MPhil in Economics and Data Science

Module: D100 Fundamentals of Data Science / D400 Research Computing

Candidate Number (BGN): 3377D

Deadline Date: Thursday 19th December, 12pm

I confirm that this is entirely my own work and has not previously been submitted for assessment, and I have read and understood the University's and Faculty's definition of Plagiarism (please see links below)

Plagiarism and Academic Misconduct

Plagiarism and correct referencing in dissertations Moodle Econ Update (003).pdf

Actual word count: 1970

Introduction

In the field of operational management, the efficient prediction and allocation of resources play a pivotal role in enhancing service delivery and customer satisfaction. This study draws inspirations from the work of Sathishkumar V E et al. (2020), and addresses the problem of forecasting hourly demand for Seoul’s bike-sharing system.

The primary objective of this study is to develop a predictive model that can capture the relationships between weather, time, and hourly bike-sharing demand. To achieve these objectives, the study follows a systematic methodology that includes:

- Data Analysis: Perform Exploratory Data Analysis (EDA) to understand the distributions and relationships of key features, guiding modelling and preprocessing.
- Feature Engineering: Transform time and weather features to improve model performance.
- Model Development: Build and refine models to optimise performance.
- Evaluation: Evaluate model effectiveness and compare different models.

This study contributes to the growing field of predictive analytics in mobility systems, showcasing how machine-learning models can support operational decisions. By building GLMs and LGBMs, this project demonstrates the trade-off between interpretability and predictive power, providing practical insights for decision-makers.

Data Overview and Data Cleaning

The Seoul Bike Sharing Demand dataset provides an hourly count of bicycles rented from the Seoul Bike Sharing System, along with corresponding weather data and time information, spanning from December 1, 2017, to November 30, 2018. Variables in the dataset include date and time variables, numeric variables: Rented Bike Count, Hour, Temperature, Humidity, Wind Speed, Visibility, Dew Point Temperature, Solar Radiation, Rainfall, Snowfall, and categorical variables: Seasons, Holiday, Functional Day. To facilitate analysis, Month, Day of Week and Week Status have been extracted from the "Date" column. This data was sourced from the UCI Machine Learning Repository and is utilised to analyse factors influencing bike rental demand in urban settings.

After an initial review, the dataset appears complete with no significant missing data or anomalies detected. However, a discrepancy was found for the date 2018-10-06, which is recorded with two different "Functioning Day" statuses, suggesting a potential error. Given that this inconsistency pertains to just one day, the most straightforward resolution is to remove this date from the dataset. Additionally, it was observed that days marked as "Functioning Day: No" consistently show zero demand for bike rentals. To enhance the accuracy of our predictive analysis, these entries will be excluded from the dataset. Following these adjustments, the dataset comprises data for 352 days, totalling 8,448 entries.

Explanatory Data Analysis

The count of rental bike users by hour is analysed across seasons, weekdays, and months to identify time trends. Figure 1(A) shows that the average count varies significantly by month, with a peak in June. Figure 1(B) highlights seasonal variations, where bike rentals are highest during summer, especially at peak commuting hours (8 AM and 6 PM), and lowest in winter. Figure 1(C) reveals weekly trends, with distinct peaks on weekdays during commuting hours and flatter patterns on weekends. Figure 1(D) further details hourly rental variations across months, showing higher demand in summer and lower activity in winter. The figures illustrate clear time-based patterns in bike rental usage. Figures 1(E) and 1(F) compare weekdays and weekends, and holidays and non-holidays separately, showing lower overall demand on weekends and holidays with more gradual trends, while weekdays non-holidays display significant spikes during commuting hours. These figures collectively illustrate strong time-based and contextual patterns in bike rental usage.

