ID5059 - Knowledge Discovery & Data Mining

Coursework Assignment 1 - Individual

Deadline: Friday 24th February 2023 (week 6), 9pm

Credit: 50% of coursework mark, 20% of overall module mark

Learning Objectives

On successful completion of this assignment you should be able to:

- investigate the properties of a real-world dataset
- construct several simple machine learning models aimed at predicting a numerical attribute from other attributes
- perform a simple evaluation of the results of those models

Requirements

It is not necessary to obtain the best possible performance by searching far and wide for the most state-of-the-art algorithm. You should instead use a reasonable model and performance measure similar, if not identical, to those discussed in our lectures and readings. Which model you use, and how you evaluate its performance, is up to you. **If your model performs poorly by your selected metric, do not worry. Your goal is to find a sensible approach and to produce clear, concise, understandable code and text documenting your effort.** Do not attempt to code everything from scratch. You are expected to use packages discussed in the lectures and readings, or similar. However, you should understand, and be capable of explaining, the packages you use.

Data

You will use data from the US Used Cars dataset on Kaggle, posted by Ananay Mital. **Do not download the data to a lab machine or School server**.

The full dataset contains about 3 million rows, and the file size is 10GB. There isn't space on School systems for everyone to have their own separate copy. Instead, the full dataset and

various subsets are provided on studres:

- The directory 4_huge contains the original file (10GB), plus an 80/20 split into training (8GB) and testing (2GB) files.
- The directory 3_large contains the same data as the original file, split into 10 equal-sized chunks (1GB each). It also contains an 80/20 split into training and testing files of one of those chunks.
- The directory 2_medium contains the same data as one of the large files, split into 10 equal-sized chunks (100MB).
- The directory 1_small contains the same data as one of the medium files, split into 10 equal-sized chunks (10MB).

All of these can be accessed via the file system from a School lab machine or server. There is a starter notebook (HTML version), showing how to do this in Python, that you can adapt if you wish (in which case you will need to download a copy of the notebook before you can change it).

Instructions

The task is to predict the list price (the attribute *price*) from a subset of the other attributes. Some, but not all, of the attributes are described on the Kaggle page.

You will need to:

- explore the structure of the data, and identify a small number of attributes that could plausibly correlate well with the price
- split into training and test sets (for small or medium files; this is already done for you for the large and huge files, to avoid you having to make separate copies)
- perform data cleaning, if necessary
- select and train a few models
- evaluate performance on the test data using an appropriate measure

It's strongly recommended that you start with one of the small data files, and gradually scale up to the larger ones once you have established your process. If you find that you can't feasibly process the full dataset this won't be a problem, as long as you clearly explain what you have done. Be aware, though, that the smaller files extracted from the larger ones are not random samples.

You can use either R or Python. If you have a strong reason for wanting to use something else, talk to the lecturers. Unless you have obtained explicit permission in advance, other languages will not be accepted.

Key Points

- We are not looking for a model that performs well: we are looking to see if you can build a sensible model and a sensible evaluation of its performance, and also if you can clearly document your effort.
- Your solution shouldn't need to include more than a couple of hundred lines of code, though you won't be penalised for more.
- If you are struggling to make something work with the volume of data present, you can use one of the smaller subsets. But explain what you have done, and why it is sensible.
- In the summary report, explaining the reasons for your decisions and showing your insight into the issues encountered is more important than including detailed descriptions of multiple models.
- Presentation counts. A concise Jupyter notebook complete with markdown annotations or some equivalent will earn more marks than an enormous raw text file full of opaque and poorly commented code.
- If you do not understand something or have questions, you are encouraged to discuss it with your peers (say, via the Teams channel) or the module staff. However, the deliverables that you submit must, of course, comply with the policy on Good Academic Practice.

Submission

A single zip file containing the following must be submitted via MMS by the deadline. Submissions in any other format will be rejected. **You are reminded of the importance of re-downloading and checking immediately after submission.**

1. The code of your solution, preferably in a Jupyter notebook with markdown annotations, or something similar built to be read with a web browser or PDF reader.
2. A clear and concise summary report (**one page maximum**) listing the models that you tried, and for the best model, giving details of the model, your measure(s) of performance, and your results, in a PDF file.

Assessment Criteria

This assignment will be marked on the standard 20-point scale using the following mark descriptors:

- **0** Nothing submitted.
- **1-3** Little evidence of any significant attempt to complete the work.

- **4-6** No substantial relevant material submitted, or no evidence of significant progress on any of the basic ML process elements.
- **7-10** A reasonable attempt at most of the basic ML process elements.
- **11-13** A competent attempt at most of the basic ML process elements.
- **14-16** A clearly presented notebook and summary report, demonstrating significant attempts at all of the basic ML process elements, and success in most elements.
- **17-18** An excellent clearly presented notebook and summary report, demonstrating successful implementation of all the basic ML process elements.
- **19-20** An exceptionally full and clearly presented notebook, with a very clear summary report, demonstrating successful implementation of all the basic ML process elements, and conveying significant insight into issues around data cleaning, model selection and evaluation.

Lateness

The Computer Science standard penalty for late submission applies (Scheme B: 1 mark per 8 hour period, or part thereof).

Good Academic Practice

The University policy on Good Academic Practice applies. This is an individual assignment. Any aspects of the submission that are not entirely your own work must be explicitly acknowledged.

Graham Kirby and Chrissy Fell, January 2023