Hi everyone,

Thank you for showing interest in our hack – we're delighted by the response! This document should provide all you need to complete the task.

Task Details

At Auto Trader one of the most common questions we are asked is "How long will a car take to sell?". This isn't just useful for private sellers selling their own car, but a vital number for professional retailers as the more stock they can sell in a month, the more revenue they generate.

You have been provided with a dataset of vehicles sold in Oct 2024. Each row is a single advert that was de-listed on Auto Trader. Often this means it was sold, however we do see adverts disappear for other reasons, such as being sent to a dealer-to-dealer auction.

Your target to predict is 'days_to_sell' which is the difference between the date first seen and last seen on Auto Trader. You also have access to many taxonomy fields, such as fuel type (e.g. Petrol), some pricing information and some internal Auto Trader metrics such as advert quality. 'Attention_grabber' is also included, which is a small piece of free text that appears on the advert's search card. Note that this is taken on the advert's last day advertised on Auto Trader.

You have access to a data dictionary with a brief description of each column in the main dataset. There is the data stored in parquet (which can be natively read in by pandas via pd.read_parquet("MY/PATH/oct_2024.snappy.parquet"). There is a data dictionary, data_dict.csv, that contains a row per column name in the main dataset and a brief description of that row. Finally there is a very short ipynb, MMU_Hack_getting_started.ipynb, which contains a simple read in of the data and covers some gotchas with the data with some examples.

This dataset has had minimal cleaning and is representative of real-world data and all its imperfections. You are encouraged to be creative in your feature engineering, and to keep in mind how practical it would be for the business to implement the proposed model. Also consider what data would be available at inference time for a production model, compared to what's present in this example dataset.

We're keen for all teams to take part and the last day to submit entries will be Friday 7th March. Please remember to submit your team name and members when you do so.

It would be great if your work was submitted in a Powerpoint presentation format rather than a notebook as this represents what would happen in a commercial business such as Auto Trader. Showcasing work is an important part of working in Data Science so maybe think about how best to show and present your findings, whilst highlighting what approaches you took and why.

Once entries have been submitted, we will share who we think the top four entries are on Monday 10th March and those teams will be able to present their work at our event on Friday 14th March. We encourage all teams to join on this day to learn how we'd approach this task. There may even be additional prizes on the day...

Please submit entries to the two email addresses below and feel free to reach out if you have any questions. Enjoy!
Alys – alys.davies@autotrader.co.uk
Tom – tom.armitage@autotrader.co.uk