University of St Andrews

School of Computer Science

ID5059 KDDM

Assignment: **P01**

Deadline: 20 August 2021 *Credits*: 20% of the module

You are expected to have read and understood all the information in this specification and any accompanying documents at least a week before the deadline. You must contact the lecturer regarding any queries well in advance of the deadline.

Aim / Learning objectives

This short project tests your ability to construct and evaluate machine learning models. You are asked to produce a simple machine learning model and a measure of its performance. 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 the UK government's land registry data. The task is to predict how much a property will sell for.

You can obtain the data at:

https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads

Download the **complete Price Paid Data-Single file** (CSV) which contains all registered purchases since 1995 as comma separated values. The data file is about 4 GB in size.

A description of the data's attribute columns can be found at:

https://landregistry.data.gov.uk/app/ppd

Instructions

Each row in the file contains a property that was purchased, the price that was paid, and features of the property and purchase. Your model should predict the price that was paid

from at most three simple features: the estate type (i.e. leasehold or freehold), the property type, and whether or not the property is in London.

In the data:

- Rows are separated by newlines, while the columns are separated by commas.
- Column 2 contains the price that was paid.
- Column 5 contains the property type (meanings of the codes can be found in the link above).
- Column 7 contains the estate type.
- Column 12 contains the town or city in which the property was located.
- You can judge a property to be in London if this field contains the word “London”.
- Ignore the remaining columns.

Testing and Training

Any purchases prior to 2017 are to be used as training data, while those made in 2017 or later are to be used as test data. The data consists of about 26 million instances, so you might want to use something like the Unix split utility:

```
split -l260000 pp-complete.csv ppsplit-
```

There are two subsets of the data on Moodle – one from 1995 the other 2020.

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.

- You should submit a solution with no more than a couple hundred lines of code.

- If you are struggling to make something work with the volume of data present, you can subsample (for instance, look at a month or a year’s worth of data). But explain what you have done, and why it is sensible.

- If you are having trouble extracting features, can you submit an evaluation of a sensible baseline on the test data?

- You can use any programming language you like to solve the problem: pick a language suited to the task, and one you are comfortable with, but your code must be presentable and understandable.

- 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 myself. However, the deliverables that you submit must comply with the policy on good academic practice.

Submission

Upload two things via moodle:

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 brief, clear and concise summary describing your model, your measure of its performance, and your result, in a single-page PDF file.

Assessment Criteria

Marking will follow the guidelines given in the school student handbook (see link in next section).

Policies and Guidelines

Marking
See the standard mark descriptors in the School Student Handbook:
http://info.cs.st-andrews.ac.uk/student-handbook/learning-teaching/feedback.html#Mark_Descriptors

Lateness penalty
The standard penalty for late submission applies (Scheme B: 1 mark per 8 hour period, or part thereof):
http://info.cs.st-andrews.ac.uk/student-handbook/learning-teaching/assessment.html#lateness-penalties

Good academic practice
The University policy on Good Academic Practice applies:
https://www.st-andrews.ac.uk/students/rules/academicpractice/

Tom Kelsey

July 2021