Università degli Studi di Padova

Department Information Engineering

Bachelor Thesis in Information Engineering

Optimizing Bike Sharing Systems in Smart Cities: A Machine Learning Forecasting Model

Supervisor
Federico Chiariotti
Università di Padova

Bachelors Candidate
Jad Kaedbey

Abstract

Bike sharing systems have emerged as a cornerstone of urban mobility in smart cities, offering numerous benefits including enhanced convenience, reduced traffic congestion, lower emissions, and improved public health. These systems also present unique opportunities for optimization through advanced data analytics, given the voluminous data they generate. This thesis presents an innovative approach to predicting bike sharing demand in London, a city with a substantial bike sharing infrastructure encompassing over 788 bike sharing stations, and over 10 million yearly bike trips. Leveraging sophisticated machine learning architectures, such as XGBoost (eXtreme Gradient Boosting) models, this research aims to accurately forecast demand patterns. By integrating comprehensive weather data and historical usage patterns, the study develops a predictive framework with a Mean Absolute Error of less than 2 that not only enhances the efficiency of bike sharing services but also contributes to the broader goals of urban sustainability and mobility. The methodologies and findings of this work hold significant implications for urban planners, policy makers, and bike sharing operators, offering a data-driven foundation for optimizing the deployment and management of bike sharing systems in London and other smart cities globally.

Summary

This thesis presents a comprehensive analysis of predictive modelling for bike-sharing systems in London, using a wide range of data sources, such as trip data, weather conditions, and urban infrastructure metrics. The study highlights the importance of integrating diverse datasets to improve the accuracy of bike rental demand predictions. It emphasizes the impact of weather conditions, time variables, and station characteristics, as demonstrated by a thorough literature review.

The study uses advanced machine learning methods, specifically XGBoost, and a custom version of XGBoost with an asymmetric loss function to create predictive models. These models outperform basic linear regression, demonstrating their ability to effectively handle complex non-linear relationships. The XGBoost model achieved impressive results with a Root Mean Squared Error (RMSE) of 3.4596 and a Mean Absolute Error (MAE) of 1.6043 on the test set, indicating strong predictive accuracy and reliability. These metrics, calculated on the original linear scale of the data following exponential back-transformation, reflect the actual differences in predicted bike demand. The findings suggest that weather conditions have a significant impact on bike-sharing usage, with higher temperatures increasing demand and rainfall decreasing it. Similarly for temporal patterns and spatial distribution of stations.

These insights provide valuable implications for urban planners and policymakers, forming a basis to improve the efficiency and satisfaction of bike-sharing systems. Strategies resulting from this research could involve optimizing station placement, adjusting bike allocations, and ensuring service reliability under varying conditions.

In essence, this thesis lays a robust groundwork for advancing the operational effectiveness of bike-sharing systems and contributing to more sustainable urban transport solutions, supported by quantitative metrics that underscore the models' high level of performance.

Contents

Listing of figures

Listing of tables

1 Introduction

During recent years, bike-sharing systems have ascended to become a pivotal component of urban mobility strategies in cities worldwide [1]. London, with its expansive infrastructure featuring over 800 bike-sharing stations and facilitating more than 10 million trips, exemplifies this global trend. These systems have been proven to reduce traffic congestion and emissions, while promoting public health and providing a flexible, eco-friendly transportation alternative [2]. However, the large scale of growth bike sharing systems have experienced introduces complex operational challenges, especially concerning the optimization of bike distribution over cities and bike availability to meet volatile and inconsistent user demand.

The essence of these challenges lies in the inherently unpredictable nature of urban mobility. Factors such as weather, traffic patterns, urban events, and even day-to-day variability in human behavior contribute to fluctuations in bike-sharing demand across different parts of the city and at different times [3]. The objective of ensuring that bikes are available where and when they're needed, without significant overstocking or shortages, poses a significant operational hurdle for bike-sharing operators. Moreover, the inefficiency in bike distribution not only degrades user experience but also undermines the broader objectives of urban sustainability and mobility that bike-sharing systems play a pivotal part in [4].

Traditional approaches to addressing these challenges have largely been reactive or based on simplistic predictive models that fail to capture the complexity and multidimensionality

of the factors influencing demand.

As operational needs grow dramatically with the growth of bike sharing systems, so does the necessity for innovative approaches that correctly exploit the wealth of data provided by the various bike sharing systems. With IoT systems becoming an integral part in many of the world's smart cities, a unique opportunity is presented to harness advanced methods in data analytics for the prediction of demand on bike sharing systems.

Machine learning emerges as a powerful tool to make solid predictions in this context. Among the plethora of machine learning methodologies, XGBoost (eXtreme Gradient Boosting) has been chosen to make this prediction. XGBoost's success in various applications, such as chronic kidney disease diagnosis [5], Gene expression value prediction [6], and in Bankruptcy prediction [7], suggest its potential to revolutionize demand forecasting in bike-sharing systems by accurately modelling the complex relationships between demand and its influencing factors. By carefully selecting adept features, and including historical usage patterns and comprehensive weather data, a predictive framework that is both accurate and actionable has been developed. The essence of XGBoost lies in its ability to create a highly efficient, flexible, and portable model by sequentially combining a set of decision trees to form a strong predictive model. Each new tree is built to correct the residual errors made by the preceding sequence of trees, with the aim of minimizing a loss function that measures the difference between the predicted and actual values.

This thesis contends with the multifaceted problem of optimizing bike-sharing systems in London's urban landscape through a data-driven approach employing XGBoost models [8]. It aims to transcend the limitations of current forecasting methodologies by developing a model that not only predicts demand with high accuracy but also provides insights into the dynamics of urban mobility. Such a model could serve as a cornerstone for dynamic rebalancing strategies, ensuring an optimal distribution of bikes across the city to meet real-time demand. Furthermore, by enhancing the efficiency of bike-sharing services, this research contributes to the broader goals of urban sustainability, reducing environmental impact, and fostering a more livable, mobile, and connected city [9].

In synthesizing a solution to these challenges, this thesis investigates the influence of various factors on bike-sharing demand. The research also examines the scalability and flexibility of the developed model, considering its applicability to other urban settings and its integration into the operational frameworks of bike-sharing systems. By doing so, it endeavors to offer urban planners, policymakers, and bike-sharing operators a robust and

data-driven foundation for enhancing the deployment, management, and strategic planning of bike-sharing systems, not just in London, but in smart cities globally [9].

2

Literature Survey

In order to discuss predictive modelling for bike-sharing systems objectively, it is important to consider the various factors that affect bike rental demand. These factors include environmental conditions, temporal patterns, and urban infrastructure characteristics. Previous research has emphasised the importance of integrating diverse datasets, such as weather conditions, time variables, and urban mobility metrics, to develop accurate predictive models. This statement introduces the examination of studies that explore the relationship between bike-sharing demand and critical influencers.

A statistical model for predicting the quantity of bikes hired every hour has been created for Lyon's Vèlo'V bike sharing systems [10]. The model included some factors like the quantity of subscribers, the time information during the week, strikes and holidays, as well as rain and temperature data. The weather data associated with each bike trip, however, was taken at a daily frequency. The average temperature over the day was used, instead of the temperature data at the time of the start of the bike trip. A similar situation avails regarding the precipitation data, as total rain volume for the entire day was utilized. The event of rain typically lasts for a few hours, sometimes minutes, and it is not a feature that would affect user behavior after the time of its termination, similarly to temperature which seems to also fluctuate drastically from morning to evening, especially throughout winter and spring seasons [11]. This study did not evaluate the possible use of machine learning

algorithms, and uses a rather miniature dataset, generated from 4000 bicycles and 334 stations.

The trip generation and trip attraction models from the city of Toronto revealed that bike ridership was positively correlated with higher temperatures, lower humidity levels, and lower amounts of ground snow. The thesis also emphasized the significance of station capacity and station-to-station proximity in providing sufficient bicycles for trip generation and docking spaces for trip attraction [12].

