A Comparative Predictive Modeling Study of Weather Impact on Bike Share Trip Volume

Christoffer Tan
University of Toronto
JSC370 Final Project

April 30, 2025

1 Introduction

Bikeshare programs have become a core part of urban transportation, particularly in cities like Toronto, where cycling provides a convenient and environmentally friendly commuting alternative. However, as with many outdoor activities, bike ridership is highly sensitive to weather conditions. Factors like rain, strong winds, and extreme temperatures significantly influence whether people decide to ride. This project aims to investigate how various weather conditions affect bikeshare usage in Toronto by building predictive models that forecast hourly trip volumes based on temperature, wind speed, humidity, cloudiness, and weather condition.

To conduct this analysis, I gathered anonymized trip data from Bike Share Toronto Ridership and matched it with hourly weather data from the OpenWeather API. The bikeshare data covers February 1 to September 30, 2024, but due to API limitations, the weather data is only available from April 18, 2024 to April 16, 2025. To ensure consistency, the study concentrates on the intersecting timeframe from April 18 to September 30, 2024, which aligns with the busiest cycling season and captures a wide range of weather variability affecting ridership.

This project explores several modeling techniques, including Linear Regression, Generalized Liner Model (GLM), Generalized Additive Models (GAM), Random Forest, and XGBoost, to compare their predictive accuracy. In addition to forecasting usage patterns, the analysis aims to identify the most influential weather factors and determine how their impacts differ throughout the day. The insights gained can help improve planning, operations, and future decision-making related to Toronto's bikeshare infrastructure.

2 Methods

2.1 Data Acquisition and Cleaning

This study combines two main datasets: hourly weather observations retrieved from the OpenWeather API and anonymized bike trip records from the Bike Share Toronto open data portal. A full description of the original features from each dataset can be found in the Appendix A. The weather dataset includes hourly measurements of temperature, humidity, wind speed, cloud coverage, precipitation, and a categorical label describing general weather conditions. These hourly readings are initially reported in Coordinated Universal Time (UTC), so they were converted to Toronto local time (Eastern Time, UTC−4) to align with the bike trip data.

The bikeshare dataset provides detailed records for each individual trip, including timestamps, station identifiers, and trip durations. To match the hourly resolution of the weather data, I aggregated the trip data by counting the number of trips initiated within each local hour. This resulted in an hourly trip count, effectively summarizing ridership demand. Finally, the hourly weather measurements and the aggregated bike trip counts were merged into a unified dataset, indexed by date and hour. The table below summarizes the features included in the final dataset used for analysis.

Table 1: Summary of Final Features in Unified Dataset

Feature	**Description**
`date`	Local date (Toronto time)
`hour`	Hour of day (0–23) in local time
`temp`	Temperature in Celsius
`humidity`	Relative humidity (%)
`wind_speed`	Wind speed in meters per second (m/s)
`cloudiness`	Cloud cover as a percentage
`weather_main`	Categorical weather condition label (e.g., Clear, Rain, Clouds)
`total_trips`	Number of bike trips started in that hour

2.2 Exploratory Data Analysis

Exploratory data analysis was conducted using ggplot2 and plotly to uncover patterns and relationships in the data. Histograms were used to examine the distributions of continuous variables such as bike trips, temperature, pressure, humidity, wind speed, and cloudiness. A pie chart highlighted the proportion of different weather conditions (`weather_main`). Relationships between weather and ridership were explored through correlation bar plots and scatterplots of each continuous weather variable against trip counts. Additionally, boxplots were used to visualize how trip volume varies by hour of the day and by weather condition. Interactive versions of these visualizations can be accessed on the project website's EDA page.

2.3 Modeling and Evaluation

All models were trained using 80% of the dataset, while the remaining 20% served as a test set to assess predictive performance. To consistently compare model performance, three widely-used metrics were used: R^2 (how much variance the model explains), RMSE (penalizes large errors), and MAE (average size of prediction errors).

