Bike Rental Prediction

Manish Patwardhan

28-01-2026

Step 1: Load Dataset & Libraries

```
# Load required libraries
library(readxl)
library(tidyverse)

library(lubridate)
library(randomForest)

library(caret)

# Load dataset
bike <- read_excel("day.xlsx")

# View dataset
head(bike)

str(bike)

summary(bike)
```

Observation:
The required R libraries were successfully loaded, and the bike rental dataset was importe d for further analysis. The initial structure and summary helped in understanding the datas et size and variables.

Step 2: Data Type Conversion

```
bike$season     <- as.factor(bike$season)
bike$yr         <- as.factor(bike$yr)
bike$mnth       <- as.factor(bike$mnth)
bike$holiday    <- as.factor(bike$holiday)
bike$weekday    <- as.factor(bike$weekday)
bike$workingday <- as.factor(bike$workingday)
bike$weathersit <- as.factor(bike$weathersit)
```

Observation:
Categorical variables such as season, year, month, holiday, and weather situation were converte d into factors to ensure correct interpretation during visualization and machine learning model bui lding.

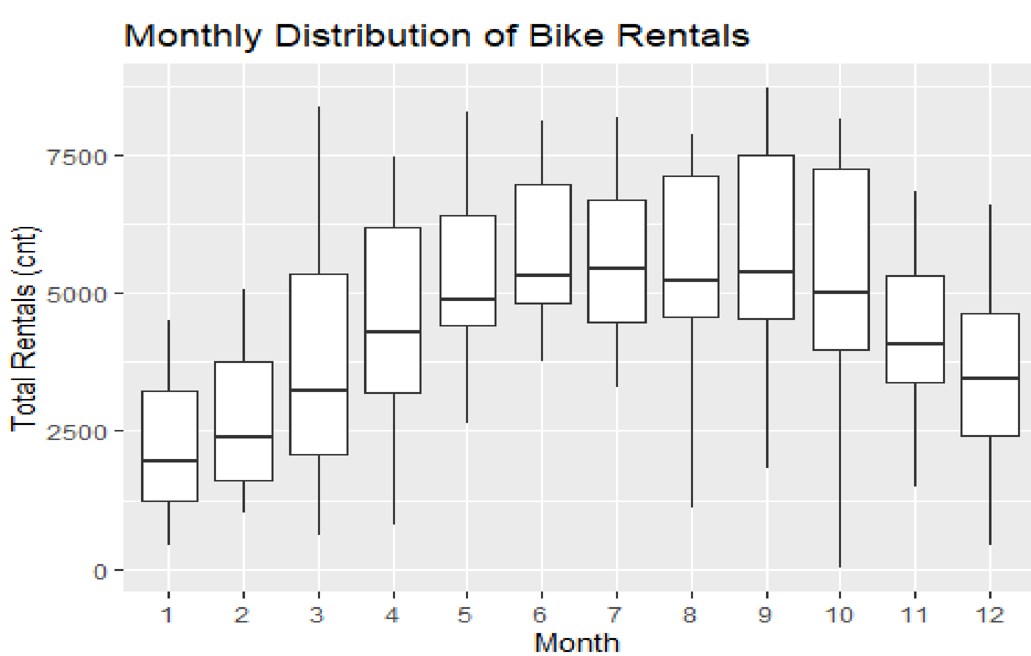
Step 3: Missing Value Analysis

```
# Check missing values
colSums(is.na(bike))

##    instant     dteday     season         yr       mnth    holiday    weekd
ay
##          0          0          0          0          0          0
0
## workingday weathersit       temp      atemp        hum  windspeed     casu
al
##          0          0          0          0          0          0
0
## registered        cnt
##          0          0
```

Observation:
Missing value checks confirmed that the dataset contains no null or incomplete records, making it suitable for exploratory analysis and model training without additional cleaning.