The investigation conducted in Daejeon also found similar results using clustering analysis to identify how weather affects groups of stations with similar properties. The study concluded that high temperature and humidity have a negative correlation with the daily demand for bicycles. A system-level examination was also conducted, which showed that certain variables had significant impacts at various times of the day. Specifically, temperature, rainfall, and whether it was a workday affected the rental bike demand at specific times [13].

In the analysis carried out by Aalborg University, k-means clustering is introduced to cluster stations based on the shape of their average daily traffic patterns. Stations were found to have different types of traffic patterns, which were then related to external spatial factors using a logistic regression model. The proposed model was able to predict demand with a Mean Absolute Error of 36.4 for the city of London and an MAE of 17.7 for Washington D.C. instead [14]. The thesis however did not indulge into any further modelling attempts.

The study by the University of Adana [15] highlights the significant role of environmental conditions, urban infrastructure, and socio-demographic characteristics in shaping the usage patterns of bike-sharing systems. The study integrates these variables to provide a holistic view of the dynamics influencing bike-sharing systems, suggesting that a multifaceted approach is essential for optimizing these systems for increased urban mobility and sustainability.

The survey carried out by Beijing University [16] explored the use of advanced machine learning techniques, specifically neural networks, to predict bike sharing usage. XGBoost is used as a baseline by many of the papers cited. The thesis also discusses the challenges associated with predictive modelling in this domain, including the integration of temporal and spatial data variations and the need for real-time prediction capabilities to adapt to rapid changes in user demand. Future directions highlighted in the thesis suggest a deeper

exploration of hybrid models that combine XGBoost with other deep learning architectures to improve predictive performance and adaptability in predicting bike sharing usage.

XGBoost was put to the test against a Deep Neural Network by the University of West Attica [17]. In the thesis, XGBoost displayed better accuracy results compared to the DNN, scoring better on both the learning rate and the prediction accuracy. Although these results are case specific, they highlight the possibility of using XGBoost as a reliable benchmark for forecasting models.

The selection of an XGBoost model over a deep learning approach for demand forecasting in bike-sharing services is advantageous for several reasons. Firstly, XGBoost is generally simpler to implement and can be faster to train compared to deep learning models. This is due to the fact that deep learning models often require large amounts of data to perform well and avoid overfitting. Secondly, XGBoost is more interpretable than deep learning models, where the contribution of each feature given to the model can be parameterised. This is crucial for business decisions and understanding demand dynamics in bike-sharing.

The decision trees, which constitute the base learners in XGBoost, provide a clear visualisation of the manner in which decisions are made [18], thereby facilitating the communication of results to stakeholders or decision makers. XGBoost frequently performs exceptionally well on structured, tabular data [8], which is a common format for historical bike-sharing usage data that includes features such as time of day, day of the week, weather conditions, and location. In contrast, deep learning models are particularly adept at handling data with complex, interrelated structures, such as image and speech recognition. XGBoost is well-suited to handle sparse data [8], which is often encountered in bike-sharing systems, where many time slots may have low or no demand. All while allowing for extensive customization of the model training process, including the optimization objective and the evaluation metric.

This approach differs from many existing studies, which predominantly employ simpler statistical models that are unable to comprehensively capture these complex interactions. Unlike some of the existing literature, which often limits analysis to basic correlations and linear regressions, this work integrates a wider range of predictive variables in a non-linear modelling context, allowing for more accurate and actionable insights, which are crucial for real-time decision-making in bike-sharing management and improvement.

3

Data Pre-Processing

3.1 London: Data Selection and Integration

The city of London was selected based on both the size and maturity of its bike share systems and the availability of other data used in the analysis. The bike sharing data is published on Transport For London (TfL) [19] cycling data portal [20]. The data has been provided in chunks of 2 weeks each, starting from the 28th of December, 2016 till the 2nd of January 2018. The excess data was dropped for homogeneity. The data, provided in Comma Seperated Value - CSV format - was then combined into one large dataframe, using the python Pandas library for ease of manipulation and formatting. Some other discrepancies were found in the data, specifically in the Trip Start Date data formatting for the months of September and October.

In addition to bike share trip data, we have collected other types of data that are not directly related to bike share systems. These types of data include geographic locations of bike share stations across the city, as well as weather data provided by Open-Meteo [21]. As these types of data are external to the bike share system.

In this project, we use data sets from 2017, as this is a recent year with normal operations prior to the COVID-19 pandemic, which also provides data completeness throughout the year. Data from the year 2019 and onwards is largely affected by the effects of the COVID-19 virus and various mobility restrictions imposed on the city. The datasets used

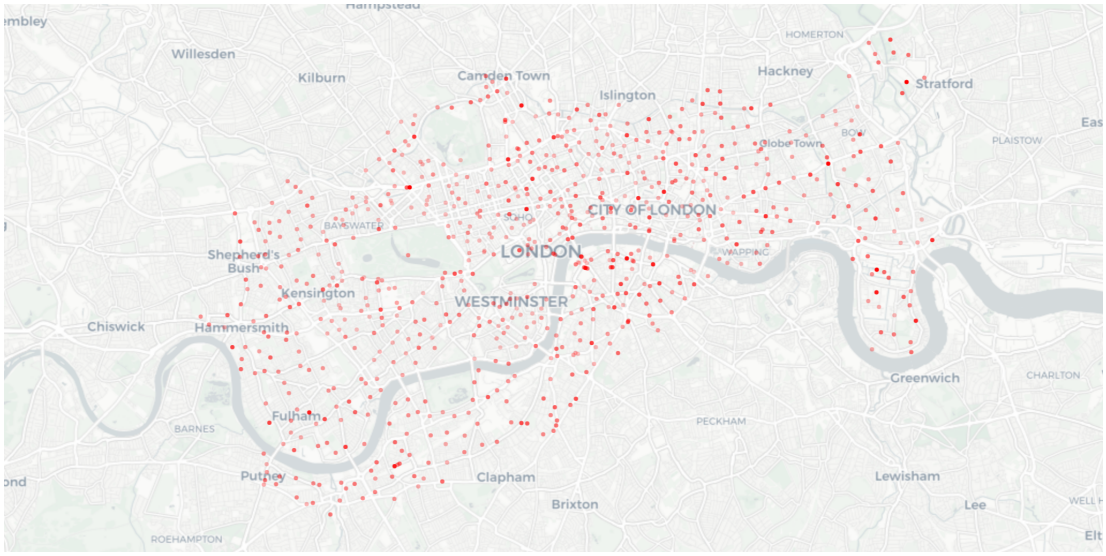

Figure 3.1: Interactive Map of London Bike Sharing Stations: Opacity-Coded Visualization Based on Bike Capacity

contain data on each individual journey made on the system, including journey duration, time of departure from the start station, start station ID, start station name, time of arrival at the end station, end station ID and end station name. The station locations were not initially provided in the dataset, but were obtained from TFL [22], who have a live "Cycle Hire Updates" feed, which lists information for each cycle hire station, updated approximately every minute. The live data was not exploited - instead only the name, ID, latitude, longitude and capacity for each cycle hire station was taken. We assume that the changes to station location and capacity over this time frame were relatively small in terms of the overall ridership of the bike share systems.

3.2 Bike-sharing Station Capacity and Demand Analysis

The locations of the stations were plotted using the Folium library in Python [23] and opacity was coded based on station capacity, as shown in Figure 3.1. The plot is interactive, showing the station name, district, and capacity when hovered over with the cursor. The interactive version can be found in the GitHub repository for this project [24].

Each station in the system was then assigned a specific service area. These service areas are determined using a Voronoi tessellation for each station location and within the defined boundary to prevent Voronoi polygons from being unnecessarily large, especially for

stations on the outer edges of the city centre [25]. Three different choropleth maps were created to showcase 3 metrics: Daily Journey Count per Station Capacity, Median Journey Duration, and Total Yearly Trips. For the sake of simplicity, the daily journey count was calculated by taking the total yearly trips for each station, then taking the average over 365 days. This plot is also interactive, and displays the exact numbers for each metric. Table 3.1 summarizes the Minimum and Maximum values for each metric measured from the plots.