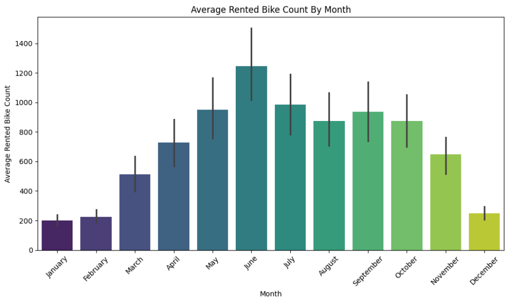


Figure 1(A)

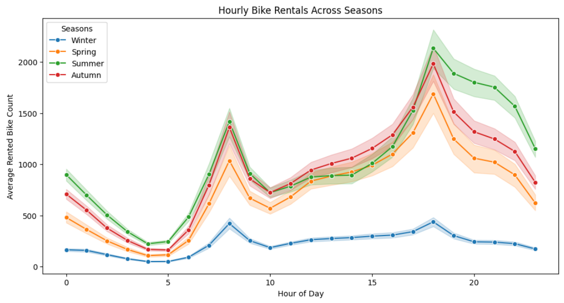


Figure 1(B)

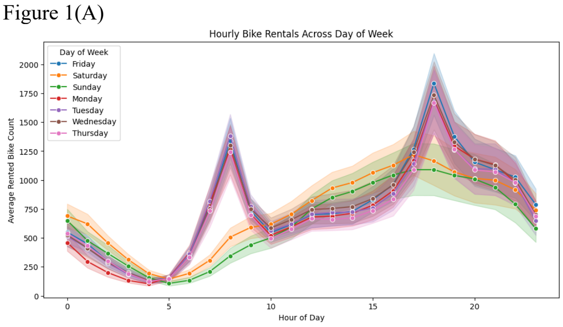


Figure 1(C)

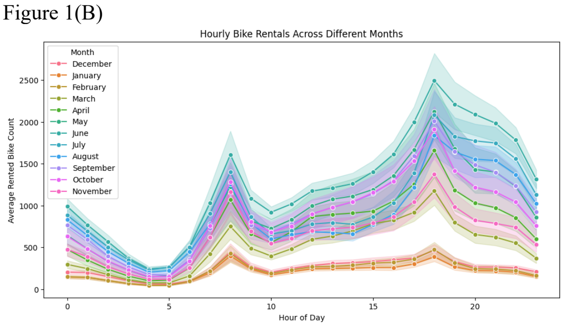


Figure 1(D)

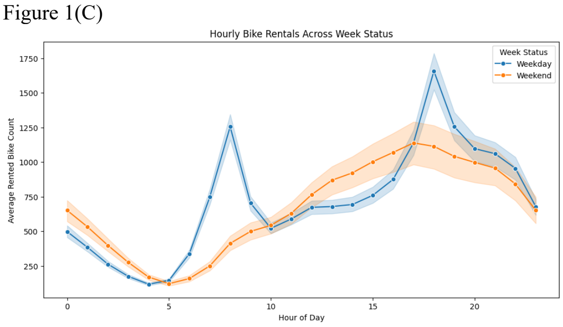


Figure 1(E)

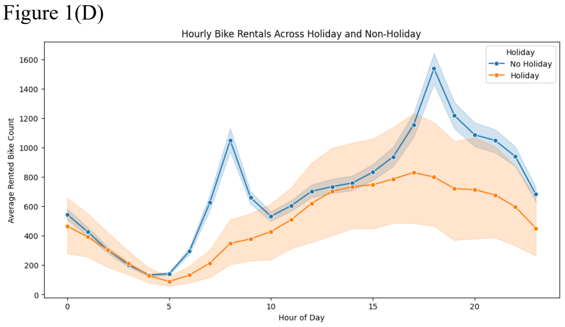


Figure 1(F)

The analysis reveals key relationships between weather features and bike rental counts. The heatmap Figure 2(A) shows that temperature has a strong positive correlation with rentals (0.56), followed by dew point temperature (0.40) and solar radiation (0.27). Visibility (0.21) and wind speed (0.13) show weaker positive correlations, while humidity (-0.20), snowfall (-0.15), and rainfall (-0.13) negatively impact bike usage.

