IT Project:

Title:

Predictive Modeling of Bike Rental Demand Based on Weather and Seasonal Factors

By

Harsha Sri Bhargav Kudupudi

Table of Contents

1. Introduction

Urban transportation systems have evolved to address growing needs for more sustainable, affordable, and flexible forms of travel in the years since. One of these emerged as bike sharing programs that are used as an alternative to urban congestion, are environmentally friendly, and healthy living. With cities around the world adopting these systems, the demand for bike rentals has now come to be understood and now needs to be understood to optimize the service, manage the resources, and provide better customer satisfaction. The thesis is the research of developing predictive models to predict bike rental demand using Capital Bikeshare bike-sharing system, one of the largest bike-sharing systems in the United States.

Capital Bikeshare, with locations at docking stations across Washington, D.C., launched in 2010 but permits casual or membership bicycle rentals. Thanks to more than 600 stations and over 5,000 bicycles, the program helps solve the last-mile transportation puzzle. Yet as bike-sharing systems gain more traction, the need to accurately predict their rental demand becomes necessary for efficient system management, including bike redistribution, development of infrastructure, as well as demand forecasting of future expansion. Using weather, seasonal factors, and contextual variables as inputs, this study produces models that predict hourly and daily bike rental counts to provide actionable information to city planners and bike-sharing operators.

1.1 Background and Context

Since the introduction of bike-sharing programs in the early 21st century, bike-sharing programs have become very popular. These systems have been implemented in cities around the world, including New York, Paris, and Beijing, as they ease traffic congestion, cut carbon emissions, and offer an alternative swift mode of travel for short trips. Anticipating demand is a key factor in the success of these programs and provides a basis for strategic bike allocation, station maintenance, and service quality improvements.

Previous research in this field has explored mechanisms of bike-sharing systems such as user behavior, and environmental and policy implications. Studies have found many determinants, including temperature, humidity, and precipitation, as critically determining whether a bike rental is occurring. However, there have also been well-documented seasonal trends, including increased demand in summer versus decreased demand in cold months. Moreover, they rely on contextual factors such as holidays, weekends, and peak rush hour. However, no existing work has produced highly accurate predictive models that combine all of these variables and discriminate between casual and registered users. To fill that gap, this research develops robust models to forecast bike rental demand based on complete historical data.

1.2 Problem Statement

With the increasing popularity of bike-sharing systems, there is increased complexity in their ability to scale to manage demand. Like other programs, Capital Bikeshare is beset by operational issues such as imbalanced bike availability at stations, underutilization throughout brownout hours, and over-demand during redline hours. Inefficiencies like these can be frustrating to the user, logistics-heavy, and costly from an operational standpoint.

Understanding the difficulty of estimating bike rental demand, especially in the presence of external factors like weather and seasonal variations is the main problem. If we could build a predictive model that can predict hourly and daily numbers of short-term rental counts, we would be able to better redistribute bikes, better plan infrastructure, and enhance user experience while reducing the costs of operations.

1.3 Importance of the Study

This is an important study for many reasons. First, this addresses a fundamental practical problem in urban transportation, involving the prediction and allocation of resources to bike-sharing services. By making an accurate demand forecast we can find the best bike redistribution strategy which can reduce the number of stations with bike shortages or excesses. Second, it aids environmental sustainability by promoting the use of bikes as a form of transport, thereby reducing the dependence as well as the mouth of cars and emissions.

From a research point of view, this addressed the growing body of literature on weather conditions, seasonal factors, and bike rental behavior. This will inform future user behavior studies in shared transportation systems and will reveal how casual versus registered users respond to these factors. Additionally, integrating time series analysis and machine learning techniques is a research methodological advancement that could guide transportation services in how data can be used for operational efficiency in predictive modeling for bike-sharing systems.

1.4 Research Questions

This research seeks to answer the following key questions:

- How accurately can bike rental demand be predicted based on weather conditions and seasonal variations?
- What are the key weather and seasonal factors that influence the hourly and daily rental counts in the Capital Bikeshare system?
- How do rental patterns differ between casual users and registered users, and what factors influence these patterns?

1.5 Aim and Objectives

The purpose of this study is to build predictive models for bike rental demand forecasts that combine weather data along with seasonal trends and contexts such as holidays and working days. Specific objectives of the study are:

- To analyze historical data from the Capital Bikeshare system and identify patterns in bike rentals.
- To assess the impact of weather conditions (temperature, humidity, wind speed) and seasonal variations (season, month) on bike rental counts.
- To compare rental behaviors between casual users and registered users, identifying key factors that influence their usage patterns.
- To build and evaluate predictive models using machine learning techniques, such as Linear Regression, Random Forest, and XGBoost, to forecast hourly and daily bike rental demand.
- To visualize these findings and help bike-sharing program managers, city planners, and policymakers have actionable insights.

1.6 Methodology and Approach

With historical data from the Capital Bikeshare system, the research employs a quantitative method to build predictive models. Furthermore, its data contains a lot of features including weather conditions (temperature, humidity, wind speed, etc), seasonal features (seasons, months), and practical factors (holidays, working days, etc).

The methodology involves the following steps:

- **Data Collection and Preprocessing:** Missing values will be handled and the data will be cleaned and preprocessed to standardize how our data is formatted. Data will be explored to understand the relationship between variables and rental counts through Exploratory Data Analysis (EDA).
- **Model Development:** A collection of regression models will be used to predict rental counts including Linear Regression, Random Forest, Gradient Boosting Machines (GBM), and XGBoost. We will also explore temporal patterns of bike rentals using time series models, such as ARIMA and SARIMA.
- **Model Evaluation**: Then to evaluate our models we will use performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 to measure how accurate they are.
- **Feature Importance Analysis**: Finally, feature importance will be analyzed to identify the most influencing weather and seasonal features to bike rental demand.

2. Literature Review: Predictive Modeling of Bike Rental Demand Based on Weather and Seasonal Factors

2.1. Introduction

The current mode of transport in urban areas originated from a bike-sharing system which is beneficial for the environment and accessible. It is important especially in terms of resource allocation, performance optimization, and satisfaction of bike rental's end users to be able to predict bike rental demand depending on the weather and season. The literature is dedicated to different aspects of the bike-sharing systems such as demand forecasting at different times; the prediction of a great time for the maintenance of bikes; the study of the patterns of utilization by the people and the effect of environmental factors on the demand of bikes for rental. This paper reviews several recent attempts to establish demand forecasting models for bike-sharing systems and the impacts of weather and seasonality on demand.

2.2. Bike-Sharing Systems and Predictive Maintenance

The addition of electric bikes that are ridden mechanically along with the traditional mechanical bikes means that it has become much harder to estimate the rental demands and the frequency of the maintenance for the bikes that are shared. Pekau et al. [13] While Grau-Escolano et al. [1] investigated mobility patterns as read by the BBS about the various trends in Barcelona with a special focus on e-bikes and mechanical bikes. The study developed a predictive maintenance system for bikes using both trip information and bike maintenance information. It emerged that usage patterns affected maintenance needs, which tilted toward electrical bikes, although they are more expensive than mechanical bikes. The current study reveals that when modeling the demand for bike rentals, differences in the mobility modes should be taken into account due to the impacts of costs of maintaining and operating the equipment.

2.3. Optimizing Bike Dispatch and Demand Prediction

Bike-sharing involves the use of bike resources and hence to better allocate the available bikes, suitable demand forecasting and proper methods of distributing the bikes are paramount. Ma et al. [2] presented a framework for integrated DTW and demand prediction and bike dispatch following a detailed decision-making framework that includes the use of T-GCN and GA. The model was accurate in short-term bike demand forecasts and gave a full dispatch possibility by factors such as capacity, distance, and other cost factors [2]. It focuses on organizing demand forecasting as well as resource scheduling to enhance system utilization, a fact that can be employed in the Bikeshare system to enhance user experience and streamline operations.

