PREDICTING HEALTH CODE VIOLATIONS IN BOSTON RESTAURANTS

Springboard Capstone Project by Roy Wright – August 2016

CLIENT AND PROBLEM

The City of Boston, like many others, conducts health inspections of food service establishments in a largely random pattern of visits. With a model for predicting which of all the potential inspections will result in violations, the inspections can be carried out in a targeted way. This could make more efficient use of inspectors' time and, more importantly, increase the chances of detecting violations.

Our aim is to find restaurants that are likely to fail an unannounced health inspection — that is, we will develop the capability of predicting whether a health code violation would be found at a given food service establishment, if it were chosen to be inspected, without prior notice, at the moment that the predictive model is used. The main factors to be used in predicting an inspection outcome are the urban conditions near the restaurant in the weeks leading up to the inspection.

DATA USED

DATASET: CITY OF BOSTON HEALTH INSPECTION RECORDS

Detailed records of health inspections throughout Boston, from 2006 to present, have been obtained from data.cityofboston.gov. Each inspection record includes many pieces of information, but those that are most relevant to our purposes are the following:

- Food service establishment's name
- Type of establishment
  - category "FS" stands for "Eating & Drinking"
  - category "FT" stands for "Eating & Drinking w/ Take Out"
  - category "MFW" stands for "Mobile Food Walk On"
  - category "RF" stands for "Retail Food"
- The result of the inspection
  - "HE_Pass" when an establishment passes an inspection with no violations (either its first yearly inspection or a re-inspection a few weeks after failure)
  - "HE_Fail" when an establishment's first inspection of the year is unsatisfactory
  - "HE_FailExt" when an establishment is re-inspected within a few weeks after a failure, and fails again
  - "HE_Filed" for the first inspection of an opening establishment
  - other result codes – for example, temporary business closure because of an emergeny – are much less common, and do not pertain to the type of inspection studied here
- The date of the inspection
- Description of each violation found, or of each previous violation that has been corrected, along with its severity, rated as * or ** or *** (one star, two stars, or three stars)
- Written comments

- Business address
- Latitude/longitude coordinates – unfortunately, this information is not always provided as part of the inspection

DATASET: CONDENSED HEALTH INSPECTION RECORDS PROVIDED BY YELP

Yelp.com has published datasets for the 2015 "Keeping it Fresh: Predict Restaurant Inspections" contest, hosted by DrivenData.org. At the heart of these datasets is a list of Boston restaurant inspection results. Each row gives the inspection date, a "restaurant ID" for the establishment inspected, and tallies of the violations found at each severity level (*, **, and ***). Note that an inspection is passed if and only if there are *zero* violations found at each of these severity levels.

Another dataset provided by Yelp gives a summary of information for each business, including its name, address, "Yelp ID" number (which is distinct from its "restaurant ID"), number of Yelp reviews available, average rating, and latitude/longitude coordinates. Lastly, a table is provided for matching Yelp ID's to restaurant ID's.

TWO SETS OF INSPECTION RECORDS: WHAT IS THE DIFFERENCE?

As can be seen above, we have two separate datasets that provide records of past health inspections. To compare these datasets, let's take a look at the contents of each one for a single day. Below, we display part of the contents of the City of Boston's records for August 12, 2014. This date is chosen mostly arbitrarily.