The histograms Figure 2(B) indicate that temperature and dew point temperature follow a roughly normal distribution, while solar radiation, rainfall, visibility, and snowfall are highly skewed with extreme values, suggesting a feature reconstruction. Rented bike count and wind speed exhibit right-skewed distributions, prompting the use of a log transformation to stabilize the variance. These findings inform feature engineering and lay the groundwork for effective predictive modelling.

.

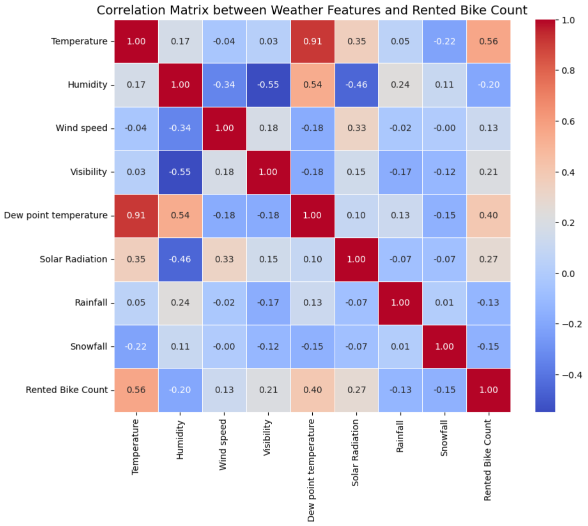


Figure 2(A)

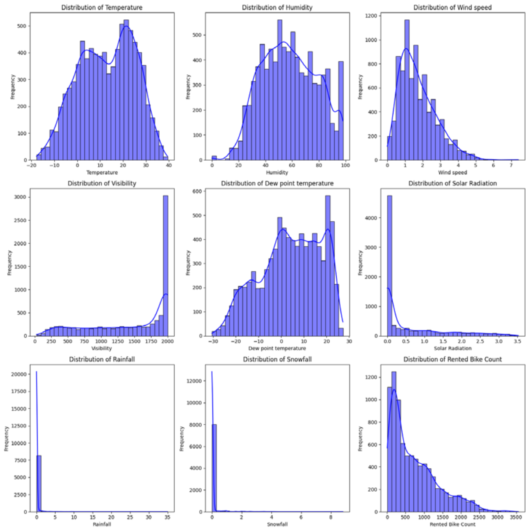


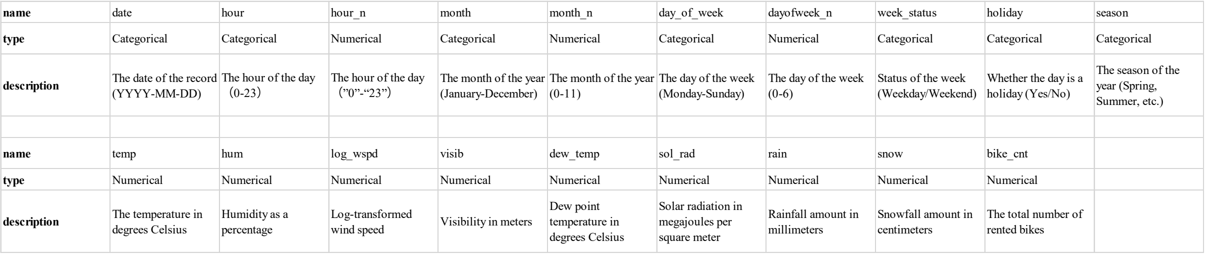
Figure 2(B)

Data Preprocessing

The steps of data preprocessing include:

- Convert 'Date' column to datetime format, ensuring day-first notation.
- Extract 'Month' and 'Day of Week' from 'Date' column, and add a column to replace numbers with names.
- Label days as 'Weekend' or 'Weekday'.
- Remove rows with inconsistent values for 'Holiday' and 'Functioning Day'.
- Remove rows with 'Functioning Day' equal to 'No'.
- Apply a logarithmic transformation to the 'Wind speed' column to create 'Log_Windspeed'.
- Add a column to convert the 'Hour' column to string type.
- Rename the columns.

