[AA] Team Assignment 2024/25

Analytics and Applications M.Sc. Course
Faculty of Management, Economics, and Social Sciences
Department of Information Systems for Sustainable Society
University of Cologne
October 30, 2024

Instructor Prof. Dr. Wolfgang Ketter **Term** WS 2025/25
TA Janik Muires **Website** `www.is3.uni-koeln.de` and ILIAS

This team assignment 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 electric vehicles (EV) and will act as an illustration of how we can use data in impactful ways to address societal issues and business at the same time.

Please download all relevant data from Sciebo: `https://uni-koeln.sciebo.de/s/59LhTdJ9c8tYgmn`

1 Introduction

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 engine (ICE) vehicles. A switch to electric vehicles (EV) is thus a key ingredient in reducing personal mobility emissions. However, the charging infrastructure and underlying power grid need to provide appropriate capacity. With EV, a lot of users are experiencing what is widely called "range anxiety", the fear of not being able to complete a trip due to uncertain battery capacity and the fact that EV cannot be recharged instantaneously, like ICE cars can. While this is largely psychologically driven and has a lot to do with experience and will fade over time as we get to know EV better (and EV technology will allow for longer trips and faster recharge times), it adds to the strain on the power grid, as users would rather plug their EV in when it is parked than not, just to "top it up". What's more, when a lot of users want to charge in the same spot (e.g. in a shopping center), they are capacity constrained, as usually not all charging stations in one hub can operate at full capacity at the same time. Thus, for operators of such charging hubs, it is of great importance to monitor, understand, plan, and potentially even steer the charging sessions of connected EV. In this project, you will take the role of a team of data scientists that work for the operator of charging hubs. You will carry out all steps of the cross industry standard process for data mining (CRISP-DM) and focus on two core aspects:

1. **Understanding System Metrics:** Gaining a deep understanding of key metrics of the charging hubs' operation and presenting relevant metrics in a refurbished way for the operator management to understand.
2. **Prediction of Utilization:** Accurately predicting future utilization is an important tool to plan the operation and facilitate new business models.

2 Description of Dataset

You have been provided with data on individual EV charging sessions at two charging sites that belong to the same operator. One site is public (at a university), the other is private (open to employees of a company). Both have approximately 50 EV charging stations that are part of the monitored charging network. The datasets have been preprocessed and are provided in CSV format to take some data wrangling off of your shoulders but have not been checked for erroneous or missing data. Additionally, we provide you with hourly historical weather from a nearby airport weather station. You are free to use this data and additional data sources you might want to include for your analysis. Some charging sessions were carried out by registered users, some by unregistered users. Registered users are able to provide requests to the charging management system, which the system can use to facilitate adaptive charging, but requests are not guaranteed to be fulfilled.

field	type	description
id	string	Unique identifier of the session record
connectionTime	datetime[a]	Time when the EV plugged in.
disconnectTime	datetime[a]	Time when the EV unplugged.
doneChargingTime	datetime[a]	Time when of the last non-zero current draw recorded.
kWhDelivered	float	Amount of energy delivered during the session.
sessionID	string	Unique identifier for the session.
siteID	string	Unique identifier for the site.
spaceID	string	Unique identifier of the parking space.
stationID	string	Unique identifier of the EVSE.
timezone	string	Timezone of the site. Based on pytz format.
userID	string	Unique identifier of the user, if provided.
userInputs	list	Inputs provided by the user. Since inputs can be changed over time, there can be multiple user input objects in the list.
WhPerMile[b]	float	Efficiency of the EV in Wh per mile.
kWhRequested[b]	float	Energy requested by the user in kWh.
milesRequested[b]	float	Number of miles requested by the user.
minutesAvailable[b]	float	Length of the session as estimated by the user.
modifiedAt[b]	datetime[a]	Time this user input was provided.
paymentRequired[b]	bool	If the user was required to pay for this session.
requestedDeparture[b]	datetime[a]	User estimated departure time.

[a] All datetimes are in UTC (GMT) see timezone field for the correct timezone of the site.
[b] Fields are optional, if user participates in app-based charging.

Table 1. Description of Fields of Charging Session Dataset

3 Task Description

1. **Data Collection and Preparation:** You have been provided with a full dataset of charging sessions in two sites for an extended timeframe. The first step should start by ingesting the data in appropriate format, checking for missing or erroneous data, and cleaning 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:** The operator of the two sites is interested in the operational performance and statistics of their charging hubs. As the company's data scientist, your task is to facilitate this. Proceed as follows:
   (a) Temporal Patterns and Seasonality: Demonstrate how the number of charging events varies during the day, the week, and between seasons. What patterns can you observe, and how do you explain them?
   (b) Key Performance Indicators (KPIs): Define three time-dependent KPIs that you would include in a dashboard for the hub operator. These KPIs must provide an immediate overview of the current hub operation and how it is doing in terms of utilization 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 and visualize them over time. Which trends do you observe? How do you explain them?
   (c) Site Characteristics: The hub operator provided you with the data set, but has seemed to forget which site was supplying which data... Can you find out which of the two sites is the public one? Try to combine data understanding from previous descriptive analytics with domain knowledge (business understanding) of how private vs. public charging hubs might differ in operation. Explain your line of thought!
3. **Cluster Analysis:** To better understand what typical charging sessions look like, carry out a cluster analysis to provide management with a succinct report of archetypical charging events. Think of an appropriate trade-off between explainability and information content and try to come up with names for these clusters. What is the value of identifying different types of charging sessions?
4. **Utilization Prediction:** The operator has tasked you with developing a model for predicting hourly utilization of the two sites. Remember to use appropriate evaluation and validation techniques to measure and ensure performance of the predictive model!
   (a) Your boss understands a little bit about statistical models, but they are not sure about these fancy new neural networks that you told them could be used for predictive tasks. **Develop two**

predictive models, one using neural networks and the other one explicitly not using neural networks, but another technique of your choice. Use cross-validation to train the models and compare the predictive performance of both models on the same holdout set. What type of model should the operator employ? Give a suggestion, keeping in mind the trade-off between explainability and performance.

(b) Now that the provider is able to (roughly) predict utilization of its charging sites, **think of a business case** you could employ the prediction for. Give an example using data and/or predictions, and make it clear how the predictive model enables the business case! Your boss is rather cautious when it comes to their money. Sketch up limitations and think about potentially dangerous fallacies. In other words: Pitch the idea, keeping in mind all eventualities!

4 Deliverables

Please adhere to Section 2 in the Syllabus, as all deliverables are detailed there. In addition, we highly recommend you to facilitate version-control systems such as git and use platforms like GitHub. This makes it easy for us to check milestone progress and also receive your final code submission. (The final report submission is still facilitated through ILIAS.) If you choose to do so, please either include a public link to your repository in your report document or invite Janik Muires (muires@wiso.uni-koeln.de) to your repository. **Remember: Failure to provide milestone progress equals failure of the team assignment portion of the portfolio exam, which in turn equals failure of this course!**