**NYC Rat Sightings Hot Spot Classification - Technical Report**

Business Problem

New York City has a major rat problem that has negative impacts on health, property values and cleanliness. The city’s 311 system collects rat complaint data, the identification of ‘hot spot’ areas would be helpful to prioritize/allocate rodent control resources. The goal of this project is building a machine learning model to classify zip codes as rat complaint hot spots, this would enable more efficient allocation of resources and provide city residents with an idea of the impacted areas.

Data Source and Preprocessing Steps

The dataset has 311 rat complaint data from 2010 to September 2017, including related location and event data. The features were engineered from create timestamps (month, year, day of week, hour and season) and response time was calculated when possible and replaced with a median value if missing. Categorical variables were one-hot encoded and numerical values were scaled using the standard scaler. Lastly, the target variable hot-spot was defined as zip codes in the top 10% of rat complaints.

Model Selection

Multiple models were ran such as logistic regression, MLPs with different architectures and test set performance was used for model selection. The logistic regression provided a baseline for accuracy with and the MLPs offered some marginal improvements.

Key Insights and Recommendations

The variables with the strongest influence on hot-spot classification were response time, borough and location type. The best model was the MLP (single hidden layer(10)). The model achieved a test accuracy of .732 and test auc of .768. Overall, the model can help NYC prioritize/allocate resources to high-risk zips and help inform city residents on which locations to avoid.

Limitations and Future Improvements

The limitation of this analysis is that the complaint data might be biased given certain areas/individuals might be more likely to report rat complaints compared to others. Another limitation is that the data is from the 2010s. A future improvement could be to use more up-to-date data and also compare/include rat complaint data from other cities.