The data that our group has chosen to utilize originates from an innovative IoT device and is sourced from the highly regarded platform Kaggle.com. Kaggle is widely recognized as a reputable site where individuals can practice and work on their machine learning skills, even publishing their own datasets. However, after looking deeper into the source of the data, we discovered that it initially stemmed from a GitHub repository, which had its initial posting in a textbook. For our specific goals, we will refer to the textbook source while deploying and training the models on the data set downloaded from Kaggle.

After we started looking at the source of our data, it becomes apparent that it is derived directly from an IoT device that captures and records various environmental metrics on a continuous basis, which records 6 types of variables, while also producing the time stamp with the recording of the event. The device's purpose is to monitor the presence of an individual in the room, and accordingly, it diligently measures seven distinct variables within the dataset. The only downfall of this data is that if the data is tracking more than one individual in the room. The device classifies the event of having one person in the room in binary but does not offer the count of the individuals in the room. Being said, the variables that are being tracked encompass a wide array of factors, ranging from the ambient humidity in the room and room temperature to even the intensity of light in the space.

The first model, which we have not decided on the specific architecture yet, but have determined will be a deep learning algorithm, aims to predict the presence of an individual in the room based on the variables recorded by the IoT device. Meanwhile, the second model, and which will be an LTSM neural network, will aim to forecast the precise times when an individual is likely to be present within the room. Nevertheless, it is worth acknowledging that the dataset itself possesses certain limitations. We believe that the limited nature of the dataset stems from the relatively limited amount of number of measurements available, and limited size. With a total of 20,560 rows of data, this quantity may be deemed sufficient for some models; however, when it comes to conducting the comprehensive time series analysis we are aiming for, we are contemplating bootstrapping techniques to augment the dataset, thereby optimizing our models by providing more training data for our models.

While the development of machine learning algorithms for our educational purposes, the applications of our IoT device extend far beyond our academic journey. Companies may find some kind of value in our machine learning models in combination of the IoT device, particularly in highly sensitive spaces where the if the presence of an individual would require the activation of recording of the individual in the room. Considering that IoT devices often serve as triggers for other actions, we believe that our device could serve as a redundant backup for security systems. In this scenario that we are thinking of, the primary motion detection security system would initiate the recording process, while our device would serve as an additional layer of assurance by leveraging environmental cues to validify the presence someone being in the room if the motion sensing system were to fail. By combining these two systems, a comprehensive and foolproof security apparatus can be established, improving the overall reliability and effectiveness of the surveillance infrastructure that the company may have deployed.

References

Room Occupancy detection data (IoT sensor). (2021, December 27). Kaggle. https://www.kaggle.com/datasets/kukuroo3/room-occupancy-detection-data-iot-sensor?resource=download

Candanedo, L. M., & Feldheim, V. (2016). Energy and Buildings. *Energy and Buildings, 112*, 28-39.