”Optimizing Urban Mobility: Machine Learning-Based Bike Sharing Demand Prediction”

Ziyan Shirin Raha[1], Khaled Hasan[2], Yasir Arafah Prince[3], Sumaiya Shejin[4]
[1,2,3,4]Department of Computer Science and Engineering,
Ahsanullah University of Science and Technology, Dhaka, Bangladesh

***Abstract*—The study addresses the contemporary challenges of traffic congestion in urban landscapes. Focusing on bike sharing services as an eco-friendly alternative, the research employs machine learning models, including Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Random Forest, Gradient Boosting, and Extreme Gradient Boosting models, to predict bike-sharing demand. Critical aspects such as data preprocessing, feature selection, and hyperparameter tuning are explored to enhance the effectiveness of the predictive models. The study aims to revolutionize fleet management in bike-sharing systems, strategically placing bicycles based on predictive models for a more responsive and sustainable urban transportation system.**

***Index Terms*—Bike Sharing, Machine Learning, Demand Prediction,Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Random Forest, Gradient Boosting, and Extreme Gradient Boosting,PCA,Regression Matrices, Adjusted R-squared.**

I. Introduction

In the rapidly evolving context of contemporary urban landscapes, addressing the challenges of traffic congestion and fostering sustainable transportation solutions have become imperative. Within this dynamic environment, bike sharing services have emerged as a transformative and eco-friendly alternative, offering a convenient mode of transportation for urban residents and visitors alike. This study, titled ”Optimizing Urban Mobility: Machine Learning-Based Bike Sharing Demand Prediction,” is driven by the overarching objective of contributing to the enhancement of urban mobility through precise predictions of bike-sharing demand. Diverging from prior studies, our approach seeks comparable outcomes with optimized machine learning models. We explore various techniques including Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Random Forest, Gradient Boosting, and Extreme Gradient Boosting models. Our research addresses critical aspects of data preprocessing, feature selection, and hyperparameter tuning, ensuring a comprehensive strategy to enhance the overall effectiveness of the predictive model. By harnessing the capabilities of these machine learning techniques, the study aims to revolutionize fleet management within bike-sharing systems. The strategic placement of bicycles, guided by predictive models, seeks to ensure their availability at high-demand locations, thereby fostering a more responsive and sustainable urban transportation system. The anticipated outcome not only aims to alleviate congestion but also aligns with broader environmental conservation goals. Through data-driven insights and advanced predictive modeling, this research endeavors to redefine urban commuting, offering a user-centric and environmentally conscious approach to mobility in contemporary cities

II. Related Work

The literature on bike share demand forecasting reveals diverse methodologies and influencing factors depending on the operational region of the program.

Alhusseini (2014) employed Support Vector Machine (SVM) and softmax regression algorithms for numerical and categorical bike demand prediction, respectively. Du et al. (2014) utilized time-by-hour data, employing generalized boosted models and random forests for demand estimation. Lee et al. (2014) modified feature sets and employed Poisson regression, neural networks, and Markov models. Kim et al. (2018) introduced a study using a graph convolutional neural network. Wang (2016) applied multiple linear regression, neural networks, decision trees, and random forests for predicting New York's CitiBike hourly demand. Godavarthy et al. (2017) explored the operational aspect and travel behavior, while Liu et al. (2019) used standard long short-term memory (LSTM) and multi-time step models for demand prediction. Pan et al. (2018) applied a recursive neural net (RNN) model, and Li (2019) used a Gaussian mixture model (GMM) for customer demand prediction. Zeng et al. (2016) proposed a station-centric model considering global features.

Kaltenbrunner et al. (2010) used an auto-regressive moving average (ARMA) model, and Yoon et al. (2012) applied a modified autoregressive integrated moving average (ARIMA) model considering spatial interaction and temporal factors. Lim and Chung (2019) modified the Holt-Winters method for demand prediction. Chen et al. (2016) predicted excess demand using a weighted correlation network model, and Li et al. (2015) proposed a hierarchical prediction model based on cluster analysis. Min et al. (2017) developed an analysis methodology for public rental bike systems in Daejeon, and Kim et al. (2012) studied the weather's effects on public bike demand. Lee et al. (2011) built a bike demand estimation model based on the number of students and passenger cars, excluding elementary school students.

III. BACKGROUND STUDY

The background study for bike share demand forecasting involves a comprehensive exploration of existing research and

methodologies employed in predicting the demand for bike sharing programs. Researchers have approached this challenge using diverse techniques, considering the unique characteristics of each program's operating region.

