UCL DEPARTMENT OF GEOGRAPHY

YEAR 2023-24

EXAM CANDIDATE ID:	GJFC7
MODULE CODE:	GEOG0178
MODULE NAME:	Machine Learning for Social Sciences with Python 23/24
COURSE PAPER TITLE:	London's Bike-share Demand Prediction Using Machine Learning (ML) Algorithms
WORD COUNT:	1993

Your essay, appropriately anonymised, may be used to help future students prepare for assessment. Double click this box to opt out of this □

London’s Santander Bike-share Demand Prediction Using Machine Learning (ML) Algorithms

1. Introduction

Bike-sharing has garnered considerable interest in recent years for its role in enhancing first/last-mile connectivity and reducing the environmental footprint of transportation (Fishman, 2016; Ricci, 2015; Shaheen et al., 2010). These benefits made it a widely embraced strategy within urban and transportation planning frameworks to advance sustainability objectives and encourage eco-friendly commuting practices.

As bike-sharing continues to gain traction, accurately predicting bike usage becomes imperative. London, where bike-sharing has become an integral component of the transportation landscape, presents an excellent case study for this task. The Barclays Cycle Hire scheme was launched in 2010 jointly by TfL and Barclays Bank to increase cycling rates and provide a convenient alternative to traditional modes of transportation (TfL, 2010). The scheme was later rebranded as Santander Cycles under Santander Bank's stewardship in 2016. Since its inception, the bike-sharing system has witnessed a steady rise in usage, surpassing 10 million trips annually by 2019 (TfL, 2020).

This research aims to develop a machine learning (ML) algorithm that accurately predicts the hourly bike-share demands using temporal and weather-related features, thereby enabling more effective management of London’s bike-share systems.

2. Literature Review

The growing popularity of bike-sharing systems sparked substantial research interest, focusing on demand forecasting (Sathishkumar et al., 2020), station-planning/network design (Faghih-Imani et al., 2014; Gu et al., 2019), and their policy implications (Ricci, 2015). Among these, bike demand prediction forms the backbone of inventory management, essential for the efficient operation and administration of bike-share systems (Ghosh et al., 2017). Central to this prediction task include capturing spatial dynamics among stations, temporal patterns, and environmental and social factors (Ashqar et al., 2020; Faghih-Imani, 2014; Lin et al., 2018; Tomaras et al., 2018; Wang, 2016).

Due to the nonlinear nature of spatio-temporal features, bike prediction most commonly uses algorithms like Decision Tree (DT), Random Forest (RF), and Gradient Boost Models (GBM). Sathishkumar et al. (2020) compared various ML models to predict hourly bike-share demand in Seoul using temporal and weather-related features. GBM was found to offer the best model, while temperature and hour are the most important predictors. Ashqar et al.'s (2020) study on network and station-level bike-sharing systems in San Francisco employed RF with least-squares boosting to predict station-level bike-share usage. Kaggle's 2014 Bike-Sharing Demand Prediction Competition also reported XGBoost as one of the most powerful algorithms for transport-related demand prediction (Kaggle, 2015). In general, various research demonstrates the superiority of nonlinear models over linear techniques in capturing the complex relationships between influencing factors and bike-share demand. More recently, Lin et al. (2018) suggested Graph Convolutional Neural Network, a deep learning technique, for bike-share demand prediction. The proposed model is successful in learning hidden pairwise heterogeneous correlations between stations to predict station-level hourly demand, but requires a significantly higher computation cost for implementation.

It's also worth noting that earlier works relied more on linear models, including Autoregressive Moving Average (ARMA) (Katenbrunner et al., 2010). Faghih-Imani et al. (2014) employed a linear-mixed model to predict bike-share demand in Montreal, using spatial features on top of temporal and weather features. Their methodology includes dividing the time of day into morning, midday, afternoon and evening, thereby handling nonlinear temporal features in linear models. Tran et al. (2015) navigated this problem using linear models, but only used built-environment features.

This research draws on the literature review to build a robust ML algorithm to predict London's bike-share demand, a research gap amongst existing literature.

3. Methodology

3.1 Data

The dataset of research is created from bike-share, weather and UK holiday data, obtained from the TfL open data, freemeteo.com, and gov.uk respectively. The raw bike-share data is grouped by hour and merged with the weather data. The few ‘na’ values are removed. The cleaned dataset contains hourly data spanning across two years from 2015 and 2016 (17414 rows). Parameters are shown in Table 1.

