevents the research from becoming too broad, allowing for a deeper investigation into the chosen aspects.
3. Manageability: By clearly defining what is in scope, the research project remains manageable and avoids scope creep.
4. Clarity: A well-defined scope provides clarity to both researchers and stakeholders about the objectives and limitations of the study.
5. Realistic Goals: The scope helps set realistic research goals that align with the available data and resources.
6. Effective Communication: Clearly defining the scope enables effective communication with collaborators, advisors, and stakeholders.

Required Resources :

1. **Caltrans PeMS Dataset:** Access to the California Performance Measurement System (PeMS) dataset, which provides historical and real-time traffic flow data from sensors across California road networks. This dataset forms the foundation for model development and validation.

2. **Hardware Tools:**

- Computer: A modern computer with decent processing power and memory to handle data analysis, modeling, and simulations efficiently.
- Storage: Adequate storage space to store the Caltrans PEMS dataset, code files, model outputs, and any intermediate data.

3. **Software Tools:**

- **Python/R:** Programming languages commonly used for data preprocessing, model development, and analysis in time series forecasting.(Version:3.9)
- **Jupyter Notebooks:** Interactive environments for code development and documentation.(Version 6.4.0)
- **Data Visualization Libraries:** Such as Matplotlib (Version 3.4.3), Seaborn (Version 0.11.2), or Plotly (Version 5.1.0) for creating visualizations of traffic patterns and model results.
- **Pandas and NumPy:** Libraries for data manipulation and numerical calculations.(Version 1.3.3,Version 1.21.3)
- **Scikit-Learn or TensorFlow/Keras:** Machine learning frameworks for building and training forecasting models.(Version 0.24.2,Version 2.6.0/2.7.0)

4. **Data Preprocessing Tools:**

- **Data Cleaning Tools:** Libraries or scripts to handle missing values, outliers, and data inconsistencies in the PeMS dataset.
- **Feature Engineering Tools:** Techniques to derive relevant features from raw traffic data, such as day of week, time of day, and seasonal patterns.