One prominent study by Alhusseini (2014) stands out for its dual methodologies. The first approach utilizes a Support Vector Machine (SVM) algorithm to predict bike demand as a numerical attribute, while the second approach employs softmax regression and SVM algorithms, treating bike share demand as a categorical attribute with five class labels.

Du et al. (2014) focused on time-by-hour data, employing a combination of generalized boosted models and random forests to estimate bike share demand. Lee et al. (2014) took a distinctive approach by modifying feature sets, transforming categorical attributes, and employing Poisson regression, a neural network, and a Markov model for demand prediction. Innovative techniques have also been explored, such as the use of a graph convolutional neural network by Kim et al. (2018) and the application of standard long short-term memory (LSTM) and multi-time step models by Liu et al. (2019). Additionally, various studies have addressed the impact of external factors. Wang (2016) focused on New York's CitiBike hourly demand, incorporating multiple linear regression, neural networks, decision trees, and random forests, with a particular emphasis on the random forest ensemble method. Godavarthy et al. (2017) delved into the operational aspects and travel behavior of bike-sharing programs.

Moreover, location-centric models, spatial interaction, and temporal factors have been considered in forecasting. Kaltenbrunner et al. (2010) used an auto-regressive moving average (ARMA) model, while Yoon et al. (2012) employed a modified autoregressive integrated moving average (ARIMA) model, taking into account spatial interaction and temporal factors.

This background study underscores the diversity and complexity of methodologies employed in bike share demand forecasting, providing a foundation for the present research to contribute to this evolving field.

IV. DATASET

The dataset is devoid of any null values.The dataset provides a comprehensive look at various factors that might influence bike rental demand. Here are some key statistics and insights from the dataset:

Data Overview: The dataset contains 8,760 entries, each representing an hour of the day across different days.

Rented Bike Count: This is our target variable for predicting bike demand.

Temporal Data: The Hour column ranges from 0 to 23, covering all hours of the day. This can be crucial for analyzing peak and off-peak hours.

Weather Conditions: Key environmental factors include:

Temperature: Ranges from -17.8°C to 39.4°C, with a mean of approximately 12.88°C.

Speed: Averages at 1.72 m/s, with a maximum of 7.4 m/s.

Visibility: Ranges widely from 27m to 2000m.

Dew Point Temperature: Varies significantly, indicating diverse weather conditions throughout the year.

Solar Radiation: Averages at 0.57 MJ/m², indicating varying levels of sunlight exposure.

Rainfall and Snowfall: These variables indicate precipitation, which can greatly affect bike rental demand.

Seasons: The dataset includes seasonal data, which is vital for understanding seasonal variations in bike rental demand.

Holiday: Indicates whether the day is a holiday or not, which can influence rental patterns. Functioning Day: Shows whether the day was a functioning day for the bike rental system.

A. DATA VISUALIZATION

A 'distplot' is helpful because it offers a rapid and simple method to examine the data distribution, spot trends or anomalies, and contrast the distributions of several variables. It also makes it possible to determine whether or not the data follows a normal distribution. In order to determine whether or not the variable distributions over the entire dataset are symmetric, we employed the histogram plot. We can observe that the distribution of the dependent variable, the leased bike, is positively skewed (right skewed) from the distribution plot above. It denotes an asymmetric distribution with respect to the mean.We discovered that the distribution of our data is not typical. Thus, normalizing the data is necessary before applying any models to it.

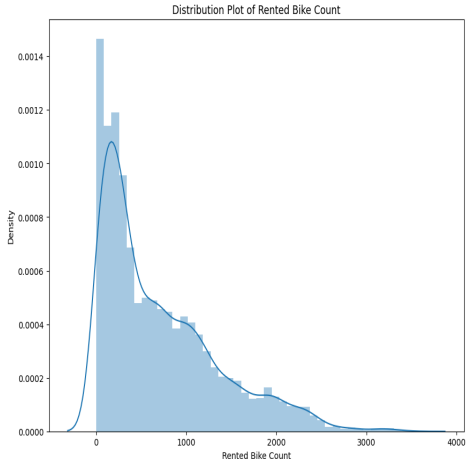
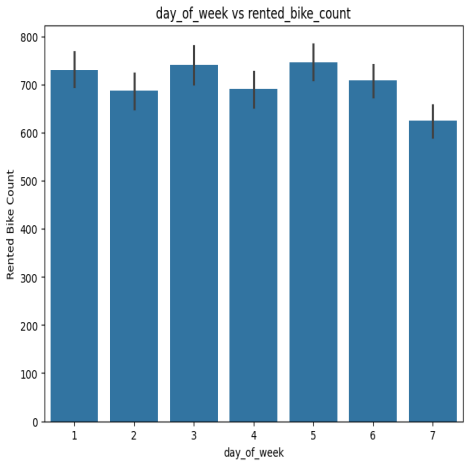


