# Executive Summary of Machine Learning-Based Estimation of Hurricane Recovery Times

Hurricane insurance is a major risk product in parts of the United States prone to tropical storms. Correctly estimating the time it takes for a region's economic activity to return to normal in the aftermath of a storm presents a large value-add for insurance providers. Broader understanding of what is at risk and the potential for damage could direct research in mitigation investment and insurance analysis. Additionally, diversifying insurance policy options is critical to spreading out risk exposure, while protecting policyholders who are potentially unaware of the changing world we live in today. Using a dataset consisting of thousands of instances of US counties being affected by hurricanes, we trained a machine learning model to predict the time between a hurricane striking and the post-hurricane resumption of normal net electricity usage patterns of a county (hereinafter referred to as the *recovery time* associated with a hurricane in a county).

Our problem is a regression problem; that is, we are trying to predict a continuous quantity, namely recovery time. Therefore, we evaluate our model's performance using mean squared error (the mean of the square of the difference between our predicted recovery time, and the actual recovery time, on data that the model has not seen). Our two-stage random forest model, which is explained in further detail below, achieved a mean squared error of 6.97 months$^2$, a significant improvement over the naive baseline of 896 months$^2$ achieved by predicting that every recovery time is equal to the overall mean recovery time.

The most salient aspect of the data is that hurricane recovery times form a heavy-tailed distribution: 32% of our dataset consists of outliers, which on average have very high recovery times (mean of 58.75 months). Meanwhile, 68% of hurricanes have a much lower mean recovery time (7.35 months). Recovery time is not the only significant difference between the outliers and the majority of the data set: averaged over outlier counties, the median house price was 47% greater than the mean of the median house price across non-outliers ($90,728 vs. $61,849).

Based on our exploratory analysis of the data, we determined that it would be best for our model to operate in two stages: in the first stage, we use the random forest technique to predict whether the recovery time will be an outlier or not, and then in the second stage, we have two separate random forests (one for outliers, one for non-outliers) that predict the recovery time itself. This two-staged approach achieved the aforementioned 6.46 months$^2$ mean squared error; a single-staged model which does not treat outliers and non-outliers separately achieved a worse mean squared error of 63.2 months$^2$.

Further work should center on examining different choices of objective function. Although the mean squared error criterion that we used in our model is a reasonable choice from a general machine learning perspective, it is plausible that in certain business contexts, an asymmetric objective function should be used. That is, if the cost of predicting a recovery time that is too high is different from the cost of predicting a recovery time that is too low by the same margin, the objective function should take that difference into account and accordingly penalize over-predictions differently from under-predictions. The resulting model will then make predictions that better optimize towards our business objectives.