Conclusions

The purpose of this project was to implement a deep learning model to predict the total collected volumes of various waste streams in New York City. In order to train, test, and tune this model, historical data for collected volumes of various waste streams, by borough, as well as population and weather data were used. From these datasets, 16 features were identified and used. The hyperparameters, including hidden layer dimension, number of layers, number of epochs, and batch size were then explored and optimized. For the dimensions and number of hidden layers, 4 hidden layers with sizes [256,128,64,32] were identified as the best performing in terms of accuracy. Number of epochs was limited to 3000, which resulted in good accuracy while retaining good run times. Mini-batching also significantly increased performance, by up to 30%.

A feature importance analysis was performed. Shapley values were extracted, and the most important features were identified as the year, population, population percentage, month, minimum temperature, and average temperature. As borough data was one-hot encoded, each individual borough was considered a feature. An interesting trend emerged, where the boroughs had a wide spread of importances. For instance, Brooklyn, Staten Island, and Queens had high importance, whereas the Bronx and Manhattan had lower importance. This could be attributed to the fact that the latter boroughs are more densely populated, with less yards, or trees, and more commercial businesses. As a result, seasonal changes in waste streams such as yard waste are much less in those two boroughs.

Different optimizers were then tried to identify the best performer, including: Adadelta, Adagrad, Adam, AdamW, Adamax, ASGD, NAdam, RAdam, RMSprop, Rprop, and SGD. Adadelta, ASGD, and SGD were identified as the most effective, with losses of .58, .576, and .575, respectively.

With the use of all the above mentioned improvements and optimizations, training losses of as low as .046 were possible, with test losses of .859. In terms of error, the average error for Refuse, Paper, and Metal/Glass/Plastics were 10.8%, 15.7%, and 19.6 percent, respectively. However, for organics waste streams, error was significantly higher, from 53-107%. This can be attributed to lack of data, stemming from the impacts of service interruptions due to budget cuts and COVID. In addition, there is a lack of history for these waste streams, owing to the fact that collection programs were only recently introduced. In addition, the relatively low volumes of organic waste serves to amplify errors. Specifically, residential organic tonnage was off by 91%, or 800 tons per week. This was mainly due to the fact that collection for this waste stream was only resumed in September of 2022. As our testing and validation was conducted on the 2022 dataset, the predictions were off by 100% for most of the year, as there was little to no volume collected during that time. For the other streams, such as school organics, leaves, and Christmas trees, the errors, while significant in percentage, were only 101, 89, and 23 tons, respectively. Given that an average garbage truck of the type used by New York City has a payload of 10-15 tons, the predictions were only off by 2-7 truckloads a week, which is relatively insignificant given the number of garbage trucks on the streets of New York.