Forecasting Ride-Hailing Demand in Urban Areas: A Deep Ensemble and Time Series Clustering Approach

Kushal Acharya[1)], Luca Quadrifoglio[2*)]

[1)] Texas A&M University, Civil and Environmental Engineering, 3136 TAMU College Station, Texas; ORCID: https://orcid.org/0009-0001-8756-4217
[2*)] Texas A&M University, Civil and Environmental Engineering, 3136 TAMU College Station, Texas; e-mail: quadrifo@tamu.edu; ORCID: https://orcid.org/0000-0002-2596-7504

http://doi.org/10.29227/IM-2024-02-52
Submission date: *17.04.2024.* | Review date: *15.05.2024*

Abstract

This paper investigates the increasingly important task of forecasting demand for ride-hailing services, which have significantly disrupted traditional transportation models. Notably, the study concentrates on New York City's Yellow Cab service, which despite the surge in popularity of app-based services, continues to serve a substantial number of commuters. The study highlights the necessity of accurate demand prediction for efficient resource allocation, reduced wait times and improved user satisfaction. Traditional forecasting methods like Historical Average, Exponential Weighted Moving Averaes, ARIMA etc., are examined, alongside the more recent machine learning and data mining techniques, CNN-LSTM and XGBoost. A novel approach, utilizing an ensemble of machine learning models – XGBoost and Convolutional Neural Network – LSTM along with creative feature engineering is proposed for real-time demand forecasting across numerous locations. Furthermore, the study also tries to understand the application of time-series clustering methods and their effectiveness in grouping similar time-series together and extracting clustering features to improve the performance of the model. Additionally, the study observes the ineffectiveness of generalized model to forecast demand in low-demand reasons and presents possible research direction for solving the issue. This study contributes to the growing literature on demand forecasting in the ride-hailing industry and provides insights into the use of time-series clustering for the same.

Keywords: *Ride-Hailing, Ensemble, CNN-LSTM, Time-Series Forecasting*

1. Introduction

The advent of ride-hailing services has not only dramatically reshaped the transportation industry but also fundamentally changed how people navigate through cities (1). This paradigm shift is marked by a significant transition from traditional taxi services to modern, app-based ride-hailing platforms, such as Uber, Lyft, and Didi. These services offer fast, efficient, and often customized transportation solutions that cater to consumers' evolving demands. The ride-hailing industry has emerged as an essential component of the global urban transportation ecosystem and has seen remarkable growt with widespread adoption of smartphones and growing digital literacy (1).

In this rapidly evolving landscape, one of the most critical and complex tasks is to accurately forecast the demand for ride-hailing services. Accurate demand prediction is crucial for various reasons. It enables optimal allocation of resources, minimizing wait times, and enhancing the user experience, contributing to the overall efficiency and profitability of the service (1). Furthermore, accurate demand prediction aids in strategic planning and guides government actions to accommodate this growing market segment.

In this study, we focus on New York City (NYC), one of the world's most populous and vibrant cities, which offers a unique environment for studying ride-hailing services. Despite the rise of app-based services, the traditional Yellow Cab service, regulated by the New York City Taxi and Limousine Commission (TLC), remains an important part of NYC's transportation system (2). The TLC's 2018 Factbook revealed that Yellow Cabs served an average of 310,950 trips per day in 2017 (3). Such a large operation emphasizes the importance of predicting the demand for Yellow Cab services.

Various methods have been employed for forecasting te demand of ride-hailing services, ranging from traditional approaches like time-series analysis and moving averages to advanced machine learning and data mining techniques. Recent years have witnessed a significant enhancement in predictive capabilities due to the introduction of deep learning techniques such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, Random Forests, etc. (4,5,6). However, despite these advancements, real-time forecasting of demand across various locations remains a challenge.

Addressing this issue, our study proposes an innovative approach that leverages an ensemble of machine learning models – XGBoost and Convolutional Neural Network – LSTM, along with capitalizing on the outputs from primitive statistical methods like historical moving averages, and exponential weighted moving averages. We aim to forecast the demand for the subsequent 30-minute time intervals across various NYC locations. Our models are compared against baseline methods, such as ARIMA, XGBoost and CNN-LSTM to highlight their superior predictive performance. Furthermore, we utilize k-shape clustering, verified using silhouette scores, to manage multi-dimensional time-series data across multiple locations, thereby enhancing the robustness of our study.

2. Literature Review

Initially, demand forecasting was tackled using traditional time series methods, such as Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA models (7,8). These methods laid the foundation for predicting demand for traditional transport services. However, the complex, multivariate nature of data from ride-hailing services and the distinct difference from traditional taxi services demanded more advanced techniques (9).

Machine learning techniques emerged as promising tools in demand prediction due to their ability to manage high-dimensional data and comprehend complex inter-variable relationships. Random Forests and Support Vector Machines demonstrated significant promise (9), where Liu et al. (10) employed them to identify and predict taxi demand in specific hotspots. Gradient boosting models, especially XGBoost, showed robust performance across varied data structures, with Poongodi et al. (11) exhibiting the superiority of XGBoost over Multi-layered Perceptron (MLP) in predicting NYC taxi trip durations.

With the advent of deep learning, the precision of time series forecasting experienced a significant improvement. Techniques like Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) offered promising results in processing spatial and temporal information, thus improving the demand forecasting task (12). Notably, Yao et al. (13) introduced the Deep Multi-View Spatial-Temporal Network (DMVST-Net), specifically for taxi demand prediction. More recent research has highlighted the potential of Graph Neural Networks (GNNs) for traffic and demand forecasting due to their capability to capture spatial dependencies (14,15). However, the computational intensity of GNNs remains a challenge.

