IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 29, No. 5, October, 2024, pp. 1903-1913

Journal homepage: http://iieta.org/journals/isi

Optimizing Urban Mobility: A Comparative Analysis of Taxi Demand Prediction Models

Ragil Saputra[1,2], Suprapto[2*], Agus Sihabuddin[2]

[1] Department of Computer Science, Universitas Diponegoro, Semarang 50275, Indonesia
[2] Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia

Corresponding Author Email: sprapto@ugm.ac.id

https://doi.org/10.18280/isi.290522

Received: 15 January 2024
Revised: 10 September 2024
Accepted: 21 September 2024
Available online: 24 October 2024

Keywords:
urban mobility, taxi demand prediction, feature engineering, Fourier transform, Random Forest, XGBoost

ABSTRACT

Urban mobility optimization is crucial in managing transportation systems efficiently. This study addresses a broad research area of urban mobility by focusing on taxi demand prediction, a key component of the transportation ecosystem. The specific problem addressed in this research is the need for accurate and efficient taxi demand prediction, especially in large, dynamic urban environments. Existing solutions, including basic time series approaches like simple moving averages and exponential weighted moving averages, while valuable, have limitations in handling the intricacies of urban taxi demand patterns. In this study, we employed a combination of data preprocessing techniques, advanced regression models, and Fourier features to predict taxi demand in dynamic urban environments. The data preprocessing techniques included data cleaning, normalization, and feature engineering. The advanced regression models used in this study were Random Forest and XGBoost, which were trained and tested using NYC taxi datasets. The Fourier features were used to capture the periodicity of the taxi demand patterns. These models are demonstrated to outperform standard solutions, effectively achieving the targeted mean absolute percentage error (MAPE) of less than 12%. Evaluation of the solution revealed its effectiveness in reducing the prediction error by more than 1%, thus highlighting the positive results of this research.

1. INTRODUCTION

Enhancing urban mobility is crucial in the continuously developing realm of transportation systems. As urban populations grow rapidly, the ability to predict taxi demand accurately and efficiently has become a key element in managing urban mobility. Improvements in predictive models and techniques are poised to substantially improve transportation services in urban centers.

In recent years, a substantial body of studies has been aimed at taxi demand prediction. Notably, studies by Rodrigues et al. [1] have explored various approaches encompassing machine learning and time series analysis to forecast taxi demand.

Chou et al. [2], Liu et al. [3], and further research conducted using Random Forest [4-6] and XGBoost [7, 8] have shown considerable advancements in the precision of predictions. These models have enhanced the predictive accuracy but still often rely heavily on large quantities of historical data, which may not always be available or reflect future conditions accurately.

Research conducted by Liu et al. [4] using Random Forest and Stadler et al. [9] using XGBoost regressor to predicts passenger demand for taxi drivers based on trip fare, distance between each region, and area of the region. Random Forest model is an algorithm for ensemble learning that builds upon bagging [10]. XGBoost, or eXtreme Gradient Boosting, is an open-source project in machine learning developed by Tianqi Chen that enhances the boosting technique originally based on GBDT [11].

Xu et al. [12] investigated the application of Recurrent Neural Networks (RNNs) for predicting taxi demand, attaining an accuracy rate close to 83%. In a similar vein, Kuang et al. [13] utilized data augmentation alongside convolutional neural networks (CNNs) to forecast short-term demands for taxis.

Despite considerable advancements in predictive accuracy by studies such as those by Rodrigues et al. [1] and Kuang et al. [13], existing models often rely heavily on large quantities of historical data and struggle to adapt to rapid changes in urban dynamics. Such models typically do not account for non-linear fluctuations in taxi demand influenced by unpredictable factors like weather or special events.

Furthermore, while sophisticated deep learning models like those introduced by Zhang et al. [14] and Ye et al. [15] offer in-depth analysis of spatiotemporal patterns, they require extensive computational resources, limiting their practical application in real-time urban settings. To overcome these limitations, our research integrates Fourier features with advanced regression models, particularly Random Forest and XGBoost, enhancing our model's ability to adapt to both regular and irregular patterns in taxi demand. This approach allows for a more robust prediction system that is responsive to real-time changes and less dependent on historical data, thereby filling a critical gap in current urban mobility management strategies.

Despite these advancements, significant challenges persist. Previous studies often focus on isolated aspects of demand prediction without a comprehensive comparison of different approaches. This has left a gap in understanding which methods are most effective under varying conditions. Additionally, while basic time series models such as Simple Moving Averages (SMA) and Exponential Weighted Moving Averages (EWMA) are effective in some scenarios, their limitations in handling the dynamic and complex nature of urban taxi demand patterns are well-noted [16].

To address these gaps, this study incorporates Fourier features and advanced regression models, particularly Random Forest and XGBoost, to enhance prediction accuracy. Fourier transform, a basic mathematical tool, proves instrumental in analyzing periodic patterns within datasets and discerning recurring temporal patterns such as daily or weekly cycles in historical taxi demand data [7, 17].

In this study, we selected Random Forest and XGBoost due to their robustness in managing large, complex datasets typical of urban mobility contexts. Random Forest is effective in preventing overfitting through its ensemble approach, which is crucial for modeling non-linear data influenced by unpredictable variables such as weather and special events. However, the reference provided [18] does not support this claim as it focuses on drought modeling rather than the specific strengths of Random Forest in urban mobility contexts.

Our methodology involved collecting and cleansing a large dataset of taxi demand records [19], followed by preprocessing and feature engineering to extract important spatial and temporal features. We then compared the performance of various regression models, including traditional time series models and advanced models, providing a comparative analysis to optimize urban mobility.

The rest of this article is organized as follows: section 2 details the methodology, section 3 presents the experiment results and discusses the findings, and suggests future directions. Section 4 concludes by providing a summary of the main findings.

2. METHODOLOGY

In this section we will present a research stages that is conducted sequentially. The research begins with data collection, preprocessing, feature engineering, modelling and ends with evaluation model. Figure 1 displays all stage of the proposed method. Our contributions is improve prediction of taxi demand by using feature engineering after preprocessing step by Fourier transform. Fourier Transform to identify repeating patterns and periodic components in taxi demand data.

In Figure 1, the modeling stage is divided into two parts. part 1 uses *SMA*, *WMA*, and *EWMA* models, while part 2 uses Linear Regression, Random Forest, and XGBoost models. Both part 1 and part 2 receive input from the preprocessing stage, while part 2 also receives additional input from feature engineering, which is performed after preprocessing. The evaluation stage involves calculating the *MAPE*.