The preprocessed data includes the features below:



name	date	hour	hour_n	month	month_n	day_of_week	dayofweek_n	week_status	holiday	season
type	Categorical	Categorical	Numerical	Categorical	Numerical	Categorical	Numerical	Categorical	Categorical	Categorical
description	The date of the record (YYYY-MM-DD)	The hour of the day （0-23）	The hour of the day （”0”-“23”）	The month of the year (January-December)	The month of the year (0-11)	The day of the week (Monday-Sunday)	The day of the week (0-6)	Status of the week (Weekday/Weekend)	Whether the day is a holiday (Yes/No)	The season of the year (Spring, Summer, etc.)
name	temp	hum	log_wspd	visib	dew_temp	sol_rad	rain	snow	bike_cnt	
type	Numerical	Numerical	Numerical	Numerical	Numerical	Numerical	Numerical	Numerical	Numerical	
description	The temperature in degrees Celsius	Humidity as a percentage	Log-transformed wind speed	Visibility in meters	Dew point temperature in degrees Celsius	Solar radiation in megajoules per square meter	Rainfall amount in millimeters	Snowfall amount in centimeters	The total number of rented bikes	

Fig 3

Feature Engineering and Feature Selection

Given that rain and snow contain a large number of zero values, these variables will be transformed into binary (rain_snow_n) and categorical (rain_snow) indicators to avoid sparsity issues. Similarly, sol_rad and visib also exhibit a large number of extreme values, leading to sparsity issues in the dataset. To address this, sol_rad and visib are divided into three categories: Low, Medium, and High, to capture the potential non-linear relationship with user behaviour. New features 'sol_rad_level' and 'visib_level' are created.

To improve interpretability 'dew_temp' is dropped, which is highly related to 'temp'. On trial, the less relevant feature 'log_wspd' and 'visib_level' are also dropped.

Further feature engineering methods vary depending on the model type. For GLM, numerical features undergo spline transformations to capture non-linear relationships, followed by standardization to ensure all features are on a similar scale. Categorical features are encoded using One-Hot Encoder. For LGBM, the feature engineering strategy relies on the Categorizer approach.

Model Training and Fine Tuning

For GLM and LGBM, different strategies are used to achieve optimal model configuration. The GLM is trained using a GeneralizedLinearRegressor with a Tweedie distribution (power=1.5), suitable for zero-inflated continuous data like hourly bike demand. The model includes an intercept and applies ElasticNet regularization, combining L1 (Lasso) and L2 (Ridge) penalties.

To optimize the model, a grid search was performed over two hyperparameters:

- alpha (regularization strength): [0.001, 0.01, 0.1, 1, 10]
- l1_ratio (L1-L2 balance): [0, 0.25, 0.5, 0.75, 1]

Using 5-fold cross-validation and Mean Absolute Error (MAE) as the scoring metric, the best parameters were alpha=0.001 and l1_ratio=0.25, achieving a train MAE of 185.68. While the model offers strong interpretability, its linear nature limits its ability to capture complex non-linear relationships in the data.

From the coefficient plot, time-related features show that commuting and nightlife hours positively impact bike demand, while late-night hours (3-6 AM) have negative effects, reflecting low demand. Weekends exhibit reduced demand, and "No Holiday" indicates increased bike usage for leisure activities. Seasonal and monthly patterns reveal that cold seasons strongly reduce demand, with month-specific effects being weaker due to overlap with seasonal trends. Weather-related features highlight that low solar radiation increases demand, while rain and snow negatively affect rentals. High humidity and extreme temperatures also reduce demand, likely due to user discomfort.