Apart from these, a few studies have explored clustering approaches for taxi demand prediction. Davis et al. (16) employed a multilevel clustering method to simulate taxi-demand density, yet time-series clustering approaches have been relatively underexplored. Dynamic Time Warping (DTW) (17), a popular method, suffers from high computational requirements for large datasets. To counter this, Paparrizos and Gravano (18) proposed the k-shape technique, a time-efficient approach for time series clustering, which this study also aims to utilize.

Ensemble techniques, as demonstrated by Moreira-Matias et al. (19), combining time-varying Poisson model and ARIMA, have shown improved prediction accuracy. A different ensemble using RNN, XGBoost, and GRUs was proposed by Vanichrujee et al. (20). However, this study differentiates itself from existing literature by focusing on the ensemble of CNN-LSTM and XGBoost for demand forecasting at multiple locations simultaneously while leveraging cluster information for feature engineering and determine its impact on overall prediction. Furthermore, current research often overlooks stable and recurrent temporal patterns of human mobility and the importance of feature engineering from existing demand data. Therefore, this study aims to fill these gaps by proposing an ensemble model leveraging time-series clustering and feature engineering to better predict ride-hailing demand.

3. Data

The primary data for our research, i.e., the taxi demand data, is obtained from the New York City Taxi and Limousine Commission (NYC TLC). The NYC TLC is responsible for regulating and licensing of taxi services within New York City. Similarly, the weather data was obtained for the weather station at Laguardia Aiport Station from the website weather underground (21). The data about Federal US holidays was obtained from the office holidays' website. (22)

The trip records contain extensive data on millions of taxi rides, including details such as pickup and drop-off times and locations, trip distances, fare amounts, and other relevant information. Data for all twelve months of 2017 were collected, however, only the data from the first three months (January till March) were used for the experimentation and final analysis. The total number of records for the first three months was 28.58 million. Each trip was assigned to one of 263 location IDs in New York city, however, most of these locations barely have any significant demand. The data was first cleaned removing any outliers, followed by aggregating the trips demand for thirty-minute time bins. The data was further divided into three sets: training, validation, and test via a 70-20-10 split for modeling purposes. Figure 1 shows the average demand for yellow cab taxis on a typical Monday.

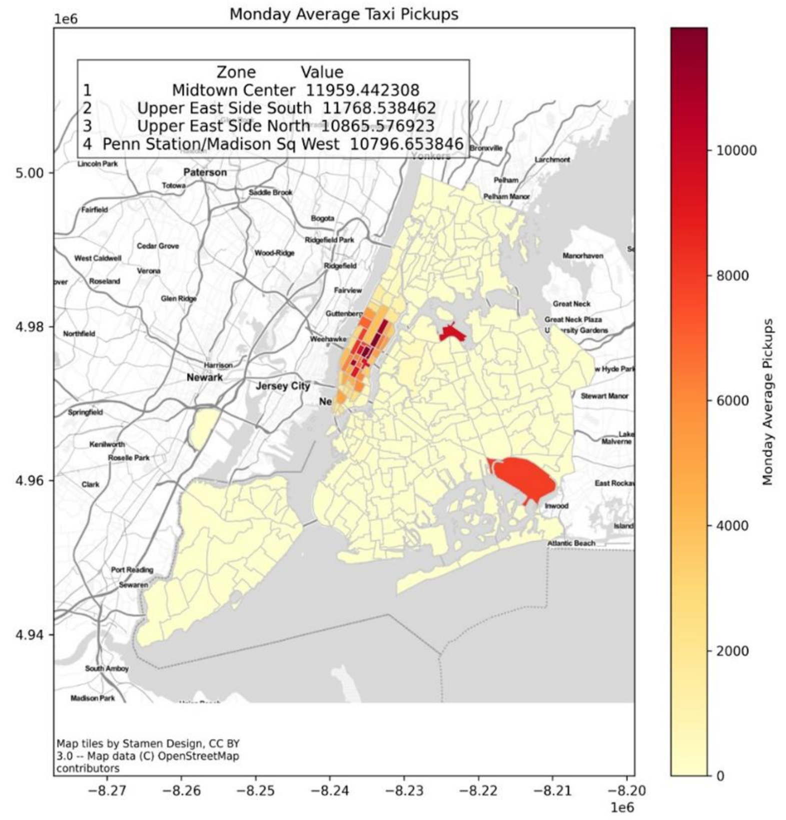


Fig. 1. Average Pickup Demand on a typical Monday.

The above figure shows that most of the locations out of the 263 locations have low to minimum demand.

Since the demand at most of the locations was sparse, the analysis was narrowed down to specific 81 location IDs which had demand for at least 75 percent of the time, i.e. these locations had at least some demand in 75 percent of the 30 minute time bins over the period of three months.

4. Methods

4.1 Time-Series Clustering

Time-series clustering is an important methodology for extracting significant insights from temporal data. In this study, a clustering analysis was conducted using the k-shape algorithm due to its efficiency, particularly with large data sets in comparison to traditional techniques like Dynamic Time Warping (DTW). Two major objectives of this analysis were to identify patterns in pickup demand time series across different locations and to enhance the performance of the predictive model using cluster-derived features.

