Industrial Engineering Journal
ISSN: 0970-2555
Volume: 52, Issue 4, April 2023

RENTAL BIKES COUNT PREDICTION USING MACHINE LEARNING APPROACH

[1*]GUDAPATI BHARGAVI, [2]DASU PRIYANKA [3]CHINCHILADA MANOHAR, [4]GANAGALA ANIL KUMAR
[1*]Assistant Professor, Dept. of Computer Science and Engineering, Raghu Engineering College,Visakhapatnam.
[2,3,4] B.Tech Students, Dept. of Computer Science and Engineering, Raghu Engineering College, Visakhapatnam.

bhargavigudapati@raghuenggcollege.in, 19981a0538@raghuenggcollege.in, 20985a0505@raghuenggcollege.in, 19981a0549@raghuenggcollege.in

ABSTRACT--

The bike rental system has gained popularity worldwide and has been attempted by various entrepreneurs in India as an alternative mode of transportation, particularly for those without cars. However, inadequate data analysis has led to failures in some cases. To improve customer retention, predicting the number of bikes available for rent at any given time in a city can be beneficial. This prediction problem is known as "Predicting the demand of rental bikes count." To tackle this issue, various machine learning techniques such as Random Forest Regression, MLP and GBM regressions are employed. The evaluation of these models performance is based on different metrics such as R^2 and RMSLE and RMSE,MAE. Generating daily forecasts can help to alleviate the issue of bike shortages, which is a major factor in customer attrition.

KEYWORDS:

Random Forest Regression, GBM Regression ,MLP Regression.

1. INTRODUCTION:

Bike-sharing programs have experienced rapid growth worldwide, with over 500 cities in 49 countries now offering them. These programs have overcome earlier issues such as theft and vandalism through specialized bicycles and monitoring technology. Bike-sharing programs offer environmental and health benefits, as well as increased connectivity to other modes of transit. There was a gradual growth of bike-sharing in the subsequent years, where only one or two new initiatives were introduced on a yearly basis[5].Predicting demand for bike-share systems is crucial for business success in the data technology era, where many decisions are made based on algorithms and data. Shared bike programs have emerged as a means of promoting sustainable transportation systems and responding to climate change and energy crises, but high fixed costs have made some cities and businesses reluctant to introduce them. Accurate demand prediction is essential for the continuous operation of shared bike programs. Shared bikes integrated with public transportation systems can extend the catchment area

of public transportation and accommodate leisure traffic during non-adjacent hours, making predicting demand essential for efficiently implementing an integrated public transportation system that includes shared bikes.

While previous studies have used statistical methodologies and given datasets, this study develops a mathematical models that integrate historical usage patterns with various factors such as weather, holidays, and weekends to estimate the demand for bike rentals. In this project, A model has been developed to anticipate the amount of bike rentals on a daily basis for the entire year by taking into account the prevailing weather conditions. Once the best-performing model is identified, it can be used to predict the bike count. This paper aims to compare multiple models and ultimately employ the most effective one for predicting the output.

2. LITERATURE REVIEW:

Dr. Shanthi Mahesh et al[4] The project reports and survey papers have referenced the Random Forest Prediction algorithm and the Neural Network., Multi Linear Regression, and K-mean Cluster Analysis, and they came to the conclusion that the utilization of a neural network model as a prediction algorithm is most appropriate for training models when there is an abundance of data and the model is not intended for a specific domain.

YouLi Feng et al[2] The authors of this paper found that using the conventional Despite having a strong linear correlation between the variables and normally distributed factors, the utilization of a multiple linear regression model for predicting bike rental demand resulted in a low level of prediction accuracy. Due to particular traits of certain factors, a multiple linear regression model used to predict bike rental demand experienced high error rates in its results. Consequently, the authors suggested a new forecasting model for bike rental demand utilizing random forest with the GBM package, which improves the decision tree's capacity in the random forest process. By integrating random decision trees into the forest, the model's generalization ability was reinforced, while maintaining accuracy through the training process with independent trees. The outcome of the proposed model showed a significant improvement in accuracy, achieving a rate of 82%.

Arthi Akilandesvari Ramesh et al[6] To predict real-time bike demand at a particular station, the Bike Demand forecasting project utilizes multiple classifiers, including random forest, gradient boosting, and linear regression. The project measures the effectiveness of these approaches and selects the best-performing one to complement the data displayed on the dashboard regarding the demand. After analyzing all three datasets, After analyzing the dataset containing all stations, the project discovered that XGBoost exhibited superior

performance compared to the other models, and was consequently selected for implementation in the web application.