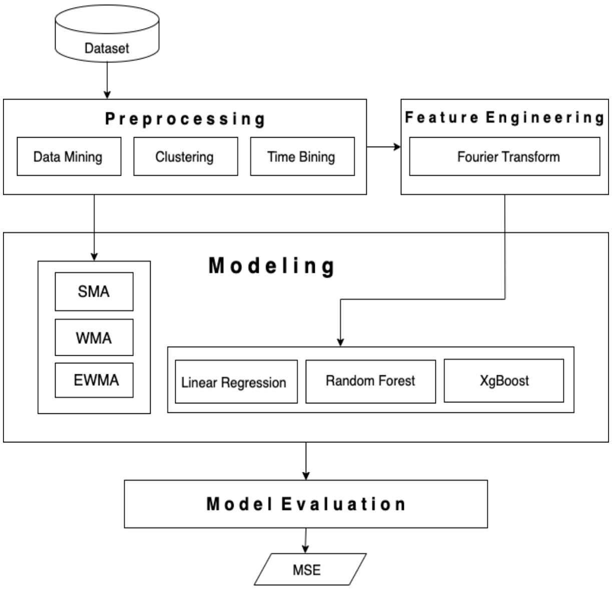


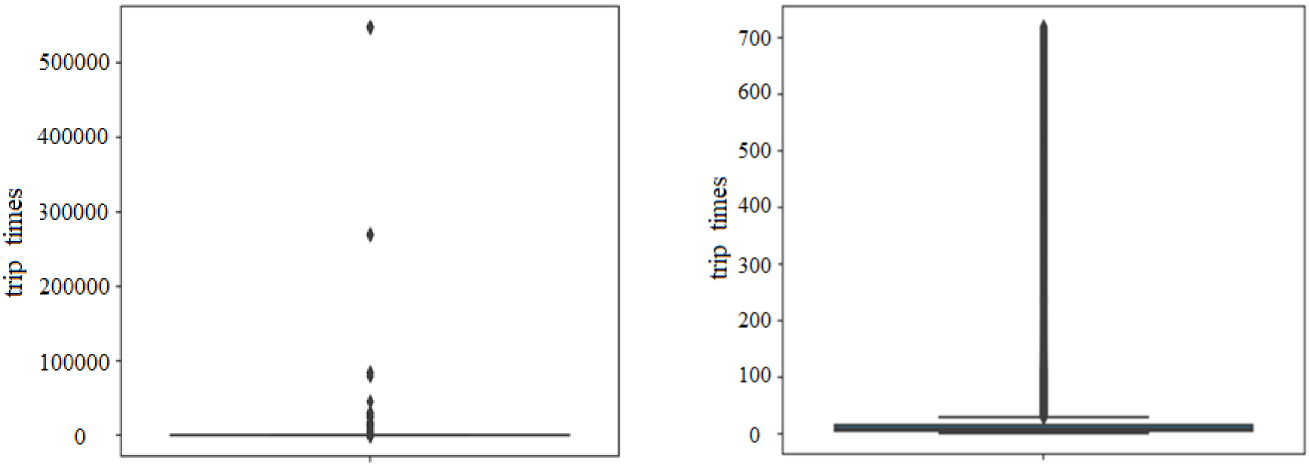
Figure 1. Proposed method

2.1 Data collection

The research makes use a comprehensive dataset from New York City, encompassing taxi demand records from January to March 2015, and January to March 2016. This data is released by the Taxi and Limousine Commission from [19]. This dataset is a valuable source of information regarding the latitude, longitude, and timestamp of each taxi pickup. Description of dataset is presented in Table 1. The resulting dataset forms the foundation for our taxi demand prediction models, offering insights into historical pickup patterns across different locations within the city.

Table 1. Dataset structure [19]

Field Name	Description
vendorID	A code that identifies the record's TPEP provider. 1: CMT / Creative Mobile Tech 2: VeriFone Systems Inc
tpepPickupDatetime	The specific date and time while the timer activated.
tpepDropOffDatetime	The specific date and time while the timer was disconnected.
passengerCount	The amount of passenger in the taxi. The amount was entered by the driver.
tripDistance	The taximeter's recorded elapsed trip distance in miles.
pickup_longitude	Longitude where the timer activated.
pickup_latitude	Latitude where the timer activated.
Rate_Code	The last amount symbol in impact at the finish at the trip. 1: Standard rate; 2: JFK; 3: Newark; 4: Nassau or Westchester; 5: Negotiated fare; 6: Shared ride
Store_and_forward_flag	This indicator specifies if the trip record was "stored and forward," or stored in the vehicle's memory prior to being transmitted to the vendor, in the event that the car was not connected to the server. Y: trip stored and forwarded. N: trip not stored or forwarded.



(a) Distribution of trip times in dataset

(b) Trip times after removing outlier

Figure 2. Trip times

2.2 Data cleaning

The data were cleaned by performing univariate analysis and dropping outlier values that could have been made due to an error. To ensure the quality and reliability of the data, preprocessing steps were undertaken. Outliers and missing values were addressed through appropriate techniques, and the data was cleaned to remove any inconsistencies.

In our preprocessing workflow, outliers were systematically identified and managed using robust statistical methods. We utilized the Interquartile Range (IQR) approach to detect outliers, where values falling more than 1.5 IQRs below the first quartile or above the third quartile were flagged for review. Depending on their impact on the model's predictive power and the likelihood of them representing true anomalies versus data errors, outliers were either adjusted or removed.

For data normalization, we employed the Min-Max scaling technique which adjusts the data to a common scale by transforming each feature to a range between 0 and 1. This normalization is crucial for maintaining consistency across input features and enhances the efficiency of the learning algorithm, especially when combining features with differing units and ranges.

2.2.1 Coordinate

New York city is bounded by the latitude and longitude coordinates (40.5774, -74.1500) and (40.9176, -73.7004) [20]. As a result, we only take into account pickups that originate in New York city, and we do not consider any coordinates that are outside of these ranges.

2.2.2 Trip duration

NYC Taxi & Limousine Commission regulations state that a trip may last no more than 12 hours in a 24-hour period [19]. Thus, those data points with trip duration more than 720 minutes were removed.

2.2.3 Trip time

We calculate the trip time by subtracting the pickup timestamp of the dropoff timestamp and divide the result by 60 to express it in minutes. The skewed box plotted visually in Figure 2(a) represents the distribution of trip times in the dataset. We systematically removed outliers from the taxi trip time dataset with univariate analysis. Percentiles of trip_times were used to identify extreme values. The range between 1 and 720 minutes was considered, following compliance with TLC regulations, to eliminate potential outliers. So the box plotted after removing outliers is presents in Figure 2(b).

2.2.4 Trip distance

To identify any aberrations in the dataset, a box plotted was employed to provide a visual representation of the distribution of trip distances. The primary objective is to pinpoint data points that deviate notably from the typical range of trip distances, which could be indicative of outliers. Figure 3(a) represents the distribution of trip distance in the dataset.

Figure 3(b) presents box plotted after removing outlier. We use interpercentile range (IPR) to calculate the percentile value of trip distance in the dataset, and find that the 99.9 percentile has a value of 22.57 miles. Any value beyond this point is considered an outlier and may significantly affect the quality and accuracy of the dataset. Therefore, we consider only the values between 0 and 23 miles.