K-Shape Clustering Algorithm. The k-shape clustering algorithm is a centroid-based algorithm designed specifically for time series data by Paparazzos and Gravano (18). Unlike traditional clustering algorithms, such as k-means, k-shape effectively manages phase shifts and amplitude variations, which are common issues with time series data. The algorithm employs several steps:

i) Data Normalization: Time series data is standardized using z-normalization, ensuring comparable amplitude, and offset by subtracting the mean and dividing by the standard deviation.

ii) Centroid Initialization: A scalable technique is used for initializing centroids, generating precise initial centroids that improve convergence and reduce the risk of falling into local optima.

iii) Assignment of Time Series to Centroids: The "shape-based distance" (SBD) metric is used to allocate time series to the nearest centroid. This metric gauges the similarity in shapes of time series, remaining invariant to phase shifts.

iv) Centroid Update: A "shape extraction" method is employed to calculate new centroids based on the average shape of time series within a cluster, without considering phase shifts.

v) Convergence: The process iterates steps iii) and iv) until reaching convergence, when time series' assignments to centroids stabilize.

K-shape clustering algorithm is preferred for its scalability, accuracy, and computational efficiency compared to other time series clustering methods (18).

Silhouette Coefficients for Cluster Selection. Silhouette coefficients, developed by Rousseeuw (23), are employed for assessing the quality of clustering results and identifying the ideal number of clusters. The coefficients range from -1 to +1, providing a quantitative measure of the suitability of a data point's assignment to its cluster in relation to other clusters. Higher values signify more appropriate cluster assignments.

Calinski-Harabasz Index for Cluster Selection. The Calinski-Harabasz (CH) index (24), or the variance ratio criterion, is another metric used for determining the quality of clustering outcomes and pinpointing the optimal number of clusters in time series data. It measures the ratio between the between-cluster variance and the within-cluster variance, with higher values indicating better-defined clusters.

Figure 2 shows the flowchart for time-series clustering approach applied in this study.

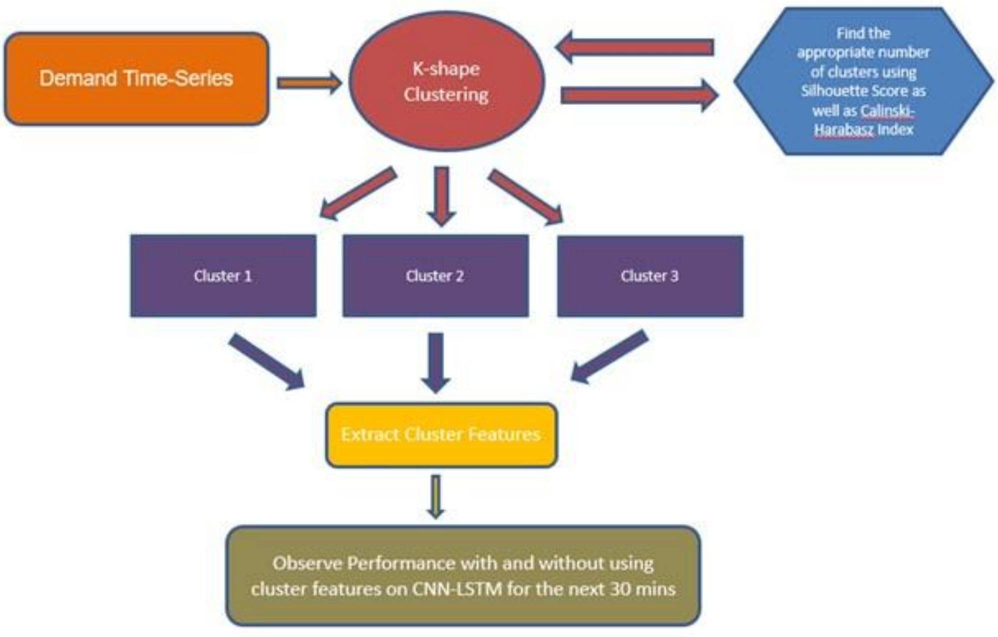


Fig. 2. Methodology: Time-Series Clustering using K-shape clustering.

K-shape clustering is applied on all 81 selected locations. The correct number of clusters are identified using Silhouette score and Calinski-Harabasz Index. One of the baseline models CNN-LSTM is applied on each cluster individually as well as on the entire 81 locations to make one step ahead prediction, i.e. for the next 30 min time bin. Cluster features are then extracted from each cluster and added as additional features to model the entire 81 locations at once. The performance of the CNN-LSTM model with cluster features is compared with the performance of the model with similar architecture without added cluster features on all 81 locations to observe improvement in performance because of added cluster features.

One of the strengths of this study is the use of neatly crafted features from the time-series data long with cluster features and exogeneous variables. Table 1 shows the general features created for the cluster analysis as well as the ensemble model, whereas Table 2 shows the features especially created using clusters. These features collectively help improve the performance of the general model as later seen in the results section.

Tab. 1. List of features created along with the exogeneous feature variables.

Features	Description
Historical Data	Demand @ Time_step_t-1, …, Time_step_t-5
Weekly historical average	t-(1 week)
Rolling_max	Max(t-1,t-2,… t-24 hrs)
Rolling_min	Min(t-1,t-2,….t-24 hrs)
Rolling_mean	Mean(t-1,t-2,…,t-24 hrs)
Rolling_std	Standard_Deviation((t-1,t-2,…,t-24 hrs)
Rolling_skew	Skewness(t-1,t-2,…t—24 hrs)
EWMA_output	Output/prediction from EWMA model for t=1
is_weekday	1 if it is a weekday, else 0
is_weekend_holiday	1 if it is a weekend or national holiday, else 1
Day_of_week	0 for Monday, …, 6 for Sunday
Weather_temperature	Temperature (Fahrenheit)
Weather_precipitation	Precipitation (mm)
Weather_humidity	Percentage (%)

Tab. 2. List of cluster features.

