Assignment

Data-X – aplikované analytické datové modely v reálných úlohách

Deadline: 30.4.2024

Submission: MS Team

# Data

* Data source: <http://insideairbnb.com/get-the-data/>
* Data dictionary: <https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit#gid=1322284596>
* Other publicly available external data are welcome.

# Submission Requirements

Consider the following problem definition:

***Predict the price per night in Prague as precisely as possible*.**

Tackle the problem using the data source provided in 1. Create a report that includes but is not limited to:

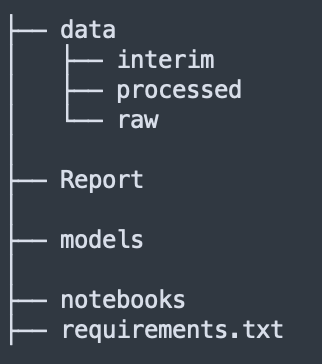
* **Data Understanding** 
  + Describe the dataset, report should include at minimum:
* Indication of missing values, type of values, unique values, outliers
* Descriptive statistics where applicable
* What can you conclude about the dataset after doing some data visualization?

Do not get stuck here. It is important to understand, that it is an absolutely necessary requirement to understand the data used for modelling and quite often also 80 % of the practical work. We, however, do not consider this as a content of the course. Therefore, even though you spend quite some time on this phase, it should not be a substantial part of your report/presentation. We just need to be sure that you did all the necessary steps before you get to modelling.

* **Data Preparation**
  + Prepare data to be ingested in the model of your choice. Provide a rationale for the features chosen as part of the training data. If you decide to remove or add columns or features explain why
* **Data Visualization**
  + Provide relevant plots that can be used to explain your model performance or rationale to choose predictors or features.
  + What can you conclude about the dataset after doing some data visualization?
* **Modelling**
  + Try different models and provide a rationale for your selected model choice and architecture. Describe your validation process. Your model report must include the following:
* Model limitations and considerations
* Ideas to improve the model
* Explain how you chose the values for the hyper-parameters of your model
* **Model interpretation**
  + Use appropriate methods to interpret the impact of your features on the predictions.
  + Try to interpret main interactions of the most influential features.
* **Bonus tasks**
  + Analyse the relation between the sentiment and price. Were people who paid more also more satisfied?
  + What high seasons did you identify? How do the seasons differ for different locations and estate types?

* **Project Structure**

Your deliverable should be structured as follows**:**



* *Report* - should include a pdf document with the items described above
* *Models -* Save and load your trained model.
* *Requirements.txt -* Execute pip freeze in your development environment and save all the dependencies utilized into this requirements file. More info about this file: [here](https://learnpython.com/blog/python-requirements-file/)
* *Notebooks -* Runnable code in jupyter notebook. This notebook will be run by the evaluators
* **Bonus**: Create a [Streamlit](https://streamlit.io/) Application and demo your model using an UI

\*\* Provide either a zip file or a link to a public repository with your solution (repo preferred). MS Teams channel will be created each team.

Do not forget that every source must be cited!

* **Results Presentations** 
  + Prepare a couple of slides to present your results, problem definition, classification reports, samples classified, learning curves, metrics chooses, or anything you consider relevant to defend your work and the rationale behind the decisions made building your project. Consider demonstrating the model using what we learned in class (Streamlit)