name	category	result	level	description	comments	latitude	longitude
A C Farm Market	RF	HE_Pass	*	Non-Food Contact Surfaces	Repair rusty shelving below prep table.	42.302302	-71.059800
A C Farm Market	RF	HE_Pass	*	Food Protection	Discontinue store vegetables in stagnant wate...	42.302302	-71.059800
A C Farm Market	RF	HE_Pass	**	Insects Rodents Animals	Evidence of fruit flies provide exterminators ...	42.302302	-71.059800
A C Farm Market	RF	HE_Pass	*	Dishwashng Facilities	Replace missing sink plugs.	42.302302	-71.059800
A C Farm Market	RF	HE_Pass	*	Food Container Labels	Provide labels for all packaged foods.	42.302302	-71.059800
Choice's by Au Bon Pain	FT	HE_Fail	*	Non-Food Contact Surfaces	bakery/replace worn door gasket to 1 door reac...	NaN	NaN
Choice's by Au Bon Pain	FT	HE_Fail	***	Cold Holding	salad bar/carrot and raisin salad 49 degrees/c...	NaN	NaN
Choice's by Au Bon Pain	FT	HE_Fail	*	Equipment Thermometers	front line/ 2 door drawer/provide internal the...	NaN	NaN

name	category	result	level	description	comments	latitude	longitude
Choice's by Au Bon Pain	FT	HE_Fail	*	Non-Food Contact Surfaces Clean	salad bar/clean interior cabinets	NaN	NaN
Choice's by Au Bon Pain	FT	HE_Fail	*	Improper Maintenance of Floors	replace damaged floor tile in front of proof b...	NaN	NaN
City Sports	RF	HE_Fail	*	Hand Cleaner Drying Tissue Signage	restroom/provide employee must wash hands signage	42.350722	-71.073709
City Sports	RF	HE_Fail	*	Installed and Maintained	repair hot water faucet handle in restroom	42.350722	-71.073709
DUNKIN DONUTS(FRANKLIN)	FT	HE_Pass	NaN	NaN	NaN	42.356510	-71.053320
Foumami	FT	HE_Pass	NaN	NaN	NaN	NaN	NaN
Great Chef	FT	HE_Pass	*	Equipment Thermometers	Provide visible thermometers where necessary.	42.379493	-71.027910
Great Chef	FT	HE_Pass	*	Soiled Linen Storage	Basement -Cover dirty laundry container	42.379493	-71.027910
Great Chef	FT	HE_Pass	*	Improper Maintenance of Walls/Ceilings	Clean cooking vent hood.	42.379493	-71.027910
Great Chef	FT	HE_Pass	*	Wiping Cloths Clean Sanitize	Keep wiping cloths in sanitizer	42.379493	-71.027910
Great Chef	FT	HE_Pass	*	Food Protection	Elevate food containers 6" off floor in walk in.	42.379493	-71.027910
Great Chef	FT	HE_Pass	*	Non-Food Contact Surfaces	Defrost french fry freezer	42.379493	-71.027910
Great Chef	FT	HE_Pass	*	Non-Food Contact Surfaces Clean	Clean underside of shelving over prep tables.	42.379493	-71.027910
Great Chef	FT	HE_Pass	*	Food Protection	Cover all open foods in reach ins.	42.379493	-71.027910
Great Chef	FT	HE_Pass	*	Non-Food Contact Surfaces Clean	Clean and refinish rusted Can opener.	42.379493	-71.027910
KANTIN	FT	HE_Fail	***	PIC Performing Duties	The time as a public health control logs are n...	42.352411	-71.125329