Abe Zeid et al[7] A machine learning model is suggested in the paper, which accurately forecasts the daily ridership of bike-sharing in urban areas. The model integrates crucial factors, including weather conditions, and employs ridge regression to achieve a high level of precision. The efficiency of the model was assessed using two separate datasets from Bluebikes in Boston and Citibike in New York City, which have confirmed its robustness and accuracy in predicting the ridership. Overall, the results suggest that this machine learning model can be a valuable tool for predicting bike-sharing ridership in urban areas.

3. METHODOLOGY:

To forecast the count of bike rentals, the suggested approach involves utilizing a range of machine learning algorithms, including Random Forest, GBM Regression and MLP Regression.

DATA SET EXPLANATION

This model employs a dataset named as “day” as the sample data[3] is shown in Fig 1, includes 16 factors, such instant,dteday,season,yr,mnth,holiday,weekday,workingday,weathersit,temp,atemp,hum,windspeed,casual,registered,cnt. which is utilized to train diverse machine learning models. techniques for predicting the bike count and have 731 records in it.

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
1	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600
5	6	2011-01-06	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.089565	88	1518	1606
6	7	2011-01-07	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
7	8	2011-01-08	1	0	1	0	6	0	2	0.165000	0.162254	0.535833	0.266804	68	891	959
8	9	2011-01-09	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.361950	54	768	822
9	10	2011-01-10	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321

Fig.1

Industrial Engineering Journal
ISSN: 0970-2555
Volume: 52, Issue 4, April 2023

Seasonal Comparison of Rental Bike Usage Across Different months--

The Fig.2 is shown below which represents a box plot to visualize the distribution of bike rental counts across different seasons and months. The plot displays the total count on the y-axis, with the x-axis representing the months of the year and the hue parameter indicating the different seasons.

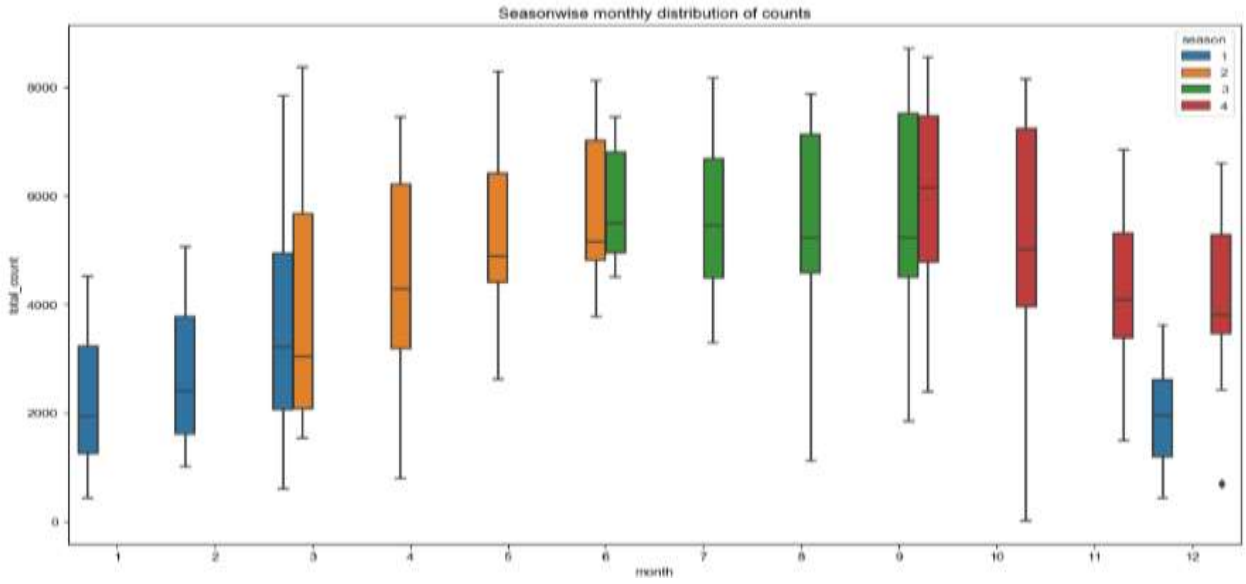
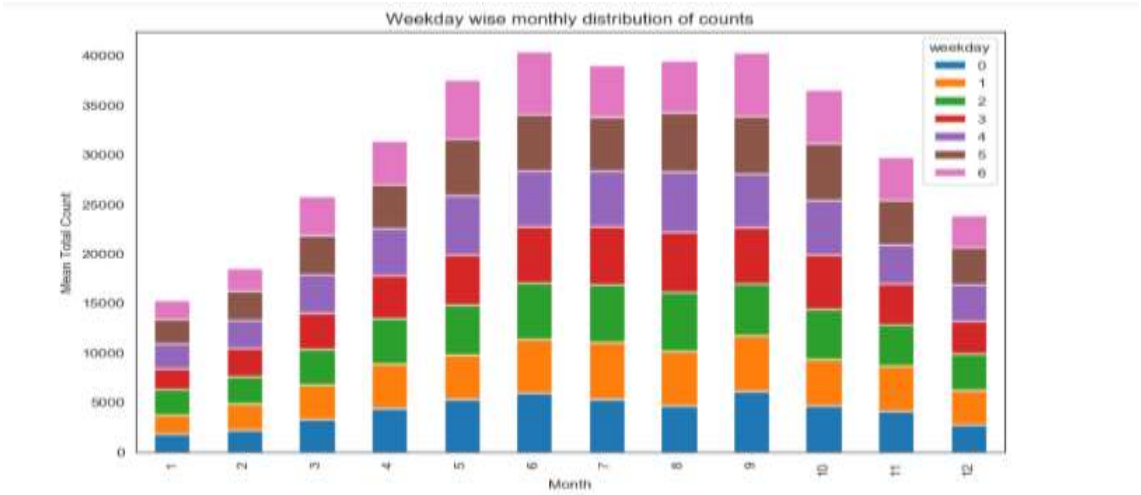


