CASE STUDY 1 – MACHINE LEARNING

Name: Course Name: Student ID:

Background and Context

Your group is the Data Science team for an e-commerce company named "True Brands". The Product team of the company wants to ensure that products published on the platform by 3rd party sellers are genuine products of the brand mentioned in the product description.

One way to do this is to build a model that can identify the brand from the product image uploaded. So your team is tasked with building a CNN model to identify the brand of a given product image. There are a few other features also available although you should not use anything with Type, SubType, Article etc since they are extremely unreliable and can be easily faked.

The data itself has not been collected well and has a lot of gaps. For some images no brand information is available. These should all be considered as fake images for the purposes of this model, since these have been uploaded with incorrect brand info by the 3rd party sellers.

Also in the data you may find certain brand info for which no product images are available. These should be ignored since the images here were lost due to an old bug in the code.

You need to build a stacked model system with two CNN models. The first one identifies if a product image seems to belong to the original brand or has been uploaded with a fake brand. The second model then looks at the ones that the first model identified as genuine and then determines which brand does the image belong to.

While the first model cannot use any additional features besides the image, the second model can use some additional features that are reliable to improve its prediction accuracy.

You also believe the management will be very impressed if you build interpretability into your system. It will help them build further trust in your model.

The dataset can be downloaded from here
https://drive.google.com/file/d/1ihgVJlqRvgvnDKI1_NoGgFCxkkKmazAE/view?usp=share_link

Objective

To predict which product image is more likely to be genuinely belonging to the brand provided. And then to classify the product image into the brand that it belongs to.

Note:

You should use CNN based techniques for this problem or compare any modelling technique you use for this problem with CNN based techniques and interpret/explain the differences. Picking the right modelling technique with an explanation will be marked. Please note CNN Models can take a significantly longer time to run, so if you have time complexity issues then you can avoid gridsearch on these techniques or sample your training data correctly. Your test data should at least include 1000 product images and you should print your confusion matrices for both your models clearly else you will lose all pts on model performance.

There are 20% bonus points for building interpretability into your system.

CASE STUDY 1 – MACHINE LEARNING

Name: Course Name: Student ID:

Best Practices for Notebook :

- The notebook should be well-documented, with inline comments explaining the functionality of code and markdown cells containing comments on the observations and insights.
- The notebook should be run from start to finish in a sequential manner before submission. It is preferable to remove all warnings and errors before submission.
- The notebook should be submitted as an HTML file (.html) and NOT as a .ipynb

Best Practices for Presentation :

Like in real-world projects, the ultimate destination of any project or work is generally an executive or decision-making meeting, where you are supposed to present your solution to the business problem, based on the project/work you have done. The purpose of this presentation is to simulate that kind of experience and to draw the attention of your audience (a business leader like CMO, COO, CFO, or CEO) to the key points of your project, which are

- Business Overview of the problem and solution approach
- Key findings and insights which can drive business decisions
- Model overview and performance summary
- Business recommendations

Please keep the following points in mind while making the presentation:

- Focus on explaining the takeaways in an easy-to-understand manner.
- Inclusion of the potential benefits of implementing the solution will give you the edge.
- Copying and pasting from the notebook is not a good idea, and it is better to avoid showing codes unless they are the focal point of your presentation.
- Please submit the presentation in PDF format only.

Submission Guidelines :

1. There are two parts to the submission:
    1. A well commented Jupyter notebook [format - .html] with 2 – 5 pages report in a .pdf format. This includes problem introduction, analysis using graphs, and conclusion. The report is Optional and a well documented Jupyter notebook will be considered as a report but you will lose 2% total pts.
    2. A presentation as you would present to the top management/business leaders [format - .pdf] (you have to export/save the .pptx file as .pdf)
2. Any assignment found copied/ plagiarized with other groups will not be graded and awarded zero marks. An ADR will be filed for the academic integrity violation

CASE STUDY 1 – MACHINE LEARNING

Name: Course Name: Student ID:

3. Please ensure timely submission as any submission post-deadline will not be accepted for evaluation
4. Kindly refer to the assessment guidelines and **make sure you check the details** of every section to get a better understanding of the expectations in this project.
5. Submission will not be evaluated if,
    1. it is submitted post-deadline, or,
    2. more than 5 files are submitted

Assessment Guidelines :

Please find below the guidelines for assessment

- Presentation - 15 pts
    - Visualization - 30%
    - Insights, discussion and conclusion - 40%
    - Spelling, Grammar, Tonality, Brevity etc. - 30%
- Report - 15 pts
    - Problem Understanding - 20%
    - Data Analysis - 40%
    - Model Building and Tuning Steps and Explanation - 20%
    - Spelling, Grammar, Tonality, Brevity etc. - 20%
- Code - 70 pts
    - Code Quality - 20 pts
        - Code runs correctly - 30%
        - Commenting and Readability - 30%
        - Code Modularity and Structuring - 40%
    - Modelling and Results - 50 pts
        - Data Preprocessing Steps and Explanations - 20%
        - Model 1 Selection/Tuning and Evaluation Steps and Explanation - 20%
        - Model 2 Selection/Tuning and Evaluation Steps and Explanation - 20%
        - Model 1 & 2 Final Results Achieved - 40%