Fig. 1. Dependent variable Distribution

From the bar charts, we learned:

There is a lot of demand around 8 a.m. and 18 p.m., according to the hour vs. rental bike chart. According to the season vs rental bike chart, summertime demand is higher and wintertime demand is lower. The day_of_week vs. rental bike chart indicates that working days have a strong demand. Based on the monthly data, it is evident that June has a significant level of demand. We learned from this that we should concentrate more on that area when there is a greater demand for bikes.



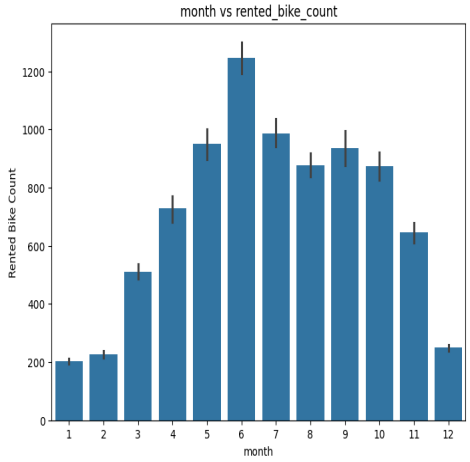
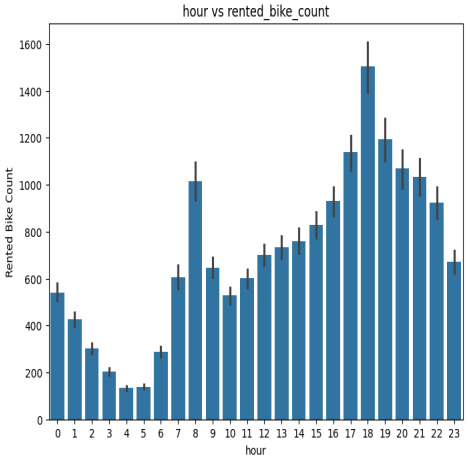


Fig. 2. Categorical variables with dependent variable



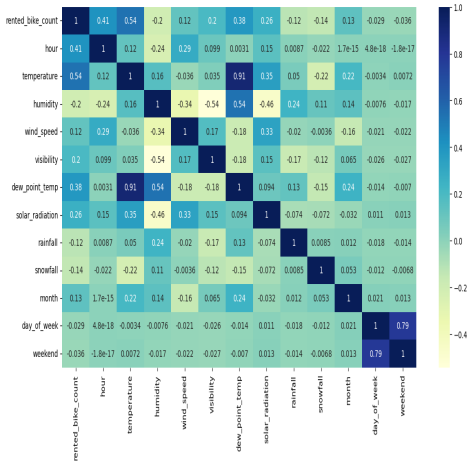
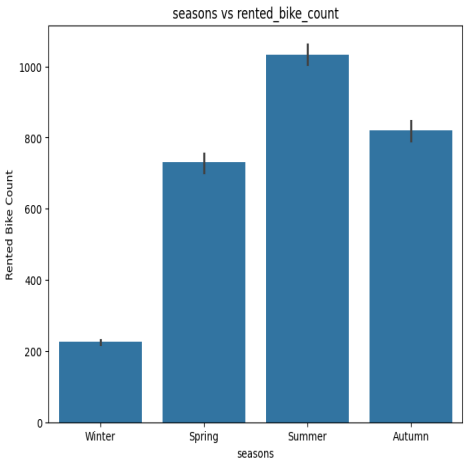


Fig. 3. Correlation Heatmap



It is evident from the correlation map above that: The independent variables (temperature, humidity and dew point temperature, weekend and day of week) have a significant degree of multicolinearity. Temperature, hour, dew point temperature, and solar radiation are correlated with the dependent variable, which is a rented bike. Apart from that, we observed no discernible association.

V. Methodology

A. Preprocessing and Model Implementation

We imported necessary libraries, loaded the dataset and checked its basic statistics. After creating a dataframe with the dataset we renamed the columns and adjusted the data column. We categorized the columns and attempted to understand the dataset through visualization using histograms and box plots. We also explored the relationship between dependent and continuous variables.

Furthermore, we examined the relationship between the dependent variable and categorical variables using line plots. Additionally, we plotted temperature and solar radiation for each observed month. We observed the average bike count for wind speed and humidity.