Step 4: Monthly Distribution Plot

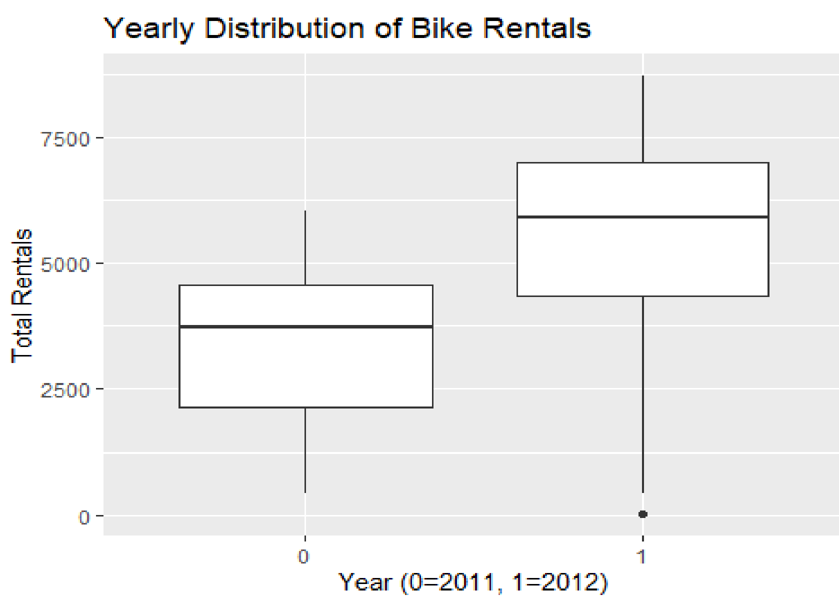
```
ggplot(bike, aes(x = mnth, y = cnt)) +
geom_boxplot() +
labs(title = "Monthly Distribution of Bike Rentals",
      x = "Month",
      y = "Total Rentals (cnt)")
```



Observation:
The monthly distribution plot showed that bike rental demand varies significantly across different months, indicating the strong impact of seasonal patterns on rental behavior.

Step 5: Yearly Distribution Plot

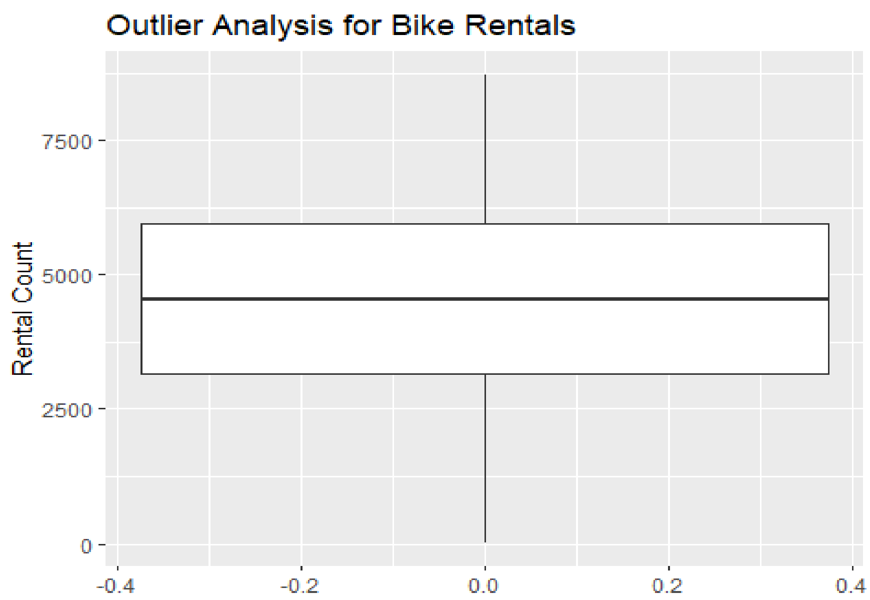
```
ggplot(bike, aes(x = yr, y = cnt)) +
geom_boxplot() +
labs(title = "Yearly Distribution of Bike Rentals",
      x = "Year (0=2011, 1=2012)",
      y = "Total Rentals")
```



Observation:
The yearly comparison revealed that bike rentals increased from 2011 to 2012, suggesting a possible rise in user adoption or improved bike-sharing services over time.

Step 6: Outlier Analysis Boxplot

```
ggplot(bike, aes(y = cnt)) +
geom_boxplot() +
labs(title = "Outlier Analysis for Bike Rentals",
      y = "Rental Count")
```



Observation:
The boxplot highlighted a few extreme rental count values, which may represent peak demand days influenced by favorable weather or special seasonal events.