Metric	Minimum		Maximum	
	Station Name	Value	Station Name	Value
Yearly Journey Count	Grant Road Central, Clapham Junction	878	Belgrove Street , King's Cross	97362
Daily Journey Count	Grant Road Central, Clapham Junction	2.41	Belgrove Street , King's Cross	266.75
Median Duration [Minutes]	Barons Court Station, West Kensington	6.0	Hyde Park Corner, Hyde Park	28.0
Journey Count per Capacity	Castalia Square, Cubitt Town	0.1	Hyde Park Corner, Hyde Park	8.6

Table 3.1: Summary of Journey Metrics by Minimum and Maximum Values

The Hyde Park station seems to dominate in terms of total bike share volume, which may indicate a need to expand its overall capacity or even introduce a new station in the area. As this station is very close to Hyde Park, we can assume that the reason for this volume of trips is leisure. Although Hyde Park station also dominates the median duration metric, we can clearly see a trend in the stations located on the outskirts of the city centre. This suggests that people who cycle from the outskirts of the city are likely to end up commuting into the city center, hence the longer journey times. The metric Journeys per Station Capacity gives a strong indication of where more bikes could be introduced into the system. The southern part of the borough of Tower Hamlets clearly shows a lower number of trips per capacity compared to Westminster.

In contrast to other similar projects, data from stations with low daily traffic were not removed from the system. This was done to ensure that the model would still be robust to different types of station.

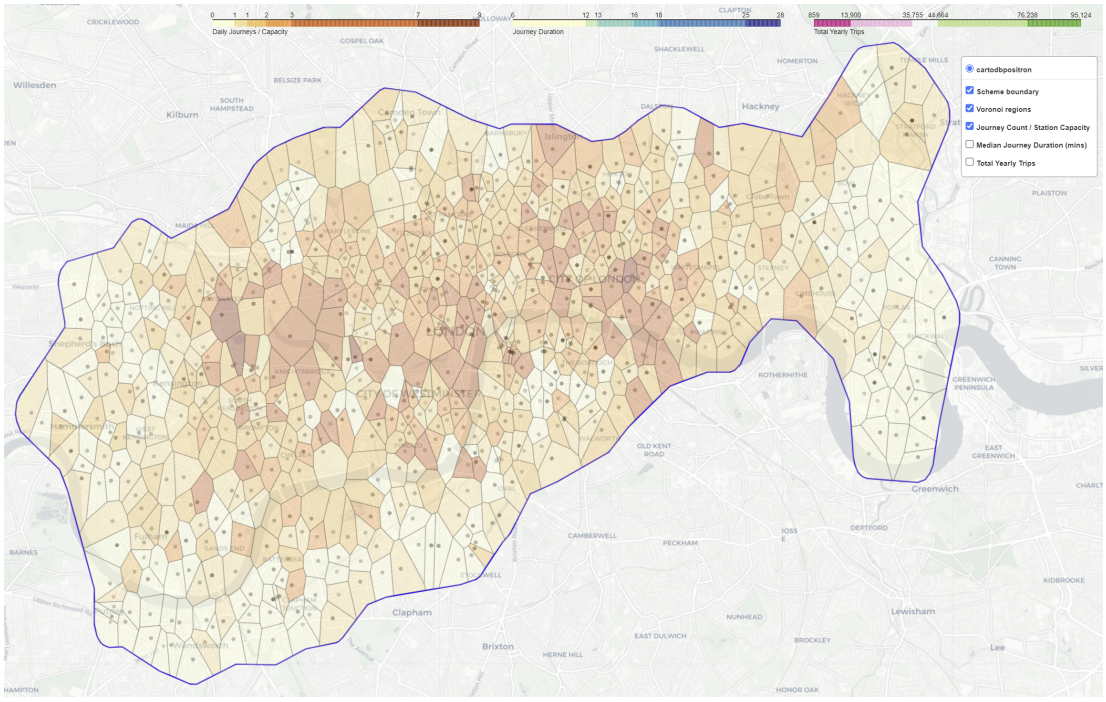

Figure 3.2: Distribution of Daily Bike Share Journeys Relative to Station Capacity Across London