The following models were applied:

- **Linear Regression (LM)**
  Linear regression provided a straightforward baseline, assuming a direct linear relationship between predictors and hourly bike trip counts. Before fitting the model, multicollinearity was assessed, and highly correlated variables were removed. Stepwise selection was then used to identify the most informative predictors. Diagnostic checks for assumptions like linearity, normality, homoscedasticity, and independence were performed. Where these assumptions were violated, continuous variables underwent Box-Cox transformations to stabilize variance and improve normality. The model was then refitted using the transformed variables, and final diagnostics were confirmed.

- **Generalized Linear Model (GLM) with Negative Binomial**
  Given that bike trips are count data, a GLM was initially considered using a Poisson distribution. However, preliminary checks revealed significant overdispersion (variance substantially exceeded the mean). To address this, the Negative Binomial distribution was used, as it includes an extra parameter to account for such variability. The model was further refined through stepwise selection to retain only predictors that significantly improved predictive performance and interpretability.

- **Generalized Additive Model (GAM) with Negative Binomial**
  Recognizing that the relationship between weather variables and bike usage might not be linear, a GAM was fitted with smooth terms for continuous predictors (e.g., temperature, wind speed, and precipitation). Each smooth term was examined visually through diagnostic plots and tested statistically for significance. Terms that appeared non-significant were compared and selectively removed using ANOVA tests, producing a simpler yet effective model. As with the GLM, a Negative Binomial distribution was employed to properly handle the count data's overdispersion.

- **Random Forest**
  A Random Forest model, which averages predictions from multiple decision trees to reduce variance and enhance accuracy, was initially built using default hyperparameters to establish a baseline. Subsequently, hyperparameters such as the number of features per split (`mtry`), the number of trees (`ntree`), and the minimum samples per terminal node (`nodesize`) were optimized through 5-fold cross-validation. Optimal hyperparameters were chosen based on the lowest average RMSE. Additionally, the model generated variable importance metrics, highlighting the most influential weather variables in predicting trip counts.

- **XGBoost (Extreme Gradient Boosting)**
  XGBoost, a powerful boosting algorithm that builds trees sequentially to correct residual errors from previous models, was explored similarly to Random Forest. Initially, default hyperparameters provided a baseline performance. Next, a thorough hyperparameter tuning was conducted using 5-fold cross-validation, optimizing parameters like learning rate (`eta`), maximum tree depth (`max_depth`), fraction of samples per tree (`subsample`), proportion of features considered at each split (`colsample_bytree`), and minimum weight needed in child nodes (`min_child_weight`). Optimal settings were selected based on the lowest RMSE. Variable importance scores from XGBoost further identified predictors with the strongest influence on bikeshare usage.

3 Results

3.1 Model Performance Comparison

Model performance metrics for all predictive models are summarized in Table 2. The Random Forest (CV-Tuned) demonstrated superior predictive accuracy on the test dataset (highest R^2 = 0.846, lowest RMSE = 330.028, lowest MAE = 236.966), closely followed by the XGBoost (CV-Tuned) model (R^2 = 0.845, RMSE = 331.139, MAE = 238.871). Detailed hyperparameter settings from cross-validation and the optimal values selected can be found in Appendix B. The best hyperparameters for Random Forest were `mtry` = 4, `ntree` = 500, `nodesize` = 10 and for XGBoost were `eta` = 0.05, `max_depth` = 6, `min_child_weight` = 3, `subsample` = 0.8, `colsample_bytree` = 1.

Table 2: Performance Comparison of Predictive Models

Model	**Train Set**			**Test Set**		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Linear Regression (LM)	0.826	360.692	251.108	0.830	347.020	245.548
GLM (Negative Binomial)	0.817	368.451	254.049	0.819	359.205	252.532
GAM (Negative Binomial)	0.829	355.891	246.153	0.833	343.998	246.577
Random Forest (Default)	0.965	174.114	122.431	0.845	336.344	250.219
Random Forest (CV-Tuned)	0.952	193.634	132.163	0.846	330.028	236.966
XGBoost (Default)	0.978	129.561	85.818	0.830	347.255	247.811
XGBoost (CV-Tuned)	0.905	268.267	184.694	0.845	331.139	238.871