Table 1: Parameters of the cleaned dataset

Parameter	Abbreviation	Type	Measurement
Timestamp	Timestamp	Datetime	-
Bike Count	cnt	Continuous	0,1,2…
Real Temperature	t1	Continuous	◦C
Feel Temperature	t2	Continuous	◦C
Humidity	hum	Continuous	%
Wind Speed	wind_speed	Continuous	km/hr
Weather Code[1]	weather_code	Categorical	Code number
Holiday	is_holiday	Categorical	0,1
Weekend	is_weekend	Categorical	0,1
Season	season	Categorical	1,2,3,4

Subsequently, the dataset is chronologically split into train and test datasets (80:20 ratio), meaning no shuffling in the data to avoid data leakage (Hyndman and Athanasopoulos, 2018), and for the data to be trained and evaluated based on the true temporal dependence structure of the time series.

3.2 Models

In this research, tree-based models are preferred as they can capture the nonlinear patterns within temporal and weather-related features. The study first employs DT as a base model for its high interpretability (Hastie et al., 2009), and subsequently uses eXtreme Gradient Boosting (XGBoost) to evaluate potential improvements to prediction accuracy brought by ensemble learning and gradient descent through the algorithm.

[1] Refer to Appendix A for detailed weather code

3.2.1 Decision Tree (DT)

DT is a popular classification algorithm widely used in many fields. The hierarchical nature of DTs makes them highly interpretable, and the model's decision-making process can be easily understood through a series of if-then-else rules (Quinlan, 1986). The algorithm’s high interpretability, while also being able to capture nonlinear relationships for both numerical and categorical variables, makes the model suitable as a baseline model for this research.

3.2.2 XGBoost

XGBoost, developed by Chen and Guestrin (2016), has become increasingly well-known for its speed, accuracy, and scalability. It’s an implementation of gradient boosting and an ensemble method combining multiple weak models (e.g. decision trees) to capture the complex, non-linear relationships within the input features. These mechanisms enable XGBoost to be potentially more robust and efficient than DT at bike-share demand predictions.

These models are optimised through feature engineering and hyperparameter tuning to obtain the best result.

4. Interpretation and discussion of results

4.1 Exploratory data analysis (EDA) and feature engineering

The exploratory data analysis is carried out to summarise and visualise the relationships between different variables.

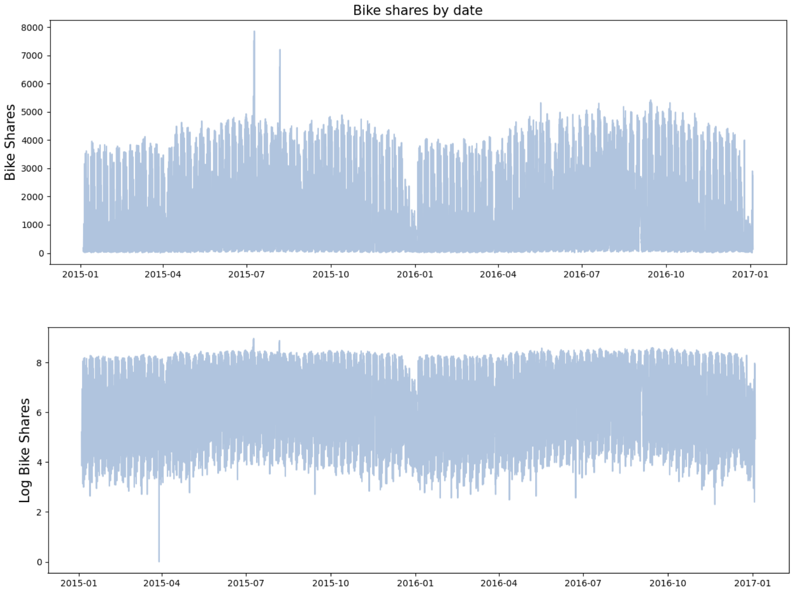

Figure 1: Bike shares and log bike shares over time.