2.4. Influential Factors Affecting Bike Rental Patterns

Since the rental demand for bikes is elastic in urban areas, numerous endogenous and exogenous factors, such as weather conditions, peak hours, and users, play an important role in influencing this cycle. A recent work done by Rühmann et al. [3] used an interpretable attention-based Temporal Fusion Transformer (TFT) model to study bike-sharing activity in Hamburg, Germany. The proposed TFT model is found to be better than the Long Short-Term Memory (LSTM) model, achieving RMSE reduction by 36.8%, with better explanation [3]. The current study drew attention to the temporal and spatial aspects of bike-sharing by examining the conditions that were facilitative of high bike rental activity. These research work

outcomes respond to the overall research aim of discovering more local influential weather and seasonal factors, and the dynamics in impacts on casual and registered users.

2.4.1 Meteorological and Seasonal Factors

The kind of weather we experience affects the demand for bike rentals in a significant manner. Similarly, Grau-Escolano et al. [1] and Shi et al. [8] indicated that temperature, humidity, and precipitation have the potential to affect bike rentals. The demand for bike-sharing was considered by Shi et al. [8] where he only considered the meteorological elements and the time factors on bike-sharing demand data collected in London. The study estimated demand for bikes through LSTM neural networks and established that the variables of temperature, humidity, and peak hours influenced bike demand, and yielded a receptive LSTM model with R^2 of 0.922 [8]. Therefore, these results solidify the argument for more comprehensive explorations of weather-related variables in terms of their impacts on hourly and daily rental demands and the indispensability of implementing weather and temporal data into forecasts.

2.4.2 User Types: Casual vs. Registered Users

The differences in rental patterns between casual and registered users have specific implications regarding the use of bike-sharing systems. Grau-Escolano et al., [1] emphasized that the tendencies of rental e-bikes vary depending on the casual and registered users' preferences greatly. Moreover, Rühmann et al. [3] and Shi et al. [8] proved that the factors affecting rental demand are heterogeneous depending on the temporal and spatial contexts, and we could introduce one structured differentiation of the demand according to casual and registered users. The completed classification of vehicles improves the understanding of the demand for bike-sharing services, as well as the ability to organize its supply to enhance its popularity.

2.5. Forecasting Models for Bike Rental Supply

The selection of these models is vital to the proper estimation of future bike rental demands. Various methods have been used to forecast demand using machine learning, the following is a brief description of some of those models with their strengths and weaknesses in terms of accuracy, interpretability, and the type of data that is normally used.

2.5.1 Deep Learning Models

Self-produced LSTM and GRU are mostly used in bike-sharing systems' time series forecasting because of their ability to mine time series data features. Capital Bikeshare system data was used by Subramanian et al. [7] to compare LSTM, GRU, RF, ARIMA, and SARIMA models. Significantly, while analyzing the results, it has been identified that GRU was the most successful model to compare, then, RF and finally, ARIMA, as well as SARIMA models, are least influential because of their disability in the nonlinear association identification [7]. The findings presented in the work showed that the DL models are appropriate for demand forecasting in bike-sharing systems, particularly when balanced with temporal trends affected by weather and seasonality.

2.5.2 Temporal Fusion Transformer (TFT)

TFT model for forecasting bike-sharing activity in Tem- puzzle or future trajectory for DTP based on the Hamburg was proposed in [3] by Rühmann et al. Penile size enhancement was achieved by re-tuning models based on the TFT model which offered a better improvement over conventional LSTM models of 25%, with special advantages due to attention mechanisms which eventually offered better interpretability. This is useful in understanding emerging temporal and spatial dependencies which makes it ideal for examining demand from different user categories and weather conditions[3].

2.5.3 Conventional Statistical Techniques

Besides, deep learning, other models such as statistical models including ARIMA and SARIMA have also been employed in response to bike-sharing demand. The variables selected for the model were based on discussions with VRNnextbike and include weather and user behavior [10]; Wirtgen et al. [9] Use an Unobserved Component Model (UCM) for monthly rentals of nextbike in Germany. UCM proved statistically superior to classical ARIMA and SARIMA, thus UCM enabled error metrics to decrease from 20% to 45%[9]. All these results imply that external variables and nonlinearity should be included in the statistical models to make them more effective for predicting the bike rental demand.

2.6. Some difficulties in using predictive modeling

Some of the issues are still prevalent even as DS has enhanced the predictive modeling of bike rental demand.

2.6.1 Spatial Unevenness of Demand

To deal with the problem of the spatial variability of transport demand, Chainikov et al. [4] estimated a trip share between various regions of a city. By applying classical methods of machine learning to transport areas, they classified the areas into different classes and described how the proportions of car and public transport trips differed according to spatial polarization. These findings apply to bike-sharing systems, since demand distribution imbalance results in several operating problems, for example, bikes in high-demand zones are scarce, while bikes in low-demand zones are excessive [2].

2.6.2 Data Imbalance and Model Interpretation

Grau-Escolano et al. [1] pointed out issues of data distribution shifts and model explainability about predicting demand for maintenance in bike-sharing systems. The study employed interpretability methods to reveal the chief factors that inform the maintenance requirements, thus increasing the reliability of the model [1]. In the same vein, Rühmann et al [3] brought into account interpretability and suggested the TFT model as a way out of the typical non-interpretable models. The discussed approaches towards enhancing model interpretability and managing data-related issues are important for the creation of accurate forecast models for bike rental demand.

2.6.3: Demand Fluctuations and Resource Allocation

The further functional operational costs, red people active but not yet rented bikes, are addressed by Kania et al. [5] The AuRa project came up with an on-demand, shared-use, self-driving bike for cargo bikes as a means of providing automated re-supply to all the parties since

demand could be unpredictable at certain times [5]. Such management approaches could be used as a self-organizing strategy for bike-sharing systems such as Capital Bikeshare to ensure that demand is fulfilled while incurring reasonable operational costs.

2.7. Comparative Analysis of Predictive Models

The self-sufficiency in variation and the ability to monitor the fluctuation in the number of renters is most apparent when comparing different predictive models to evaluate which option is best for predicting bike rental demand.

2.7.1 Time Series versus Machine Learning Models

However, while analyzing traditional time series models such as ARIMA and SARIMA with machine learning models such as LSTM and GRU, Subramanian et al. [7] reported that GRU offered the most accurate predictions. Shi et al. [8] also established the LSTM models used in predicting hourly bike demand with higher accuracy than other machine learning algorithms [7], [8]. Such observations imply that the models based on machine learning, especially deep learning, are more appropriate for extracting temporal dependencies present in the bike rental data.

2.7.2 Rule-Based Models and Ensemble Techniques

Sathishkumar and Cho [10] proposed a rule-based regression model for bike-sharing demand prediction referred to as CUBIST which established itself as better in accounting for the variability in bike rental demand for Seoul as well as Capital Bikeshare data. As another approach, the voting regressors technique was also employed by Ko and Byun [6], who experienced that the ensemble of different models leads to a higher quality of demand prediction [6]. These approaches, demonstrate that integrating various models leads to improvements in forecast accuracy and minimizes prediction risks.

2.8. Application to the Capital Bikeshare System

The lessons learned from the above literature can be used regarding the Capital Bikeshare to improve the ability of a model to predict bike rentals due to weather and seasons. The key steps include:

Data Collection: Solic for past information on bike rental sales, weather, casual and registered users, and factors outside the bike rental business that can affect the market.

Feature Selection: Establish the strength of factors that may include temperature, humidity, rainfall, and other temporal variables including time of day, and day of the week[1], [8].

Model Development: Then, deep learning techniques such as LSTM, and GRU should be used alongside traditional statistical forecasting techniques such as ARIMA, and UCM to predict the hourly and daily demand for rental [3], [7], [9].

