42577 Introduction to Business Analytics course

Project statement

Welcome to this year's challenge!

The topic this year is *energy (and mobility)*. At a time when the world is facing unprecedented challenges of different kinds, including climate change, pandemics, social inequality, and degrading biodiversity, electric vehicles (EVs) present themselves as essential for reducing greenhouse gas emissions and combating climate change, as they produce zero tailpipe emissions and can be powered by renewable energy sources. Additionally, EVs offer long-term economic benefits by reducing reliance on fossil fuels, lowering fuel costs, and improving air quality, contributing to healthier communities.

In this project, we invite you to step in the shoes of an EV charging provider and use your best Business Analytics skills to help the company manage its operations and plan its future developments, thus providing a better service to its users and enhancing its business model. We want you to address the mandatory questions (below) but should also seek new questions, new data, and new insights.

You have access to a dataset that provides comprehensive insights into the usage patterns of electric vehicle (EV) charging stations in the city of Palo Alto, California. This dataset offers valuable information about the availability, utilization, and demand for EV charging infrastructure, allowing analysts to gain a deeper understanding of the local EV market and sustainable transportation practices. The data includes details on charging station locations, types of charging connectors available, charging session durations, energy consumption, and other relevant metrics. This information enables users to analyze trends, identify peak usage times, assess charging infrastructure performance, and develop data-driven solutions to enhance the efficiency and accessibility of EV charging services.

Project

The project has three components:

- *<u>Mandatory component</u>*: All groups need to address the same problem(s).
- *<u>Exploratory component</u>*: Each group is invited to choose their own research questions and explore the data accordingly.
- *<u>Report:</u>* Each group should deliver one jupyter-notebook, which should be self-explanatory in each step (or block). This will function as a report, so it should have introduction and conclusions, besides comments and reflections. However, there are some rules about the structure of the report, which should follow the 4-part outline shown below:

Section 1: Introduction + Data analysis and visualization

Section 2: Mandatory Component

Section 3:Exploratory Component

Section 4:Conclusions

At the end of this document you will find a list of practical information, which will include details on what is expected in each task, and how these aspects contribute to the final grade.

- **Introduction to the data**

Each row in the dataset corresponds to an EV charging event by a user. Therefore it will have a user ID associated, a charging station (location), the particular plug used (a charging station can have multiple plugs), a start time when the EV is plugged in, an end time when the charging stops and the actual time when the EV is unplugged (these two can be very different due to "hoarders"), duration of the charging, energy consumption, etc. The full list of variables that you will have in this dataset is the following: Station Name, MAC Address, Org Name, Start Date, Start Time Zone, End Date, End Time Zone, Transaction Date (Pacific Time), Total Duration (hh:mm:ss), Charging Time (hh:mm:ss), Energy (kWh), GHG Savings (kg), Gasoline Savings (gallons), Port Type, Port Number, Plug Type, EVSE ID, Address 1, City, State/Province, Postal Code, Country, Latitude, Longitude, Currency, Fee, Ended By, Plug In Event Id, Driver Postal Code, User ID, County, System S/N, Model Number. The names of the variables should be more or less self-explanatory but don't hesitate to ask for clarification if you have any doubt about them.

The data is provided as a CSV file. Notice that the variables require some treatment in order to be usable (e.g. Dates, categorical, strings, different scales, IDs).

The *mandatory component* consists of two tasks:

1. Cluster the users based on their usage patterns - users with similar usage behaviors (e.g., charging more on the weekends, or more in daytime vs. nighttime) should be grouped together. The goal is to segment the user base, such that you can try to identify a target audiency for the company's upcoming ad campaigns. (Important: note that the dataset provided contains charging events - you will have to transform that data into hourly energy use and hourly charger occupation for this analysis and also for Task 2). In the end, you should be able to provide recommendations to the company – you will not be evaluated on that recommendation, but on the appropriate usage of the techniques that you learned in class to achieve that recommendation. Make sure to clearly motivate your choices along the way!

