Regression Models to Predict Bike Sharing Demand.

G. Naga Saikarthik

Lovely Professional University.

KEYWORDS

Data Mining, Linear Regression, Correlation Analysis, Bike Sharing Demand Prediction, Hyperparameter tuning

ABSTRACT

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return has become automatic. Through these systems, user can easily rent a bike from a particular position and return at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousand bicycles.

Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues. Apart from interesting real-world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

I.INTRODUCTION

Bike-sharing systems represent a transformative approach to urban mobility, offering a convenient and sustainable means of transportation. These systems operate through a network of automated kiosks strategically placed throughout cities, allowing users to rent bikes on a short-term basis for commuting, leisure, or exercise. The concept of bike-sharing originated in the 1960s in Europe, but it gained widespread adoption globally starting in the mid-2000s, particularly in densely populated urban areas where congestion and pollution are significant challenges.

In North America, bike-sharing programs are often implemented in partnership with municipal governments, aimed at reducing traffic congestion, promoting healthier lifestyles, and decreasing carbon emissions. However, they are also prevalent in smaller communities and college towns, where they serve as a convenient mode of transportation within campus or local areas.The core features of a typical bike-sharing system include station-based bikes, automated payment systems, membership options, and per-hour rental fees. These systems are designed to be user-friendly, allowing both occasional users and regular commuters to easily access and utilize the service. Despite variations in program specifics from city to city, the overall structure and functionality remain consistent enough to prevent confusion among users transitioning between different bike-sharing programs.

The integration of Internet of Things (IoT) technologies with bike-sharing systems has been a game-changer, particularly with the advent of Industry 4.0. Real-time data tracking enables operators to monitor bike availability, analyse usage patterns, and optimize fleet management. Factors such as weather conditions, traffic density, and user preferences can be dynamically accounted for, leading to more efficient operations and enhanced user experiences.

From an environmental standpoint, bike-sharing plays a crucial role in promoting sustainable transportation and reducing the carbon footprint associated with traditional commuting methods. By encouraging cycling as a mode of transport, cities can mitigate air pollution, alleviate strain on public

transit systems, and contribute to overall urban liability.

Moreover, bike-sharing initiatives have significant public health implications. Regular cycling not only improves physical fitness but also reduces the risk of chronic diseases such as obesity, diabetes, and cardiovascular ailments. By integrating physical activity into daily commuting routines, bike-sharing fosters healthier habits among the population, addressing sedentary lifestyle challenges prevalent in modern society.

In terms of market dynamics, the global Bike-Sharing market has exhibited substantial growth, with a market size of USD 2570.9 million in 2019. Projections indicate that the market is poised to exceed USD 13780 million by 2026, showcasing a robust Compound Annual Growth Rate (CAGR) of 26.8% during the period of 2021-2026, according MarketWatch analysis.

For our project, we accessed data from the UCI Machine Learning Repository, focusing on bike rental counts per day. The dataset comprised 8760 entries with 14 attributes, including weather variables (temperature, humidity, windspeed, visibility, dewpoint, solar radiation, snowfall, rainfall), bike rental counts, and date information. While the date variable itself may not directly contribute to predicting bike rental counts, it serves as a crucial factor in understanding seasonal trends, holiday effects, and other temporal patterns that influence bike usage.

By leveraging advanced regression analysis techniques, we aimed to develop a predictive model that correlates weather conditions, time of day, and other factors with bike rental demand. This analytical approach enables us to uncover insights into the factors driving bike-sharing usage, thereby informing strategies for optimizing system performance, enhancing user experiences, and contributing to the broader goals of sustainable urban mobility.

II.LITERATURE REVIEW

1.Bike-sharing Systems: Evolution and Adoption
This section delves into the historical evolution of bike-sharing systems, tracing their origins from early experiments in Europe to their global proliferation in the 21st century. Key factors driving the adoption of bike-sharing programs in urban areas, such as congestion mitigation, last-mile connectivity, and environmental concerns, are discussed. Various models of bike-sharing programs, including station-based, dockless, and hybrid systems, are analysed in terms of their operational efficiency and user accessibility.

2. ***Urban Mobility and Transportation Planning*** - Examining the role of bike-sharing within the broader context of urban mobility, this section explores how bike-sharing complements existing public transit infrastructure and contributes to more sustainable transportation systems. Studies on the integration of bike-sharing with public transit networks, urban planning policies promoting cycling infrastructure, and the impact on modal shift behaviour are reviewed. The potential synergies between bike-sharing, ride-sharing, and other innovative mobility solutions are also discussed.

3. ***Environmental Benefits and Sustainability***:
This section focuses on the environmental benefits of bike-sharing, including reductions in carbon emissions, air pollution, and traffic congestion. Research studies measuring the environmental impact of bike-sharing programs, life cycle assessments of shared bicycles, and strategies for promoting cycling as a sustainable mode of transport are reviewed. The role of bike-sharing in achieving climate change mitigation goals at the city and regional levels is also examined.

4. ***Public Health Implications***: Addressing the public health dimensions of bike-sharing, this section highlights the positive impact of regular cycling on physical fitness, mental well-being, and disease prevention. Studies on the health benefits of active transportation modes, including cycling, walking, and micromobility, are synthesized. The role of bike-sharing in promoting physical activity, reducing sedentary lifestyles, and addressing public health challenges related to obesity, diabetes, and cardiovascular diseases.

