Using real-time data to monitor and optimize fleet operations

Group: Team zehn

Members: Ercan Tazegül, Simon Knecht, Diego , Marc Asbach, Vicent Sedaletec

Instructor: Prof. Dr. Wolfgang Ketter, Nastaran Naseri

Subject: Analytics and Application

University of Cologne

31.01.2022

Contents

List of Tables

List of Figures

1 Summary

The objective of this project is to study the 2018 Divvy Bikes Chicago bike ride dataset, which comprises two datasets: one containing data on Chicago bike rentals in 2018, and the other containing hourly weather data for 2018 obtained through the weather.com API. In order to understand and optimize the performance of the bike fleet, we have defined key performance indicators (KPIs) and analyzed the datasets for temporal and spatial demand patterns. Cluster analysis was used to identify recurring patterns and inform business decision-making. Furthermore, we have applied predictive analysis techniques, such as scientific forecasting models, to forecast future demand and optimize operations.

2 Problem Description

Transport-related greenhouse gas emissions account for a large share of total emissions in the EU, and it is widely recognized that our approach to mobility needs to change in order to achieve our decarbonization goals. Traditional urban mobility is mainly based on internal combustion engine vehicles, which have four negative impacts: Contribution to global greenhouse gas emissions, pollution with serious health risks for urban populations, high accident rate with nearly 1.3 million fatal accidents annually worldwide, and inefficient use of motor vehicles with low occupancy and high space requirements for roads and parking, and traffic congestion. The need for a major transformation of the mobility system has been recognized, and the mobility landscape is changing rapidly, with the important trend of Mobility-as-a-Service (MaaS) and On-Demand (MoD), as well as the use of bikesharing platforms and similar platforms for other modes such as cars, mopeds, and e-scooters. This project examines 2018 data from the fleet operator of bike-sharing company Divvy Bikes Chicago. Real-time data streams are used to monitor and optimize operations, increase profitability, and improve service levels. The goal is to demonstrate that Data Science can benefit society. To this end, two core aspects are focused on. The first core aspect is system monitoring to make sustainable and profitable business and operational decisions. The second core aspect is demand forecasting to improve service levels by repositioning bikes or providing additional bikes. The objective is to optimize bike-sharing for sustainability and efficiency, enabling urban populations to utilize the mode of transportation for their commuting needs, thus reducing air pollution in cities and promoting sustainable living.

3 Data Collection and Preparation

3.1 Bikesharing Data

The dataset contains bike sharing data from Divvy Bikes Chicago from 2018. The overview of the variables in the dataset can be seen in Table 1.

Table 1: Description of bikeshare dataset columns

Variable name	**Format**	**Description**
start time	datetime	Day and time trip started
end time	datetime	Day and time trip ended
start station id	int	Unique ID of station where trip originated
end station id	int	Unique ID of station where trip terminated
start station name	name	str Name of station where trip originated
end station name	name	str Name of station where trip terminated
bike id	int	Unique ID attached to each bike
user type	User	membership type

In the preprocessing phase of the data, we first eliminated duplicate entries by keeping only the last occurrence and discarding the first. Then we filtered out cases where the starttime was later than the endtime. We also ensured that each stationname was uniquely associated with a single stationid and vice versa, and that the maximum number of stationid associated with a single stationname and vice versa did not exceed 1. To merge cases where a single starting station name was associated with multiple starting station IDs (and vice versa), we implemented a merging procedure. We then calculated the average duration and the total duration of all trips, including the 0.999 quantile of trip times to mitigate the effects of unrealistic outliers. Finally, we added two new columns to the data: 'nextride' and 'nextbike', which indicate whether a particular bike ID is used during a given period, and the identity of the next ride for each bike, respectively. After that, we still included the trips that occurred before and after.

3.2 Geological Data

3.3 Wheater Data

In order to improve the quality of the weather data, we removed all rows containing NaN values. The hottest and coolest temperatures recorded in the dataset appeared to be reasonable, so no further removal was necessary. We then identified 1328 duplicates in the dataset and decided to retain only the last recorded entry for duplicates, as this is generally considered the most reliable in such situations. There were also several rows with data for the same time, so we chose to take the average and remove the duplicates. The earliest recorded date in 2018 was January 1 at midnight, while the latest was December 31 at

The weather data is provided by the wheater.com API and contains the variables listed in Table 2.

Table 2: Description of weather dataset columns

Variable name	**Format**	**Description**
date time	datetime	Day and time of measurement
max temp	float	Maximum temperature recorded in degC
min temp	float	Minimum temperature recorded in degC
precip	int	Binary indicator for precipitation (1=yes,0=no)

11pm. During this time period, 623 hours of data were missing, which is almost 26 days. Ultimately, we decided to estimate the missing data and found that there are missing data every month, distributed throughout the year. There are only a few sequences that are longer than 1, with a maximum length of 6. In the worst case, we therefore do not have data for a period of 6 hours. Taking into account the above arguments, we have decided that it should be possible to estimate the weather for the missing data without making overly inaccurate estimates.

4 Descriptive Analysis

4.1 Bikesharing Data

4.2 Bikesharing Data

4.3 Bikesharing Data

Z

5 Cluster Analysis

5.1 Bikesharing Data

5.2 Bikesharing Data

Z

6 Predictive Analytics

z

7 Conclusions

Z