Comprehensive Analysis and Predictive Modeling of Bike-Sharing Usage Patterns

Anand Gajaria

November 15, 2023

Contents

1 Introduction

Ride-sharing companies like Uber and Curb have undoubtedly revolutionized urban transportation, offering customers a convenient, affordable, and efficient alternative to traditional car ownership. However, as the number of automobiles on the road continues to rise, these car-based ride-sharing services may face challenges, particularly in densely populated areas like city centers. That's where bike-sharing emerges as a brilliant solution, providing people with an additional short-range transportation option. Bike-sharing allows individuals to navigate through congested streets without worrying about traffic jams, all while enjoying the cityscape and possibly even incorporating a bit of exercise into their journey.

The focus of this project is to delve into the bike-sharing rental data from "Capital Bikeshare," which has been serving Washington D.C. and its surrounding areas since 2010. When Capital Bikeshare first launched, it set the precedent for bike-sharing services in the United States, predating Uber's ride-sharing program by about 15 years. Starting with just 10 stations and 120 bicycles in Washington D.C., Capital Bikeshare has since expanded into an extensive bike-sharing network, boasting over 700 stations and 5,400 bicycles. This network not only serves the nation's capital but also extends it reach to other cities of United States of America.

My objective of the analysis is to find out the determining factor that drives the demand on bike share rentals, construct Machine Learning models and then try to make prediction on rentals based on the information I have. My exploration and the analysis of the data will be performed in Python.

2 Data Description

The dataset utilized in this study has been sourced from the UCI Machine Learning Repository. This particular dataset is focused on bike sharing rentals and comprises hourly records spanning from January 1, 2011, to December 31, 2012. In total, the dataset encompasses 17,379 individual records. Within this dataset, there are 15 independent features, with the hourly count of bike rentals serving as the dependent feature that I aim to predict.

Variable	Description
`datetime`	Hourly date in timestamp format
`season`	Integer values (1 to 4) representing seasons: 1 - Winter 2 - Spring 3 - Summer 4 - Fall
`holiday`	Boolean value (1 or 0) indicating whether it is a holiday
`workingday`	Boolean value (1 or 0) indicating whether it is a working day
`weather`	Integer values (1 to 4) for different weather conditions: 1 - Clear or cloudy 2 - Mists 3 - Light rain or snow 4 - Heavy rain, snow, or worse weather
`temp`	Temperature values at the given time
`atemp`	Feeling temperature values at the given time
`humidity`	Relative humidity levels (1 to 100) at the given time
`windspeed`	Wind speed values in mph (miles per hour)
`casual`	Count of non-registered user rentals across all stations
`registered`	Count of registered user rentals across all stations
`count`	Total count of rentals at the given hour across all stations.

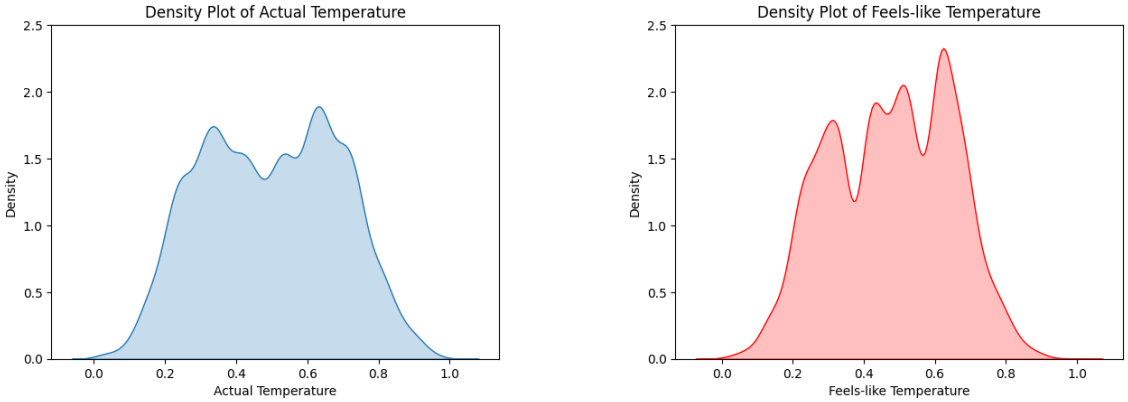
Table 1: Description of Variables in the Bike Sharing Dataset

3 Exploratory Data Analysis

This section is dedicated to the exploratory data analysis of the dataset, where I delve into a comprehensive examination of the data and explore relationships among its various features. My analysis encompasses both univariate and multivariate approaches, allowing me to gain insights into the dataset's characteristics, distributions, and interdependencies.