name	category	result	level	description	comments	latitude	longitude
KANTIN	FT	HE_Fail	*	Installed and Maintained	The cold water at the back handwash sink is no...	42.352411	-71.125329
KANTIN	FT	HE_Fail	*	Non-Food Contact Surfaces	There is duct tape on the handle of the rice c...	42.352411	-71.125329
KANTIN	FT	HE_Fail	*	Premises Maintained	There is excess clutter in the upstairs storag...	42.352411	-71.125329
...	...	...	...	...	...	...	...
Samurai Kuang Eatery	FT	HE_Pass	***	Cold Holding	Sushi grade salmon 51F White fish 50F / Provi...	42.355795	-71.058451
Samurai Kuang Eatery	FT	HE_Pass	*	Equipment Thermometers	Dish machine gauge is broken / Repair.	42.355795	-71.058451
Samurai Kuang Eatery	FT	HE_Pass	*	Improper Maintenance of Floors	Floors under cookline around handsink heavily...	42.355795	-71.058451
Samurai Kuang Eatery	FT	HE_Pass	*	Non-Food Contact Surfaces	Back door opened without screen / Provide scr...	42.355795	-71.058451
Samurai Kuang Eatery	FT	HE_Pass	*	Improper Maintenance of Walls/Ceilings	Hood vents with visible grease build up / Clea...	42.355795	-71.058451
Shaw's Supermarket No. 586	RF	HE_Fail	*	Food Contact Surfaces Design	Sponge being used at the 3 bay sink in the pro...	42.271930	-71.069700
Shaw's Supermarket No. 586	RF	HE_Fail	*	Non-Food Contact Surfaces Clean	Interior of the chicken freezer near the rotis...	42.271930	-71.069700
Shaw's Supermarket No. 586	RF	HE_Fail	*	Improper Maintenance of Floors	Floor under the storage cabinet near the rotti...	42.271930	-71.069700
Shaw's Supermarket No. 586	RF	HE_Fail	*	Improper Maintenance of Walls/Ceilings	Portion of the wall in the meat walk-in cooler...	42.271930	-71.069700
Shaw's Supermarket No. 586	RF	HE_Fail	*	Installed and Maintained	Pipe under the hand sink in the meat preparati...	42.271930	-71.069700
Shaw's Supermarket No. 586	RF	HE_Fail	*	Non-Food Contact Surfaces	Salad bar unit operating at around 50F. PIC (B...	42.271930	-71.069700

name	category	result	level	description	comments	latitude	longitude
Shaw's Supermarket No. 586	RF	HE_Fail	***	Cold Holding	All foods inside of the salad bar withg temper...	42.271930	-71.069700
WORLD TRADE CENTER	FT	HE_Fail	*	Walls/Ceilings Designed Constructed Installed	replace stained and heavily soiled ceiling til...	NaN	NaN
WORLD TRADE CENTER	FT	HE_Fail	*	Floors Designed Constructed Installed	clean floor of paper goods storage room. also ...	NaN	NaN
WORLD TRADE CENTER	FT	HE_Fail	*	Non-Food Contact Surfaces	remove ice buildup from pipes inside walk i8n ...	NaN	NaN
WORLD TRADE CENTER	FT	HE_Fail	*	Premises Maintained	remove shoes from tops of lockers in both chan...	NaN	NaN

By contrast, here are the contents of Yelp's dataset for August 12, 2014, once they have been assembled from the three sources mentioned before.

date	restaurant_id	violations			name	latitude	longitude
		*	**	***			
2014-08-12	we39j9ok	4	0	0	Dunkin' Donuts	42.349264	-71.042474
2014-08-12	lnORdd3N	0	0	0	Dunkin' Donuts	42.356527	-71.053353
2014-08-12	KAoK8ZOg	0	0	0	Fóumami	42.356039	-71.053455
2014-08-12	0ZEDGWOD	9	0	0	Great Chef Chinese Food Day Square	42.379525	-71.027940
2014-08-12	njoZ1D3r	3	0	1	Kantin	42.352744	-71.125447
2014-08-12	eVOBLr3j	1	0	1	Lollicup	42.352444	-71.125403
2014-08-12	B1oXNlEV	11	1	3	Max Brenner	42.349491	-71.080588
2014-08-12	B1oX4boV	4	0	1	Samurai Kuang Eatery	42.355741	-71.058335
2014-08-12	8xExZeo0	4	1	1	South Boston Chinese Restaurant	42.336483	-71.047309

At a glance, we can see that the inspection records maintained by the City of Boston are more detailed than those provided by Yelp. It is easier to compare the two if we summarize the city's records for the day in question in a way that looks more like Yelp's version. This can be seen on the next page.

