Introduction to Data Science and Machine Learning - Summer Semester 2023

Bachelor of Science WI / IS
EM Wirtschaftsinformatik II [1277BEWIF2]
Faculty of Management, Economics, and Social Sciences
Department of Information Systems for Sustainable Society
University of Cologne
Version - April 26, 2023

Instructor Vertr.-Prof. Dr. Friedrich Chasin **Term** SS 2023
TA Julius Kuhmann & Philipp Kai 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 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. **Network Understanding**: A deep understanding of the network of docking stations and how bikes are used to address daily transportation needs can direct marketing efforts and focus operational priorities of fleet operators.
2. **Idle Time Prediction**: An accurate prediction of how long bikes are parked at a docking station is an important step towards providing a high service level (e.g. by scheduling maintenance accordingly or by re-positioning vehicles from less to more utilized stations, etc.).

2 Description of Dataset

You have been allocated datasets of bikesharing rentals in five major US cities (see Section 4 for details on the 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. Note that these bikeshare operators exclusively offer a docked bikesharing services, meaning that customers can rent a bicycle from a docking station and can only return it at a docking station. This is different compared to what you might be used to from KVB Bikes or Call-a-Bike in Cologne, which let you rent and drop off a bike at any place within the service area. 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_lat	float	Latitude of start station
start_station_lon	float	Longitude of start station
end_station_lat	float	Latitude of end station
end_station_lon	float	Longitude of end station
bike_id	int	Unique ID attached to each bike

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**
city	string	Location where weather data was measured
timestamp	datetime	Day and time of measurement
temperature	float	Actual temperature recorded in degC
felt_temperature	float	Felt temperature recorded in degC
cloud_cover	int	Numerical measure of cloudiness
cloud_cover_description	string	Description of weather conditions
pressure	float	Atmospheric pressure recorded in hPa
windspeed	float	Wind speed recorded in km/h
precipitation	float	Precipitation recorded in mm

Table 2: Description of weather dataset columns

Next to the provided datasets, you are free to include further data sources in your analysis. Since we are interested in the network of docking stations, it might be particularly interesting to consider the location of relevant places in your allocated city. This information could be used to calculate the distance of docking stations to points of interest, such as the closest train station or the beach.

3 Description of Tasks

1. **Data Collection and Preparation**: You have been provided with a full dataset of bike sharing rentals. Select the city and year(s) you have been allocated and clean your dataset for use in later stages of your project. Briefly describe how you proceeded and how you dealt with possible missing/erroneous data. Here are some further steps you might want to follow:
   - Remember that we are interested in the idle time of bikes. This could be operationalized in two ways: first, the idle time of a particular bike (i.e. the time period from when bike with $bike\_id = x$ was dropped off at station y until bike with $bike\_id = x$ is booked again) or second, the idle time at a particular station (i.e. the time period from when bike with $bike\_id = x$ was dropped off at station y until any other bike at station y is booked). Consequently, you should compute a variable $idle\_time$ reflecting this information for each trip. You can select one of the two measures defined above and justify your choice.
   - Location & Weather Data: You may want to enrich your dataset with further information, including the provided weather data and the locations of points of interest to generate further relevant trip features.

2. **Descriptive Analytics**: As a fleet operator it is crucial to have access to close to real-time information on the operational performance of the vehicle fleet. As a data scientist your task is to facilitate this.
   - Overall System Performance: The marketing department would like to launch a social media campaign featuring the success of the city's bikesharing system. You are tasked to support by providing relevant insights.
   - Station-Level Insights: The operations manager has made it a priority to optimize the network of docking stations. You were asked to support him by delivering relevant insights on the state and performance of the station network.
   - Preparation for Predictive Task: Going on, we aim to predict $idle\_time$. Make sure to conduct a descriptive analysis to prepare for this next step.

3. **Predictive Analytics**: Idle time 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 idle time, either of a bike or of a particular station depending on your choice as discussed in Section 3.1**. Fleet operators can use this information to, for example, plan maintenance activities without losing out on bookings. You are tasked to develop a prediction model that predicts idle time as a function of suitable features available in or derived from the datasets (incl. the weather data, location data, etc.). Proceed as follows:
   - Feature Engineering: Develop a rich set of features that you expect to be correlated with your target. For example, you can use information on the time and weather at the time of drop-off, the location of stations, etc. Additionally, you could consider the distance of the drop-off location to certain points of interests, such as train stations or sights. 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 two 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?

4. **Discussion & Outlook**: Discuss the implications of your results for the fleet operator. Which further analysis would you consider useful and could be conducted on the given dataset?

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.

– **A note on the usage of ChatGPT**: ChatGPT is one among many tools that can provide useful assistance when writing code. This is a course on data science and machine learning, so we do not prohibit the usage of AI tools as such. However, we do expect you to be explicit about it. Therefore, it is mandatory to clearly indicate any code or text that was generated by tools such as ChatGPT (just like you would reference any other external sources). This includes providing the specific prompt that you used to generate the tool's output. In case you utilized AI tools, please include a short section in your report describing how you made use of them.

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 the Table below. All data can be downloaded via the following link: `https://uni-koeln.sciebo.de/s/ZOdCfXORDTeGnlF`.

Group	**Datasets (City, Year)**
Team 1	Los Angeles, 2017-2022
Team 69	Philadelphia, 2018-2020
ChatGPT 5.0	Philadelphia, 2021-2022
The Feature Engineers	San Francisco (Bay Area), 2018
The Learning Machines	San Francisco (Bay Area), 2019
TPJJAM	Boston, 2015-2016
Data Dream Team	Boston, 2017-2018
Team 8	Boston, 2019
Baby Cobras	Boston, 2021
Data Detectives	Boston, 2022
Team 11	Chicago, 2017
Team 12	Chicago, 2018
Team 13	Chicago, 2019

Table 3: Dataset allocation

As the 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. The report is the main deliverable of your project and should contain the following sections:
  1. *Cover page* with informative title, team number and member names
  2. *One-page executive summary:* summarizes the entire report for a non-technical manager (the business problem, data, the analytics solution, implications and recommendations)
  3. *Detailed report:*
     (a) Problem description (business goal and data science goal)
     (b) Data description
     (c) Brief data preparation details (how your data were created from the raw data) and key charts. Details can be provided in an Appendix.
     (d) Data analytics: Analytical methods applied (with sufficient detail and screenshots; use Appendix if needed) and appropriate performance evaluation (proper choice of measures, benchmarking).
     (e) Conclusions (advantages and limitations) and business recommendations

  As presentation is a key component of a successful data science project we will consider it in our evaluation. We therefore advise you to write your report in LaTeX. To facilitate the writing process, we provide a **LaTeX-template** on Overleaf. Using Overleaf's **documentation**, you can easily learn how to use LaTeX for scientific projects. Alternatively you may use Microsoft Word using similar typeset and line spacing.
– 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 **23:59h on 23rd of July, 2023**. Please refer to the course syllabus (Section G.4) for further information on how to submit your work.