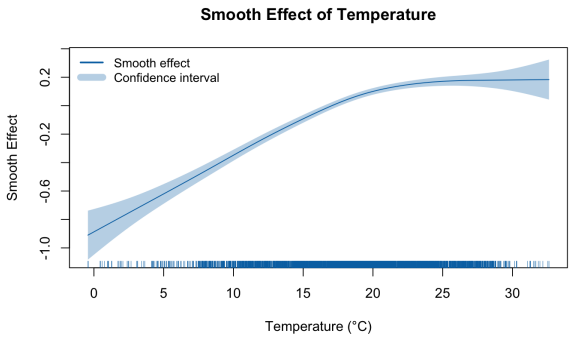
3.2 Overview of Traditional Statistical Models (LM and GLM)

To complement the predictive analysis performed using machine learning models, traditional statistical methods were also explored to provide interpretability and assess underlying assumptions. A Linear Regression (LM) model was initially fitted, achieving good predictive accuracy ($R^2 = 0.830$ on the test set). However, when a Poisson Generalized Linear Model (GLM) was tested, significant overdispersion was detected (dispersion parameter = 101.62), suggesting a violation of the Poisson distribution assumption.

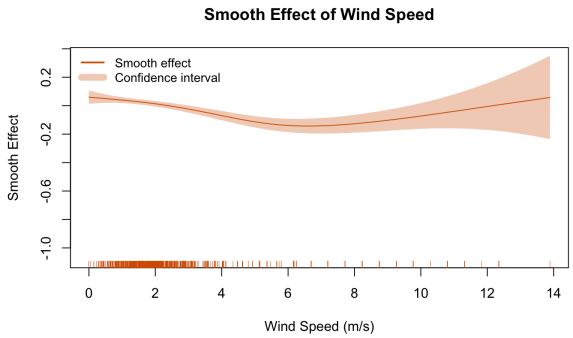
To address this issue, a Negative Binomial GLM was employed, which corrected for overdispersion and provided improved performance. Trip volumes were significantly higher during morning and evening peak hours (e.g., 8 AM and 5 PM), while adverse weather conditions such as rain and thunderstorms showed negative associations with trip counts. Summaries of the model fits and key statistical outputs are provided in Appendix C.

3.3 Interpretation of Nonlinear Weather Effects (GAM)

Based on the Negative Binomial GAM model, Figure 1 shows the nonlinear relationships between weather conditions (temperature and wind speed) and trip volumes. The model highlights significant nonlinear trends, particularly showing that trip volumes rapidly increase with temperature until around 20°C, after which they plateau, suggesting diminishing effects at higher temperatures. Wind speed exhibits a milder nonlinear trend, with decreasing trips at higher wind speeds. Summaries of the model fit and key statistical outputs are also provided in Appendix C.



(a) Temperature smooth effect

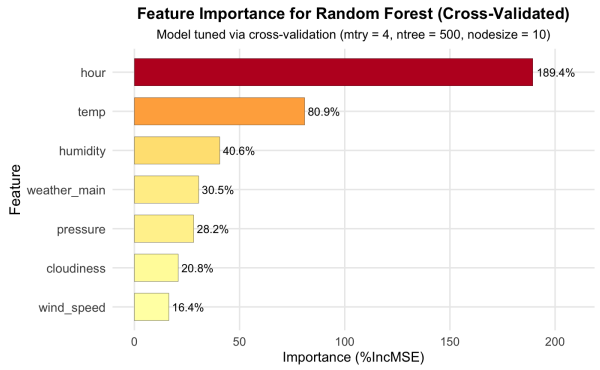


(b) Wind speed smooth effect

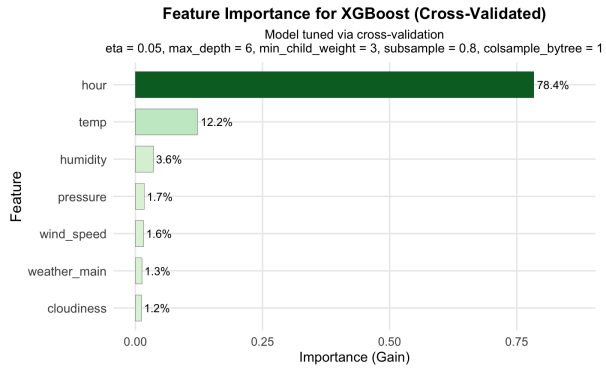
Figure 1: Estimated smooth effects from the GAM (Negative Binomial) model for temperature and wind speed. Shaded regions represent 95% confidence intervals, capturing uncertainty in the estimated smooth terms.