Note that some businesses present in the city's records do not appear in Yelp's data. In fact, it seems that no business of the "retail food" type appears in Yelp's data. On the other hand, every inspection listed in Yelp's data can be seen represented in the city's data, although not always clearly. For example, the first Dunkin' Donuts inspection listed in Yelp's data (see above), with four 1-star violations and no 2- or 3-star violations, is listed at the bottom of Boston's data (shown below), with the name "WORLD TRADE CENTER" – which is actually the location of that particular Dunkin' Donuts. But aside from that, each of the other inspections shown in Yelp's dataset can be found without much difficulty in Boston's dataset.

name of business	violation level	number of violations found or corrected	inspection result	establishment type
A C Farm Market	*	4	HE_Pass	Retail Food
	**	1	HE_Pass	Retail Food
Choice's by Au Bon Pain	*	4	HE_Fail	Eating & Drinking w/ Take Out
	***	1	HE_Fail	Eating & Drinking w/ Take Out
City Sports	*	2	HE_Fail	Retail Food
DUNKIN	none		HE_Pass	Eating & Drinking w/ Take Out
Foumami	none		HE_Pass	Eating & Drinking w/ Take Out
Great Chef	*	9	HE_Pass	Eating & Drinking w/ Take Out
KANTIN	*	3	HE_Fail	Eating & Drinking w/ Take Out
	***	1	HE_Fail	Eating & Drinking w/ Take Out
LOLLICUP TEA ZONE	*	1	HE_Fail	Eating & Drinking w/ Take Out
	***	1	HE_Fail	Eating & Drinking w/ Take Out
Max Brenner	*	11	HE_Fail	Eating & Drinking
	**	1	HE_Fail	Eating & Drinking
	***	3	HE_Fail	Eating & Drinking
Pho Viets	*	1	HE_Filed	Eating & Drinking w/ Take Out
Pollos A La Brasa Beto's	*	7	HE_Filed	Eating & Drinking w/ Take Out
SOUTH BOSTON CHINESE	*	4	HE_Pass	Eating & Drinking w/ Take Out
	**	1	HE_Pass	Eating & Drinking w/ Take Out
	***	1	HE_Pass	Eating & Drinking w/ Take Out
Samurai Kuang Eatery	*	4	HE_Pass	Eating & Drinking w/ Take Out
	***	1	HE_Pass	Eating & Drinking w/ Take Out
Shaw's Supermarket No. 586	*	6	HE_Fail	Retail Food
	***	1	HE_Fail	Retail Food
WORLD TRADE CENTER	*	4	HE_Fail	Eating & Drinking w/ Take Out

There are some important mistakes in Yelp's records. To see why, note for example that in the city records above, the inspection of Samurai Kuang Eatery is marked as passing. For that inspection, four 1-star and one 3-star violations are noted, but this is only because that inspection was a follow-up to an inspection from one week before, which found those violations. Unfortunately, the Yelp data treats these inspections identically, marking four 1-star and one 3-star violations for each of the two inspections, which is quite misleading. This mistaken double-entry of health violations happens frequently throughout the Yelp dataset. We will see later that this issue, once we make an appropriate correction for it, will not impede the central purpose of this project.

In light of the comparison above, we can now consider the advantages of each dataset of inspection results. The City of Boston's dataset is kept continually up-to-date, with new results being entered as they occur, while Yelp's data ends in mid 2015. Boston's dataset is also more complete in the sense that each violation is categorized in much finer detail than a simple a three-level severity rating. However, it should be noted that Yelp’s dataset was developed with the express support of the City of Boston, working toward a purpose

very similar to the purpose of the present project. Most crucially, this dataset provides latitude/longitude coordinates for *every* inspection location listed.

