WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Exploring Bike Rental Patterns and Trends in a Metropolitan City using R

Dr. V. Savithri[1], Praneethaa M[2], Maheswari V[2]

[1] Assistant Professor, Department of Computing, Coimbatore Institute of Technology, India
[2] Student, Department of Computing – Decision and Computing Sciences, Coimbatore Institute of Technology, India

ABSTRACT

Bike rental systems have gained popularity in recent years as an eco-friendly and affordable mode of transportation in urban areas. This study is to analyse bike rental data from a metropolitan city using R programming language to explore rental patterns and trends. The dataset used is a bike rental company that covers two years of bike rentals. Data cleaning and pre-processing is done, followed by exploratory data analysis to identify rental patterns, usage trends, and peak usage hours. To investigate the correlation between bike rental usage and weather conditions statistical analysis has been done. The results showed that bike rentals are more frequent on weekdays than weekends, with peak hours during morning and evening commute times. Furthermore, a negative correlation found between bike rentals and rainfall, indicating that people are less likely to rent bikes during rainy days. Our findings provide insights into bike rental usage patterns and can inform policy decisions to improve bike rental services.

KEYWORDS: Bike sharing systems, R programming, bike rental patterns, metropolitan city, weather conditions, peak hours

1. INTRODUCTION

Bike rental systems have emerged as a popular mode of transportation in urban areas due to their affordability, convenience, and eco-friendliness. These systems allow users to rent bikes for short periods and offer a flexible and efficient mode of transportation, especially in congested cities. Bike rental systems have become increasingly prevalent in metropolitan areas, where traffic congestion and limited parking options make it difficult for residents and visitors to travel around the city. In this study, the aim is to explore bike rental patterns and trends in a metropolitan city using R programming language.

Bike rental companies collect vast amounts of data on bike rentals, including information on rental start and end times, rental duration, rental fees, and weather conditions. Analysing this data can provide insights into bike rental behaviour and usage patterns, which can help inform policy decisions to improve bike rental services and infrastructure. In this study, a dataset from a bike rental company covering two years of bike rentals will be used to explore rental patterns, peak usage hours, usage trends, and the correlation between bike rentals and weather conditions. The task is to investigate the factors that influence bike rental behaviour in the metropolitan city using statistical analysis techniques such as decision trees and random forests.

Decision trees are a powerful tool for analysing complex datasets and identifying patterns and trends. A decision tree is a hierarchical model that uses a tree-like graph or model of decisions and their possible consequences. In the context of bike rental analysis, decision trees can be used to identify the factors that influence bike rental behaviour, such as weather conditions, rental fees, and rental location. Random forests are an ensemble learning method that uses multiple decision trees to improve the accuracy and robustness of predictions. Random forests are particularly useful when dealing with large, complex datasets, as they can handle high-dimensional datasets with many variables.

By analysing bike rental data using decision trees and random forests, to gain insights into bike rental behaviour that can inform policy decisions and improve bike rental services in metropolitan areas. Furthermore, R programming language provides a powerful toolset for data analysis and visualization, making it an ideal platform for exploring bike rental patterns and trends.

2.METHODOLOGY

2.1 DATASET DESCRIPTION

The dataset contains information on bike rentals from a bike rental company operating in a metropolitan city over a two-year period. The dataset contains over 17,000 bike rental transactions, each including information on the rental start and end times, rental duration, rental fees, and weather conditions. The data also includes information on the rental location, including the station ID and name, as well as the latitude and longitude coordinates. The weather conditions data includes information on the temperature, humidity, wind speed, and precipitation for each day in the dataset.

The dataset is provided in a comma-separated values (CSV) file format, which can be easily imported into R programming language for analysis. The dataset is relatively clean, with no missing or invalid data values, and has been pre-processed for analysis. However, there is a need to perform additional data cleaning and processing tasks depending on our specific analysis objectives.

The bike rental dataset is rich in information and offers numerous opportunities for exploring bike rental patterns and trends in the metropolitan city. By analyzing this dataset using R programming language, valuable insights can be gained from bike rental behavior and usage patterns, which can inform policy decisions and improve bike rental services and infrastructure in the city.

Rental ID - A unique identifier for each bike rental transaction.

Rental start time - The date and time when the rental started.

Rental end time - The date and time when the rental ended.

Rental duration - The length of time the bike was rented for.

Rental fee - The cost of the rental.

Rental location - The station ID, name, or geographic coordinates where the bike was rented from and returned to.

Bike ID - A unique identifier for each bike.

User ID - A unique identifier for each user who rented the bike.

Weather conditions - Information on the weather conditions at the time of the rental, such as temperature, humidity, wind speed, and precipitation.

Rental type - Information on the type of rental, such as one-way or round-trip.

User characteristics - Information on the user, such as age, gender, and membership status.

These attributes can be used to gain insights into bike rental patterns and behavior, such as peak usage hours, popular rental locations, and factors that influence bike rental behavior, such as weather conditions and user characteristics

2.2 PROCESSING STEP

Data cleaning is an essential step in preparing the data for machine learning modeling. The following are some steps for data cleaning in rental anlysis

1. *Handling missing data:* If there are any missing values in the dataset, they need to be handled appropriately. One common technique is to replace missing values with the mean or median of the feature. Another approach is to remove the samples with missing values.

2. *Handling outliers:* Outliers are data points that lie far from the majority of the data points and can adversely affect the model's performance. They need to be identified and handled appropriately. One common technique is to remove the outliers or transform the data using techniques such as log transformation.