2. Build a prediction model that, at the end of a day, allows one to predict what the energy consumption for each charging station will be over the next 24 hours with a 1-hour resolution – i.e. not the total energy consumption for the next day, but how the time-series of the consumption (with 1h resolution) is expected to look like for the next day (e.g., given demand data until midnight of day 1, predict the consumption for all 1h intervals (6-7am, 7-8am, …, 11-12pm) in day 2). This information is crucial for your company to manage the energy grid properly. You can choose to use a single model that predicts consumption for all stations at once or multiple independent models. It is up to you to decide how to best formulate this problem as a machine learning problem. Whatever modeling approach you choose, your forecasts should have a lower error than a naïve baseline model that makes a prediction based on the

historical average – e.g., use the time series of how an "average Monday" looks like (according to historical data in the training set) as the prediction for the next Monday in the testset. You are expected to implement this baseline approach and compare your prediction model to it. You should **not shuffle the data**. You should instead use the first 70% of the data to train your model, and the remaining data as a test set.

The ***exploratory component*** focuses on a strategic planning component. You are expected to leverage the data and modeling techniques covered in the course to provide recommendations to the company that help it improve on a more long-term horizon. Each group needs to address at least one new research question. We expect that you use at least one modeling technique (e.g. dimensionality reduction, clustering, classification, time series) in this component (i.e., you should not just use data handling and visualization to answer all your research questions). We expect you to formulate your own question, and follow the data sciences cycle. Feel free to make assumptions, but state them clearly. Some example topics that you may focus on are:

- Infrastructure planning. For example, where to build new charging stations or where to increase the capacity of the existing ones? What is the impact of land use on the charging demand (e.g., proximity to bus/metro station, shops, residential area vs. business district.)?
- Campaigns to increase revenue or improve resource utilization. This could be, for example, by providing discounts for charging at particular hours or at particular stations or by discouraging certain behaviors such as "hoarding" (how much potential revenue is the company losing due to hoarding?).

Evaluation

The evaluation of the report will be based on the following criteria:

- Clarity - self-explanatory nature of the notebooks
- Thoroughness - Each research question deserves to be explored to the right amount of depth
- Insightfulness - It's important to go beyond the surface of the conclusions
- Technical aspects:
  - Data have been properly analyzed (data cleaning data preparation, data pre-processing).
  - Which model has been used (only one model, multiple models, only linear models, or non-linear models)?
  - Is the model and approach appropriate?
  - Which performance metrics were used (how performances were evaluated)? Were they appropriate?
  - How was the approach benchmarked (how conclusions were drawn)?
- Honesty - While it's fine to use others' code (as a starting point), these shouldn't generally be the actual deliverable **and** the appropriate ethical practice is to **always** reference the source of that code you used.

Rules

- Each group should consist of 4 students.
- The submission of the project shall be a zip file with all the notebooks. This zip file should contain the surnames of the group members (for example, for Pablo, Anders, Suarez, and Mila, it should be Pablo_Anders_Suarez_Mila.zip).
- At the end of the report, there must be a section where **individual contributions are clearly clarified**. In case of doubts on individual contributions or authenticity of the report, the teachers will call the group for an oral defence. This section should **not be part of the page counts.**
- Meeting the deadlines for the milestones is important, including for non-evaluated milestones. A penalty of 10% is given for each extra day of delay

PLEASE INDICATE NAME, SURNAME, and STUDENT NUMBER IN THE REPORT

Report length

The report must be in the form of a jupyter notebook. The structure should be the one described on page 1. The overall lengt of the report should be 2500 words (see more details below). As a rule of thumb, the project (description of the research questions and results) should not exceed 4 pages. This limit does not apply to figures and codes.

To be more precise, the report can include unlimited figures, and there is no limit to the length of the code. The 4 pages limit only applies to markdown cells. As a reference, you can should use this code[1] to make the word count of your markdown cells. One document page is about 500 words (3000 characters including space). The project should be approximately 2000-2500 words. Again, this applies only to markdown. **Excessively long reports will be penalized**. A penalty of 10% is given for each extra 500 words, and anyway the report should under no circumnstance exceed the 3500 words.

- October 7 – Announcement of this challenge statement
- October 23 – Communication of group members (through DTU learn)
- Project Milestone – November 11

December 9 – Final submission – all materials, including report notebook. Submit through DTU learn.

[1] https://stackoverflow.com/questions/71194571/word-count-of-markdown-cells-in-jupyter-notebook