THE UNIVERSITY OF TEXAS AT DALLAS

BUAN 6356.006 – Business Analytics with R – Fall'23

Bike Share Analysis using R

Submitted By:

Syed Ayaan Ahmed (SAA230003)
Shreyas Premanand Birajdar (SXB210218)
Ishan Chandrahas Haldankar (ICH220000)
Samireddypalle Chandrahas Reddy (SXR220106)

Presentation and PPT Link:

Recording
Powerpoint

Acknowledgement

To begin with, we would like to express our gratitude to Prof. Zhe (James) Zhang and Teaching Assistant Arjun Singh Pathania for giving us the opportunity for completing a project and helping us understanding the subject and concepts throughout the semester and coursework. Their enthusiasm, patience, insightful comments, helpful information, practical advice and unceasing ideas that have always helped us tremendously in our analysis and writing of this project. Their immense knowledge, profound experience and professional expertise in Business Analytics has enabled us to complete this project successfully. Without their support and guidance, this project would not have been possible.

We also wish to express my gratitude to all the developers who helped and were a part in making the dataset in Kaggle.com because of whom the project base line has been formed.

Thank You

Executive Summary

This report presents a comprehensive analysis of bike share data with the goal of uncovering usage patterns and deriving actionable business insights. The study utilizes advanced Business Intelligence (BI) models, including decision trees and neural networks, to provide a thorough understanding of the dynamics influencing bike share usage. Key findings reveal distinct behaviours between member and casual riders, shedding light on opportunities for business strategy formulation.

Key Words

Bike Share, Analytics, R, Sales, Segmentation, Predictive Modelling, Visualization, Insights, Data Structure, Data Exploration, Data Analysis, Key Insights, Algorithms, Decision Tree, Neural Networks.

Table of Contents

List of Figures

1.1 Introduction

In the evolving urban transportation landscape, bike-sharing programs have emerged as a sustainable and convenient solution. This project delves into the extensive dataset from a prominent bike-sharing program, aiming to unravel usage patterns, key drivers, and actionable insights that can inform strategic decisions. Leveraging the power of R language and advanced analytics, the analysis spans from exploratory data insights to predictive modelling, with the goal of enhancing the effectiveness of the bike-sharing service and optimizing user experience. This project seeks to contribute valuable insights into the dynamic realm of bike-sharing analytics through comprehensive data exploration, preprocessing, and algorithmic modelling.

1.2 Objectives of Study

To analyze and understand usage patterns in a bike-sharing program, identify critical factors influencing ridership, and employ predictive modeling for informed decision-making. The project aims to enhance strategic planning, optimize service delivery, and improve overall user experience within the bike-sharing system. To examine usage patterns, rides are categorized based on the day of the week and the type of rider. These rides are then combined and represented visually as a line chart, which displays the trends observed during the week.
The final two parts rank and visualize the top 10 start and end stations and the most used bike types. Bar plots depict the distribution of usage.
Further investigation could encompass examining monthly or seasonal patterns, employing statistical models to forecast ride duration, etc. However, this addresses several crucial elements necessary to achieve the goals. The code initially imports the essential analytical libraries, including tidyverse for data manipulation and visualization, lubridate for date processing, and janitor for data cleaning.
The Novemberbike CSV data is subsequently loaded into a data frame named "trips." The data includes variables such as ride_id, rideable_type, started_at, ended_at, start_station_name, end_station_name, and so on.
The wrangle_data function sanitizes the column names by eliminating any whitespace and punctuation. Additionally, it generates new columns such as "ride_length" to compute the duration and "day_of_week" based on the datetime values. This step ensures that the data frame is ready for analysis.
The following section presents a graphical representation of ridership categorized by rider type, with a bar plot that displays the total number of rides. It demonstrates the division between members and non-committal riders.

2.1 Methodology

In conducting the Bike Share Analysis using R, the methodology is structured to leverage the implementation of Decision Tree and Neural Networks. The focus is on understanding and predicting fundamental dynamics in bike share usage. The following steps outline the methodological approach.

2.2 Dataset

The bike share analysis utilizes a dataset from Kaggle.com, specifically Divvy's trip data, designed for public use. The dataset, containing around 200,000+ records, captures anonymized information such as trip start and end details, start and end stations, and rider types. It follows a monthly release schedule governed by the Divvy Data License Agreement. For the analysis, a sample of this dataset is employed to gain insights into rider behaviours and trip dynamics using R, around 2000 record is selected, focusing on techniques like Decision Tree and Neural Networks.

