Journal of IoT in Social, Mobile, Analytics, and Cloud (ISSN: 2582-1369)
www.irojournals.com/iroismac/

Machine Learning Driven Smart Transportation Sharing

N. P. Shangaranarayanee[1], V. Aakashbabu[2], M. Balamurugan[3], R. Gokulraj[4]

[1]Assistant Professor, Department of Computer Science and Engineering, Erode Sengunthar engineering College, Perundurai, Erode, Tamilnadu, India.

[2,3,4] Student, Department of Computer Science and Engineering, Erode Sengunthar engineering College, Perundurai, Erode, Tamilnadu, India.

E-mail: [1]npsnarayaneeesec@gmail.com, [2]aakashbau114@gmail.com, [3]balamano243@gmail.com, [4]rgr.official30@gmail.com

Abstract

In many urban areas, traffic congestion has become one of the most challenging issues of modern life, resulting in detrimental effects on the environment, productivity loss, fuel wastage, and longer travel times. As a solution, people are increasingly turning to shared transportation modes due to the convenience of multi-modal journeys facilitated by smart transportation systems. The last mile problem refers to the fact that, in large cities, buses and trains deliver passengers to transit stations close to retail and job areas, leaving them needing another form of transportation to reach their final destination. By promoting the use of public transportation and addressing this issue, a smart bike-sharing system can contribute to reducing traffic congestion. The study presents a review of various methods that are associated with the designing of the bike sharing system and suggests a model incorporating various methods to derive solutions, with a focus on utilizing clustering algorithms for the analysis of the provided time series dataset. The study reveals that the application of algorithms such as the K-Means algorithm, Fuzzy C-means, etc. would be very effective in visualizing the resulting clusters and improve the forecasting accuracy.

Keywords: Demand Forecasting, Collaborative Computing, Bike Sharing System, Machine Learning

DOI: https://doi.org/10.36548/jismac.2024.1.001
Received: 11.12.2023, received in revised form: 18.01.2024, accepted: 01.02.2024, published: 13.02.2024

1. Introduction

A station is connected to the neighboring stations and is not isolated. We decided to include clustering analysis and meticulous feature analysis in this work because of the following discoveries. When a station is close to an office building but does not have enough bikes, more bikes must be sent to the station in order to meet demand. Typically, users would pick up bikes from adjacent stations in such a scenario. Because a single station's demand cannot account for the assistance that users seek from nearby stations, the demands of a cluster of nearby stations may more accurately reflect the overall demand of users. As a result, estimating the bike demand for individual stations only yields low accuracy. Mobile bike-sharing services are a green and innovative form of transportation that enhance the variety of municipal transportation options. They also serve as a practical means of commuting, assisting in the solution of the first-and-last mile issue in urban public transportation and significantly reducing traffic congestion. People can take advantage of the sharing service without having to purchase or ride their own bikes thanks to mobile bike-sharing services. Although mobile bike-sharing services bring much convenience, they also face many challenges. Bikes are frequently used in an erratic manner at many bike stations, and high demand frequently arises from a variety of circumstances, including time, place, and surroundings. When the demand for motorcycles is not met and the quality of service (QoS) declines, there are two common scenarios [15].

1.1 Forecasting the Demands

Demand forecasting is the process of predicting future demand based on customer needs over a specified time period. It often integrates historical data and additional analytical insights to generate the most accurate estimates. More precisely, demand forecasting involves leveraging predictive analytics on past data to predict and comprehend customer demand. This process aids in making critical supply decisions necessary for corporate success and understanding significant economic situations. Demand forecasting approaches can be broadly classified into two main groups: quantitative and qualitative [15]. Major cities all around the world have enacted public bike-sharing (PBS) programs to cut down on air pollution, ease traffic congestion, and minimize injury rates. Bicycles can be picked up and dropped off at any station by users, which may cause inventory imbalances. To improve system efficiency, system operators should therefore create appropriate repositioning strategies based on accurate demand estimates for bicycles. The project's goal is to estimate demand for bicycle pickups and drops at the station level by using station activity data. The number of pickups and drops

at a station one to three hours before to the prediction, in addition to time and weather data, will be utilized as a predictor [14]. Demand forecasting is crucial for companies across a range of industries, particularly when it comes to mitigating operational risks Nevertheless, it is widely acknowledged that businesses often struggle due to the complexities involved, especially in quantitative analysis. However, understanding customer needs is a vital component for every company, enabling more effective implementation of business plans and better alignment with market demands. Businesses stand to gain various advantages by embracing the concept of demand forecasting, including but not limited to reducing waste, optimizing resource allocation, and potentially experiencing significant increases in revenue and sales.

