42577 Introduction to Business Analytics course

Challenge statement

Welcome to this year's challenge!

The topic this year is *mobility*. At a time when the world is facing unprecedented challenges of different kinds, including climate change, pandemics, social inequality, and degrading biodiversity, sharing mobility services offer and emission-free mobility option that is both efficient and attractive. In this project, we invite you to use your best Data Sciences skills to help operators to manage their fleet, providing better a service and enhancing their business model. We do not expect you to discover revolutionary business models with a single Data Science project, instead, we want you to address the mandatory questions (below) but also seek yourself for new questions, new data, new insights.

You have access to data from Citi Bike (New York)[1], one of the biggest station-based bike-sharing systems in the United States. The dataset includes more than 900 stations and 14000 bikes, and it contains over 17 million bike rides observed during 2018 (yes, it's huge). This dataset has the general objective of helping City Bike operating at its best and of making bike sharing more attractive. You can read more about it here[2]. The data is in itself an interesting Data Sciences exploration.

Project

The project has three components:

- *Prediction challenge*: All groups need to address the same problem.
- *Exploratory component*: Each group is invited to choose their own research questions and explore the data accordingly.
- *Report:* 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: Prediction Challenge

  Section 3: Exploratory Component

  Section 4: Conclusions

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

[1] https://s3.amazonaws.com/tripdata

[2] https://ride.citibikenyc.com/system-data

- **Introduction to the data**

  Figure 1 shows the variables you will have in this dataset. The data is provided as a CSV file. Notice that the variables require a lot of treatment in order to be usable (e.g. Dates, categorical, strings, different scales, IDs).

	0	1	2	3	4	5	6
Unnamed: 0	0	1	2	3	4	5	6
tripduration	970	723	496	306	306	1602	722
starttime	2018-01-01 13:50:57.4340	2018-01-01 15:33:30.1820	2018-01-01 15:39:18.3370	2018-01-01 15:40:13.3720	2018-01-01 18:14:51.5680	2018-01-01 21:31:54.1920	2018-01-02 07:54:53.6460
stoptime	2018-01-01 14:07:08.1860	2018-01-01 15:45:33.3410	2018-01-01 15:47:35.1720	2018-01-01 15:45:20.1910	2018-01-01 18:19:57.6420	2018-01-01 21:58:36.3530	2018-01-02 08:06:55.8720
start_station_id	72	72	72	72	72	72	72
start_station_latitude	40.7673	40.7673	40.7673	40.7673	40.7673	40.7673	40.7673
start_station_longitude	-73.9939	-73.9939	-73.9939	-73.9939	-73.9939	-73.9939	-73.9939
end_station_id	505	3255	525	447	3356	482	228
end_station_latitude	40.749	40.7506	40.7559	40.7637	40.7747	40.7394	40.7546
end_station_longitude	-73.9885	-73.9947	-74.0021	-73.9852	-73.9847	-73.9993	-73.9719
bikeid	31956	32536	16069	31781	30319	30106	32059
usertype	Subscriber	Subscriber	Subscriber	Subscriber	Subscriber	Subscriber	Subscriber
birth_year	1992	1969	1956	1974	1992	1968	1978
gender	1	1	1	1	1	1	1

Figure 1. Dataframe view

For the *prediction challenge*, you are expected to predict the **demand for the bike-sharing system (number of dropoffs and pickups).** You should do the predictions for clusters of stations. This challenge consists of three tasks:

1. Cluster the stations spatially (nearby departing stations should be grouped together) in no less than 20 clusters. Tasks 2 and 3 will be based on the results of this clustering, and analysis should be performed on at least one cluster (e.g., the one with the largest demand). More is preferable.
2. You are expected to build a prediction model that, at the end of a day, allows to predict what the demand for a cluster of stations will be over the next 24 hours – i.e. not the total demand for the next day, but how the time-series of the demand will look like for the next day (e.g., given demand data until midnight of day 1, predict the number of pickups for all 1h intervals [6-7am, 7-8am, …, 11-12pm] in day 2). You should predict both the arrivals (i.e., bicycle dropoffs) and the departures (pickups). You should use a time aggregation of one hour or less. You can choose to use two different models or a single one to predict both. It is up to you to decide how to best formulate this problem as a machine learning problem. You should **not shuffle the data**. You should instead use the data from January to October (included) to train your model, and the data from November and December as a test set. As a reference,

a good model should be able to predict the test set with an R^2 of at least 0.60. You can use any sklearn regression model you want.

3. Overnight, the bike-sharing company manually repositions their bikes in order to ensure that the demand for the next day can be satisfied. You are expected to use the outputs from the prediction model above to compute the required number of bicycles to be placed in each cluster of stations analyzed in Task 2 at the beginning of the next day. To compute this number, you can use the cumulative of the arrivals and departures. The goal is to ensure that, over the duration of the next day, there will never be a shortage of bikes – or, if there is, the goal is to minimize the number of bikes in deficit. The number of bicycles required can be extrapolated as the maximum difference between departures and arrivals.

In the *exploratory component*, each group needs to address at least one new research question. Here, we expect you to formulate your own question, and follow the data sciences cycle. The project will be positively valued with one or more of the following extensions:

- Extension of the dataset with other relevant data (weather data, national holidays, special events, etc.)
- Generation and analysis of insightful visualizations;
- Usage of the breadth of techniques from the class beyond regression and data preparation (e.g. dimensionality reduction, clustering, classification, time series)

Some example research questions:

- How to make predictions for each station? What about a cluster of stations?
- Are there periodic and seasonal trends (winter, summer, …), and how can we model them?
- What is the impact of land use (e.g. proximity to bus/metro station, shops, residential area vs. business district.)?
- What stations are more uncertain in terms of their expected pickups and drop-offs, and how can we try to ensure that, with a predefined confidence level (e.g. 90% confidence), there will be no imbalances between the pickups and drop-offs that result in a shortage of bikes for next day?

Note: The ordering of tasks we mention is **not** mandatory. In other words, if you prefer to start with the exploratory component, and then go to the prediction challenge, this is also acceptable. You can mention that in the report (or invert Sections 2 and 3). Similarly, data analysis might appear after the introduction (if relevant). However, please be aware that a simple descriptive analysis of the data is not sufficient to complete the task. Make sure to go one step forward and try at least one of the techniques discussed in the course.

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 3 to 5 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_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 is important. 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 should use this code[3] 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 (above 2500), and anyway the report should under no circumnstance exceed the 3500 words.

- October 9 – Announcement of this challenge statement
- October 23 – Communication of group members (through DTU learn)

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

- Project Milestone – November 6

December 4 – Final submission – all materials, including report notebook. Submit through DTU learn