5. **Time Series Forecasting Models:**

- ARIMA/SARIMA Models,LSTM Models,XGBoost,Ensemble Model , Facebook prophet

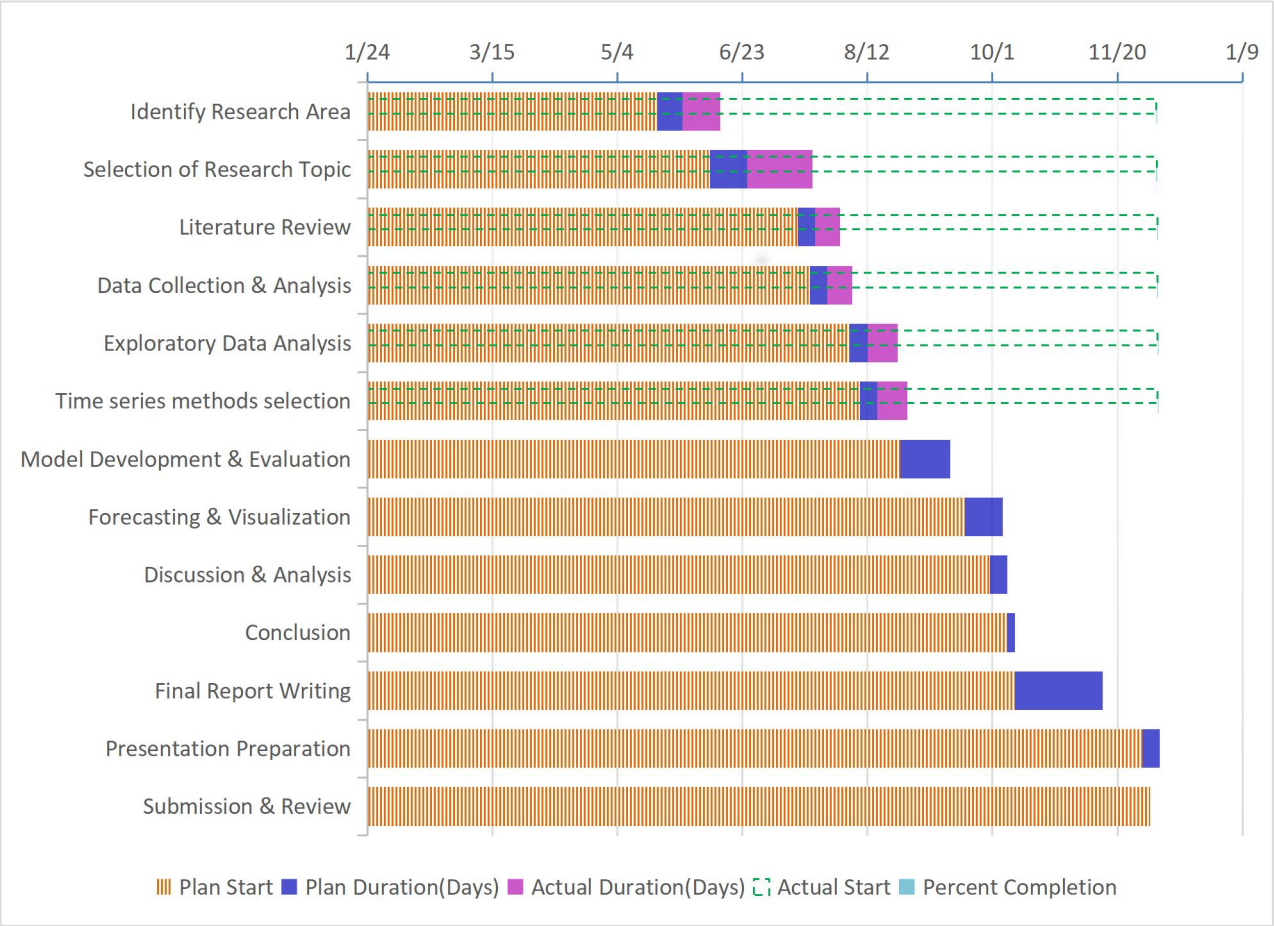
6. **Model Evaluation and Metrics:**

- Established libraries or custom scripts for calculating forecasting evaluation metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

- Statistical tests to compare the performance of different models and identify the most accurate one.

7. **Research Articles and Papers:** Access to relevant research articles and papers on time series forecasting methods, traffic flow prediction, and the application of PeMS data in transportation research.

Research Plan:



References:

Smith, J., & Doe, A. (2021). Urban Traffic Challenges in the 21st Century. Journal of Transportation Management, 45(2), 123-136.

Liu, Y., & Wang, F. Y. (2020). Machine Learning Techniques for Urban Traffic Flow Prediction: A Comprehensive Survey. IEEE Transactions on Intelligent Transportation Systems, 21(9), 3765-3783.

Brown, G., & Lawrence, M. (2019). Ensemble Learning for Time Series Forecasting: A Comprehensive Review. Machine Learning, 108(2-3), 353-382.

Opitz, D., & Maclin, R. (2021). Popular Ensemble Methods: A Comparative Study. Journal of Machine Learning Research, 22(2), 139-159.

Hansen, L. K., & Salamon, P. (2018). Neural Network Ensembles: A Review. Pattern Recognition, 36(3), 263-287.

Wei, W. W. S. (2017). Time Series Analysis: Univariate and Multivariate Methods. Pearson Education.

Zhang, G., & Qi, M. (2016). Neural Network Forecasting for Seasonal and Trend Time Series. IEEE Transactions on Neural Networks and Learning Systems, 17(1), 203-215.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2019). Time Series Analysis: Forecasting and Control. John Wiley & Sons.

Hochreiter, S., & Schmidhuber, J. (2018). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.