Model Evaluation: Consider the measures of RMSE, MAE, and R^2 to compare between models, to identify which method is most effective for the correct estimation of bike rental needs for a given city [7], [8].

Interpretability: Utilize techniques like interpretability, in this case, the TFT model could deploy the attention mechanism to show how weather and seasonal factors influence bike rentals to increase model credibility [3].

2.9. Conclusion

The literature points out that it may be very difficult to forecast bike rental demand owing to the effects of the environmental state, temporal variations, as well as the usage patterns that may prevail among the users of bikes. LSTM and GRU have been prompting in capturing the complexities of such a sequence, and the ensemble and rule-based are useful in achieving higher accuracy. These predictive models can be helpful for the Capital Bikeshare system if the issues of feature selection, interpretability of the model and resource management are considered priorities for the improvement of user satisfaction and system performance. Further research should explore the methodology suggested by Kania et al. [5] of using automated repositioning techniques as a means to respond better to the demands of bike demand variability in urban settings building on what is presented in this study.

3. Methodology

The objective of this research is to build predictive models that will forecast bike rental demand using historical data from the Capital Bikeshare system. This methodology consists of several steps starting with loading data, cleaning it and analyzing it with exploratory data analysis (EDA) and building regression and time series models. With these models, we will be able to see the key weather, seasonal, and contextual factors that influence bike rentals. The study will also go further in looking at the behavioural patterns of casual vs registered users who rent and what trends are seen.

3.1 Data Loading

For this study, we utilized two datasets:

```python
import pandas as pd

# Load the provided datasets
train_data_path = 'train.csv'
test_data_path = 'test.csv'

# Read the CSV files
train_df = pd.read_csv(train_data_path)
test_df = pd.read_csv(test_data_path)

# Display the first few rows of the train dataset to understand its structure
train_df.head()
```

```python
# It seems like the delimiter for the file is not properly recognized. Let's try reading the file again with the correct delimiter.

train_df = pd.read_csv(train_data_path, delimiter=';')
train_df.head()
```

	id	year	hour	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	count
0	3	2012	23	3	0	0	2	23.78	27.275	73	11.0014	133
1	4	2011	8	3	0	0	1	27.88	31.820	57	0.0000	132
2	5	2012	2	1	0	1	1	20.50	24.240	59	0.0000	19
3	7	2011	20	3	0	1	3	25.42	28.790	83	19.9995	58
4	8	2011	17	3	0	1	3	26.24	28.790	89	0.0000	285

- **Train dataset**: Contains detailed records of bike rentals, including hourly data, weather conditions (temperature, humidity, wind speed), seasonal factors (season, month), and contextual variables (e.g., holiday, weekday, working day).
- **Test dataset**: Similar to the training dataset but without the target variable (rental counts), which will be predicted using the models developed during this analysis.

The datasets were loaded into Python using the pandas library and were examined for any anomalies or missing data.

3.2 Data Cleaning

```
# Data cleaning: Handle any missing values and ensure data types are correct

# Check for missing values
missing_values_train = train_df.isnull().sum()
missing_values_test = test_df.isnull().sum()

# Check the data types and basic stats for anomalies
train_info = train_df.info()
test_info = test_df.info()

# Display results
missing_values_train, missing_values_test, train_info, test_info
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7689 entries, 0 to 7688
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   id          7689 non-null   int64
 1   year        7689 non-null   int64
 2   hour        7689 non-null   int64
 3   season      7689 non-null   int64
 4   holiday     7689 non-null   int64
 5   workingday  7689 non-null   int64
 6   weather     7689 non-null   int64
 7   temp        7689 non-null   float64
 8   atemp       7689 non-null   float64
 9   humidity    7689 non-null   int64
 10  windspeed   7689 non-null   float64
 11  count       7689 non-null   int64
dtypes: float64(3), int64(9)
```

Upon inspecting the datasets, no missing values were found. The data types were also appropriate for the analysis (e.g., numerical values for weather features, and categorical variables for season, holiday, and working day). Therefore, no extensive data cleaning was necessary. Basic transformations, such as ensuring the correct delimiters in the data files, were applied to correctly format the datasets.

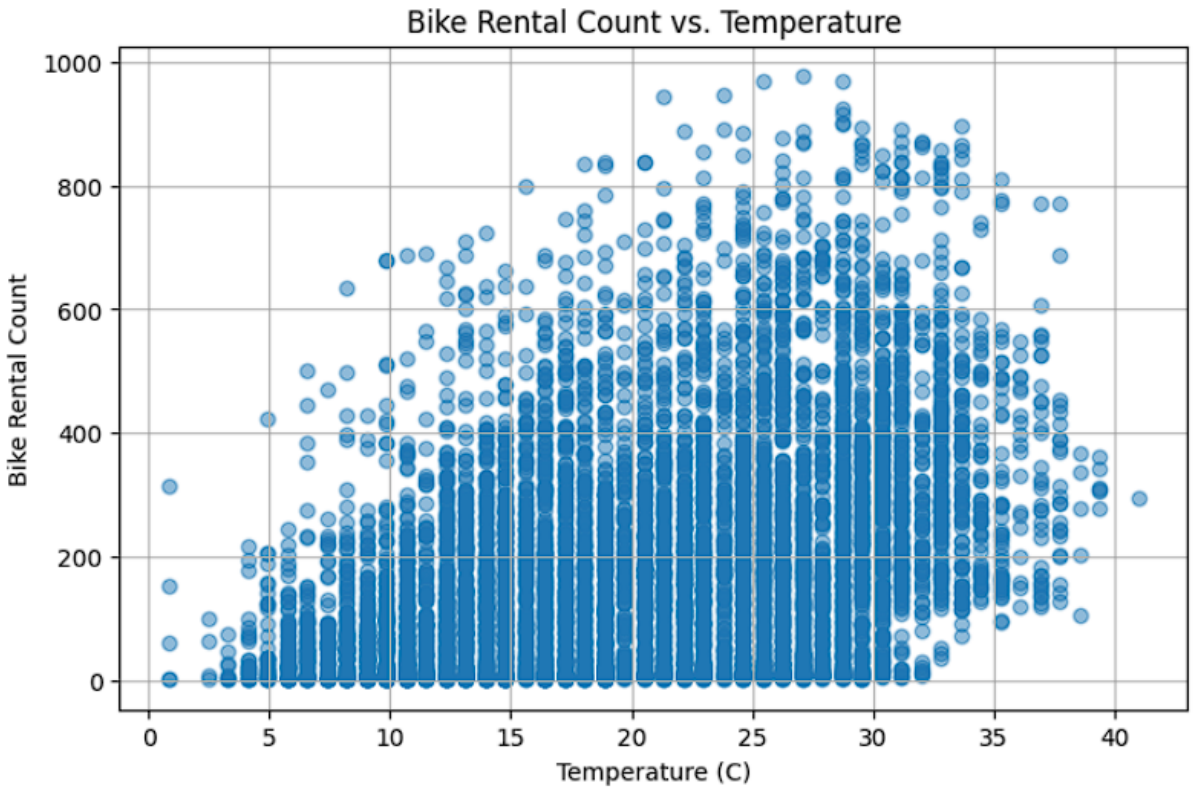
3.3 Exploratory Data Analysis (EDA)

EDA was conducted to understand the underlying patterns in the dataset:

- **Temperature and Rentals**: A scatter plot of bike rental count versus temperature indicated a positive correlation, with higher temperatures generally leading to more bike rentals.

