Bike-Sharing Demand Prediction in Montreal: A Data-Driven, Interpretable Approach

Keywords: bike-sharing, sustainable transportation, machine learning, demand prediction.

1 INTRODUCTION AND BACKGROUND

Transportation is one of the major sources of GHG emissions and climate change, and it contributes to nearly 24% of CO_2 emissions around the world. To address this, a shift from car use (the dominant mode in North America) to more sustainable transportation modes is required (*1*). One of the possible approaches could be promoting bike-sharing systems since they provide us with many advantages, such as more physical activity, reduced emissions, flexible mobility, reduced fuel dependency and congestion, multimodal transport connections, and financial savings (*2*). Interestingly, bike-sharing can also promote the transition to cycling. Station-based bike-sharing systems include many ready-to-use bicycles in different stations (docks), and passengers can pick up a bicycle at a station and return it at any station near their destination. These systems can promote cycling through convenience and facilities such as easy access to bicycles, ease of payment, and simple membership procedures (*3*).

As a crucial responsibility in managing bike-sharing systems, the travel demand should be predicted, which can help facilitate the relocation of bicycles and optimize the location of new bike-sharing docks (*4*). Therefore, researchers have developed prediction models to predict bike-sharing demand, and some of their recent studies are summarized in Table 1.

Table 1 – *A summary of recent studies on bike-sharing demand prediction*

Reference	The most accurate method	Location	Independent variables					Metrics			
			Historical data	Infrastructure	Weather	Accessibility	Socio-demographics	MAE (trip/h)	RMSE (trip/h)	R^2	Prediction level
Li et al. (*4*)	IrConv+LSTM	Chicago	x	-	-	-	-	2.204	4.335	-	Adjacent cells
Li et al. (*4*)	IrConv+LSTM	Washington, D.C.	x	-	-	-	-	1.969	3.359	-	Adjacent cells
Li et al. (*4*)	IrConv+LSTM	New York	x	-	-	-	-	5.776	10.814	-	Adjacent cells
Li et al. (*4*)	IrConv+LSTM	London	x	-	-	-	-	3.578	6.296	-	Adjacent cells
Sathishkumar et al. (*5*)	Gradient Boosting	Seoul	x	-	x	-	-	109.78	172.73	0.92	District
Yang et al. (*6*)	Deep learning	New York	x	-	x	-	-	-	8.114	-	Regional
Yang et al. (*6*)	Deep learning	Chicago	x	-	x	-	-	-	5.268	-	Regional
Pan et al. (*7*)	Deep LSTM	New York City and Jersey City	x	-	x	-	-	-	2.712	-	Station
Hulot et al. (*8*)	Gradient Boosting	Montreal	x	-	x	-	-		1.394	0.59	Station
Sathishkumar and Cho (*9*)	CUBIST	Seoul	x	-	x	-	-	78.45	139.64	0.95	District

A recent review demonstrated that previous trips, weather conditions, land use, built environment, access to public transport, suitable cycling infrastructure, and socio-demographic variables strongly influence bike-sharing demand (*3*). However, these variables have not yet been used together to predict bike-sharing demand. The objectives and contributions of this study are as follows:

- Incorporate all key variables (e.g., historical data, weather conditions, land use, built environment, accessibility measures, deprivation measures, cycling infrastructure, and socio-demographics) into a single model to achieve the highest possible accuracy in predicting bike-sharing demand.

- Prioritize and assess the relative impact of these variables on bike-sharing demand using an interpretable ensemble learning approach.
- Capture the non-linear relationship between the key variables and bike-sharing demand.