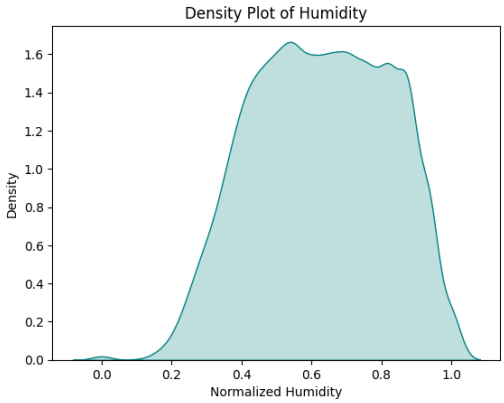
3.1 Univariate Analysis

3.1.1 Data Distribution of Actual and Feels-Like Temperature



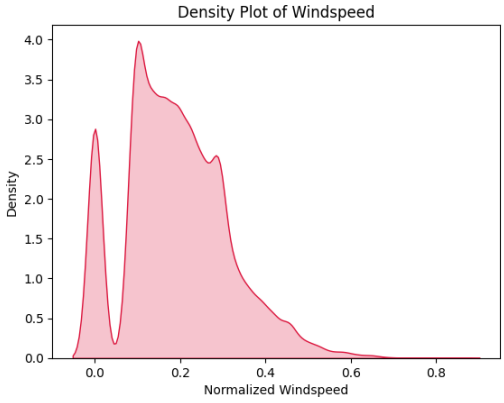
From the above graphs it can be observed that certain temperature ranges are preferred for bike rentals, possibly reflecting environmental variables such as seasonal changes or daily thermal cycles. A notable concentration of rentals occurs when temperatures fall within the 0.6 to 0.8 normalized range for both actual and 'feels-like' readings, implying a user preference for these conditions, possibly due to their association with optimal biking comfort. An additional, albeit less prominent, clustering is observed between 0.2 to 0.4, indicating a secondary range of favorability, perhaps during cooler periods. The acute peaks in 'feels-like temperature' suggest a more defined consensus among users regarding comfortable biking conditions, indicating a tighter clustering around certain perceived temperatures. On the other hand, the actual temperature data reveal a broader and more gentle distribution, reflecting a more varied set of temperatures at which users are inclined to rent bikes. This insight could be particularly useful , as it highlights the critical temperatures that drive rental demand.

3.1.2 Data Distribution of Humidity



The density plot of normalized humidity levels from the dataset indicates a preference for moderate humidity conditions among bike renters. With values stretching from 0 to 1, the graph peaks slightly above 0.6, signifying that the majority of bike rentals occur under these conditions. The bell-shaped curve suggests that both low and high extremes of humidity are less favorable for bike rentals, with a sharp decrease in rentals under high humidity, potentially due to discomfort or adverse weather conditions. This trend reflects a typical user tendency to avoid biking during excessively dry or damp weather, favoring a comfortable median humidity level for their rides.

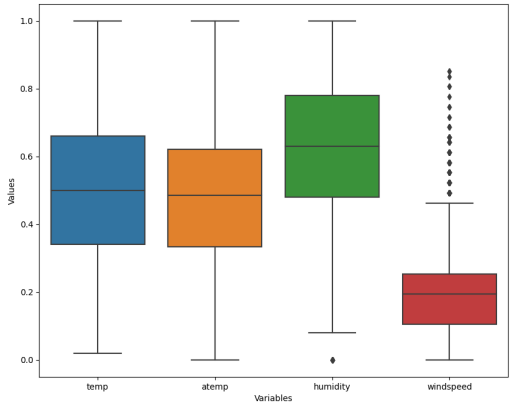
3.1.3 Data Distribution of Windspeed



The density plot for normalized windspeed from the dataset reveals a distinctive distribution, with the majority of bike usage occurring at lower wind speeds. The peak density around the lower quartile indicates that calm conditions are highly preferred by cyclists, aligning with the idea that high winds can be a deterrent to biking due to increased exertion and safety concerns. The plot also shows an unexpected spike near zero, which could represent exceptionally calm days. As the wind speed increases, the density sharply declines, suggesting a strong aversion to biking as conditions become windier. This pat-

tern underscores the significance of wind conditions on biking behavior and the potential influence of wind on the decision to rent bikes.

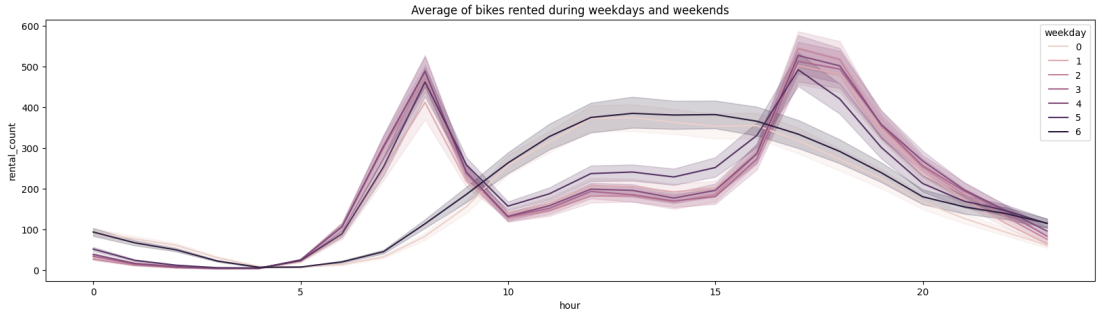
3.1.4 Boxplot of all the Continous variables



This box plot visualizes the distribution of four variables—temperature, feels-like temperature, humidity, and wind speed. Temperature and feels-like temperature show a similar distribution with median values around the 0.5 mark, suggesting moderate conditions are common. Notably, the feels-like temperature has a slightly tighter interquartile range (IQR), indicating less variability in how warm or cold it feels compared to the actual temperature. Humidity is also centered around the median value of 0.6 with a relatively small IQR, showing a less varied distribution. Wind speed, however, displays a much wider range and a number of outliers, as indicated by the dots beyond the whiskers. The outliers suggest that high wind speeds are less common but can vary greatly when they do occur. The distributions suggest that temperature and humidity in the data are relatively stable, while wind speed varies more significantly, which could have implications for bike rental frequency and user comfort.