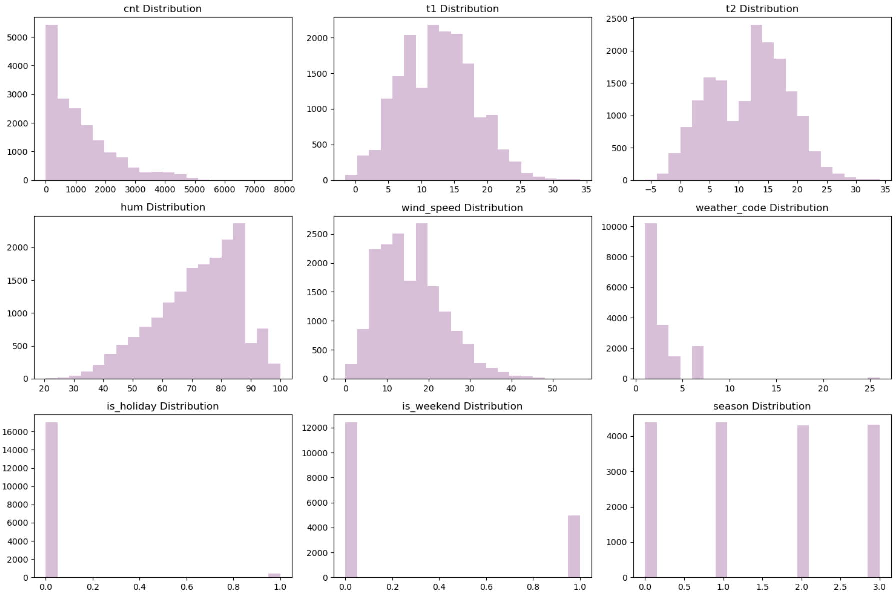

Figure 2 : Distribution of dataset features.

Figure 1 highlights multiple seasonal patterns of bike-share usage. Cross-referring to metadata, the extremely high/low values of bike count are not systematic errors but rather random occurrences. Hence, the values are kept for model training.

The bike-share count is also observably skewed to the left (Figure 2 top-left), resembling an exponential-logarithmic distribution. While tree-based models don't strictly require normal distribution of variables, log transformations are shown by literature review to improve model performances (Wang, 2016; Zheng et al., 2019). Hence, the bike-share count variable is transformed using 'numpy.log1p' function. Other variables were also tested using the skew function, all of which indicated that transformations are not necessary.

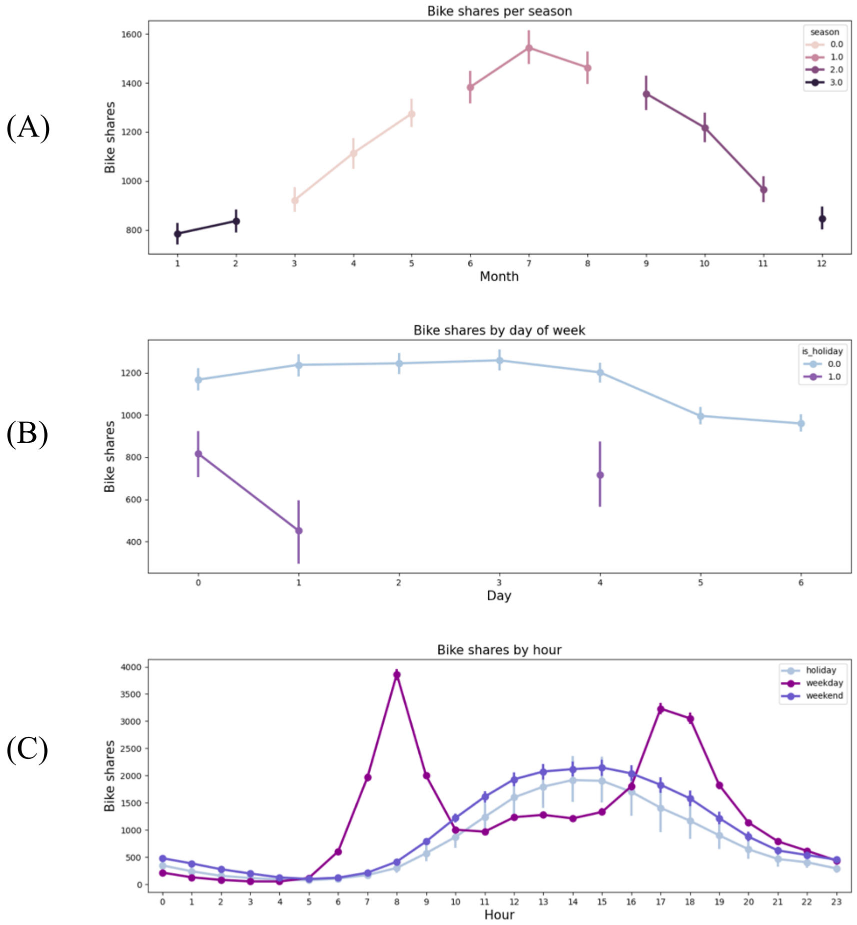

Figure 3: Relationship between bike-share count and temporal features.