Authors	Dataset used	Size	Pre Processing techniques	Model used	R2 score
Zhi Feng Wang	**Hours.csv**	5900	Data Collection Data Description Data Analysis	Linear Regression Model, Extreme learning machine	P>|t|=0.00 P>|t|=0.00
Zijin Xu	**Days.csv**	12000	Check for null/missing values. Removing unnecessary columns	Linear Regression Empirical Analysis	0.8203
Aditya Singh Kashyap	**Bikes.csv**	8760	Scatter distribution	Hypothesis Formation Regression Parameters	0.567

III.DATA

3.1 Variables

3.1.1 Variables and definitions

In this project, we are working with a dataset sourced from Kaggle that provides information on daily demands across American markets for a two-year period. The dataset contains 16 columns, including various variables that describe different aspects of the data. Let's delve into the variables and their definitions:

instant: This variable represents a unique identifier for each record in the dataset.
date: Indicates the date for each record.
season: Represents the season of the year, with values 1, 2, 3, and 4 corresponding to seasons as follows: 1 for January to March 2 for April to June 3 for July to September, and 4 for October to December.
year: Represents the year of the record, with values 0 and 1 corresponding to the years 2018 and 2019, respectively.
month: Indicates the month of the year, with values ranging from 1 to 12, representing January to December.
holiday: This binary variable has values 0 and 1, where 0 signifies "No" holiday and 1 signifies "Yes" holiday.

weekday: Represents the day of the week, with values ranging from 0 to 6, corresponding to

working day: Another binary variable with values 0 and 1, where 0 indicates a "No" working day and 1 indicates a "Yes" working day.
weather situation: Indicates the weather conditions, with values 1, 2, and 3 representing three different weather situations.
lowest temperature: Represents the lowest temperature recorded for the day.

Highest temperature: Indicates the highest temperature recorded for the day.
humidity: Represents the humidity level for the day.
wind speed: Indicates the wind speed for the day.
casual user: Represents the number of casual users of a bike-sharing system on that day.

registered user: Indicates the number of registered users of a bike-sharing system on that day.
total demand: Represents the total demand for bike rentals on that day, combining casual and registered users.

To incorporate categorical variables like 'month,' 'weekday,' 'season,' and 'weather situation' into our regression analysis, we will create dummy variables. Dummy variables are binary (0 or 1) variables created to represent categorical variables in quantitative analysis. For example, for the 'season' variable, we will create three dummy variables (season 2, season 3, and season 4), where each dummy variable will take a value of 1 if the corresponding season is true and 0 otherwise.
Creating dummy variables allows us to include

categorical information in regression models and estimate the effects of these categorical variables quantitatively. It helps capture the influence of qualitative factors on our analysis while maintaining compatibility with numerical analysis techniques.

data preprocessing steps, and exploratory data analysis (EDA) look thorough and well-structured. Let's break down each section for better understanding:

3.2 Data Preprocessing

3.2.1 Check for Null/Missing Values

You correctly pointed out the importance of checking for null values in data analysis. Missing data can significantly impact analysis results and models. It's good to hear that there are no missing values in your dataset, ensuring the reliability of your analysis.

3.2.2 Removing Unnecessary Columns

You identified and justified the removal of certain columns from further analysis based on their redundancy or lack of relevance to the study's goals. This step helps streamline your dataset and focus on the variables that are most impactful for your analysis.

3.3 Data Analysis

3.3.1 EDA

Exploratory Data Analysis is indeed a crucial step in understanding the dataset's characteristics, patterns, and relationships. It sets the foundation for subsequent analysis and modeling tasks by providing valuable insights and detecting potential anomalies.

3.3.2 Visualizing Numeric Variables

Creating pair plots for numerical variables is an effective way to visualize relationships and identify linear correlations. Your observation of a linear relationship between "temp," "atemp," and "cnt" is valuable for modelling purposes.

3.3.3 Visualizing Categorical Variables

Using box plots to visualize the impact of categorical variables on the target variable ("cnt") is a standard practice in EDA. Your analysis of different categorical variables (season, month, weathersit, holiday, weekday, working day) provides valuable insights into their potential predictive power.

3.4 Rescaling the Features

Rescaling features using Min-Max Scaler is a common preprocessing step, especially when dealing with numerical variables with varying scales. It ensures that all features contribute equally to the analysis and prevents biases due to feature magnitudes.

Overall, your data preprocessing and EDA steps demonstrate a structured approach to understanding and preparing your dataset for further analysis and modelling. These steps are essential for ensuring the accuracy, reliability, and interpretability of your results.

IV.PROPOSED METHODOLOGY

4.1 Data collection

the Capital Bikeshare System in Washington D.C., covering the years 2011 to 2012. With 15 attributes including weather conditions, date, weekday/public holiday information, bike rental counts, and temperature, you have a rich source of data for your analysis. This dataset can provide valuable insights into bike-sharing patterns, the impact of weather on bike rentals, and the overall usage trends of the bikeshare system in Washington D.C.

The inclusion of weather variables is particularly important as weather conditions can significantly influence people's decision to use bike-sharing services. Factors such as temperature, precipitation, wind speed, and weather situations (e.g., sunny, rainy, cloudy) can all affect ridership levels.

Additionally, having information about dates, weekdays, and public holidays allows for a detailed analysis of temporal patterns in bike rentals. For example, you can investigate whether there are differences in bike usage between weekdays and weekends, or how bike rentals vary during public holidays or seasonal events.

The count of bikes rented on each day serves as your target variable or dependent variable, which you can predict, or model based on the other attributes in the dataset. This can be done using various statistical and machine learning techniques to understand the factors driving bike rental demand and to make predictions about future usage patterns.

Overall, dataset from the Capital Bikeshare System provides a rich and diverse set of variables that can yield valuable insights into bike-sharing dynamics in Washington D.C. It's a great foundation for

conducting thorough analysis and deriving actionable conclusions to improve bike-sharing services and urban mobility.

4.2 Data Description

The bike sharing data is recorded on daily basis. The records are about 731 days. All the data information can be shown in Table 1.

Table 1. Bike sharing dataset.