Finally, we presented a heatmap to better understand the correlation within the dataset. Categorical encoding has been applied, and a dummy dataset has been created with a range of features. The correlation between temperature and dew point has been examined. To mitigate overfitting, feature such as temperature was not considered. Multicollinearity was assessed using the VIF technique. Additionally, date, weekend, and seasons were dropped from the features. The dependent variable was normalized using the square root, and independent variables were assigned to X, while the dependent variable was assigned to Y.

PCA was applied for feature extraction with scaled value of X, and the dataset was split into 70-30 ratio. Seven models, including Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Random Forest, Gradient Boosting, and Extreme Gradient Boosting were implemented. Hyperparameters of these models were tuned, and regression matrices along with the models were presented. The adjusted R-squared was also plotted.

Furthermore, an attempt was made to run the dataset with 14 features. Square root transformation was applied to the dependent variable, and both with and without square root transformations were considered in different files to understand different regression matrices. PCA was used for feature extraction, and Anova , RFE and k-Best methods were employed for feature selection. Twelve best features were selected for K-Best , eleven best features were selected for Anova and six for the RFE. The data were then split into 70-30 ratio, and the same seven models were applied. Hyperparameters of some models were tuned and performed, and regression matrices along with the models were presented. The adjusted R-squared was plotted.

VI. Result Analysis & Conclusion

We have tested the adjusted r2 for 7 models. For Kbest with sqrt y Random Forest performed the best with 87.08%, for Kbest without sqrt Extreme Gradient Boosting performed the best with 79.8%, for Anova Extreme Gradient Boosting performed the best with 83.75%, for RFE Extreme Gradient Boosting performed the best with 82.237%, for manual selection Extreme Gradient Boosting performed the best with 89.53%.

Selection techniques	Linear regression	Lasso regression	Lasso with alpha	Ridge regression	Decision tree	Random forest	Gradient Boosting	Extreme Gradient Boosting
K –best with Square rooted Y	59.07%	44.401%	59.06%	59.055%	77.23%	87.08%	84.74%	87.43%
K –best without Square rooted Y	39.55%	38.97%	39.55%	39.52%	64.52%	79.44%	76.1%	79.8%
Anova	50.36%	43.83%	50.298%	50.35%	66.95%	81.57%	75.914%	83.75%
RFE		32.26%	48.60%	48.59%	67.66%	80.479%	79.378%	82.237%
Manual	77.57%	50.91%	77.65%	77.652%	75.12%	89.03%	84.24%	89.53%