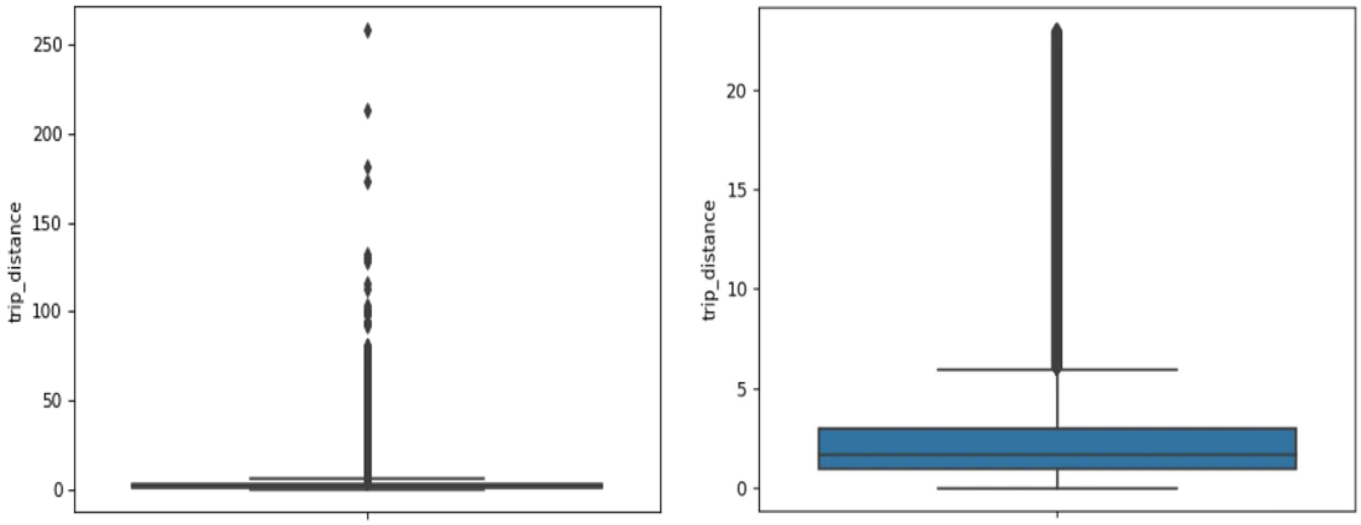
2.2.5 Speed

After removal of outliers in trip duration, we proceed to examine the distribution of the speed feature. This feature characterizes the speed of a taxi trip, calculated as the ratio of the distance traveled to the duration of the trip, with the result multiplied by 60 to express the speed in miles per hour. Box plots in Figure 4 allow identification of outliers or any extreme values in the Speed feature, this may significantly affect the data set's quality and accuracy.

We use the IPR to remove speeds with outliers, thus we adopt the percentile method. In this method, we find 0–100th percentile values to detect beyond what point the outliers occur. Finally, we filtered out records with speeds less than 0 or greater than 45.31 miles per hour.

2.2.6 Total fare

An examination of the dataset's fare values concentrated on the last 50 data points, excluding the last two, revealed a significant surge in value at around the 1000 fare value. This focused analysis provides insights into potential anomalies or irregularities in the dataset, ensuring data consistency and reinforcing the research's overall accuracy and credibility. So, the fare is less than or equal to 0 or greater than or equal to 1000. The distribution of fare in dataset is presented in Figure 5.



(a) Distribution trip distance in dataset

(b) Trip distance after removing outlier

Figure 3. Trip distance

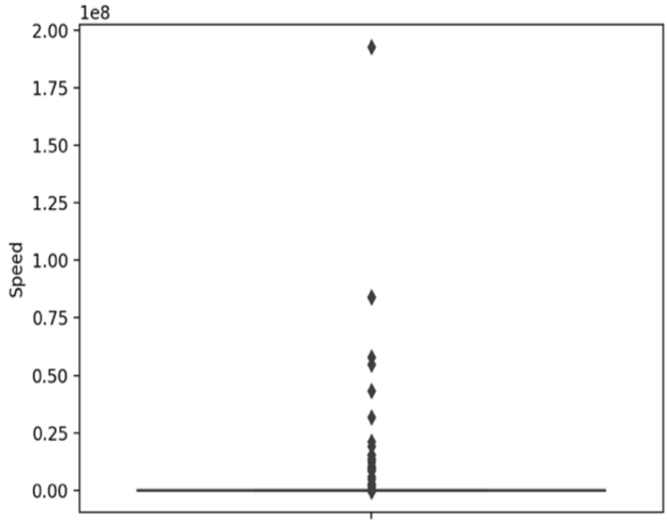


Figure 4. Distribution of speed in dataset

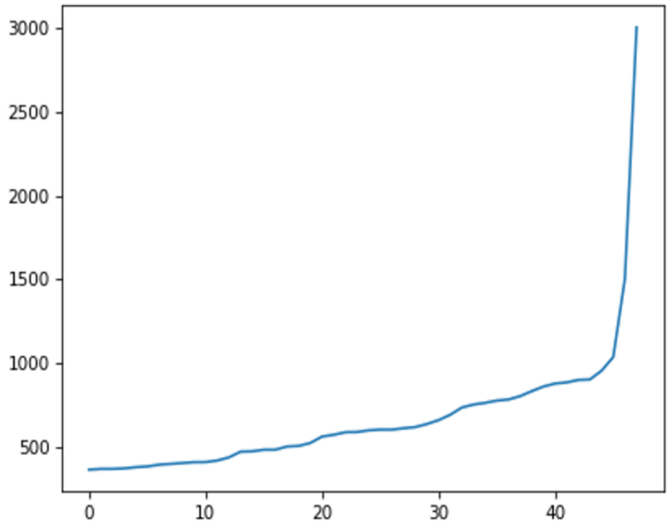


Figure 5. Distribution of fare in dataset

2.3 Data preprocessing

This stage begins with clustering, continued with organizing and aggregating the data using time-binding approach.

2.3.1 Clustering

In this study, a clustering process was carried out using the K-means algorithm, used to identify the optimal number of clusters (K), that contributes a significant component in spatial data analysis. The dataset used here consists of the GPS coordinates of taxi pick-up locations, and the goal is to group these coordinates into clusters for further analysis. The algorithm is designed to systematically evaluate different cluster sizes by performing iterations from 10 to 90 with increments of 10.

Cluster evaluation to assess the quality of the clusters based on two key factors: (1) the average number of clusters inside the area where the inter-cluster the distance is less than 2 units. This indicates how tightly the clusters are formed, and (2) the average number of clusters outside the area where the inter-cluster distance is greater than 2 units. This reflects the spread of clusters.

Furthermore, the minimum inter-cluster distance is calculated. It represents the minimum spatial separation between any two clusters. This metric is essential to ensure that the clusters are well-separated. The clustering results are presented in Figure 6.

Figure 6 illustrates an evaluation that was conducted for various cluster sizes (K) in cluster analysis. The graph demonstrates that as the cluster size (K) increases, the average number of clusters both within and outside a specific region also rises. However, on the flip side, the minimum distance between clusters decreases. This implies that larger K values lead to more complex and dispersed clusters. Therefore, the selection of the K value should consider a balance between the number of clusters and the distance between them, aiming to find a K value that yields sufficiently concentrated clusters within a certain region while maintaining reasonable inter-cluster distances. Because the main goal is to choose the optimal minimum distance, so our optimal number of clusters is 30. The plotting results in map form are present in Figure 7.