4.3 Data Analysis .

In order to figure out which factors have impact on the number of bikes. The bar plot, box plot and scatter plot will be drawn to show the influence on the number of bikes.

Figure 1.a shows that there is a relationship between temperature and number of users, the temperature is 0.8, which has the highest number.

Figure 1.b shows that there is a correlation between Season and the number of users. Summer and Autumn has the most users.

Figure 1.c: The weather and number of bikes bar plot shows that when the weather sit is in 1 means that the weather is clear, few clouds mist or broken cloud, the total number of bikes is around 2,250,000. When the weather sit is in means that the weather is mist and cloudy, mist and broken or mist, the total number of bikes is around 996,000. When the weather sit is in 3 means that the weather is light snow, light rain and thunderstorm or light rain and scattered clouds, the total number of bikes is around 37,000. No users use bikes in weather sit 4 which are heavy rain and ice pallets and thunderstorm and mist or snow and fog.

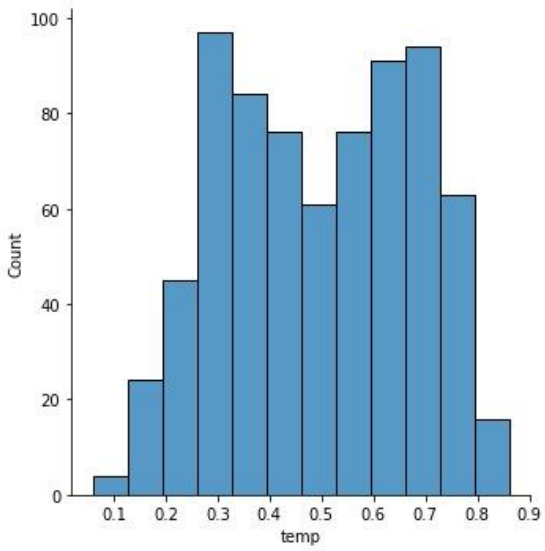


Figure1 : Temp& Count

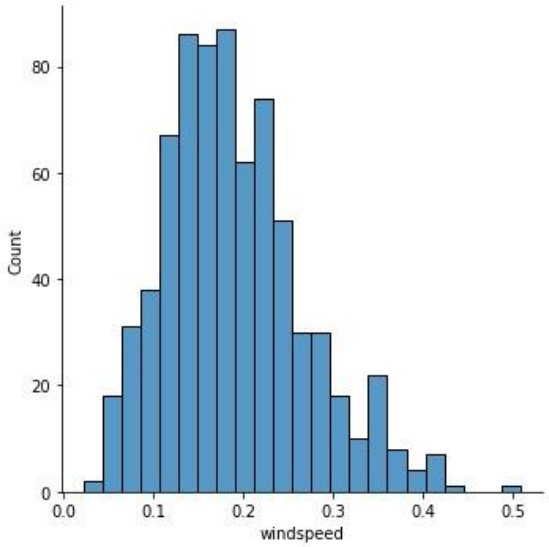
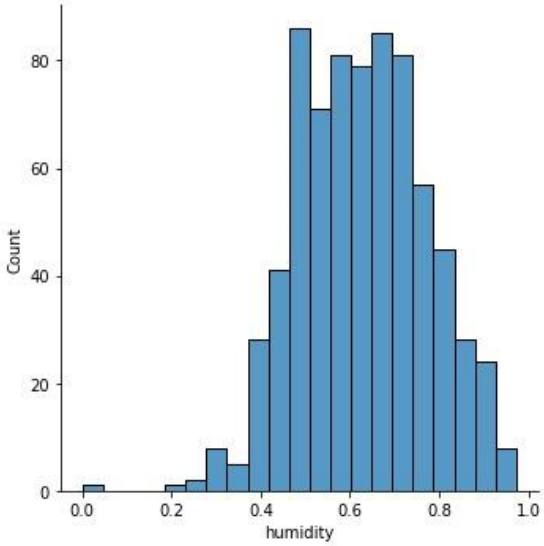