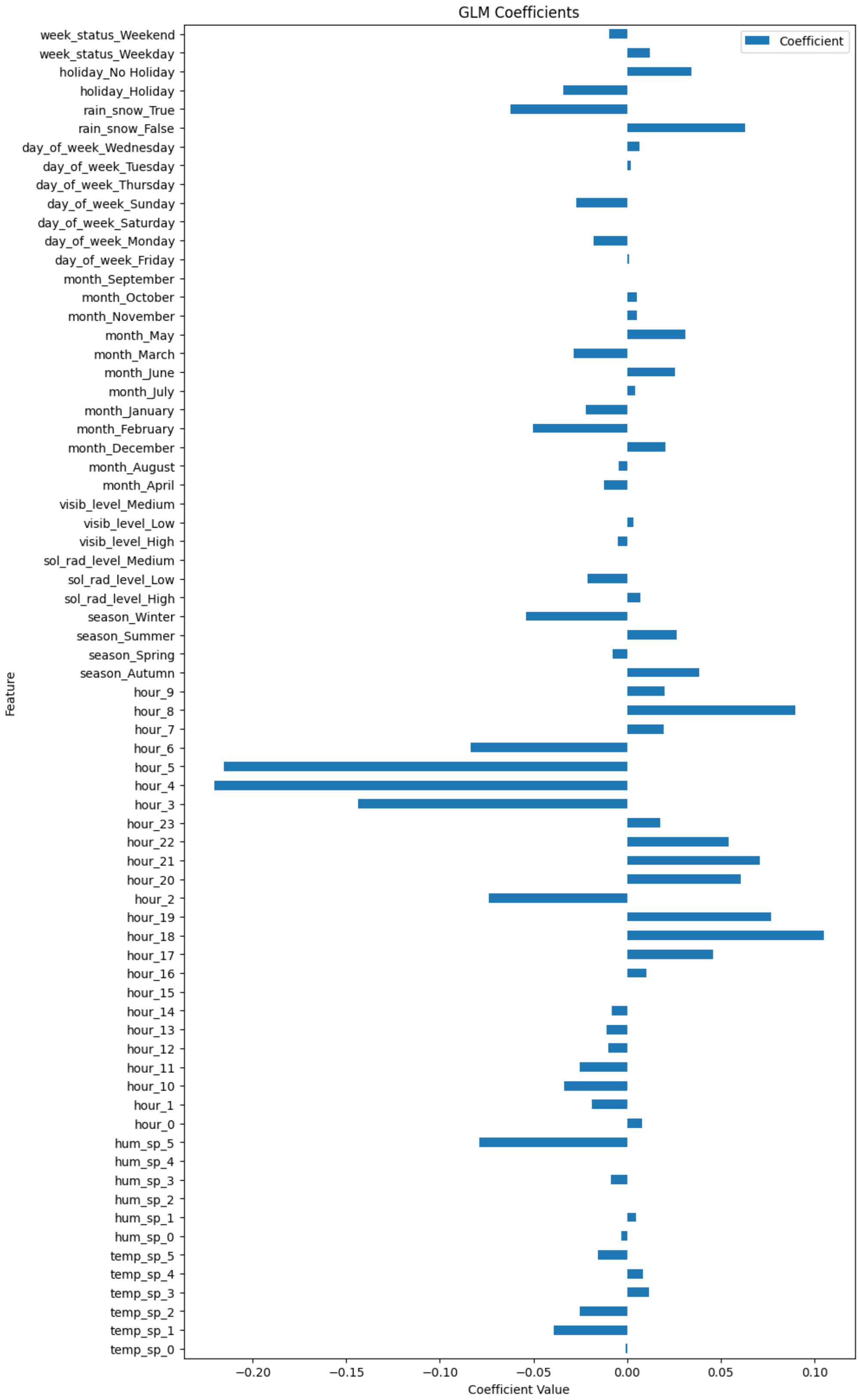

GLM Coefficients
Coefficient
Feature
week_status_Weekend
week_status_Weekday
holiday_No Holiday
holiday_Holiday
rain_snow_True
rain_snow_False
day_of_week_Wednesday
day_of_week_Tuesday
day_of_week_Thursday
day_of_week_Sunday
day_of_week_Saturday
day_of_week_Monday
day_of_week_Friday
month_September
month_October
month_November
month_May
month_March
month_June
month_July
month_January
month_February
month_December
month_August
month_April
visib_level_Medium
visib_level_Low
visib_level_High
sol_rad_level_Medium
sol_rad_level_Low
sol_rad_level_High
season_Winter
season_Summer
season_Spring
season_Autumn
hour_9
hour_8
hour_7
hour_6
hour_5
hour_4
hour_3
hour_23
hour_22
hour_21
hour_20
hour_2
hour_19
hour_18
hour_17
hour_16
hour_15
hour_14
hour_13
hour_12
hour_11
hour_10
hour_1
hour_0
hum_sp_5
hum_sp_4
hum_sp_3
hum_sp_2
hum_sp_1
hum_sp_0
temp_sp_5
temp_sp_4
temp_sp_3
temp_sp_2
temp_sp_1
temp_sp_0
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
Coefficient Value

Fig 4

The LGBM model uses a LightGBM Regressor with the Tweedie objective, ideal for zero-inflated data. Hyperparameter tuning is performed using RandomizedSearchCV with 5-fold cross-validation and Mean Absolute Error (MAE) as the scoring metric. The tuning was performed over four hyperparameters:

- learning_rate (learning rate for tree updates): [0.01, 0.05, 0.1]
- n_estimators (the number of boosting rounds): [500, 1000, 1500]
- num_leaves (the maximum number of leaves in a tree): [30, 40, 50]
- min_child_weight (the minimum weight of data points for a leaf node): [0.001, 0.002, 0.004]

The search process found the following optimal combination: {'estimate__num_leaves': 40, 'estimate__n_estimators': 1000, 'estimate__min_child_weight': 0.001, 'estimate__learning_rate': 0.05}, and achieved a Mean Absolute Error (MAE) of 94.25, which is significantly better than the GLM's performance.

According to Figure 5(A) and Figure 5(B) , temperature (temp) is the most influential feature, with higher temperatures positively impacting demand. Humidity (hum) has a strong negative effect at higher levels, reducing bike usage. Hour (hour_n) captures daily patterns, showing peaks during commuting hours and drops during late-night periods. While Rain/Snow (rain_snow_n) is less significant in feature importance, SHAP reveals its clear negative impact on demand during adverse weather. Temporal factors like Day of the Week and Month show moderate importance, reflecting seasonal and weekly variations. Overall, both plots confirm that temperature, humidity, and hour are the primary drivers of bike demand.

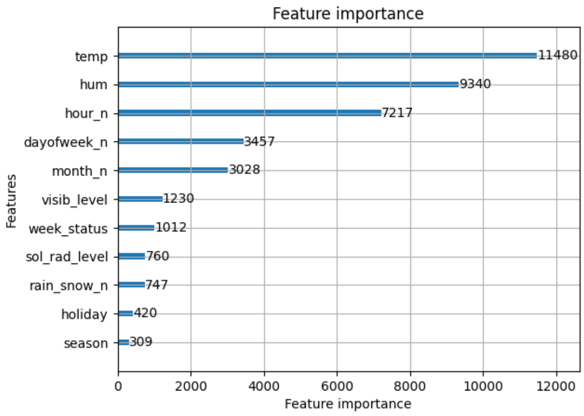


Figure 5(A)