Fig. 4. Result Analysis of Model

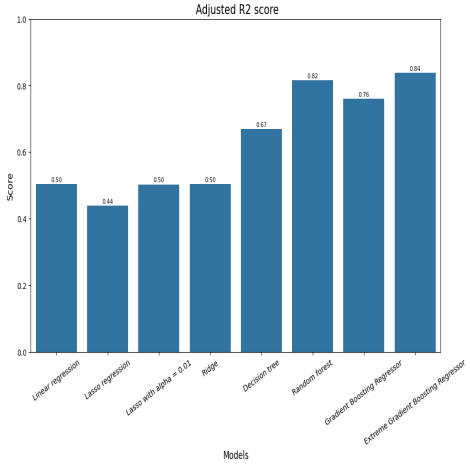


Fig. 5. PCA Anova

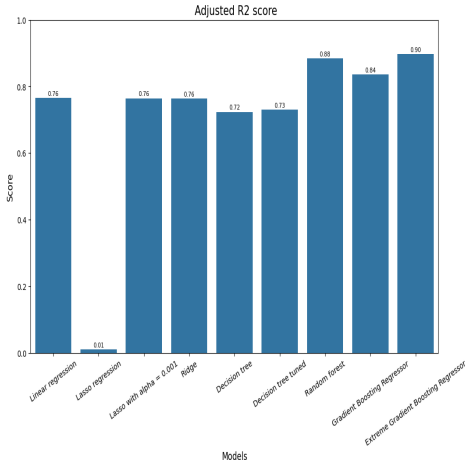


Fig. 6. PCA Manual

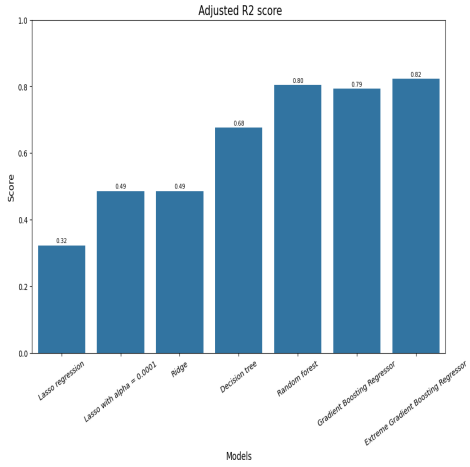


Fig. 7. PCA Wrapper

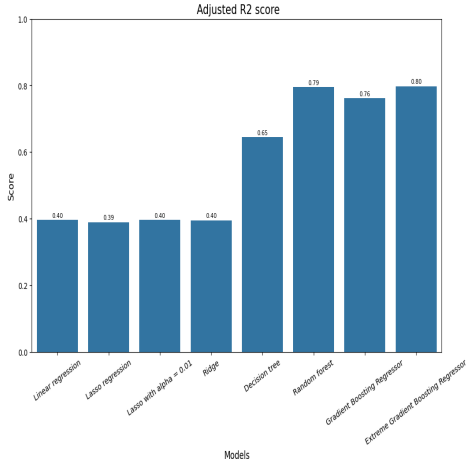


Fig. 8. sqrt wihout kbest

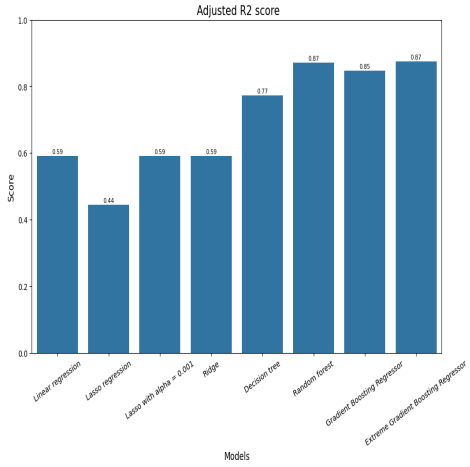


Fig. 9. sqrt wih kbest

VII. LIMITATIONS

There are certain limitations to this study that need to be recognized. Outliers might occur in real-world situations, which would complicate the model's ability to forecast the future. Regional variances may impact the recently introduced feature, which is the daily count of recorded accidents. Moreover, biases may be introduced since certain nations or jurisdictions do not publicly release daily accident records. Through careful hyperparameter optimization, it is also possible that different machine learning or deep learning models might produce better results.

VIII. FUTURE WORK

There are a number of directions that might be investigated in order to improve the validity and relevance of subsequent research. Geographical data may provide important insights; examples of this type of data include the longitude and latitude of bike rental docks, subway stations, and reported accident sites. A more nuanced interpretation of the data may be provided by examining the distances between these geographical points.Reliable sources of hourly weather data during the bike-sharing season may be gathered, and when combined, the resulting larger training set has adequate granularity to enable hourly forecasting. By adding station-specific characteristics to the collection of independent variables and using different training sets for each station, station-level demand prediction may be achieved. Subsequent studies might go beyond forecasting the demand for bike shares to include additional modes of transportation such as vans, electric motorbikes, and vehicle pools.It is important to use caution when extrapolating results to other bicycle sharing programs throughout the world. For the purpose of developing models that can capture a variety of features seen in urban landscapes, a thorough understanding hence necessitates the examination of data from numerous cities.

REFERENCES

[1] Alhusseini, M. (2014). Prediction of Bike Sharing Systems for Casual and Registered Users. Academic Press.

[2] Du,J,He,R.,Zhechev,Z.(2014).Forecasting Bike Rental Demand. In CS 229 Machine Learning Project. Stanford University

[3] Lee, C., Wang, D., Wong, A. (2014). Forecasting utilization in city bike-share program (Vol. 254). Technical report, CS 229 2014 Project.

[4] Kim, Won-Kyung, Son. (2018). Bike Sharing Demands Prediction based on GCN. In Proceedings of the 2018 KISS Conference (pp. 832- 834). Korea Information Science Society

[5] Wang, W. (2016). Forecasting bike rental demand using New York Citi Bike data (MSc thesis). Technological University Dublin, School of Computing College of Science of Health.

[6]] Godavarthy, R., Mattson, J., Taleqani, A. R. (2017). Evaluation Study of the Bike Share Program in Fargo, North Dakota (No. SURLC 17-005). SURLC.

[7] Liu, X., Gherbi, A., Li, W., Cheriet, M. (2019). Multi Features and Multi-time steps LSTM Based Methodology for Bike Sharing Availability Prediction. Procedia Computer Science,