2 METHODS

2.1 Data

Data from nine different sources are fused to generate the final dataset. These sources included different information at the dissemination area (DA) level. A DA is a small area with at least a neighboring dissemination block and Canada's smallest conventional geographic area. The population of DAs across Canada was targeted to be 400 to 700 persons (*10*). The applied data sources in this study include bike-sharing trips (Bixi), Canada Census data, Canada proximity measure data, deprivation index dataset, walk Score, DA dataset, the Canadian Bikeway Comfort and Safety (Can-BICS), Open Street Map, and Canada Weather Stats. Bixi is a station-based bike-sharing system in Montreal, Canada (the case study of this study), and its dataset includes the start time, origin station, end time, and destination station of all trips. From this data, all the trips from April to October in the recent three years (i.e., 2022, 2023, and 2024) are considered. The number of trips in the mentioned duration was over 19.6 million. For more information about Canadian Census data, Canada proximity measure data, deprivation index dataset, and walk Score, please read Naseri et al. (*1*). Can-BICS is a measure of cycling infrastructure indicating the weighted length of bikeways within a one km buffer. In this measure, the lengths of high-comfort, medium-comfort, and low-comfort bikeways are multiplied by 3, 2, and 1 (*11*). Open Street Map is applied to calculate the average cycling distance of Bixi docks in each DA to the Montreal city center. Canada Weather Stats (www.weatherstats.ca) is a database, which stores the historical weather data in Canada.

The data is prepared at the weekly level. For example, the temperature is the average temperature of the week, and the temperature variation is the average temperature variation of seven days in the week. The dependent variable is the number of trips per hour in each DA. To calculate this, the number of trips during weekday peak hours, weekday off-peak hours, and weekend hours are counted for each week, and these values are divided by the respective number of hours in each period (30, 90, and 48 hours). The generated dataset contained 30 variables. The correlation between variables is tested using the Pearson correlation coefficient. Seven variables are excluded from the dataset due to high correlation, and the final dataset includes 23 variables: week (represents *i*th week of the year), day type (weekday peak hours, weekday off-peak hours, and weekend hours), Walk Score, cycling distance to the city center, number of stations (Bixi docks) in the DA, CanBICs, proximity to parks, proximity to transit stations, material deprivation index, social deprivation index, the average income of households, percentage of French speakers, the percentage of people whose highest level of education is a high school diploma, the percentage of people whose highest level of education is postsecondary certificate below bachelor, employment rate, population density, population, year (2022, 2023, or 2024), temperature variation, average temperature, average precipitation, percentage of people aged below 15, and percentage of people aged over 64. The final dataset includes 23 independent variables and 95,874 data observations.

2.2 Modeling

This study proposes an accurate model to predict bike-sharing demand and identify the determinants of using the bike-sharing system. To this end, the Light Gradient Boosting Machine (LightGBM) is used for modeling. LightGBM is a new ensemble learning technique developed by Microsoft. This method is super-fast and appropriate for big data analyses since it supports parallel learning using lower memory usage (*12*). Moreover, LightGBM outperformed many other machine learning techniques and statistical analyses when comparing prediction accuracy and computational cost (e.g., (*1*).). In the modeling, 80% of the data is considered training data and the remaining 20% is used as testing data to evaluate the prediction power of the model. Further, k-fold cross-validation (considering k=5) and Optuna were simultaneously used to tune the hyperparameters of LightGBM. The performance of the model is evaluated using root mean square error, mean absolute error (MAE), the coefficient of determination (R^2), mean absolute percentage error (MAPE), and the percentage of data observations with an error of less than 30% (A30). For interpreting the results of LightGBM, SHapley Additive exPlanations (SHAP) and Partial Dependence Plots (PDP) are employed. For more details about these methods, please read Naseri et al. (*13*).

3 RESULTS AND DISCUSSIONS

The performance of the model on testing and five-fold validation data and the testing data error histogram are presented in Figure 1. As shown, the performance of testing data is approximately the average performance of validation datasets, indicating that the model performs consistently across different datasets. This trend confirms that the model generalizes well, and its predictive power is stable. The final model predicts the testing data with an RMSE of 0.769 trip/hour, an MAE of 0.383 trip/hour, and an R^2 of 0.971. These values demonstrate the superior performance of the proposed model compared to the literature (please see Table 1). In this table, the minimum RMSE, minimum

MAE, and maximum R^2 are 1.394 trip/hour, 1.969 trip/hour, and 0.95. Hence, it can be postulated that adding land use, built environment, accessibility measures, deprivation measures, cycling infrastructure, and socio-demographic variables to the conventional bike-sharing demand prediction models can significantly improve their predictive performance. Further, the error histogram indicates that all the testing data points are located near the equity line suggesting most predictions align closely with the actual outcomes.

