**A SYMPHONY OF INSIGHTS: PREDICTIVE MODELLING FOR BIKE RENTAL DEMANDS.**

**Introduction:**

* The growing trend of urbanization has led to an increased demand for efficient and sustainable transportation solutions. Among these solutions, bike-sharing systems have emerged as a viable alternative to traditional transport methods. By providing easy access to bicycles, these systems not only alleviate traffic congestion but also promote a healthier lifestyle among urban dwellers.
* Understanding the dynamics of bike rental demand is essential for the effective management of these systems. Predictive modelling offers a powerful approach to analysing historical data and identifying the factors that influence bike usage. This project seeks to create a predictive model for bike rentals, taking into account various variables such as weather conditions, time of day, and seasonal patterns.
* Utilizing machine learning techniques, specifically Linear Regression and Random Forest algorithms, this project will analyse historical bike rental data to forecast demand accurately. The insights gained from this analysis will assist city planners and bike-sharing operators in optimizing their services, enhancing user satisfaction, and promoting the use of sustainable transportation options.

**Objectives:**

The primary objectives of this project are as follows:

* **Understand Demand Drivers:** This objective focuses on identifying and analysing the key factors that influence bike rental demand. Factors such as weather conditions, including temperature and humidity, seasonal variations, time of day, and special events will be explored. By understanding these drivers, we can gain insights into how they affect the number of bike rentals, which is crucial for effective planning and resource allocation.
* **Develop Predictive Models:** The project aims to create and evaluate various machine learning models to accurately forecast bike rental counts. Specifically, Linear Regression and Random Forest models will be implemented to assess their predictive capabilities. This objective emphasizes the importance of selecting the right algorithms to ensure reliable predictions.
* **Assess Model Performance:** A critical aspect of this project involves measuring the effectiveness of the predictive models. We will employ performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared Score to determine which model yields the best predictions. This objective underscores the need for robust evaluation techniques to validate the models.
* **Visualize Data Relationships:** This objective highlights the use of visual tools to represent relationships between variables and the distribution of bike rentals. By utilizing correlation heatmaps, scatter plots, and boxplots, we can better understand how different factors interact and influence bike rental demand. Visualization will facilitate a clearer interpretation of the data and results.
* **Provide Recommendations:** Based on the analysis and model results, the project will offer actionable insights and recommendations for bike-sharing operators. These recommendations will focus on optimizing fleet management and improving service availability during peak demand periods. This objective aims to translate data insights into practical strategies for enhancing operational efficiency.
* **Explore Limitations and Future Work:** Finally, this project will discuss the limitations encountered during the analysis and modelling process. By acknowledging these limitations, we can provide suggestions for potential improvements or areas for future research in bike-sharing demand forecasting. This objective ensures that the study remains relevant and sets the stage for continued exploration in this field.

**Data Overview**

The dataset utilized for this project comprises two primary files: **hour.csv** and **day.csv**. Each dataset contains various features that contribute to understanding bike rentals across different time periods.

**1. Hourly Data (hour.csv)**

* **Description**: This dataset records bike rental counts on an hourly basis, allowing for a detailed analysis of rental patterns throughout the day.
* **Key Features**:
  + instant: Unique identifier for each record.
  + dteday: Date of the record.
  + season: Categorical variable representing the season (1: winter, 2: spring, 3: summer, 4: fall).
  + yr: Year of the record (0: 2011, 1: 2012).
  + mnth: Month of the record (1 to 12).
  + hr: Hour of the day (0 to 23).
  + holiday: Indicates if the day is a holiday (0: no, 1: yes).
  + weekday: Day of the week (0: Sunday, 6: Saturday).
  + workingday: Indicates if the day is a working day (0: no, 1: yes).
  + temp: Normalized temperature in Celsius (ranging from 0 to 1).
  + atemp: Normalized feeling temperature in Celsius.
  + humidity: Normalized humidity (ranging from 0 to 1).
  + windspeed: Normalized wind speed.
  + casual: Number of casual users.
  + registered: Number of registered users.
  + cnt: Total bike rentals (casual + registered).