3.4 Variable Importance Analysis (Random Forest and XGBoost)

The variable importance plots for the tuned Random Forest (CV-Tuned) and tuned XGBoost (CV-Tuned) models, shown in Figure 2, illustrate the most influential predictors for bike trip volumes. The Random Forest model (with hyperparameters described previously) identifies `hour` and `temp` as most important, but with substantial contributions from other weather factors such as humidity and weather conditions. The XGBoost model emphasizes the strong importance of the hour of the day, with a more concentrated dependence compared to the Random Forest model, followed by temperature.



(a) Random Forest importance (%inc. MSE)



(b) XGBoost importance (gain)

Figure 2: Comparison of variable importance between Random Forest (CV-Tuned) and XGBoost (CV-Tuned) models.

3.5 Predictive Accuracy of the Best Model

Figure 3 illustrates the predictive performance of the Random Forest (CV-Tuned) model, comparing predicted and actual bike trips on the test dataset. Predictions largely cluster around the diagonal line, indicating reliable predictive accuracy for typical conditions. However, predictions become less

accurate during peak usage hours, suggesting areas for future model refinement to handle extreme usage scenarios.

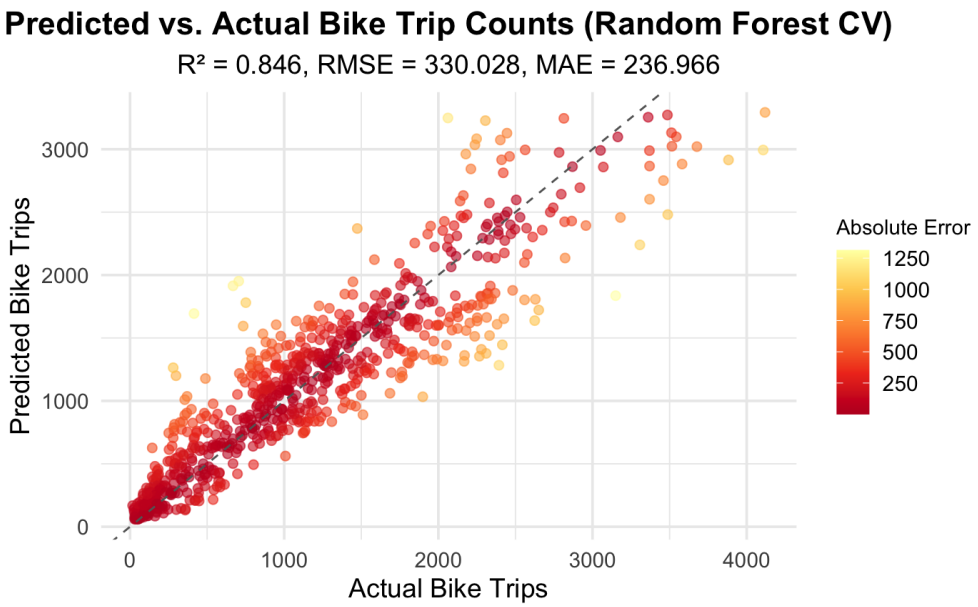


Figure 3: Predicted vs. actual bike trip counts on the test set (Random Forest CV-Tuned).

4 Conclusion and Summary

This study investigated the relationship between weather conditions and bikeshare usage in Toronto by developing and comparing several predictive models. Through extensive analysis, the tuned Random Forest model emerged as the best-performing model, achieving the highest predictive accuracy on the test dataset. This indicates that ensemble tree-based models like Random Forests, which can capture complex nonlinear interactions, are particularly well-suited for forecasting urban bike trip volumes based on meteorological data.