Features	Description
Diff	(time_series_@ t-1 - represntative _time_series @t-1)
Mean_diff	Mean (time_series_i – represntative _time_series)
Std_diff	Std (time_series_i – representative_time_series)
Intra_cluster_similarity	Cluster Similarity within cluster
Inter_cluster_similarity	Cluster Similarity between clusters
Rolling_mean	Mean(t-1,t-2,…,t-24 hrs)
Rolling_std	Standard_Deviation((t-1,t-2,…,t-24 hrs)
Rolling_skew	Skewness(t-1,t-2,…t—24 hrs)

4.2 Baseline Models

Historical Average (HA). Historical average is a simple method of time series prediction that involves using the average of a variable's past values to predict its future values. For example, if we are to predict the demand at the period between 9:00 am and 9:30 am on Tuesday, we take an average of demand across all Tuesdays for the same time.

Exponential Weighted Moving Average (EWMA). Exponential weighted moving average (EWMA) is a time-series forecasting technique that assigns exponentially decreasing weights to the historical observations. One advantage of EWMA is that it can capture short-term fluctuations effectively. However, it is not suitable for handling long-term trends or seasonality in the data.

AutoRegressive Integrated Moving Average (ARIMA). ARIMA is a popular statistical model used for time series forecasting. It is a combination of three components: Autoregression (AR): a model that uses the past values of a variable to predict its future values. Integration (I): a model that uses the difference between the current value and the past value of a variable to make a prediction. Moving Average (MA): a model that uses the average of the past error terms to predict the future value of a variable.

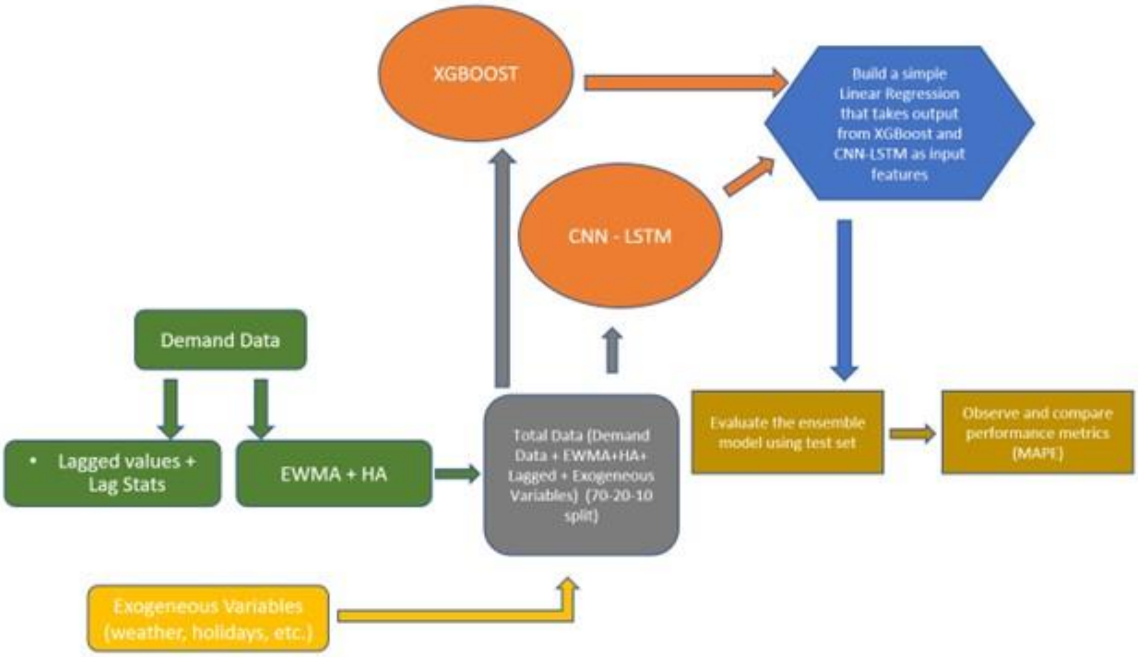
Extreme Gradient Boosting (XGBoost). Extreme Gradient Boosting (XGBoost) is a highly efficient machine learning algorithm that uses gradient boosting to improve decision tree models. It enhances model performance by adjusting the weights of training samples, focusing on those misclassified by the previous decision tree. XGBoost outperforms Gradient Boosting Decision Trees with its faster training and built-in regularization, which helps prevent overfitting by adding a penalty term to the objective function (25).

CNN-LSTM. Convolutional Neural Networks (CNNs) and Long Short-Term Memory Neural Networks (LSTM NNs) offer distinct strengths in processing spatially indexed time-series data and temporal data respectively. CNNs are effective at identifying spatial features and learning hierarchical patterns, ideal for pinpointing high-demand taxi locations. They apply filters to specific input regions, allowing learning of complex features. LSTM NNs, on the other hand, excel at handling sequential data and capturing long-term dependencies in time series. The CNN-LSTM hybrid model is one of the various methods of integrating the two architectures. It employs CNN to extract spatial features, which are then processed by LSTM for temporal dynamics modeling (26). This approach is especially appropriate for spatio-temporal prediction tasks, including taxi demand forecasting. ConvLSTM models offer another alternative, fusing convolutional and recurrent operations within a single layer for a more compact representation of the problem. Both the models can capture both spatial and temporal dynamics, yielding more accurate predictions for taxi demand forecasting.

