Introduction to Data Science and Machine Learning - Summer Semester 2022

Bachelor of Science WI / IS
Faculty of Management, Economics, and Social Sciences
Department of Information Systems for Sustainable Society
University of Cologne

Instructor Prof. Dr. Wolfgang Ketter **Term** SS 2022
TA Karsten Schroer, Philipp K. Peter **Website** `www.is3.uni-koeln.de` and ILIAS

Team Assignment

This DSML team project is designed to test a representative cross-section of the data analytics and machine learning approaches we cover during this course. It is based on a real-world problem with high relevance to the current hot topic of smart mobility systems and will act as an illustration of how we can use data in impactful ways to address pressing societal issues.

1 Background

Transport-related greenhouse gas emissions make up for the second largest chunk of total EU emissions. It has thus long been recognized that in order to meet decarbonization targets our approach to mobility will have to change. To this day traditional urban mobility relies primarily on internal combustion (IC) engine vehicles. This mobility setup brings with it four well-known social negatives. First, traditional road transport contributes substantially to the global GHG emission balance sheet. Second, pollution in the form of NOx, HC, PM and other emissions poses serious health hazards to urban populations. Third, road traffic is a major safety concern with close to 1.3m people dying in road accidents each year across the globe. Finally, road transport is highly inefficient, as utilization of passenger cars is low, thus requiring many cars to provide mobility to comparatively small numbers of passengers. This results in massive space requirements for roads and parking as well as traffic congestion. The need for a comprehensive transformation of the mobility system has been recognized and the mobility landscape is changing fast. A crucial trend in this newly emerging ecosystem is the consumption of mobility as-a-service (MaaS) and on-demand (MoD) heralding in the age of shared, fleet-based transportation companies. Bikesharing platforms are an excellent manifestations of a MaaS and MoD. Similar platforms are also getting traction for other transport modes such as cars, mopeds and more recently e-scooters (e.g. Lime and Bird).

In this project we investigate how fleet operators can make use of increasingly ubiquitous real-time data streams to monitor and optimize their operations, boost profitability and increase service level. The underlying assumption is that by enabling fleet operators to do well in their operations, data science can enable them to do good for society (”Doing well by doing good”).

We focus on two core aspects that are of interest to fleet operators:

1. **Customer Understanding**: A deep understanding of customer behavior and how bikes are used to address daily transportation needs can direct marketing efforts and focus operational priorities of fleet operators.
2. **Demand prediction**: An accurate prediction of future demand is an important step towards providing a high service level (e.g. by deploying additional bikes or by re-positioning vehicles towards areas of high demand, etc.)

2 Description of Dataset

You have been allocated datasets of bikesharing rentals in four major US cities for a period of one year each (see Section 4 for details on dataset allocation). This data was collected via the open trip history data of Bay Wheels in the Bay Area (San Francisco, Berkeley, etc.), Blue Bikes Boston, Divvy Bikes Chicago, Ride Indego Philadelphia and Bikeshare Metro in Los Angeles. More details on the datasets can be found on their respective websites:

- https://www.lyft.com/bikes/bay-wheels/system-data
- https://www.bluebikes.com/system-data
- https://www.divvybikes.com/system-data
- https://www.rideindego.com/about/data/
- https://bikeshare.metro.net/about/data/

These datasets have been pre-processed by us but have not been fully cleaned. Table 1 provides a brief description of variables included in this pre-processed dataset.

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	str	Name of station where trip originated
end_station_name	str	Name of station where trip terminated
bike_id	int	Unique ID attached to each bike
user_type	str	User membership type

Table 1: Description of bikeshare dataset columns

In the predictive analytics part of your assignment you should also draw on weather data to improve your prediction. For this purpose we have provided you with hourly weather data for the relevant cities and time periods. This data has been collected from the weather.com api. You can engineer features from this data as you see fit.

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 whether precipitation (snow or rainfall) was recorded in the respective period (1=yes,0=no)

Table 2: Description of weather dataset columns

Note that additional data, such as the locations of individual bike stations may be available from the operator websites. You can incorporate those in your analyses (e.g., for visualization purposes) for extra marks but it is not a requirement.