Generally, demand tends to be higher in summer (Figure 3A), and during peak hours (Figure 3C). The distribution of demand is similar across workdays (Figure 3B).

From Figure 2 (bottom-left), the 'is_holiday' value counts are highly imbalanced and may cause higher inaccuracy for tree-based models. As bike-share demand on holidays and weekends exhibits similar trends (Figure 3C), these two columns are combined into the 'is_workday' feature.

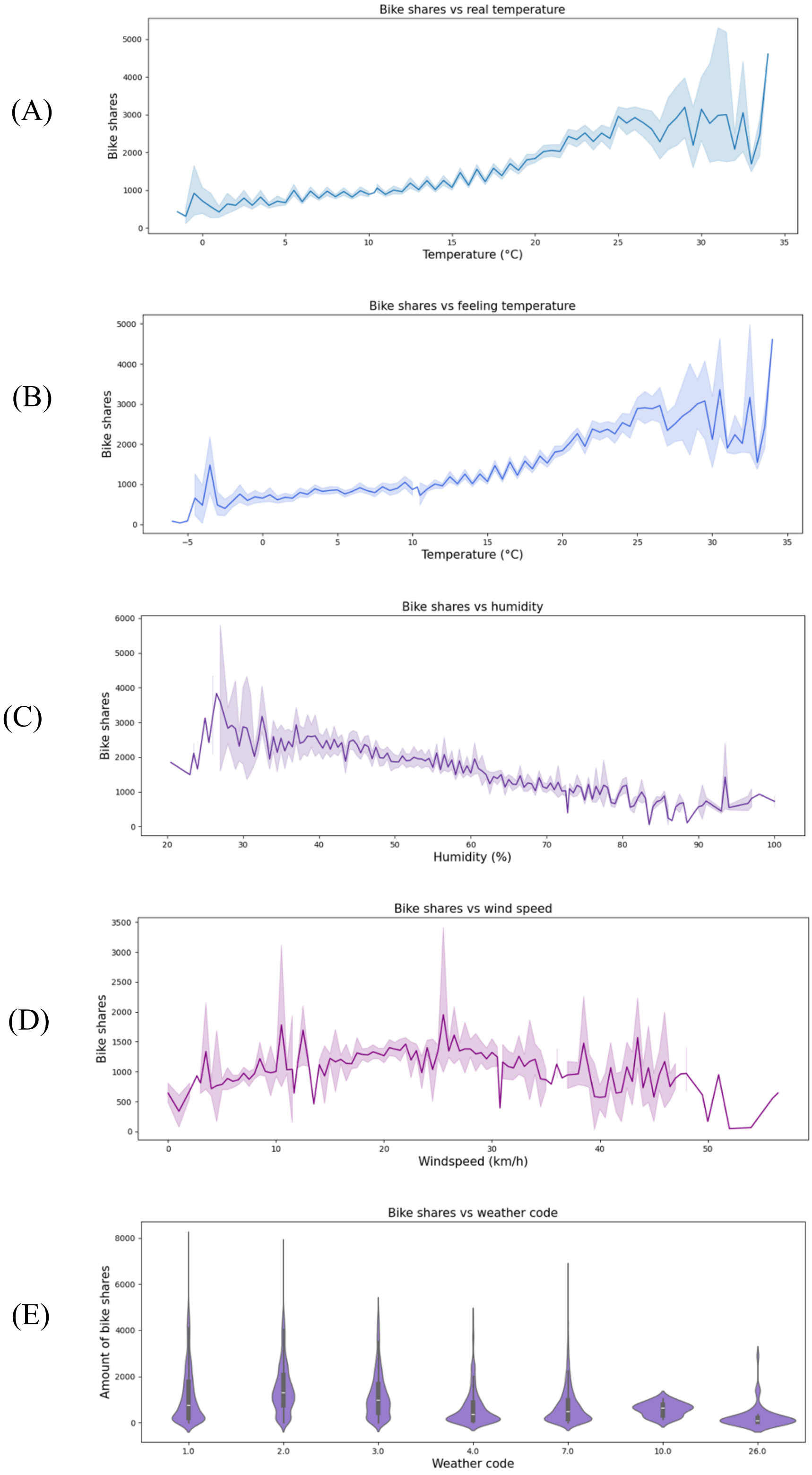

Figure 4: Relationship between bike-share count and weather-related features

For weather-related features, the real and feeling temperatures (which are highly correlated (*r=0.988)*), show similar positive trends with bike-share demands (Figure 4A-B). Only feeling temperature is retained for model training, as it intuitively influences commuters' cycling decisions more. Humidity shows a negative relationship with bike-share demand (Figure 4C), while bad weather appears to with decreased demand (Figure 4E). The impact of wind speed on bike-share demand appears inconclusive (Figure 4D).