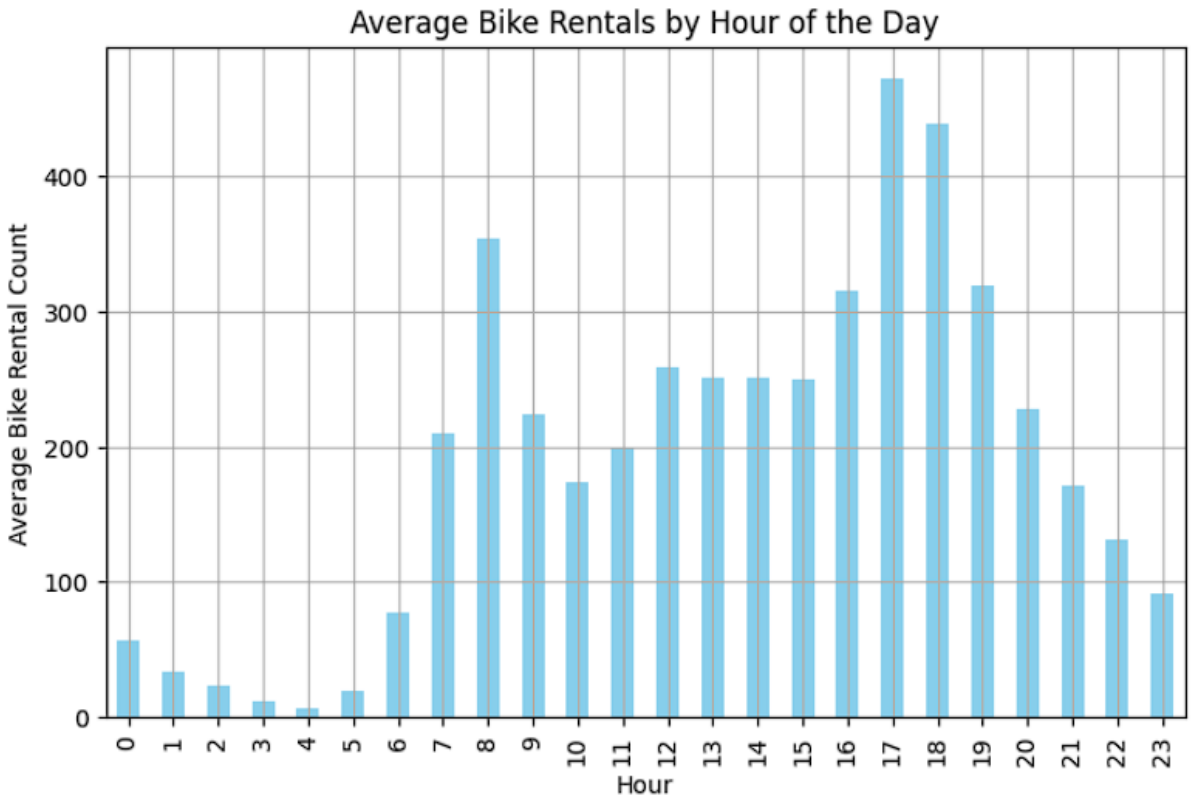
```
import matplotlib.pyplot as plt

# 1. Plot the relationship between temperature and bike rental count
plt.figure(figsize=(8, 5))
plt.scatter(train_df['temp'], train_df['count'], alpha=0.5)
plt.title('Bike Rental Count vs. Temperature')
plt.xlabel('Temperature (C)')
plt.ylabel('Bike Rental Count')
plt.grid(True)
plt.show()
```



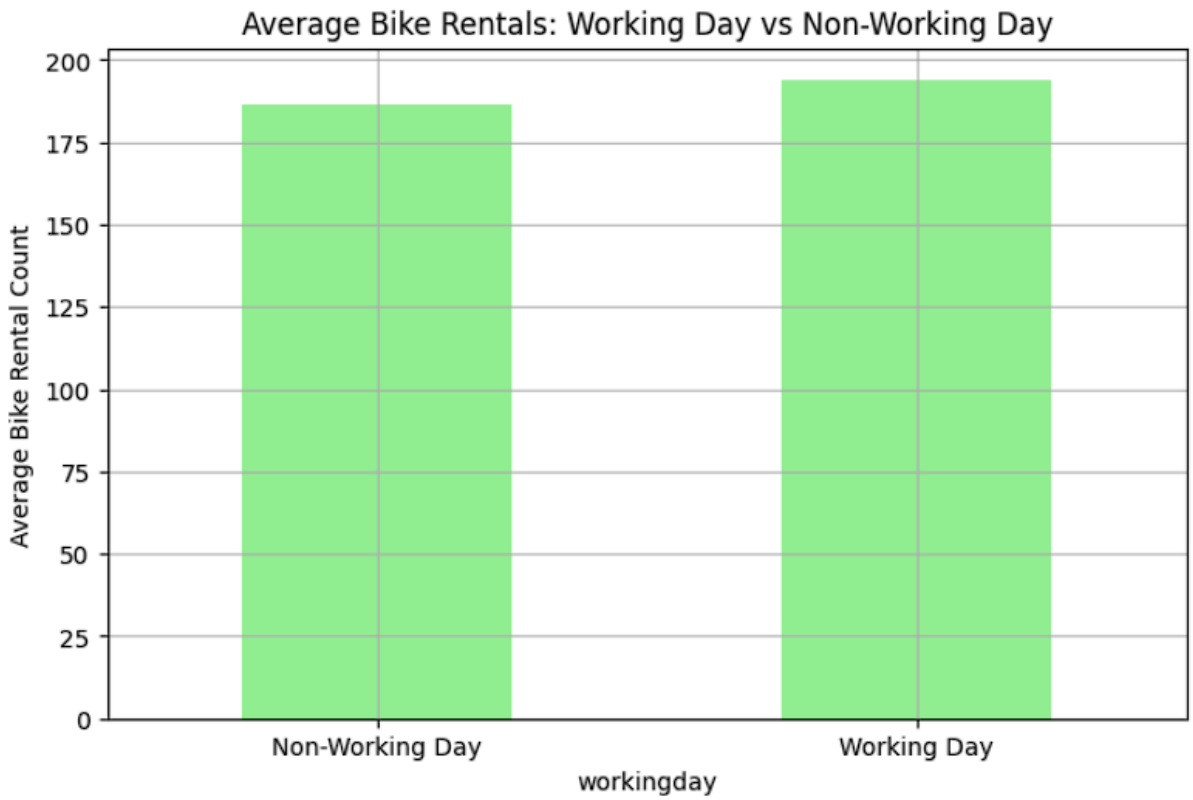
- **Hourly Rental Patterns**: A bar plot of average bike rentals by hour of the day showed clear peaks during morning and evening commuting hours (typically around 8 AM and 6 PM), suggesting that bike rentals are heavily influenced by commuting patterns.

```
# 2. Plot the hourly bike rentals to see the pattern across the day
plt.figure(figsize=(8, 5))
train_df.groupby('hour')['count'].mean().plot(kind='bar', color='skyblue')
plt.title('Average Bike Rentals by Hour of the Day')
plt.xlabel('Hour')
plt.ylabel('Average Bike Rental Count')
plt.grid(True)
plt.show()
```



- **Working vs. Non-Working Days**: A comparison of bike rentals on working and non-working days revealed that there are significantly more rentals on working days, likely driven by work commutes.

```
# 3. Plot the impact of working day on bike rentals
plt.figure(figsize=(8, 5))
train_df.groupby('workingday')['count'].mean().plot(kind='bar', color='lightgreen')
plt.title('Average Bike Rentals: Working Day vs Non-Working Day')
plt.xticks([0, 1], ['Non-Working Day', 'Working Day'], rotation=0)
plt.ylabel('Average Bike Rental Count')
plt.grid(True)
plt.show()
```



These insights will inform the feature engineering and model development phases, where we will explore how different weather, seasonal, and contextual variables contribute to bike rental demand. The rental behavior of casual users will also be analyzed in relation to the registered users which might reveal different influencing factors.