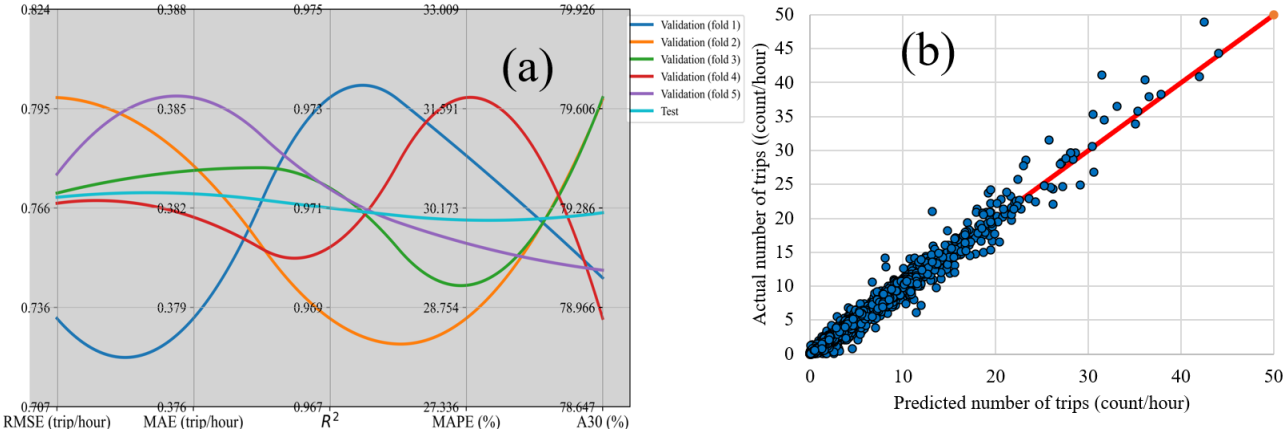

Figure 1 – *(a): the model's performance (b): error histogram of testing data*

Then, the tuned LightGBM model is synchronized with SHAP to examine the relative impact of independent variables on bike-sharing demand and the outcomes are presented in Figure 2.

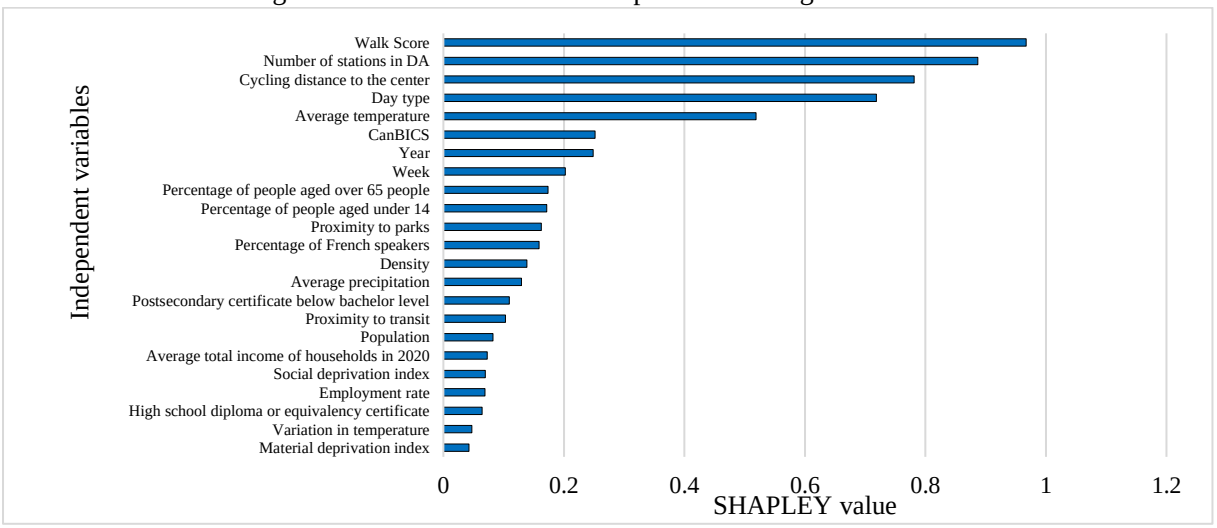

Figure 2 – *The relative influence (SHAPLEY value) of variables on bike-sharing demand*