4.2 Model Building

Based on EDA and feature engineering, the first set of selected features is shown in Table 2. The hyperparameters for DT and XGB Regressors are tuned using GridSearchCV to find the best parameters that optimise model performances (Table 3).

The determination coefficient (R^2) is employed to understand goodness of fit, while root mean squared error (RMSE) is used to measure the scale of error, penalising large errors more heavily (Hastie, 2009). The MAE is also calculated to understand the prediction errors in actual scale.

Table 2: First set of selected features

Parameter	Abreviation	Type	Measurement
Feel Temperature	t2	Continuous	◦C
Humidity	hum	Continuous	%
Wind Speed	wind_speed	Continuous	m/s
Hour	hour	Continuous	0, 1, …, 23
Month	month	Continuous	1, 2, …, 12
Weather Code	weather_code	Categorical	Code number
Workday	is_workday	Categorical	0, 1
Season	season	Categorical	1, 2, 3, 4

Table 3: Best hyperparameters for models

DT Regressor 1	'max_depth': 12, 'min_samples_leaf': 10, 'min_samples_split': 7
XGB Regressor 1	'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 150
DT Regressor 2	'max_depth': 12, 'min_samples_leaf': 7, 'min_samples_split': 10
XGB Regressor 2	'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 150

Table 4: Model Performances

Feature Set	Model	RMSE	MAE (in actual scale)	R^2
1	DT 1 Training	0.310	-	0.939
	DT 1 Testing	0.332	215	0.932
	XGB 1 Training	0.271	-	0.954
	XGB 1 Testing	0.300	192	0.944
2	DT 2 Training	0.306	-	0.941
	DT 2 Testing	0.322	206	0.935
	XGB 2 Training	**0.274**	**-**	**0.952**
	XGB 2 Testing	**0.299**	**192**	**0.945**

*** Best model is XGB 2 (bolded)*

Results indicate that XGB does better than DT with an R^2 value of 0.944 vs 0.932, RMSE of 0.300 vs 0.332, and MAE (in actual scale) of 192 vs 215 bikes (Table 4).

For model enhancement, feature importance is evaluated using SHapley Additive exPlanations (SHAP). SHAP is preferred over the built-in feature importance functions of DT/XGB, as it offers finer-grained interpretations accounting for both the individual effects of features and their interactions, thereby providing a more accurate and consistent way of measuring feature importance (Molnar, 2020).