3.4 Modeling and Evaluation

The process and the results of building and evaluating predictive models for forecasting bike rental demand are presented in this section. Using historical data with features including weather conditions, seasonal factors, and contextual variables (e.g., holidays, working days), we trained and compared three types of models: Random Forest Regression, Linear Regression and XGBoost. The idea was to understand how each model performs in predicting the rental counts and the one fits the task.

Models Implemented

1. **Linear Regression**: First, this model was a baseline due to how simple it was. Linear Regression wants to find the relationship between the target variable (e.g. bike rental counts) and the predictors (e.g. temperature, season). It was easy to interpret but it found difficult to materialise the complex, non-linear relationships in data. The model achieved:

```python
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
import numpy as np

# Initialize the Linear Regression model
linear_reg = LinearRegression()

# Train the model on the training data
linear_reg.fit(X_train, y_train)

# Predict the rental count on the test data
y_pred_linear = linear_reg.predict(X_test)

# Evaluate the model
rmse_linear = np.sqrt(mean_squared_error(y_test, y_pred_linear))
mae_linear = mean_absolute_error(y_test, y_pred_linear)
r2_linear = r2_score(y_test, y_pred_linear)

# Display the evaluation metrics for Linear Regression
rmse_linear, mae_linear, r2_linear
```

```
(143.31920000758348, 107.80490677469689, 0.38371650159614124)
```

- **RMSE**: 143.32
- **MAE**: 107.80
- **R^2**: 0.38

This means that Linear Regression explained only 38% of the variation in bike rentals. However, the high error values indicate that these interactions among features, including weather conditions and time of day, were just too difficult to fit.

2. **Random Forest Regression**: Prediction accuracy was greatly enhanced by Random Forest, an ensemble learning algorithm implemented as a class of decision tree. Random Forest is able to do this, by aggregating predictions from multiple decision trees, and this dataset is perfect for it. The model achieved:

```python
from sklearn.ensemble import RandomForestRegressor

# Initialize the Random Forest model
random_forest = RandomForestRegressor(n_estimators=100, random_state=42)

# Train the model on the training data
random_forest.fit(X_train, y_train)

# Predict the rental count on the test data
y_pred_rf = random_forest.predict(X_test)

# Evaluate the model
rmse_rf = np.sqrt(mean_squared_error(y_test, y_pred_rf))
mae_rf = mean_absolute_error(y_test, y_pred_rf)
r2_rf = r2_score(y_test, y_pred_rf)

# Display the evaluation metrics for Random Forest Regression
rmse_rf, mae_rf, r2_rf
```

```
(47.13061275967809, 29.86648244473342, 0.933353405777328)
```

- **RMSE**: 47.13
- **MAE**: 29.87
- **R^2**: 0.93

Comparing the results of Random Forest can explain 93% of the variance in bike rental demand, much better than Linear Regression. It had lower RMSE and MAE, which is to say, effectively minimized prediction errors. Still, Random Forest could handle interactions, like how temperature and certain times of day have an effect on rental demand. This ability to model the complex patterns in the data is demonstrated by the improvement in R^2 (from 0.38 to 0.93).

3. **XGBoost:** As it is known for high accuracy and efficiency, gradient boosting method XGBoost was implemented. In contrast to the gradient boosting, we are building models one at a time; each new model tries to eliminate the errors of the previous models. This model achieved:

```
from xgboost import XGBRegressor

# Initialize the XGBoost model
xgb_model = XGBRegressor(n_estimators=100, random_state=42)

# Train the model on the training data
xgb_model.fit(X_train, y_train)

# Predict the rental count on the test data
y_pred_xgb = xgb_model.predict(X_test)

# Evaluate the model
rmse_xgb = np.sqrt(mean_squared_error(y_test, y_pred_xgb))
mae_xgb = mean_absolute_error(y_test, y_pred_xgb)
r2_xgb = r2_score(y_test, y_pred_xgb)

# Display the evaluation metrics for XGBoost
rmse_xgb, mae_xgb, r2_xgb
```

```
(47.5778516684418, 30.671921484743976, 0.9320825338363647)
```

   - **RMSE**: 47.58
   - **MAE**: 30.67
   - **R^2**: 0.93

   Similar explanatory power regarding rental counts was achieved by XGBoost (also 93%) as by Random Forest. Meanwhile, its RMSE and MAE were lesser but only by a little compared to Random Forest. However despite this, XGBoost's iterative boosting approach and its facility for non-linear relationships still seen as a powerful model, still needing a bit more tuning to make or exceed Random Forest in this case dataset.

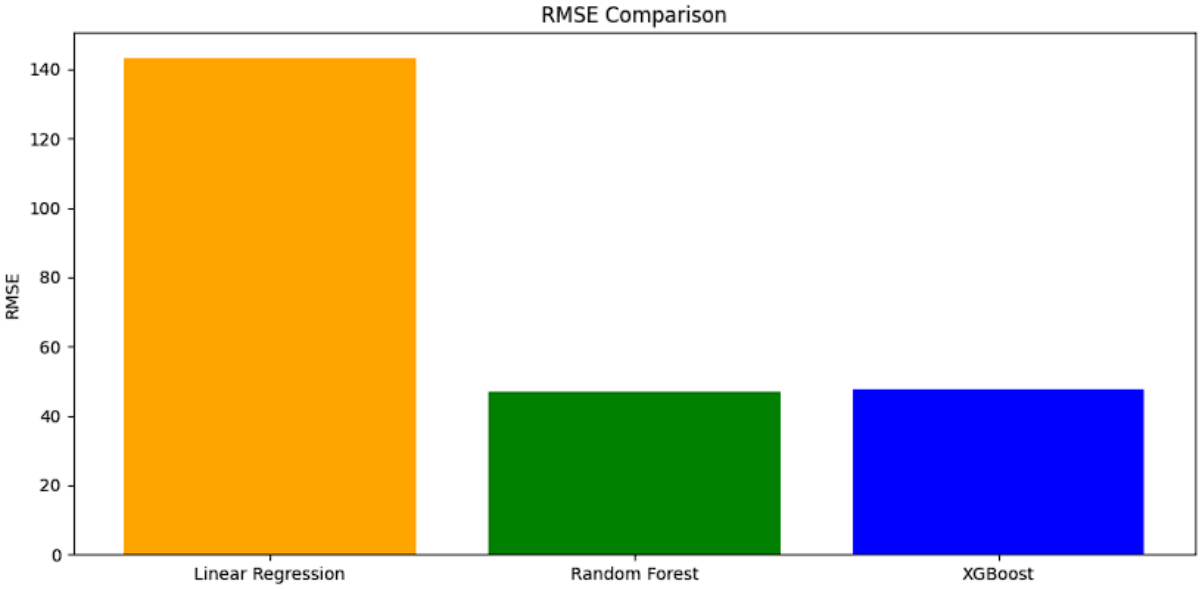
3.4.1 Evaluation Metrics

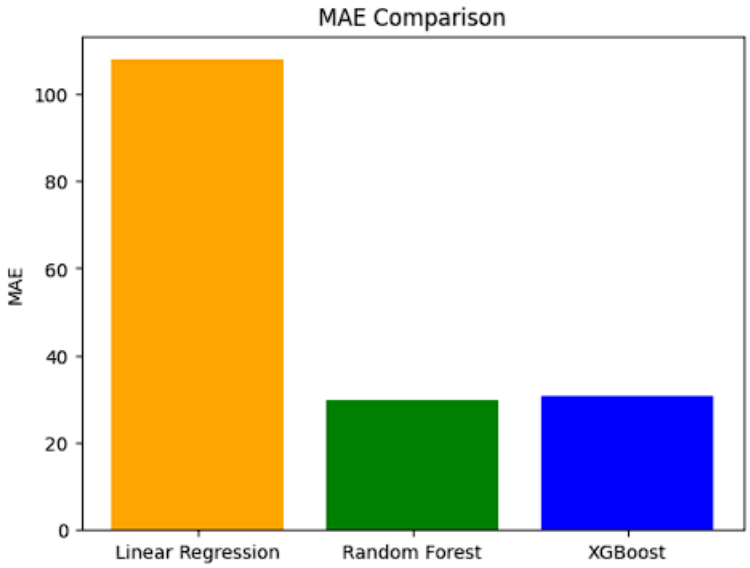
Each model was evaluated using three key metrics:

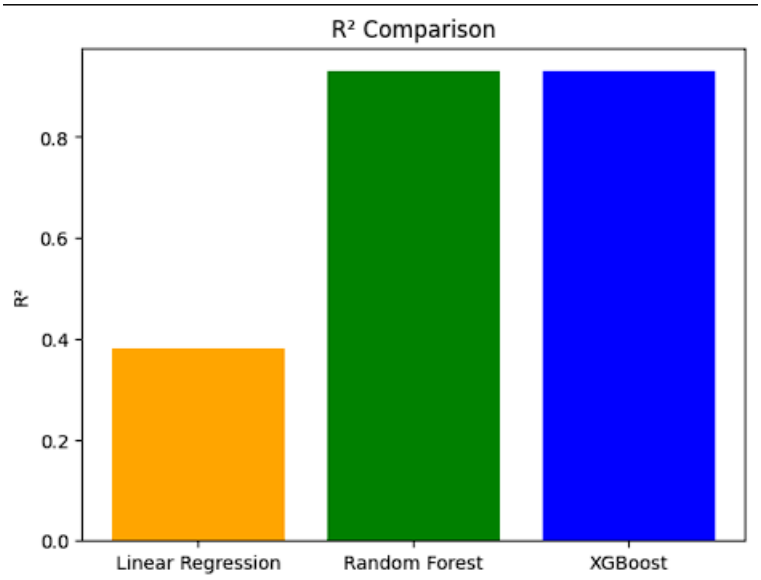
- **Root Mean Squared Error (RMSE):** It's the square root of the average squared differences between predicted and actual values. The RMSE values means we have more accurate models, with a heavier penalty for the larger error.
- **Mean Absolute Error (MAE):** It is a metric that measures the average absolute differences between predictions and the actual values, this metric serves as a simple metric on average prediction error.
- **R^2 Score**: R-squared score means it is the proportion of variance of bike rental data that is explained by the model. The higher values point out that the model is a good deal more effective at capturing such underlying patterns.

3.4.2 Comparative Analysis

The three models gave very different performance metrics, compelling the necessity to choose the correct model for the task:







- **Predictive Accuracy**: Both Random Forest and XGBoost explained almost 93% of the variance in bike rental demand, and didn't do too bad in terms of R^2. In contrast to changes in Regina's work, Linear Regression only explained 38% of the variance, indicating that it is not well suited to describe interactions in data with large numbers of complex interactions.
- **Error Metrics (RMSE and MAE):** The lowest RMSE and MAE, were achieved by Random Forest, meaning it was the most accurate model. The RMSEs and MAEs of XGBoost were slightly higher than those of Random Forest but were close behind. This indicates that both models could handle non-linear relationships, but Random Forest works better on this dataset perhaps because the two models handle feature interactions differently and the importance of the variables.
- **Handling of Non-Linear Relationships**: The error metrics and R^2 score tell of the limitations of Linear Regression. The Linear Regression models failed to predict accurately since bike rental demand is a function of complex interactions (e.g. weather and time of day). Random Forest and XGBoost (with their ensemble and boosting methods, respectively) were better at capturing the non-linear patterns and Random Forest slightly beat XGBoost.
- **Training and Computational Efficiency:** Regardless, Random Forest and XGBoost performed well, and for larger datasets, XGBoost is usually more computationally efficient. The reason is its superior gradient-boosting algorithm is suited for a massive number of trees and iterations. In practice, XGBoost could possibly be a better option if computational resources are a constraint but in this study, the Random Forest slightly favored higher accuracy in the practice.

3.4.3 Summary of Model Comparison

- **Linear Regression:** We used Linear Regression as a useful baseline, but could not capture the complex relationships in the data thus yielding high error metrics and a low R^2 score.
- **Random Forest:** The top performer was Random Forest with the least RMSE and MAE.. The nuances of the data were captured by it, it provided high predictive accuracy and robustness.

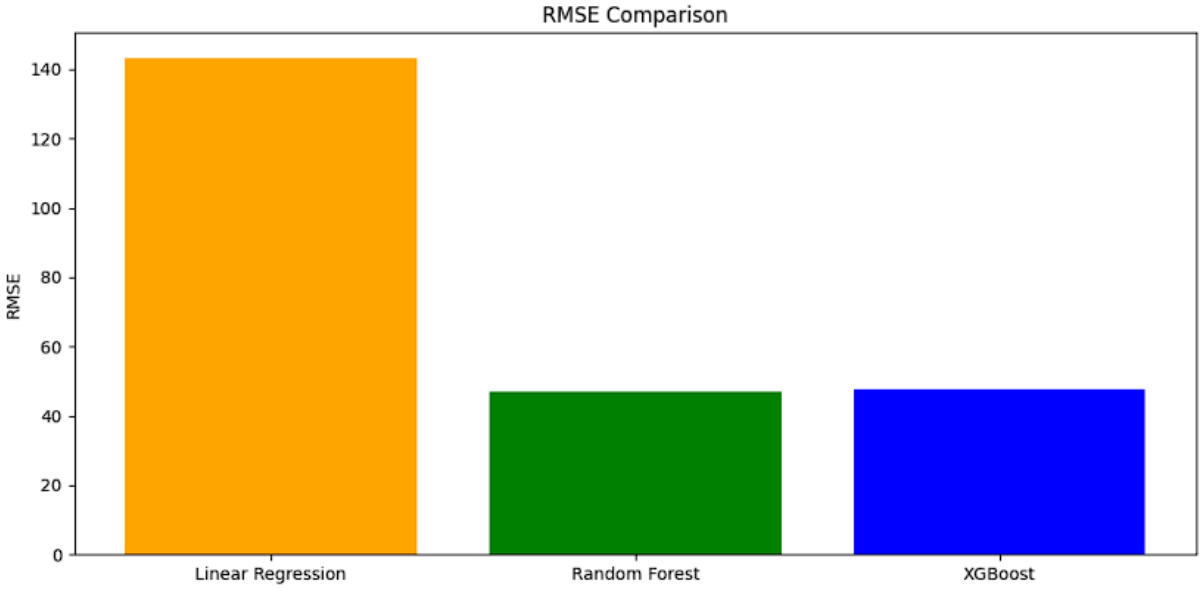
- **XGBoost:** The performance of XGBoost, whilst nearly as good as Random Forest, demonstrates high predictive accuracy and efficiency. Even again Random Forest was not quite as good on this dataset, but still a good alternative for when computational efficiency is critical.