Fig.2

Weekday wise monthly distribution of counts-

The Fig.3 is shown below represents a stacked bar chart that shows the mean total count of bike rentals distributed across different weekdays and months. The x-axis represents the months, the y-axis represents the mean total count, and the different colors in each bar represent the distribution of the data across different



weekdays.

Fig.3

Yearly distribution of counts-

The Fig.4 is shown below which represents a pie chart to display the distribution of bike rentals across different years. The percentage of the total count that each year represents is displayed as a label on each slice. This plot can help us identify any changes or trends in the distribution of bike rentals across different years

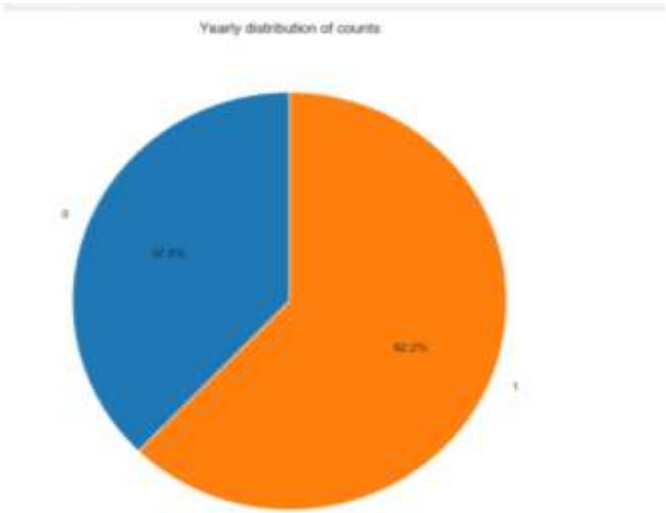


Fig.4

Holiday wise distribution of counts-

The Fig.5 is shown below, which represents a violin plot .The x-axis represents whether a day is a holiday or not, and the y-axis represents the total count of bike rentals. The different colors in each violin plot represent the distribution of the data across different seasons.