3.2 Multi-variate Analysis

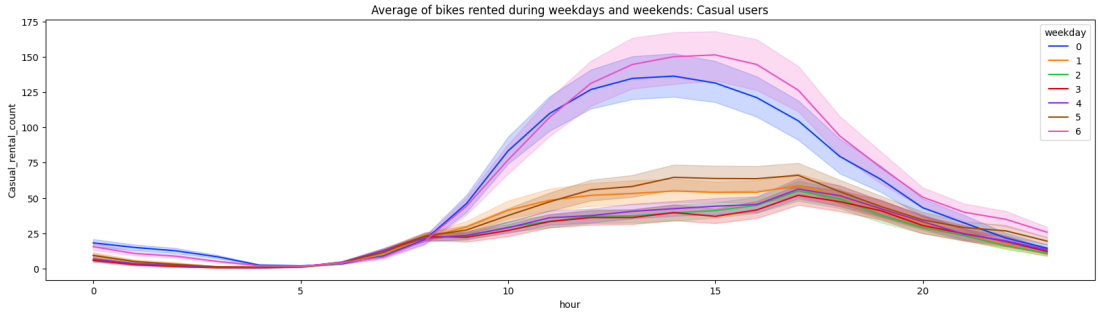
3.2.1 Bike Rentals on Weekdays and Weekends



The above line graph depicts the average count of bikes rented during different hours of the day, separated by weekdays and weekends. Each line corresponds to a different day of the week, with the legend indicating the specific day through numbers, where 0 stands for Sunday and 6 for Saturday.

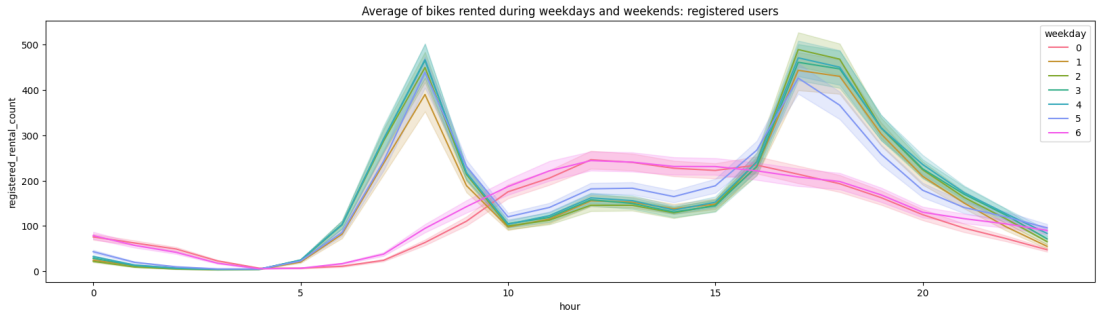
The graph shows two distinct patterns: during weekdays, there is a noticeable peak in bike rentals during the morning hours and another significant peak in the evening hours . On weekends, the pattern is different, showing a more gradual increase in rentals starting mid-morning, reaching the highest point in the early afternoon, and then declining towards the evening. This suggests a more recreational or leisurely use of bike rentals on weekends, as opposed to the commuter pattern observed during weekdays. The variation in rental counts across different hours of the day indicates the impact of daily routines and work schedules on bike rental habits

3.2.2 Bike Rentals on Weekdays and Weekends:Casual Users



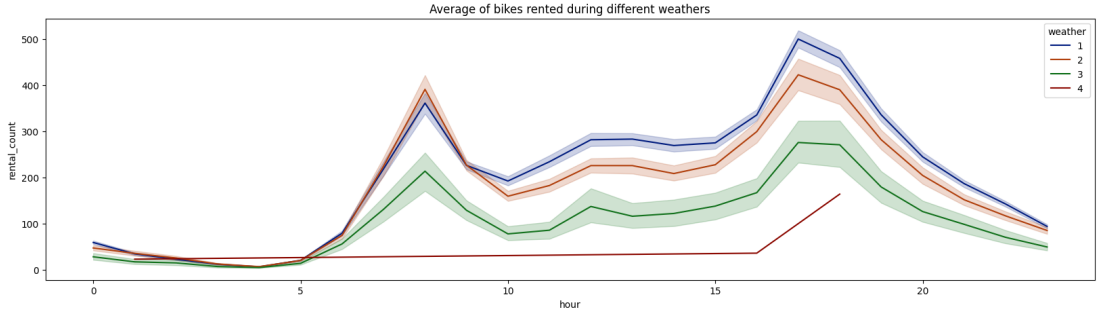
From the above graph it can be observed that unregistered users tend to rent bikes most frequently in the mid to late afternoon, particularly on weekends. Starting with Sunday, there's a noticeable peak in bike rentals during the early afternoon hours, which tapers off as the evening approaches. This trend is significantly higher on weekends compared to weekdays. Throughout the weekdays, the graph maintains a relatively stable and lower pattern of rentals with slight increases around midday, indicating a consistent but reduced level of casual usage.

