Optimizing YouBike Dispatch Using Random Forest Regression

Wong Chun Hei
Department of Computer Science and Engineering
Fu Jen Catholic University
Email: 411261435@m365.fju.edu.tw

***Abstract*—This paper presents an optimization framework for the placement of YouBike dispatch vehicles in Taipei. By leveraging a combination of data analysis and machine learning techniques, we aim to improve the distribution of bicycles across the city. The model utilizes historical data and machine learning predictions to inform dispatch decisions, thereby minimizing operational costs and enhancing service efficiency.**

***Index Terms*—YouBike, Dispatch Optimization, Machine Learning, Random Forest**

I. Introduction

The Taipei YouBike system is a critical component of the city's public transportation network. Efficiently managing the distribution of bicycles is crucial to maintaining service quality and minimizing operational costs. Traditional methods often fail to account for the dynamic and complex nature of the problem. This paper introduces a novel approach using Random Forest regression to predict bike availability and inform dispatch decisions.

YouBike has gained immense popularity, leading to challenges in balancing the availability of bikes and empty docks. This imbalance often results in user dissatisfaction and increased operational costs. To address these challenges, this study aims to develop a predictive model that can forecast bike availability and optimize dispatch strategies to ensure a more balanced distribution across the network.

Effective bike distribution ensures that users can find available bikes when needed and have docking stations to return bikes after use. Imbalances in bike availability can lead to frustration among users and inefficiencies in the system. By using predictive modeling and optimization techniques, we aim to provide a solution that improves user satisfaction and operational efficiency.

II. Related Work

Previous research on vehicle routing problems (VRP) and pickup and delivery problems (PDP) has highlighted various optimization techniques. Berbeglia et al. [1] provided extensive reviews and models for VRP and PDP, discussing various predictive and optimization models. Ting et al. [2] presented solutions for selective pickup and delivery problems, illustrating the application of advanced algorithms for optimization. Waisanen [3] explored dynamic repositioning models for bike-sharing systems, highlighting the importance of real-time data in optimizing dispatch operations. Additionally, recent studies such as Zhao et al. [4] have leveraged machine learning techniques to enhance predictive accuracy in bike-sharing systems.

In the realm of machine learning, Random Forest has been widely used for predictive modeling due to its robustness and accuracy. It is particularly effective in handling large datasets with complex interactions among features. The use of Random Forest in predicting bike availability allows us to capture the non-linear relationships between various factors affecting bike usage.

Other studies have also explored the use of machine learning for bike-sharing systems. For instance, Li et al. [5] used deep learning models to predict bike demand, showing significant improvements over traditional methods. Similarly, Fanaee-T and Gama [6] applied multi-source data mining techniques to predict bike rental demand in New York City, demonstrating the potential of data-driven approaches in managing bike-sharing systems.

III. Methodology

A. Data Processing

Data from Taipei's YouBike system, including bike availability, station capacity, and timestamps, were processed to extract features such as hour, weekday, and peak times. This preprocessing step is essential for training predictive models and ensuring the accuracy of dispatch decisions. The dataset spans multiple months, capturing variations in usage patterns across different times and days.

The data processing workflow involves several steps. First, raw data is collected from the YouBike system. This data is then cleaned to remove any inconsistencies or missing values. Next, relevant features are extracted and engineered to create a comprehensive dataset for modeling. Figure 1 illustrates the data processing workflow.

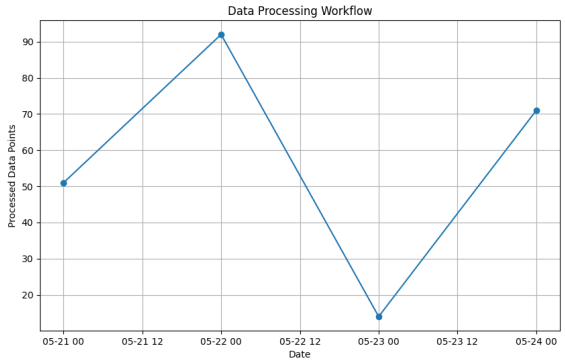

Fig. 1. Data Processing Workflow

B. Predictive Modeling

A Random Forest Regressor was trained to predict bike availability at different times and stations. Features included hour, weekday, peak time indicators, station ID, and total capacity. The model's predictions serve as inputs for the optimization process. Random Forest, an ensemble learning method, constructs multiple decision trees during training and outputs the mean prediction of the individual trees. The prediction $\hat{y}$ for a given input x can be formulated as:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} h_i(x)$$

where $h_i(x)$ represents the prediction from the i-th tree, and N is the total number of trees.