DATASET: SERVICE REQUEST CALLS

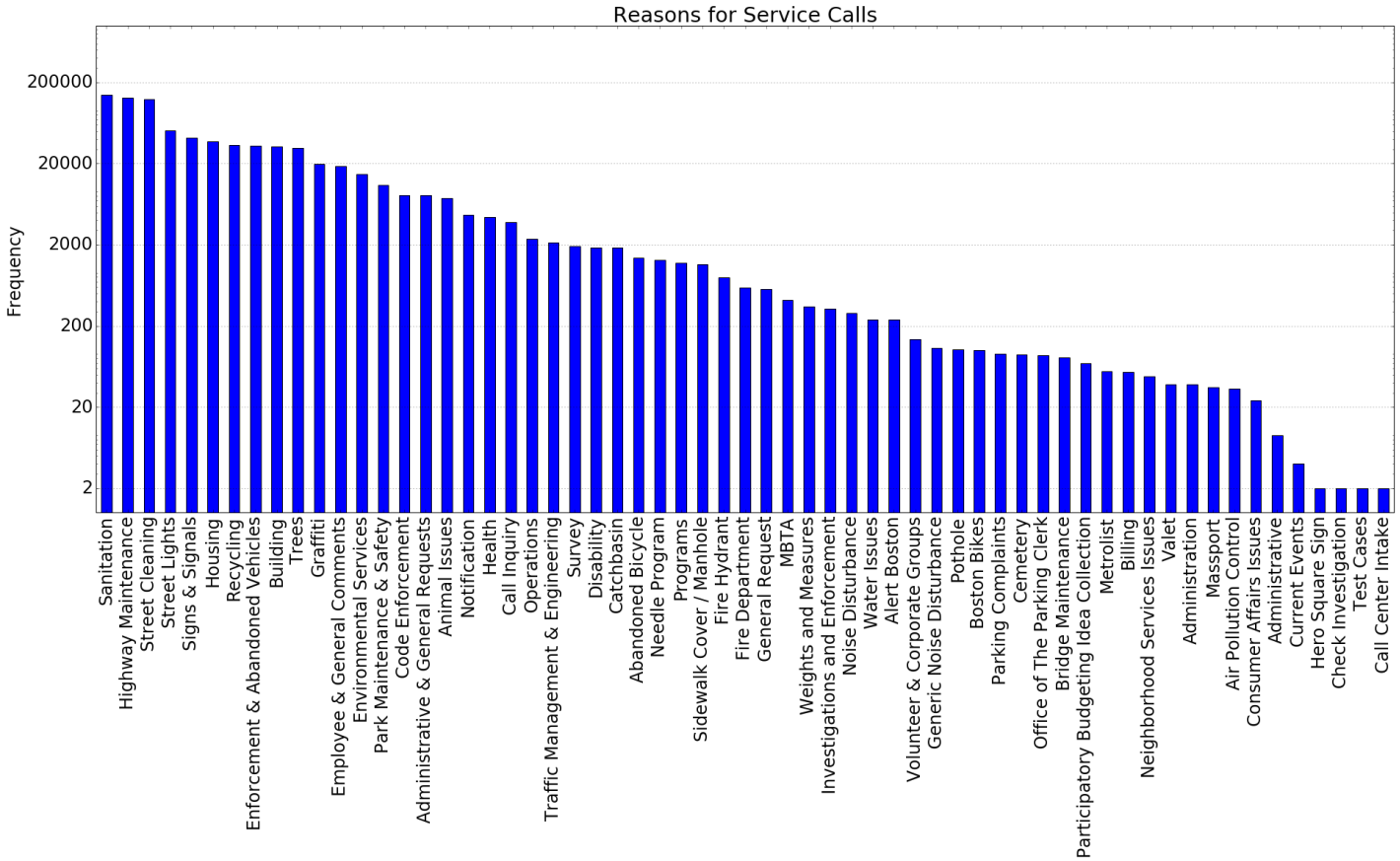
Boston residents are able to dial "311" and report public property issues such as rodent sightings, unsanitary conditions, streetlight outages, and so forth. Records of these 311 service requests, from July 2011 to present, are available from data.cityofboston.gov.

