CSP 571 Data Preparation and Analysis

Project Report on

**Analyzing Divvy Bike Usage Patterns in Chicago**

|  |  |  |
| --- | --- | --- |
| Mohammad Hamza Piracha  mpiracha@hawk.iit.edu | Poshan Pandey  Ppandey6@hawk.iit.edu | Usman Matheen Hameed  uhameed@hawk.iit.edu |
| A20554741 | A20519852 | A20564338 |

Abstract -

# Introduction

## Objective

## Literature Review

# Methodology

## Data Preparation

Our dataset covers ride data for the year 2023, with individual datasets downloaded monthly. These datasets were combined into a single file (merged\_data.csv) to facilitate thorough analysis. The combined dataset was then prepared for further cleaning and analysis to extract insights into ridership patterns and trends.

## Data Cleaning and Transformation

### Data Cleaning

The dataset for the year 2023 contains comprehensive ride data, encompassing details from each month. To ensure the dataset's quality and relevance, several key actions were undertaken during the cleaning process:

* Unnecessary columns were removed, focusing on essential variables such as ride ID, rideable type, timestamps, station locations, and rider type (member or casual).
* Missing values were systematically addressed through the removal of incomplete data points.
* Timestamps for ride start and end times were standardized into datetime format to facilitate accurate time-based analysis.
* Essential features like trip duration, day of the week, and hour of the day were derived to enhance the dataset's analytical depth.
* Outliers, including negative durations and trips exceeding 24 hours, were filtered out to maintain dataset integrity.
* Checks for data consistency ensured that each ride's end time logically followed its start time.

These actions collectively ensure that the cleaned dataset is robust and ready for in-depth analysis of bike-sharing behaviors and trends in 2023.

### Outlier Detection

Apart from general analysis and missing value reduction, Outlier detection and removal is also performed using these following approaches:

#### **Z-Score Method**

The Z-score method standardizes the dataset and identifies outliers based on a threshold, usually 3 standard deviations from the mean.

#### Isolation Forest

Isolation Forest is a machine learning algorithm that isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

### Data Transformation

Following approaches were employed for data transformation:

#### Normalization and Standardization

**Normalization**: Rescales the data to a range of [0, 1]. **Standardization**: Centers the data to have a mean of 0 and standard deviation of 1.

#### Log Transformation

Log transformation can help stabilize variance and make the data more normally distributed.

#### Date and Time Features

Extracting useful features from datetime columns, such as the day of the week, month, hour, etc.

#### Feature Engineering

**Trip Duration in Different Units**: Convert trip duration to hours, days, etc.

**Time of Day Segmentation**: Segment the day into periods (e.g., morning, afternoon, evening, night).

**Time of Year Segmentation**: Segment the ride into periods (e.g., summer, winter, evening, fall, spring).

### Distribution Analysis

Distribution analysis refers to the process of examining the statistical distribution of data within a dataset. Extensive usage of histograms and graphs for distribution analysis.

Analysis Used in Distribution Analysis:

1. Violin Plot for Member Type vs Trip Duration
2. Rug Plot of Trip Duration
3. Scatterplot Matrix for time duration
4. Member vs Casual Riders by Ride Type for Each Month
5. Member vs Casual Riders by Seasons
6. Time of Day Distribution
7. Starting Station Distribution (Top 10)
8. End Station Distribution (Top 10)

## Data Analysis and Modelling

### Clustering

### Dimensionality Reduction

### Feature Selection

Feature selection is a critical process in data analysis and modeling, aimed at identifying the most relevant features that contribute to the predictive power of a model. Initially, a correlation analysis is conducted to identify and remove highly correlated features, reducing redundancy in the dataset. Subsequently, recursive feature elimination (RFE) is applied to further refine the feature set. RFE works by recursively fitting a model and removing the least important features, as determined by the model's performance. This iterative process continues until the optimal subset of features is identified, balancing model accuracy and complexity. To handle large datasets efficiently and manage memory usage, the data.table package is used. data.table provides optimized and fast data manipulation capabilities, which are particularly useful when dealing with large volumes of data. By selecting a subset of the most significant features and using data.table for efficient data processing, we enhance the model's interpretability and efficiency, ultimately improving its predictive performance.