**2. Daily Data (day.csv)**

* **Description**: This dataset summarizes bike rentals on a daily basis, providing insights into broader trends over time.
* **Key Features**:
  + instant: Unique identifier for each record.
  + dteday: Date of the record.
  + season: Categorical variable representing the season.
  + yr: Year of the record (0: 2011, 1: 2012).
  + mnth: Month of the record.
  + holiday: Indicates if the day is a holiday.
  + weekday: Day of the week.
  + Working day: Indicates if the day is a working day.
  + temp: Normalized temperature in Celsius.
  + atemp: Normalized feeling temperature in Celsius.
  + humidity: Normalized humidity.
  + windspeed: Normalized wind speed.
  + casual: Number of casual users.
  + registered: Number of registered users.
  + cnt: Total bike rentals.
  + total: Total rentals for the day.

**3. Data Characteristics**

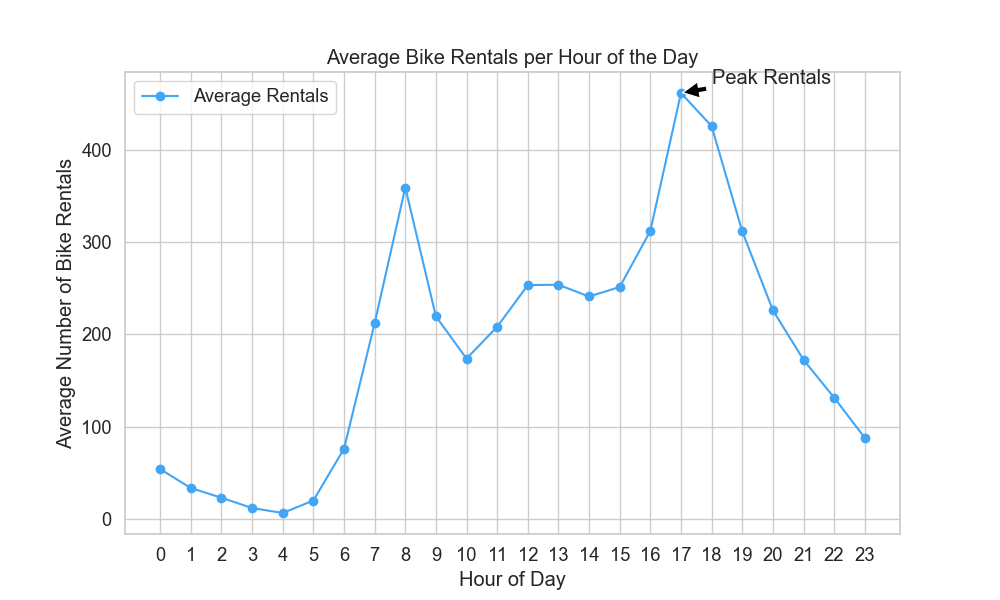
* **Size**: The **hourly dataset** consists of over 17,000 records, while the **daily dataset** contains approximately 7,000 records.
* **Temporal Coverage**: The data spans from January 2011 to December 2012, offering insights into bike rental trends over two full years.
* **Data Quality**: The dataset appears to be clean with minimal missing values, as indicated during the exploratory data analysis.

**Exploratory Data Analysis (EDA)**

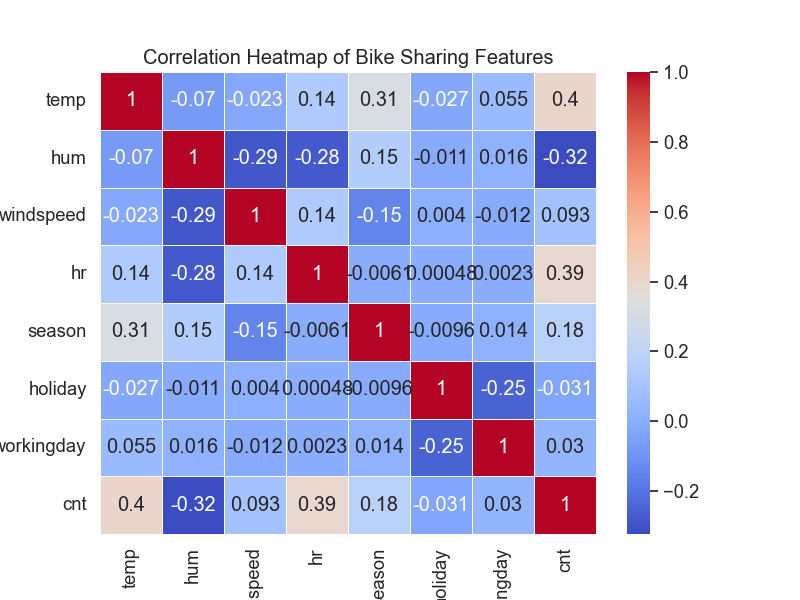
* Exploratory Data Analysis (EDA) is a critical step in understanding the characteristics and patterns within the dataset. In this analysis, we will examine various visualizations that provide insights into bike rental trends based on different features. This process helps in identifying relationships, anomalies, and key factors influencing bike rentals.

