Forecasting Cycling Volume Measurements:
A Cross-city Comparative Analysis Using Machine Learning and Deep Learning Techniques

Aditya Narayan Rai
Supervisor: Prof. Lynn Kaack
Hertie School, MSc Data Science for Public Policy

Abstract

This study addresses the growing importance of cycling in urban mobility and the complexity of predicting bicycle volumes due to various factors. A harmonized machine learning and deep learning framework was developed and applied to forecast daily bicycle volumes in Berlin and New York City. The models, including decision trees, ensemble methods, and neural networks, were evaluated using long-term automated bicycle counts integrated with diverse datasets under temporal holdout and leave-one-group-out cross-validation. The results demonstrated that machine learning and deep learning models significantly outperformed baseline time series approaches in both cities.

Introduction

Cycling is an essential component of sustainable urban mobility, providing substantial environmental and public health benefits. Despite its growing importance, accurately forecasting cycling volumes remains complex due to significant variability influenced by local infrastructure, weather conditions, land use, and traffic patterns. Traditional predictive methods typically fall short in addressing the intricate, non-linear, and spatially diverse nature of cycling behaviour. This thesis addresses these gaps by employing advanced ML and DL models in a structured comparative analysis across two major cities, Berlin and New York, to offer insights into effective forecasting strategies and factors influencing cycling volume predictions.

