Probabilistic Bike-Sharing Demand Forecasting under Changing Weather and Seasonal Regimes with Transformer-Based Models

Abstract

Bike-sharing systems must plan fleet and rebalancing under changing weather and seasonal regimes. We conduct reproducible probabilistic one-hour-ahead demand forecasting on two public datasets (UCI Bike Sharing hourly data and Seoul Bike Sharing Demand) under forward-chaining temporal splits. We compare tree-based Random Forest and XGBoost quantile baselines with N-BEATS, a Transformer encoder, an Informer-lite encoder, and a Temporal Fusion Transformer (TFT). Across both cities, N-BEATS and tree models yield the lowest MAE, while TFT produces the best calibrated 80% prediction intervals (81.1% coverage in Seoul). Error and coverage degrade under extreme weather, motivating uncertainty-aware operations.

Keywords

bike-sharing demand; probabilistic forecasting; quantile regression; Temporal Fusion Transformer

Questions

Bike-sharing demand is often forecast with statistical or tree-based regression baselines that can achieve strong point accuracy but do not always provide well-calibrated uncertainty under regime changes. Transformer-based models and N-BEATS can fuse exogenous weather and calendar signals with seasonal regimes to learn nonlinear interactions that may improve robustness. We therefore ask: (RQ1) Under our forward-chaining temporal split that evaluates next-period generalization (UCI: 2011 → 2012; Seoul: Dec–May → Jul–Nov), how do TFT, Informer-lite, Transformer, and N-BEATS compare with RF/XGBoost for one-hour-ahead demand forecasting? (RQ2) Do 80% prediction intervals (P10–P90) remain calibrated during holidays/weekends and under extreme weather? (RQ3) Which calendar and weather covariates (including season/month and weather variables) are most influential according to TFT variable selection?

Methods

Datasets. We used the UCI Bike Sharing Dataset (hour.csv; Washington, DC) and the Seoul Bike Sharing Demand dataset (hourly). For UCI, the target is total rentals (cnt); we used the exogenous covariates season, holiday, workingday, weathersit, and the continuous weather variables temp, atemp, hum, and windspeed (excluding casual and registered). For Seoul, the target is Rented Bike Count; we retained only records with Functioning Day = “Yes” and used the available calendar covariates (Hour, Seasons, Holiday) and continuous weather variables (Temperature,

Humidity, Wind speed, Visibility, Dew point temperature, Solar Radiation, Rainfall, and Snowfall). Table 1 summarizes sizes and time spans.

Forecasting task and splits. We performed one-hour-ahead forecasting with a fixed lookback window of 24 hours. To stress temporal generalization under changing seasonal and weather regimes, we used forward-chaining splits that evaluate next-period performance (i.e., next-period forecasting rather than a controlled seasonal-shift experiment): UCI train < 2011-11-01, validation 2011-11-01 to 2011-12-31, and test year 2012; Seoul train < 2018-06-01 (winter/spring), validation June 2018, and test July– November 2018 (summer/fall). All reported metrics are computed on the held-out test periods only.

Features. We used calendar and weather covariates available in each dataset. Categorical time variables (hour-of-day, weekday, month, season, and weather category where provided) were one-hot encoded. Continuous weather variables were standardized using training-set statistics; binary indicators (holiday, working day) remained 0/1. The target was transformed with $\log(1+y)$ to stabilize variance (Hyndman & Athanasopoulos, 2021). For deep models, each training sample is a sequence of length 25: the past 24 observed demand values (standardized log-demand) with covariates, plus the target-time covariates with the demand value masked and an explicit observation flag.

Models. We compared: (i) Random Forest quantile regression using the empirical distribution over trees (Breiman, 2001); (ii) XGBoost quantile models using the quantile loss objective (Chen & Guestrin, 2016); (iii) N-BEATS (Oreshkin et al., 2020) with a three-quantile output head; (iv) a Transformer encoder baseline (Vaswani et al., 2017); (v) an Informer-lite encoder with temporal distilling for efficiency (Zhou et al., 2021); and (vi) a simplified Temporal Fusion Transformer (TFT) with variable selection, an LSTM encoder, and attention (Lim et al., 2021). All neural models were trained with the pinball loss for quantiles $q \in \{0.1,0.5,0.9\}$ (Koenker & Bassett, 1978). We did not implement a heterogeneous ensemble baseline (e.g., simple averaging across models) to keep the study focused on comparable single-model uncertainty; evaluating such ensembles remains future work.