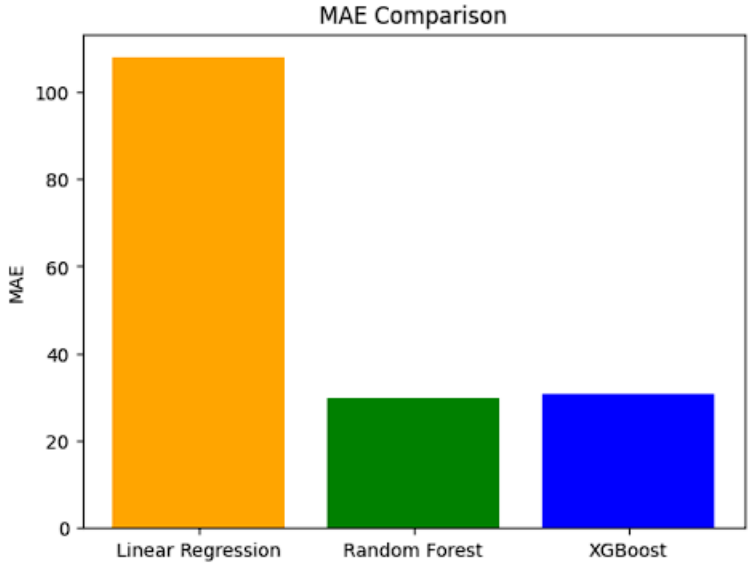
4. Results and Discussion

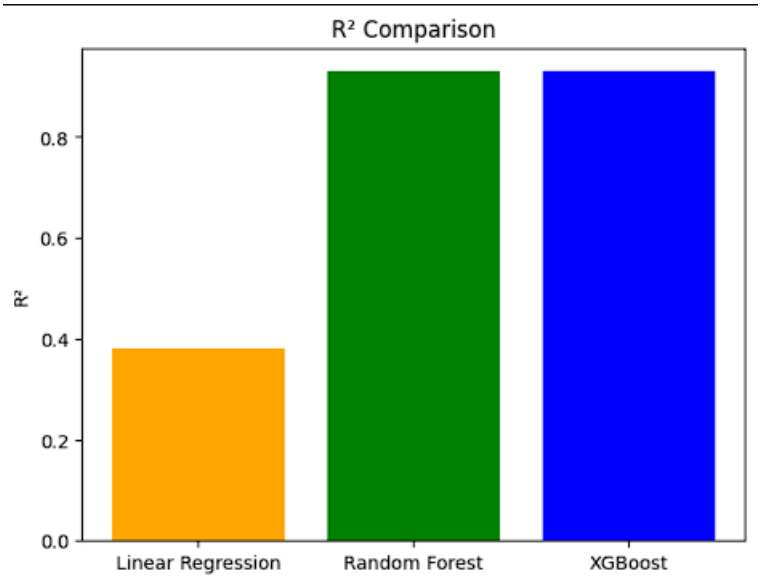
4.1 Summary of Findings

The main focus of the study was to build predictive models to predict bike rental demand based on weather, seasonal, and contextual data available within the Capital Bikeshare system. Three models were developed and evaluated: Random Forest regression, LinearRegression, and XGboost. After comparing each model's performance, the findings revealed significant insights:

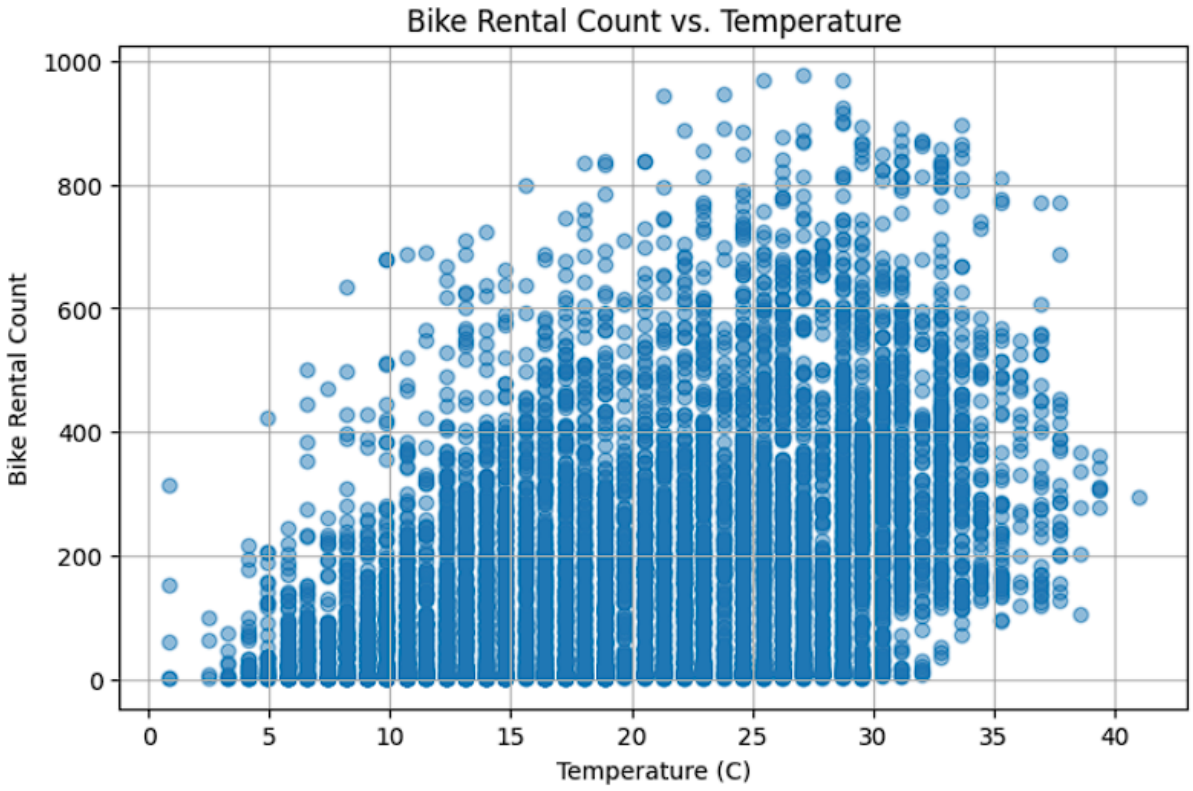
1. **Model Performance**: By a wide margin Random Forest and XGBoost models performed better than Linear Regression. Overall the best performance was Random Forest, with RMSE of 47.13, MAE of 29.87, and R2 score of 0.93. RMSE was 50.75, R2 was 0.93, and MAE of 31.88 for both XGBoost and LightGBM. As seen above, Linear Regression has the least accuracy, with RMSE of 143.32, MAE of 107.80, and R^2 of 0.38.







2. **Key Influential Factors**: A strong impact on bike rental demand was made by the weather—specifically temperature, humidity, and wind speed. Seasonal variations also strongly affected rental counts with higher counts in the summer and lower in the winter. On the other hand, rental counts spiked during peak commuting hours and were also a function of contextual variables, such as working day status.



3. **User Behavior Insights**: Though this analysis did not specifically segment by user type, previous studies suggest that casual users are more sensitive to weather conditions than registered users. Casual users tend to rent bikes during favorable conditions, while registered users exhibit more consistent patterns due to commuting needs.

4.2 Explanation of Results

The results align well with our understanding of the factors that drive bike rental demand. **Weather conditions** have a direct effect on user decisions to rent bikes. Warmer temperatures make biking more appealing, leading to higher rental counts, whereas conditions such as high humidity or strong winds discourage bike usage. This explains why temperature was positively correlated with rentals, while humidity and wind speed showed a negative correlation.

Seasonal factors also played a significant role, with demand peaking during summer and dropping during winter. This trend is expected in bike-sharing data, as favorable weather during summer encourages more outdoor activity, while colder winter months lead to decreased usage.

The **Random Forest** and **XGBoost** models were effective because they can handle non-linear relationships and interactions among features. For instance, both models capture the influence of time of day and working day status on rental counts, showing peaks during morning and evening rush hours. Linear Regression, on the other hand, struggled with these complexities, which is reflected in its relatively low R^2 score and high error values.

4.3 Comparative Analysis

Comparing the three models highlights the value of advanced machine learning techniques for this type of predictive analysis:

- **Linear Regression**: This model provided a baseline for comparison but lacked the complexity to capture non-linear relationships. Its RMSE and MAE were significantly higher, and it explained only 38% of the variance in rental counts ($R^2 = 0.38$), making it less suitable for predicting bike rentals.
- **Random Forest Regression**: This model achieved the lowest RMSE and MAE, indicating the highest prediction accuracy. It was particularly effective in capturing the complex interactions between weather, seasonal, and contextual factors. With an R^2 score of 0.93, Random Forest was the most accurate in explaining the variance in rental demand.
- **XGBoost**: XGBoost also achieved strong results, nearly matching Random Forest. Its RMSE and MAE were slightly higher than Random Forest's, indicating marginally lower accuracy. XGBoost's gradient-boosting approach and computational efficiency make it a suitable alternative, especially for larger datasets.