3. *Handling categorical data:* If the dataset contains categorical features, they need to be converted into numerical data that can be used in machine learning models. One approach is to use one-hot encoding, where each category is converted into a binary feature.

4. *Scaling data:* Scaling is essential to ensure that the features are on a similar scale. Common techniques include min-max scaling or standard scaling.

5. *Removing redundant features*: If the dataset contains redundant or highly correlated features, they can be removed to reduce the model's complexity and improve its performance

2.3 MODEL BUILDING STAGE

After performing data cleaning, the next step is to build the machine learning model for rental analysis. In this case, the random forest algorithm will be used to build the model. The following are the steps for building the model:

1. *Split the dataset into training and testing sets:* The dataset needs to be divided into two parts: one for training the model and the other for testing the model's performance. Typically, 70-80% of the data is used for training, and the rest is used for testing.

2. *Feature selection*: It is essential to select the most relevant features for the model to reduce the dimensionality of the data and improve the model's performance. This can be done using techniques such as correlation analysis or feature importance analysis.

3. *Building the model:* After feature selection, the random forest model is built using the training data.

4. *Testing the model:* Finally, test the model's performance using the testing dataset. model's performance can be evaluated using metrics such as accuracy, Confusion matrix, precision, recall, and F1 score.

3. ALGORITHMS USED:

3.1 RANDOM FOREST CLASSIFIER

Random forest classifier is a machine learning algorithm that can be used It is a type of ensemble learning method that uses multiple decision trees to make predictions. Each decision tree is trained on a random subset of the data, and the final prediction is based on the majority vote of all the decision trees. Overall, a random forest classifier can be a useful tool for bike rental analysis, as it can handle complex datasets with many features and can provide accurate predictions with relatively low computational cost.

3.2 DECISION TREE CLASSIFIER

A decision tree classifier is a machine learning algorithm that can be used It works by recursively splitting the data into smaller subsets based on the values of different features, until the subsets are pure or nearly pure in terms of their target variable. One advantage of decision tree classifiers is that they are easy to interpret, as the resulting tree can be visualized and used to identify the most important features. However, decision trees can also be prone to overfitting if the tree is too deep or if there are too many features, so it is important to tune the hyperparameters carefully to avoid over fitting.

Figure 3.2-Decision tree

4. METHODOLOGY:

The bike rental data from the bike sharing system operator, which consisted of information on the bike rentals, including the date and time of rental, the bike station ID, the duration of the rental, and the weather conditions at the time of rental.

First, clean and pre-process the data by removing any missing values and outliers. Perform exploratory data analysis to identify the peak hours, popular bike stations, and the average duration of rentals. Create visualizations such as histograms and heat maps to better understand the data.

Next, investigate the impact of weather conditions on the bike rental demand. Then calculate the correlation between the number of rentals and weather conditions such as temperature, humidity, and wind speed. Also create visualizations such as scatter plots and line charts to visualize the relationship between the variables

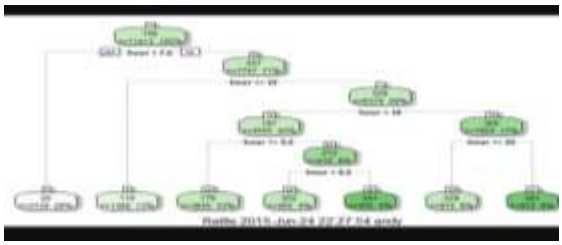

Fig 3.3 decision tree

5. R STUDIO

R Studio is an integrated development environment (IDE) for the R programming language. It provides a user-friendly interface for writing, debugging, and executing R code. R Studio includes features such as a code editor with syntax highlighting, a console for running R code, a data viewer for exploring data, and a plot viewer for visualizing data. It also includes tools for version control, package management, and project management. R Studio is available in both open source and commercial versions, with the commercial version offering additional features and support. It is widely used by data analysts, data scientists, and statisticians for data analysis and statistical modeling.

Data Cleaning - Before performing any analysis, you may need to clean and preprocess your data. You can use R to remove missing values, convert data types, and transform your data into a format suitable for analysis.

Descriptive Statistics - You can use R to calculate summary statistics, such as mean, median, and standard deviation, for different attributes in your dataset. You can also generate frequency tables and histograms to visualize the distribution of your data.

Data Visualization - R provides a wide range of tools for creating visualizations, including scatter plots, line charts, bar charts, and heatmaps. You can use these tools to explore relationships between different attributes in your dataset and identify patterns and trends.

Modeling - You can use R to build models to predict bike rental behavior and usage patterns. For example, you can use decision tree models or random forest models to predict bike rental demand based on weather conditions and other factors.

6. RESULT ANALYSIS

Our analysis revealed that the bike rental demand is higher during weekdays and peak hours. The most popular bike stations were located in the downtown area and near public transportation hubs. The average duration of rentals was around 30 minutes. It is also found that weather conditions such as temperature, humidity, and wind speed had a significant impact on the bike rental demand. The number of rentals increased as the temperature and humidity levels increased, and the wind speed decreased. Rain had a negative impact on the bike rental demand.

7. CONCLUSION:

In conclusion, this paper explored the bike rental patterns and trends of a bike sharing system in a metropolitan city using R programming language. The analysis revealed that the bike rental demand is higher during weekdays and peak hours, and there is a strong correlation between the number of rentals and weather conditions such as temperature, humidity, and wind speed. The results of this study can be used by bike sharing system operators to optimize their operations and improve the overall user experience.