Industrial Engineering Journal
ISSN: 0970-2555
Volume: 52, Issue 4, April 2023

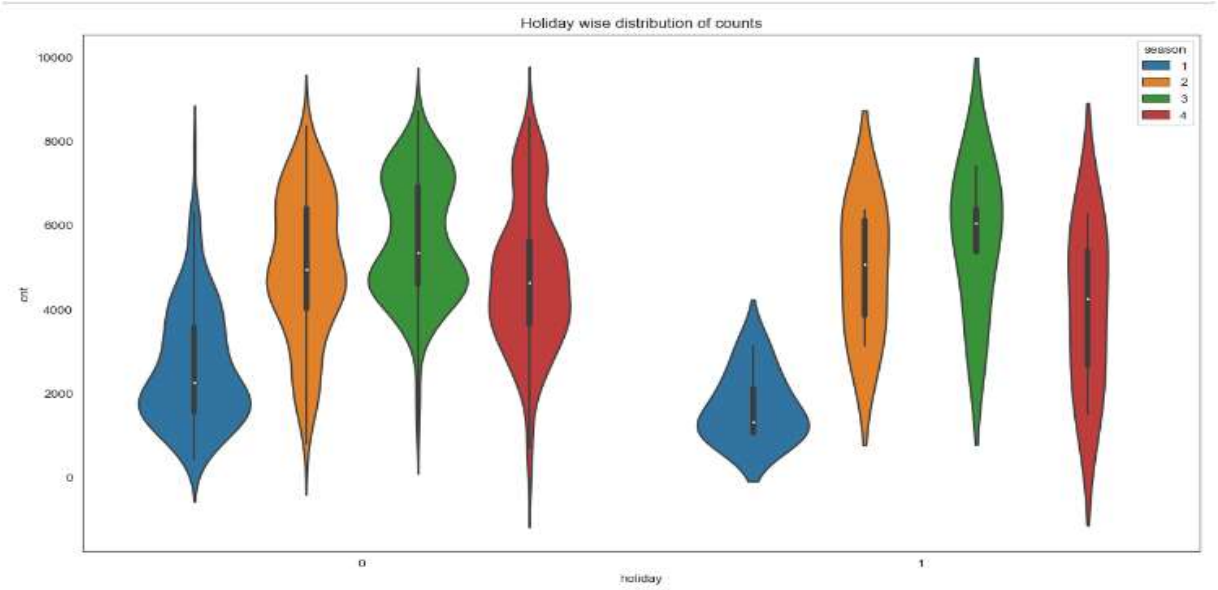


Fig.5

Workingday wise distribution of counts by season—

The Fig.6 is shown below which represents a pie chart shows the distribution of total counts of bike rentals by season for each workingday category. The original data is grouped by workingday and season, and the sum of total counts is calculated.

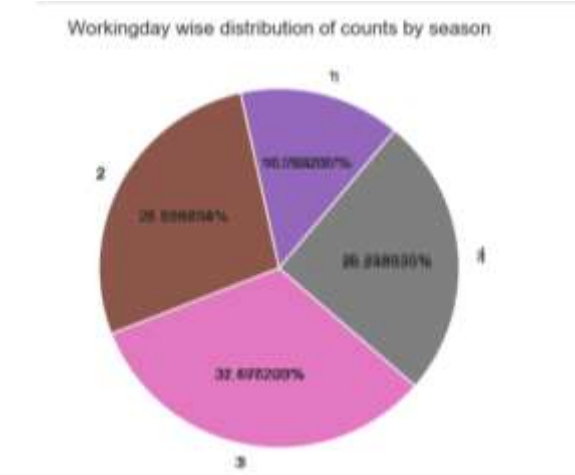


Fig.6

Industrial Engineering Journal
ISSN: 0970-2555
Volume: 52, Issue 4, April 2023

Weather condition wise monthly distribution of counts—

The Fig.7 is shown below which represents a bar plot to display the distribution of bike rentals based on different weather conditions for each month. The x-axis represents the different weather conditions, the y-axis represents the mean total count of bike rentals, and the bars represent the distribution of bike rentals for each month.

