Datium Data Science Test

Section A

The goal of this exercise is to predict vehicle sale prices (*Sold_Amount*). Prepare a report documenting your thought process in building out the prediction model. The process should display some of the following considerations:

- Data understanding including highlighting errors and concerns with the data.
- Features to select and/or engineer. Take note that you cannot use the following fields: *AvgWholesale, AvgRetail, GoodWholesale, GoodRetail, TradeMin, TradeMax, PrivateMax*
- Experimentation with various feature encoding and modeling techniques.
- Different ways of evaluating the performance of the model and diagnosing the model for areas where it may be underperforming.

You can use Jupyter Notebooks for the report, but feel free to experiment with different Python libraries apart from the generic Scikit-learn and scientific libraries. For the sake of managing your experiments, feel free to play with tools like MLFlow. If you wish to focus the exercise more on engineering and deploying the model, you can keep the modeling process basic but demonstrate the use of tools like Seldon and Kubeflow for training and productionizing models.

Please place all code and logs in a .git repository with README files for executing the code or any demo if necessary.

Section B

We have an app that allows users to take full-sized images of a vehicle to be valued and sold accordingly. The goal of this exercise is to build an algorithm to identify damaged areas of the vehicle. You are expected to spend more time researching and understanding existing deep-learning architectures and research work in the field. In the report, you should be explicit about:

- Preprocessing techniques
- Trade-offs between different deep-learning architectures and/or transfer learning options
- Flowchart or plan on the detection architecture and the number of models required. For example, the images may contain multiple vehicles or no vehicle at all.

The following links have demonstrated some work in this area, but none of them are fully comprehensive and/or do not employ the latest developments in the field:

- https://beta.vu.nl/nl/Images/stageverslag-deijn_tcm235-882561.pdf
- https://github.com/neokt/car-damage-detective
- https://arxiv.org/pdf/1804.11207.pdf
- https://www.analyticsvidhya.com/blog/2018/07/building-mask-r-cnn-model-detecting-damage-cars-python/

You have the option of preparing a research report without actually implementing the algorithms, as the primary intent of this exercise is for you to be accustomed to literature research and display thorough understanding of the problem instead of blindly using existing libraries. If you wish to also implement it, you will have to source for the training datasets yourself (some of the links above will point you to publicly available datasets). Please include all references and links in your report.