**1. Average Bike Rentals by Hour**

* **Graph**: A line graph displaying average bike rentals across different hours of the day.
* **Explanation**: This graph illustrates how bike rentals vary by hour. Typically, bike usage peaks during morning and evening commute hours (around 8 AM and 5 PM) and is lower during the night. This trend suggests that bike rentals are heavily influenced by daily commuting patterns, which can guide resource allocation for peak hours.



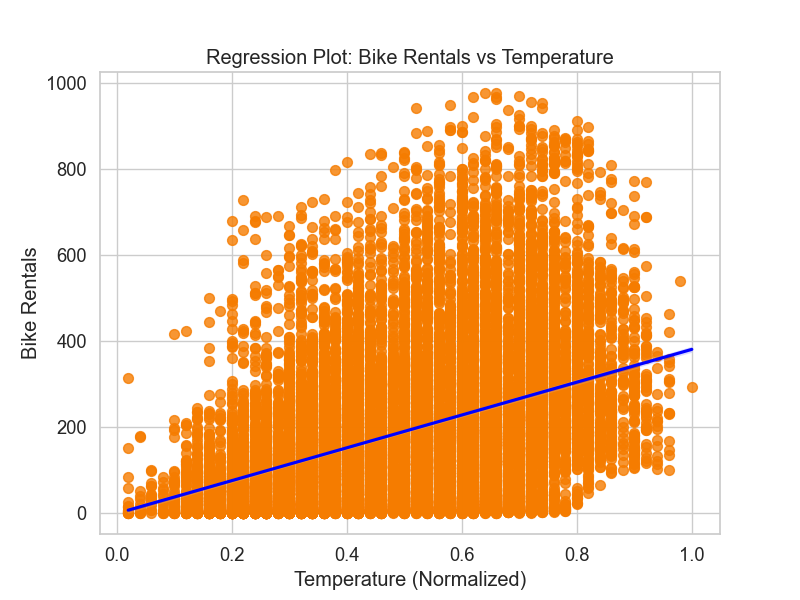
**2. Correlation Heatmap**

**Graph**: A heatmap showing the correlation between various features in the dataset.

**Explanation**: The heatmap indicates the strength of relationships between features. For instance, we may observe a strong positive correlation between temperature and bike rentals, suggesting that higher temperatures lead to increased bike usage. Conversely, wind speed may show a negative correlation with rentals, indicating that higher winds could deter bike users. Understanding these correlations is vital for predictive modelling

**3. Regression Plot: Rentals vs. Temperature**

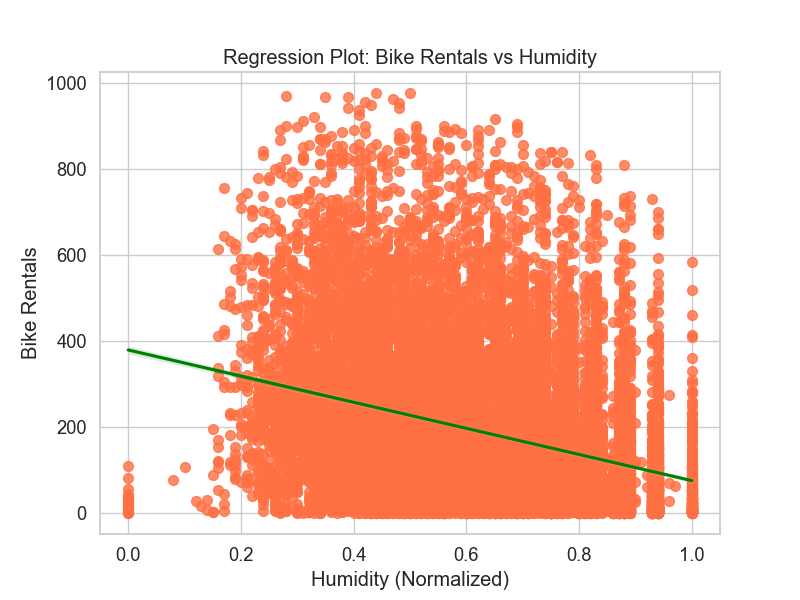
**Graph:** A regression plot illustrating the relationship between bike rentals and temperature.

