Introduction to Data Science and Machine Learning - SS 2021

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 2021
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. Bikesharing platforms are an excellent manifestations of a 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. **System 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 four major US cities for a period of one year each. This data was collected via the open trip history data of Blue Bikes Boston, Divvy Bikes Chicago, Ride Indego Philadelphia and Bikeshare Metro in Los Angeles. More details on the datasets can be found on their respective websites:

- 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_name	str	Name of station where trip originated
end_station_name	str	Name of station where trip terminated
bike_id	int	Unique ID attached to each bike
user_type	str	User membership type

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
date_time	datetime	Day and time of measurement
max_temp	float	Maximum temperature recorded in degC
min_temp	float	Minimum temperature recorded in degC
precip	int	Binary indicator for whether precipitation (snow or rainfall) was recorded in the respective period (1=yes,0=no)

Table 2: Description of weather dataset columns

Note that additional data, such as the locations of individual bike stations may be available from the operator websites. You can incorporate those in your analyses (e.g., for visualization purposes) for extra marks but it is not a requirement.

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. Briefly describe how you proceeded and how you dealt with possible missing/erroneous data.
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 it to facilitate this. Proceed as follows:
   - Temporal Demand Patterns and Seasonality: Demonstrate how fleet usage varies during a day, a week and the year. What patterns do you observe? Explain.
   - Geographical Demand Patterns: Which stations are particularly popular and which are not? Provide a rationale as to why you observe these patterns.
   - Key Performance Indicators (KPIs): Define at least (!) three KPIs that you would include in a dashboard for a fleet operators. These KPIs must provide an immediate overview of the current fleet operations and how well the fleet is doing in terms of utilization, revenue, coverage and/or other business-related aspects. Briefly explain the rationale behind selecting each KPI, explain why you have chosen it and where needed provide references. Calculate hourly values for the selected KPIs for the city/year in your dataset and visualize them over time. Which trends do you observe? How do you explain them?
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. For the purpose of this assignment we will be interested in forecasting **total system-level demand in the next hour**. 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).

  - Feature Engineering: Develop a rich set of features that you expect to be correlated with your target. In this process you can draw on your domain knowledge and/or conduct additional research around the topic of demand prediction in vehicle rental networks. Justify your selection of features.
  - Model Building: 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 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?
  - Outlook: 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 for managers/decision makers.

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 Table 4. All data can be downloaded via the following link: `https://uni-koeln.sciebo.de/s/gL1lM6FjSqTYdBT`.

Group	**Datasets (City, Year)**
[DATA]Miners	Boston, 2019
Team Viper	Boston, 2018
Error 404: Group not found	Boston, 2017
MMA_TEE	Boston, 2016
colasechs	Boston, 2015
NeuerOrdner	Chicago, 2019
Standard-Normalverpeilt	Chicago, 2018
The Predictors	Chicago, 2017
Prescriptive Struggle	Chicago, 2016
6pack	Chicago, 2015
Hacket	Philadelphia, 2019
Team12	Philadelphia, 2018
Data Scientology	Philadelphia, 2017
Plotttwist	Philadelphia, 2016
Silicon Valley	Los Angeles, 2019
Information in Formation	Los Angeles, 2018

Table 3: 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 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]

[1] To share the github repository please include the link in your report. Also make sure to set the visibility of your repository to private while you are working on your code and only publish it shortly prior to your submission.

- 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 **12:00h on 21st of July, 2020**. Your work will then be graded as per the guidelines set out in the course syllabus.