As shown, the Walk Score has the strongest influence on bike-sharing demand, which has been overlooked in previous studies. Distance to the center can indirectly represent the built environment, and it is the third top variable. CanBICS and the percentage of people aged over 65 are the sixth and ninth top variables, representing the importance of infrastructure and socio-demographics in bike-sharing demand prediction models. Hence, adding land use, built environment, cycling infrastructure, and socio-demographic variables to conventional models is essential.

Subsequently, PDP is applied to capture the non-linear relationship between independent variables and bike-sharing demand, and the outcomes are presented in Figure 3. In this figure, only the PDP of some variables (top 6) are presented due to the page count limitation, but the full results and discussions will be presented at the conference. According to the results, the bike-sharing demand is maximum in DAs with a walk score of over 95, and hence, walking complements bike-sharing trips in Montreal. Increasing the number of stations up to five maximizes the number of trips, and after this threshold, the increase rate is not significant. The highest demand is in regions closer to the city center (with a cycling distance of less than 5.5 km). Bike-sharing demand is much lower in weekday peak hours than in weekday off-peak hours and weekends. In weeks with an average temperature of over 15°C, in neighborhoods with more nearby bikeways, and weeks 25 to 39 (June to September), Bixi use is likely to reach its highest level. In DAs with the lowest elderly (over 65) and children (under 15) population percentage, the demand is

expected to be higher. In DAs with minimum accessibility to parks, minimum percentage of French speakers, and maximum density, the likelihood of bike-sharing use is much higher.