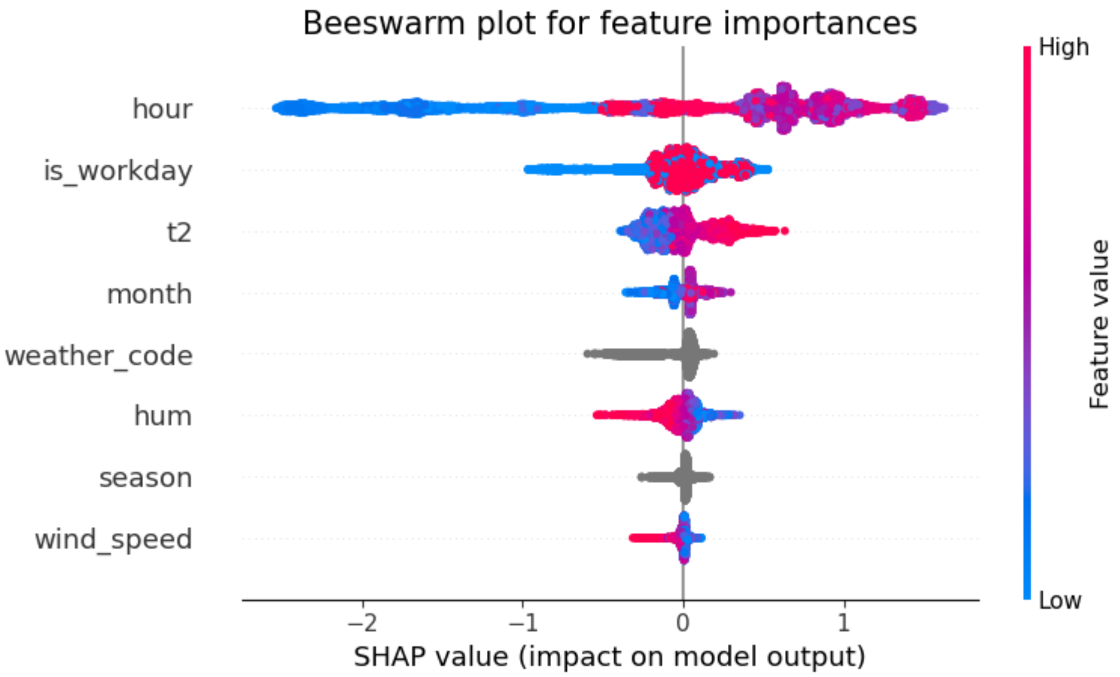

Figure 5: Beeswarm plot for feature importance of the first set of features.

The SHAP analysis shows that ‘wind_speed’ and ‘season’ are the 2 least important predictors (Figure 5). The minimal impact of 'wind_speed' can be attributed to London's high urban density where wind speeds may not deter cyclists, whereas information captured in ‘season’ is somewhat repeated in the ‘month’ variable. These two features are therefore removed in subsequent models to reduce model complexity. Moreover, ‘weather_code’ is simplified to bad weather (‘weather_code’=7,10,26) and otherwise (‘weather_code’=1-4) based on evidence from EDA and SHAP. The combination of ‘temperature’, ‘month’, and ‘bad_weather’ features should, in theory, encompass the information previously captured in ‘weather_code’.

These adjustments reduce model complexity and avoid potential overfitting by focusing on more influential factors driving bike-share demand.

Table 5: Second set of selected features

Parameter	Abreviation	Type	Measurement
Feel Temperature	t2	Continuous	◦C
Humidity	hum	Continuous	%
Hour	hour	Continuous	0, 1, …, 23
Month	month	Continuous	1, 2, …, 12
Bad Weather	bad_weather	Categorical	Code number
Workday	is_workday	Categorical	0, 1

A DT and an XGB model are trained by the second set of selected features (Table 5). These models generally perform equally well as their counterparts in experiment 1 (Table 4), and their simplicity renders them more preferred. Again, XGB’s ensemble learning and gradient descent mechanisms enable the XGB model to perform better than the DT model.

XGB Regressor 2 is chosen as our best model based on its performance metrics and model simplicity. The MAE of this model suggests the prediction of the model, on average, has an error margin of 192 bikes. From Figure 6, the most important features are by far ‘hour’, followed by ‘is_workday’ and feeling temperature.

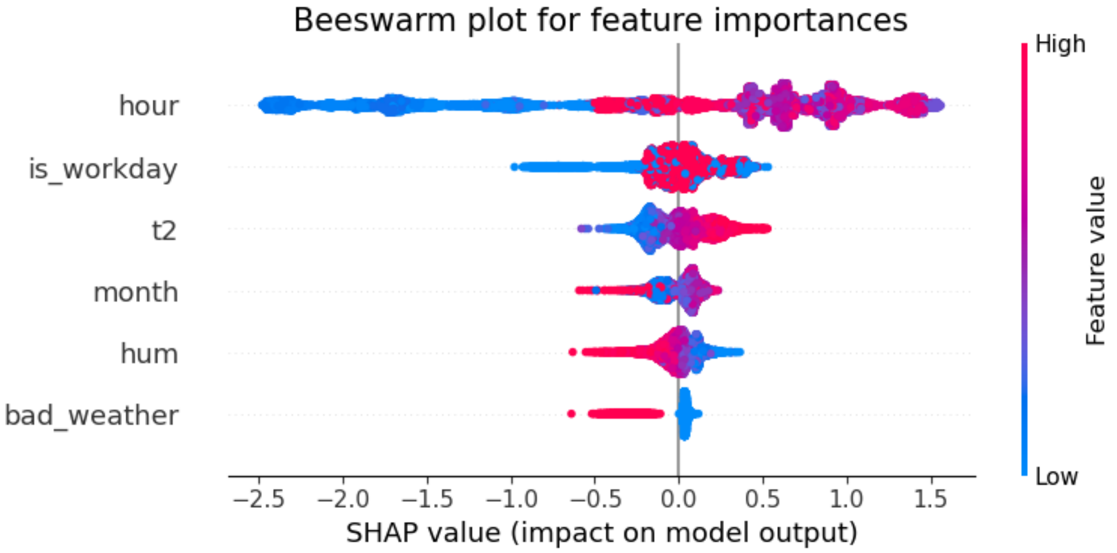

Figure 6: Beeswarm plot for feature importance of the second set of features.