2. Related Work

The proposed study employs deep learning, particularly Bi-LSTM networks, to enhance short-term bike availability forecasts in urban sharing systems. By analyzing various factors, such as weather and past usage patterns, the model improves predictions, addressing challenges like irregular distribution and uncertainty, notably benefiting e-bike stations. [1]. The proposed method uses TOPSIS combined with entropy-based evaluation to optimize bike-sharing station rebalancing. This innovative approach enhances system efficiency by systematically assessing station importance and dynamically optimizing rebalancing strategies based on criteria such as capacity and user demand [2]. The method uses the Spatial-Temporal Memory Network (STMN) to forecast short-term bicycle demand in bike-sharing systems. STMN combines Convolutional Long Short-Term Memory models with a feature engineering framework to capture spatial-temporal correlations in usage data. It outperforms baseline models across various cities and usage rates, aiding in system optimization and resource allocation. [3-7]. employed Discrete Wavelet Transform (DWT) to reduce dimensionality and filter noise in bike-sharing system (BSS) time series data, enhancing pattern recognition. Additionally, Dynamic Time Warping (DTW)-based k-means clustering aids in identifying distinct bike usage patterns across different stations, contributing to a deeper understanding of user behavior in bike-sharing systems [7-9]. The study utilizes machine learning techniques to forecast bike-sharing demand, incorporating car accident data to improve prediction accuracy, particularly in urban areas where traffic incidents impact transportation patterns significantly. By integrating external factors like daily vehicle accidents, the research enhances the precision of predictive models such as random forest, XGBoost, and LightGBM, underscoring the

importance of incorporating external data for improved predictive modeling in the sharing economy [10-15].

3. Proposed Work

The proposed idea begins with dataset collection and preprocessing, loading the dataset for specified date range plotting, and analyzing station attributes, time series length, and usage statistics. A visual representation of year-wise time series data aids in understanding patterns within the dataset. Subsequently, the application of clustering algorithms to segment data points and the. Iterative model refinement and validation procedures ensure the reliability of clustering outcomes along with the potential optimizations and extensions for urban transportation planning based on clustering results, which are discussed for future consideration.

3.1 Architecture Diagram

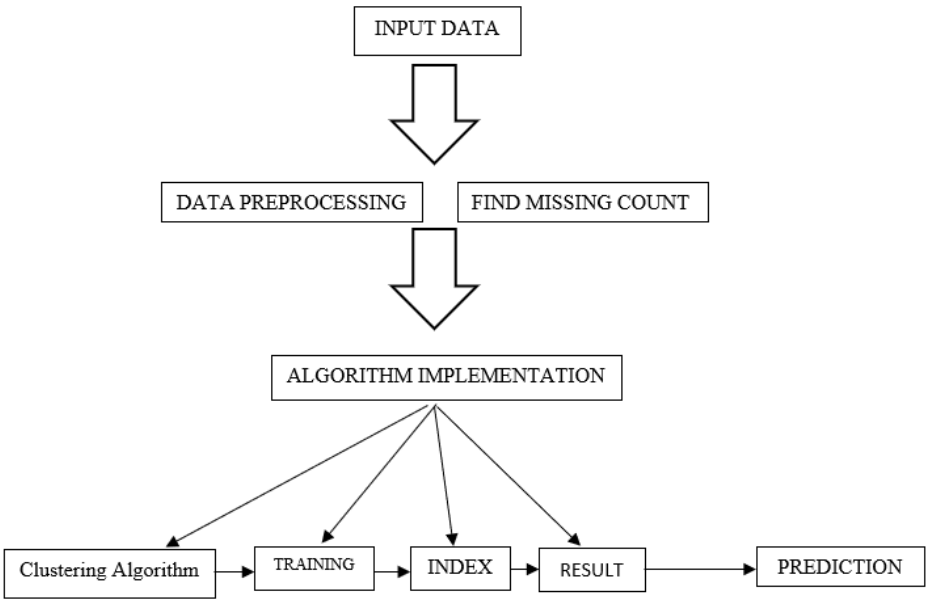

Figure 1. Architectural Diagram

3.2 Modules

3.2.1 Data preprocessing

- Data preprocessing is an essential stage in machine learning, encompassing the manipulation and refinement of raw data to prepare it for the construction and training of machine learning models. Data preprocessing can improve the quality and efficiency of the

data analysis and can also enhance the performance and accuracy of the machine learning models.