Figure1a:windspeed&count

Figure1b:humidity&count



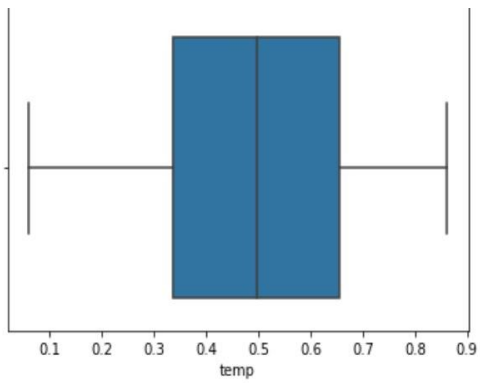


Figure 2 : temp

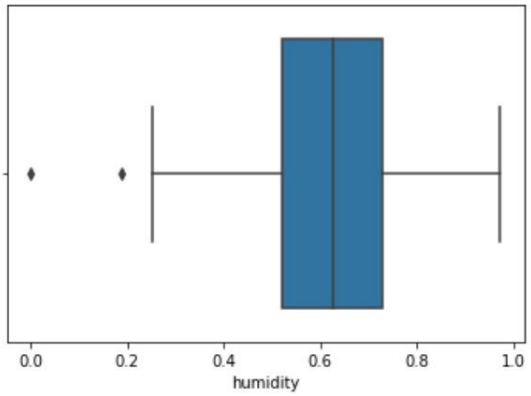


Figure 2a.Humidity

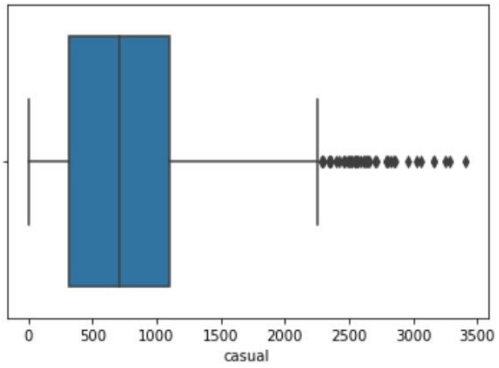


Figure2b.Casual

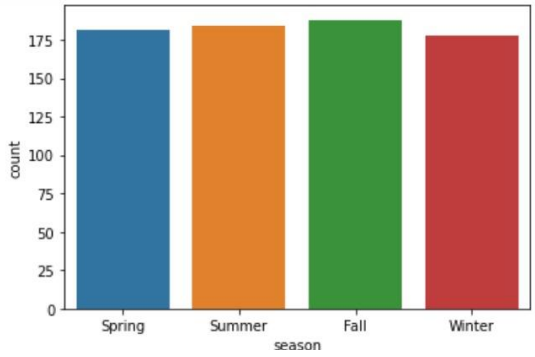


Figure 3.Season &Count

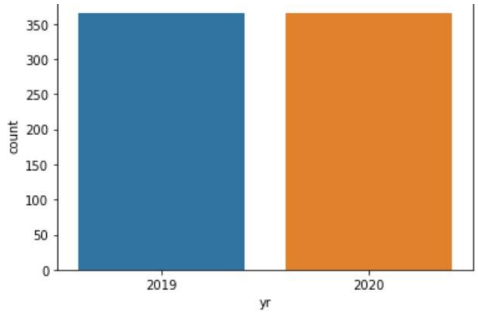


Figure 3a.Year &count

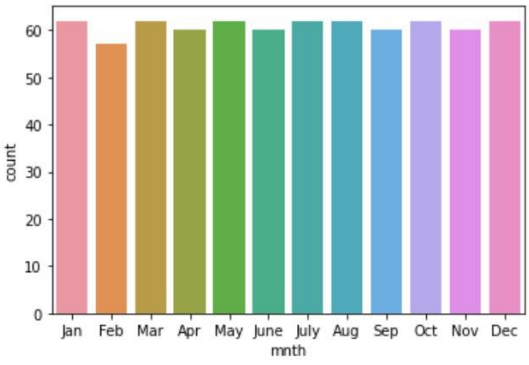


Figure 3b. Month &count

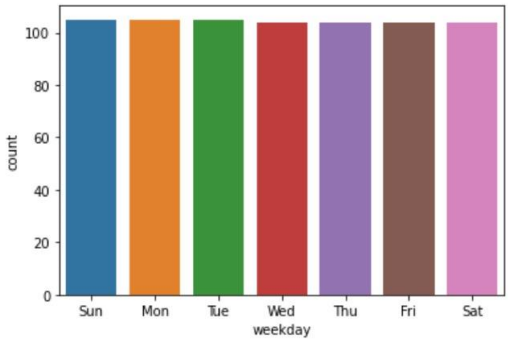


Figure 3c.Weekday&count

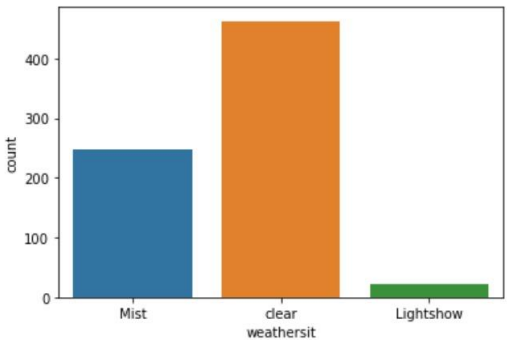


Figure 4. Weathersit &count

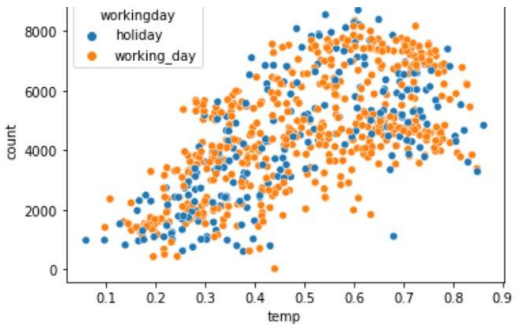


Figure 5.temp&count

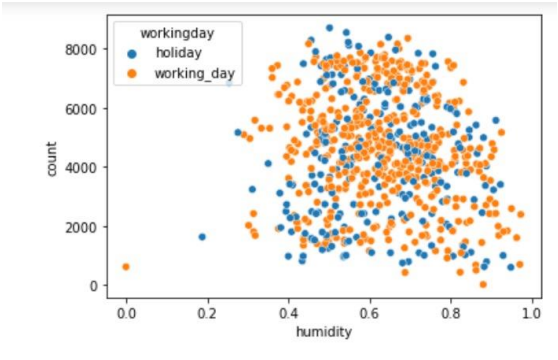


Figure 5a.humidity &count

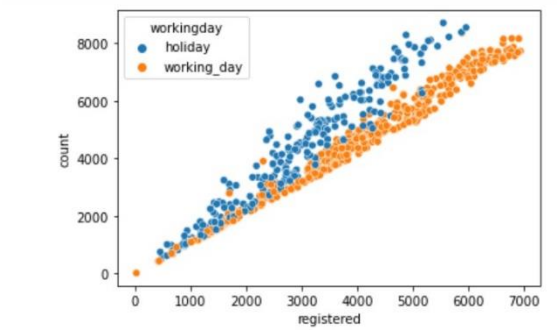


Figure 5b.registered &count

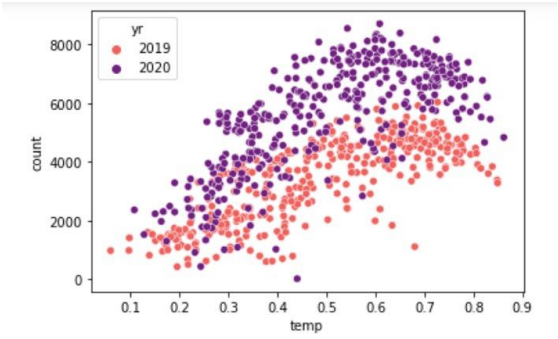


Figure 6. temp&count

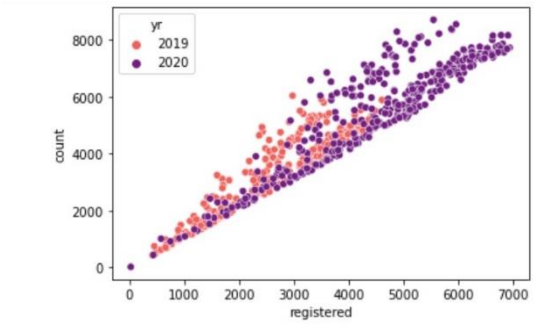


Figure 6a.registered &count

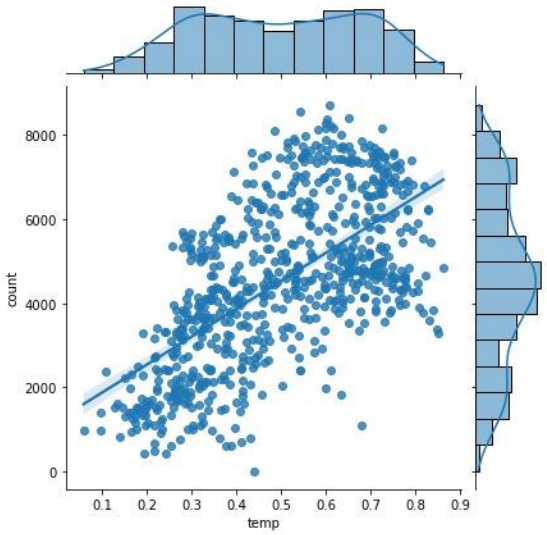


Figure 7 temp&count

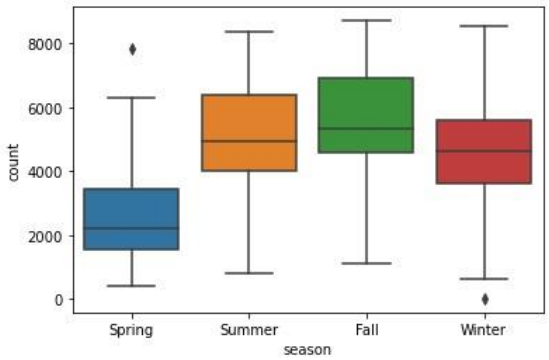


Figure 8.season &count

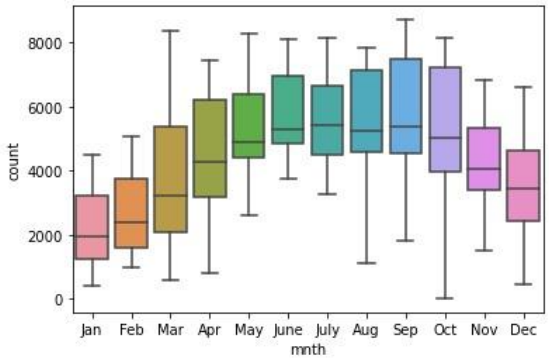