4.3 Proposed Ensemble

The proposed model is an ensemble that linearly combines the output generated by the XGBoost model which takes the exogeneous variables as well as the generated features as input and a multivariate CNN-LSTM model, which also takes specialized generated features for improved performance. XGBoost model can capture the effect of exogeneous variables on the time-series as well as capture the short-term fluctuations through the generated new features, while CNN-LSTM is able to better capture the spatial patterns as well as long-term dependencies of the demand time-series.

Both XGBoost and CNN-LSTM are tuned in to get the best performance. The created ensemble is tested throughout the second cluster of time -series which consists of 41 locations. This cluster is less zero inflated and represents the area which had high demand throughout various times of the day as well as throughout the year. Several features are created from the existing time-series including lag values, as well as exogeneous variables like weather and holidays which are all tabulated in Table 1. The use of the outputs from Historical Average and Exponential Weighted Moving Average which in themselves have a decent performance really helps the proposed model perform better. Figure 3 shows the flowchart of the proposed ensemble model.



Fag. 3. Proposed Ensemble Model

5. Results and Discussion

K-shape clustering was applied to the data in selected 81 locations. Two metrics were used to come up with the right number of clusters for the clustering analysis, namely, Silhouette coefficients and Calinski-Harabasz Index. Figures 4, 5 and 6 show the scores for various cluster sizes. From the analysis, it was found that dividing the locations into three clusters was most appropriate.

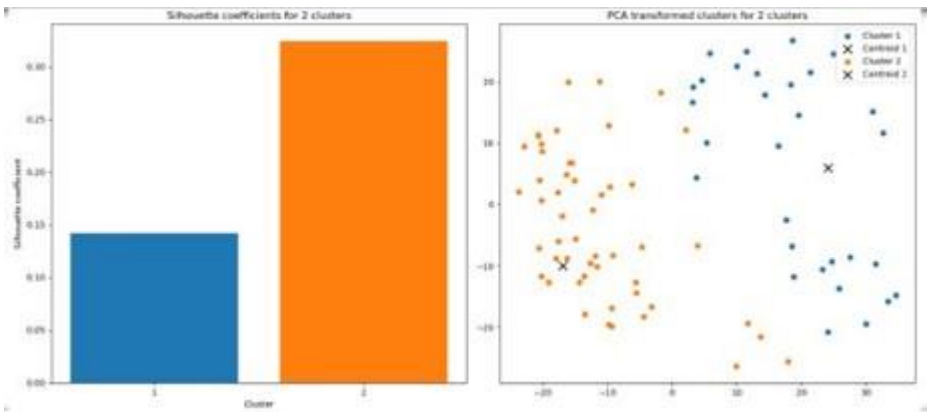


Fig. 4. Silhouette coefficient and PCA transformed cluster representation for two clusters

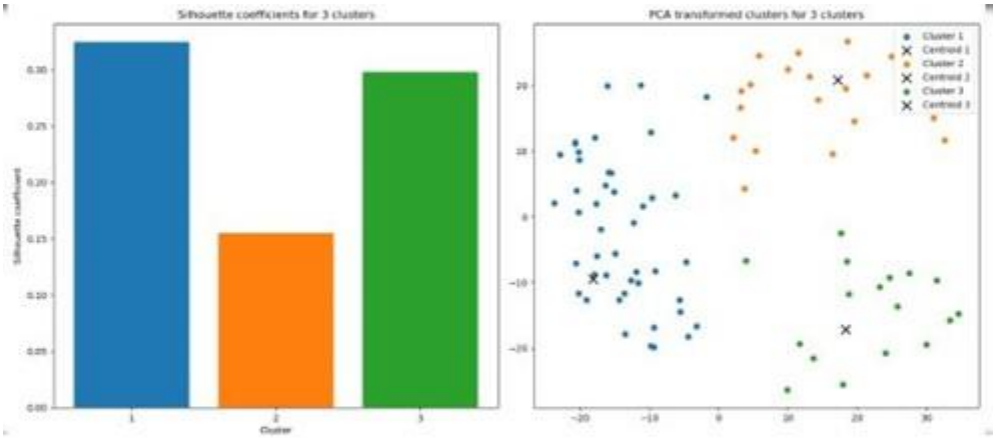


Fig. 5. Silhouette coefficient and PCA transformed cluster representation for three clusters

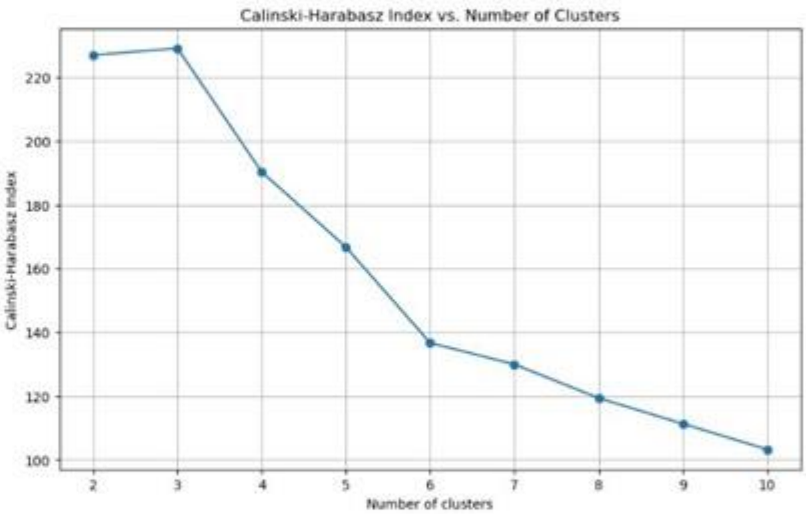


Fig. 6. Calinski-Harabasz Index for various cluster sizes

Figure 7 visualizes these clusters on map.