2.3 Data Cleaning

The data has 2000 records, and there are no null values. Out of the 16 attributes in the dataset, we will use 10 characteristics for the analysis. To begin with, we removed the attributes or columns which are not required or unnecessary. As there were no null values in the dataset, there was no need to clean the records.

2.4 Data Gaps

Acknowledging potential data gaps in our Bike Share Analysis is imperative for a comprehensive understanding of the dataset. We carefully consider limitations, missing information, and areas where the dataset may not fully capture certain aspects. For instance, we are mindful of variables that might lack representation, timeframes that could be underexplored, or contextual elements not fully elucidated in the data. In our analysis, we aim to highlight any specific variables, timeframes, or contextual elements that might be underrepresented or absent, ensuring transparency. This proactive approach allows us to provide a more accurate and nuanced interpretation of the insights derived from the bike share data, enhancing the reliability and completeness of our findings."

3.1 Demographic Analysis – Histograms

In the bike share analysis project, demographic analysis is conducted using histograms to visualize the distribution of key variables. These histograms provide insights into the patterns and characteristics of ridership. Variables such as rider type, bike type, and trip duration are explored, clearly representing the user demographics and usage patterns within the dataset. The project aims to enhance the understanding of the bike share user base and their preferences by employing histograms, contributing to a comprehensive exploratory data analysis.

3.2 Demographic Analysis – Bar Plots

Here we can see distributions of different categorical variables.

