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 various kinds, including climate change, pandemics, social inequality, and declining biodiversity, shared mobility services offer an emission-free mobility option that is both efficient and attractive. In this project, we invite you to use your best data science skills to help operators manage their fleet, providing better service and enhancing their business model. We do not expect you to discover revolutionary business models with a single Data Sciences project; instead, we want you to address the mandatory questions (below) but also seek out new questions, new data, and 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 operate at its best and of making bike sharing more attractive. You can read more about it here[2]. The data itself is an interesting exploration in data science.

Project

The project has three components:

- *Prediction Challenge*: All groups must address the same problem.
- *Exploratory component*: Each group is invited to choose its own research questions and explore the data accordingly.
- *Report:* Each group should deliver one report in a paper format – use the IEEE double-column paper template (Manuscript Templates for Conference Proceedings | IEEE – or the one uploaded with this description on DTU learn). The maximum length is 6 pages. Besides this paper, you must also submit one (or multiple, if you prefer) well-structured and well-documented Jupyter notebook with all the code on which the report was based. The notebook must be self-explanatory and easy to follow and navigate. There should be a clear link between the sections/results in the paper and the corresponding source code in the notebook. The paper should follow the 5-part outline shown below. The suggested length of each section is indicative and intended as a guideline.

Section 1: Introduction + motivation (0,5 pages)

Section 2: Data analysis and visualization (1 page)

Section 3: Prediction Challenge (2 Pages)

Section 4: Exploratory Component (2 Pages)

Section 5: Conclusions (0,5 Pages)

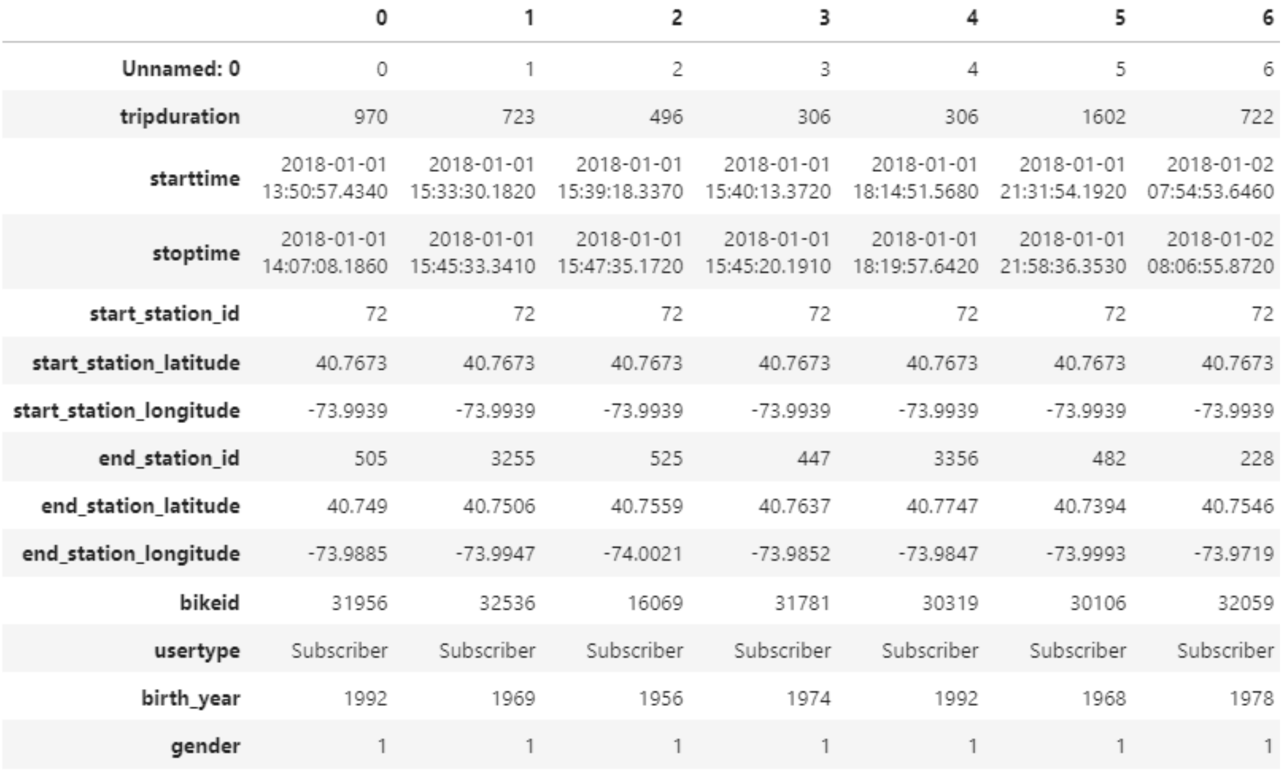
[1] https://s3.amazonaws.com/tripdata

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

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. Section 2 and 3 can be further divided into methodology and result sub-sections.

- **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 extensive treatment 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 this clustering, and they should be completed for at least two clusters (more is preferable) so that you can compare their respective results and discuss them.
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 (12-1am, 1-2am, 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 determine the most effective way to 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. You can use any model you want.

3. Overnight, the bike-sharing company manually repositions its bikes to ensure that demand for the next day can be met. 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 estimated by extrapolating the maximum difference between the number of 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 additional relevant data (such as weather data, national holidays, and special events).
- 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 (e.g., 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 ensure that, with a predefined confidence level (e.g., 90% confidence), there will be no imbalances between pickups and drop-offs that result in a shortage of bikes for the 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 may be presented after the introduction (if relevant). However, please note that a simple descriptive analysis of the data is insufficient to complete the task. Make sure to go one step forward and try at least one of the techniques discussed in the course. Finally, in the exploratory component, you are expected to use at least one of the techniques listed in class (classification, regression, supervised, unsupervised, dimensionality reduction, …). Ideally, you should also benchmark multiple techniques (e.g., multiple regressors in the case of regression).

