Milestone #5

We have a variety of ideas left to explore at more length. These include:

- **Multi-class classifications:** Given that this information is contained within the data, it is important and interesting to predict a variety of classes, not just whether they pass or fail. It will also be telling if certain classes are easier to predict than others. Creating a useful multi-class classification model will be particularly valuable given the fact that there are a limited number of possible inspections done in any given period of time. We therefore want to measure the severity of likely violations for a given restaurant (as well as the probability). Using multiple classes is a good way of including the severity within the model, as long as the classes match up.

- **Geographical data:** Rather than geohashing to convert location attributes to numerical values, we think it could be more valuable to incorporate data regarding the neighborhood layout of Chicago, then group restaurants by neighborhood rather than simply by their latitude and longitudes. While more involved, our guess is that using this knowledge of the city's layout and of how a city actually functions could be more representative of the underlying data (i.e. even if restaurants are in practice only a few blocks away, they may have much more in common with the other restaurants in their respective neighborhoods than with each other. This would in essence geographically cluster the data, and could be due to a variety of factors including different neighborhood groups controlling rules and behaviors for establishments, or different clientele frequenting different areas).

- **Sanitation Code Complaints:** This is an important dataset to include to improve our baseline model. The dataset comes organized by location, so by cross referencing the dates and locations of health code violations with the dates and locations of the Sanitation Code Complaints, we can measure both whether local Sanitation Code Complaints affect the result of an inspection (i.e. maybe we can get a sense of whether the people running the restaurant have a cleanliness proclivity as exhibited by their lack of Sanitation Code Violations), as well as whether a recent Sanitation Code Complaint affects the likelihood of being followed up by a failed health code inspection.

- **Yelp data:** We've slowly been collecting data on Chicago restaurants using Yelp's python API. While we've encountered some cross-referencing issues between Yelp and our dataset, after accumulating more information we hope to test a variety of ways in which we can incorporate this Yelp data into our model. Intuitively, we believe it should be fairly predictive, as Yelp is many people's first resort when complaining about a restaurant's hygiene or cleanliness, not the health department.

- **Restaurants with repeat inspections:** Does a failing grade on an inspection incentivize restaurants to clean up their act? Do restaurants who pass become complacent and relax their hygeine standards?

- **Improving the Performance Metric:** In the future, it could be valuable to implement a function incorporating costs as our performance metric. Here we'd estimate not only the cost of an inspection, but also the respective costs (both to the city and to society) of false positives, true positives, true negatives, and false negatives. This function would allow us to choose which location on the ROC curve minimized the costs, and pick our thresholds accordingly. Estimating these costs, however, may involve a significant amoung of hand-waving, so for now we will defer to our models' sensitivity.