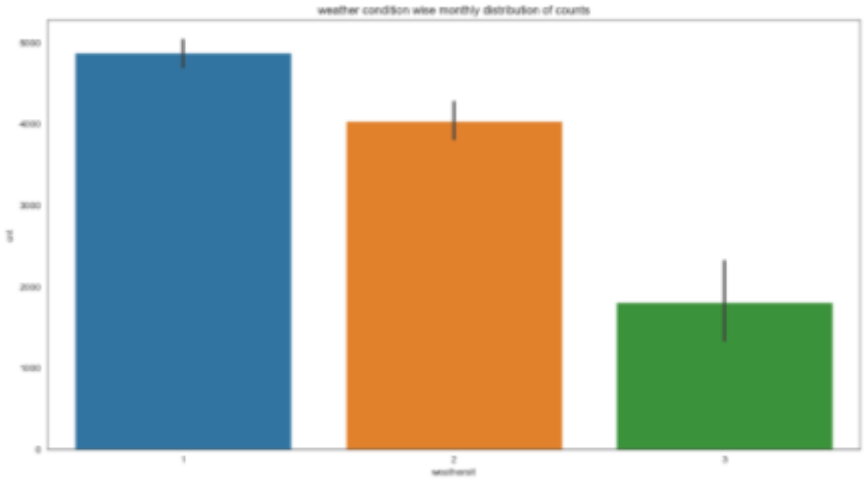


Fig.7

DATA PREPROCESSING:

First we perform checking missing values by using Bike.isnull().sum() calculates the number of missing values for each variable in the Bike dataset .Next checking of outliers.These are data points that lie far away from the majority of other data points and can have a significant impact on the results of a statistical analysis or machine learning model. Outliers can be caused by various factors such as measurement error, data entry errors, or rare events.Next feature selection is performed .Here dropping of unnecessary columns such as atemp,hum.

One can use the 'train_test_split' function provided by the 'sklearn' library to split the dataset into separate subsets for training and testing purposes. This function allows for the data to be split based on a designated ratio. It is recommended to use an 80:20 split ratio, which means that 80% of the data will be used for training and 20% for testing. depending on whether the value is present or not.

IMPLEMENTATION TECHNIQUES

RANDOM FOREST REGRESSION:

Industrial Engineering Journal
ISSN: 0970-2555
Volume: 52, Issue 4, April 2023

An Ensemble learning is utilized in Random Forest Regression, which is a technique that generates a robust prediction model by amalgamating several decision trees. The individual trees are trained on a randomly chosen subset of data, and the ultimate prediction is an average of all the trees' forecasts. [9]. Compared to single decision

tree models, Random Forest Regression offers superior generalization, less overfitting, and increased accuracy.

GRADIENT BOOSTING REGRESSION:

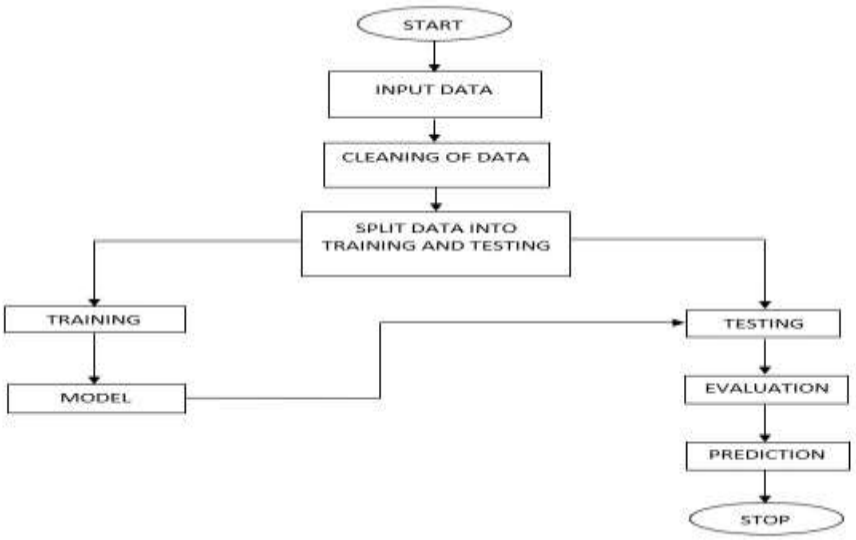
Supervised learning regression problems can be addressed with the use of Gradient Boosting Regression, which is a machine learning technique. It belongs to the ensemble method family, which combines several weak learners, usually decision trees, to create a more robust model. Initially, the algorithm fits a model to the data, and subsequently, it adjusts the model iteratively to reduce errors[10]. It is often utilized for high-dimensional data and can manage both linear and nonlinear relationships between the target variable and features.

MLP REGRESSION:

Multilayer perceptron regression, or MLP regression, is an artificial neural network that is utilized in regression analysis to predict a continuous output variable from a set of input variables. This model consists of multiple layers of nodes or neurons that perform nonlinear transformations on the input data.

