Moksh Jain (Unity ID - *mrjain*)
Pratik Devnani (Unity ID - *prdevnan*)
Rahil Sarvaiya (Unity ID - *rhsarvai*)

Project Group 14

Dataset: The dataset that we have decided to work on is the ECONet Dataset. The reduced dataset has 4 qc flags along with location, timestamp and 2 other properties. The dataset is highly imbalanced so we will have to perform some transformations for pre-processing the data before we can train a model.

Project Idea: Since the dataset is highly imbalanced in terms of there being a minority of tagged anomalies, we will explore certain techniques that will help us in tackling this issue. One of the methods that we have considered for this problem is SMOTE. SMOTE (Synthetic Minority Oversampling Technique) is an approach that is used to tackle imbalanced datasets by oversampling the minority class. New examples are synthesized from existing examples. Another approach for tackling the imbalance is using Penalized Models. Penalized classification models such as penalized-SVM and penalized-LDA impose an additional penalty cost on the model for making incorrect predictions on the minority class during training. This option will be explored in our implementation.

After literature review, we have found the following approaches that we will consider for anomaly detection. One of the approaches would be to train a Bi-LSTM with an attention layer. Since the data is temporal and sequential in chunks, we can use this property to our advantage and train an LSTM for the task. This approach would be computationally expensive but we anticipate that it will generate good results. The second approach that we are considering is using a GAN. Using a GAN is also very computationally intensive but the idea of forecasting and then comparing with the actual is a well-rounded approach. The third approach that we are planning to potentially implement is an unsupervised machine learning approach called Clustering. The approach is pretty straightforward. Data instances that fall outside of defined clusters could potentially be marked as anomalies. Clustering does have some drawbacks in terms of slow predictions for unseen data, however in this case, it might be a productive approach.

Software: Python3, Google Colab, NumPy, Scikit-Learn, Pandas, TensorFlow, Keras, Matplotlib, Jupyter Notebooks

Papers:
1) https://ieeexplore-ieee-org.prox.lib.ncsu.edu/stamp/stamp.jsp?tp=&arnumber=9378139&tag=1
2) https://link.springer.com/chapter/10.1007/978-3-030-86362-3_11
3) https://ieeexplore.ieee.org/document/8692127

Team and Work Division: For the first phase, we plan to ensure that our data is accurately sampled. As mentioned above we plan to implement models that are computationally intensive, we must ensure that we are working with the correctly represented dataset. Once we are sure about the dataset, we will implement all the three models to and compare which model performs the best with respect to anomaly detection.

Midterm Milestone: By the midterm, we plan to have sampled the data for our training and explore unsupervised clustering algorithms and penalized classification models to see if the data has been sampled correctly.

After the midterm, we plan to implement the bidirectional LSTMs first and then implement the GANs.