Project Part 1:

Small Data Problem Analysis Report

Complete this document and submit it with your

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.	**Transfer learning:** The model of interest is a supervised learning task, and the type of data is Image data. Transfer learning is the methodology of choice: is very applicable in this scenario as the dataset of training images that is available to the company is small, and therefore it is benefited from transfer learnings knowledge transfer of image recognition of the **pretrained models layers**. As the first layers of CNNs recognise rudimentary features like edges and color frequencies, these filters are identical in most image data and hence, the weights and biases of these layers are identical, we will only adapt a classifier that is costumed to our data at hand, and **freeze the pretrained models layers,** to only train the classifier.

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.	**Synthetic Data:** In the loan funding prediction problem we are working with structured data, and it is a small dataset. Further the dataset is **imbalanced,** with the Loan Status class (feature) of value 1, being underrepresented, and therefore it is beneficial for us to generate additional synthetic data based on the underrepresented class of the target, **Loan Status ==1.** We will generate this data, and set a reasonable threshold to expand the original dataset with this synthetic data. Finally we will test precision, recall, and F1 score to see how the expanded dataset, including synthetic data, performs.