4. PROPOSED MODEL

The Fig.8 shows the proposed approach which is given below .To build a machine learning model, the data needs to be input into the program, cleaned and preprocessed to ensure that it is in a suitable format, and then split into a training and testing set. The subsequent stage is to choose a fitting algorithm and train the model



using the training data by adjusting its parameters to improve its performance. Once the model is trained, it is tested on the testing set to assess its performance and metrics such as R-Squared,RMSE,MAE,RMSLE score are computed. If the model performs well, it can be used to make predictions on new data, but if not, it may need to be refined or a different algorithm may need to be tried. Finally, the model can be deployed in a production environment, where it may ensure its sustained optimal performance, it is necessary to monitor and update it over time..

Fig.8 Proposed Architecture

EVALUATION METRICS

i. R-Squared: The formula for R-squared, also called the coefficient of determination, is expresse:

R^2 = 1 - (SS_res / SS_tot)

In this equation, SS_res represents the sum of squared residuals or errors, and SS_total is the total sum of squares. The R-squared metric takes values between 0 and 1, with 0 indicating that the model cannot account for the variation in the response variable around its mean, while 1 implies that the model accounts for all the variability.

ii. RSME: The formula for Root Mean Squared Error (RMSE) is:

RMSE = sqrt(mean((y_true - y_pred)^2))

The RMSE formula calculates the square root of the average of the squared differences between the predicted and

actual values of the target variable. It involves summing up the squared differences between each predicted value

and its corresponding actual value and dividing by the total number of model predictions. The resulting value is

the mean squared error (MSE), which is then square rooted to yield the RMSE. This metric measures the average

deviation between the predicted and actual values, allowing for the quantification of the model's performance.

The square root operation is applied to ensure that the RMSE is expressed in the same units as the target variable.

iii. MAE: The formula for Mean Absolute Error (MAE) is:

MAE = mean(|y_true - y_pred|)

The mean absolute error (MAE) formula provides an average measure of the absolute differences between the predicted and actual values of the target variable. To calculate it, the absolute difference between each predicted value and its corresponding actual value is obtained. These absolute differences are then summed up and divided

by the total number of model predictions.

iv. RMSLE: The formula for Root Mean Squared Log Error (RMSLE) is:

RMSLE = sqrt(mean((log(1+y_true) - log(1+y_pred))^2))

In the given context, the vector y_true denotes the actual values, whereas y_pred represents the predicted values. The mean is calculated by considering all the forecasts present in the dataset. When the target variable has a broad range of values, the Root Mean Squared Log Error (RMSLE) is a frequently employed metric for evaluating regression model performance. It measures the ratio between the predicted and true values, rather than the absolute difference. The logarithmic transformation is applied to both the predicted and true values to reduce the impact of large differences between them.

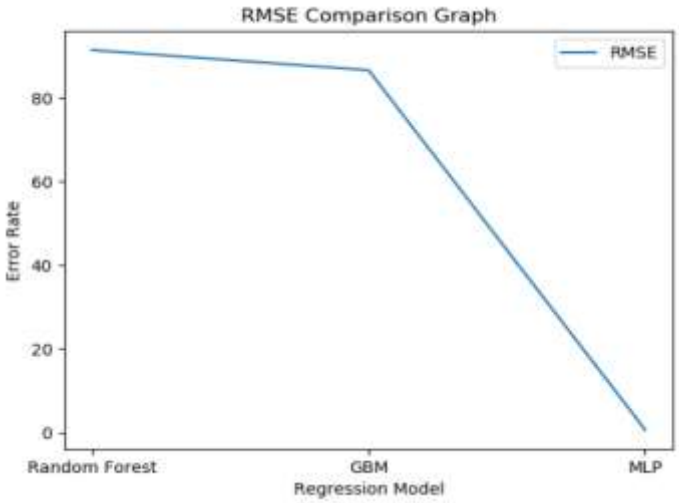
5. RESULT AND ANALYSIS:

To assess the efficacy of each regression model, the study utilized multiple performance metrics, including R-squared, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Root Mean Squared Log Error. These metrics were employed to evaluate different aspects of the models' performance, such as the accuracy of the predictions and the magnitude of errors between predicted and actual values. By employing a variety of metrics, the study was able to gain a more comprehensive understanding of each model's effectiveness and make informed decisions regarding which models to use for further analysis or deployment. The table below provides a detailed overview of the error rates for each model across the three different performance measures.