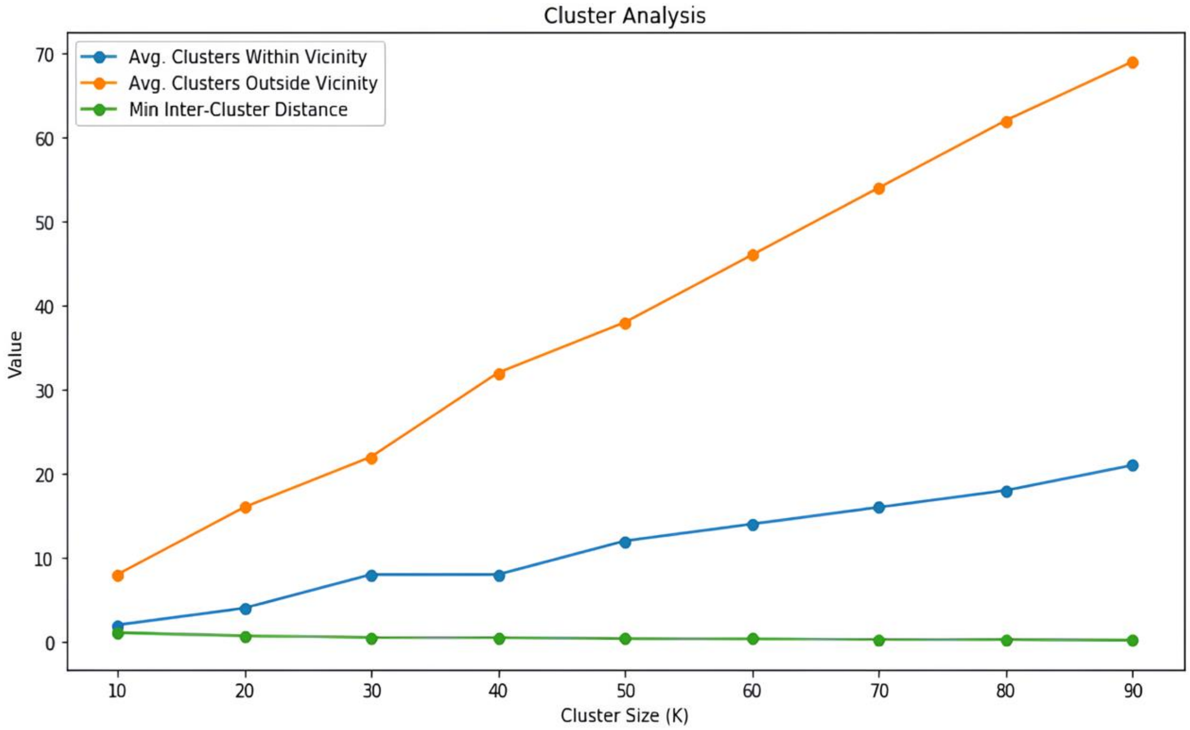


Figure 6. Cluster analysis

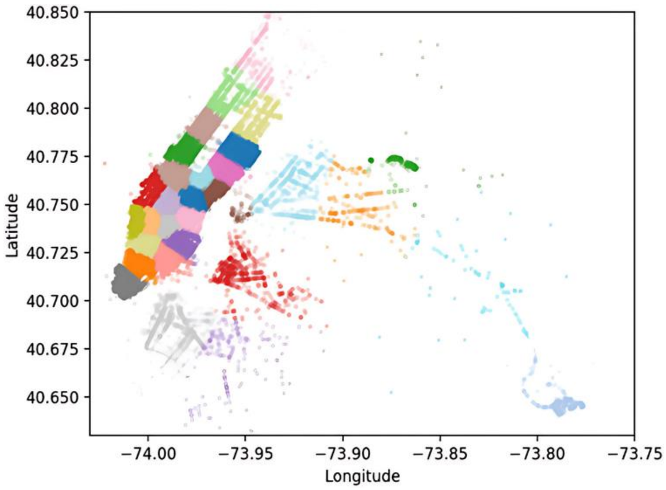


Figure 7. Clustering result plot

2.3.2 Time binning

In the context of our study for the month of January 2015, we employed a time-binning approach to organize and aggregate the data, specifically the number of taxi pickups that occurred within 10-minute intervals. The resulting dataset was structured with two key indices. The primary index pertained to the pickup cluster, indicating the specific cluster to which a pickup location was assigned. These clusters were derived using k-means clustering techniques based on the geographical coordinates of the pickup locations. The secondary index was known as pickup bins. For example, pickup bins data for each cluster are shown in Table 2.

In the context of predicting taxi demand for the year 2016, our data preparation efforts extended to the early months of the year, encompassing January, February, and March. We meticulously executed a series of steps to ensure the datasets were optimized for in-depth analysis. Initially, we carefully selected and filtered the relevant columns from the raw data, subsequently enhancing the dataset by incorporating significant trip-related attributes such as trip durations, speeds, and Unix timestamps of pickup times.

Table 2. Pickup bins each cluster

	Pickup_ Longitude	Pickup_ Latitude	Pickup_ Cluster	Pickup_ Bins
0	-73.993896	40.750111	14	2130
1	-74.001648	40.724243	25	1419
2	-73.963341	40.802788	8	1419
3	-74.009087	40.713818	21	1419
4	-73.971176	40.762428	28	1419

Building on our prior work involving spatial clustering, we retained our approach to assign each trip to specific clusters based on their pickup locations. This clustering method effectively grouped trips that shared similar geographical characteristics, enabling the exploration of spatial patterns. We also introduced the concept of pickup_bins, which represented 10-minute intervals within a day, aiding in the temporal segmentation of the data. Consequently, we derived two invaluable datasets for each of the target months in 2016, furnishing detailed trip-level information along with the clustered, time-segmented data. These datasets empowered us to uncover cluster-specific taxi demand patterns over time and enhance our predictions regarding urban mobility and taxi service demand throughout 2016.

Given that there are 24 hours in a day, 31 days in January, and each hour consists of 60 minutes, there were a total of 4,464 unique time bins created for this temporal segmentation. The visual representation of our findings provides a comprehensive insight into the temporal dynamics of taxi service demand, as presented in Figure 8. In Figure 8, we present cluster 1st representative plot among a collection of 30 cluster. This illustrative plot offers a comprehensive insight into the temporal dynamics of taxi service demand within a specific geographic area.