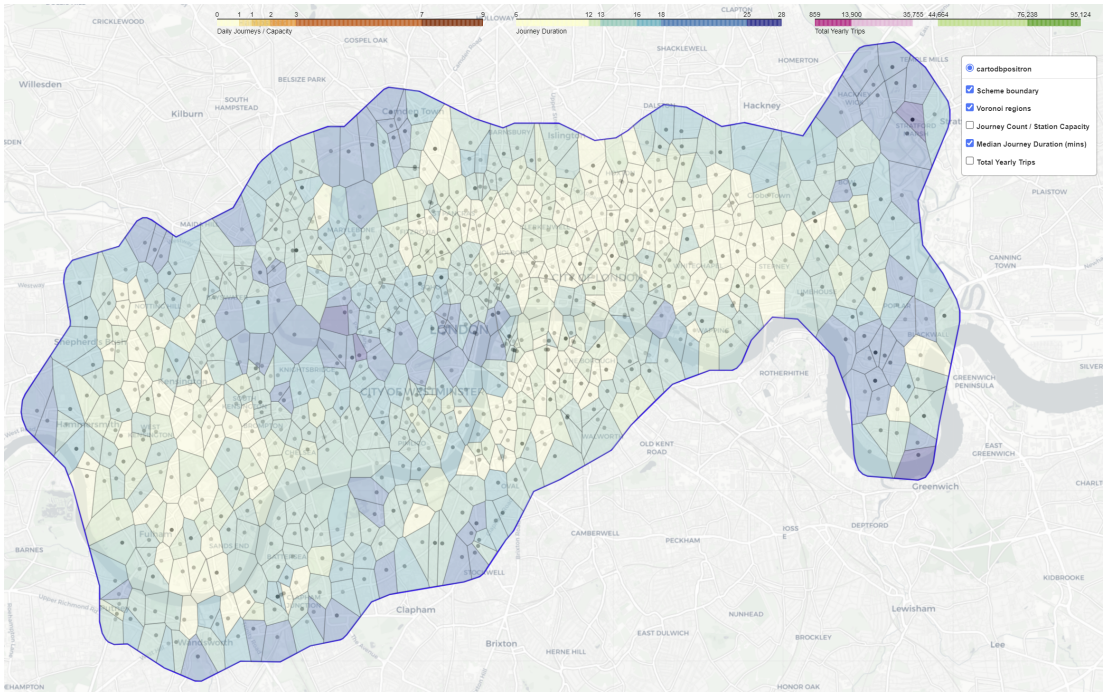

Figure 3.3: Median Duration of Bike Share Journeys in London

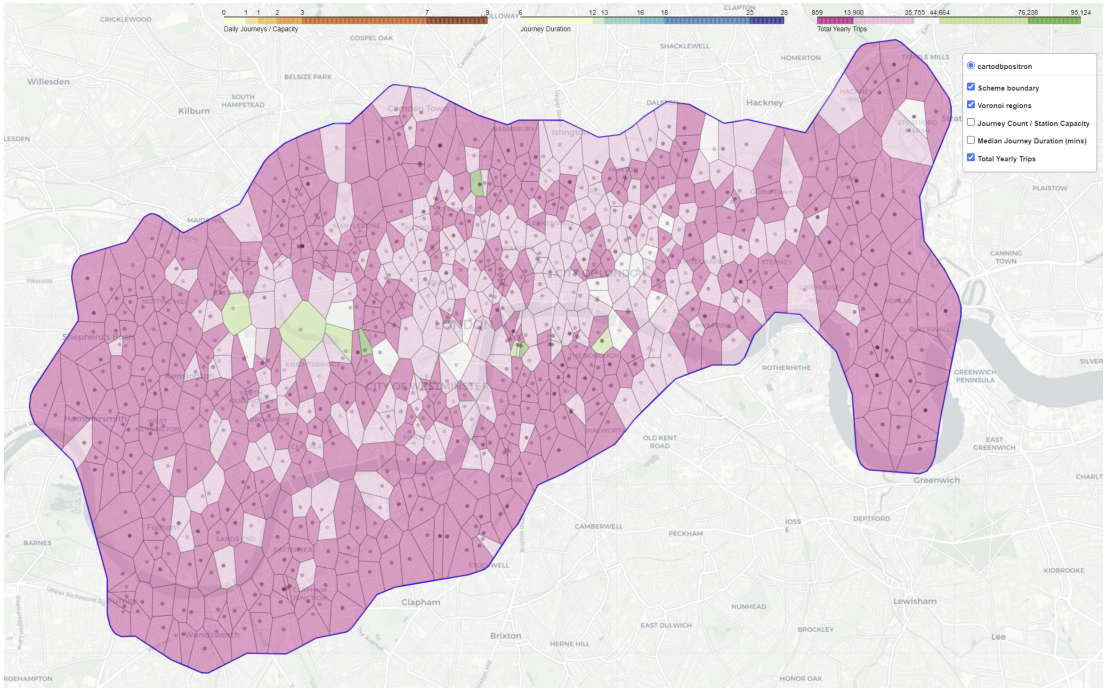

Figure 3.4: Annual Bike Share Usage: Total Yearly Trips by Station in London

3.3 Data Integration and Feature Engineering

The weather data used with the cycling trip data was obtained using the Open-Meteo API. The data was provided as 8760 entries, each corresponding to every hour of the year 2017, and included the parameters in table 3.2

As the data provided by TFL has each cycle trip recorded in the system as a separate row, the data has been grouped by Start Station Id and Start Station Date. The dataset had 10379323 rows and 10 columns before grouping. The grouped dataset was then merged with that of the weather data to form the complete dataset used to train the machine learning models. Some other parameters were then added to the model to further broaden the content of the dataframe and allow the model to better learn the patterns of the many bike share stations, resulting in a dataframe of shape: 3256360 rows and 24 columns. Such data is listed in table 3.3

Data	Description
temperature_2m (°C)	Air temperature at 2 meters above ground
relative_humidity_2m (%)	Relative humidity at 2 meters above ground
apparent_temperature (°C)	Perceived feels-like temperature combining various factors
precipitation (mm)	Total precipitation (rain, showers, snow) sum of the preceding hour
rain (mm)	Rain from large scale weather systems of the preceding hour
snowfall (cm)	Snowfall amount of the preceding hour in centimeters
weather_code (WMO code)	Numeric code representing weather condition (WMO code)
cloud_cover (%)	Total cloud cover as an area fraction
wind_speed_100m (km/h)	Wind speed at 100 meters above ground

Table 3.2: Summary of Weather Parameters

Table 3.3: Summary of Added Dataframe Columns

Column	Description
count_log	Logarithm of the count, used to normalize skewed data
is_holiday	Indicator (0 or 1) for whether the day is a public holiday
is_weekend	Indicator (0 or 1) for whether the day is part of the weekend
month	Month of the year (1-12)
day	Day of the month
hour	Hour of the day (0-23)
hour_sin	Sine transformation of the hour to capture cyclical nature in daily data
hour_cos	Cosine transformation of the hour to capture cyclical nature in daily data
month_sin	Sine transformation of the month to capture cyclical nature in annual data
month_cos	Cosine transformation of the month to capture cyclical nature in annual data
is_night	Indicator (0 or 1) for whether it is nighttime

3.4 Bike-Sharing Trends Across Workdays, Weekends, and Holidays

One striking feature of figure 3.5 is the pronounced peak hours on weekdays, which are likely to correspond to typical commuting times, possibly indicating a significant reliance on bike sharing for transport to and from work. In contrast, the pattern on weekends and holidays suggests a more leisurely use of the service, with a broader distribution of bike shares throughout the day. Furthermore, there is a noticeable shift in the timing of peak

usage during weekends and holidays, generally later in the day, which may reflect a more relaxed pace of life on non-working days. In addition, the average number of cycle trips on holidays tends to be slightly higher than on weekdays and weekends, suggesting a tendency towards leisure activities on off days. Furthermore, the evening peak on weekdays is quite pronounced compared to the relatively moderate increase during the same hours on weekends and holidays. This observation may suggest that evening leisure use does not compensate for the decrease in commuting on non-working days.

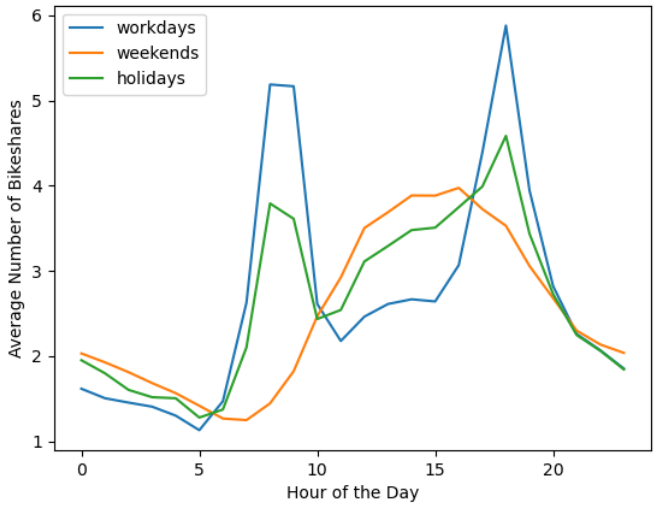

Figure 3.5: Comparative Analysis of Average Bike Shares by Hour on Workdays, Weekends, and Holidays

3.5 Dataset Variable Correlations

The main findings of figure 3.6 include the presence of strong positive correlations between the temperature variables (real_temp, feel_temp) and the number of bike trips (cnt), indicating that warmer conditions are conducive to higher cycle use. There is a noticeable negative correlation between rainfall variables (rain, precipitation) and the number of bike-share users, highlighting the negative impact of rainfall on bike-share demand. Temporal factors such as time of day also show significant correlations, reflecting usage patterns linked to daily human activities.

XGBoost is an excellent choice for this study as it is very effective at handling the non-linear relationships and interactions shown in the figures 3.5 & 3.6 results. This machine

learning algorithm excels in scenarios where both the strength and direction of variable relationships are critical. XGBoost's ability to handle model complexity and regularisation helps prevent overfitting, which is critical given the varying influences of weather and temporal factors on bike share demand. In addition, XGBoost's feature importance scores can provide deeper insight into which variables are most important in influencing bike share usage, allowing the algorithm to be further refined based on empirical evidence. XGBoost also implements tree-based learning algorithms, which are inherently good at capturing interactions between features. This is particularly useful in this case where the model needs to generalise well across different weather conditions, as well as a varying demand patterns throughout different parts of the weeks.

Another notable advantage of XGBoost is its efficiency in terms of computational resources [8]. Unlike more computationally intensive models such as deep neural networks, XGBoost is relatively lightweight, requiring less processing power and memory. This characteristic makes it especially suitable for scenarios where rapid model training and validation are beneficial. The speed of XGBoost does not compromise its performance, making it an excellent choice for iterative testing and real-time data analysis. Furthermore, the simplicity of deployment and the lesser need for high-performance hardware render XGBoost particularly attractive for practical applications where budget or hardware capabilities are constrained [26]. This ability to operate efficiently on limited resources, coupled with its robust performance, underscores its applicability in dynamic and resource-sensitive environments.

