Predicting Network Degradation

Our group would like to predict future network degradation using NetMicroscope's historical speed test data. Over the summer, Chase and Luca used EMA models to detect real-time anomalies in ping, upload, and download statistics. While this helps triage network issues, it cannot warn network operators of future issues. Therefore, we'd like to build an ML pipeline that flags high-risk regions of a network before degradation occurs. We will focus specifically on latency.

The project will be split into two classification tasks: first, we need to determine our labels for when devices have degraded performance. We would ideally do this manually, but NetMicroscope's historical datasets span thousands of records. Instead, we will need to build models to classify this status, likely using an EMA model (reconstructing the current system, but possibly considering variations of EMA or other models). This will create a dataset of previous network metrics with labeled degradation statuses. Finally, we will need to build a predictive model that, given a current network state, predicts the likelihood of new or continued anomalies (degredation) over a range of time periods (1 hour, 1 day, etc.).

We will primarily evaluate the final predictive model from part 2. This supervised model will have concrete ground-truth labels, allowing us to measure accuracy. We plan to use backtesting to evaluate our models on previous data, possibly in addition to data collected while the project is in progress. While we would ideally measure the accuracy of our classification model from part 1 as well, that evaluation would require defining an anomaly. Instead of this ambitious goal, we will consider some naive unsupervised evaluation techniques that are scheduled for lecture in the coming weeks.

Our learning objective for this project is primarily to experience the process of creating an ML pipeline end-to-end, using ML to solve a real-world industry networking problem. We plan to divide our work across the following members:

- Luca will prepare the data, implement the EMA, and consider other unsupervised learning models, ultimately creating a labeled anomaly dataset. This will include visualization tools and infrastructure to run our models as needed.
- Clarice and Kasey will be individually responsible for half of the models we want to test[1] for predicting future anomalies. They will then evaluate the models they create on a number of conditions, including: accuracy over different durations, F1 scores, ROC/AUC curves, and other metrics from class.

Our final analysis will be a completed code notebook that includes model recommendations. While we will consider our models' accuracies, we will also prioritize other considerations (e.g., time to inference and recall) that affect real-world applications.

[1] We plan to test: `Linear Regression and variants (Ridge Regression, HistBoostRegressor), logistic regression, SVM, Random Forest, Naive Bayes, and Time-Series models`