3 Description of tasks

1. **Customer Behavior Analytics**: As a fleet operator it is crucial to have a deep understanding of user preferences. A standard approach in customer analytics is clustering to identify user types and/or recurring behavioral patterns. It is your task to apply these methodologies to the case of bikesharing and identify archetypical trip types and infer their purpose. Proceed as follows:
   - Feature Engineering: A typical trip is described by three main dimensions: (1) day and time of departure, (2) location of departure/destination[1] (3) trip qualities such as duration, net distance[2] or net speed[3]. Create new trip-level features that capture these three dimensions and visualize their distributions. Which patterns do you observe?
   - Model Building/Evaluation: Select and apply a single unsupervised clustering algorithm to identify clusters of typical trip archetypes. In doing so, try to adhere closely to the CRISP data mining standard covered extensively in the lectures. At each point in the process clearly defend/argue your modeling choices with hard data/analyses where possible (e.g., why you have opted for the specific clustering algorithm, how you prepared the data, the number of clusters your used, etc.)
   - Managerial Implications and Interpretation of Results: Provide a comprehensive description of the trip archetypes you have identified by analyzing how they vary across the above-mentioned three core trip dimensions. Assign a meaningful label per each trip archetype that can be readily interpreted by a non-expert in a meaningful way.

2. **Predictive Analytics**[4]: Future demand is a key factor that will steer operational decision making of a shared rental network. As a data scientist it is your responsibility to facilitate this type of decision support. For the purpose of this assignment we are interested in forecasting **the expected number of trips in the next hour at the most popular rental station in the network**. Fleet operators can use this information to assess whether the current amount of vehicles available at this station will be sufficient to satisfy demand. To do so, develop a prediction model that predicts bike rental demand for your target station as a function of suitable features available in or derived from the datasets (incl. the weather data). Again, follow the core steps of the CRISP data mining process.
   - Feature Engineering: Develop a rich set of features that you expect to be correlated with your target. You can draw on your domain knowledge and/or conduct additional research around the topic of demand prediction in vehicle rental networks. Justify your selection of features.
   - Model Building: Select two regression algorithms that are suitable for the prediction task at hand. Explain and justify why you selected the three algorithms and describe their respective advantages and drawbacks.
   - Model Evaluation: How well do the models perform? Evaluate and benchmark your models' performance using suitable evaluation metrics. Which model would you select for deployment?
   - Outlook: How could the selected model be improved further? Explain some of the improvement levers that you might focus on in a follow-up project.

Notes and tips

- Make generous use of visualization techniques to clearly illustrate your findings and present them in an appealing fashion.
- Evaluate your methodology and clearly state why you have opted for a specific approach in your analysis.
- Relate your findings to the real world and interpret them for non-technical audiences (e.g. What do the coefficients in your regression model mean?, What does the achieved error mean for your model?, etc.)
- Make sure to clearly state the implications (i.e. the ”so what?”) of your findings for managers/decision makers.

[1] You are not required to perform in-depth geographical analyses here. Instead use simple metric such as distance from city center, or similar

[2] You are not required to compute actual road network-based distances. Instead use simpler approaches such as the commonly-used Haversine distance

[3] Net speed is an indicator of whether the trip was a roundtrip or a point-to-point trip

[4] Note that you can start this task independently from the clustering task above

4 Team allocation, deadlines and formats

The class has been divided into equally sized teams consisting of ca. 6 students each (see ILIAS for group composition). Please coordinate the work independently in your teams. To keep things interesting, different teams will focus on different datasets. Please find the allocation in Table 4. All data can be downloaded via the following link: `https://uni-koeln.sciebo.de/s/gL1lM6FjSqTYdBT`. Note that we are using data up to 2019 only, since we want to avoid COVID-related shocks that may significantly impact the performance/generalizability of prediction/clustering models.

Group	**Datasets (City, Year)**
Data Scientists	San Francisco (Bay Area), 2019
Rakete	San Francisco (Bay Area), 2018
Data Lake Divers	Boston, 2019
Group 4	Boston, 2018
The Erasmus Team	Boston, 2017
Data Vipers	Boston, 2016
Google Gurus	Chicago, 2019
Intim im Team	Chicago, 2018
DataCondas	Chicago, 2017
Powerpoint Rangers	Chicago, 2016
Mapa Ivdomaka	Philadelphia, 2019
Group 12	Philadelphia, 2018
Data Dons	Philadelphia, 2017
Deadly Python Assassination Squad	Philadelphia, 2016
Remainder Group	Boston, 2015

Table 3: Dataset allocation

As the main deliverable of this group project you are expected to submit the following documents:

- A 7-page report (excl. figures, references and appendices) in .pdf format detailing your answers to task 1-3 as well as any additional findings
- A single well-structured and clearly annotated Jupyter notebook (.ipynb format) with your code detailing your analysis and including executable Python code.
- A 1-page supplementary document (not counting toward the page limit) detailing the individual contributions of each team member (i.e. who did what).

Please make sure to submit these electronically via the upload link in ILIAS no later than **12:00h on 13th of July, 2020**. Your work will then be graded as per the guidelines set out in the course syllabus.