Training protocol. Neural models used Adam (learning rate 1e-3), batch size 256, gradient clipping (1.0), and early stopping on validation pinball loss (patience 2, max 6 epochs for Transformer/Informer/TFT and max 8 for N-BEATS). RF used 200 trees (min_samples_leaf=2, max_features=0.7). XGBoost trained three separate models (one per quantile) with 200 estimators, max_depth=5, learning_rate=0.08, subsample=0.8, and colsample_bytree=0.8. We selected these hyperparameters by validation tuning: each model family was evaluated on the validation period and we report the configuration with the lowest validation pinball loss. All experiments used a fixed random seed (42).

Evaluation. Point accuracy was measured on the median (P50) using MAE, RMSE, and MAPE; we emphasize MAE because it is in bikes/hour and is directly interpretable for operations and capacity planning, while RMSE and MAPE are reported for completeness. Probabilistic quality used mean pinball loss and empirical 80% interval coverage and width (P10–P90). Because counts are nonnegative, negative predictions after inverse transform were clipped to 0. We also report stratified results for holidays and extreme weather to quantify shift robustness.

Table 1. Dataset summary and temporal splits.

Dataset	Time step	Period	Target	N (total)	Train/ Val/Test split	Covariates (count)
UCI Bike Sharing (Washington, DC)	Hourly	2011-01-01 to 2012-12-31	cnt (total rentals)	17379	Train: <2011-11-01; Val: 2011-11-01..2011-12-31; Test: 2012	57
Seoul Bike Sharing Demand	Hourly	2017-12-01 to 2018-11-30	RENTED_BIKE_COUNT	8465	Train: <2018-06-01; Val: 2018-06; Test: 2018-07..2018-11	57

Findings

Overall accuracy and calibration. Table 2 reports test performance, where lower MAE/RMSE/MAPE indicates more accurate point forecasts, and P80 coverage closer to the nominal 80% indicates better calibrated uncertainty. On UCI, the lowest MAE is obtained by N-BEATS (51.33 bikes/hour) and XGBoost (51.79), followed by RF (52.70); the same ordering holds for RMSE and MAPE. TFT has higher MAE (55.90) but the best-calibrated intervals among neural models (P10–P90 coverage 72.14% vs 44.84% for N-BEATS and 46.14% for the Transformer), while XGBoost is sharp but under-covered (59.14% coverage with width 74.84). On Seoul, RF and N-BEATS achieve the lowest MAE (115.86 and 115.07), whereas Transformer-based encoders underperform (MAE ≥ 211.01) under this winter/spring → summer/fall split. In Seoul, train+validation contain 5,040 functioning-day hours drawn only from Winter/Spring, while the test period contains 3,425 hours drawn only from Summer/Autumn (mean temperature shifts from 5.24°C in train to 20.09°C in test). By contrast, UCI trains on a full year (2011) and tests on the next year (2012). Answering RQ1, tree models and N-BEATS provide the most accurate point forecasts across the two cities; answering RQ2, TFT yields the closest-to-nominal 80% coverage (81.11%) but with wider intervals (width 903.08), while RF is conservative (89.26% coverage).

Shift and regime analysis. Table 3 focuses on holidays and extreme weather, where “Normal” denotes non-holiday hours with no precipitation (UCI: weathersit≤2; Seoul: Rainfall=0 and Snowfall=0) and “Extreme weather” denotes precipitation regimes (UCI: weathersit≥3; Seoul: Rainfall>0 or Snowfall>0). In UCI, TFT coverage drops from 72.90% in normal conditions to 58.90% under extreme weather, indicating under-coverage when weather regimes are rare. XGBoost remains under-covered across all UCI regimes (58.60% normal, 62.10% holiday, 64.70% extreme). In Seoul, TFT is near nominal in normal conditions (81.80%), becomes

conservative on holidays (95.80%), and under-covers during precipitation (68.00%). These results show that regime shifts primarily affect tail uncertainty even when MAE changes modestly (Gneiting & Raftery, 2007).

Case study and hourly structure. Figure 1 shows a holiday week in July 2012 where TFT's interval expands around demand peaks and tightens during low-demand periods. Figure 2 decomposes MAE by hour-of-day in UCI: for every model, errors are higher at commuting (≈08:00) and evening (≈17–18:00) peaks than during overnight hours, reflecting heteroskedastic demand. N-BEATS exhibits the lowest MAE at most hours, suggesting it captures strong diurnal seasonality more effectively than the Transformer encoders in this one-hour-ahead setting.

