Chris Phillips 10/25/2023

Project Part 1:

Small Data Problem Analysis Report

Complete this document and submit it with your project.

Match the scenario with the most appropriate solution and explain your choice

Scenario #1: Travel Planner Problem

A travel planning company asks customers to share pictures of past vacations/holidays so their staff can identify what kind of trips they enjoy. The company offers three basic categories of trips:

- Exploring in the Forest
- Adventure in the Desert
- Relaxing on the Beach

As part of a new online trip planning software, the company is creating an AI bot that will automatically figure out from the uploaded photos which category is likely to be most appealing to the customer. The challenge is the company has fewer than 500 photos that are categorized, and they feel it will be difficult to train a model using such little data.

Scenario #1: Travel Planner Problem Should you use transfer learning or a synthetic data approach to solve this problem? Please explain your answer in a short paragraph containing 3-5 sentences.	For the Travel Planner image set, I would use transfer learning. Training models for high accuracy requires a lot of data, which can be expensive to acquire and time consuming to train. We could train a small data set from scratch, but we would likely run into overfit issues. With pre-trained models, the weights obtained from the models can be reused in other related image classification tasks. This will allow us to use the small image training set and save time training from scratch.

Scenario #2 Loan Funding Prediction Problem

A loan company has a fairly large dataset that they want to use to train a model that predicts whether or not a loan should be funded. The problem they face is the dataset they are using has a large class imbalance... they don't have enough examples of loans that were denied. This is creating a model that doesn't perform well, particularly for loans that probably should be denied.

Scenario #2: Loan Funding Prediction Problem Should you use transfer learning or a synthetic data approach to solve this problem? Please explain your answer in a short paragraph containing 3-5 sentences.	For the loan dataset, I would use augmented data, which can help to resolve the issue with imbalanced data - essentially, the data doesn’t include enough examples of loans that were denied. The data is biased towards approved loans, which can lead to Type II errors – false negatives – or loan applications that should be rejected. We could eliminate some of the approved loans (downsample), but a small data set would likely lead to overfitting and poor performance. So, by generating synthetic data to add to the denied loan category, we can better train the model to detect loan applications that are risky. The set will be balanced between classes, thus reducing overfitting and improving predictive performance.