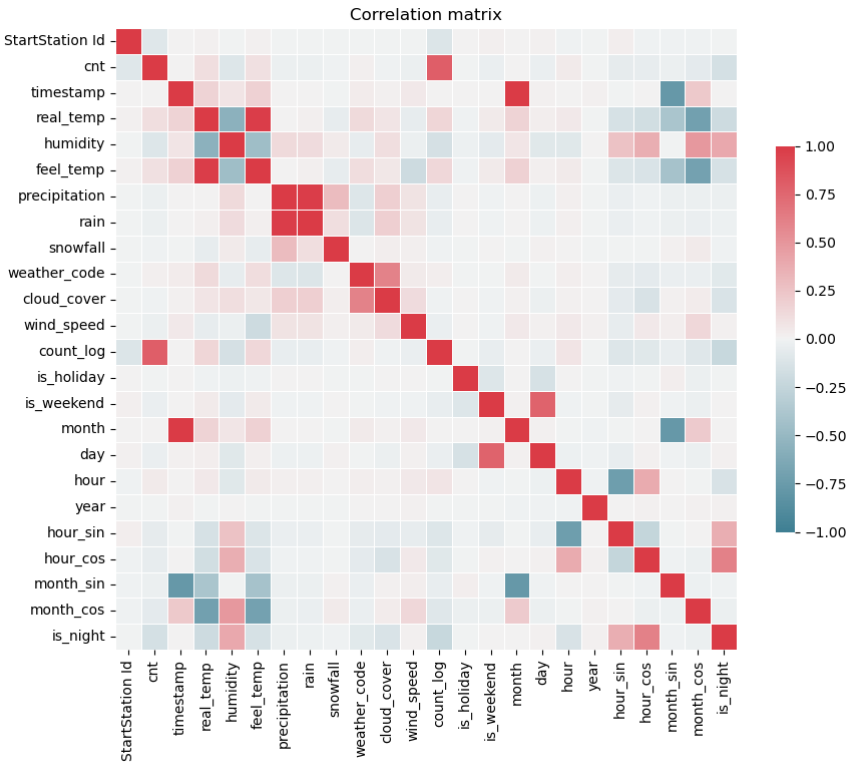

Figure 3.6: Comprehensive Correlation Heatmap of Bike-Sharing Variables

4

XGBoost Modelling and Results

The choice of models was based on their ability to handle regression tasks. Linear regression was chosen for its simplicity and interpretability. XGBoost was chosen for its robustness and efficiency in dealing with large datasets and complex feature spaces [8]. The Asymmetric XGBoost model was included specifically to investigate the effects of an asymmetric loss function, where over- and under-predictions are penalized differently, reflecting their different impacts on the business context of bike sharing schemes.

The evaluation of the models showed clear differences in performance. Linear regression served as a baseline, providing a benchmark while being rather fast and simple to implement. XGBoost improved on this baseline by handling non-linear relationships more effectively, with a Root Mean Squared Logarithmic Error (RMSLE) of 0.485 compared to RMSLE of 0.5494 for the Linear Regression. The Asymmetric XGBoost, with its custom loss function, scored a RMSLE of 0.7318. Although its RMSLE was worse than that of the Linear Regression, it still provided a useful insight into how different penalties for over- and under-prediction could affect model performance. The results from figure 4.1 show that XGBoost is the dominant model for default settings.

In addition to training with the default settings, the models were then tuned and evaluated with the test set. The Grid Search Cross-Validation - GridSearchCV - function was performed on these 3 algorithms. GridSearchCV is a method for tuning the parameters of

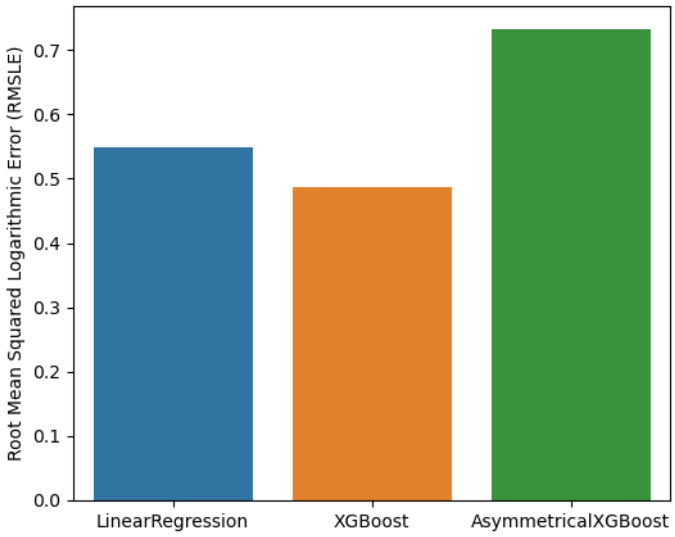

Figure 4.1: Comparative Analysis of Machine Learning Model Accuracy by RMSLE

a model to find its best possible configuration. It tests all combinations of the given hyperparameters to find the combination that produces the best model performance, based on a specific evaluation metric, which in this research is the Root Mean Squared Error - RMSE.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

A custom scoring function is implemented - root mean squared logarithmic error - RMSLE. The Root Mean Squared Logarithmic Error (RMSLE) is defined by modifying Scikit-Learn's [27] mean_squared_log_error function, which is itself a modification of the familiar Mean Squared Error (MSE) metric.

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(a_i + 1))^2}$$

Where:

n *is the total number of observations in the dataset*

p_i *is the prediction of the target*

a_i *is the actual target for* i.

The RMSLE metric is chosen because the dataset contains both very large and very small values. With this metric, any outliers will have a smaller effect on the scoring and the predictions are evaluated with percentage error. Another effect of this metric is that it adds an extra slight penalty to underestimates.

TimeSeriesSplit was used to split the data. TSS is a type of cross-validation specifically designed for time series data. Unlike standard cross-validation methods, which randomly shuffle the data, TimeSeriesSplit preserves the temporal order of the observations. This is crucial for time series analysis because the prediction for a given time may depend on previous times, making random shuffling and splitting inappropriate.

For each combination of parameters given to GridSearchCV, the model is trained and validated five times, each time using a different segment of the data as described in the TimeSeriesSplit approach. This ensures that the sequential nature of the data is respected, which is crucial for time series forecasting to avoid lookahead bias. Lookahead bias in machine learning occurs when a model unintentionally incorporates information that would not have been available at the time of prediction in a real-world scenario. [28]

To save time on future use of the algorithms, the best estimators are exported to individual files using the Python Pickle package. The cross validation results were also saved to CSV files.

4.1 Cross Validation Analysis

Table 4.1 shows the top 3 hyperparameter combinations, their RMSLE score and training time of every algorithm. The analysis shows that Asymmetrical XGB exhibits slightly more variation in performance than standard XGBoost, particularly as the number of estimators decreases. This might indicate a higher sensitivity of Asymmetrical XGB to changes in model complexity. The RMSLE values are consistently higher for Asymmetrical XGBoost, which suggests that when errors occur, they are more significant. The table illustrates a trade-off between achieving higher accuracy and maintaining consistency in performance across different configurations. Asymmetrical XGBoost shows potential benefits in handling specific data characteristics, but at the cost of increased error magnitudes when predictions are inaccurate.