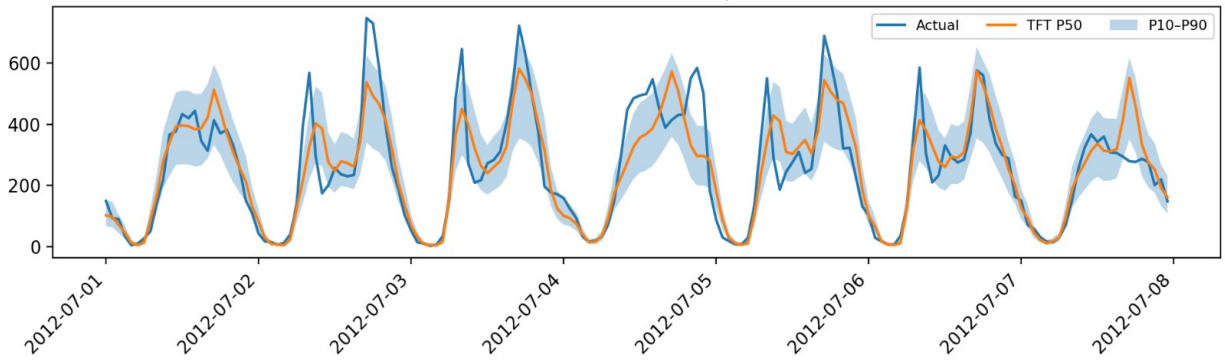

Figure 1. UCI-DC: TFT probabilistic forecast (P10–P90), 2012-07-01–2012-07-07.

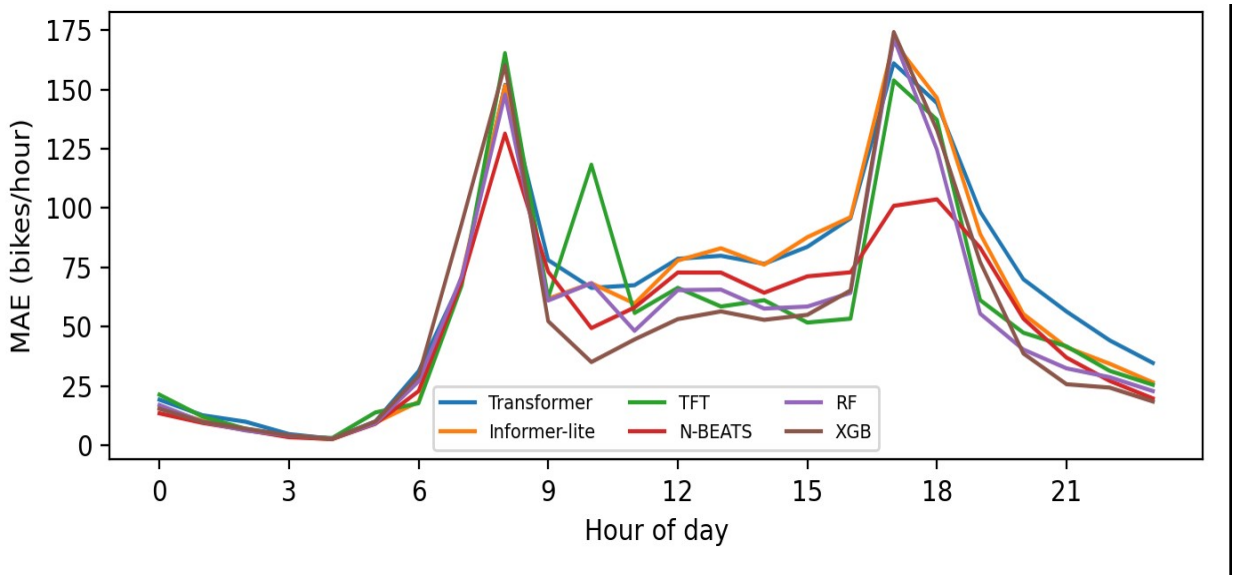

Figure 2. UCI: MAE by hour-of-day on the 2012 test set.

Interpretability. Figure 3 summarizes TFT variable selection weights. The most influential covariates include hour-of-day indicators (e.g., hr_17 denotes the 17:00 one-hot feature), seasonality indicators (month and season), and adverse weather categories (e.g., weathersit_3 denotes light rain/snow), along with continuous humidity (hum) and temperature variables (temp, atemp). For local explanation of the tree baselines, SHAP (SHapley Additive exPlanations) is directly applicable (Lundberg & Lee, 2017), while TFT's built-in selection provides a compact global ranking.

