Advanced Analytics and Application – SS 2020

Master 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 Muhammed Demircan, Nastaran Naseri **Website** www.is3.uni-koeln.de and ILIAS

Team Assignment

This AAA team project is designed to test a representative cross-section of the data analytics and machine learning approaches we will 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

Dear students,

Nice to hear from you, and that you are taking up the challenge. As you have already noticed, the micro-mobility market is highly competitive. Many large companies and small start-ups are entering this industry. The latest trend is shared e-scooters that can be rented and returned at any point (shared free-floating e-scooters). The market for this seems huge. However, many established companies lack the know-how on how to integrate this trend into their existing range of products.

For the following team project, we put you in the position of a top-tier management consultancy. As the Data Science Team of McBoston & Company (MBC) you are experts in the field of Machine Learning and Data Science. **Background**: A renowned German car company is in the process of establishing a platform for mobility. On this platform the company wants to offer its customers several different transport options. The latest trend is so called *Shared E-Scooter* – such as Lime, Tier and Bird. However, the customer lacks tactical and strategic know-how in the area of shared e-scooters. *Therefore, our client asks for our help and would like to better understand the dynamics in the field of micro-mobility for several German cities in temporal and spatial resolution.*

However, we lack trip data for shared e-scooters, so we have to rely on shared bike data. **We assume that shared e-scooters are a substitute for shared bikes.** [**Side note**: If you are interested in these or similar questions around shared e-scooters, bikes, cars, we can investigate this in the context of a master thesis; we do have plenty of data for shared vehicles for many cities in Europe and USA].

2 Description of Dataset

You have been provided with a dataset of bike sharing rentals in a range of German cities for the period from Jan 1st, 2019 to Dec 31st, 2019. This data was made available by the bike sharing platform operator NextBike. Unfortunately, we were not able to get trip data but availability data. In other words, we only have data about the locations of vehicles while they were not rented. You can also imagine this as if you opened the app several times (many, many times) and wrote down the positions of the available vehicles. To make matters worse, a meaningful description of the dataset was not provided by the customer. However, the following documentation is available: https://api.nextbike.net/api/documentation#nextbike_api.

For your assignment you should also draw on weather data to improve your 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 **here**. Hourly weather observations for Germany can be easily downloaded via their FTP server.

Also, you should consider to harness land-use data for your project. Many researchers have shown that land use is a highly influential factor effecting customers of shared mobility. A possible source for land-use data might be https://land.copernicus.eu/pan-european/corine-land-cover. However, this is **not!** mandatory; you should decide based on your skills, resources and time plan whether to consider land-use data.

3 Description of Tasks

1. **Data Collection and Preparation**: You have access to the bike sharing location data. Select the city and year that have been assigned to you 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). We need trip data to analyze the movement behavior of customers in German cities. Thus, compute trips based on the availability data. Also, provide a detailed description of both datasets (i.e., availability data and trip data) such that there are no pending questions. After submitting the project, it is important for the client to understand how you computed trips based on available data. To better analyze location data in subsequent steps, discretize your city with the help of suitable tools as a matrix of hexagons (such as H3-Uber). This discretization is crucial for the analysis of the spatial resolution.
2. **Descriptive (Spatial) Analytics**: Analyze both the availability of bikes and the bike rental demand patterns for the relevant one-year period and city (please check carefully which city your team has been allocated). Specifically show how these patterns (such as availability of vehicles, start time, trip length, start and end location, average "IDLE" time between trips) for the given sample varies in spatio-temporal resolution. Give possible reasons for the observed patterns. An example of how the spatial resolution can be varied, is the examination of hexagons with different edge lengths.
3. **Cluster Analysis**: Based on the bike rental demand patterns, can you identify clusters of trip types and/or customer types? How would you label these clusters? Can you identify hot spots for vehicle availability and demand in spatial resolution?
   Methods: Identify clusters with soft-clustering and visualize your results. Compare your results to a hard-clustering method of your choice.
4. **Predictive Analytics with Support Vector Machines**: Develop two prediction models that predict a) bike rental demand and b) bike availability in spatio-temporal resolution. In other words, your method should predict for each hexagon and time-basket (e.g., 08am-11.59am) the bike demand and number of available vehicles.
   - Simply start without a kernel. Then, gradually make your model complex by integrating different kind of kernels. Also, use grid search to find optimal values for your hyperparameters.
   - How good is your model? Evaluate your model's performance and comment on its shortfalls.
   - Show how you model's performance varies as you increase or decrease temporal resolution for the following period length:1h, 2h, 6h, 24h. Also, vary the length of the hexagon edges.
   - How could the model be improved further? Explain some of the improvement levers that you might focus on in a follow-up project.
5. **Predictive Analytics with Deep Learning:** Repeat the steps from subtask 4, but this time use a feedforward neural network.
   - Is the performance very different from the previous approach?
   - With this realization, do you think it is worth to employ a deep-learning approach?
6. **Discussion & Outlook**: Discuss the implications of your results for the potential fleet operator (client). Which further analysis would you consider useful and could be conducted on the given dataset? Based on this analysis, could you draw the boundaries of an operating area for the fleet? Which other external data sources might be interesting to consider? Do you think that the client could profitably realize such a business model?

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.

- Do not forget that your goal is to convince the customer of your results.

4 Team allocation, deadlines and formats

The class has been divided into equally sized teams consisting of 5 students each. Please coordinate the work independently in your teams. The data can be downloaded here (PW: **AAA_SS_20**):

https://uni-koeln.sciebo.de/s/aPDzAW9PxPsZTBO

To keep things interesting, different teams will focus on different cities. Please find the allocation in Table 2:

	Team Name	City
1	Jupyter Explorers	Dresden
2	Analytics Anacondas	Bonn
3	Significantly Different	Leipzig
4	LupoAnalytica	Cologne
5	Quaranteam	Berlin

Table 2: Group allocation

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

- A 10-page report (excl. figures, references and appendices) in .pdf format detailing your answers to task 1-6 as well as any additional findings
- An annotated Jupyter notebook (.ipynb format) detailing your analysis and including executable Python code. For the sake of readability, you can split the Jupyter notebook into multiple Jupyter notebooks (e.g., 01_prep.ipynb, 02_descriptive, …)

Please make sure to submit these electronically via ILIAS no later than **23:59h on Jul 31st**. Your work will then be graded as per the guidelines set out in the course syllabus.