Each call record includes the following relevant information:

- The date the complaint was made
- Various descriptions of the nature of the complaint
- Latitude/longitude coordinates of the issue

Each complaint is described, often redundantly, by a "title," a "subject," a "reason," and a "type." In the dataset, there are 7837 different titles, 18 different subjects, 61 different reasons, and 215 different types. It is convenient to use "reasons" as a natural way to categorize complaints, since they strike a balance between being overly specific (as in the thousands of different "titles") and not being descriptive enough (like the handful of vague "subjects").

The following graph shows the prevalence of each of the various service call "reasons." Note that there is a swift decline in the frequencies of the least common reasons, with a handful of the very least common reasons only appearing in the dataset a few times. Because of this, in the work that follows we will disregard these extremely rare service reasons.



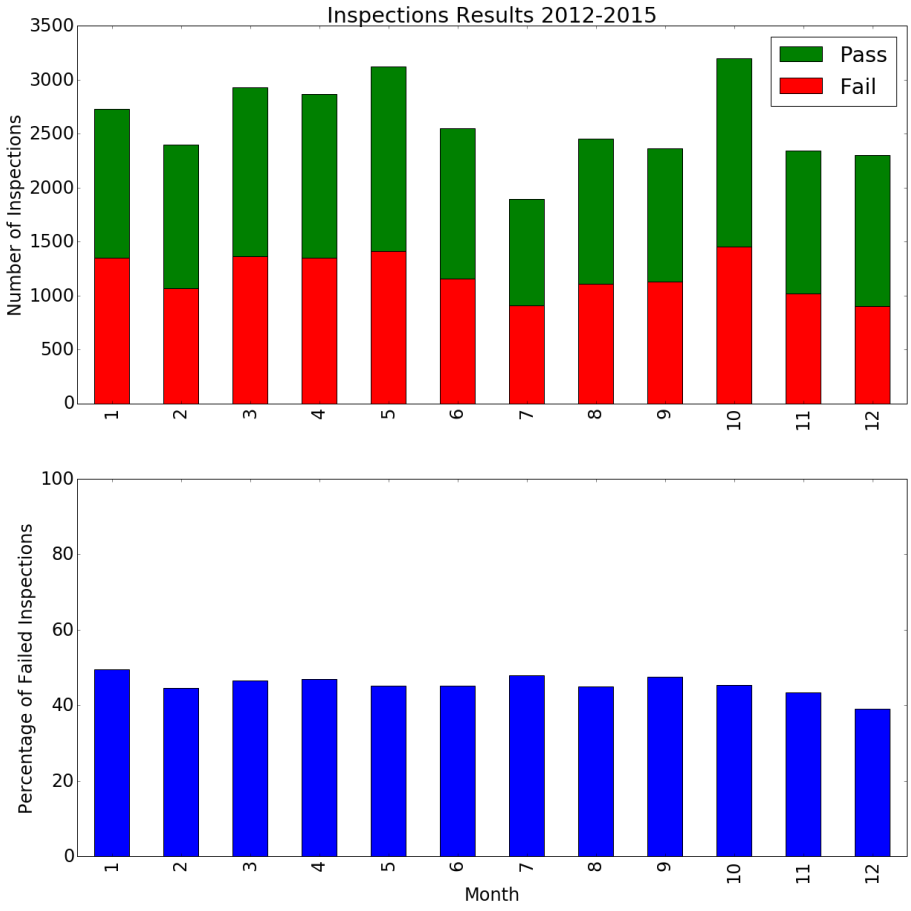
This dataset provides a wealth of details about reported environmental conditions near restaurant inspections. From these details we derive a model for predicting the outcomes of those inspections, later in this report.

DATA ANALYSIS

PRELIMINARY EXPLORATION

Since service call data is only available from July 2011 onward, we will only consider inspection results from August 2011 onward. In Yelp's dataset, 1634 different businesses were inspected in total, and 1327 of them experienced a failed inspection at some point. Meanwhile, in the more complete raw inspection data from the City of Boston, 5604 businesses were inspected, with 3876 of them experiencing at least one failed inspection. For this first exploration, we will focus on the raw City of Boston dataset, since it can provide us with some interesting insights not available from the more condensed and selective Yelp dataset.