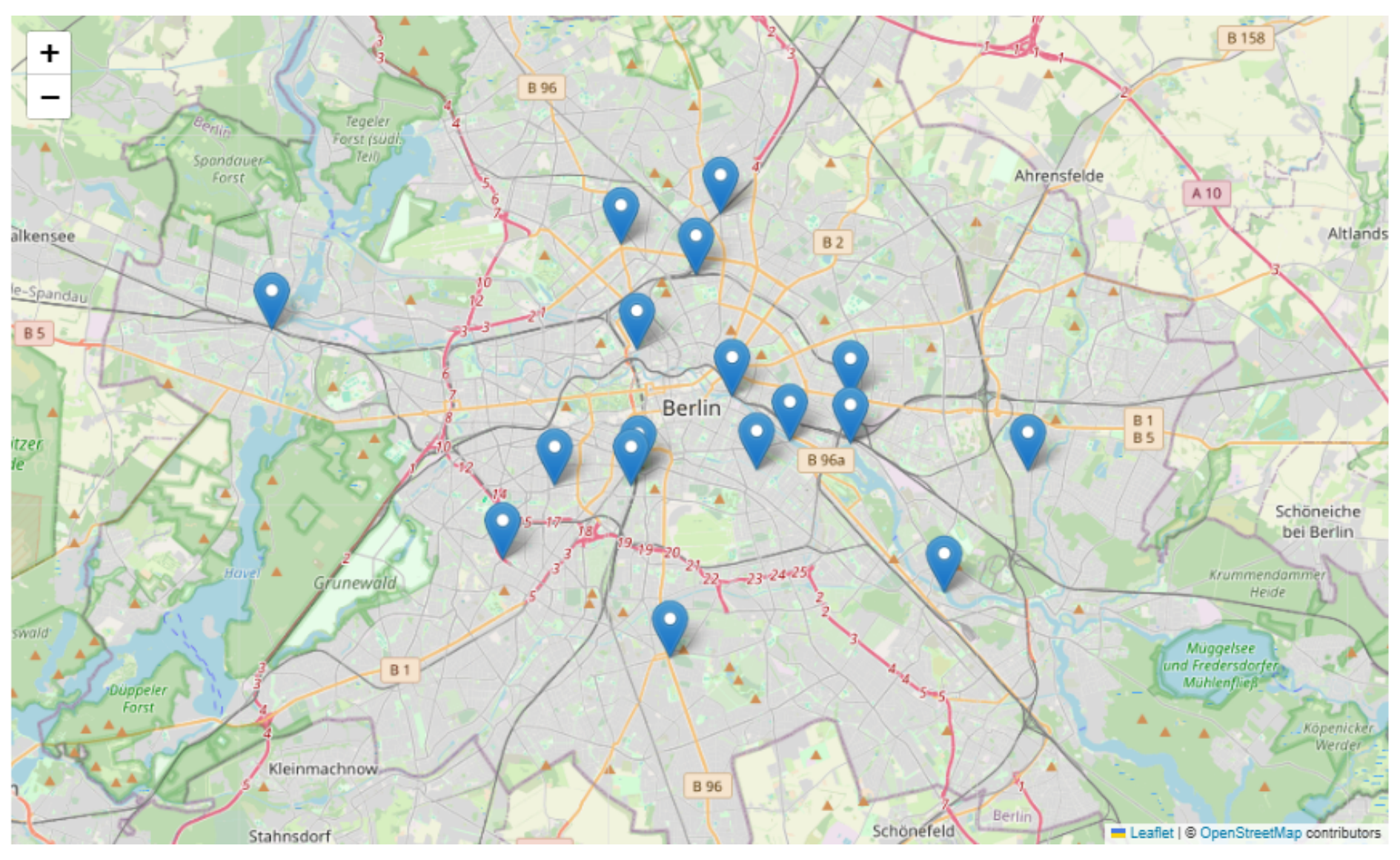

Figure: Berlin - Location of the Counting Stations

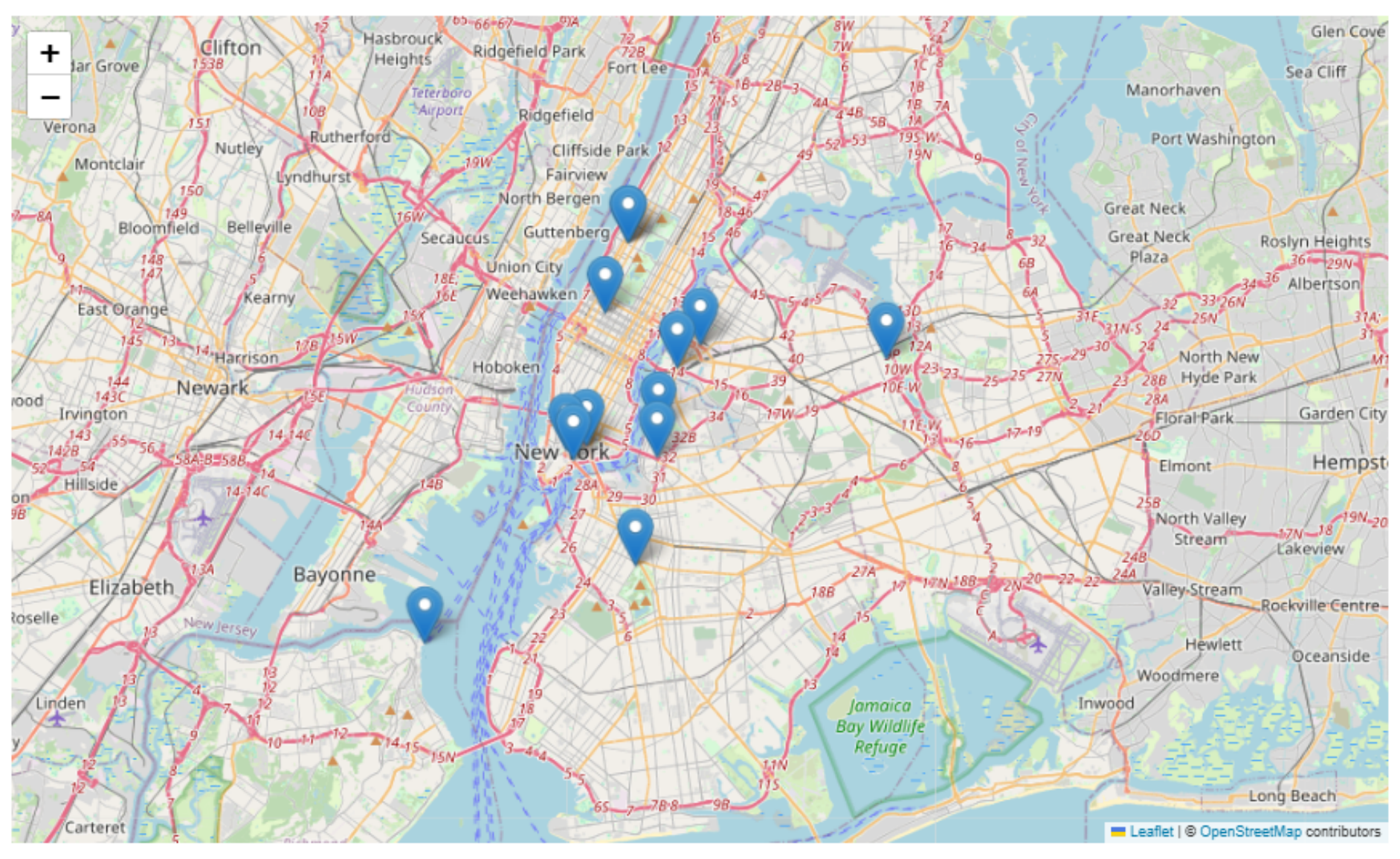

Figure: New York - Location of the Counting Stations

GitHub Repository

Geographic Scope and Datasets

- **Geographic Scope:** The study focuses on two distinct urban environments: Berlin, Germany, and New York City, USA.
- **Bicycle Counts:** Long-term hourly measurements of permanent bicycle counters.
- **Weather Data:** Daily weather information (e.g., temperature, precipitation, wind).
- **Land Use, Traffic, Infrastructure:** Various datasets representing urban characteristics, traffic patterns, and cycling infrastructure.
- **Temporal Data:** Information such as day of the week, seasonality, and holidays.