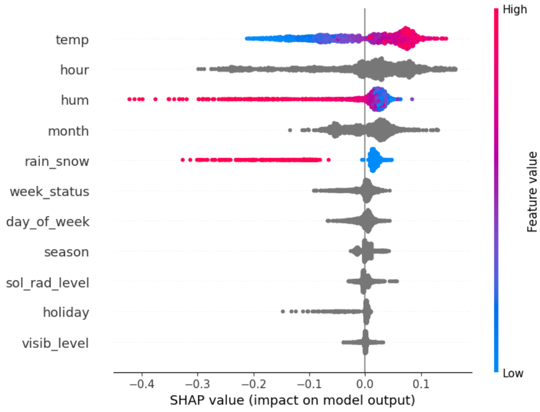


Figure 5(B)

Model Evaluation and Interpretation

By analysing the metrics in Figure 6, I find that LGBM outperforms GLM across all metrics, showcasing lower prediction error (MAE), better accuracy (MSE), and a stronger model fit (R^2 Score). However, there is some degree of overfitting for LGBM. The performance of the training set is significantly better than that of the test set.

Metric	GLM (Train)	GLM (Test)	LGBM (Train)	LGBM (Test)
MAE	183.55	180.93	50	82.95
MSE	78,551.11	77,727.08	8,217.79	22,331.73
R^2	0.8127	0.8007	0.9804	0.9428

Figure 6

A further comparison of predicted and actual bike demand is conducted. While GLM captures the overall trend, it struggles with higher actual values, where predictions increasingly underestimate the actual values. This highlights the model's limited ability to capture complex, non-linear relationships, particularly for moderate to high demand (1000+). In contrast, LGBM predictions align closely with the diagonal line across the range. For lower values near zero, LGBM demonstrates minimal scatter and accurately captures the trend. At higher values (above 2000), predictions remain consistent but show slight deviations, suggesting minor underfitting or challenges in modelling extreme values. Overall, LGBM performs better across all ranges.

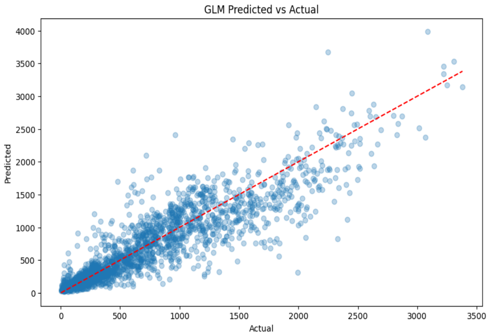


Figure 7(A)

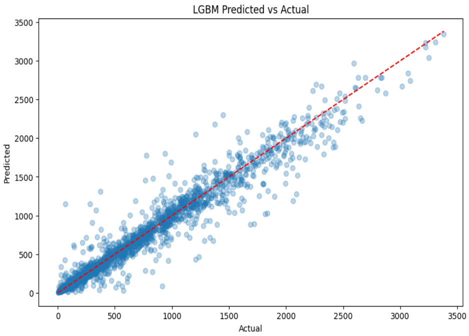


Figure 7(B)

The Lorenz curve compares the predictions of GLM and LGBM models against a random baseline. The **LGBM model** (Gini index: 0.471) slightly outperforms the **GLM model** (Gini index: 0.466), as shown by its closer alignment to the top-right corner. Both models significantly outperform the random baseline, but the LGBM curve shows slighly better predictive power, especially for higher cumulative outcomes. This indicates that LGBM captures more variance and produces more accurate predictions overall.

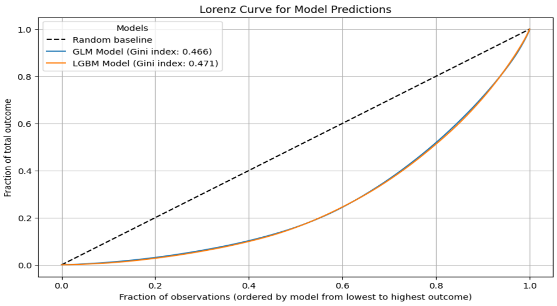


Figure 8

The charts below show the importance of variables in the GLM and LGBM model based on drop-out loss, which measures how much the loss increases when a specific variable is removed. The hour, temperature, humidity and rain or snow showed high importance in both models. While season is more important for GLM and month is more important for LGBM.