Machine Learning Models	R-Squared	RMSE	MAE	RMSLE
Random Forest Regression	0.99	91.3	66.21	0.034
Gradient Boosting Regression	0.99	86.5	62.50	0.028
Mlp Regression	0.99	0.70	0.56	0.000

Industrial Engineering Journal
ISSN: 0970-2555
Volume: 52, Issue 4, April 2023

Table.1 Error Rate Comparison



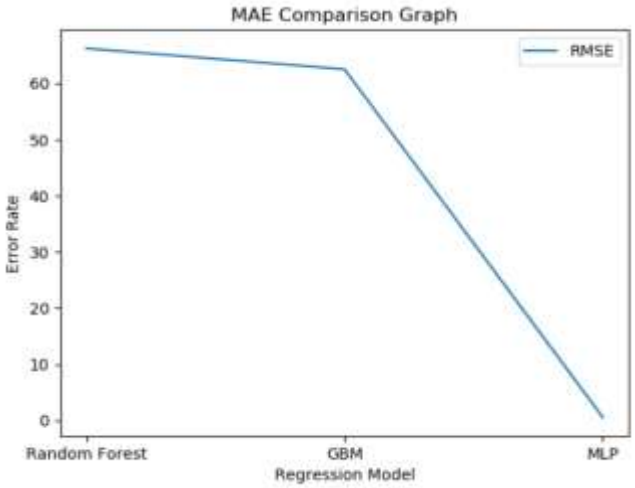


Fig.9 RSME Graphical Comparison

Fig.10 MAE Graphical Comparison

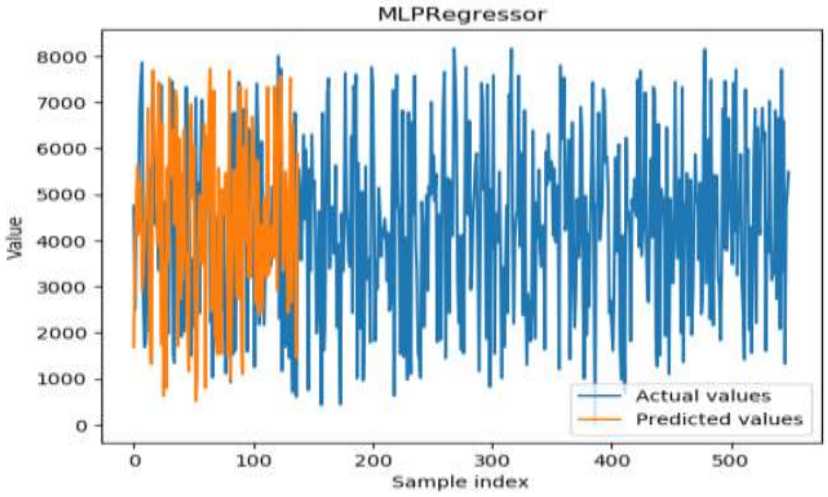


Fig.11 Line chart which shows actual and predicted values based on MLP

Based on the results presented in Table 1,the Random Forest Regression model has an R2 value of 0.99, an RMSE value of 91.3, a MAE value of 66.21, and an RMSLE value of 0.034. The Gradient Boosting Regression model has an R2 value of 0.99, an RMSE value of 86.54, a MAE value of 62.5, and an RMSLE value of 0.028. Finally, the MLP Regression model has an R2 value of 0.99, an RSME value of 0.70, a MAE value of 0.56, and

an RMSLE value of 0.00. In conclusion, According to the results, the MLP model is the most accurate and reliable method for forecasting bike rental counts.

Figure 9 displays a chart comparing the R-squared errors of three models, namely Random Forest Regression, Gradient Boosting Regression, and MLP Regression, providing a comprehensive evaluation of their performance. This graph facilitates a side-by-side comparison of the predictive performance of these models for bike count estimation.

Fig. 10 displays a comparison graph of the Root Mean Squared Log Error (RMSLE) for three different models, namely Random Forest Regression, Gradient Boosting Regression, and MLP Regression. The graph allows for a comparative assessment of the performance of each model in terms of RMSLE.

The Fig.11 represents a line chart which gives us the actual count and predicted count of bikes given by MLP with x-axis as sample index and y-axis as values.

