Analytics and Application - WS2020-21

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 Nastaran Naseri
Term WS 2020-21
Website www.is3.uni-koeln.de and ILIAS

Team Assignment

This AA 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

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. Car2Go and DriveNow, the world's leading free-floating carsharing platform operators are excellent manifestations of MaaS and MoD service providers. Similar platforms are also getting traction across other transport modes such as bikes and e-bikes (e.g., Nextbike) 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 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 good by doing well"). Accurately predicting rental demand becomes an important step towards providing a high service level (e.g., by providing additional assets, etc.). This project will act as a first introduction to this topic.

2 Description of Dataset

You have been allocated datasets of bike sharing rentals in two German cities for a period of up to 5 months. This data was retrieved and collected via an open API of nextbike, Europe's largest bike sharing operator. More details on nextbike's API can be found **here**. 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 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.

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).
2. **Descriptive Analytics**: Analyze the bike rental demand patterns for the given period and cities (please check carefully which cities your team has been allocated).
   - Define up to five key performance indicators (KPIs) that provide indications of the current 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?
   - Show how rental patterns (such as start time, trip length, start and end location) for the given sample varies on a seasonal, weekly and daily level. Give possible reasons for the observed patterns.
   - Find explanations for any differences between cities. Which city performs better/worse and why?
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 cluster the locations based on their demand pattern?
4. **Predictive Analytics**: Develop a prediction model that predicts bike rental demand as a function of suitable features available in or derived from the datasets.
   - Select three prediction models, at least one regression algorithm, that are suitable for the prediction task at hand.
   - Why did you choose a specific regression type? Clearly justify your choice of methods and describe its advantages over other methods.
   - How good is your model? Evaluate your model's performance and comment on its shortfalls.
   - 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 model be improved further? Explain some of the improvement levers that you might focus on in a follow-up project.
5. **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 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 5 students each. Please coordinate the work independently in your teams. To keep things interesting, different teams will focus on different cities. All data can be downloaded from **here**. Please find the allocation in Table 2:

Group number	**City 1**	**City 2**
Team 1	Bonn	Essen
Team 2	Koeln	Potsdam
Team 3	Berlin	Frankfurt
Team 4	Leipzig	Bochum
Team 5	Karlsruhe	Giessen
Team 6	Dortmund	Kassel
Team 7	Duisburg	Freiburg
Team 8	Marburg	Heidelberg
Team 9	Bremen	Duesseldorf
Team 10	Mannheim	Kaiserslautern

Table 2: Group 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 1-5 as well as any additional findings
- An annotated Jupyter notebook (.ipynb format) 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 ILIAS no later than **12:00h on Feb 10th**. Your work will then be graded as per the guidelines set out in the course syllabus.