Some of the common steps involved in data preprocessing are:

- **Acquiring the Dataset:** This involves collecting or obtaining the relevant data from various sources, such as databases, files, web pages, sensors, etc. The data can be structured, unstructured, or semi-structured, depending on the format and organization of the data.

- **Importing the Libraries:** This involves importing the necessary libraries or modules in Python that can help with data preprocessing, such as pandas, numpy, scipy, sklearn, seaborn, matplotlib, etc. These libraries provide various functions and methods for manipulating, processing, and visualizing the data.

- **Importing the Dataset:** This involves loading or reading the data into a Python environment, such as a dataframe or an array, using functions like pd.read_csv(), np.loadtxt(), etc. The dataset can be in various formats, such as CSV, JSON, XML, etc.

- **Handling Missing Values:** This involves dealing with the missing or incomplete values in the dataset that can affect the quality and reliability of the data analysis. Missing values can be handled by various methods, such as deleting the rows or columns with missing values, replacing them with mean, median, mode, or a constant value, or using more advanced techniques like interpolation or imputation.

- **Handling Categorical Data:** This involves dealing with the categorical or nominal data in the dataset that can represent different classes or categories of information, such as gender, color, country, etc. Categorical data can be handled by various methods, such as encoding them into numerical values using label encoding or one-hot encoding, or using more advanced techniques like feature hashing or embedding.

- **Splitting the Dataset:** Dataset splitting is a critical process that entails partitioning the dataset into multiple subsets, which may include a training set and a test set, or a training set, validation set, and test set. The training set is utilized for constructing and training the machine learning model, whereas the test set is employed to assess

the model's performance and accuracy. The validation set aids in fine-tuning the model's hyperparameters to prevent overfitting or underfitting. This division of the dataset can be performed randomly or according to specified criteria using functions such as train_test_split() available in the sklearn library.

- **Feature Scaling:** This involves standardizing or normalizing the numerical values in the dataset to bring them to a common scale or range. Feature scaling can improve the speed and efficiency of the data analysis and can also enhance the performance and accuracy of some machine learning models that are sensitive to the scale of the features. Feature scaling can be done by various methods, such as min-max scaling, standard scaling, z-score scaling, etc. using functions like MinMaxScaler(), StandardScaler(), etc. from sklearn library.

3.2.2 Finding Missing Count

This module is focused on identifying and quantifying missing or null values within the dataset. It determines the number of missing values for each attribute or feature, essential for understanding data completeness before analysis or modeling.

The module takes the raw dataset containing bike-sharing system records as input. It systematically scans through each attribute or feature in the dataset to identify missing or null values. For each attribute, the module checks for any missing or null entries using appropriate functions or methods. It then counts the number of missing values for each attribute and compiles this information into a summary report. The module produces a comprehensive report detailing the count of missing values for each attribute, which can be used to guide further data preprocessing steps such as imputation or feature selection.

By systematically identifying and quantifying missing values within the dataset, the "Finding Missing Count" module ensures that subsequent analyses are based on complete and reliable data, thereby enhancing the robustness of insights derived from the bike-sharing system data.

3.2.3 Clustering Algorithm

The suggested model based on the study intends to use the Clustering algorithms such as K-means as a pivotal segmentation tool, for partitioning bike-sharing system data into distinct clusters based on similarities among data points. This algorithm uncovers inherent patterns and structures within the dataset, facilitating the identification of usage patterns,

station behaviors, and temporal trends. The resulting clusters will offer a valuable insight for decision-making in urban transportation planning, resource allocation, and system optimization. As K-Means scalability and computational efficiency make it well-suited for handling large volumes of time series records or station attributes encountered in bike-sharing datasets. The clustering outcomes undergo rigorous evaluation to ensure coherence and effectiveness, with scope for iterative refinement to optimize model parameters and accurately represent meaningful patterns within the data.

Steps in K-Means Implementation:

- **Initialization:**
  - Randomly select k initial centroids.
- **Assignment:**
  - Assign each data point to the nearest centroid.
- **Update Centroids:**
  - Recalculate centroids based on the mean of data points in each cluster.
- **Iteration:**
  - Repeat the assignment and centroid update steps until convergence.