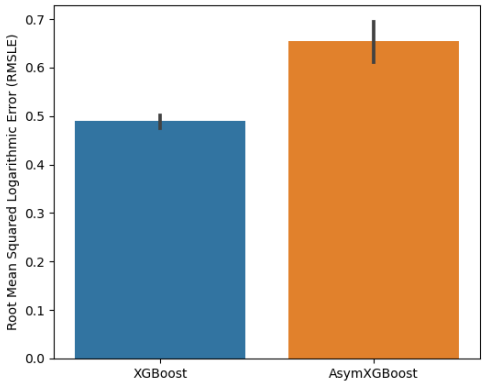

Figure 4.2: Comparison of Prediction Accuracy between XGBoost and AsymXGBoost Models Using RMSLE

Rank	Learning Rate	Max Depth	N Estimators	Mean Fit Time		Mean Test Score		Std Test Score		RMSLE	
				Asym XGB	XGB	Asym XGB	XGB	Asym XGB	XGB	Asym XGB	XGB
1	0.1	10	1200	704.83	163.97	-0.6546	-0.4901	0.0466	0.0140	0.581	0.465
2	0.1	10	1000	589.87	190.09	-0.6569	-0.4911	0.0488	0.0134	0.632	0.485
3	0.1	10	700	416.69	133.24	-0.6634	-0.4915	0.0500	0.0143	0.648	0.505

Table 4.1: Cross-validation Results for the top 3 Ranked Models

4.2 Comparative Analysis of Model Generalization and Robustness

By analyzing the results from table 4.2 we immediately notice that for the XGBRegressor model, the root mean square error (RMSE) decreased from 3.71 on the training set to 3.45 on the testing set. Furthermore, for the Asymmetric XGBRegressor, the RMSE decreased from 4.67 to 4.21. These results suggest that both models are capable of generalizing beyond the training data. However, the reduction is more significant in the case of XGBRegressor. This reinforces its effectiveness and robustness as a predictive tool. This demonstrates that while Asymmetric XGBRegressor might be specialized for certain types of datasets, XGBRegressor provides more stable and reliable predictions across different datasets, making it a preferable choice in scenarios where robustness and reliability are critical.

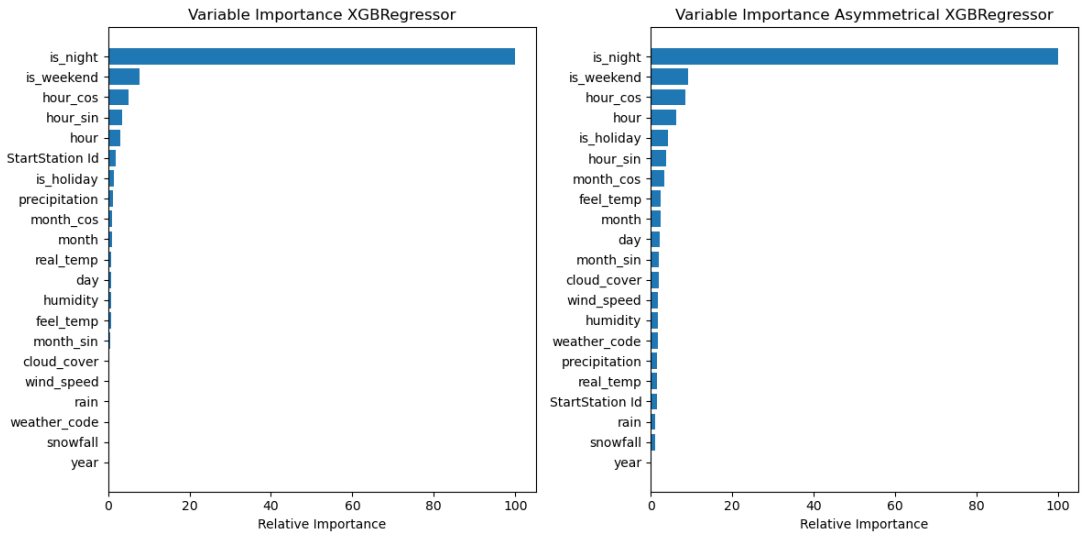

Figure 4.3: Comparison of Variable Importance for XGBRegressor vs Asymmetrical XGBRegressor

Metric	XGBRegressor		Asymmetric XGBRegressor	
	Train Set	Test Set	Train Set	Test Set
Root Mean Squared Error	3.7177	3.4596	4.6753	4.2119
Mean Absolute Error	1.7204	1.6043	2.1199	1.9128
R^2 Coefficient of Determination	0.2662	0.2289	-0.1604	-0.1429
RMSLE	0.4618	0.4606	0.7056	0.6557

Table 4.2: Comprehensive Model Performance Evaluation for Training and Testing Sets

4.3 Predictive Performance Assessment

Both models show admirable predictive efficiency, particularly in areas where data points are densely packed. This reflects a strong alignment with the underlying patterns for a significant proportion of the dataset, which is a positive indication of the models' effectiveness in these regions. Although the predictions do not always match the true values, they show an impressive range, suggesting that the models are robust to different scenarios.

Despite the seemingly lower accuracy of the Asymmetric XGBoost model in terms of the metrics presented in Table 4.2, it may still be valuable in practical applications. For example, a bike-sharing company may prefer a model that slightly overestimates demand to

ensure sufficient supply, rather than underestimating it, which could result in a shortage of bikes and lost revenue. The Asymmetric XGBoost model has a mean absolute error (MAE) of 1.91, indicating that its overestimation falls within an acceptable range for this context. Being off by approximately two bikes, it is likely a tolerable error margin for operational planning. An evaluation of this nature is crucial in determining the appropriate model for deployment in a real-world scenario. It highlights the significance of aligning model selection with strategic objectives and operational tolerances.

By examining figures 4.4 & 4.5 we see a large discrepancy between the ground truth values and the predicted values by the models. Given that the test set comprises nearly one million entries, such disparities raise concerns regarding the model's accuracy. To further investigate, a focused analysis on a smaller subset of 800 entries was conducted to determine whether the observed discrepancies were outliers or indicative of a broader issue. The analysis of these additional plots reveals that the model generally predicts values accurately. This is evidenced by the more aligned peaks and troughs between the true and predicted values in the smaller subset. The Mean Absolute Error (MAE) of 1.6, observed across the larger dataset, appears justified based on this targeted examination.

Furthermore, D'Agostino's K2ˆ Test was performed to check the nature of the distribution of residuals [29]. By observing plots in figure 4.8, the distribution of the residuals appears to be Gaussian. However, D'Agostino's test fails and the hypothesis is rejected with a p-value of 0.

The scatter plot of predicted values versus residuals shows a pattern where the residuals appear to fan out as the predicted values increase. This pattern could indicate that the variance of the residuals is not constant, otherwise known as heteroscedasticity [30]. The histogram of the residuals suggests a distribution with a slight right skew, as indicated by a skewness of 0.4747. The peak is close to zero and most of the data seem to cluster around the mean residual, which is zero. The kurtosis [31] is 0.47, indicating a distribution that is slightly flatter than normal, but not extremely so. The Q-Q plot shows that the quantiles of the residuals deviate from the theoretical normal distribution in the tails. Specifically, the upper tail is heavier, indicating the presence of outliers that are higher than what a normal distribution would predict [32].

There appear to be more pronounced outliers for the asymmetric model. The histogram for these residuals has a more pronounced peak and a stronger right skew, with a skewness value of 0.8821, and a higher kurtosis of 1.12, which indicates a more peaked dis-

tribution with heavier tails when compared to the XGBoost.

In general, these plots show that while the model has an average residual close to zero and a reasonable standard deviation, there are signs of non-normality in the residuals' distribution, particularly with right skewness and heavy tails. These findings suggest that while the model may predict the average trend well, as indicated by a low MAE, there are systematic patterns in the prediction errors that could potentially be improved upon.