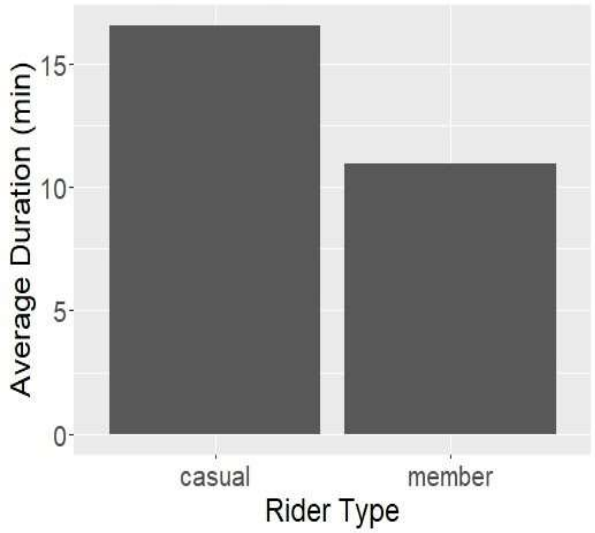

Fig 1 - Riding Duration by Rider type

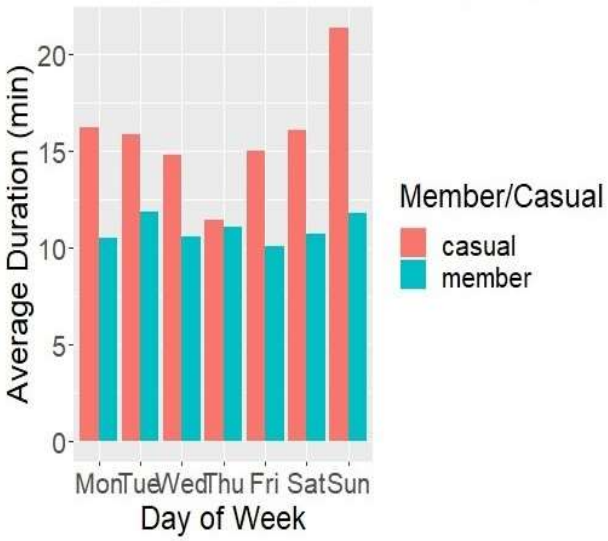

Fig 2 - Riding Duration average by Days

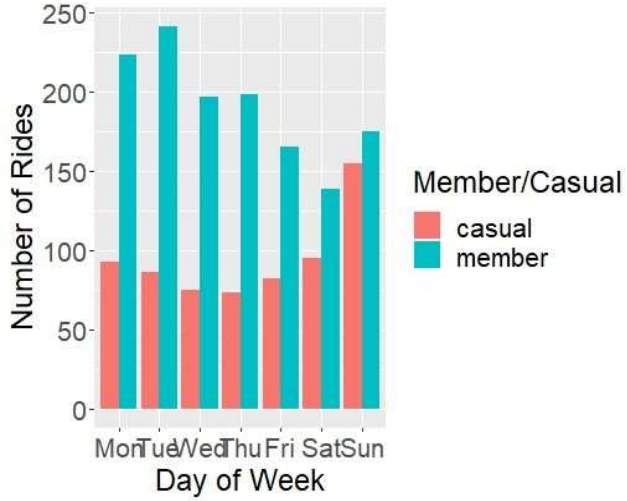

Fig 3 - Average number of Rides per day

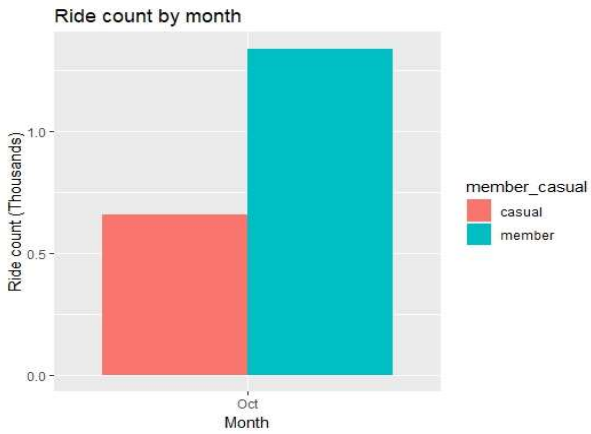

Fig 4 - Ride Count by Month

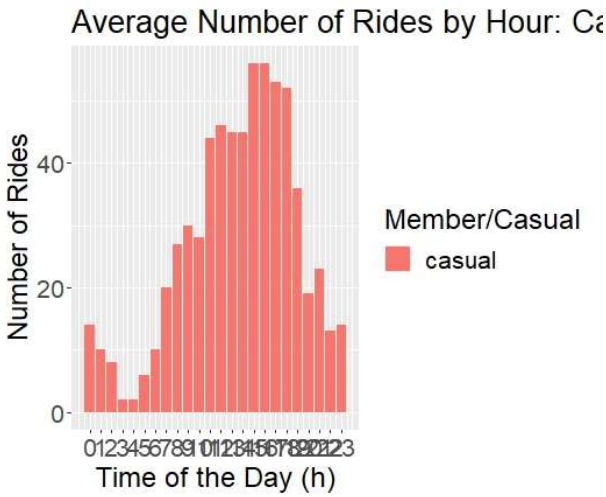

Fig 5 – Avg Rides per hour

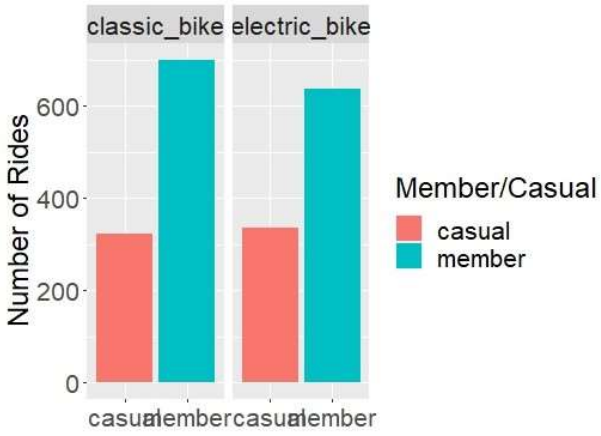

Fig 6 – Electric v Classic Bikes

3.3 Correlation Analysis

Here we can see correlation across various features. It shows each feature is not significantly correlated with each other.