Figure 8a.mnth &count

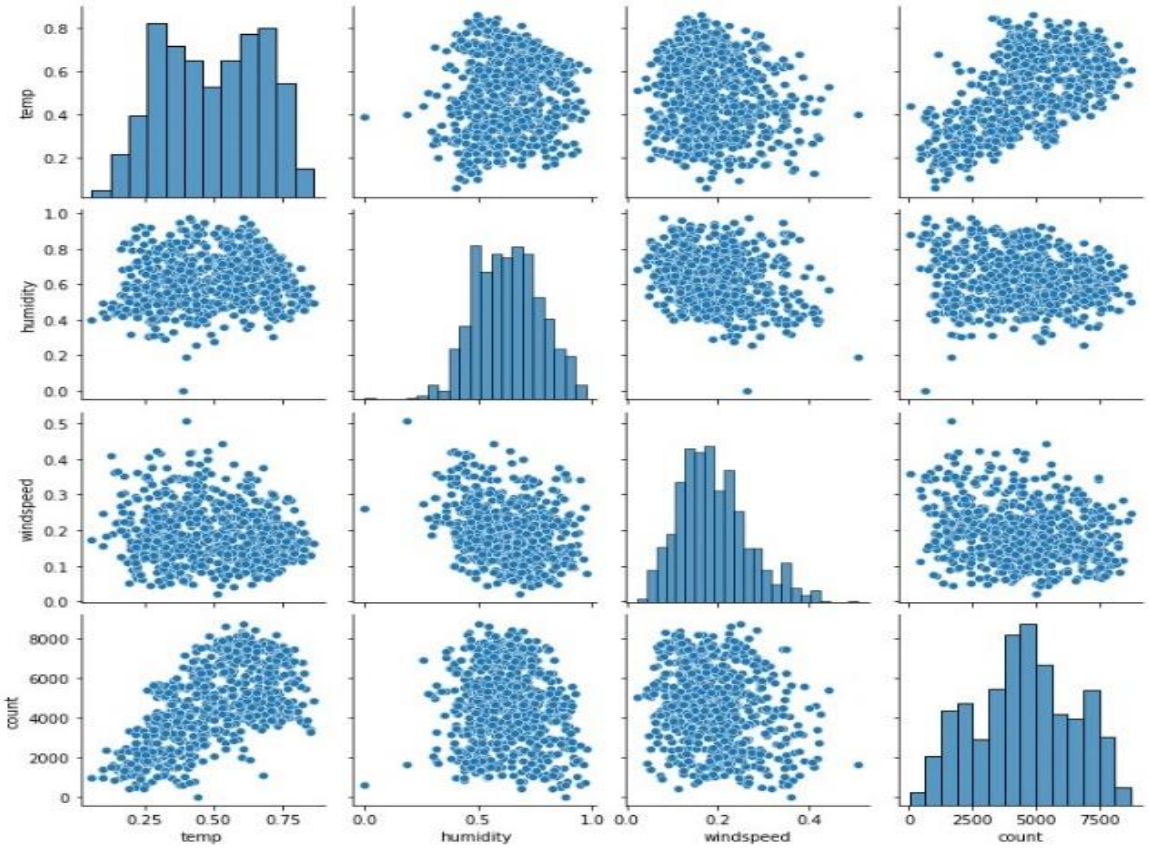


Figure 10.Pairplot(day.csv)

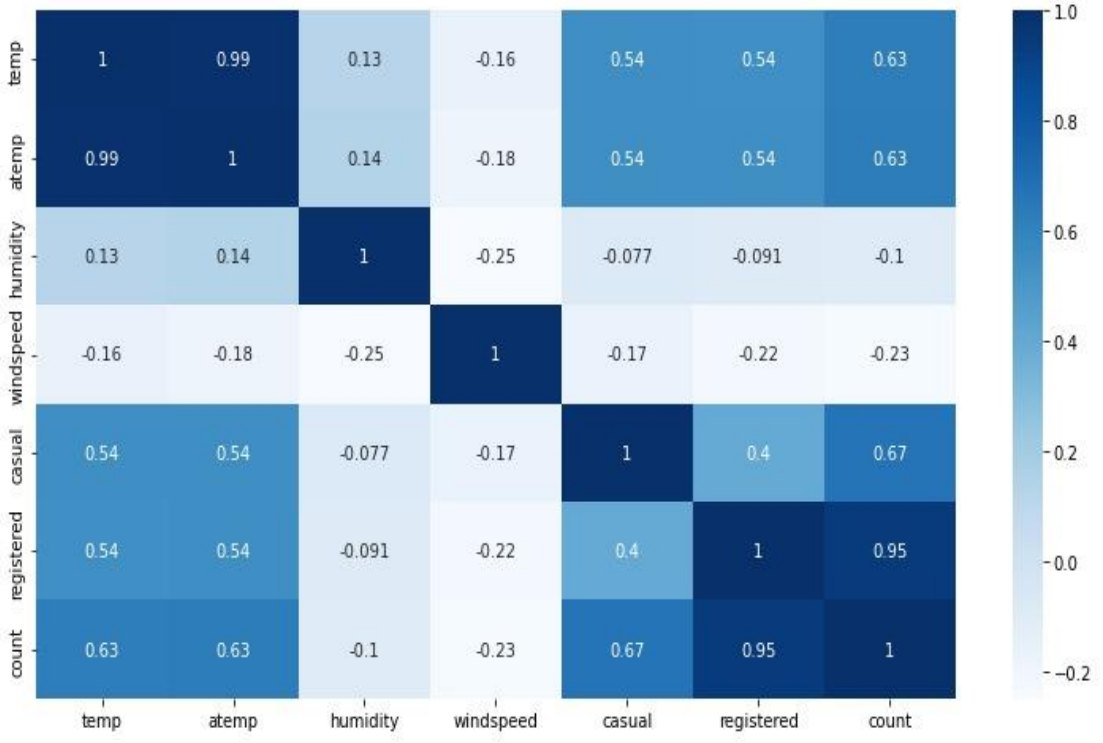


Figure 11.correlation matrix(day.csv)

Models Used :

a. Linear Regression Model

b. Decision Tree Regressor

c. Random Forest Regressor

a. Linear Regressor Model:

Linear regression is a fundamental statistical method used for modelling the relationship between a dependent variable (target or outcome variable) and one or more independent variables (predictor variables). It assumes a linear relationship between the independent variables and the dependent variable, meaning the relationship can be represented by a straight line.

b. *Decision Tree Regressor :*

Decision tree regression is a machine learning algorithm used for regression tasks. Unlike classification trees that predict categorical labels, decision tree regressors predict continuous numerical values. Here are the key points about decision tree regressors:

Objective: Predict a continuous target variable based on input features.

Model Structure: Decision tree regressors create a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a predicted numerical value. while a shallow tree may underfit the data.

Predictions: To predict a target value for a new data point, the algorithm traverses the tree from the root node to a leaf node based on the feature values of the data point.

c. Random Forest Regressor:

Random Forest Regressor is a versatile machine learning algorithm used for regression tasks. It belongs to the family of ensemble methods, combining multiple decision trees to make more accurate predictions.

Objective: Predict a continuous target variable based on input features.

Ensemble Method: Random Forest Regressor is an ensemble of decision trees. It builds multiple decision trees during training and averages their predictions to improve accuracy and robustness.

Hyperparameters: Important hyperparameters to tune include the number of trees (n_estimators), maximum depth of each tree (max_depth), minimum number of samples required to split a node (min_samples_split), and minimum number of samples required to be at a leaf node (min_samples_leaf).