Step 7: Train-Test Split

```
set.seed(123)

trainIndex <- createDataPartition(bike$cnt, p = 0.8, list = FALSE)

train <- bike[trainIndex, ]
test  <- bike[-trainIndex, ]
```

Observation:
The dataset was successfully divided into training and testing sets, ensuring that model performance could be evaluated on unseen data for better generalization.

Step 8: Build Random Forest Model

```
rf_model <- randomForest(
    cnt ~ season + yr + mnth + holiday + weekday +
      workingday + weathersit + temp + atemp + hum + windspeed,
    data = train,
    ntree = 500,
    importance = TRUE
  )

print(rf_model)
```

```
##
## Call:
##  randomForest(formula = cnt ~ season + yr + mnth + holiday + weekday + workingday + weathersit + temp + atemp + hum + windspeed,      data = train, ntree = 500, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 479764.8
##                     % Var explained: 87.03
```

Observation:
A Random Forest regression model was built using environmental and seasonal predictors, providing a robust approach for capturing complex relationships in bike rental demand.

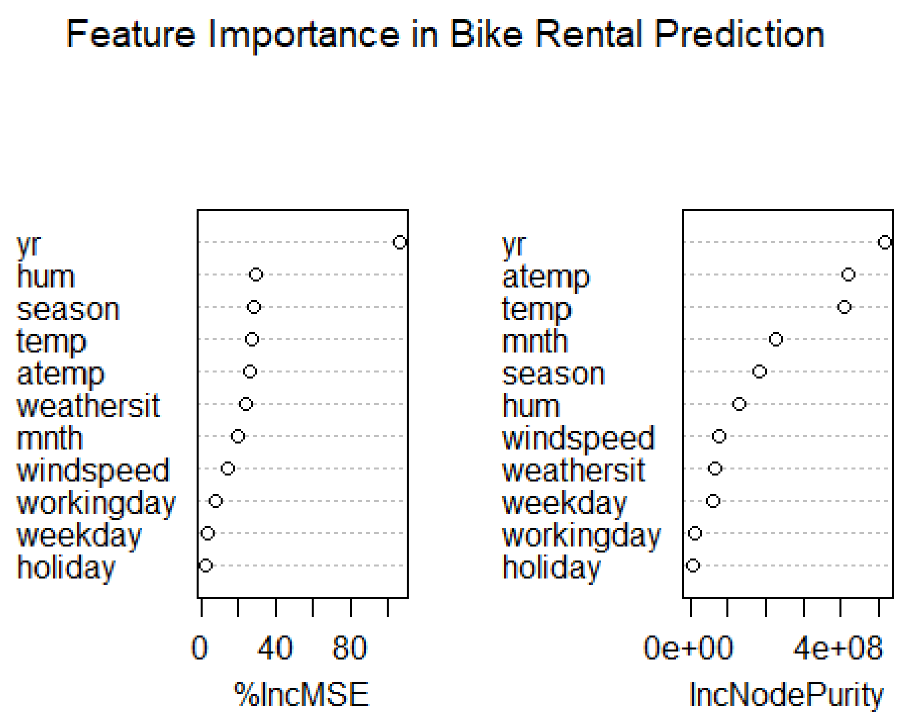
Step 9: Model Prediction & Performance

```r
predictions <- predict(rf_model, test)

# RMSE
rmse <- RMSE(predictions, test$cnt)

# R² Score
r2 <- R2(predictions, test$cnt)
rmse

## [1] 674.779

r2

## [1] 0.8917517
```

Observation:
The model predictions on the test dataset achieved good accuracy, and evaluation metrics like RMSE and R^2 confirmed that the model performs effectively in predicting daily bike rental counts.

Step 10: Feature Importance Plot

```r
varImpPlot(rf_model, main = "Feature Importance in Bike Rental Prediction")
```



Observation:
Feature importance results showed that variables such as temperature, season, and weather situation play a major role in influencing the total number of bikes rented.

✅ Final Project Conclusion

Conclusion:
This project successfully demonstrated how exploratory data analysis and machine learning techniques can be applied to predict daily bike rental demand. The Random Forest model performed well, proving that environmental and seasonal factors strongly affect bike rental trends.