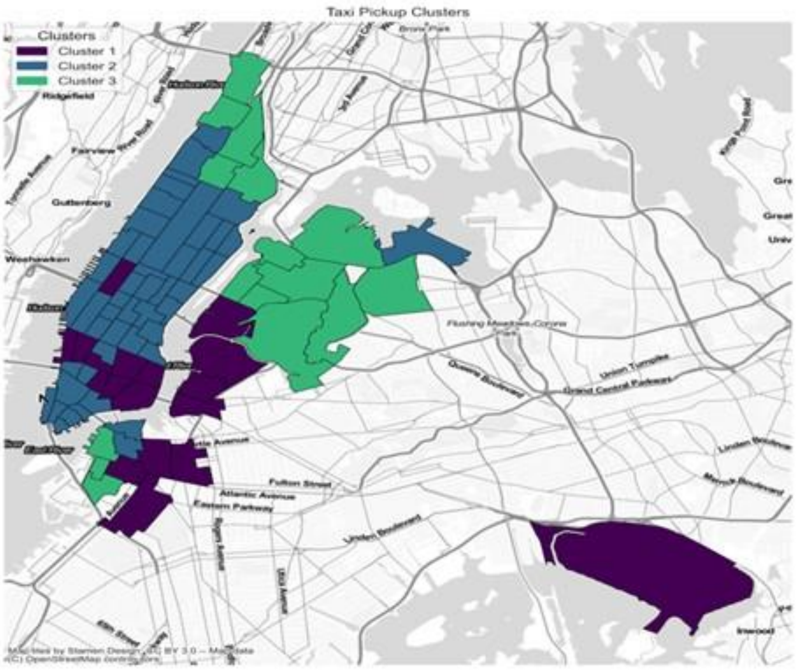


Fig. 7. Clusters visualized geographically formed by using k-shape clustering

It can be observed that cluster 2 is mostly centered around Manhattan Island where the average demand for the trips is high and is the largest cluster. Cluster 1 and 3 on the other hand have relatively lower demands.

The results showed that the time series cluster with more sparse values are difficult to predict as they create an abrupt change in the time series. The created cluster features however, improved the performance of the CNN-LSTM model to some degree as seen in Figure 8. This is promising in that the clustering approach can be further explored to improve the performance of standard models while modeling time-series that are very different from each other and improving the prediction in areas with low demand.

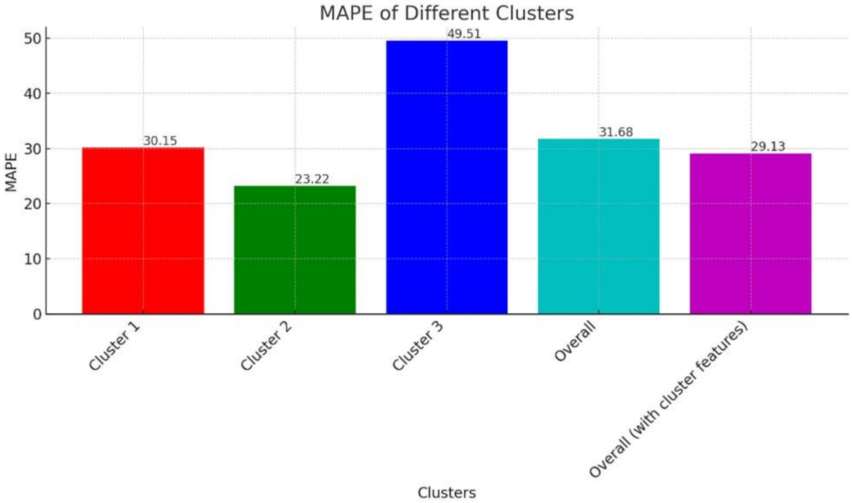


Fig. 8. Mean Absolute Percentage Error (MAPE) for various clusters with and without cluster features.

Similarly, the proposed ensemble model seems to perform better than all the baseline models. Both the component models used CNN-LSTM and XGBoost have an MAPE of over 22 but combining them reduced the MAPE to below as observed in Figure 9. This shows that combining models that are different in nature from each other can improve upon the accuracy of individual models. The proposed ensemble is applied to cluster 2, which has the highest demand to compare its relative performance in a higher demand setting.

