🚲 Predicting Bike Rental Demand Using Generalized Linear Models

By Supreet Mutsuddi | Predictive Analyst | Actuarial Science & Machine Learning

Business Problem Statement

Urban mobility services, particularly bike-sharing systems, operate in dynamic environments where demand fluctuates significantly by hour, weather, and season. Poor demand forecasting leads to:

* **Empty stations** during high demand, causing customer frustration and lost revenue
* **Overstocked stations** during low demand, wasting operational effort and reducing efficiency

The key business question is:

Can we accurately predict hourly bike rental demand using weather and time-based features, to support smarter inventory allocation and operational decisions?

Business Value:

* **Inventory Management** - Optimize bike distribution across stations
* **Staffing and Scheduling** - Align workforce and rebalancing operations with peak hours
* **Strategic Planning** - Inform future infrastructure investment decisions based on demand patterns
* **Customer Experience** - Reduce service failures by anticipating demand shifts

This analysis aims to develop a statistically robust and interpretable forecasting model to support real-time and long-term decision-making for a bike-sharing platform.

Project Overview & Data Understanding

To forecast hourly demand, I used a publicly available dataset containing 17,379 hourly observations of bike rentals. Each record included:

* **Environmental features**: temperature, humidity, windspeed
* **Temporal features**: hour, weekday, month, season
* **Weather conditions**: coded categorical levels

The target variable was cnt, representing the total number of bikes rented per hour.

To prepare the data:

* **Dropped leakage features**: casual and registered (components of cnt)
* **Removed identifiers**: instant, dteday (non-predictive)
* **Converted** time-based variables (hr, weekday, season) to factors
* **Normalized** continuous variables (temp, hum, windspeed) for stability in GLM fitting

The goal was to create a clean, interpretable dataset consistent with GLM assumptions, ready for diagnostics and modeling.

Exploratory Data Analysis

Before modeling, I conducted exploratory analysis to understand the distribution of the target variable and its relationship with key predictors — both numeric and categorical.

Target Variable: cnt

The distribution of cnt was **right-skewed**, with frequent high-demand spikes during peak commuting hours and favorable weather.

📸 Image 1: Hourly Bike Rental Counts (cnt): Right-Skewed with Demand Spikes

Most rentals cluster below 200, with several hours exceeding 400 rentals

A graph of rentals

AI-generated content may be incorrect.

Numeric Predictors vs Target

I examined continuous variables for their influence on demand:

* **Temperature** had a strong positive association, with demand increasing until a saturation point
* **Humidity** and **windspeed** showed mild negative relationships, suggesting environmental discomfort suppresses usage

📸 Image 2: Bike Rentals Rise with Temperature, Then Plateau

A graph showing the temperature

AI-generated content may be incorrect.

📸 Image 3: High Humidity Slightly Reduces Rental Volume

A graph showing the normalized humidity

AI-generated content may be incorrect.

📸 Image 4: Higher Windspeeds Associate with Fewer Rentals

A graph showing a normalized windspeed

AI-generated content may be incorrect.

These results supported the inclusion of all three numeric predictors, with temperature expected to be the most influential.

Categorical Predictors vs Target

I explored categorical features to capture cyclical and seasonal patterns:

* **Hour of day** revealed distinct usage peaks:
  + Weekdays: sharp peaks at **8 AM** and **6 PM**
  + Weekends: broader afternoon peak
* **Seasonality** influenced demand:
  + **Fall and Summer** saw the highest usage
  + **Spring** consistently had the lowest
* **Weather conditions** impacted demand significantly:
  + **Clear** weather boosted rentals
  + **Rainy or misty** weather suppressed usage

📸 Image 5: Commuter vs Leisure Behavior: Hourly Rental Trends

A graph showing a number of hours

AI-generated content may be incorrect.

📸 Image 6: Seasonal Demand Variation in Bike Rentals

A graph showing different seasons

AI-generated content may be incorrect.

📸 Image 7: Demand Dips with Poor Weather, Especially on Humid Days

A graph showing the weather

AI-generated content may be incorrect.

