IS709: Introduction to Data Science Term Project

Group Submission Deadline: 28.12.2020

Project Submission Deadline: 24.01.2021

Dataset: You are going to work with ads listing data for this project. The data was scraped from Hurriyet Emlak for Ankara, Turkey covering both residential and commercial places between June 2019 and July 2019. You are not obligated, but if you want, you can use any additional datasets to enrich your analyses. For instance, you can use Google Geolocation API service to obtain more information about commercial places, schools or etc. Don't forget that it has limitations to retrieve data[1].

ad_id - Unique identification number for each ad listing
ad_title - Title for the ad listing
lat - Latitude of the location of the ad listing
lng - Longitude of the location of the ad listing
currency - Currency of the price
m2 - Area of the house/building advertised
type - Type of the listing
posted - Publishing date of the ad listing
price - Price of the ad listing

Important Reminder:
Mehmet Ali Akyol created the data set for his academic research. You cannot use this dataset for any purposes except for this assignment.

Instructions:

- Generate research questions upon your dataset. It should be clearly described what you hypothesize to find as a result of your analyses.
- You are expected to prepare a literature review section related to your research questions. A minimum of three papers needs to be included per person in the group. You should follow ACM author guidelines[2] for preparation. If you have time, it is better to learn Latex[3][4][5] while writing up (It is tidier. It also produces better equations, facilitates writing up and structuring the paper).
- Provide general information and descriptive statistics of your data set.
- Please feel free to use any preprocessing, feature selection, and machine learning tools you have found in the literature. Python environment provides several new packages on these topics. However, if you are comfortable with R, you can use it as well.
- Do not hesitate to create some derived attributes or manipulate or drop some features for better analyses.
- Divide your data set into a training, test, and validation sets. Explain how much of the data you included in each data set (training, test, and validation) and state briefly the reasons why you selected that proportion.

[1] https://developers.google.com/maps/documentation/geolocation/usage-and-billing#:~:text=While%20you%20are%20no%20longer,(QPS)%2C%20per%20user.
[2] https://www.acm.org/publications/authors/submissions
[3] https://www.overleaf.com/
[4] The Not So Short Introduction to Latex, https://tobi.oetiker.ch/lshort/lshort.pdf
[5] https://www.texniccenter.org/

- You are required to build at least four models in total. One of the models has to be a clustering model, and one of them has to be a classification model. The remaining two models could be any model you select (e.g., regression, association rules, statistical tests). You can stick to the models taught during the course or choose any other one you learned in different resources. Don't forget that you can use clustering models such as Self Organizing Maps for visualization tasks as well. The models don't need to be used together to solve a certain task/research question.If you group has two members, you should build six models in total (again one for classification and one for clustering and the remaining can be carried out with anything you would prefer).
- Justify your selection of parameters you used in your modeling. You can try different parameters in your training data set and use validation data set for parameter selection.
- Compare the results obtained with each model using various metrics and comment on your results in the report.
- You are expected to write a report explaining what you have done and what you have discovered. Each step should be carefully explained and justified in your report. You can support your findings with visualization techniques at any step.
- Don't forget to present a conclusion and discussion section.
- Respect academic integrity and use your own words while preparing your report.

Deliverables:

- Each project group may comprise at most two people.
- You are required to inform us about your group information latest by **28.12.2020 in ODTUClass**.
- At the end of the term project, you need to submit a report (in pdf or Docx format) and your Python codes (in ipynb format) or R codes. The paper/report should be a maximum of 6 pages in ACM format.
- You can check the papers below for writing styles. Pay attention to how they posed their research questions, explained the experiments and presented the conclusion and discussion:
    - Visuri, A., van Berkel, N., Luo, C., Goncalves, J., Ferreira, D., & Kostakos, V. (2017, September). Predicting interruptibility for manual data collection: a cluster-based user model. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (p. 12). ACM.
    - Mark, G., Iqbal, S. T., Czerwinski, M., & Johns, P. (2014, April). Bored mondays and focused afternoons: the rhythm of attention and online activity in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3025-3034). ACM.
- Your project report: Must include all the analyses you performed in detail, justifications, and comparisons. **You must explain the reasoning behind your studies. Each number you used needs to have a justification.**
- Your results **must be reproducible**. Therefore, please be sure that you used seeds and your codes are executable in your ipynb files.
- You are required to upload your files to the ODTU-Class with a compressed file named as "**YourID_Project.rar**".
- One submission per each group is sufficient.