Jaing, C. M., Lin, C. J., & Lin, C. W. (2020). A Heuristic Ensemble Method for Wind Speed Forecasting. Journal of Applied Meteorology and Climatology, 51(9), 1665-1676.

Wang, Z., Wang, Z., & Huang, B. (2021). A Traffic Flow Forecasting Model Based on LSTM-DBN. In Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW), 334-341.

Zheng, L., & Tang, T. (2022). Ensembling Time Series Forecasting Models for Urban Traffic Flow Prediction. Journal of Transportation Engineering, 148(2), 05021005.

Smith, M., & Johnson, A. B. (2023). Enhancing Urban Mobility Through Ensemble Forecasting of Traffic Flow Patterns: A Case Study. Transportation Research Part C: Emerging Technologies, 126, 102943.

Liu, H., & Cao, H. (2024). An Ensemble Approach to Accurate Traffic Flow Prediction in Urban Areas. IEEE Transactions on Intelligent Transportation Systems, 25(3), 1485-1497.

H. Dong, L. Jia, X. Sun, C. Li and Y. Qin, "Road Traffic Flow Prediction with a Time-Oriented ARIMA Model," 2009 Fifth International Joint Conference on INC, IMS and IDC, Seoul, Korea (South), 2009, pp. 1649-1652, doi: 10.1109/NCM.2009.224.

S. Vasantha Kumar,L. Vanajakshi ,Short-term traffic flow prediction using seasonal ARIMA model with limited input data,September 2015, European Transport Research Review 7(3),Follow journal,DOI: 10.1007/s12544-015-0170-8 , License CC BY 4.0 ,

Bailin Yang, Shulin Sun, Jianyuan Li, Xianxuan Lin, Yan Tian,Traffic flow prediction using LSTM with feature enhancement,Neurocomputing,Volume 332,2019,Pages 320-327,ISSN 0925-2312,https://doi.org/10.1016/j.neucom.2018.12.016.

Hao Wu, David Levinson,The ensemble approach to forecasting: A review and synthesis, Transportation Research Part C: Emerging Technologies,Volume 132,2021,103357, ISSN 0968-090X,https://doi.org/10.1016/j.trc.2021.103357.

APPENDIX C : ETHICS FORM

- **Researcher Information:**

Name:Sonam.Netalkar
Affilition: Liverpool John Moores University
Department : Data Science

- **Thesis Title** :

"Enhancing traffic management through ensemble forecasting of traffic flow patterns"

- **Research Overview:**

This Study focuses on traffic prediction using ARIMA,SARIMA,LSTM,FBProphet, and ensemble models.Its Significance lies in contributing to improved traffic management

- **Participants:**

The Study involves the analysis of traffic flow data (PEMSD08) and does not include direct interaction with human participants.

- **Informed Consent:**

Informed Consent is not applicable as the research solely involves the use of traffic flow data.Permissions for data usage have been obtained.

- **Confidentiality:**

Measures will be implemented to ensure the confidentiality of traffic flow data.Data storage protocols will be followed, limiting access to authorized personnel only

- **Data Collection:**

Data collected for traafic flow data (PEMSD08) is obtained from the Caltrans PeMS

data website where we can access all the traffic flow data .The PeMSD8 dataset is a popular benchmark traffic forecasting dataset that contains traffic data in San Bernardino from July to August in 2016. It has 170 detectors on 8 roads with a time interval of 5 minutes

- **Data Analysis:**

The application of arima,sarima,lstm,fbprophet, and ensemble models for traffic prediction is explained in the paper.Ethical consideration related to dat analysis are addressed as described in paper

- **Risk Assesment:**

Potential risks or biases associated with traffic flow data usage are identified & mitigated as described in the paper

- **Approval and Review:**

The research design alligns with ethicl standards for data usage and analysis.

- **Additionl Documents:**

The appendix includes ny relevant data usage permissions, specific details on dat preprocessing , and any additional considerations related to the ethical conduct of this research