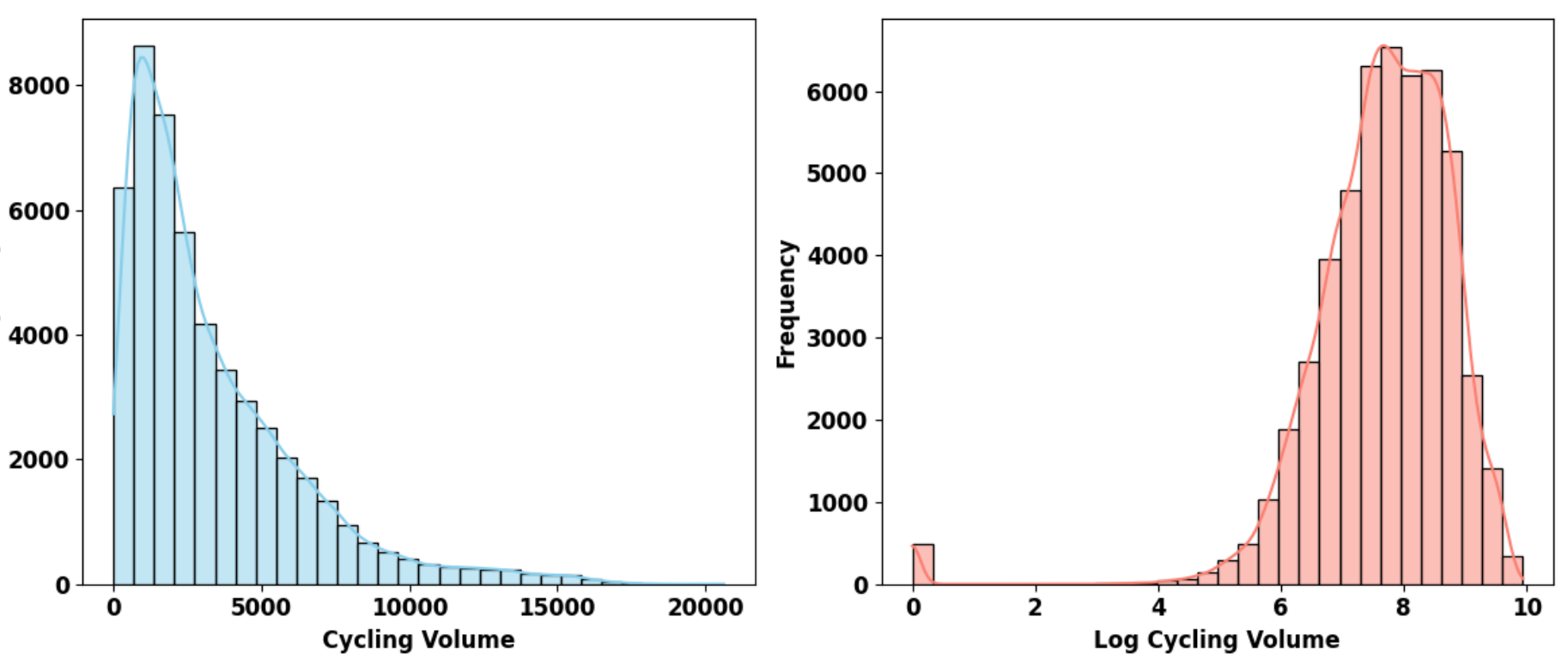

Figure: Berlin: Distribution of Cycling Volume

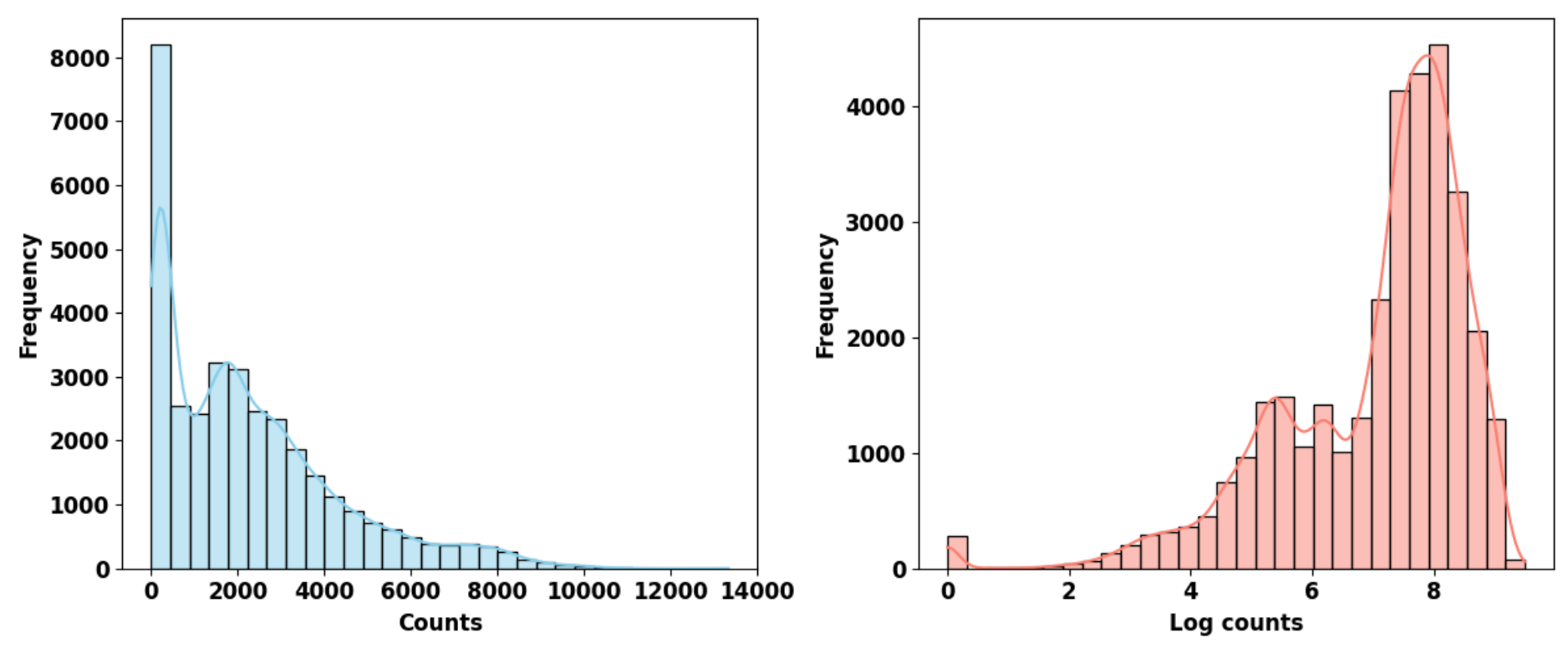

Figure: New York: Distribution of Cycling Volume

Methods

To evaluate model effectiveness, this study employed two distinct predictive tasks. **Temporal forecasting** utilized a station-specific temporal holdout strategy, where models predicted the most recent year based on historical data to assess predictive accuracy. **Spatial generalization** was evaluated using a Leave-One-Group-Out Cross-Validation (LOGO-CV) approach, testing each model's ability to generalize predictions to unseen locations within the same city.

The research compared three categories of predictive models to capture the complexity of cycling patterns: Baseline models (Naïve Mean, Exponential Smoothing, ARIMA), Machine Learning models (Linear Regression, Decision Trees, Random Forest, Gradient Boosting, XGBoost), and Deep Learning models (Shallow Neural Networks, Deep Neural Networks, Long Short-Term Memory Networks).

All models underwent systematic hyperparameter tuning via grid search and were supported by robust feature selection techniques, including Recursive Feature Elimination, Sequential Feature Selection, SelectKBest, and Principal Component Analysis. Performance was measured using standard metrics at the daily and the Annual Average Daily Bicycle (AADB) levels: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Symmetric Mean Absolute Percentage Error (SMAPE).

Results

Predictive performance varied considerably by city and evaluation strategy, indicating strong city-specific contexts influencing cycling volumes. For **temporal forecasting (station-specific)**, Decision Trees were the top-performing models in both cities. In Berlin, they achieved excellent predictive accuracy with a daily SMAPE of 26.42%. In New York, while still the best among tested models, their performance showed higher variability, resulting in a daily SMAPE of 47.01%.