****

**Explanation:** The regression line indicates a positive trend, highlighting that as the temperature increases, the number of bike rentals also tends to rise. This is an essential finding, suggesting that warmer weather encourages more people to rent bikes, likely due to the comfort and appeal of biking in pleasant conditions.

**4. Regression Plot: Rentals vs. Humidity**

**Graph:** A regression plot showing the relationship between bike rentals and humidity.

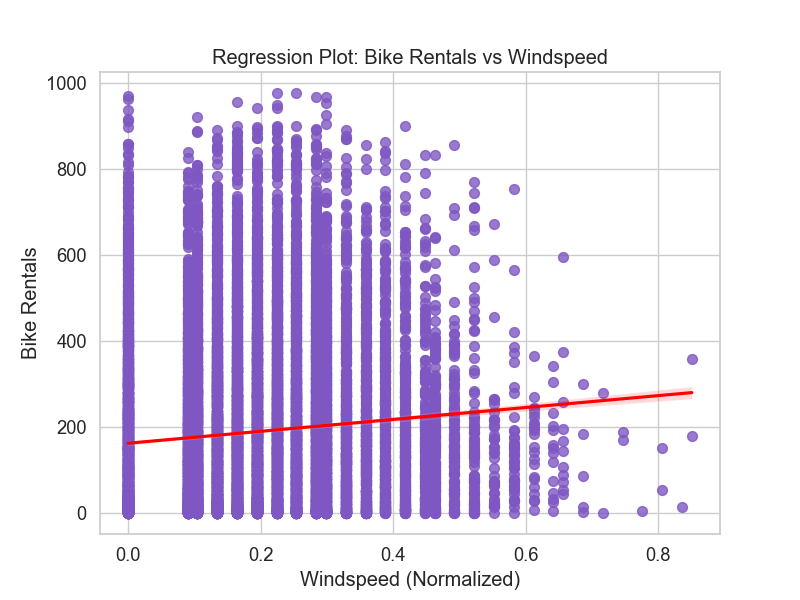
****

**Explanation:** This plot may reveal a slight negative correlation, indicating that higher humidity levels could reduce bike rentals. High humidity might deter users from biking due to discomfort, suggesting that weather conditions play a significant role in biking habits.

**5. Regression Plot: Rentals vs. Windspeed**

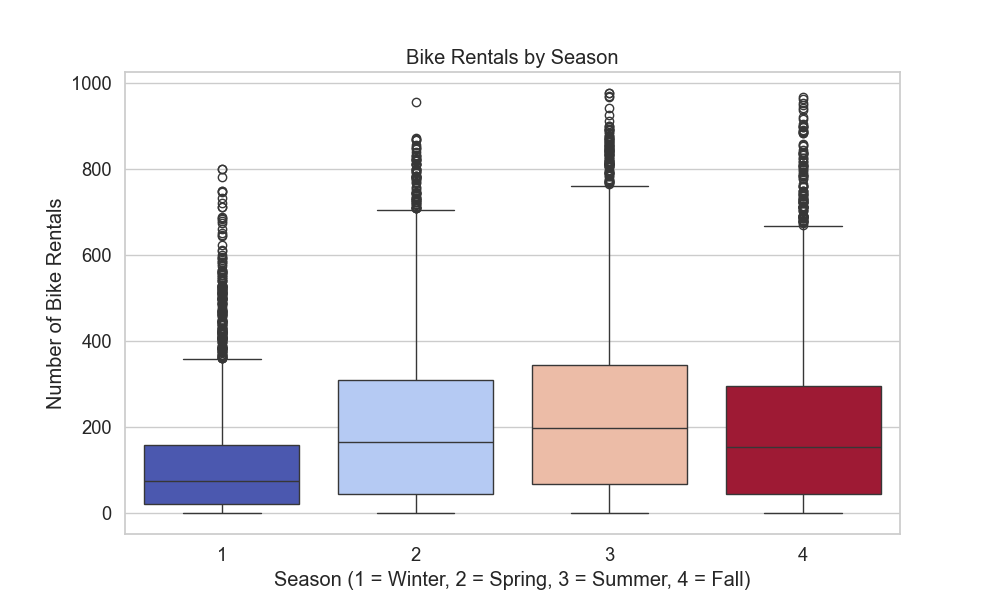
**Graph**: A regression plot depicting the relationship between bike rentals and windspeed.