4.3 Model evaluation and limitations

A 5-fold time series cross-validation strategy is applied to the best model, with each fold containing 2 months of hourly data. Unlike standard cross-validation which randomly partitions the dataset, time series cross-validation preserves the chronological order of the data to preserve the temporal dependencies of the data, simulating the real-world scenario where the models would be required to make forecasts on future, unseen data.

The cross-validation results reveal a mean MAE of 176.3 (±24.1), indicating consistent performance across the five test folds. The low standard deviation implies the model's stable performance and its capacity to generalise well to unseen data. It also suggests that the model is not overly sensitive to specific periods or data distributions, further demonstrating the effectiveness and robustness of the chosen model.

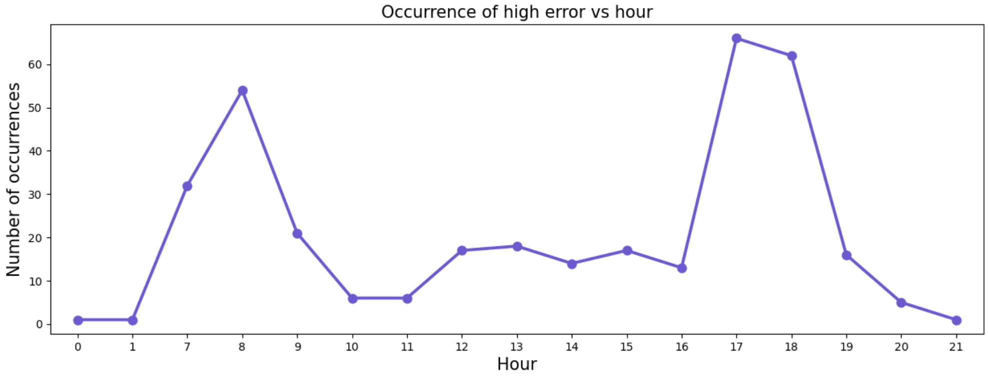

Figure 7: Occurrence of high error term vs hour of the day.

The model's limitations are explored by analysing residuals across various data subgroups. Notably, large residuals cluster mostly during peak hours (Figure 7), indicating the model's tendency to underestimate demand during these periods when bike-share usage is high. This discrepancy is likely because peak hours involve greater unpredictability in demand, suggesting the need to incorporate a larger error term in the chosen model to better accommodate unexpected spikes in demand during peak hours.

The bike-share count data is also tested using the Augmented Dickey-Fuller test to examine its stationarity (Dickey and Fuller, 1979). The obtained p-value is below the critical value, indicating strong evidence of data stationarity. However, periodic reassessment of data stationarity is advisable to detect potential drifts, which could necessitate model retraining.

Another notable limitation of this study is its exclusive focus on city-level bike demand, which restricts its ability to offer insights into station-level forecasting and the dynamics of bike-share demands among neighbouring docking stations.

5. Conclusion

This study developed a robust ML model that predicts bike-share demand in London, utilising temporal and weather-related features. The results indicate that XGBoost models consistently perform better, with the best model achieving an R^2 value of 0.946 and an MAE of 192. The model also indicates that the most important influencers for bike-share demand are 'hour', 'is_workday' and feeling temperature.

With higher granularity data (station-level bike-share count), future research can investigate station-level bike-share demand prediction and account for the interactions of bike-share demands in adjacent docking stations to aid bike-share rebalancing problems.

Additionally, given the autoregressive nature of time series data, it's valuable to include lagged variables in the model to account for temporal dependencies [2](Hamilton, 2020). It's suggested to use a substantial lag value (e.g. one week), which could enhance demand estimation and facilitate improved bike inventory management by providing sufficient lead time.

[2] In this research, it's found that the inclusion of lag 1 value (previous hour's bike count) significantly improves the model. However, larger lagged values, like lag=168 (1 week), lead to decreased model accuracy. It's worth investigating how lagged values can be engineered to enhance the model.

