Blue Bike Demand Forecasting

Yashwanth Reddy Vutukoori, Hariharasudan Savithri Anbarasu, Mithun Dineshkumar

Abstract

We develop a regression-based model to forecast **station-level** BlueBikes demand in Metro Boston using 2020 trip data (~2M records). We implement and compare **Linear Regression with polynomial expansion**, **Random Forest Regressor**, and a **Deep Neural Network (DNN)**. Features include **temporal signals** (cyclical encodings), **spatial context** (station coordinates/IDs), **historical patterns** (selected lags/rolling means), and **external factors** (hourly weather, holidays). The goal is **short-term, actionable forecasts** that support proactive rebalancing, improve availability, and enhance user experience. We evaluate with **time-aware cross-validation** using **RMSE/MAE**, and present a concise **comparative analysis** of models. We also outline a lightweight **interactive dashboard prototype** for visualizing predicted demand to aid operational planning.

Introduction

Bike-sharing helps address last-mile mobility while supporting sustainable transport. BlueBikes' growth has intensified a core operational challenge: **spatiotemporal imbalance**—some stations face shortages while others hold surpluses. We target **accurate short-term, station-level demand prediction** to enable **proactive redistribution** and better capacity planning.

Key challenges include: (1) **highly dynamic demand** driven by time-of-day, weather, and occasional events; (2) **spatial dependencies** where neighboring stations interact; (3) **sufficient lead time** for field operations; and (4) **balancing accuracy and computational efficiency** for near real-time use. Solving this problem can **boost availability and user satisfaction**, **lower redistribution costs**, and **inform planning** for station siting and capacity. Applications include **real-time fleet management**, **strategic expansion**, **multi-modal integration**, and **availability cues** inside rider apps.

Proposed Project

Problem Definition and Approach

We frame demand prediction as a **regression** task: forecast the **hourly rentals per station** for upcoming intervals. Compared with categorical (high/medium/low) classification, regression yields **numerical forecasts** that are more directly actionable for scheduling and rebalancing.

Dataset Description

We use BlueBikes 2020 trip data (~**2M records**) covering **Jan–Dec 2020** and **300+ stations** in Metro Boston. We aggregate trips to **station × hour** to produce supervised training examples and join **hourly weather** and **holiday** indicators.

Feature Engineering Strategy

We transform raw data into compact, informative features across four buckets:

- **Temporal** (from start/stop times): hour-of-day, day-of-week, month (all with cyclical encodings), weekend flag, and peak-hour indicators.
- **Spatial** (from station data): station ID (compact encoding), latitude/longitude; optional proximity/cluster features where helpful.
- **Historical demand**: selected lags (**t-1, t-24, t-168**), **7-day rolling mean**, and "same hour previous week."
- **External**: hourly **temperature**, **precipitation**, **wind**, **holiday** flag; a lightweight **special-event** indicator when available.
- **User/demographic aggregates (when present)**: subscriber/customer share and typical trip duration; coarse age-group distribution if reliably derivable.

Preprocessing Pipeline

Clean trips (remove extreme durations; fix missing coords), **aggregate to station×hour**, **join weather/holiday**, apply **appropriate imputation** for sparse gaps, scale continuous features as needed, and create **chronological splits** (e.g., 70/15/15 train/val/test) to **avoid leakage**.

Model Implementation

- **Linear Regression + Polynomial Features**: degree-2 basis with **ridge** regularization; serves as a transparent baseline and reveals linear/low-order interactions.

- **Random Forest Regressor**: non-linear ensemble tuned over core hyperparameters (e.g., depth, min_samples_split, n_estimators) to capture interactions with minimal feature engineering.

- **Deep Neural Network (DNN)**: compact feedforward network (3–5 layers, decreasing widths) with ReLU activations, batch norm, dropout, **Adam** optimization, and **early stopping**. All models use **validation-guided tuning** and identical training/validation splits for fair comparison.

Evaluation Methodology

We report **RMSE** (penalizes large errors) and **MAE** (typical error magnitude) under **time-series cross-validation** that preserves temporal order. We control for **leakage** by constructing lag/rolling features strictly from prior timestamps. **Ablations** remove feature buckets to gauge their contribution. We optionally contrast **peak vs. off-peak** performance to understand operational reliability during high demand.

References

1. Li, Y., Zheng, Y., Zhang, H., & Chen, L. (2015). Traffic prediction in a bike-sharing system. *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 1-10.
2. Chen, P. C., Hsieh, H. Y., Sigalingging, X. K., Chen, Y. R., & Leu, J. S. (2020). Predicting station level demand in a bike-sharing system using recurrent neural networks. *IET Intelligent Transport Systems*, 14(5), 554-561.
3. Collini, E., Nesi, P., & Pantaleo, G. (2021). Deep learning for short-term prediction of available bikes on bike-sharing stations. *IEEE Access*, 9, 124337-124347.
4. Sathishkumar, V. E., & Cho, Y. (2020). Season wise bike sharing demand analysis using random forest algorithm. *Computational Intelligence*, 36(4), 1456-1473.
5. Xu, C., Ji, J., & Liu, P. (2020). The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. *Transportation Research Part C*, 95, 47-60.
6. Wang, B., & Kim, I. (2018). Short-term prediction for bike-sharing service using LSTM and GRU. *International Journal of Engineering and Technology*, 10(3), 231-236.
7. Ai, Y., Li, Z., Gan, M., Zhang, Y., Yu, D., Chen, W., & Ju, Y. (2019). A deep learning approach on short-term spatiotemporal distribution forecasting of dockless bike-sharing system. *Neural Computing and Applications*, 31(5), 1665-1677.
8. Lin, L., He, Z., & Peeta, S. (2018). Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. *Transportation Research Part C*, 97, 258-276.
9. Yang, Z., Hu, J., Shu, Y., Cheng, P., Chen, J., & Moscibroda, T. (2016). Mobility modeling and prediction in bike-sharing systems. *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, 165-178.
10. Albuquerque, V., Sales, F., Casanova, M. A., & Soares, A. (2021). Machine learning approaches to bike-sharing systems: A systematic literature review. *ISPRS International Journal of Geo-Information*, 10(2), 62.