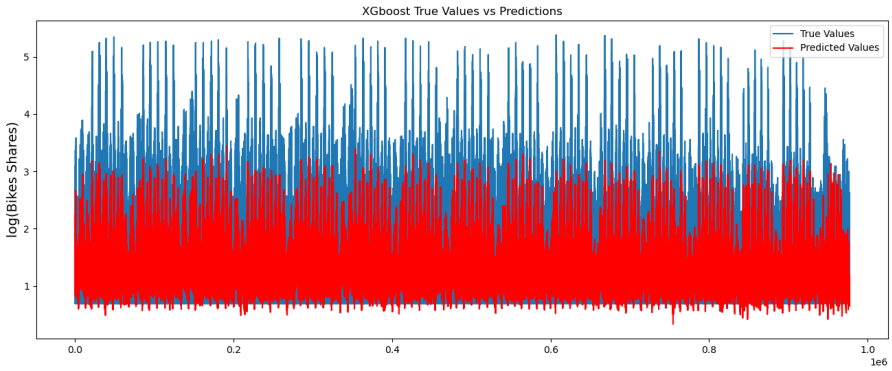

Figure 4.4: True vs Predicted Values with XGBoost Model

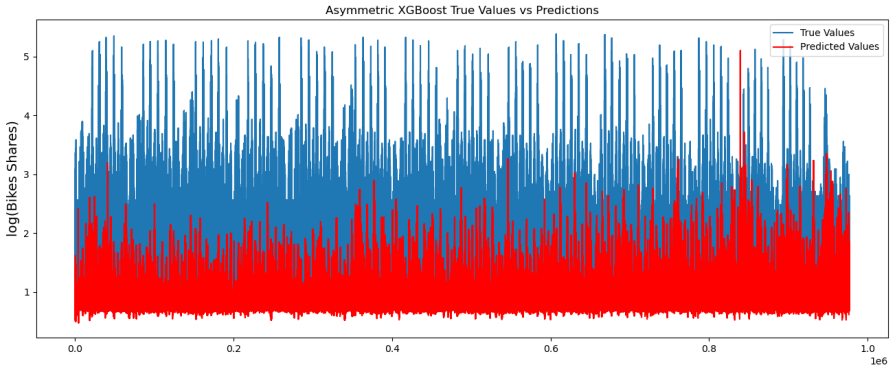

Figure 4.5: True vs Predicted Values with Asymmetric XGBoost Model

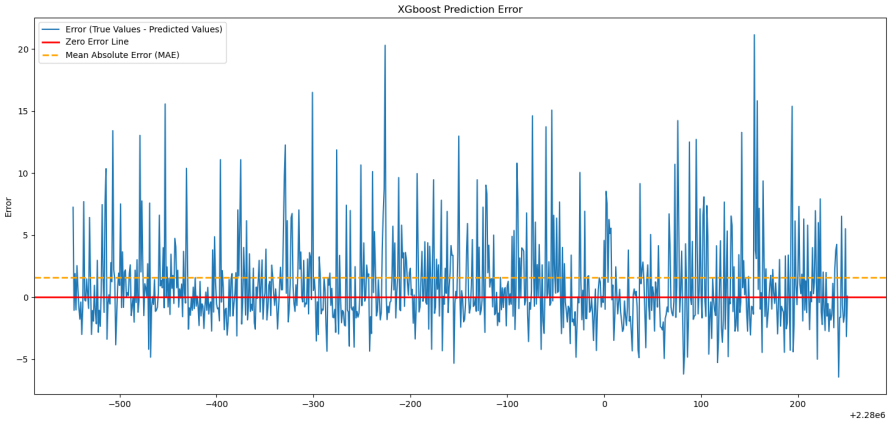

Figure 4.6: XGBoost Linear Prediction Error over 800 Samples

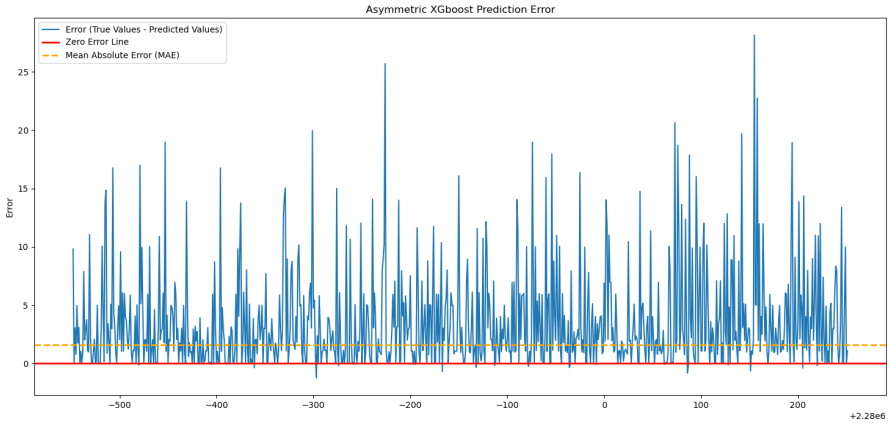

Figure 4.7: Asymmetric XGBoost Linear Prediction Error over 800 Samples

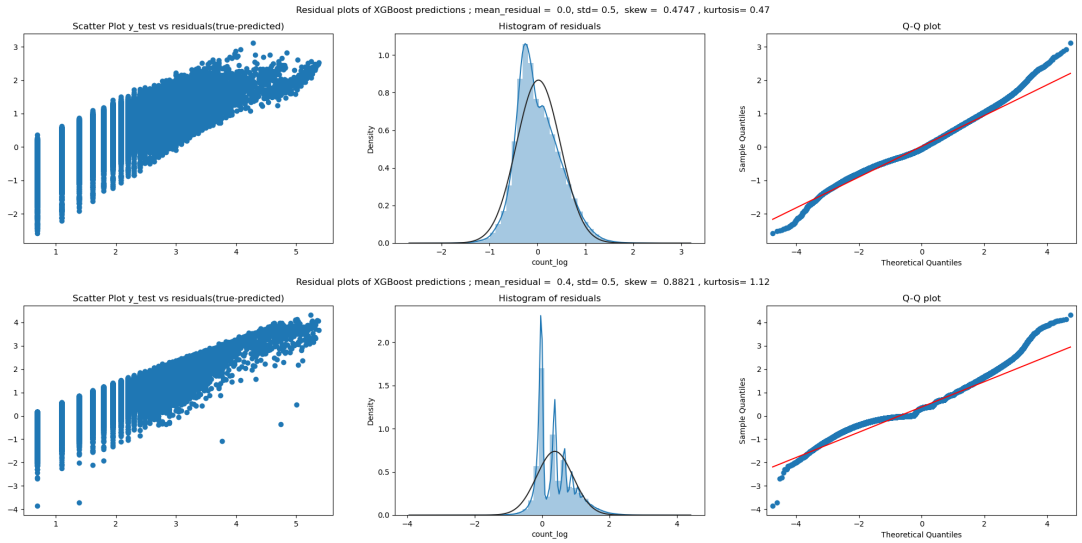

Figure 4.8: Comparative Residual Analysis for XGBoost and AsymXGBoost Models

5 Conclusion

This thesis has presented a comprehensive investigation into predictive modelling for bike-sharing systems, with a particular focus on the bike-sharing infrastructure in London. The research utilised a range of data sources, including trip data, weather conditions and urban infrastructure metrics, to develop robust predictive models using advanced machine learning techniques such as XGBoost and custom XGBoost with an asymmetric loss function.

The literature review highlighted the importance of integrating different datasets to accurately predict bike rental demand, highlighting factors such as weather conditions, time variables and station characteristics. This was supported by the data pre-processing and analysis phases, which showed that factors such as temperature and rainfall have a significant impact on bike-sharing usage patterns.

The models developed during this research demonstrated that XGBoost provided superior predictive accuracy over linear regression, effectively handling complex non-linear relationships. The Asymmetric XGBoost model introduced a novel approach by incorporating an asymmetric loss function, which proved particularly insightful in understanding the different effects of over- and under-prediction in a bike-sharing context.

5.1 Key findings from the modelling efforts

1. Weather conditions play a critical role, with higher temperatures increasing bike rental demand and rainfall decreasing it.
2. Temporal patterns, such as time of day and whether a day is a working day, weekend or holiday, have a significant impact on demand.
3. The spatial distribution of stations and their capacity are critical to balancing supply and demand across the network.

The practical implications of these findings are significant for urban planners and policy makers seeking to improve the efficiency and user satisfaction of bike sharing systems. By understanding the key factors influencing demand, strategies can be developed to optimize station placement, adjust bike allocation and improve overall service reliability under different weather conditions and times of day.

