Rotten Tomatoes: Produce Classification Models for More Productive and Efficient Farms

Farming is not for the faint of heart. External pressures from climate change, seasonal variability, crop disease, and market volatility all impact the tight margins of the people responsible for the food on our table. Recently, mounting economic pressures from recent trade policies and tariffs have left U.S. farmers worried for a future of decreased export value. Retaliatory tariffs imposed by foreign governments target U.S. agricultural exports, making American goods more expensive and less competitive abroad. These can leave farmers with shrinking markets, surplus goods, and smaller revenues. Especially for smaller farmers, already at odds with a highly concentrated market, these tariffs can have detrimental effects.

These pressures are being felt acutely in California, a hub of U.S. agriculture. Farmers in California can expect an estimated loss of 6 billion dollars this year on products such as oranges, almonds, and wine. The U.S. government has floated potential aid packages for farmers, but no concrete action has been taken. In response, farmers are left to consider potential cost cutting measures.

You are a data scientist tasked with addressing this cost-cutting problem. Your task is to develop an image analysis model capable of classifying produce based on its physical characteristics, deeming it fit or unfit to sell. By helping to automate this sorting process you'll investigate the potential to cut labor costs, reduce food waste, and develop additional revenue streams for struggling farmers. Further instructions and the data you'll need to complete this case can be found on the case GitHub page linked here.

Case Study Rubric

Submission Format: Upload GitHub repo to Canvas

Purpose: This study is an opportunity to showcase your problem solving and technical skills through a hands-on project. As you work through this case you will be exposed to a practical application of image analysis and encouraged to think about creative real-world applications of your model.

Task: The link to this case study can be found in the 'Hook' Document. You will find a Google Drive link to the image data you will need to complete this task. After downloading the images and their respective folders (ripe, unripe, damaged, old), you will upload them to Google Colab and conduct an initial EDA. This will give you information about the distribution of RGB pixels across the images and allow you to identify initial trends across categories of tomatoes. You will then train a convolutional neural network model (CNN) off the existing images through a 80/20 train test split. This model extracts visual features from the images, detects patterns and existing structures, and then classifies images producing a prediction based on the training data.

You will ultimately produce a deliverable that covers the requirements outlined below. The deliverable be a GitHub Repository that includes:

- README: Executive Summary of Project and Purpose of file in the repository.
- SCRIPTS: Well documented code and data used.
- OUTPUT: At least 3 EDA plots with brief descriptions.
- REFERENCES: Any references used

Criteria for Success:

Category	**Details**
Formatting	- One GitHub repository (submitted via link on Canvas) - Repository should be titled 'CS3_Rotten_Tomatoes' that contains - README.md - LICENSE.md - SCRIPTS - DATA - OUTPUT - REFERENCES.md
README.md	- Brief summary of the repository and its contents. This doesn't have to be detailed, but should orient a user to the content and goals of your work. - Map of your documentation: - List what is in each folder - Instructions for reproduction - Order by which you used the materials and completed the case.
SCRIPTS file	- Well documented Colab notebook that contains the code used: - To execute your EDA - Create the CNN model - Evaluate model accuracy - Test example of the model showing it correctly classifying an image
DATA file	- Includes the Google Drive link to access the image data.
OUTPUT file	- In PDF format upload at least 3 EDA plots showing trends in the data set. - Code is provided for several, choose the ones you believe to be most interesting and impactful towards the project goal.
REFERENCES.md	- Please include any references used in IEEE citation styles that were **not** already included in the provided reference list.
LICENSE.md	- Use MIT as default