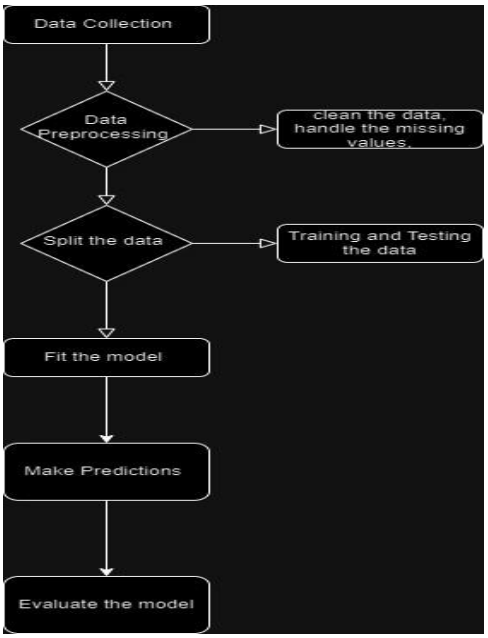
Feature Importance: Random Forests provide a measure of feature importance, indicating which features contribute the most to the model's predictions.

Random Forest Regressor is a powerful and widely-used algorithm due to its ability to handle complex datasets, reduce overfitting, and provide insights into feature importance. Proper tuning of hyperparameters and careful feature selection can further improve its performance.

V .Experimental Analysis :

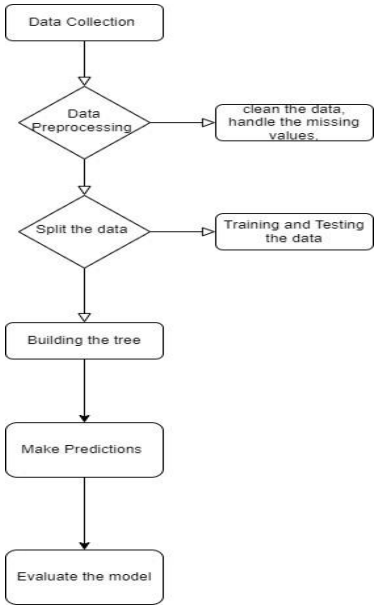
a. ***Linear Regression :*** The steps involved in to finding the accuracy and the r2 scores of the model

Type of Model	Models Used	Accuracy	R2 scores	Root Mean Score Error	Mean absolute Error
Regression	Linear Regression	0.816	0.800	782.40	594.50
Regression	Decision Tree Regression	0.808	0.729	925.79	667.81
Regression	Random Forest Regression	0.981	0.847	638.62	424.73



b. Decision Tree Regressor

These are the following steps to find out the decision tree regressor.



C. *Random Forest Regressor:*

These are the following steps to find out the Random Forest Regressor .

Hyperparameter tuning is a crucial step in machine learning to optimize the performance of your model. For a Random Forest Regressor, several hyperparameters can be tuned to achieve better results. Here's a guide on hyperparameter tuning specifically for a Random Forest Regressor:

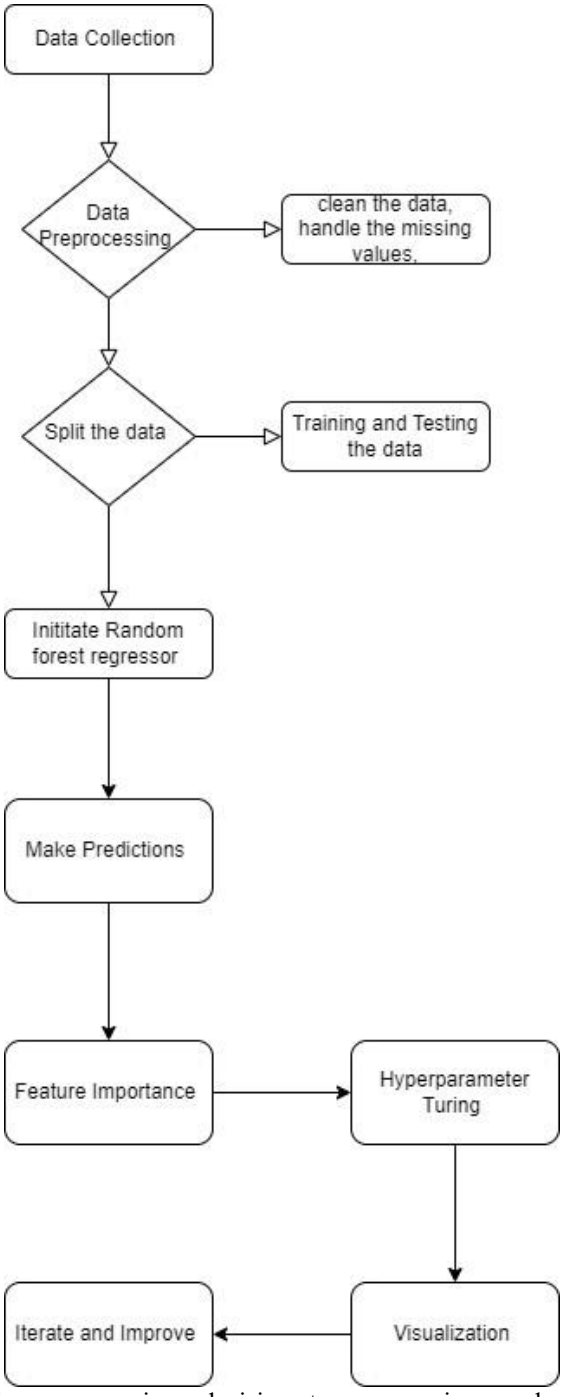
Define Hyperparameters to Tune:

the hyperparameters that significantly impact the performance of your Random Forest Regressor. Some key hyperparameters for Random Forests include:

n_estimators: Number of trees in the forest.
max_depth: Maximum depth of each tree.
min_samples_split: Minimum number of samples required to split an internal node.
min_samples_leaf: Minimum number of samples required to be at a leaf node.
max_features: Number of features to consider when looking for the best split.