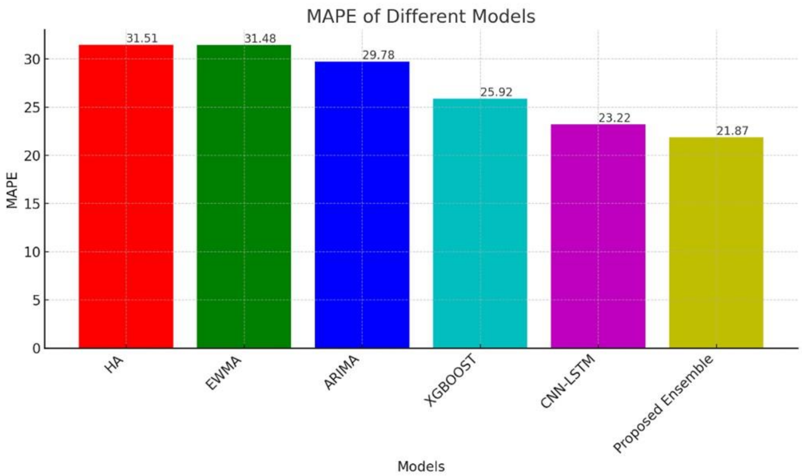


Fig. 9. Mean Absolute Percentage Error (MAPE) for baseline models vs. the proposed ensemble for Cluster 2.

Figure 10 shows the test versus prediction results comparison for a random location using the ensemble prediction.

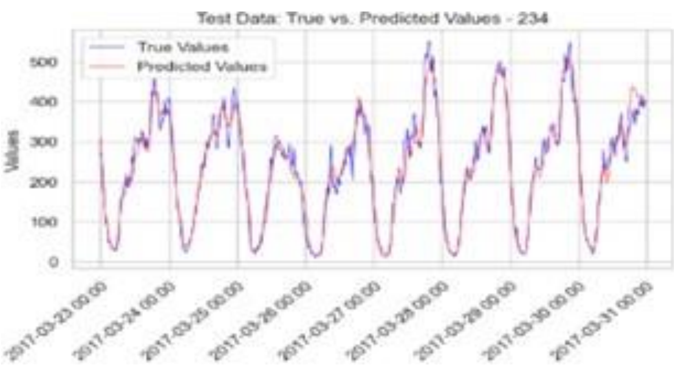


Fig. 10. Test Actual Values versus Test Prediction values for Location ID : 234.

Similarly, Figure 11 shows the distribution of test error (MAPE) across all 81 locations.

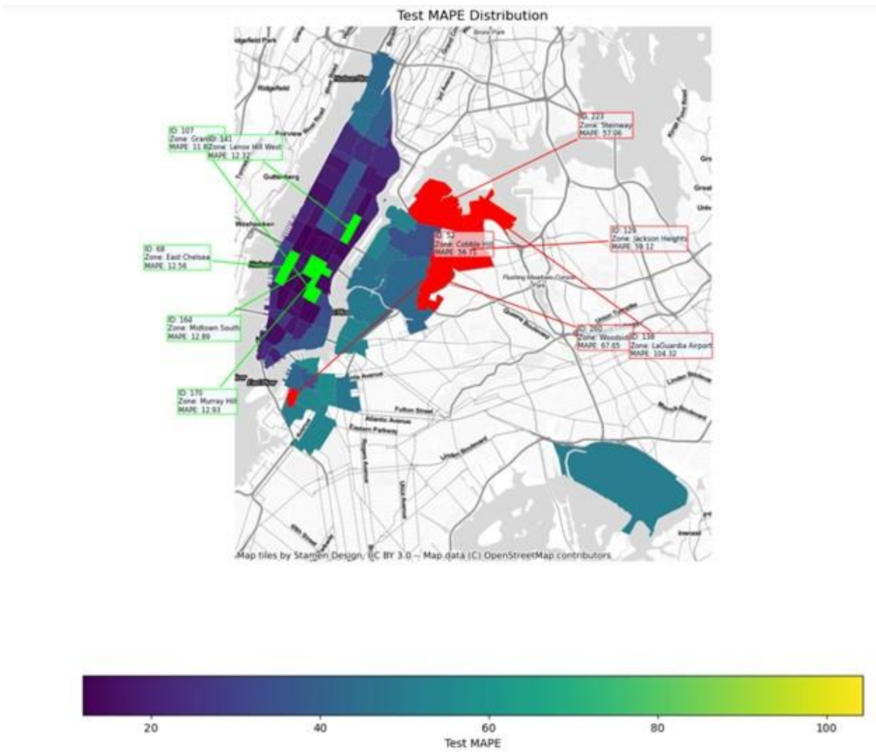


Fig. 11. Distribution of MAPE across various regions for the selected 81 locations, showing the top five locations with best prediction results and the bottom five locations with worst prediction results.

The green highlighted locations are the top five locations with least error or MAPE while the red highlighted locations are the top five locations with maximum error or MAPE. It can be observed that the locations with dense pickup demand also have a high prediction accuracy, whereas less dense areas and LaGuardia Airport which have irregular demand have the least prediction accuracy.

6. Conclusion

The study was set out to understand and predict the demand for the Yellow Cab Taxi service in New York City. The study proposed an ensemble model composed of XGBoost, a gradient boosting algorithm, and a convolutional neural network – long short-term memory (CNN-LSTM) model, with finely engineered features. Moreover, the study also investigated the use of time-series clustering using k-shape clustering, and whether features created out of those clusters improved the prediction accuracy in the presence of sparse demand.

The proposed approach has its origin in the development of time series forecasting techniques, which started with conventional statistical models and progressed to advanced deep learning techniques in the end. The combination of XGBoost and CNN-LSTM enabled us to capitalize on the distinct advantages of both models, namely XGBoost's robust performance across a variety of data structures and CNN-LSTM's proficiency in processing spatial and temporal data.