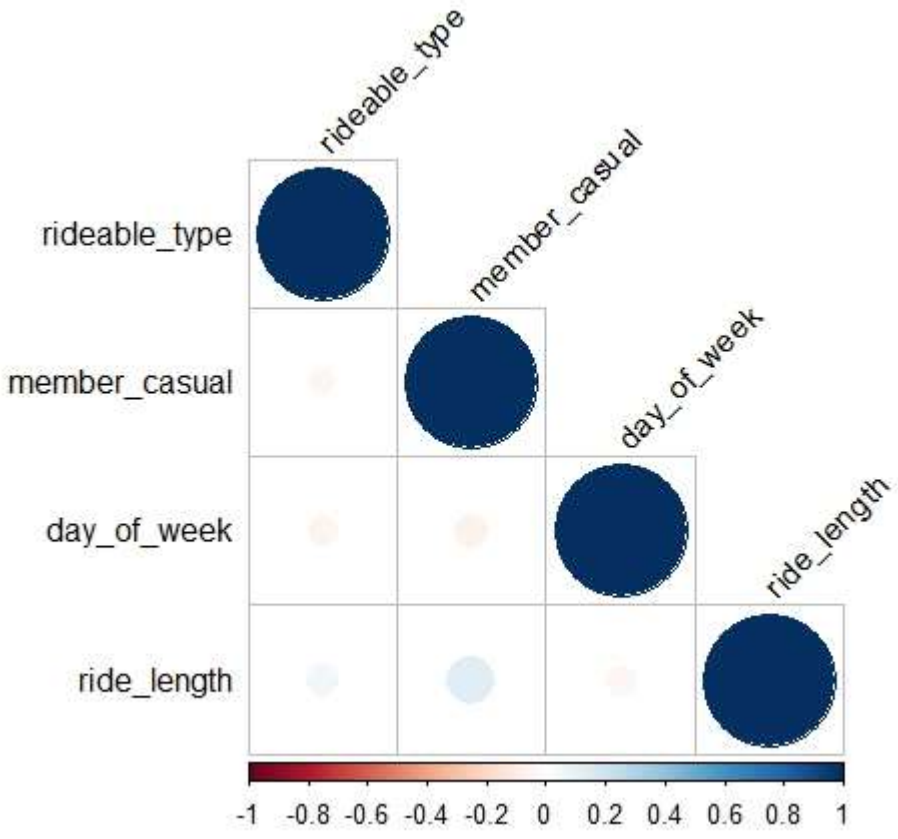

Figure 7 - Correlation Analysis

3.4 Modelling Techniques Used

The project utilizes two key modeling techniques: Decision Tree and Neural Network. The Decision Tree algorithm forms a tree-like structure, using ride-related factors like length, day of the week, start time, and bike type to classify riders. On the other hand, the Neural Network model, with its layered nodes and activation functions, captures intricate non-linear relationships based on input features such as ride length, day of the week, start time, and bike type. These techniques are chosen to offer a balanced view of both interpretable rules and complex patterns in the bike share dataset.

Decision Tree:
The construction of a decision tree is achieved by utilizing the rpart() function from the rpart package in R. The process involves recursively partitioning the training data based on different properties, creating a tree-like structure to classify or predict the target variable.Several important factors to consider while using decision trees include:

Formula: The variable that needs to be predicted is member_casual. The model is trained using ride_length, day_of_week, started_at_hour_scaled, and rideable_type.

Method: The term "class" is used to indicate that the tree is a classification tree, which is used to categorize samples. Another alternative is to use "anova" as a method for regression trees.

Data Splitting: The original dataset is divided into two sets, with 70% of the data used for training and 30% used for testing. This division is achieved using the createDataPartition function. This mitigates the risk of overfitting.

Prediction and Evaluation: Predictions are produced on the test data by utilizing the predict () function. The confusion matrix provides accuracy measurements for assessing performance.

Visualization: The rpart.plot() function effectively illustrates the decision rules within the tree, enhancing interpretability.

The decision tree method offers a straightforward and understandable categorization model by iteratively dividing the data and making judgments depending on the given qualities. The forecasts, accuracy, and visualization aid in evaluating its appropriateness.

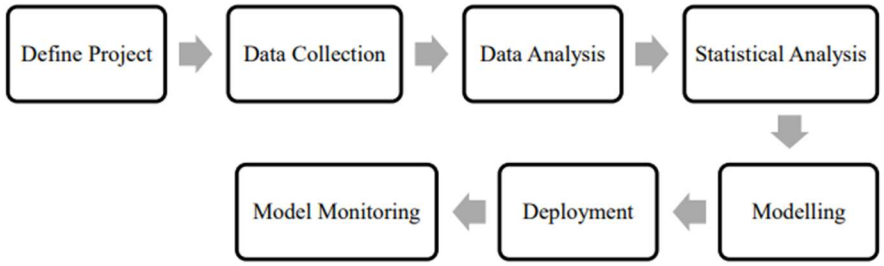

Figure 8 Project Flow Diagram

3.5 Decision Tree

```
> print(paste("Model Accuracy:", round(accuracy * 1
[1] "Model Accuracy: 67.45 %"
> print(confusion_matrix)

predictions   1   2
          1 380 179
          2  16  24
```