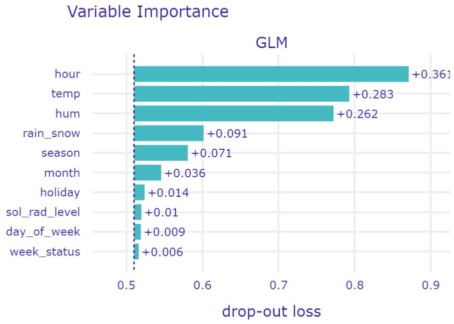


Figure 9(A)

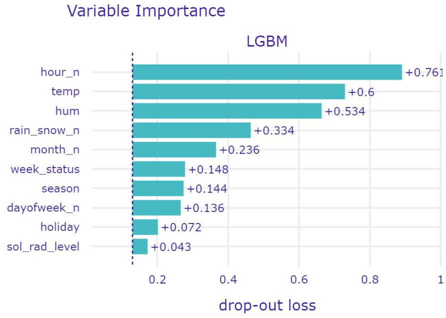


Figure 9(B)

Therefore, the partial dependences of hour of the day (hour_n), temperature (temp), humidity (hum), and temporal features such as day of the week (dayofweek_n) and month (month_n) will be assessed. Temperature shows an increase in demand, peaking around 20-25°C before stabilizing. Humidity causes a sharp drop in demand when it exceeds 75%, likely due to discomfort. Hour of the day highlights clear peaks during commuting hours (8-9 AM and 6-7 PM) and minimal demand during late night and early morning. Rain or Snow exhibits a linear decline, confirming reduced demand under adverse weather conditions. Month shows peak demand in spring and summer, with a decline toward year-end.

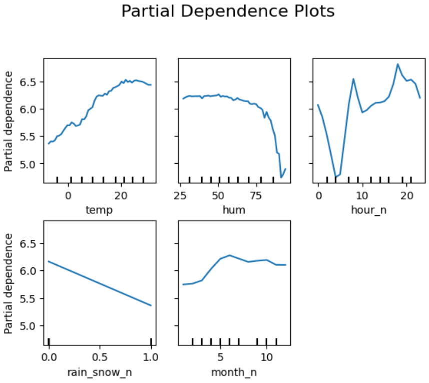


Figure 10

Outlook for Improvement

The model effectively captures the impact of weather conditions, time patterns, and seasonal variations on bike demand, but there is room for improvement.

To start with, integrating additional data sources like real-time traffic congestion, road closures, and event schedules (e.g., concerts or festivals) could help account for external factors influencing demand. Including user demographics, such as age or occupation, may also uncover hidden usage patterns.

Refining feature engineering can improve the robustness of the analysis. For instance, creating interaction terms between weather and time-based features could help uncover hidden non-linear relationships, leading to better predictions.

To make the process more efficient, leveraging optimized algorithms or distributed computing can speed up large-scale analysis. Automating data preprocessing and model tuning would enable real-time demand forecasting.

With these enhancements, the model can provide a strong foundation for smarter urban mobility and more sustainable transportation systems.

Reference

- "Seoul Bike Sharing Demand." UCI Machine Learning Repository, 2020, https://doi.org/10.24432/C5F62R.
- E, Sathishkumar V, et al. “Using data mining techniques for bike sharing demand prediction in Metropolitan City.” *Computer Communications*, vol. 153, Mar. 2020, pp. 353–366, https://doi.org/10.1016/j.comcom.2020.02.007.
- OpenAI. ChatGPT. Version 4, OpenAI, 2024, https://openai.com/chatgpt. Used for debugging tasks.