Findings from study can be summarized as below:

- Time series at various locations have different characteristics, and the performance of the model varies based on the type of time-series. Regions with sparse demand are difficult to model in comparison to the regions with higher demands.
- Time-series clustering using methods like k-shape and using indexes like Silhouette scores and Calinski-Harabasz index to find the correct number of clusters and can group similar time series together. These grouped time series can then be treated separately according to their nature to improve the forecast accuracy on each group.
- Cluster based features like inter-cluster similarity, intra-cluster similarity, cluster centroid mean, cluster standard deviation etc. can help the performance of the model when modeling different time-series together.
- Good feature engineering based on existing time-series data along with tree based boosting models can give competitive performance in comparison to the deep learning models.
- Ensemble learning leverages the strengths of different models and can improve the overall performance on the forecast.

All in all, this study holds promise for enhancing the operational efficiency of the Yellow Cab Taxi services and other similar ride-hailing services in urban environments. The generated cluster features as well as features like minimum and maximum values help capture some of the trends in location with sparse demand. However, further attention is required in this direction to help better capture the states of pickup demand in less dense locations. Another direction would be to explore other models and create a different ensemble with the recent deep learning models like transformers, although they are computationally more expensive. Future research can be done trying new time-series clustering methods and creating better cluster features to improve the model performance while modeling different types of time series together.

Acknowledgments

The study is part of the author’s original MS Thesis. ChatGPT version 4 is used for summarizing and paraphrasing certain portions to fit the manuscript to the TRB submission guidelines.

References

1. Hossain, M., & Muurlink, O. (2020). Current Issues in the Sharing Economy: The Debate about Uber. International Journal of Social Science Studies, 2020.
2. New York City Taxi and Limousine Commission. "About TLC". Retrieved 2023-06-03.
3. New York City Taxi and Limousine Commission. "2018 TLC Factbook". Retrieved 2023-06-03.
4. Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., & Damas, L. Predicting taxi–passenger demand using streaming data. IEEE Transactions on Intelligent Transportation Systems, 2013.
5. Yuan, J., Zheng, Y., Xie, X., & Sun, G. T-drive: Enhancing driving directions with taxi drivers' intelligence. IEEE Transactions on Knowledge and Data Engineering, 2011.
6. Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., Gong, P., Ye, J., & Li, Z. Deep multi-view spatial-temporal network for taxi demand prediction. AAAI, 2018, pages 2588–2595.
7. Hyndman, R. J., & Athanasopoulos, G. Forecasting: principles and practice. OTexts., 2018.
8. Liu, L., Andris, C., & Ratti, C. . Uncovering cabdrivers' behavior patterns from their digital traces. Computers, Environment and Urban Systems, 34(6)., 2010, pages 541-548.
9. Breiman, L. Random forests. Machine learning, 45(1), 2001, pages 5-32
10. Z. Liu, H. Chen, Y. Li, and Q. Zhang, Taxi demand prediction based on a combination forecasting model in hotspots, J. Adv. Transp, 2020.
11. Poongodi M, Malviya M, Kumar C, Hamdi M, Vijayakumar V, Nebhen J, Alyamani H. New York City taxi trip duration prediction using MLP and XGBoost. Int J Syst Assur Eng Manag. 2022;13(Suppl 1):16–27. doi: 10.1007/s13198-021-01130-x. Epub 2021 Jul 1. PMCID: PMC82482
12. Zhang, J., Zheng, Y., Qi, D. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), 2017.
13. Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., Gong, P., Ye, J., & Li, Z. Deep multi-view spatial-temporal network for taxi demand prediction. AAAI, 2018 pages 2588–2595.
14. Wu, Z., Pan, S., Long, G., Jiang, J., & Zhang, C. Graph wavenet for deep spatial-temporal graph modeling. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19., 2019
15. J. Choi, H. Choi, J. Hwang, and N. Park, “Graph neural controlled differential equations for traffic forecasting,” in Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22), 2022
16. Davis, Neema & Raina, Gaurav & Jagannathan, Krishna. A multi-level clustering approach for forecasting taxi travel demand. 223-228., 2016 10.1109/ITSC.2016.7795558.
17. Berndt, D. J., & Clifford, J. Using dynamic time warping to find patterns in time series. AAAIWS'94: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA., 1994.
18. Paparrizos, J., & Gravano, L. k-shape: Efficient and accurate clustering of time series. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015
19. Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., & Damas, L. Predicting taxi–passenger demand using streaming data. IEEE Transactions on Intelligent Transportation Systems, 14(3), 1393-1402. 2013
20. U. Vanichrujee, T. Horanont, W. Pattara-atikom, T. Theeramunkong and T. Shinozaki, "Taxi Demand Prediction using Ensemble Model Based on RNNs and XGBOOST," 2018 International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES), Khon Kaen, Thailand, 2018, pp. 1-6, doi: 10.1109/ICESIT-ICICTES.2018.8442063.
21. Weather Underground. Historical Weather: New York City, New York. Retrieved June 03, 2023, from https://www.wunderground.com/history/daily/us/ny/new-york-city/KLGA.
22. Office Holidays. (n.d.). United States Public Holidays 2017. Retrieved from https://www.officeholidays.com/countries/usa/2017
23. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65., 1987.
24. Calinski, T., & Harabasz, J.. A Dendrite Method for Cluster Analysis. Biometrics, 30(1), 101-107, 1974
25. Chen, T., & Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining., 2016.
26. Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in Neural Information Processing Systems, 802-810., 2015.