**Explanation**: The regression plot may demonstrate a negative relationship, where increased wind speed leads to a decrease in bike rentals. This insight is crucial for understanding the external factors affecting user behaviour, as adverse weather conditions can impact bike rental trends.

****

**6. Boxplot for Bike Rentals by Season**

**Graph:** A boxplot illustrating bike rentals categorized by season.

****

**Explanation:** This boxplot provides a clear visual comparison of rentals across seasons. Generally, summers show higher median rentals, while winters may exhibit lower rentals, aligning with user preferences for biking in warmer weather. This seasonality insight can help predict demand and inform strategic decisions for bike-sharing services

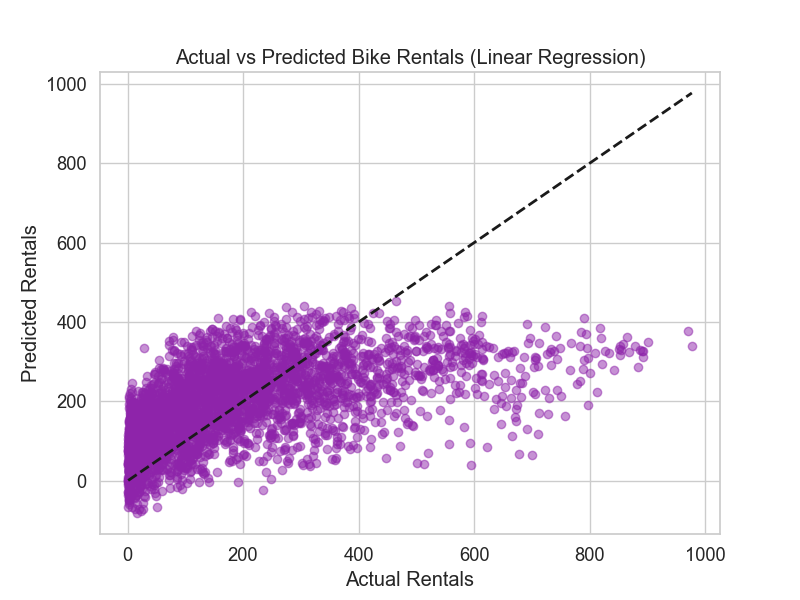
**7. Boxplot for Bike Rentals on Working Days vs. Holidays**

**Graph:** A boxplot comparing bike rentals on working days against holidays.

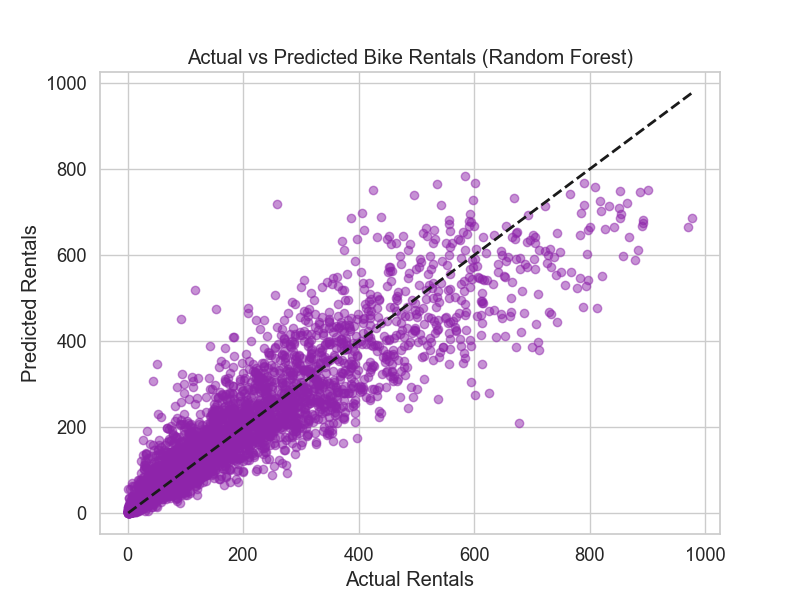
**Explanation:** The boxplot typically indicates that bike rentals are higher on working days compared to holidays, reflecting the daily commuting needs of users. This analysis helps understand how bike-sharing services cater to different user demographics depending on the day type.