3.2.3 Bike Rentals on Weekdays and Weekends:registered Users



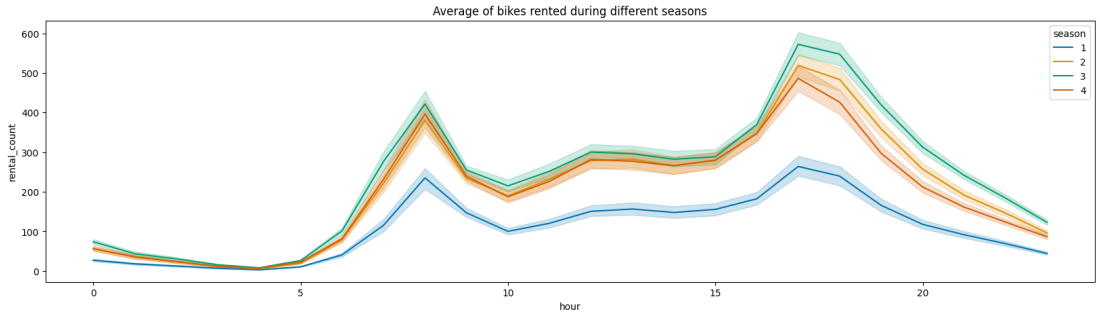
The above line graph shows the average number of bikes rented by registered users across different hours of the day for each day of the week. The graph indicates two prominent spikes in rentals during weekdays: one in the morning hours and another in the evening, suggesting a pattern consistent with commuting to and from work or school. In contrast, the weekend days display a different pattern, with rentals gradually increasing throughout the morning, peaking in the early afternoon, and then tapering off through the evening. Overall, the graph implies that registered users' rental patterns on weekdays are dictated by work or school schedules, while weekend use is more recreational.

3.2.4 Bike Rentals during different weathers



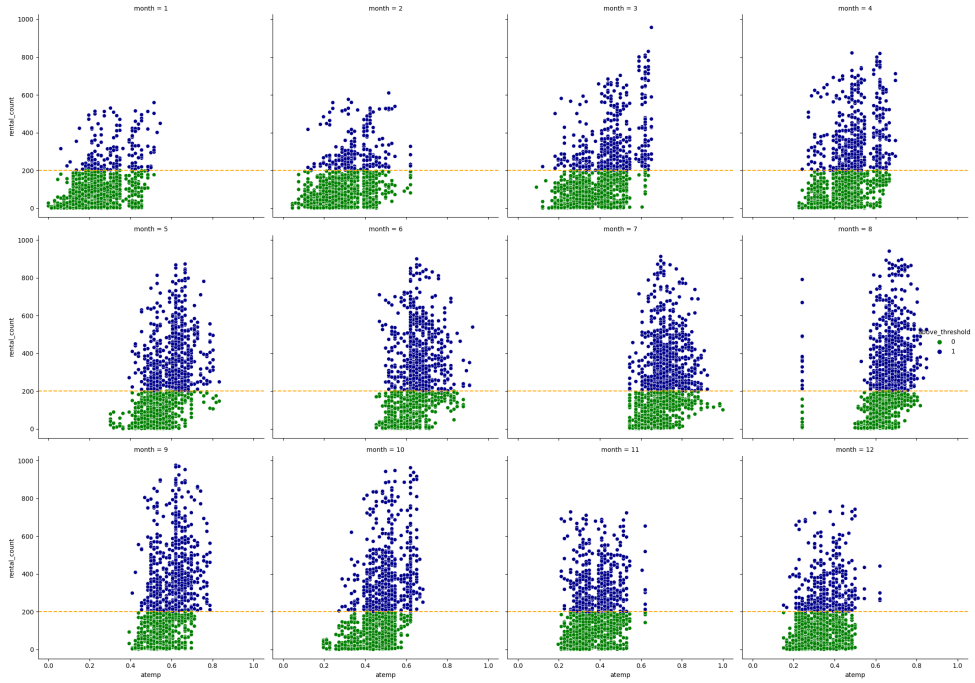
The above line graph depicts the relationship between different weather conditions and the average number of bike rentals throughout the day. Clear skies (1) correlate with the highest rental numbers, showing two pronounced peaks during typical morning and evening commute times, suggesting that users prefer to bike when the weather is favorable. As weather conditions become less ideal, from mist (2) to light precipitation (3), there's a noticeable decrease in rentals, especially during commute hours. In the presence of heavy precipitation (4), there's a significant drop in bike rentals, with the demand curve flattening and shifting downwards throughout the day. This trend illustrates a strong preference for biking in clear weather, while precipitation, especially heavy, is a deterrent, likely due to safety and comfort considerations.