### Model Selection and Training

In our analysis, we focused on two robust machine learning algorithms: Random Forest and XGBoost. These models were chosen for their effectiveness in handling complex datasets and their proven track record in delivering high predictive accuracy.

Random Forest: an ensemble learning method, constructs multiple decision trees during training and outputs the average prediction of the individual trees. This approach reduces overfitting and improves generalization, making it suitable for our dataset, which includes numerous features and potential interactions.

XGBoost, on the other hand, is a powerful gradient boosting algorithm known for its speed and performance. It builds models in a stage-wise manner and optimizes the model by minimizing the loss function, thus enhancing predictive accuracy and robustness.

Both models were evaluated using a subset of the selected features from our dataset. The performance of the models was assessed using Root Mean Squared Error (RMSE) and R-squared metrics. Random Forest demonstrated an RMSE of [rf\_rmse] and an R-squared value of [rf\_r2], indicating its capability to explain a significant proportion of the variance in the trip duration data. XGBoost, with an RMSE of [xgb\_rmse] and an R-squared value of [xgb\_r2], also performed admirably, showcasing its efficiency in capturing complex patterns within the data. The comparison revealed that while both models provide strong predictive power, Random Forest slightly outperformed XGBoost in terms of RMSE and R-squared, making it the preferred choice for this analysis.

## Software Packages and Tools

* **R Language:** The primary programming language for analysis.
* **R Studio:** Integrated development environment for R.

**Libraries:**

* dplyr for data manipulation
* ggplot2 for data visualization
* tidyr for data tidying
* dbscan for outlier detection
* isotree for outlier detection
* lubridate for date and time handling
* caret for machine learning model training and evaluation (hopefully)
* more on implementation

This project will leverage these tools and methodologies to provide actionable insights into the Divvy bike-sharing system in Chicago, ultimately aiming to enhance its efficiency and user satisfaction.

# Conclusion

# Future Scope

* **Integration with Public Transit:** Analyze the relationship between Divvy bike docks and CTA bus stops to optimize the multimodal transportation network in Chicago.
* **Population Analysis:** Correlate Divvy usage data with population density data to identify areas with unmet demand for bike docks and cycles.
* **Expansion to Other Cities:** Apply the analysis methodology to other cities with bike-sharing programs to compare and improve overall urban mobility.
* **Real-Time Data Integration:** Incorporate real-time data feeds to provide dynamic recommendations for bike redistribution and dock availability.
* **User Experience Enhancement:** Analyze user feedback and usage patterns to improve overall user experience.
* **Sustainability Impact:** Evaluate the environmental benefits of the bike-sharing program and suggest improvements for increasing its positive impact on urban sustainability.

# References:

L. Czarlnski, "Exploratory Data Analysis (EDA) of the Chicago Divvy Bikes Dataset," Medium. [Online]. Available: <https://medium.com/@leonczarlnski/exploratory-data-analysis-eda-of-the-chicago-divvy-bikes-dataset>.

"Exploring variations in Divvy bike station usage volume: from historical trip records to Google Street view images," MACS 37000 (Spring 2021) Thinking with Deep Learning for Complex Social & Cultural Data Analysis, uchicago.edu. [Online]. Available: <https://uchicago.edu/macs37000/divvy-bike-station-usage>.

"Divvy Trips," City of Chicago, Data Portal. [Online]. Available: <https://data.cityofchicago.org/Transportation/Divvy-Trips>.

Shivaniwac, "Quarterly Success: Divvy Bike’s 2024 Growth Analysis," Medium, May 2024. [Online]. Available: <https://medium.com/@shivaniwac/quarterly-success-divvy-bikes-2024-growth-analysis-e49927841eaf>.

"Index of bucket 'divvy-tripdata'," [Online]. Available: [https://divvy-tripdata.s3.amazonaws.com/index.html.](https://divvy-tripdata.s3.amazonaws.com/index.html.%20)

A screenshot of a computer

Description automatically generated