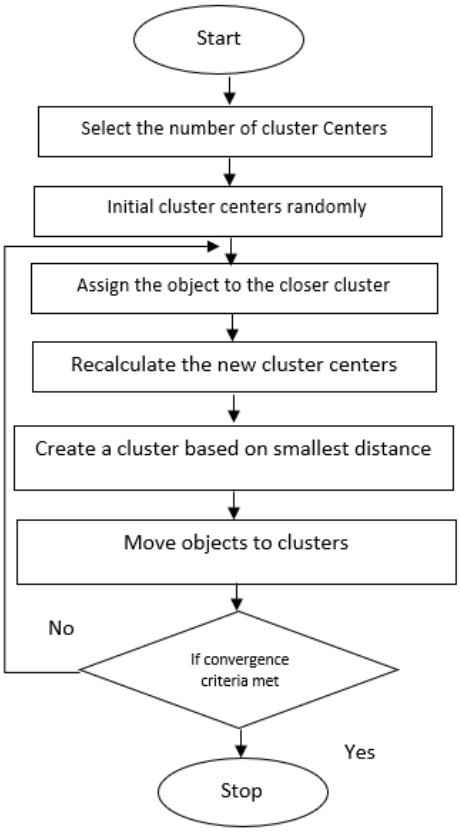

Figure 2. K-means Algorithm

4. Discussion

The discussion section, mentions the current phase of the research. Based on the study the research has decided with the dataset to be used in the proposed model. The proposed model uses the dataset that is manually collected from the website using the key search terms like Bike-sharing data, Bike-sharing usage, Bike-sharing system analysis, Bike-sharing stations, Bike-sharing trip data, Bike-sharing app data etc. and has decided to preprocess the data in Python programming language using the libraries like pandas for handling missing data and removing duplicates, MinMaxScaler from sklearn.preprocessing for normalization and scaling, SMOTE from imbalanced-learn to handle imbalanced data and apply the Recursive Feature Elimination from sklearn.feature_selection to select the appropriate features to train the model. Further the proposed method utilizes the K-means algorithm to cluster the collected dataset and ease the process of prediction.

4.1 Dataset Description

The dataset was retrieved from the web source and consists of a time series dataset with 8749 rows and attributes such as time and usage. The reindex value for the time series dataset is 8760, based solely on the time attribute. There are a total of 2000 CSV files, each containing 8760 rows.

In the evaluation of the dataset, three key parameters were utilized: the Silhouette Index, Calinski Index, and Davies Bouldin Index. The Silhouette Index provides insight into the consistency of data clusters by visually depicting each object's classification accuracy. Following this, the Calinski-Harabasz (CH) Index, devised in 1974 by Calinski and Harabasz, emerges as a valuable tool for assessing clustering models when ground truth labels are unavailable. It evaluates clustering quality based on intrinsic dataset characteristics. Additionally, the Davies Bouldin Index serves as an internal evaluation technique, leveraging the dataset's inherent quantities and features to validate clustering quality. It's important to note that while a favorable result from this method is indicative of good clustering, it may not always correlate with the best information retrieval.

By assessing clustering quality based on dataset features rather than external benchmarks, these parameters offer a robust evaluation approach. Furthermore, the comprehensive evaluation facilitated by multiple parameters ensures a nuanced understanding of clustering performance, empowering researchers and practitioners to make informed decisions in data analysis and pattern recognition tasks.

4.2 Benefits

- It provides computationally efficient and scalable for large datasets. Clustering process would help in identifying usage patterns within bike sharing data, enabling the system to recognize trends and regularities in demand based on factors like time of day, weather, or events.

- Clustering divides data into distinct groups, allowing focused analysis and prediction for specific station clusters. This segmentation enhances the accuracy of predictions by customizing models for each cluster's unique characteristics.

- Accurate demand prediction within clusters enables optimal resource allocation, including redistributing bikes based on predicted demand. This ensures stations are stocked appropriately, enhancing operational efficiency and user satisfaction.

- Clustering in bike sharing systems is that it can easily adjust to changes in the locations of bike stations.

4.3 Future Work

The future work of the research will involve implementing and evaluating the performance of the proposed model, followed by its deployment for real-world use. Additionally, we aim to develop a user interface to address the challenges of the bike-sharing system and provide a more reliable application for users

5. Conclusion

The study presents details of the bike-sharing system and proposes a model utilizing machine learning to enhance shared transportation. The suggested model includes data collection processes, preprocessing techniques, and evaluation of datasets using the Silhouette Index, Calinski Index, and Davies Bouldin Index. Furthermore, the model employs the K-means clustering algorithm for dataset clustering. In future iterations, the research will involve implementing and evaluating the model's performance, followed by its deployment for real-world use. Additionally, we aim to develop a user interface to address the challenges of the bike-sharing system and provide a more reliable application for users.

References

[1] The Meddin Bike-Sharing World Map, Bike-Sharing Word Map, 2021. https://bike sharingworldmap.com. (Accessed 18 April 2021).