3.2.5 Bike Rentals during different weathers



The above line graph presents the variation in bike rentals throughout different hours of the day across the four seasons, denoting a clear influence of seasonal changes on rental habits. In winter (1), bike usage is notably lower, with subdued peaks that suggest only the hardiest of riders braving the cold. As the seasons progress to spring (2) and then summer (3), there is a marked increase in rental activity, peaking during summer afternoons and evenings, reflecting the optimal conditions for both commuting and leisure rides. The pattern begins to taper off in fall (4), still mirroring summer's shape but with reduced frequency, as cooler temperatures and diminishing daylight hours gradually deter riders.

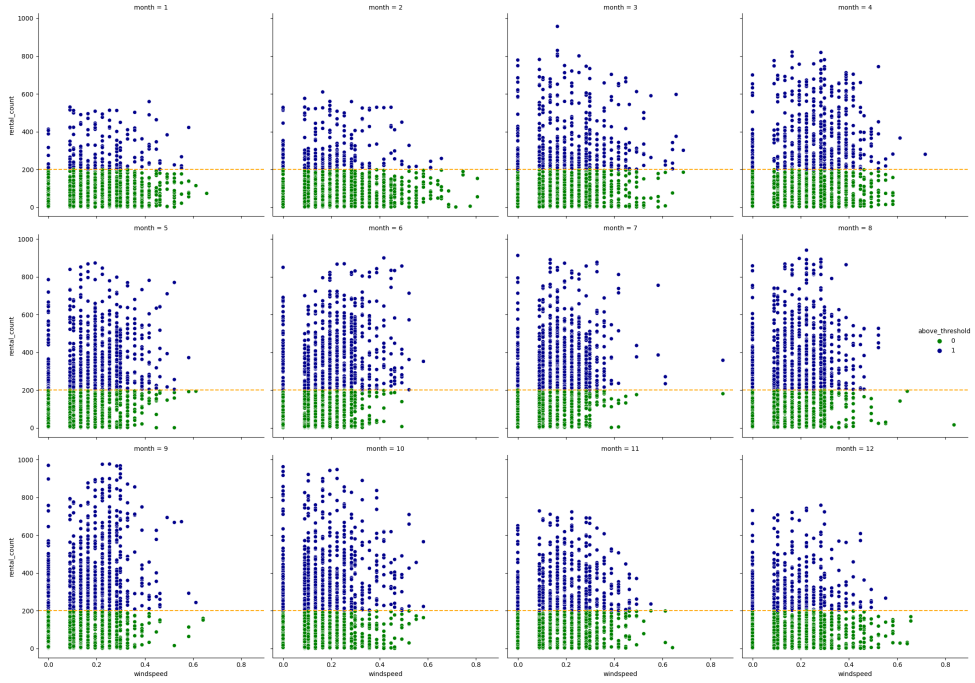
3.2.6 Analysing the Bike rental demand with change in Feels-like temperature



The scatter plot illustrates the relationship between the perceived temperature (atemp) and bike rental counts across different months, with blue points indicating above-average rental counts and green points indicating below-average counts, relative to an overall average threshold denoted by the yellow line. Observing the data, there is a discernible

increase in bike rentals with higher feels-like temperatures, evident from the density of blue points as the atemp value increases. This trend is more pronounced during the warmer months (May to September), where we see a higher concentration of blue points above the threshold line, implying more frequent above-average rentals. Conversely, the colder months (January to April and October to December) show a dominance of green points, especially at lower temperature values, suggesting that cooler weather correlates with a decline in bike rentals. The yellow line serves as a reference for the average rental count, and it's clear that during periods of moderate temperatures, the counts fluctuate around this average. The data points also exhibit a wide spread, indicating variability in rental counts that could be attributed to factors other than temperature, such as day of the week, holidays, or other weather conditions. Overall, the demand for bike rentals exhibits a strong positive correlation with warmer perceived temperatures, with significant seasonal variations.

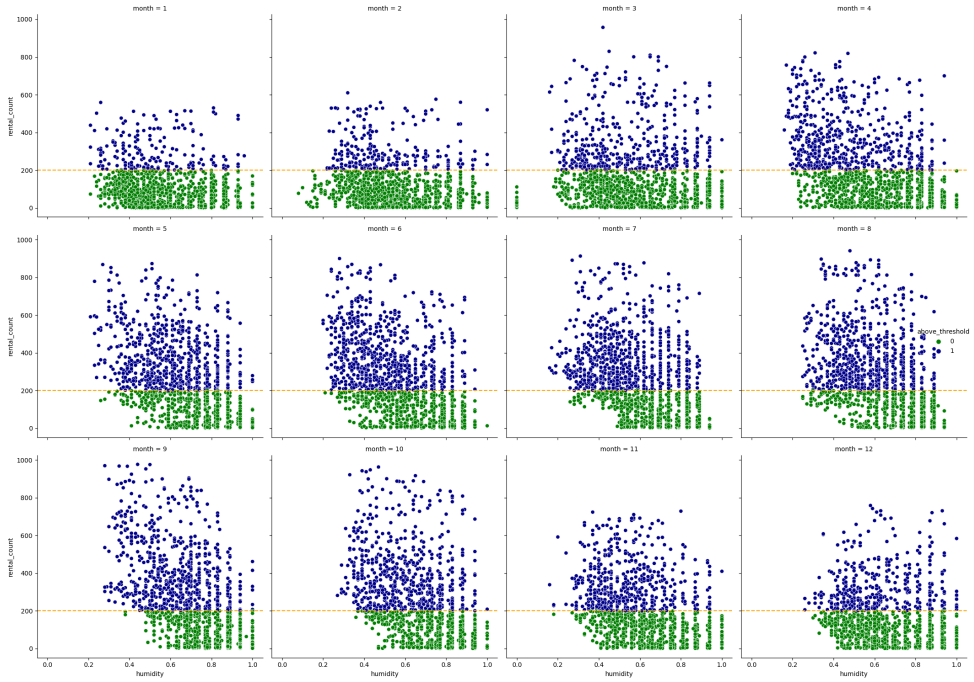
3.2.7 Analysing the Bike rental demand with change in Windspeed