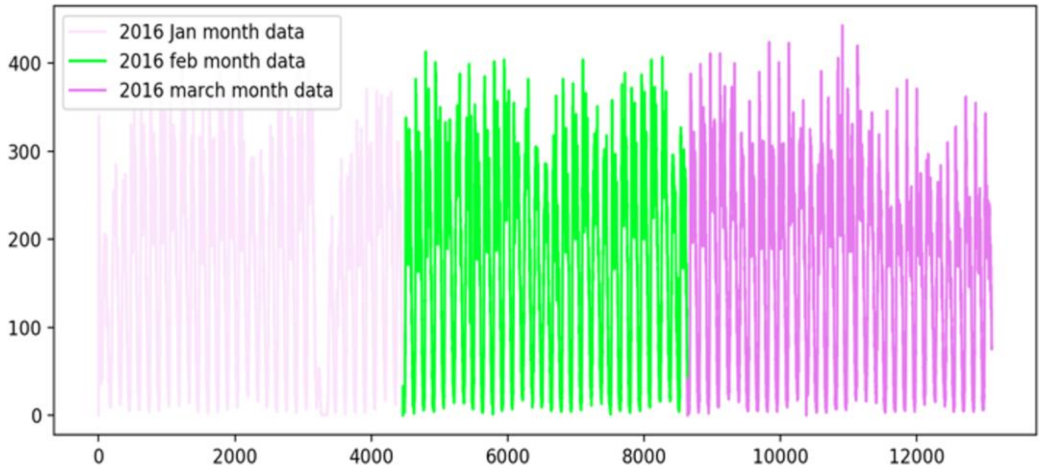


Figure 8. Plot temporal dynamics for 1st cluster

2.4 Modelling approaches

Our methodology comprises a variety of prediction models, including baseline models and advanced regression techniques. The baseline models, which include SMA, WMA, and EWMA, serve as the foundational benchmarks for evaluating the performance of more sophisticated models. These basic models are inspired by traditional time series forecasting techniques.

SMA model is the first one to be employed, it forecasts the next value by utilizing the n previous values. Ratio value using in Eq (1).

$$R_t = \frac{R_{t-n} + R_{t-n+1} + \cdots + R_{t-2} + R_{t-1}}{n} \tag{1}$$

Next, we use Eq. (2) to forecast the future value using the Moving Averages of the 2016 values itself.

$$P_t = \frac{P_{t-n} + P_{t-n+1} + \cdots + P_{t-2} + P_{t-1}}{n} \tag{2}$$

WMA are used in the second model. All of the data in the window were given equal weight by the Moving Averages Model, but we know deep down that the most recent values will probably be more comparable to the future than the earlier values. WMA with ratio values from Eq. (3).

$$R_t = \frac{\sum_{i=0}^{n-1}(n-1) \times R_{t-i-1}}{n \times \left(\frac{n+1}{2}\right)} \tag{3}$$

Use Eq. (4) to calculate WMA based on prior 2016 data.

$$P_t = \frac{\sum_{i=0}^{n-1}(n-1) \times P_{t-i-1}}{n \times \left(\frac{n+1}{2}\right)} \tag{4}$$

EWMA are employed in the third model. We have met the analogy of assigning greater weights to the most recent value and decreasing weights to the subsequent ones through weighted averages, but we are still unsure of the best weighting scheme due to the infinite number of ways we can adjust the hyperparameter window-size and assign weights in a non-increasing order.

We utilize a single hyperparameter, α, for exponential moving averages. Its value ranges from 0 to 1, and the weights and window sizes are set up according to Eq. (5):

$$R'_t = \alpha R_{t-1} + (1-\alpha)R'_{t-1} \tag{5}$$

Next, we employ Eq. (6) to predict the future value using the EWMA of the 2016 values themselves.

$$P'_t = \alpha P_{t-1} + (1-\alpha)P'_{t-1} \tag{6}$$

The advanced regression models, namely Random Forest and XGBoost, are central to our prediction strategy. These models have proven effective in capturing the complex relationships within the data. They were chosen based on their documented success in taxi demand prediction tasks [9].

2.5 Proposed method

The proposed method in our research is improve prediction of taxi demand by using feature engineering. In spite of the temporal and spatial characteristics inherent in the dataset, we incorporated Fourier features to capture periodic patterns and seasonality in taxi demand. This enhancement enables our models to better capture recurring trends, such as daily and weekly variations in demand. The incorporation of Fourier features was inspired by recent work in time series analysis by Xu [7], which demonstrated the effectiveness of this approach in capturing periodic patterns.

In addition, we leverage Fourier Transforms to decompose the time series data into its constituent frequency components. The Fourier Transforms help us identify recurring patterns and underlying frequency signals within the time series [21]. One of the fundamental formulas of Fourier Transforms is as follows Eq. (7):

$$H(f) = \int_{-\infty}^{\infty} h(t)e^{-i2\pi ft}\,dt \tag{7}$$

where:

$H(f)$ represents the frequency domain representation of the time series H.

$h(t)$ is the actual time series data at time t.

f stands for frequency in the frequency domain.

i represents the imaginary unit.

By applying Fourier Transforms to our time series data, we gain insights into recurring patterns and the presence of periodic components in the taxi demand, allowing us to capture and model the influence of various frequencies on demand. To characterize further the temporal patterns of taxi service demand in the selected regions, we conducted a comprehensive frequency analysis. The analysis involved computing the top 5 frequencies ($F1$ to $F5$) and their corresponding amplitudes ($A1$ to $A5$) for each region, which allowed us to identify the most prominent cyclic patterns in taxi pickups. The frequencies represent the temporal cycles at which the demand for taxi services exhibits substantial variations, while the amplitudes quantify the magnitude of these variations.

Then, proceeded with predictive modeling to estimate the temporal patterns of taxi service demand. To achieve this, we prepared a dataframe that featured the $x(i)$ values as the smoothed data from January 2015 and the $y(i)$ values as the corresponding data from January 2016. This dataframe enabled us to calculate the ratios between the observed demand in January 2016 (P_t^{2016}) and that of January 2015 (P_t^{2015}) for each time bin as Eq. (8). These ratios served as essential indicators for assessing how the demand patterns had evolved over time, allowing us to make insightful predictions about future taxi service requirements. This predictive aspect of our research has significant implications for optimizing taxi fleet management and resource allocation to meet the dynamic demand patterns in different geographical regions.

$$R_t = \frac{P_t^{2016}}{P_t^{2015}} \tag{8}$$

In addition to the existing features, we have introduced five new features, denoted as Ft_1 to Ft_5. Ft_1 represents the number of pickups that occurred during the previous four 10-minute intervals, from *t*-2 to *t*-5. Table 3 presents feature Fourier in the data test, and Table 4 presents feature Fourier in the data test. These features capture the periodicity of taxi demand patterns and provide valuable information for our predictive models. Specifically, we combine the amplitudes $A1$ to $A5$ and frequencies $F1$ to $F5$ into dataset. Meanwhile, Figure 9 presents a frequency and amplitude graph, here we will see the difference in amplitude at a certain time.

Figure 9 illustrates the amplitude versus frequency plot from the Fourier Transform applied to taxi demand data. On the x-axis, frequency is shown where lower frequencies correspond to daily patterns in taxi demand, while the y-axis represents amplitude, quantifying the strength of these patterns. The most significant peak, reaching an amplitude of approximately 350,000, occurs near the 0.01 frequency mark, highlighting the dominance of daily cycles in taxi usage. The sharp decrease in amplitude as the frequency approaches 0.5 indicates that high-frequency, short-term fluctuations are considerably less impactful. This graph clearly demonstrates the periodic nature of taxi demand, which is pivotal for optimizing our prediction models.