These patterns suggested that both **hour** and **weather** could interact meaningfully with other variables, such as weekday, and should be considered for potential interaction terms in the model. This formed the foundation for enhancing model complexity in the next stage.

**Multicollinearity Checks & Feature Selection**

With the cleaned dataset in place, I conducted a multicollinearity analysis to ensure the reliability and stability of the regression estimates. This step is critical in any Generalized Linear Model (GLM), where collinear predictors can inflate standard errors and distort inference.

**Identifying Correlated Predictors**

Pairwise correlation analysis revealed a near-perfect linear relationship between temp and atemp (correlation coefficient ≈ 0.99). While both variables measure temperature, atemp represents “feels-like” temperature. Including both would violate the assumption of predictor independence.

📸 Image 8: GGPairs matrix showing temp and atemp correlation

A screenshot of a graph

AI-generated content may be incorrect.

To resolve this:

* I dropped atemp and retained temp, which was more directly interpretable in operational terms.

Redundant Feature: workingday

Using logical checks and the alias() function in R, I discovered that workingday was a deterministic function of weekday and holiday. In other words, it provided no additional information beyond those two variables.

* Including it would introduce perfect multicollinearity, leading to instability in the design matrix.
* I therefore excluded workingday from the model.

📌 These actions helped ensure that the GLM would operate on a well-conditioned, non-redundant feature set, supporting clearer interpretation and stronger diagnostics.

📸 Image 9: A logic diagram of feature dependencies

A diagram of a workday

AI-generated content may be incorrect.

**Model Development & Comparison**

With a clean and well-specified dataset, I began model development using **Generalized Linear Models (GLMs)** — starting with a **Poisson regression**, followed by a **Negative Binomial GLM** to address overdispersion.

**Poisson GLM**

Given that our target variable cnt is a **count of rentals**, the Poisson distribution is a natural starting point.

The Poisson model assumes:

* **Variance = Mean**
* A log link: **log(E[cnt]) = β₀ + β₁X₁ + ⋯ + βₖXₖ**

The model was fitted as:

poisson\_glm <- glm(cnt ~ ., data = df\_train, family = poisson(link = "log"))

However, diagnostic checks revealed a high **dispersion ratio** (> 30), where the residual deviance far exceeded the residual degrees of freedom. This violates a core Poisson assumption and suggests **overdispersion** — a case where variance exceeds the mean due to unobserved heterogeneity or structural noise.

📌 Insight: Overdispersion inflates Type I error and leads to underestimated standard errors. A more flexible distribution was needed.

**Negative Binomial GLM**

To address overdispersion, I used the **Negative Binomial GLM**, which introduces a dispersion parameter (θ), allowing the variance to exceed the mean:

Var(cnt) = μ + μ² / θ

This model was fitted using:

nb\_glm <- MASS::glm.nb(cnt ~ ., data = df\_train)

**Result**:

* Dispersion parameter (θ ≈ 3.94) confirmed **overdispersion**
* Residual deviance and AIC improved vs. Poisson
* Test **RMSE**: increased slightly (Poisson: **91.15** vs NB: **104.20**)

Although the Poisson GLM had lower test RMSE (91.15 vs. 104.20), it violated a core modeling assumption — equidispersion. The Negative Binomial model better captured variance structure and yielded more statistically valid coefficient estimates.

📌 *Interpretation*: The Poisson model showed better predictive accuracy on unseen data, but the Negative Binomial model **provided more reliable parameter estimates and inference**. This reflects a common tradeoff in statistical modeling between **predictive performance and model validity**.

📸 Image 10: Predicted vs Actual plot for Poisson and NB GLM side by side

A graph with a dotted line

AI-generated content may be incorrect.

A graph showing a line of rentals

AI-generated content may be incorrect.