VI. CONCLUSION :

This study aimed to develop predictive models using



linear regression, decision tree regression, and random forest regression for forecasting bike demand in a bike-sharing system. Key factors impacting bike demand, including temperature, season, month, weekday, and weather conditions, were identified as significant variables.

The model building process involved data cleaning, feature selection, and rescaling of features. Linear regression analysis was used to create a baseline model, capturing linear relationships between predictors and bike demand.

Decision tree regression was then employed to model non-linear relationships and capture complex interactions among variables. Finally, random forest regression, an ensemble method of decision trees, was utilized to improve prediction accuracy and handle overfitting.

The findings of this research have practical implications for bike-sharing system operators and urban transportation planners. By understanding the key factors influencing bike demand and comparing the performance of different regression models, operators can optimize resources, improve service quality, and enhance user satisfaction.

Additionally, this research contributes to promoting sustainable transportation solutions by encouraging the use of bike-sharing systems as eco-friendly alternatives to traditional transport modes. However, it's important to acknowledge the limitations of each model.

Linear regression assumes a linear relationship, while decision tree and random forest regressors may overfit if not properly tuned. Further research and validation are needed to ensure the generalizability of the findings across diverse geographical and cultural contexts.

Nevertheless, this study provides a comprehensive framework for predicting bike-sharing demand and offers valuable insights for practitioners, regulators, and researchers interested in the field of urban transportation planning and sustainable mobility.

VII. Future scope :

The future scope for bike-sharing demand forecasting encompasses several exciting areas that can further enhance the efficiency, sustainability, and user experience of bike-sharing systems. Here are some key future directions:

Smart Mobility Integration*:* Integrate bike-sharing systems with emerging smart mobility technologies such as Internet of Things (IoT), real-time data analytics, and smart city platforms. This integration can enable dynamic optimization of bike availability, routing algorithms, and user notifications based on real-time demand and traffic conditions.

Predictive Analytics and Machine Learning: Explore advanced predictive analytics techniques, including machine learning models such as deep learning, reinforcement learning, and ensemble methods, to enhance the accuracy and responsiveness of bike-sharing demand forecasts. Incorporate factors like user behavior patterns, event-based demand spikes, and environmental factors into the predictive models.

Demand-Responsive Operations*:* Implement demand-responsive operations strategies that dynamically adjust bike deployment, rebalancing efforts, and pricing structures based on predicted demand patterns. Utilize predictive analytics to optimize fleet management, minimize idle bikes, and improve overall system efficiency.

Multi-Modal Integration: Foster seamless integration of bike-sharing systems with other modes of transportation, such as public transit, ride-sharing services, and micro-mobility options (e.g., e-scooters). Develop interoperable platforms, mobile applications, and payment systems that facilitate multi-modal journeys and encourage sustainable transportation choices.

User-Centric Innovations: Focus on user-centric innovations by leveraging data-driven insights, user feedback mechanisms, and gamification elements to enhance the user experience. Introduce personalized recommendations, loyalty programs, and incentives that promote bike usage, address user preferences, and foster community engagement.

VIII.REFERENCES

[1]. Sathishkumar V E, Jangwoo Park, Yongyun Cho (2020), ‘Using data mining techniques for bike sharing demand prediction in metropolitan city’, The International Journal for the Computer and Telecommunications Industry.

[2]. Sathishkumar V E and Yongyun Cho (2020), ‘A rule based model for Seoul Bike sharing demand prediction using weather data’, European Journal of Remote Sensing.

[3]. ‘Bike Sharing: It’s about the community’, Cycling Industries Europe - https://cyclingindustries.com/fileadmin/content/documents/170707_Benefits_of_Bike_Sharing_UK_AI.pdf

[4]. Seoul Bike Sharing Demand Data Set, UCI Machine Learning Repository -

https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Shar Ing Demand

[5]. Hadi Faunae-T, and Joao Gama (2013), 'Event labelling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence.

[6]. Bike-Sharing Service Market Size 2021-2026, MarketWatch - https://www.marketwatch.com/press release/bike-sharing-service-market-market-size-2021 2026-comprehensive-study-development-status opportunities-future-plans-competitive-landscape-and growth-2021-01-11.

[7]. Data Source :http://data.seoul.go.kr/

[8]Y. Zhang, Z. Mi. Applied energy, 220, (2018)

[9]M. Ricci. Research in Transportation Business & Management, 15, (2015)

[10] P. Vogel, T. Greiser, D. C. Mattfeld. Procedia-Social and Behavioral Sciences, 20, (2011)

[11] T. D. Tran, N. Ovtracht, B. F. d'Arcier. Procedia Cirp, 30, (2015)

[12] kaggle datasets. Available at https://www.kaggle.com/datasets/gauravduttakiit/bike-sharing

[13]. D. B. Suits. Journal of the American Statistical Association, 52, 280(2012)

[14] S. Morgenthaler. Wiley Interdisciplinary Reviews: Computational Statistics, 1, 1 (2009)

[15] X. Su, X. Yan, C. L. Tsai. Wiley Interdisciplinary Reviews: Computational Statistics, 4, 3 (2012)

[16] X. W. Chen, J. C. Jeong. In Sixth international conference on machine learning and applications (ICMLA 2007) . IEEE (2007)

17]. R. M. O'brien. Quality & quantity, 41, (2007)

[18]. E. C. Alexopoulos. Hippokratia, 14, 23 (2010)

[19]. J. W. Osborne. Practical Assessment, Research, and Evaluation, 7, 2 (2000)

[20]. T. J. Smith,C. M. McKenna. Multiple Linear Regression Viewpoints, 39, 2 (2013)

GOOGLE COLAB LINK FOR THE WORKING CODE :

https://colab.research.google.com/drive/10sQMKxWMEmLnrSaFK-0PRa7jbvMLN96J?usp=sharing