References

Ashqar, H. I., Elhenawy, M., Rakha, H. A., Almannaa, M., & House, L. (2022). Network and station-level bike-sharing system prediction: A San Francisco bay area case study. *Journal of Intelligent Transportation Systems*, *26*(5), pp. 602-612.

Brownlee, J. (2018). Time Series Forecasting with XGBoost in Python. Machine Learning Mastery. [online] Available at: https://machinelearningmastery.com/time-series-forecasting-with-xgboost-in-python/ (Accessed: 15 Apr 2024).

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366a), pp. 427–431.

Fishman, E. (2016). Bikeshare: A review of recent literature. Transport Reviews, 36(1), pp. 92-113.

Faghih-Imani, A., Eluru, N., El-Geneidy, A. M., Rabbat, M., & Haq, U. (2014). How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal. *Journal of transport geography, 41,* pp. 306-314.

Fishman, E., Washington, S., & Haworth, N. (2014). Bike share's impact on car use: Evidence from the United States, Great Britain, and Australia. Transportation Research Part D: Transport and Environment, 31, 13-20.

Freemeteo.com. (2024). Historical Weather Data. [Online] Available at: https://freemeteo.co.uk/ (Accessed: 2 Apr 2024).

Gov.uk. (2024). Bank holidays in the UK. [Online] Available at: https://www.gov.uk/
bank-holidays (Accessed: 2 Apr 2024).

Gu, T., Kim, I., & Currie, G. (2019). To be or not to be dockless: Empirical analysis of dockless bikeshare development in China. *Transportation Research Part A: Policy and Practice*, *119*, pp. 122-147.

Hamilton, J. D. (2020). *Time series analysis*. Princeton university press.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice.* OTexts.

Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., & Banchs, R. (2010). Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, *6*(4), pp. 455-466.

Lin, L., He, Z., & Peeta, S. (2018). Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. *Transportation Research Part C: Emerging Technologies, 97,* pp. 258-276.

Met Office. (2024). Code Definitions. [Online] Available at: https://www.metoffice.gov.uk/services/data/datapoint/code-definitions (Accessed: 2 Apr 2024).

Molnar, C. (2020). Interpretable Machine Learning. Lulu.com.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, pp. 81-106.

Ricci, M. (2015). Bike-sharing: A review of evidence on impacts and processes of implementation and operation. *Research in Transportation Business & Management, 15*, pp. 28-38.

Sathishkumar, V. E., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. *Computer Communications*, *153*, pp. 353-366.

Shaheen, S.A., Guzman, S. and Zhang, H. (2010). Bikesharing in Europe, the Americas, and Asia: past, present, and future. *Transportation research record*, *2143*(1), pp.159-167.

Tomaras, D., Boutsis, I. and Kalogeraki, V. (2018, March). Modeling and predicting bike demand in large city situations. In *2018 IEEE international conference on pervasive computing and communications (PerCom)*. pp. 1-10.

Tran, T. D., Ovtracht, N., & d'Arcier, B. F. (2015). Modeling bike sharing system using built environment factors. *Procedia Cirp*, *30*, pp. 293-298.

Transport for London (TfL) (2010). Mayor's flagship cycling scheme - 'Barclays Cycle Hire' - opens for business. [Press Release] 30 July. Available from: https://tfl.gov.uk/info-for/media/press-releases/2010/july/mayors-flagship-cycling-scheme--barclays-cycle-hire--opens-for-business [Accessed 14 April 2024].

Transport for London (TfL) (2020) Travel in London: Report 13. [Online] Available from: https://content.tfl.gov.uk/travel-in-london-report-13.pdf [Accessed 14 April 2024].

Wang, W. (2016). Forecasting Bike Rental Demand Using New York Citi Bike Data.

Zheng, Z., Zhou, Y. and Sun, L. (2019). A multiple factor bike usage prediction model in bike-sharing system. In *Green, Pervasive, and Cloud Computing: 13th International Conference, GPC 2018, Hangzhou, China, May 11-13, 2018, Revised Selected Papers 13* (pp. 390-405). Springer International Publishing.

Appendix A

"weather_code" category description, based on Met Office:

1 = Clear ; mostly clear but have some values with haze/fog/patches of fog/ fog in vicinity

2 = scattered clouds / few clouds

3 = Broken clouds

4 = Cloudy

7 = Rain/ light Rain shower/ Light rain

10 = rain with thunderstorm

26 = snowfall

94 = Freezing Fog