**Key Predictors Identified**

The NB model yielded several important effects (based on IRR):

* Temperature: IRR ≈ 3.8 → warmer temperatures significantly boost rentals
* Humidity: IRR ≈ 0.79 → high humidity slightly reduces demand
* Hour of day and weather conditions showed strong temporal and environmental patterns

**Model Diagnostics**

* Cook’s Distance: No observations exceeded the standard threshold (>1), indicating no single data point was unduly influencing the model
* Deviance residuals: Randomly scattered, suggesting appropriate functional form

📸 Image 11: Cook’s Distance plot

A graph showing the number of people

AI-generated content may be incorrect.

**Interaction Effects & Model Enhancement**

While the base Negative Binomial model provided reasonable accuracy and interpretability, exploratory analysis suggested that **interactions between time-based variables** could further improve the model’s predictive power.

**Observed Pattern**

From earlier visualizations, it was clear that **hourly rental patterns differ significantly between working and non-working days**:

* Weekdays showed sharp peaks around **8 AM and 6 PM**, reflecting commuter behavior.
* Weekends exhibited a **broader midday peak**, consistent with recreational usage.

📸 **Suggested image: Line plot of average rentals by hour, colored by working day status**

This behavioral insight led to the inclusion of an **interaction term between hr and weekday** in the Negative Binomial GLM:

nb\_glm\_interact <- MASS::glm.nb(cnt ~ . + hr:weekday, data = df\_train)

**Impact on Predictive Accuracy**

Adding the interaction term substantially improved model performance:

* **Test** **RMSE** dropped from **104.2** to **62.6**, a ~**40%** improvement
* Captured dual-peak vs single-peak behavior across different day types
* Retained interpretability by modeling interaction effects explicitly rather than through

black-box methods

📌 This highlights the power of incorporating domain knowledge — in this case, commuter vs leisure usage — into statistical models through thoughtfully constructed interaction terms.

📸 Image 12: Predicted vs Actual Rentals (with interaction model)

A graph with a dotted line

AI-generated content may be incorrect.

**Model Diagnostics (Post-Interaction)**

* **Cook’s Distance** values were rechecked. No points exceeded the influence threshold.
* The inclusion of the interaction did not introduce multicollinearity or convergence issues.

📸 Image 13: Cook’s Distance plot after interaction model

A graph showing the number of numbers

AI-generated content may be incorrect.

Including hr \* weekday was a business-aware enhancement that led to tangible predictive improvements while preserving model explainability — an ideal outcome in any operational modeling scenario.

**Recap & Conclusion**

This project began with a straightforward business challenge:

**How can we accurately predict hourly bike rental demand to optimize urban mobility operations?**

Through a structured modeling approach, I developed an interpretable and statistically robust solution using Generalized Linear Models (GLMs), supported by exploratory analysis and domain-driven enhancements.

**🔁 Key Steps Recapped**

* **Data Understanding**: Cleaned, transformed, and normalized 17,000+ hourly observations.
* **EDA Insights**: Identified strong predictors (temperature, hour, weather) and distinct usage patterns across weekdays and weekends.
* **Multicollinearity Checks**: Removed redundant variables like atemp and workingday to stabilize estimates.
* **Modeling**: Compared Poisson and Negative Binomial GLMs. Despite lower RMSE in Poisson, overdispersion justified the use of NB for better inference.
* **Interaction Terms**: Incorporated hr \* weekday, improving RMSE from 104.2 to 62.6 — capturing real behavioral variation.
* **Diagnostics**: Checked for influential points and confirmed model assumptions post-adjustment.

**🎯 Business Takeaways**

* **Smarter inventory allocation** is possible with interpretable, data-driven forecasting.
* **Behavioral patterns** (commute vs leisure) matter — and can be modeled effectively.
* **GLMs still offer strong performance** when supported by diagnostics, interaction modeling, and business intuition.

**Final Thoughts**

In an era of black-box algorithms, this project demonstrates the value of interpretable, diagnostics-driven modeling — particularly where business context matters.

Whether you're managing a mobility platform, building demand forecasts, or designing smart infrastructure, blending statistical rigor with domain insight is essential.

If you're working on demand modeling, transportation analytics, or operational forecasting, I’d love to connect and collaborate.