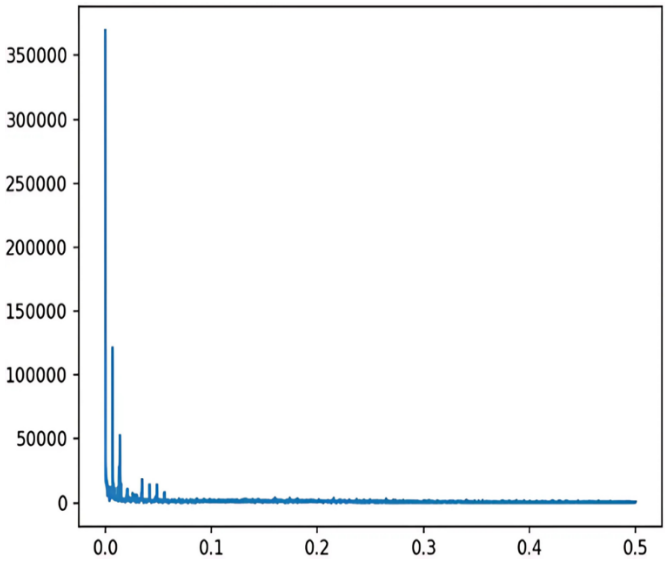


Figure 9. Plot of amplitude vs frequency

2.6 Evaluation metrics

Our methodology employs performance metrics with MAPE. MAPE evaluates the percentage variance between estimated and actual demand, providing an intuitive understanding of prediction accuracy. These metrics were chosen based on their widespread use in taxi demand prediction research by Zhang et al. [22]. The formula of MAPE is shown as in Eq. (9).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{X}_i - X_i}{X_i}\right| \times 100\% \tag{9}$$

Table 3. Feature Fourier in data test

	Ft_5	Ft_4	Ft_3	Ft_2	Ft_1	Lat	Lon	Weekday	exp_avg	wei_avg
0	240	213	243	222	234	40.777809	-73.954054	4	231	690
1	213	243	222	234	291	40.777809	-73.954054	4	273	816
2	243	222	234	291	256	40.777809	-73.954054	4	261	803
3	222	234	291	256	266	40.777809	-73.954054	4	264	788
4	234	291	256	266	268	40.777809	-73.954054	4	266	802

Table 4. Feature Fourier in data train

	Ft_5	Ft_4	Ft_3	Ft_2	Ft_1	Lat	Lon	Weekday	exp_avg	wei_avg
0	0	106	243	299	328	40.777809	-73.954054	4	309	955
1	106	243	299	328	340	40.777809	-73.954054	4	330	1008
2	243	299	328	340	316	40.777809	-73.954054	4	320	972
3	299	328	340	316	327	40.777809	-73.954054	4	324	970
4	328	340	316	327	323	40.777809	-73.954054	4	323	973

3. RESULT AND DISCUSSION

The results of our taxi demand prediction study encompass the performance of both baseline models and advanced regression models, with a specific focus on $MAPE$ as the primary evaluation metric.

3.1 Baseline MODEL PERFORMANCE

Table 5. Prediction baseline model

Methods	MAPE	
	Ratios	**Values**
Simple moving averages	0.22785	0.15583
Weighted moving averages	0.22707	0.14795
Exponential weighted moving averages	0.22755	0.14754

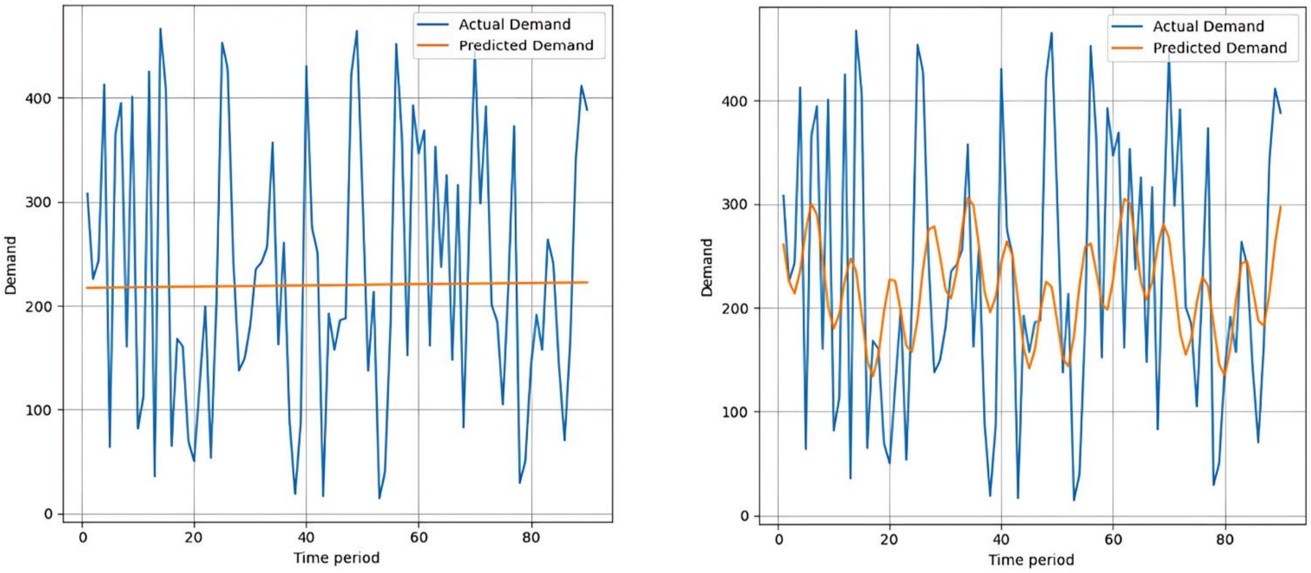
Our study began by evaluating the performance of basic time series models, namely SMA, WMA, and EWMA. These models served as essential benchmarks for our more complex approaches. The performance of the model is presented in Table 5.

These baseline models, though straightforward, provide a crucial baseline for evaluating the effectiveness of advanced regression models.

3.2 Impact of feature engineering

The incorporation of Fourier features, which capture periodic patterns and seasonality, significantly enhanced the models' ability to capture recurring trends. This was especially important in urban transportation, where demand patterns exhibit strong temporal dependencies.