After analyzing various regression models to predict bike rental counts, it was determined that the MLP (Multi-Layer Perceptron) Regression model outperformed other models, including Random Forest Regression, and Gradient Boosting Regression. The MLP Regression model demonstrated superior performance in terms of its ability to accurately predict bike rental counts compared to the other models evaluated. The findings suggest that the MLP Regression model may be a promising approach for predicting bike rental counts in similar contexts. The MLP model had the highest R-squared value, the lowest RSME value, the lowest MAE value, and the lowest RMSLE value, indicating superior accuracy and reliability in predicting bike rental counts. Therefore, the MLP model can be considered as the most suitable and effective approach for predicting bike rental counts and can help bike rental businesses to make informed decisions about managing their inventory and resources efficiently.

6. CONCLUSION AND FUTURE SCOPE

In summary, the project of utilizing machine learning models to predict bike rental counts holds significant potential for bike rental businesses. Among the various regression models analyzed, the MLP (Multi-Layer Perceptron) model emerged as the most promising. After comparing its performance with other models, including Random Forest Regression, and Gradient Boosting Regression, it was evident that the MLP model demonstrated superior accuracy and reliability in predicting bike rental counts. These findings indicate that implementing the MLP model may be a viable approach for bike rental businesses seeking to improve their bike rental count prediction accuracy. Based on the analysis, the results suggest that the regression model may be the most suitable approach for accurately predicting bike rental counts in similar scenarios.. The use of machine learning models can help bike rental businesses make informed decisions about managing their inventory and resources efficiently, leading to increased customer satisfaction and profitability.

In terms of future scope, the project can be extended by incorporating additional features such as weather data and events data, which can further improve the accuracy of the models. The project can also be expanded to cover a wider range of locations and timeframes, enabling bike rental businesses to make predictions for various types of bikes in different locations. Additionally, the project can be integrated with mobile applications and sensors, allowing bike rental businesses to access real-time bike rental count predictions and make timely decisions. Overall, the project has significant potential to contribute to the development of smart mobility solutions and improve the efficiency and profitability of bike rental businesses.

7. REFERENCES:

[1] G. Balachandran, S. Dinesh Kumar, S. Nageshwaran, R. Nithish Ram, Mr. R. Karthik M.E. “BIKE RENTAL DEMAND FORECAST USING ML TECHNIQUES”, International Research Journal of Modernization in Engineering Technology and Science, Volume:04/Issue:07/July-2022.

[2] YouLi Feng , ShanShan Wang, “A Forcast for Bicycle Rental Demand Based on Random Forests and Multiple Linear Regression”,icis,2017

[3] https://www.kaggle.com/c/bike-sharing-demand/

[4] Dr. Shanthi Mahesh , Aditya Warrier , Deepthi P, Dimple K H “Prediction of Bike Rentals” , International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, Vol. 7, Issue 2, February 2019.

[5] P. DeMaio, "Bike-sharing: History, Impacts, Models of Provision, and Future," Journal of Public Transportation, vol. 12, no. 4, p. 3, 2009.

[6] Arthi Akilandesvari Ramesh; Sai Pavani Nagisetti; Nikhil Sridhar; Kenytt Avery; Doina Bein, "Station-level Demand Prediction for Bike-Sharing System", 2021 IEEE 11th Annual Computing and Communication Workshop and Conference, **DOI:** 10.1109/CCWC51732.2021.9375958

[7] Abe Zeid, Trisha Bhatt, and Hayley A. Morris," Machine Learning Model to Forecast Demand of Boston Bike-Ride Sharing", European Journal of Artificial Intelligent and Machine Learning www.ej-ai.org, ISSN: 2796-0072 DOI : 10.24018/ejai.2022.1.3.9

[8] V. Cherkassky and Y. Ma, "Practical Selection of SVM Parameters and Noise Estimation for SVM Regression," Neural Networks, vol. 17, pp.113-126, 2004

[9] Joelsson, S. R., Benediktsson, J. A., & Sveinsson, J. R. (2005). Random forest classifiers for hyperspectral data. In Geoscience and Remote Sensing Symposium, 2005. IGARSS 05. Proceedings. 2005 IEEE International (Vol. 1, p. 4–pp). IEEE.

[10]] GeeksforGeerks. Python – Coefficient of Determination-R2,RSMLE,MAE,RMSE score [Internet]. 2020 [cited 2022 Feb 22]. Available from: https://www.geeksforgeeks.org/python-coefficient-of-determinationr2-score/.