5.2 Future research directions

Future studies could extend this work by exploring the integration of a larger quantity of data, or even more granular weather data, allowing for dynamic adjustments in bike sharing operations. Additionally, experimenting with other forms of asymmetric penalty functions could yield improvements in model performance tailored to specific business needs. Further research could examine the long-term effects of urban development and changes in public transport infrastructure on bike-sharing patterns. Finally, exploring machine learning models that incorporate predictions of user behaviour based on demographic and psychographic data could provide deeper insights into demand fluctuations and user preferences in bike-sharing systems.

In conclusion, this thesis provides valuable insights into the dynamics of bike-sharing systems and lays a solid foundation for further research that could lead to more sustainable and user-friendly urban transport solutions.

References

[1] E. Fishman, S. Washington, and N. Haworth, "Bike Share's Impact on Car Use: Evidence from the United States, Great Britain, and Australia," *Transportation Research Part D: Transport and Environment*, vol. 31, pp. 13–20, 2014.

[2] S. A. Shaheen, S. Guzman, and H. Zhang, "Bikesharing in Europe, the Americas, and Asia: Past, Present, and Future," *Transportation Research Record*, vol. 2143, no. 1, pp. 159–167, 2010.

[3] Y. Zhang, T. Thomas, M. Brussel, and M. Van Maarseveen, "Expanding Bicycle-sharing Systems: Lessons Learnt from an Analysis of Usage," *PLoS one*, vol. 11, no. 12, p. e0168604, 2016.

[4] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A Tale of Many Cities: Universal Patterns in Human Urban Mobility," *PLOS ONE*, vol. 7, no. 5, pp. 1–10, 05 2012. [Online]. Available: https://doi.org/10.1371/journal.pone.0037027

[5] A. Ogunleye and Q.-G. Wang, "XGBoost Model for Chronic Kidney Disease Diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 6, pp. 2131–2140, 2019.

[6] W. Li, Y. Yin, X. Quan, and H. Zhang, "Gene Expression Value Prediction Based on XGBoost Algorithm," *Frontiers in Genetics*, vol. 10, p. 484931, 2019.

[7] S. Ben Jabeur, N. Stef, and P. Carmona, "Bankruptcy Prediction Using the XGBoost Algorithm and Variable Importance Feature Engineering," *Computational Economics*, vol. 61, no. 2, pp. 715–741, 2023.

[8] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[9] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart Cities of the Future," *The European Physical Journal Special Topics*, vol. 214, pp. 481–518, 2012.

[10] P. Borgnat, P. Abry, P. Flandrin, and J.-B. Rouquier, "Studying Lyon's Vélo'V: A Statistical Cyclic Model," in *ECCS'09*, University of Warwick. Warwick, United Kingdom: Complex System Society, Sep. 2009. [Online]. Available: https://ens-lyon.hal.science/ensl-00408147

[11] Weather Spark, "Historical Weather Spring 2008 in Lyon, France," https://weatherspark.com/h/s/50604/2008/0/Historical-Weather-Spring-2008-in-Lyon-France, 2008, accessed: 2024-04-10.

[12] W. El-Assi, M. Salah Mahmoud, and K. Nurul Habib, "Effects of Built Environment and Weather on Bike Sharing Demand: A Station Level Analysis of Commercial Bike Sharing in Toronto," *Transportation*, vol. 44, pp. 589–613, 2017.

[13] K. Kim, "Investigation on the Effects of Weather and Calendar Events on Bike-Sharing According to the Trip Patterns of Bike Rentals of Stations," *Journal of Transport Geography*, vol. 66, pp. 309–320, 2018.

[14] N. A. Weinreich, D. B. van Diepen, F. Chiariotti, and C. Biscio, "Automatic Bike Sharing System Planning from Urban Environment Features," *Transportmetrica B: Transport Dynamics*, vol. 11, no. 1, p. 2226347, 2023.

[15] "A Review on Bike-sharing: The factors Affecting Bike-sharing Demand, author=Eren, Ezgi and Uz, Volkan Emre," *Sustainable cities and society*, vol. 54, p. 101882, 2020.

[16] W. Jiang, "Bike Sharing Usage Prediction with Deep Learning: A Survey," *Neural Computing and Applications*, vol. 34, no. 18, pp. 15 369–15 385, 2022.

[17] F. Giannakas, C. Troussas, A. Krouska, C. Sgouropoulou, and I. Voyiatzis, "Xgboost and Deep Neural Network Comparison: The case of Teams' Performance," in *Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings 17*. Springer, 2021, pp. 343–349.

[18] S. S. Azmi and S. Baliga, "An Overview of Boosting Decision Tree Algorithms Utilizing AdaBoost and XGBoost Boosting Strategies," *Int. Res. J. Eng. Technol*, vol. 7, no. 5, pp. 6867–6870, 2020.

[19] Transport for London, "Transport for London," https://tfl.gov.uk/, accessed: 2024-04-11.

[20] T. for London, "TfL Cycling Data," https://cycling.data.tfl.gov.uk/, accessed: 2024-04-11.

[21] "Open-Meteo," https://open-meteo.com/, accessed: 2024-04-11.

[22] Transport for London, "Live cycle hire updates," https://tfl.gov.uk/tfl/syndication/feeds/cycle-hire/livecyclehireupdates.xml, accessed: 2024-04-11.

[23] "Folium Documentation," https://python-visualization.github.io/folium/latest/#, accessed: 2024-04-11.

[24] J. Kaedbey, "Smartcitybikedemandml," https://github.com/JadKaedBey/SmartCityBikeDemandML, May 2024, code Repository for Predicting Bike-Sharing Demand in Smart Cities Using Machine Learning.

[25] Kevin. (2014) Drawing Boundaries In Python. Accessed: 2024-04-11. [Online]. Available: https://thehumangeo.wordpress.com/2014/05/12/drawing-boundaries-in-python/

[26] P. Wang, M. Guo, Y. Han, L. Zhao, X. Zhou, and D. Zhang, "Ensemble Learning-based Hierarchical Retrieval of Similar Cases for Site Planning," *Journal of Computational Design and Engineering*, vol. 8, no. 6, pp. 1548–1561, 2021.

[27] "scikit-learn: Machine Learning in Python," https://scikit-learn.org/stable/, accessed: 2024-04-14.

[28] A. Matuozzo, P. Yoo, A. Provetti, and M. Kim, "Machine Learning Methods for Equity Time Series Forecasting: A Compendium," 2022.

[29] R. B. D'Agostino, "Tests for the Normal Distribution," in *Goodness-of-fit-techniques*. Routledge, 2017, pp. 367–420.

[30] Q. M. Zhou, P. X.-K. Song, and M. E. Thompson, "Profiling Heteroscedasticity in Linear Regression Models," *Canadian journal of statistics*, vol. 43, no. 3, pp. 358–377, 2015.

[31] L. T. DeCarlo, "On the Meaning and Use of Kurtosis." *Psychological methods*, vol. 2, no. 3, p. 292, 1997.

[32] J. I. Marden, "Positions and QQ plots," *Statistical Science*, pp. 606–614, 2004.

Acknowledgments

I am immensely grateful to my family, whose unwavering support has been my cornerstone throughout the journey of completing this degree. To my mother, whose courage continually inspires me to pursue my passions fearlessly. To my father, the wisest and most exemplary figure in my life, whose guidance has always steered me towards the right path. To my sister, thank you for the relentless support and belief in my abilities, which has been a great source of motivation. To little Joey, who brought immense happiness to our family.

A heartfelt thank you goes to my partner, Emilija, who has been a constant source of support and happiness throughout my academic journey. Her presence and encouragement have been invaluable in every step of this process.

I extend my gratitude to my supervisor, Federico Chiariotti, for the valuable insights and tips that have subtly yet significantly shaped the direction and execution of this research.

Lastly, I would like to express my appreciation to the Università degli Studi di Padova for providing an environment conducive to learning and growth. The opportunities and resources available have been fundamental in the completion of this degree.