The analysis also highlighted the critical weather factors influencing bike ridership. Hour of the day and temperature were consistently identified as the most influential predictors, with nonlinear effects clearly captured by the GAM model. For instance, bikeshare demand increased rapidly with temperature up to approximately 20°C, after which the relationship plateaued. Wind speed also showed a measurable but milder negative impact on ridership. These insights can help city planners and bikeshare operators optimize operations, infrastructure planning, and targeted marketing strategies to encourage ridership, especially during favorable weather conditions.

Despite promising results, this study has several limitations. Firstly, the analysis period from late April to September 2024 predominantly captures the summer and early fall months, characterized by relatively stable and warm weather. Consequently, these findings might not generalize well to other seasons, especially winter months, when harsher conditions might alter biking behavior significantly. Additionally, the current models do not account explicitly for special events, holidays, or temporary infrastructure changes, factors known to influence bike usage independently of weather. Future research could expand the temporal scope of the data collection to encompass year-round conditions and integrate additional contextual variables to build more robust and generalizable predictive models.

Appendix

A Feature Descriptions

The table below provides detailed descriptions of features from both the weather and bikeshare datasets used in the project.

Table 3: Detailed Feature Descriptions from Source Datasets

Weather Data Features

Feature	Description
`date`	UTC date
`hour`	UTC hour
`temp`	Temperature in Celsius
`humidity`	Humidity (%)
`wind_speed`	Wind speed (m/s)
`cloudiness`	Cloud cover (%)
`weather_main`	General weather label

Bike Share Data Features

Feature	Description
`trip_id`	Unique trip identifier
`trip_start_time`	Start time of trip
`trip_stop_time`	End time of trip
`trip_duration_seconds`	Duration of trip (s)
`from_station_name`	Origin station
`to_station_name`	Destination station
`user_type`	Type of user (member/casual)

B Hyperparameter Grid Search

The table below summarizes the hyperparameter grid used during cross-validation for both the Random Forest and XGBoost models.

Table 4: Hyperparameter grid considered for Random Forest and XGBoost.

Model	Parameter	Description	Values Considered
Random Forest	`mtry`	Number of variables randomly sampled as candidates at each split	{2, 4, 6, 8}
	`ntree`	Number of trees to grow in the forest	{100, 250, 500}
	`nodesize`	Minimum number of observations in each terminal node	{1, 5, 10}
XGBoost	`eta`	Learning rate, controlling the contribution of each tree	{0.05, 0.1}
	`max_depth`	Maximum depth of individual trees	{4, 6}
	`min_child_weight`	Minimum sum of instance weights (hessian) in a child node	{1, 3}
	`subsample`	Fraction of training instances randomly sampled for each tree	{0.8, 1}
	`colsample_bytree`	Fraction of features randomly sampled for each tree	{0.8, 1}

C Traditional Statistical Model Summaries

Linear Regression (LM)

```
Call:
lm(formula = total_trips_bc ~ temp + pressure + humidity_bc +
wind_speed_bc + cloudiness_bc + hour + weather_main, data = lg_train)

Multiple R-squared:  0.8626, Adjusted R-squared:  0.8611
Residual standard error: 5.184 on 3152 degrees of freedom
F-statistic: 565.4 on 35 and 3152 DF, p-value: < 2.2e-16
```

Negative Binomial GLM

```
Call:
MASS::glm.nb(formula = total_trips ~ temp + humidity + wind_speed +
hour + weather_main, data = glm_train)

AIC: 44898
Residual deviance: 3297.2 on 3154 degrees of freedom
Theta: 6.564 (SE: 0.165)
```

Negative Binomial GAM

```
Formula:
total_trips ~ s(temp) + s(wind_speed) + hour + weather_main

R-sq.(adj) = 0.827; Deviance explained = 83.9% -REML = 22454

Approximate significance of smooth terms:
s(temp): edf = 4.335, p < 2e-16
s(wind_speed): edf = 3.522, p < 2e-16
```