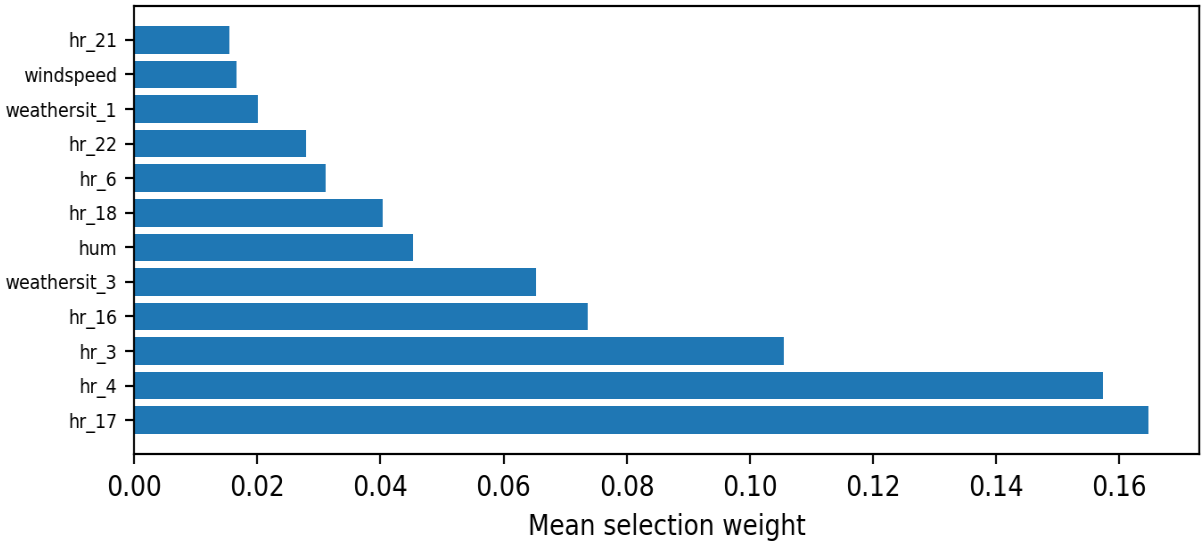

Figure 3. UCI: TFT variable selection weights (top 12 covariates).

Implications. Across two cities, strong point forecasts do not guarantee calibrated uncertainty: models with the lowest MAE often under-cover. For capacity planning and rebalancing, we recommend training with quantile losses and explicitly auditing coverage under holidays and extreme weather.

Table 2. Overall test performance (median P50 for MAE/RMSE/MAPE; P10–P90 for coverage/width).

Dataset	Model	MAE	RMSE	MAPE (%)	P80 Cov. (%)	P80 Width
UCI-DC	N-BEATS	51.33	75.88	26.72	44.84	57.47
UCI-DC	XGB	51.79	89.13	27.62	59.14	74.84
UCI-DC	RF	52.70	88.21	28.66	70.72	123.05
UCI-DC	TFT	55.90	90.87	32.24	72.14	142.95
UCI-DC	Informer-lite	61.20	94.99	30.72	57.74	98.03
UCI-DC	Transformer	64.69	95.14	31.99	46.14	89.74
Seoul	N-BEATS	115.07	166.97	19.67	75.42	339.33

Seoul	RF	115.86	185.20	20.85	89.26	555.78
Seoul	XGB	172.68	274.39	22.70	72.64	368.87
Seoul	Informer-lite	211.01	279.28	29.61	68.50	568.68
Seoul	TFT	306.98	454.15	42.07	81.11	903.08
Seoul	Transformer	317.72	436.86	44.14	74.01	904.88

Table 3. Robustness under holidays and extreme weather (MAE and empirical P10–P90 coverage).

Dataset	Model	Condition	MAE	P80_Coverage
UCI-DC	TFT	Normal (non-holiday & no precip)	56.71	72.90
UCI-DC	TFT	Holiday	40.86	82.00
UCI-DC	TFT	Extreme weather	51.92	58.90
UCI-DC	XGB	Normal (non-holiday & no precip)	52.82	58.60
UCI-DC	XGB	Holiday	37.29	62.10
UCI-DC	XGB	Extreme weather	44.75	64.70
Seoul	TFT	Normal (non-holiday & no precip)	324.45	81.80
Seoul	TFT	Holiday	246.79	95.80
Seoul	TFT	Extreme weather	145.81	68.00
Seoul	XGB	Normal (non-holiday & no precip)	179.66	73.20
Seoul	XGB	Holiday	140.34	72.50
Seoul	XGB	Extreme weather	117.06	66.50

Code and Data Availability

https://github.com/bifoli/bike-demand-experiments

Acknowledgments

During the preparation of this work, the author(s) used ChatGPT in order to improve readability and language style. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication. All results, code, and citations have been rigorously verified by the authors for accuracy and integrity.

References

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.

Fanaee-T, H., & Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence, 2(2–3), 113–127.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359–378.

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd ed.). OTexts.

Koenker, R., & Bassett, G. (1978). Regression quantiles. Econometrica, 46(1), 33–50.

Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774).

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2020). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In International Conference on Learning Representations.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998–6008).

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 35(12), 11106–11115.