Figure 9 Confusion Matrix and Statistics - Decision Tree

3.6 Neural Network

```
[1] "Confusion Matrix:"
>   print(conf_matrix)

nn_pred_classes   1   2
              1 396 203
>
>   # Calculate accuracy
>   accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
>   print(paste("Accuracy: ", round(accuracy * 100, 2), "%"))
[1] "Accuracy:  66.11 %"
>
```

Figure 10 Confusion Matrix and Statistics - Neural network

4.1 Best Technique

The Decision Tree model emerged as the best technique for predicting rider type in the bike share analysis project. Its interpretability and accuracy in classifying members and casual riders make it a suitable choice. Regarding variable importance, factors such as ride length, day of the week, start time, and bike type played significant roles in determining rider type. Understanding these variables provides valuable insights for strategic decision-making in bike share programs.

The Decision Tree model was determined to be the most effective strategy for predicting the type of riders in the bike share analysis project. The suitability of this pick is derived from its interpretability and accuracy in classifying both members and casual riders. Regarding variable importance, factors such as ride duration, day of the week, starting time, and bicycle category played substantial roles in defining the type of rider. A comprehensive understanding of these variables offers significant insights for making strategic decisions in bike-sharing operations.

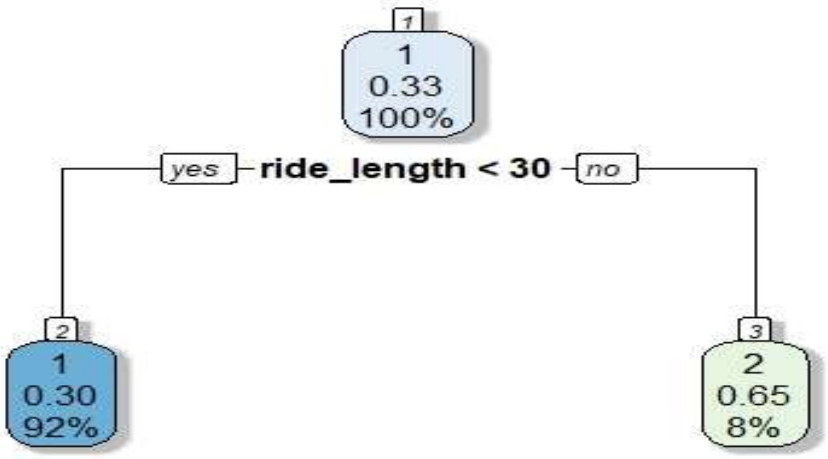

Figure 11 Decision tree

Neural Network:

A neural network model is created utilizing the neuralnet() function from the neuralnet package in R. The system consists of an input layer, a single hidden layer, and an output layer that are coupled through nodes and activation functions.