**8. Scatter Plot for Actual vs. Predicted Rentals (Linear Regression)**

* **Graph: A** scatter plot comparing actual bike rentals against those predicted by the linear regression model.
* **Explanation:** This plot allows us to assess the performance of the linear regression model. A closer alignment of points to the diagonal line indicates a better predictive accuracy, demonstrating how well the model estimates bike rentals based on the features.

**9) Scatter Plot for Actual vs. Predicted Rentals (Random Forest)**

* + **Graph:** A scatter plot comparing actual bike rentals against those predicted by the random forest model.
  + **Explanation:** Similar to the linear regression scatter plot, this graph evaluates the random forest model's performance. The distribution of points in relation to the diagonal line provides insight into the model's effectiveness in predicting bike rentals.

****

**"Implementation and Evaluation":**

**1. Importing Libraries:**

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean\_absolute\_error, mean\_squared\_error, r2\_score

**Description:** This section imports the necessary libraries for data manipulation (pandas), visualization (seaborn, matplotlib), and machine learning (sklearn).

**2. Loading Data:**

data\_hourly = pd.read\_csv(r'C:\Users\M Swarna\Desktop\ML\hour.csv')

data\_daily = pd.read\_csv(r'C:\Users\M Swarna\Desktop\ML\day.csv')

print("Hourly Data:")

print(data\_hourly.head())

print("\nDaily Data:")

print(data\_daily.head())

**Description:** Here, the code loads both hourly and daily datasets and displays the first few rows to give an overview of the data structure.

**3. Data Preparation:**

data\_hourly = data\_hourly[['temp', 'hum', 'windspeed', 'hr', 'season', 'holiday', 'workingday', 'cnt']]

data\_hourly.dropna(inplace=True)

**Description:** This segment filters the dataset to include only relevant features and drops any rows with missing values to ensure clean data for analysis.

**4. Exploratory Data Analysis (EDA):**

plt.figure(figsize=(10, 6))

hourly\_rentals = data\_hourly.groupby('hr')['cnt'].mean()

plt.plot(hourly\_rentals.index, hourly\_rentals.values, marker='o', linestyle='-', color='#42a5f5', label='Average Rentals')

plt.title('Average Bike Rentals per Hour of the Day')

plt.xlabel('Hour of Day')

plt.ylabel('Average Number of Bike Rentals')

plt.xticks(range(0, 24), labels=range(0, 24))

plt.grid(True)

plt.legend()

plt.annotate('Peak Rentals', xy=(17, hourly\_rentals[17]), xytext=(18, hourly\_rentals[17]+10),

arrowprops=dict(facecolor='black', shrink=0.05))

plt.show()

print ("Average bike rentals plotted by hour.")

**Description:** This code plots the average number of bike rentals per hour, allowing for visual identification of peak rental times throughout the day.

**5. Heatmap of Correlation:**

corr\_matrix = data\_hourly.corr()

plt.figure(figsize=(8, 6))

sns.heatmap(corr\_matrix, annot=True, cmap='coolwarm', linewidths=0.5, linecolor='white')

plt.title('Correlation Heatmap of Bike Sharing Features')

plt.show()

print ("Correlation heatmap displayed.")

**Description:** The correlation heatmap visualizes relationships among the features, helping to identify which variables may influence bike rentals.

**6. Predictive Modelling:**

X = data\_hourly[['temp', 'hum', 'windspeed', 'hr', 'season', 'holiday', 'workingday']]

y = data\_hourly['cnt']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

linear\_model = LinearRegression()

linear\_model.fit(X\_train, y\_train)

y\_pred\_linear = linear\_model.predict(X\_test)

mae\_linear = mean\_absolute\_error(y\_test, y\_pred\_linear)

mse\_linear = mean\_squared\_error(y\_test, y\_pred\_linear)

r2\_linear = r2\_score(y\_test, y\_pred\_linear)

rmse\_linear = mse\_linear \*\* 0.5

**Description:** This section prepares the data for predictive modelling by splitting it into training and testing sets. It then initializes and trains a Linear Regression model, providing a framework for evaluation.