[2] Citi Bike System, Citi bike system data. https://www.citibikenyc.com/system-data, 2021. (Accessed 21 April 2021).

[3] F. Chiariotti, C. Pielli, A. Zanella, M. Zorzi, A bike-sharing optimization framework combining dynamic rebalancing and user incentives, ACM Trans. Autonom. Adapt. Syst. 14 (11) (2020) 1–30.

[4] Contardo, Claudio, Catherine Morency, and Louis-Martin Rousseau. Balancing a dynamic public bike-sharing system. Vol. 4. Montreal: Cirrelt, 2012.

[5] Schuijbroek, Jasper, Robert C. Hampshire, and W-J. Van Hoeve. "Inventory rebalancing and vehicle routing in bike sharing systems." European Journal of Operational Research 257, no. 3 (2017): 992-1004.

[6] J. Liu, L. Sun, W. Chen, H. Xiong, Rebalancing bike sharing systems: a multi-source data smart optimization, in: Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2016, pp. 1005–1014.

[7] Y. Liang, S. Ke, J. Zhang, X. Yi, Y. Zheng, Geoman: multi-level attention networks for geo-sensory time series prediction, in: Proceedings of International Joint Conference on Artificial Intelligence, IJCAI, 2018, pp. 3428–3434.

[8] Y. Li, Y. Zheng, H. Zhang, L. Chen, Traffic prediction in a bike-sharing system, in: Proceedings of the 23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2015, pp. 1–10.

[9] Gebhart, Kyle, and Robert B. Noland. "The impact of weather conditions on bikeshare trips in Washington, DC." Transportation 41 (2014): 1205-1225.

[10] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[11] Frade, Inês, Anabela Ribeiro, Daniela Dias, and Oxana Tchepel. "Bike sharing systems implementation impact on emissions, for cyclist preferred routes in urban areas." International Journal of Sustainable Transportation 16, no. 10 (2022): 901-909.

[12] Faghih-Imani, R. Hampshire, L. Marla, N. Eluru, An empirical analysis of bike sharing usage and rebalancing: evidence from barcelona and seville, Transport. Res. Pol. Pract. 97 (2017) 177–191.

[13] J. Bao, T. He, S. Ruan, Y. Li, Y. Zheng, Planning bike lanes based on sharing-bikes' trajectories, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 1377–1386.

[14] Tekouabou, Stephane Cedric Koumetio. "Intelligent management of bike sharing in smart cities using machine learning and Internet of Things." Sustainable Cities and Society 67 (2021): 102702.

[15] Yang, Xiaoxian, Yueshen Xu, Yishan Zhou, Shengli Song, and Yinchen Wu. "Demand-aware mobile bike-sharing service using collaborative computing and information fusion in 5G IoT environment." Digital Communications and Networks 8, no. 6 (2022): 984-994.

Author's biography

Er. Shangaranarayanee N. P. working as Assistant Professor in Erode Sengunthar Engineering College, Thudupathi. Worked as Associate in Software Development at US based company in Freelance, Freelance developer on RCP Plugin development environment and Game strategy planning on Android. She Completed Master of Engineering in Computer Science and Engineering and completed Bachelor of Computer Science and Engineering at Angel College of Engineering and Technology, Tirupur. Have published 2 National Conference papers, 2 International conference papers, 2 article in International Journal and performed 4 Technical Paper Presentations. Her research interests include Data Structures, Artificial Intelligence and Data Science.

V. Aakashbabu is presently pursuing an undergraduate degree in Computer Science and Engineering at Erode Sengunthar Engineering College. He is interested in the web development domain and has developed a project. Additionally, he is currently taking a Data Analytics course on Coursera. He has also participated in symposiums and obtained certificates. Furthermore, he participated in a CYBERSECURITY WORKSHOP and a

MOBILE DEVELOPMENT WORKSHOP, earning certificates for both. He completed a JavaScript course at Infosys Springboard. He also likes to learn new technologies and gain knowledge from them.

M. Balamurugan is presently pursuing an undergraduate degree in Computer Science and Engineering at Erode Sengunthar Engineering College. He has completed a CLOUD COMPUTING course at Infosys Springboard. Additionally, he participated in an AWS WORKSHOP and a MOBILE DEVELOPMENT WORKSHOP, earning certificates for both.

R. Gokulraj is presently pursuing an undergraduate degree in Computer Science and Engineering at Erode Sengunthar Engineering College. He has completed a JAVA course at Infosys Springboard. Additionally, he participated in an AWS WORKSHOP and obtained certificates.