Evaluation

The evaluation of the paper+notebook(s) will be based on the following criteria:

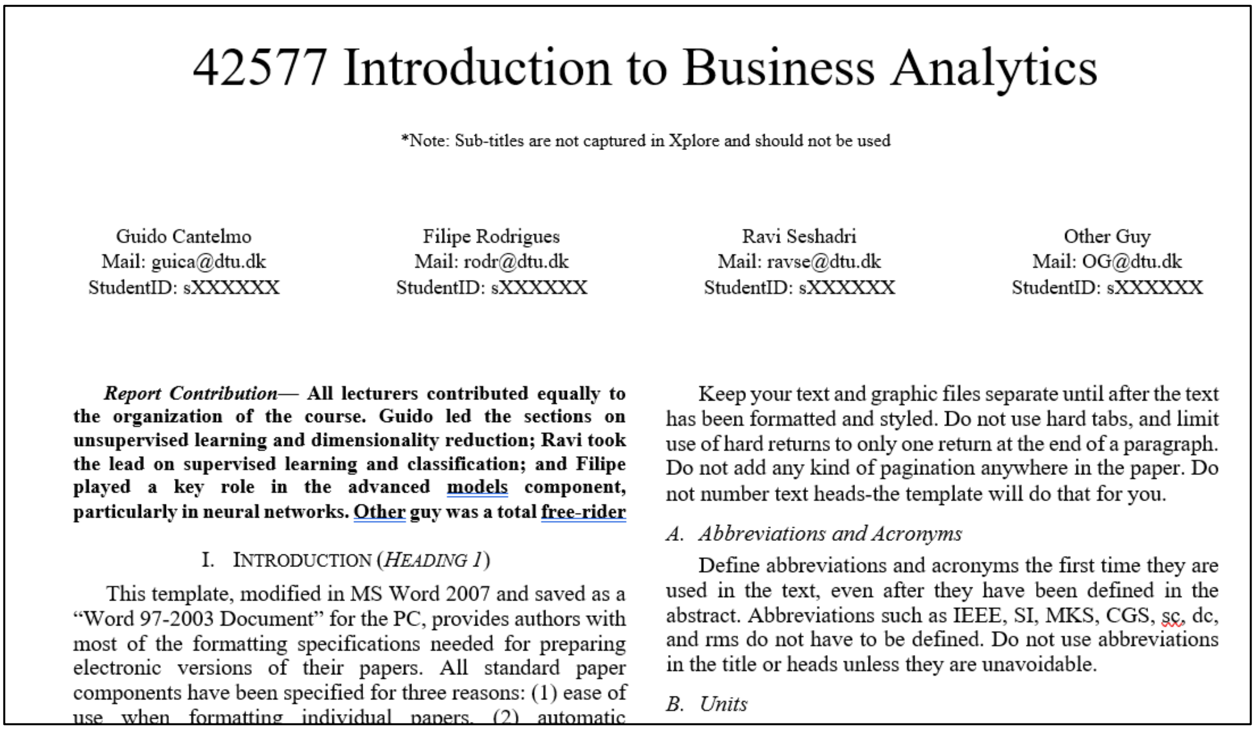
- Clarity – clarity of the paper and the self-explanatory nature of the notebooks
- Appendices – We will mainly review the paper and only consult the notebook if needed. If space constraints prevent including certain plots or analyses, they may be added as appendices in the notebook. These must be clearly referenced in the paper and, where appropriate, briefly explained.
- Thoroughness - Each research question should be examined to a suitable depth. The explanatory component is expected to be developed with a level of detail comparable to the prediction challenge.
- Insightfulness - Don’t stop at summarizing your findings — explain what kind of business insights or practical implications your results could provide.
- 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. Please add the relevant source material as a reference to the paper. **References do count towards the page limit**.

Rules

- Each group should consist of 4 students.
- The project submission shall be a zip file containing the paper PDF + 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 beginning of the paper, the abstract section should be used to **clearly clarify individual contributions**. In the event of doubts regarding individual contributions or the authenticity of the report, the teachers will call the group for an oral defense. **Individualized contribution section does count to the page limit**.

PLEASE INDICATE NAME, SURNAME, and STUDENT NUMBER IN THE REPORT. This should be indicated at the top of the article (author section). See the image below for an example

Deadline: December 5



42577 Introduction to Business Analytics

*Note: Sub-titles are not captured in Xplore and should not be used

Guido Cantelmo
Mail: guica@dtu.dk
StudentID: sXXXXXX

Filipe Rodrigues
Mail: rodr@dtu.dk
StudentID: sXXXXXX

Ravi Seshadri
Mail: ravse@dtu.dk
StudentID: sXXXXXX

Other Guy
Mail: OG@dtu.dk
StudentID: sXXXXXX

***Report Contribution*— All lecturers contributed equally to the organization of the course. Guido led the sections on unsupervised learning and dimensionality reduction; Ravi took the lead on supervised learning and classification; and Filipe played a key role in the advanced models component, particularly in neural networks. Other guy was a total free-rider**

I. Introduction (*Heading 1*)

This template, modified in MS Word 2007 and saved as a "Word 97-2003 Document" for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

Figure 2. Example of report, with names, emails, student ID, and report contribution.