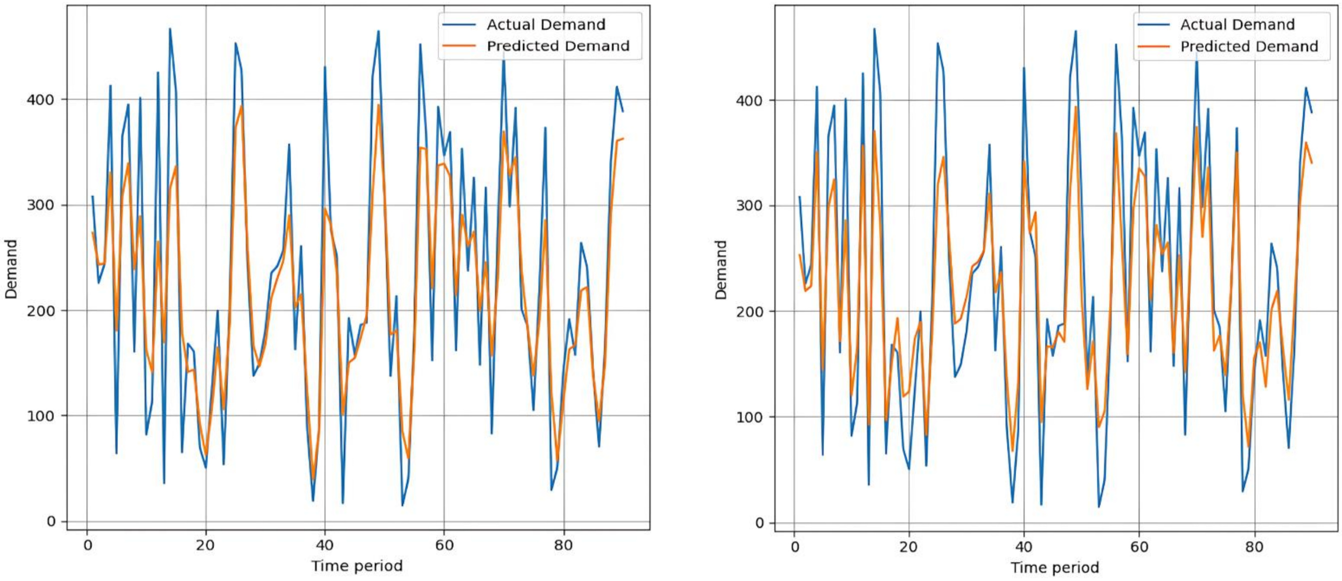
For example, we used daily taxi demand data for three months and compared three models (i.e., Liner Regression, Random Forest and XGBoost). The model without Fourier feature is presented in Figure 10. Meanwhile, the model with fouririer feature is presented in Figure 11.



(a) Without Fourier

(b) With Fourier

Figure 10. Plot demand prediction without Fourier feature



(a) Without Fourier

(b) Without Fourier

Figure 11. Plot demand prediction with Fourier feature

Figure 11 presents the improved performance of Linear Regression, Random Forest, and XGBoost models in predicting taxi demand when enhanced with Fourier Transform features, in contrast to the predictions shown in Figure 10 without such enhancements. This figure illustrates that the integration of Fourier features significantly refines the models' accuracy: the predictions (orange line) closely mirror the actual demand (blue line) across various time periods. Specifically, Linear Regression, typically less adept at capturing complex patterns, shows marked improvement in aligning with the actual demand curves, suggesting effective capture of cyclic demand variations. Similarly, Random Forest and XGBoost models exhibit enhanced responsiveness to sudden demand changes, with XGBoost demonstrating particularly notable precision in tracking the intricacies of demand fluctuations. This visual comparison underscores the value of Fourier Transform features in augmenting predictive models to better understand and anticipate the dynamics of taxi demand, thereby offering a more robust tool for urban mobility planning and management.

3.3 Model performance

Moving beyond the baseline models, our study incorporated advanced regression models to improve prediction accuracy, Linear Regression, Random Forest Regression, and XGBoost Regression were among the models used. We take three months of intake data from 2016 and split it so that each region is 70% training and 30% testing.

To ensure the reproducibility of our findings, detailed configurations of the Random Forest and XGBoost models are provided. The Random Forest model was configured with 100 trees, each with a maximum depth of 10, using the Gini coefficient to measure split quality. For XGBoost, we set a learning rate of 0.1, maximum depth of 6, and ran 150 training rounds with a subsample ratio of 0.8 to prevent overfitting.

The robustness of these models was validated using a *k*-fold cross-validation approach with k set to 5, ensuring the models were tested across diverse subsets of data. Parameter optimization was performed using grid search, focusing on minimizing the root mean square error to fine-tune the models for optimal performance.

In Table 6 shows that the accuracy of the demand prediction model with Fourier feature is obviously improved. Linear Regression exhibited a MAPE of 0.1156, indicating a substantial improvement over the baseline models. Random Forest Regression and XGBoost Regression further enhanced the performance, with MAPE values of 0.1137 and 0.1161, respectively. These advanced models showcased their capability to capture complex patterns in urban taxi demand data, leading to more accurate predictions.

Table 6. Prediction model

Methods	Without Fourier Feature		With Fourier Feature	
	Train	Test	Train	Test
Linear Regression	0.1333	0.1290	0.1198	0.1156
Random Forest	0.1281	0.1271	0.1142	0.1137
XGBoost Regression	0.1296	0.1267	0.1197	0.1161

The integration of Fourier Transform features significantly improved the adaptability of the Random Forest and XGBoost models, making them more effective than traditional time series methods like SMA and EWMA in handling unpredictable changes in urban demand. This was particularly evident in scenarios with sudden shifts, such as weather changes, where the Fourier Transform helped to capture and adjust for high-frequency variability in the data. XGBoost, enhanced with these features, demonstrated a substantial reduction in MAPE, proving its efficacy over traditional models in accurately forecasting taxi demand under dynamic conditions.

3.4 Implications

The practical implications of our study's findings for urban mobility and taxi services are significant. The achievement of a MAPE of less than 12% in predicting taxi demand holds promise for optimizing the efficiency of transportation systems.

The incorporation of Fourier features played a crucial role in enhancing these models' performance, allowing them to effectively model periodic demand patterns. This emphasizes the significance of feature engineering in improving the accuracy of taxi demand predictions, aligning with the findings of Xu [7]. These results align with prior research in urban computing by Faghih et al. [23], which emphasized the significance of machine learning models in urban mobility analysis. The comparison revealed that advanced regression models are well-suited for addressing the complexities of urban taxi demand prediction, showcasing their potential for optimizing transportation services in metropolitan areas.

The success of Random Forest Regression and XGBoost Regression in outperforming baseline models can be attributed to their robustness in handling complex, nonlinear patterns in urban taxi demand data. Random Forest Regression leverages the power of ensemble learning, effectively combining several decision trees to improve accuracy in predictions. XGBoost Regression, on the other hand, employs gradient boosting to iteratively improve the model's performance. Our findings are in line with the work of Stadler et al. [9], which emphasized the significance of machine learning models in urban computing.

3.5 Limitations of the study

Despite the promising results, our study is not without limitations. One limitation lies in the dataset's temporal scope, which covers only the period from January to March. While this period provides insights into seasonal trends, a more extended dataset could offer a more comprehensive understanding of long-term demand patterns.

Another limitation is the assumption that the historical data reflects future taxi demand patterns. External factors such as economic changes, special events, or unexpected incidents can significantly impact demand and were not considered in our models. Future research could explore methods for integrating real-time external data sources to enhance prediction accuracy.

3.6 Future research directions

Further research should delve into refining the signal processing techniques used in urban mobility prediction by integrating Fourier Transform features with wavelet transforms. This approach can enhance the models' capacity to analyze non-stationary data, improving adaptability to abrupt changes in urban conditions. Exploring hybrid models that combine machine learning with econometric analyses can also provide a more holistic understanding of factors influencing taxi demand, integrating traffic data with socio-economic indicators.