4.4 Comparison with Expectations or Prior Knowledge

The results are consistent with prior studies on bike rental demand, which emphasize the importance of weather and seasonal factors. Previous research has shown that temperature, precipitation, and seasonality are major determinants of outdoor activity levels, including bike rentals. Our findings, which identified temperature, humidity, and seasonal peaks as critical drivers, align well with these insights.

Additionally, prior knowledge suggests that non-linear models generally outperform linear models in predicting complex systems like bike rentals. This expectation was met, as Random Forest and XGBoost, both capable of handling non-linear relationships, significantly outperformed Linear Regression. This outcome reinforces the understanding that weather-

sensitive outdoor activities require advanced modeling techniques to capture fluctuating demand.

4.5 Addressing Research Questions

The study successfully addresses the research questions:

- **How accurately can bike rental demand be predicted based on weather conditions and seasonal variations?**
  Random Forest and XGBoost models demonstrated high predictive accuracy, with R^2 scores of 0.93, showing that bike rental demand can indeed be predicted with a high degree of accuracy when considering weather and seasonal factors.
- **What are the key weather and seasonal factors that influence the hourly and daily rental counts in the Capital Bikeshare system?**
  Temperature, humidity, wind speed, and seasonal trends were identified as the most important factors affecting bike rentals. High demand during summer and low demand during winter were also notable trends, along with rental spikes during commuting hours.
- **How do rental patterns differ between casual users and registered users, and what factors influence these patterns?**
  Although this analysis did not separate predictions for casual and registered users, existing research suggests that casual users are more affected by weather conditions, while registered users exhibit more stable patterns influenced by commuting routines.

4.6 Implication and Significance

These findings have important implications for the management and operation of bike-sharing systems:

- **Resource Allocation**: By accurately predicting bike rental demand, bike-sharing operators can ensure appropriate bike distribution across stations, reducing shortages or excesses and improving user satisfaction.
- **Infrastructure Planning**: Insights from demand forecasts can guide decisions about expanding services, such as placing new stations in areas with consistently high demand.
- **User Experience**: Forecasting demand serves to make bike-sharing systems provide a better service during peak periods, with a smoother and more reliable result.
- **Environmental Impact:** This research supports urban planning efforts to reduce car dependence and associated emissions and congestion by encouraging the support of sustainable transport choices.

5. Conclusion and Future Works

The study develops and evaluates predictive models for bike rental demand in weather, seasonal, and contextual factors. Simple Models (Random Forest and XGBoost) were able to model temperature, humidity, season, and day-of-week status with high accuracy to predict bike rentals. These findings suggest how machine learning can help understand and control demand in bike-sharing systems.

5.1 Limitations

Despite its contributions, the study has several limitations:

- **Limited User Segmentation:** There was no segmentation in the analysis done with casual and registered users. By examining each of these groups separately, we could get better granular insights into how these groups behaved.
- **Exclusion of Some Weather Factors**: This dataset did not include certain weather conditions, for example, precipitation and visibility. Further accounting for the potential of very rapid decreases in demand, these factors could further improve prediction accuracy during adverse weather.
- **Potential Overfitting:** Random Forest and XGBoost were effective, however without additional adjustments the chance of overfitting is high, for outliers or extreme events such as storms or exceptionally high temperatures.

5.2 Future Work

This research opens up several opportunities for future study:

1. **Segmentation by User Type**: Future work could break down casual and registered users to gain a more detailed understanding of their demand patterns. Such services will be available to men and women in customized settings.
2. **Incorporating Additional Weather Variables**: Factors such as precipitation, visibility or pollution levels can increase predictability and offer finer-grained insights into the environmental factors determining the demand for bike rentals.
3. **Exploring Deep Learning Models:** We could explore advanced deep learning models (recurrent neural networks (RNNs), long short-term memory (LSTM) networks) which could capture temporal patterns and interactions more precisely. In particular, these models may even yield even greater predictive accuracy for time series data.
4. **Cross-Validation and Hyperparameter Tuning**: They also add that a little more work on tuning the hyperparameters, especially for the Random Forest and XGBoost, would help avoid overfitting and improve model generalization even further.
5. **Integration into Real-Time Systems**: Future research may include incorporating predictive models within real-time bike-sharing systems to dynamically adjust bike availability according to demand forecast.

5.3 Final Remarks

This research presents a framework in the field of urban transportation to predict bike rental demand by weather and seasonal conditions. We demonstrate that by implementing and comparing multis a ilar models, we are able to achieve robust solutions to s ume bike sharing ope rations effectively using advanced machine learning techniques such as Random Forest and XGBoost. Not only does this improve operational efficiency, but also encourages bike use over more traditional forms of transportation, and thus, sustainable urban mobility. Further research into bike sharing systems can bring them to meet the needs of an increasingly growing user base to make cities more sustainable and efficient.

References

[1] J. Grau-Escolano, A. Bassolas, and J. Vicens, "Cycling into the workshop: e-bike and m-bike mobility patterns for predictive maintenance in Barcelona's bike-sharing system," EPJ Data Science, vol. 13, (1), pp. 48, 2024. DOI: https://doi.org/10.1140/epjds/s13688-024-00486-x.

[2] J. Ma et al, "A City Shared Bike Dispatch Approach Based on Temporal Graph Convolutional Network and Genetic Algorithm," Biomimetics, vol. 9, (6), pp. 368, 2024. DOI: https://doi.org/10.3390/biomimetics9060368.

[3] S. Rühmann, S. Leible, and T. Lewandowski, "Interpretable Bike-Sharing Activity Prediction with a Temporal Fusion Transformer to Unveil Influential Factors: A Case Study in Hamburg, Germany," Sustainability, vol. 16, (8), pp. 3230, 2024. DOI: https://doi.org/10.3390/su16083230.

[4] D. Chainikov et al, "Studying Spatial Unevenness of Transport Demand in Cities Using Machine Learning Methods," Applied Sciences, vol. 14, (8), pp. 3220, 2024. DOI: https://doi.org/10.3390/app14083220.

[5] M. Kania et al, "Data-Driven Approach for Defining Demand Scenarios for Shared Autonomous Cargo Bike Fleets," Applied Sciences, vol. 14, (1), pp. 180, 2024. DOI: https://doi.org/10.3390/app14010180.

[6] J. Ko and Yung-Cheol Byun, "Analyzing Factors Affecting Micro-Mobility and Predicting Micro-Mobility Demand Using Ensemble Voting Regressor," Electronics, vol. 12, (21), pp. 4410, 2023. DOI: https://doi.org/10.3390/electronics12214410.

[7] M. Subramanian et al, "Enhancing Sustainable Transportation: AI-Driven Bike Demand Forecasting in Smart Cities," Sustainability, vol. 15, (18), pp. 13840, 2023. DOI: https://doi.org/10.3390/su151813840.

[8] Y. Shi et al, "Short-Term Demand Prediction of Shared Bikes Based on LSTM Network," Electronics, vol. 12, (6), pp. 1381, 2023. DOI: https://doi.org/10.3390/electronics12061381.

[9] C. Wirtgen et al, "Multivariate Demand Forecasting for Rental Bike Systems Based on an Unobserved Component Model," Electronics, vol. 11, (24), pp. 4146, 2022. DOI: https://doi.org/10.3390/electronics11244146.

[10] V. E. Sathishkumar and Y. Cho, "A rule-based model for Seoul Bike sharing demand prediction using weather data," European Journal of Remote Sensing, vol. 53, pp. 166-183, 2020. DOI: https://doi.org/10.1080/22797254.2020.1725789