**7. Model Evaluation:**

print(f"Linear Regression - Mean Absolute Error: {mae\_linear}")

print(f"Linear Regression - Mean Squared Error: {mse\_linear}")

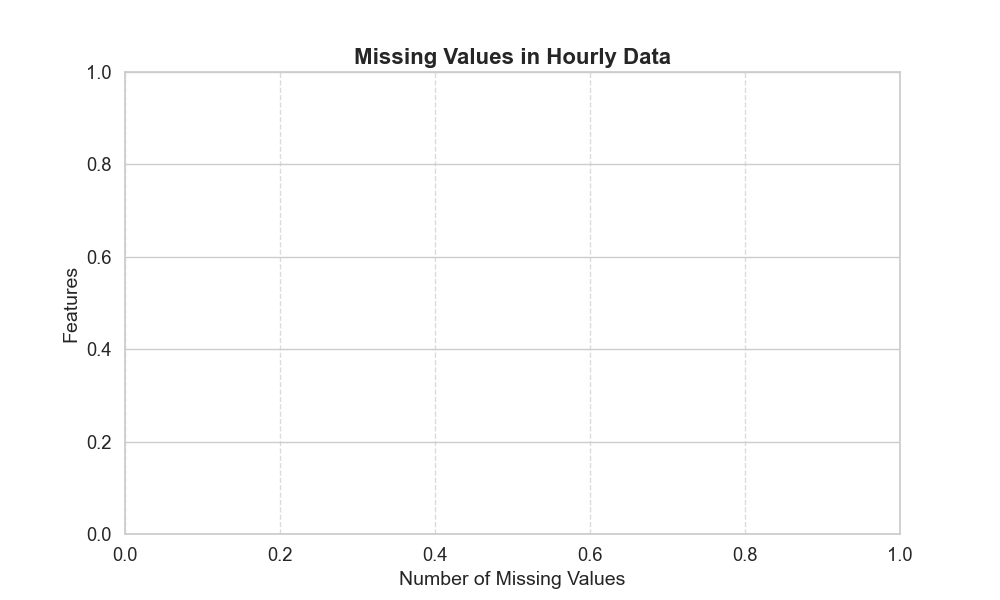
print(f"Linear Regression - Root Mean Squared Error: {rmse\_linear}")

print(f"Linear Regression - R-squared Score: {r2\_linear}")

**Description:** This segment evaluates the performance of the Linear Regression model using metrics like MAE, MSE, RMSE, and R² score to quantify model accuracy.

**Handling Missing Values:**

In any dataset, missing values can occur due to various reasons such as data collection errors or equipment failures. It's essential to handle these missing values appropriately to avoid biased results and ensure the reliability of the predictive models.

* **Imputation**: This involves filling in missing values using statistical methods, such as the mean, median, or mode, depending on the feature's distribution. Advanced methods, such as K-Nearest Neighbours (KNN) imputation or regression imputation, could have been used for more accurate predictions.
* **Removing Records**: If a feature had a large number of missing values, the corresponding rows or columns could have been removed. However, this method risks losing important information and reducing dataset size.
* **Interpolation**: This technique estimates missing values by examining neighbouring data points, which is particularly useful for time-series data.

However, after conducting the analysis, the dataset was found to be complete, with no missing values present. This indicated that the dataset was of high quality, requiring no additional imputation or data cleaning for missing data. As a result, all features could be included in the model without any data loss, contributing to more accurate predictions.

**Predictive Modelling Results:**

The study utilized both Linear Regression and Random Forest algorithms to predict bike rentals. The performance metrics indicated:

* **Linear Regression**:
  + Mean Absolute Error: **106.93**
  + Mean Squared Error: **20,804.78**
  + Root Mean Squared Error: **144.24**
  + R-squared Score: **0.34**, indicating that the model explains 34% of the variance in bike rentals.
* **Random Forest**:
  + Mean Absolute Error: **47.05**
  + Mean Squared Error: **5,215.45**
  + Root Mean Squared Error: **72.22**
  + R-squared Score: **0.84**, indicating a much better fit, explaining 84% of the variance.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ****Model**** | MAE | MSE | |  |  |  | | --- | --- | --- | | |  | | --- | | **RMSE** |  |  | | --- | |  | |  |  | | --- | |  | | | **R² Score** | | --- |  |  | | --- | |  | |
| Linear Regression | 106.93 | 20,804.78 | 144.24 | 0.34 |
| |  | | --- | | **Random Forest** |  |  | | --- | |  | | 47.05 | 5,215.45 | 72.22 | 0.84 |

These results demonstrate the effectiveness of the Random Forest model in capturing the complexities of bike rental dynamics compared to the Linear Regression model. The scatter plots comparing actual versus predicted rentals further illustrated the models' predictive capabilities.