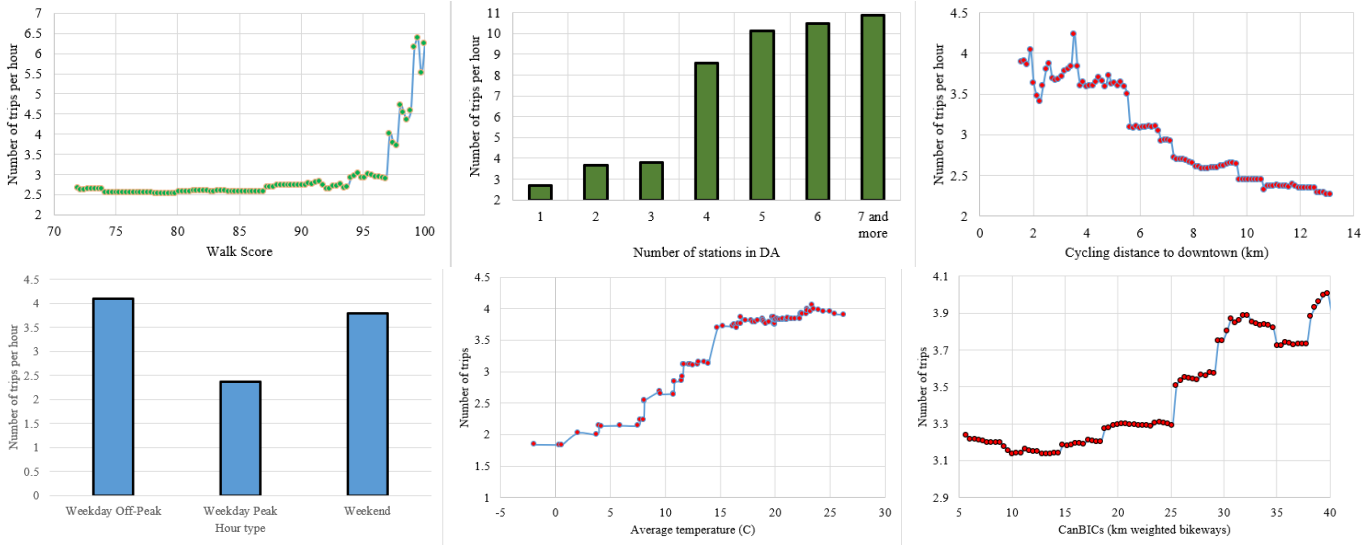

Figure 3 – *The direction influence of top variables on bike-sharing demand*

References

1. Naseri, H., E. O. D. Waygood, Z. Patterson, M. Alousi-Jones, and B. Wang. Travel Mode Choice Prediction: Developing New Techniques to Prioritize Variables and Interpret Black-Box Machine Learning Techniques. *Transportation Planning and Technology*, 2024. https://doi.org/10.1080/03081060.2024.2411611.
2. Fishman, E., S. Washington, and N. Haworth. Bike Share: A Synthesis of the Literature. *Transport Reviews*. 2. Volume 33, 148–165. https://www.tandfonline.com/doi/abs/10.1080/01441647.2013.775612. Accessed Oct. 28, 2024.
3. Eren, E., and V. E. Uz. A Review on Bike-Sharing: The Factors Affecting Bike-Sharing Demand. *Sustainable Cities and Society*. Volume 54. https://www.sciencedirect.com/science/article/pii/S2210670719312387?casa_token=FnAAc1Fs_UkAAAAA:h8hlaXtSc07kXhZoog4IAaXZXIO9zwhUE91wpWUNrT_iaRWSFyy6VVM1vte3zUAD5ORjJcB62886. Accessed Nov. 20, 2021.
4. Li, X., Y. Xu, X. Zhang, W. Shi, Y. Yue, and Q. Li. Improving Short-Term Bike Sharing Demand Forecast through an Irregular Convolutional Neural Network. *Transportation Research Part C: Emerging Technologies*, Vol. 147, 2023, p. 103984. https://doi.org/10.1016/j.trc.2022.103984.
5. E, S. V., J. Park, and Y. Cho. Using Data Mining Techniques for Bike Sharing Demand Prediction in Metropolitan City. *Computer Communications*, Vol. 153, 2020, pp. 353–366. https://doi.org/10.1016/j.comcom.2020.02.007.
6. Yang, Y., A. Heppenstall, A. Turner, and A. Comber. Using Graph Structural Information about Flows to Enhance Short-Term Demand Prediction in Bike-Sharing Systems. *Computers, Environment and Urban Systems*, Vol. 83, 2020, p. 101521. https://doi.org/10.1016/j.compenvurbsys.2020.101521.
7. Pan, Y., R. C. Zheng, J. Zhang, and X. Yao. Predicting Bike Sharing Demand Using Recurrent Neural Networks. No. 147, 2019, pp. 562–566.
8. Hulot, P., D. Aloise, and S. D. Jena. Towards Station-Level Demand Prediction for Effective Rebalancing in Bike-Sharing Systems. No. 18, 2018, pp. 378–386.
9. V E, S., and Y. Cho. A Rule-Based Model for Seoul Bike Sharing Demand Prediction Using Weather Data. *European Journal of Remote Sensing*, Vol. 53, No. sup1, 2020, pp. 166–183. https://doi.org/10.1080/22797254.2020.1725789.
10. Statistics Canada. Dissemination Block (DB), Dictionary, Census of Population, 2021. https://www12-2021.statcan.gc.ca/census-recensement/2021/ref/dict/az/definition-eng.cfm?ID=geo014.
11. Winters, M., J. Beairsto, C. Ferster, K. Laberee, K. Manaugh, and T. Nelson. The Canadian Bikeway Comfort and Safety Metrics (Can-BICS): National Measures of the Bicycling Environment for Use in Research and Policy. *Health Reports*, Vol. 33, No. 10, 2022, pp. 3–13. https://doi.org/10.25318/82-003-x202201000001-eng.
12. Microsoft. LightGBM - Microsoft Research. *Microsoft*. https://www.microsoft.com/en-us/research/project/lightgbm/. Accessed Sep. 16, 2024.
13. Naseri, H., E. O. D. Waygood, Z. Patterson, and B. Wang. Which Variables Influence Electric Vehicle Adoption? *Transportation*, 2024, pp. 1–38. https://doi.org/10.1007/s11116-024-10525-1.