Introduction to Data Science and Machine Learning - SS 2020

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 2020
TA Karsten Schroer **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. Nextbike, Europe's leading leading bikesharing platform is an excellent manifestations of a MaaS and MoD service provider. 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. **Real-time monitoring**: A deep understanding of the (real-time) operational performance of the fleet is core to inform business and operational decisions.
2. **Demand prediction**: Accurately predicting future demand is an important step towards providing a high service level (e.g. by deploying additional bikes or by re-positioning vehicles etc.)

2 Description of Dataset

You have been allocated datasets of bikesharing rentals in two German cities for a period of up to one year. This data was retrieved and collected via an open API of nextbike, Europe's largest bikesharing operator. More details on nextbike's API can be found here `https://api.nextbike.net/api/documentation#nextbike_api`.

The datasets you have received from us have been pre-processed but not fully cleaned. The main pre-processing work that has been performed is the extraction of trip data (incl. start and end positions)

from raw position data. Table 1 provides a brief description of variables included in this pre-processed dataset.

Variable name	Format	Description
day	datetime	Start day of rental
time	datetime	Start time of rental
b_number	int	Unique ID of vehicle
city	str	Name of city
trip_duration	timedelta	Duration of trip
orig_lat	float	Latitude of rental start point
orig_lng	float	Longitude of rental start point
dest_lat	float	Latitude of rental end point
dest_lng	float	Longitude of rental end point

Table 1: Description of variables

In the predictive analytics part of your assignment you should also draw on weather data to improve you prediction. Part of the work of a data science is to obtain relevant datasets independently. For this purpose we would like you to collect weather data independently. There are many resources but we would recommend using the open data portal of the German Weather Service (DWD), which can be accessed at `https://cdc.dwd.de/portal/201912031600/searchview`.

3 Description of tasks

1. **Data Collection and Preparation**: You have been provided with a full dataset of bike sharing rentals. Select the cities you have been allocated and clean your dataset for use in later stages of your project. To obtain weather data, access the open data portal of the German Weather Service (DWD) and retrieve hourly data for the relevant cities and periods.
2. **Real-time monitoring and 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 it to facilitate this. Proceed as follows:
   - Define up to five key performance indicators (KPIs) that provide indications of the current fleet operations and how well the fleet is doing in terms of utilization, revenue, coverage and other business-related aspects.
   - Briefly explain the rationale behind each KPI and why you have chosen it
   - Calculate hourly KPIs for the two cities in your dataset and visualize them over time. Which trends do you observe? How do you explain them?
   - Find explanations for any differences between cities. Which city performs better/worse and why?
3. **Predictive Analytics**: 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. To do so, develop a prediction model that predicts bike rental demand as a function of suitable features available in or derived from the datasets (incl. the weather data).
   - Select three 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 over other methods.
   - How well do the models perform? Evaluate and benchmark your models' performance using suitable evaluation metrics. Which model would you select for deployment?
   - 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 tipps

- 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.

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 cities. Please find the allocation in Table 4. All data can be downloaded via the following link: `https://uni-koeln.sciebo.de/s/Nv5tueVBupfAYgi`.

Group	**City 1**	**City 2**
ML Technologies in Sust. Envs.	Koeln	Essen
Floatastic	Bonn	Potsdam
Data Wizards	Leipzig	Frankfurt
Group[0]	Berlin	Bochum
Programming Pandas	Dortmund	Giessen
PoseidonMind	Karlsruhe	Kassel
Atad Ecneics	Duesseldorf	Heidelberg
Rhino	Bremen	Freiburg
Magma	Marburg	Duisburg
The Big Six	Mannheim	Kaiserslautern
GREEN	Nuernberg	Hannover

Table 2: Dataset allocation

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

- A 5-page report (excl. figures, references and appendices) in .pdf format detailing your answers to task 1-3 as well as any additional findings
- A private well-structured git repository including your coding work in the form of annotated Jupyter notebooks (.ipynb format) detailing your analysis and including executable Python code. [1]
- 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 ILIAS no later than **12:00h on 15th of July, 2020**. Your work will then be graded as per the guidelines set out in the course syllabus.

[1] To share a private github repository create a full-access dummy user for your repository and include the log-in credentials in your final report.