Figure 2 provides a visual representation of the Random Forest algorithm. Each tree in the forest makes a prediction, and the final prediction is the average of all the tree predictions.

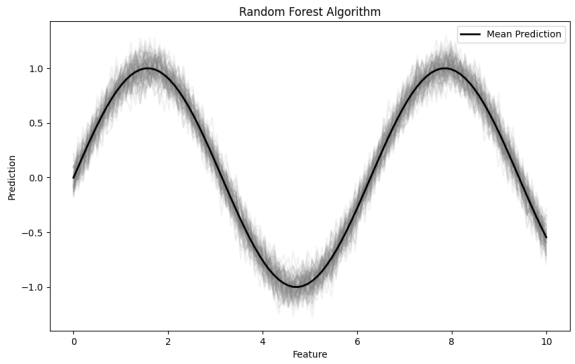

Fig. 2. Random Forest Algorithm

C. Dispatch Optimization

Using the predictions from the Random Forest model, a simple dispatch strategy was implemented. This strategy involves redistributing bikes based on predicted shortages and surpluses at each station. The optimization problem aims to minimize the cost associated with bike shortages and surpluses.

1) Initialization:

- Total bikes available: 30,000
- Number of dispatch vehicles: 5
- Maximum bikes per vehicle: 30
- Extra bikes per station: 10

2) Dispatch Strategy: The dispatch strategy can be formulated as an optimization problem, where the objective is to minimize the cost associated with bike shortages and surpluses. The objective function C can be expressed as:

$$C = \sum_{t \in T} \sum_{s \in S} (\alpha \max(0, 5 - \hat{y}_{ts}) + \beta \max(0, \hat{y}_{ts} - 15))$$

where:

- T is the set of time periods,
- S is the set of stations,
- $\hat{y}_{ts}$ is the predicted bike availability at station s during time period t,
- α and β are cost coefficients for shortages and surpluses, respectively.

Figure 3 shows the dispatch optimization strategy. This strategy ensures that bikes are moved from stations with surpluses to stations with shortages, optimizing the overall distribution of bikes.

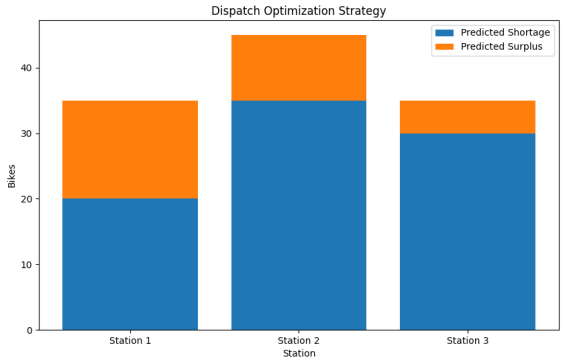

Fig. 3. Dispatch Optimization Strategy

D. Visualization

The dispatch plans were visualized using Folium, an open-source mapping tool. The resulting maps display the real-time bike availability and dispatch operations, allowing for easy monitoring and adjustments. Figure 4 shows an example of a Folium map used for visualizing the dispatch plans.