The scatter plot showcases the influence of wind speed on bike rental demand across various months. The distribution of points suggests that as wind speed increases, the number of bike rentals generally decreases. This trend is consistent across all months, with a higher concentration of green points (below-average rentals) in regions of higher wind speed. However, during certain months, particularly in the warmer seasons such as May to August, there are still a notable number of blue points (above-average rentals) even at higher wind speeds, which may indicate that the pleasant temperatures offset the negative impact of wind on cyclists' comfort. Conversely, during colder months like January and December, the density of blue points is visibly lower at increased wind speeds, implying that cold wind may discourage bike rentals more strongly than warmer wind. Overall, the plot indicates a negative correlation between wind speed and bike rental demand, with the

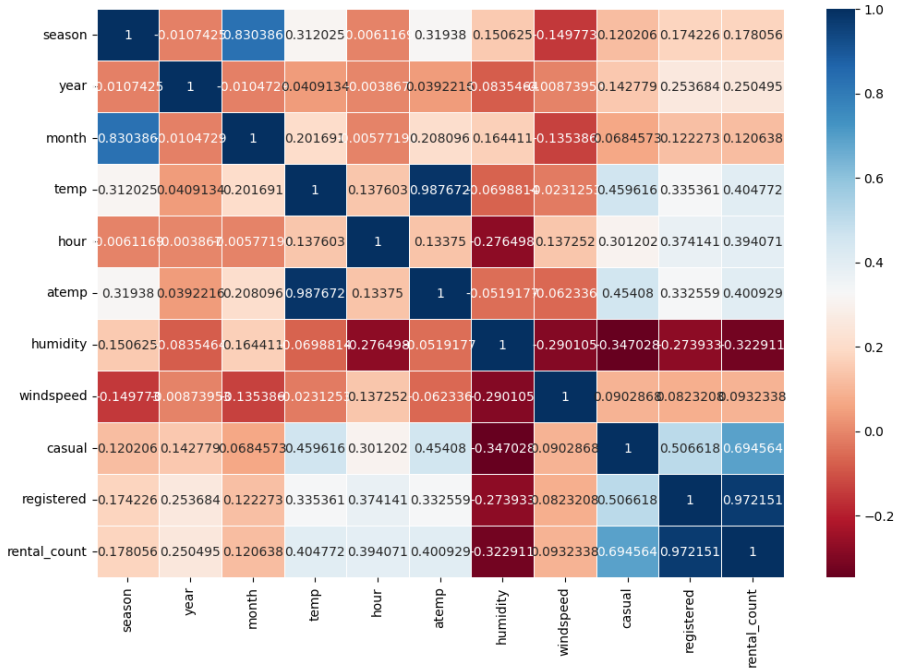
strength of this relationship potentially moderated by the seasonal context.

3.2.8 Analysing the Bike rental demand with change in Humidity



This scatter plot represents the relationship between humidity levels and bike rental demand over various months. A distinct pattern emerges across the graph where higher humidity levels correlate with a lower frequency of above-average rentals. This trend is illustrated by a greater concentration of green points at higher humidity levels across all months. Particularly in months with more extreme temperatures, such as January, February, and December, high humidity seems to have a more pronounced negative effect on bike rentals. During the warmer months from May to September, despite the presence of high humidity, there is still a substantial number of blue points, suggesting that warmer temperatures might mitigate the deterrent effect of high humidity to some extent. The data also exhibits a decline in bike rentals as humidity approaches very high levels, which could be indicative of uncomfortable weather conditions for biking. Overall, the demand for bike rentals shows an inverse relationship with humidity, although the intensity of this relationship can vary with seasonal changes.

3.2.9 Correlation Matrix



The correlation heatmap analysis for bike rentals reveals a strong positive correlation between rental counts and registered users, indicating that regular users are the backbone of rental frequency. Conversely, rental counts have a negative correlation with humidity and windspeed, suggesting adverse weather conditions deter bike usage. Temperature variables show a positive correlation, highlighting that pleasant weather boosts rentals. Time of day is also positively correlated, reflecting peak rental times that may align with commuting patterns. Seasonal factors such as month and season display a moderate negative correlation, possibly due to less favorable weather in certain times of the year. The negligible negative correlation with year suggests stable rental patterns over the time period analyzed. Overall, user type, weather conditions, and time are key influencers of bike rental dynamics.