Additionally, enhancing the interpretability of complex models like Random Forest and XGBoost is crucial for their application in urban planning and policy-making. Developing methods for model decomposition and applying explainable AI frameworks could increase transparency and stakeholder trust. Comparative studies across diverse urban environments would also be valuable, testing the scalability and adaptability of predictive models to different urban layouts and mobility systems, ensuring their effectiveness and generalizability.

4. CONCLUSION

This study achieves the initial objectives by significantly enhancing the accuracy of taxi demand prediction models through the integration of Fourier Transform features with Linear Regression, Random Forest, and XGBoost. This approach has markedly improved model performance in capturing complex urban demand patterns, leading to practical applications in urban mobility management.

Enhanced predictive capabilities enable transportation authorities and service providers to better align resources with demand fluctuations, improving operational efficiency and passenger satisfaction. This methodology can also be applied to other urban planning areas, such as public transportation and emergency services, where accurate predictions are essential.

Furthermore, these advancements support sustainable urban development by enabling more efficient transportation systems that contribute to reduced congestion and lower emissions. Future research could focus on incorporating real-time data to refine these predictions further, promoting smarter, more responsive urban growth.

ACKNOWLEDGMENT

This work was partially funded by the Department of Computer Science and Electronics, Universitas Gadjah Mada under the Publication Funding Year 2024.

REFERENCES

[1] Rodrigues, F., Markou, I., Pereira, F.C. (2019). Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach. Information Fusion, 49: 120-129. https://doi.org/10.1016/j.inffus.2018.07.007

[2] Chou, K.S., Wong, K.L., Zhang, B., Aguiari, D., Im, S.K., Lam, C.T., Tse, R., Tang, S.K., Pau, G. (2023). Taxi demand and fare prediction with hybrid models: Enhancing efficiency and user experience in city transportation. Applied Sciences, 13(18): 10192. https://doi.org/10.3390/app131810192

[3] Liu, T., Wu, W., Zhu, Y., Tong, W. (2020). Predicting taxi demands via an attention-based convolutional recurrent neural network. Knowledge-Based Systems, 206: 106294. https://doi.org/10.1016/j.knosys.2020.106294

[4] Liu, Z., Chen, H., Li, Y., Zhang, Q. (2020). Taxi demand prediction based on a combination forecasting model in hotspots. Journal of Advanced Transportation, 2020(1): 1302586. https://doi.org/10.1155/2020/1302586

[5] Liu, Z., Chen, H., Sun, X., Chen, H. (2020). Data-driven real-time online taxi-hailing demand forecasting based on machine learning method. Applied Sciences, 10(19): 6681. https://doi.org/10.3390/APP10196681

[6] Ibrahim, M.M., Mubarek, F.S. (2023). Improving prediction for taxi demand by using machine learning. In 2023 15th International Conference on Developments in eSystems Engineering (DeSE), Baghdad & Anbar, Iraq, pp. 451-456. https://doi.org/10.1109/DeSE58274.2023.10099731

[7] Xu, T. (2022). Demand analysis of taxi passenger-carrying hot spot areas based on XGBoost algorithm. In 2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, pp. 682-686. https://doi.org/10.1109/ICAIBD55127.2022.9820001

[8] Vanichrujee, U., Horanont, T., Pattara-atikom, W., Theeramunkong, T., Shinozaki, T. (2018). Taxi demand prediction using ensemble model based on RNNs and XGBoost. In 2018 International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES), Khon Kaen, Thailand, pp. 1-6. https://doi.org/10.1109/ICESIT-ICICTES.2018.8442063

[9] Stadler, T., Sarkar, A., Dünnweber, J. (2021). Bus demand forecasting for rural areas using XGBoost and Random Forest algorithm. In Computer Information Systems and Industrial Management: 20th International Conference, CISIM 2021, Ełk, Poland, pp. 442-453. https://doi.org/10.1007/978-3-030-84340-3_36

[10] Ristin, M., Guillaumin, M., Gall, J., Van Gool, L. (2015). Incremental learning of Random Forests for large-scale image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(3): 490-503. https://doi.org/10.1109/TPAMI.2015.2459678

[11] 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, San Francisco, USA, pp. 785-794. https://doi.org/10.1145/2939672.2939785

[12] Xu, J., Rahmatizadeh, R., Bölöni, L., Turgut, D. (2017). Real-time prediction of taxi demand using recurrent neural networks. IEEE Transactions on Intelligent Transportation Systems, 19(8): 2572-2581. https://doi.org/10.1109/TITS.2017.2755684

[13] Kuang, L., Yan, X., Tan, X., Li, S., Yang, X. (2019). Predicting taxi demand based on 3D convolutional neural network and multi-task learning. Remote Sensing, 11(11): 1265. https://doi.org/10.3390/rs11111265

[14] Zhang, J., Zheng, Y., Qi, D., Li, R., Yi, X., Li, T. (2018). Predicting citywide crowd flows using deep spatio-temporal residual networks. Artificial Intelligence, 259: 147-166. https://doi.org/10.1016/j.artint.2018.03.002

[15] Ye, J., Sun, L., Du, B., Fu, Y., Tong, X., Xiong, H. (2019). Co-prediction of multiple transportation demands based on deep spatio-temporal neural network. In Proceedings of the 25th ACM SIGKDD International

Conference on Knowledge Discovery & Data Mining, Anchorage, USA, pp. 305-313. https://doi.org/10.1145/3292500.3330887

[16] Rajak, S., Baruah, U. (2020). An ensemble model for predicting passenger demand using taxi data set. In Machine Learning, Image Processing, Network Security and Data Sciences: Second International Conference, MIND 2020, Silchar, India, pp. 336-346. https://doi.org/10.1007/978-981-15-6318-8_28

[17] Zhao, K., Khryashchev, D., Vo, H. (2019). Predicting taxi and uber demand in cities: Approaching the limit of predictability. IEEE Transactions on Knowledge and Data Engineering, 33(6): 2723-2736. https://doi.org/10.1109/TKDE.2019.2955686

[18] Gul, E., Staiou, E., Safari, M.J.S., Vaheddoost, B. (2023). Enhancing meteorological drought modeling accuracy using hybrid boost regression models: A case study from the Aegean region, Türkiye. Sustainability, 15(15): 11568. https://doi.org/10.3390/su151511568

[19] TLC Trip Record Data. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[20] Wang, Y., Ren, J. (2021). Taxi passenger hot spot mining based on a refined k-means++ algorithm. IEEE Access, 9: 66587-66598. https://doi.org/10.1109/ACCESS.2021.3075682

[21] Brigham, E.O. (1988). The Fast Fourier Transform and its Applications. Prentice-Hall, Inc., United States.

[22] Zhang, C., Zhu, F., Lv, Y., Ye, P., Wang, F.Y. (2021). MLRNN: Taxi demand prediction based on multi-level deep learning and regional heterogeneity analysis. IEEE Transactions on Intelligent Transportation Systems, 23(7): 8412-8422. https://doi.org/10.1109/TITS.2021.3080511

[23] Faghih, S., Shah, A., Wang, Z., Safikhani, A., Kamga, C. (2020). Taxi and mobility: Modeling taxi demand using ARMA and linear regression. Procedia Computer Science, 177: 186-195. https://doi.org/10.1016/j.procs.2020.10.027