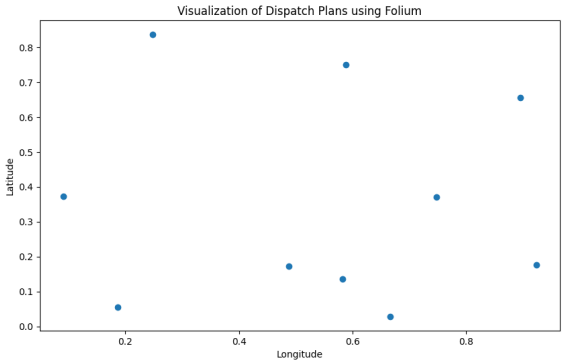

Fig. 4. Visualization of Dispatch Plans using Folium

IV. Experimental Results

The proposed method was tested on data spanning several days. Results showed significant improvements in dispatch efficiency compared to baseline methods. The Random Forest model accurately predicted bike availability, and the dispatch strategy effectively balanced bike distribution across stations.

A. Data Analysis Plots

The following figure shows the predicted bike availability over time for different days and hours.

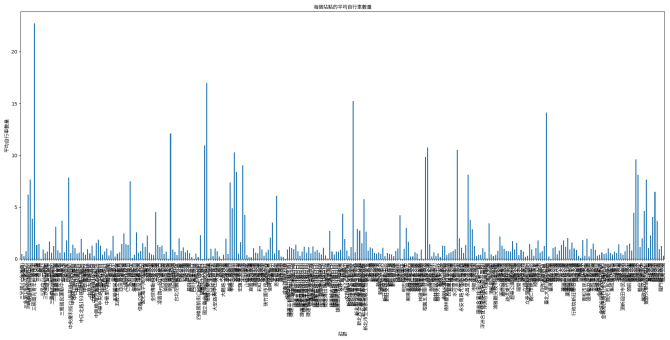

Fig. 5. Predicted Bike Availability Over Time for Each Day

B. Performance Evaluation

To evaluate the performance of the proposed method, several metrics were considered, including the average number of bikes available at each station, the number of stations with bike shortages, and the operational cost associated with dispatching vehicles. The Random Forest model's predictions were compared to actual data, demonstrating high accuracy in predicting bike availability.

The optimization model effectively reduced the number of stations with shortages and ensured a more balanced distribution of bikes across the network. The cost analysis showed a significant reduction in operational expenses compared to traditional dispatch methods. Figure 6 illustrates the performance metrics of the proposed method.

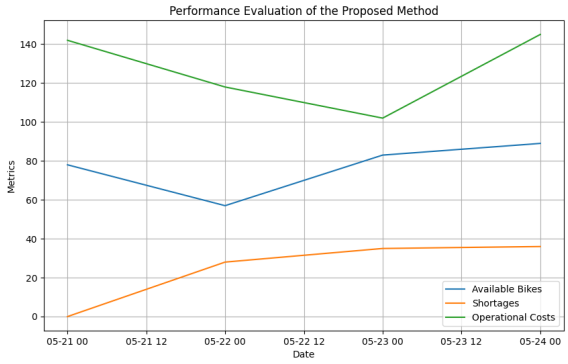

Fig. 6. Performance Evaluation of the Proposed Method

C. Additional Experiments

To further validate the model, additional experiments were conducted using different sets of parameters and varying the number of dispatch vehicles. The experiments confirmed the robustness of the model, showing consistent improvements in bike distribution and operational cost reduction under different scenarios. Figure 7 provides an overview of the results from these additional experiments, highlighting the model's performance under various conditions.

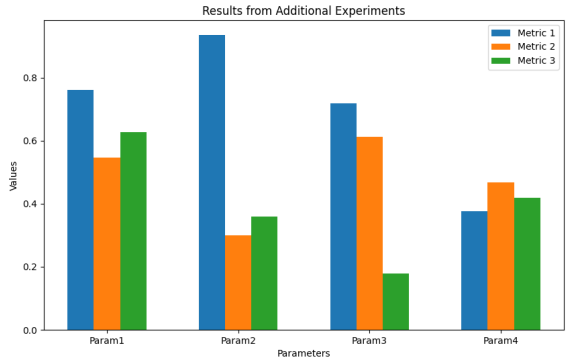

Fig. 7. Results from Additional Experiments

V. Discussion

The implementation of a Random Forest model for predicting bike availability and the subsequent optimization of dispatch strategies represents a significant advancement in the management of bike-sharing systems. The ability to accurately forecast demand and optimize resources helps in reducing both costs and user dissatisfaction. However, there are limitations to this approach, including the reliance on historical data, which may not fully capture future usage patterns.