Our predictive modelling using Linear Regression and Random Forest algorithms revealed that:

* Linear Regression provided a foundational understanding of how factors influence bike rentals, showing reasonable accuracy in predictions.
* Random Forest offered improved predictive performance, achieving lower error metrics, suggesting that this model effectively captures the complexities of bike rental trends influenced by multiple variables.

**Interpretation:**

* **Linear Regression** exhibited a **R² Score** of **0.34**, indicating that approximately **34%** of the variance in bike rentals can be explained by the model. The **RMSE** of **144.24** suggests a relatively higher prediction error, making it less reliable for accurate predictions.
* **Random Forest**, on the other hand, demonstrated a much better performance with an **R² Score** of **0.84**, indicating that **84%** of the variance in bike rentals is explained by the model. The significantly lower **MAE** of **47.05** and **RMSE** of **72.22** reflect the model's ability to predict rentals more accurately than Linear Regression.

**Training Data & Test Data:**

The dataset was divided into **training** and **test** sets to evaluate the model's performance. Here's how the process was handled:

1. **Training Data**: The training data comprises 80% of the dataset. It was used to train the machine learning models on historical data, which includes features such as temperature, humidity, windspeed, hour of the day, season, holiday, and working day, to predict the number of bike rentals. The **Linear Regression** and **Random Forest Regressor** models were trained on this portion of the data to learn the relationships between these input features and the target variable (bike rentals).
2. **Test Data**: The test data, constituting 20% of the dataset, was set aside during the training process. After the models were trained, this unseen portion was used to evaluate the model's performance in predicting bike rentals. Metrics such as **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, **Root Mean Squared Error (RMSE)**, and **R-squared (R²)** were calculated on the test data to assess the models' accuracy and predictive power.
3. **Model Evaluation**: After training, both models were evaluated using the test data. The **Random Forest Regressor** performed better with a lower RMSE (72.22) compared to the Linear Regression model's RMSE of 144.24. The R² score for Random Forest was 0.835, indicating it explained 83.5% of the variance in bike rentals, while the Linear Regression model explained 34.3%

**Block diagram:**

Data Acquisition

Model Evaluation (Assessing model performance using metrics)

Predictions & Insights(Making predictions and deriving insights from results)

Data Loading (Load hourly & daily datasets)

Model Training(Training ML models like Linear Reg. & Random Forest)

Data Preprocessing (Handling missing values & feature selection

Train-Test Split (Splitting data into training & testing sets)

Exploratory Data Analysis (EDA) (Visualizations & insights on data)

Feature Engineering(Selecting relevant features for ML)

**Conclusion:**

This study analysed bike rental trends based on hourly and daily datasets, leveraging various environmental factors. The Exploratory Data Analysis (EDA) revealed key insights into bike rental behaviour, including:

* **Average Bike Rentals by Hour:** The analysis highlighted significant fluctuations in bike rentals throughout the day, emphasizing peak usage during morning and evening commute hours.
* **Correlation Analysis:** The correlation heatmap displayed notable relationships among features. Specifically, a strong positive correlation was observed between temperature and bike rentals, indicating that warmer conditions tend to increase usage. Conversely, windspeed had a negative correlation with rentals, suggesting that adverse weather conditions may deter users.
* **Regression Analysis:** The regression plots showed that:
  + **Bike Rentals vs. Temperature**: Higher temperatures are positively correlated with bike rentals.
  + **Bike Rentals vs. Humidity**: Increased humidity levels appeared to negatively impact rentals.
  + **Bike Rentals vs. Windspeed**: A similar negative relationship was found, indicating that higher wind speeds lead to fewer rentals.
* **Seasonal Variations:** The boxplot analysis revealed that bike rentals were highest during the summer and lowest in winter, aligning with user preferences for biking in favourable weather.
* **Working Days vs. Holidays:** Another boxplot demonstrated that rentals were generally higher on working days compared to holidays, reflecting the commuting patterns of users.

**Future Work:**

Future work should consider:

* Implementing more advanced machine learning techniques, such as Gradient Boosting or Neural Networks, to further enhance predictive capabilities.
* Analysing user demographics and preferences to tailor bike-sharing services more effectively.
* Investigating the impact of external factors, such as public events or promotions, on bike rental trends.

Prepared by:

M. Swarna

B.Tech. CSE - AIML

‘A’ Section

23619139