Advanced Analytics and Application – SS 2024

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
TA Ramin Ahadi

Term SS 2024
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 smart mobility market is highly competitive. Many large companies and small start-ups such as Uber, Lyft, FreeNow, Lime, Tier, - you name the rest - are active in this industry. Current reports and projects indicate that particularly ride-hailing (in combination with autonomous and electric vehicles) will dominate the market in the near future. The market for this business seems to be huge. Car manufacturers realized that they need to offer comparable mobility services to their customers, otherwise they might be disrupted by tech titans from Silicon Valley (or soon perhaps from China).

For the following team project, we put you in the position of a top-tier management consultancy. As the Data Science Team of Bane & Wayne Partners (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 ride-hailing mobility services. Initially, the client wants to offer its US customers (as customers from the US market are more receptive for novel business models) a ride-hailing platform using a fully electrified vehicle fleet. However, the customer lacks tactical and strategic know-how in the area of operations management of electrical vehicle fleets. Therefore, our client asks for our help and would like to better understand the dynamics in the field of ride-hailing for a US city in temporal and spatial resolution.

2 Description of Dataset

You have been provided with a historical dataset of taxi data – which can be assumed to be a good approximation of ride-hailing data - from Chicago, USA. This data was made available by the Chigaco Data Portal and contains taxi trips from 2013 onwards. Detailed documentation related to the taxi trip data can be found: https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew#column-menu.

For your assignment you should also draw on weather data. 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 by your own. There are many resources but we would recommend using open data. You might have a look at the official open data government website (here) or on Kaggle (here), or any other source you find valuable for your research. Bear in mind, the temporal resolution of your chosen weather data might impact the predictive

power of your models. We highly encourage you – based on our experience - to harness hourly weather data (i.e., Kaggle).

Also, you should consider to harness Point-Of-Interest data for your project. Many researchers have shown that POI is a highly influential factor effecting customers of shared urban mobility. A possible source for POI data is OpenStreetMap. However, this is **not!** mandatory; you should decide based on your skills, resources and time plan whether to consider POI data.

3 Description of Tasks

1. **Data Collection and Preparation**: You have access to the taxi trip data. Select the year(s) that have been assigned to you and clean your dataset for use in later stages of your project. As the CSV file might be too large for your computer to open it with Python, you could preprocess the file first with tools like "sed" or "xsv" in order to filter out not needed rows and columns. To obtain hourly weather data, access the links provided above (or reach out to us). Also, provide a detailed description of the trip dataset such that there are no pending questions. Due to privacy reasons, spatial data is given only on census tract level. To better analyze location data – specifically in lower dimensions - further discretize the city in scope with the help of suitable tools (such as a matrix of hexagons using h3-Uber). This discretization is crucial for the analysis of the spatial resolution. Furthermore, you should also consider different temporal discretization (e.g., hourly, 4-hourly, daily) etc.
2. **Descriptive (Spatial) Analytics**: Analyze taxi demand patterns for the relevant one-year period and city (please check carefully which year your team has been allocated). Specifically show how these patterns (start time, trip length, start and end location, price, average idle time between trips, and so on) for the given sample varies in different spatio-temporal resolution (i.e., census tract vs. varying hexagon diameter and/or temporal bin sizes). Give possible reasons for the observed patterns.
3. **Cluster Analysis**: Getting a deep understanding of how customers use your mobility service is crucial. As an example, for marketing purposes, or managerial decisions. One widely adopted approach to understand customers' behavior is to find prevalent usage/trip clusters. **Tasks:** Based on the taxi trip patterns, can you identify clusters of trip types and/or customer types? How would you label these clusters? ***Methods***: Identify clusters with soft-clustering and visualize your results. Compare your results to a hard-clustering method of your choice. You can use additional features like "distance to city center", expressive hourly resolutions (e.g., "bar hours", "morning commuting"), or even land-use/POI data.

   Furthermore, can you identify spatial hot spots for trip demand using Gaussian Mixture Models (i.e., using Spatial Kernel Density Estimation)?

4. **Predictive Analytics with a) Support Vector Machines and b) Neural Networks**: Develop two prediction models that predict taxi trip demand using a) support vector machines and b) neural networks (deep learning) in spatio-temporal resolution (i.e., spatial-unit and time buckets). In other words, your method should predict for each spatial unit (hexagon and census tract) and time-basket (e.g., 08am-11.59am) the taxi demand. Also adevise a reasonable validation strategy for your prediction model (i.e., definition of test, training data etc).

   Approach for SVM:
   - 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 does your performance change when you only use census tract as spatial units?
   - How could the model be improved further? Explain some of the improvement levers that you might focus on in a follow-up project.

   Approach for Deep Learning:
   - Repeat the steps from subtask 4a), 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?

5. **Smart Charging Using Reinforcement Learning:**
   Consider an electric taxi driver who can charge her vehicle at home. To simplify the problem, we assume that the vehicle always arrives at home at 2 p.m. and leaves the garage at 4 p.m. each day. We want to design an intelligent charging system (an automated agent). Therefore, instead of a flat charging rate, the charging agent adjusts the charging power every 15 minutes, which is bounded between 0 kW and the highest rate (e.g., 22 kW). Also, the vehicle's battery has a capacity that cannot be exceeded. After leaving the garage, the taxi needs enough energy to complete its working day. The energy demand is a stochastic value following a normal distribution (you should choose the parameters, e.g., μ= 30 kWh, σ = 5 kWh) and must be generated exactly when the driver wants to leave. The agent's goal is to avoid running out of energy (you should consider a very high penalty for running out of energy) and to minimize the recharging cost. The recharging cost follows an exponential function of the power (i.e., charging cost $(\mathrm{t}, \mathrm{p}) = \sum_{t \in T} \alpha_t e^{p}$), where α_t is the time coefficient and p is the charging rate.

   The task is to create the environment (a very simple discrete event simulation) that receives the agent's decisions and returns the reward. In addition, you must define a Markov decision process, including states, actions, and reward function, and solve it using a reinforcement learning algorithm (e.g., deep q-network) to find optimal charging policies. To allow the use of discrete action methods, you can consider only limited charging options such as zero, low, medium, high.

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? Which other external data sources might be interesting to consider? Is it sensible to install private charging stations, or should the operator draw on public charging stations?

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:

https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew#column-menu

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

	Team Name	Year	Alternative Year
1	A^3nalysts	2013	2023
2	AAAllstars	2014	2022
3	The Algorithm Avengers	2015	2021
4	Duracell AAA 1,5v	2016	2020
5	Algorithm Amigos Anonymous	2017	2019
6	AIbertus Magnus	2017	2018
7	The Copy-Pasters	2016	2023
8	DataCrafters	2015	2022

Table 2: Group allocation

Note: Since historical weather data for the years between 2018 and the present is not easily available, we consider the years 2013-2017 as the main case, but you can work with alternative options (see Table 2). If you use alternative years and prepare the weather data, you can get bonus points (7 out of 100), which could be very helpful for your final score. All teams could earn bonus points by providing weather data through website scraping and API requests. Here are some websites where you can access weather data:

https://www.ncei.noaa.gov/support/access-data-service-api-user-documentation

https://www.wunderground.com/history/daily/gb/doncaster/EGCN/date/2022-5-3

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, …). If you do so, please provide instructions on the order of executing notebooks and make sure that your code is runnable! If you employ third-party libraries, please also include an environment specification file (requirements.txt or environment.yml) in your submission.

Please make sure to submit these electronically via ILIAS no later than **11:59h (noon) on Aug 15th**. Your work will then be graded as per the guidelines set out in the course syllabus.