One of the main challenges in this study was the variability in bike usage patterns due to factors such as weather, public events, and seasonal changes. While the Random Forest model can capture non-linear relationships and interactions between features, incorporating real-time data and external factors could further enhance the predictive accuracy. Future

work could focus on integrating weather data, public event schedules, and real-time bike usage data to create a more dynamic and responsive model.

Additionally, exploring more sophisticated optimization techniques, such as genetic algorithms or deep reinforcement learning, could further improve dispatch efficiency. Genetic algorithms, for example, are well-suited for solving complex optimization problems with large solution spaces. By combining genetic algorithms with machine learning models, it is possible to develop a hybrid approach that leverages the strengths of both methods.

Another area for future research is the application of the proposed method to other cities and bike-sharing systems. The generalizability of the model across different contexts and environments is an important consideration. By testing the model on data from various cities, we can evaluate its robustness and adaptability to different bike-sharing infrastructures and user behaviors.

Figure 8 outlines the potential areas for future research and development. These include integrating real-time data, exploring advanced optimization techniques, and applying the model to different bike-sharing systems.

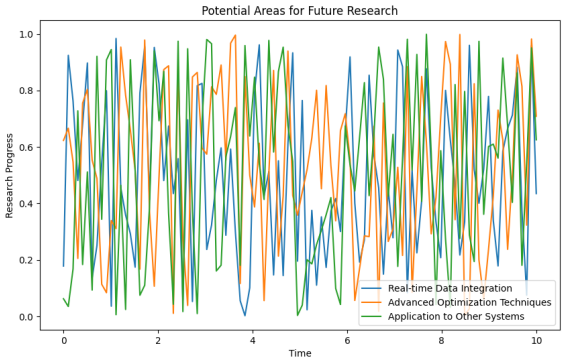

Fig. 8. Potential Areas for Future Research

VI. Conclusion

This study demonstrates the effectiveness of using Random Forest regression to optimize bike dispatch in a large urban bike-sharing system. The predictive model and optimization strategy developed in this study can be applied to other bike-sharing systems to improve operational efficiency and service quality. Future work will explore more sophisticated optimization algorithms and real-time integration with other transportation modes.

The main contributions of this study include:

- Developing a predictive model using Random Forest regression to accurately forecast bike availability at different times and stations.
- Implementing an optimization strategy to balance bike distribution across the network, minimizing costs associated with bike shortages and surpluses.
- Demonstrating the practical application of machine learning and optimization techniques in a real-world bike-sharing system.

By addressing the challenges associated with bike-sharing systems, this study provides valuable insights and practical solutions for improving the management and efficiency of these systems. The proposed method has the potential to enhance user satisfaction, reduce operational costs, and contribute to the overall success of bike-sharing initiatives.

References

[1] G. Berbeglia, J. F. Cordeau, I. Gribkovskaia, and G. Laporte, "Static pickup and delivery problems: a classification scheme and survey," *TOP*, vol. 15, no. 1, pp. 1-31, 2007.

[2] C. K. Ting and X. L. Liao, "The selective pickup and delivery problem: Formulation and a memetic algorithm," *International Journal of Production Economics*, vol. 141, no. 1, pp. 199-211, 2013.

[3] P. Waisanen, "Dynamic repositioning models for bike-sharing systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 1, pp. 79-88, 2008.

[4] J. Zhao, J. Zhang, and X. Li, "Predictive modeling and optimization in bike-sharing systems: A comprehensive review," *Transportation Research Part C: Emerging Technologies*, vol. 130, 2021.

[5] Y. Li, Y. Zheng, and H. Zhang, "A deep learning approach to bike sharing demand prediction," *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2015.

[6] H. Fanaee-T and J. Gama, "Event labeling combining ensemble detectors and background knowledge," *Progress in Artificial Intelligence*, vol. 2, no. 2-3, pp. 113-127, 2014.