4 Modeling

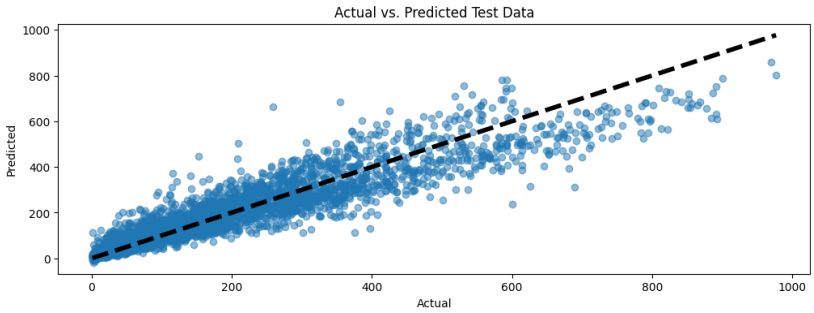
4.1 Model Selection

LightGBM is particularly effective for the UCI Bike Sharing dataset due to its proficiency in handling diverse data types, an essential capability given the mix of categorical and numerical inputs within the dataset. Its advanced tree-based algorithms efficiently handle the non-linear relationships and complex interactions between features, such as weather conditions and rental patterns. LightGBM's gradient boosting approach incorporates techniques like Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), which enhance its ability to deal with large datasets by optimizing memory usage and reducing computational time—key advantages when processing the extensive records typical of bike sharing systems. The model's fine-tuning capabilities, through hyperparameter optimization, allow for custom tailoring to the dataset's unique demands. Post-training, LightGBM offers insightful feature importance metrics, which are instrumental in understanding the driving factors behind bike rental frequencies, aiding in both predictive accuracy and strategic decision-making for bike fleet management.

4.2 Model Training

In the process of configuring the LightGBM model for the Bike Sharing dataset, I started with the feature set, which encompasses categorical variables like 'season', 'month', 'hour', 'holiday', 'weekday', and 'workingday', alongside continuous variables such as 'atemp', 'humidity', and 'windspeed'. With the features prepared, I split the data into distinct training and testing sets, a measure taken to validate the model's efficacy on data it has not encountered before. Initial hyperparameters, including the learning rate and metrics for model evaluation, were set up with the intent to fine-tune them through LightGBM's iterative training process. This approach aids in mitigating overfitting and in fine-tuning the model by optimizing the hyperparameters to enhance the model's predictive accuracy. During training, the model underwent numerous rounds, guided by the loss metrics on the validation set, to determine the best iteration.

4.3 Model Evaluation



The scatter plot visualizes the predictive performance of LightGBM regression model on Test data, with actual values on the x-axis and predicted values on the y-axis. The

concentration of data points around the dashed trend line indicates a general alignment between predictions and actual values, signaling a decent model fit. Notably, the plot reveals a tendency for the model to underpredict at higher actual values, as evidenced by the data points trailing below the trend line in the upper region of the graph. This pattern suggests that while the model predicts moderately well across most of the value range, its accuracy diminishes for higher actual values, which could be a focal point for further model refinement.

4.4 Model Statistics

Metric	Train Data	Test Data
RMSE	62.3	65.1
R^2 Score	0.88	0.87
MAD	43.2	44.6

Table 2: Model Performance Statistics

The above model statistics indicate a reasonably strong predictive performance, as evidenced by the R^2 scores and the error metrics. The R^2 score, which measures the proportion of the variance in the dependent variable that is predictable from the independent variables, is 0.88 for the training data and 0.87 for the test data. These values are quite close, suggesting that the model generalizes well to unseen data. The Root Mean Squared Error (RMSE), which quantifies the average squared root difference between the predicted values and the actual values, is slightly higher for the test data (62.3) compared to the training data (65.1). This slight increase in RMSE from training to testing is expected and indicates a small degree of overfitting, but it is not significant enough to raise major concerns about the model's generalizability. The Mean Absolute Deviation (MAD) values, measuring the average absolute difference between the predicted and actual values, are 43.2 for the training data and 44.6 for the test data, again showing a modest increase in the test data but remaining relatively close to the training performance. Overall, these statistics demonstrate that the model has achieved a good balance between fitting the training data and generalizing to new data, with a relatively high degree of predictive accuracy.

5 Scaling Properties of LightGBM

- **Histogram-based Splitting:** LightGBM uses histogram-based algorithms for decision tree learning which groups continuous feature values into discrete bins. This reduces memory consumption significantly and speeds up the training process because the algorithm operates on bin counts rather than on individual data points.
- **Efficient Use of Memory:** LightGBM does not need to load the entire dataset into memory. It can handle data that is stored on disk, which is essential for dealing with several terabytes of data, as this would not fit into the main memory of most systems.
- **Parallel and Distributed Computing:** LightGBM supports multi-core parallel processing to speed up training and can also be run in a distributed manner across a cluster of machines, which is crucial for handling terabyte-sized datasets.
- **Network Communication Optimization:** During distributed learning, LightGBM reduces the amount of data that needs to be communicated between nodes by aggregating histograms. This is a critical feature for scalability as network communication can become a bottleneck in distributed systems.