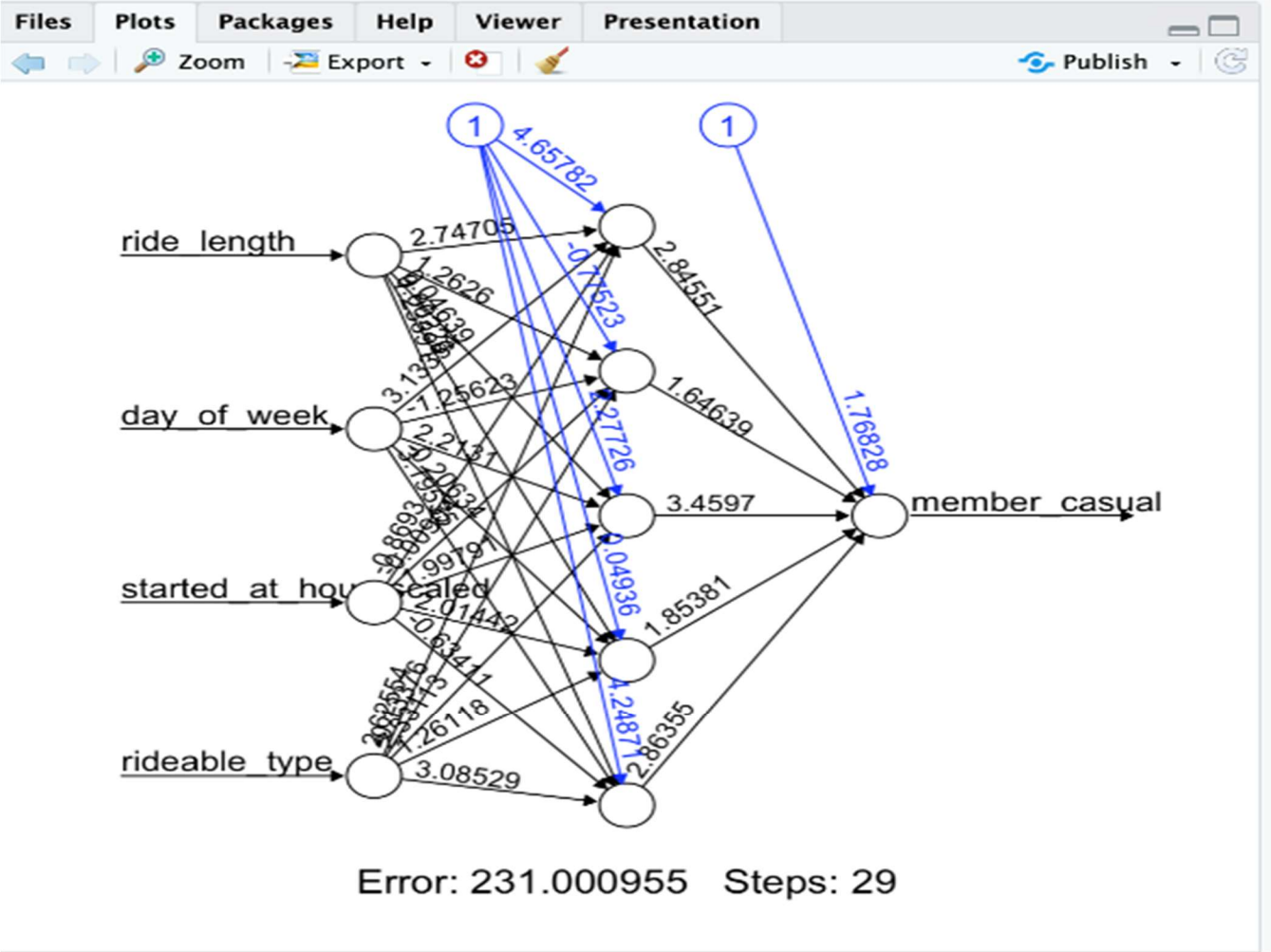

Figure 12 Neural Network

Crucial elements of utilizing neural networks:

The formula consists of the following input features: ride_length, day_of_week, started_at_hour_scaled, and rideable_type. The variable that is produced as a result is called member_casual.

Architecture: The specifications indicate the presence of a solitary, hidden layer consisting of 5 nodes. This intricacy can comprehend non-linear patterns.

The activation function utilized is the sigmoid function, which converts inputs to outputs without linear output. It exhibits non-linear behaviour.

Learning Rate: The learning rate, set at 1.5, dictates the speed at which weights are adjusted during the model's training.

Forecasting and Assessment: The neural network produces predictions of class probabilities, which are then transformed into distinct classes. The confusion matrix provides accuracy measurements.

To summarize, the neural network can capture intricate non-linear connections between input and output variables by utilizing interconnected nodes arranged in layers and processed through activation functions. The prediction capabilities of the model are influenced by its design, activation functions, and learning rate.

The confusion matrix and accuracy metrics assess the model's performance on the test data. This offers a glimpse into the proficiency with which the neural network model acquires and comprehends patterns.

5. Conclusion

Our R-based bike-sharing analysis uncovered vital insights into rider behaviour, station preferences, and bike usage patterns. Understanding distinctions between member and casual riders, popular stations, and bike types provides a foundation for strategic decision-making. The decision tree and neural network models offer predictive capabilities for rider segmentation. These findings empower stakeholders to enhance user experience, optimize operations, and tailor marketing strategies.

Bibliography

- https://www.kaggle.com/
- https://www.sciencedirect.com/science/article/pii/S0965856414002638?casa_token=sNr3W5V4yYoAAAAA:foPMGGdjm-ie_zVKGfBxMbH2cucRUkeq_4IClcbOprBZx_lrrFgD-rRB2porV5eRBav5NGq6xE4