DI501: Introduction to Data Informatics

Term Project

Deadline: 30 January 2022

With this project, you will be able to practice what you have learnt during the class.

We will be using a dataset from Kaggle (link = https://www.kaggle.com/c/house-prices-advanced-regression-techniques). This dataset consists of 79 explanatory variables that describe different aspects of houses in a city of Iowa, US. Each of the attribute is explained provided in “data_description.txt” in the ZIP file, you should definitely check them before doing any kind of analysis. You have two datasets provided, “train.csv” and “test.csv”. You should use train.csv to train your models and test.csv to evaluate your model’s performances. Although this dataset is mainly for predicting house prices, we will approach it differently and try to understand individual variables as well.

At the end of the project, you are required to produce an IEEE format paper and a Jupyter notebook for demonstrating your implementation. Note that you will not submit the dataset with your notebook. Please put your dataset folder in the same folder as your notebook and use relative paths while accessing any file.

The required tasks are as follows:

Step1 (10 pts): Develop at least two interesting, creative and feasible research questions based on the dataset.

You should write research questions considering the following factors[1]:

“A good research question should be:

- *Clear and focused.* In other words, the question should clearly state what the writer needs to do.
- *Not too broad and not too narrow*. The question should have an appropriate scope. If the question is too broad it will not be possible to answer it thoroughly within the word limit. If it is too narrow you will not have enough to write about and you will struggle to develop a strong argument (see the activity below for examples).
- *Not too easy to answer.* For example, the question should require more than a simple yes or no answer.
- *Not too difficult to answer.* You must be able to answer the question thoroughly within the given timeframe and word limit.
- *Researchable*. You must have access to a suitable amount of quality research materials, such as academic books and refereed journal articles.

[1] Taken from https://www.monash.edu/rlo/research-writing-assignments/understanding-the-assignment/developing-research-questions

- *Analytical rather than descriptive.* In other words, your research question should allow you to produce an analysis of an issue or problem rather than a simple description of it."

You should pay attention how you use your words while formulating your questions. There are several guidelines regarding how to pose a research question. For instance: https://writingcenter.gmu.edu/guides/how-to-write-a-research-question

Step2 (10 pts): Find at least three related papers from the literature related to your research questions and explain them briefly in your paper under "Introduction" section.

You should not choose a random paper returned from your query in Google Scholar. The paper should have been published in a respectable journal or conference proceeding. There are many ways to check the quality of the paper. Some of them are:

1. It might be indexed by Web of Science: https://www.webofscience.com/wos/woscc/basic-search
2. It might have been published by respectable publishers such as IEEE, ACM, Elsevier or AAAI.
3. Check the paper or the conference/journal name from https://www.sciencedirect.com/ or https://www.scimagojr.com/journalrank.php

You are also expected to reference them properly in your report in IEEE format. All the fields should be complete and proper. Do not rely on what returns from the Google Scholar as they are often missing. Check the origin of the paper. Although some appear to be published in arxiv, they might have been published in a respectable journal or conference afterwards. Hence you need to update the details accordingly.

Step3 (10 pts): Apply data quality checks (data profiling). Show the descriptive statistics on the features you have selected for your research questions.

Explain how you checked the quality of the dataset. Give sufficient information about the procedures you followed in your paper. Your Jupyter notebook should include all the codes you ran for this purpose while the paper should show the results in sufficient detail (this is valid for your further analyses). If there are missing values, explain how you dealt with them.

Step4 (10 pts): Apply feature selection, feature elimination or feature generation using a systematic analysis (at least two applications of them).

You should explain which features are important or redundant using a feature selection algorithm. You can aggregate certain features in order to make it more informative for your analysis. As an alternative, you can apply a dimensionality reduction algorithm. However, you are expected to show two different analysis and you are expected to relate them to your method and/or your research questions. In addition, you are required to explain how you pre-process your dataset (normalization, discretization or any other issues).

Step5 (10 pts): Apply at least one machine learning model.

You are expected to apply at least one machine learning model (classification or clustering). In addition to that, you can prefer to apply any statistical test if your research question requires you to do that. But if you want to do it, you should check any assumptions the test needs and you need to state and show them clearly in the paper (in the Jupyter notebook). The method you have chosen can be selected among the methods covered during the class or any other.

Step6 (10 pts): Model tuning.

Explain how you constructed/tuned your model in detail. For instance, any parameter you have chosen (you cannot simply say that they are default parameters; you will get no points if you do that), architecture selection etc. Explain whether you used a validation dataset for model tuning and if so explain how you selected your validation dataset. Justify your selections.

Step7 (10 pts): Comparison.

Choose at least two different metrics for comparing your model with an alternative one. Justify why you have selected them. You can create a baseline model if it is applicable. Put your results in a table and discuss your findings. Comment on whether your model is statistically superior to the one that you are comparing with (model from the literature, or a baseline). You are expected to conduct a statistical test for that.

Step8 (5 pts): Conclusion and discussion.

Explain whether you were able to answer your initial research questions with your models and how. Explain the limitations/advantages of your approach. State your future work.

Important Issues to Consider (for Grading!):

1. Your whole analyses should be connected/related to each other. You should not do something for the sake of doing it. For instance, if you do apply a feature engineering and do not use the results in somewhere else or do not relate it to your research questions, your mark will be degraded accordingly.
2. The report should be in IEEE format and have a maximum of 6 pages including the visualizations and references. You can find the template at: https://template-selector.ieee.org/ (choose publication type: Conferences, Original Research, Word or Latex format). You will be awarded 10 pts for it.
3. Use at least one visualization method. But it should be related/relevant to the content of your paper (5 pts).
4. You are not allowed to use any ready Kaggle scripts in your analyses. If we identify it, you will get no points from the step you used the code for.
5. Do not leave the documentation to the last minute. Many students, leaving this task to the last minute, lose a lot of points needlessly.
6. Creativity/ Novelty: We appreciate the hard and creative work that you put in your project. If you demonstrate the work is novel/creative (based on literature etc.), you will be awarded with a bonus up to 20 pts.
7. Your Jupyter notebooks should be self-explanatory and easy to run. We should see all the steps you followed for your analyses in the notebook. You cannot use any other platform (partial runs are not allowed such as using Excel or R). Because it will be difficult for use to check the codes.
8. Storytelling: You are expected to conduct your analysis by referring to your research questions. You can pose sub research questions. It is important to tell your story in a nice easy to follow

flow.(+10 pts). Spell and grammar check before the submission of your report. You can use Grammarly.

9. The projects should be prepared on an individual basis. You are not allowed to work together. You are expected to follow academic integrity.
10. The deadline is strict and it is within the final exam week. Hence, start working on your project and documentation as soon as possible. You can’t finish it if you leave it to the last minute.