6 Probable problems related to these properties

6.1 Parallel and Distributed Computing:

6.1.1 Resource Management and Scalability:

- Efficiently managing and scaling computational resources across multiple machines or clusters.
- Inefficient resource utilization can lead to increased costs and reduced performance.

6.1.2 Data Partitioning and Load Balancing::

- Ensuring effective data partitioning and load balancing across different nodes.
- Imbalanced loads can lead to some nodes being overburdened while others are underutilized, affecting overall efficiency.

6.2 Network Communication Optimization:

6.2.1 Communication Overhead:

- In distributed training, there's significant overhead associated with communication between nodes.
- Excessive communication can negate the benefits of parallel processing, especially for algorithms requiring frequent data exchange.

6.2.2 Synchronization Costs:

- Keeping the model updates synchronized across different nodes.
- Synchronization issues can lead to inconsistent model states and affect convergence and performance.

6.3 Histogram-based Splitting:

6.3.1 Loss of Precision:

- Binning continuous features into histograms can result in a loss of granularity.
- This loss of precision might affect the model's ability to find the most effective splits, potentially impacting accuracy.

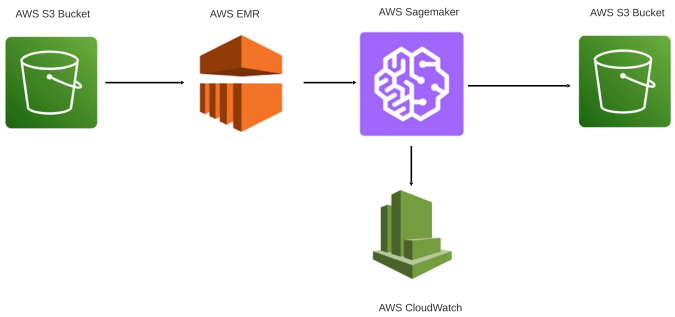
6.3.2 Handling Sparse and Skewed Data:

- Histogram-based methods can struggle with features that have sparse or highly skewed distributions.

- Can lead to inefficient bin usage and potentially impact model performance.

7 Probable solutions to these problems

Amazon Web Services (AWS) can offer a comprehensive solution to the challenges associated with large-scale machine learning projects, like those involving LightGBM. AWS provides a wide array of services that can be leveraged for distributed computing, efficient data storage and processing, and machine learning.



- **Data Storage and Collection: Amazon S3**
  - Store raw data in Amazon S3. It's a scalable and secure object storage service.
  - Reason: S3 is ideal for storing large datasets and can serve as a central repository for all data.
- **Data Preprocessing: Amazon EMR**
  - Use use Amazon EMR with frameworks like Apache Spark for preprocessing and handling large-scale data.
  - Reason: Amazon EMR is optimized for big data processing. This combination is effective for transforming and preparing data for training.
- **Model Training: Amazon SageMaker**
  - Use Amazon SageMaker for training LightGBM model. SageMaker supports distributed training and offers various instance types suitable for different scales of data.
  - Reason: SageMaker simplifies the machine learning workflow and provides tools for easy scaling, monitoring, and optimization of model training.
- **Hyperparameter Tuning: Amazon SageMaker Automatic Model Tuning**
  - Utilize the Automatic Model Tuning feature in SageMaker to optimize your LightGBM model's hyperparameters.

  - Reason: This service automates the complex task of hyperparameter tuning, saving time and improving model performance.

- **Model Deployment: Amazon SageMaker Endpoints**

  - Deploy the trained model using Amazon SageMaker Endpoints for real-time or batch predictions.
  - Reason: SageMaker Endpoints facilitate easy deployment of models and provide a scalable and managed environment for inference.

- **Monitoring and Logging: Amazon CloudWatch**

  - Use Amazon CloudWatch for monitoring the performance of your machine learning models and the health of your AWS resources.
  - Reason: CloudWatch provides insights into operations and performance, allowing for proactive management and optimization.

- **Amazon SageMaker Model Registry**

  - Use Amazon sagemaker model registry for for model versioning and management.

8 Key Challenges with AWS

- **Complexity:** The vast AWS ecosystem requires significant expertise to manage, posing challenges for small teams or those without AWS knowledge.
- **Vendor Lock-in:** Dependence on AWS-specific tools can limit flexibility and complicate migration to other platforms.
- **Performance Optimization:** Optimal performance demands resource and configuration tuning, impacting costs and efficiency if not managed well.
- **Integration with Existing Systems:** Integrating AWS with on-premises systems can be complex and may increase costs and complexity.

9 Personal Experience

- **AWS Sagemaker:** I have 2 years of experience on working with AWS Sagmekaer. I used it for developing and deploying machine learning models.
- **AWS EMR:** I have 1 year of expericence on working AWS EMR. I utilized it Data preprocessing, Tranformation and Analysis