For **spatial generalization (LOGO-CV)**, model performance declined across both cities, reflecting the difficulty of predicting cycling volumes at unseen locations. In Berlin, XGBoost showed the strongest spatial generalization with a SMAPE of 45.01%. In New York, all models struggled due to high spatial variability, with Gradient Boosting achieving the best—but still high—SMAPE of 85.55%.

Table: SMAPE at AADB Level for Temporal Forecasting and Spatial Generalization

Temporal Forecasting		
Model	**Berlin**	**New York**
Naïve Mean	25.47	46.88
Exponential Smoothing	77.86	63.79
ARIMA	81.04	65.49
Linear Regression	44.13	98.85
Decision Tree	8.10	32.72
Random Forest	18.05	42.24
Gradient Boosting	19.56	47.69
XGBoost	23.96	44.56
Shallow Neural Network	20.40	38.20
Deep Neural Network	14.67	40.66
LSTM	38.73	83.41
Spatial Generalization (LOGO-CV)		
Model	**Berlin**	**New York**
Naïve Mean	65.24	89.45
Linear Regression	51.67	97.33
Decision Tree	42.02	87.05
Random Forest	48.37	85.42
Gradient Boosting	43.21	84.37
XGBoost	38.69	66.27
Shallow Neural Network	106.15	126.35
Deep Neural Network	76.10	111.85
LSTM	131.54	138.45

Discussion & Conclusion

This study highlights the strong performance of advanced machine learning models, especially Decision Trees and XGBoost, in forecasting cycling volumes, outperforming traditional methods. However, spatial generalization remains a major challenge, particularly in heterogeneous cities like New York. The results emphasize the importance of urban context, with infrastructure and geography driving accuracy in Berlin, and land cover playing a larger role in New York.

Tailored, city-specific models consistently outperformed generalized ones, reinforcing the need for localized approaches. Incorporating short-term or crowdsourced data, such as Strava, could further improve spatial predictions. These findings support data-driven planning and targeted investment in cycling infrastructure.