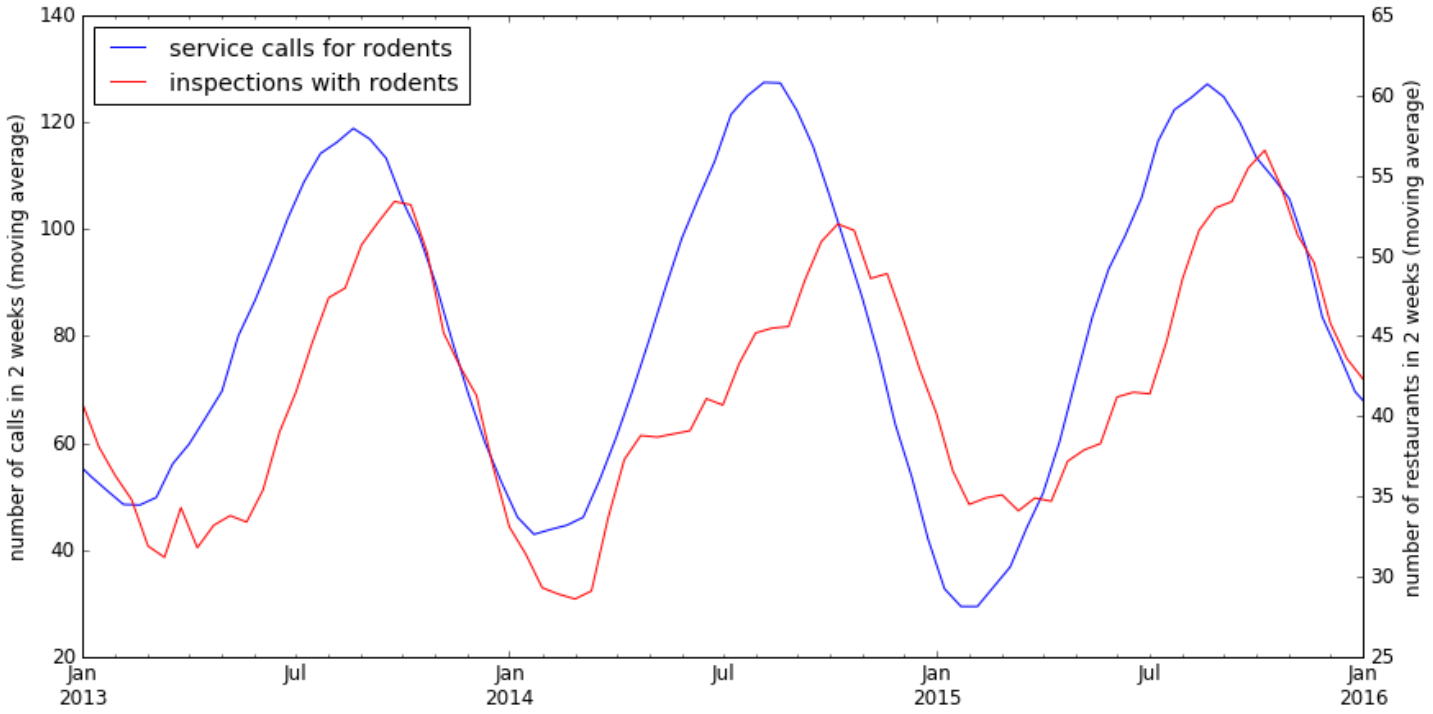
Looking at the years 2012 through 2015, the number of inspections performed varies a great deal from month to month, but the *percentage of inspections that fail* is consistently between about 40 and 50 percent*.



* It should be noted that in the more focused dataset provided by Yelp, the failure rate is about 64%.

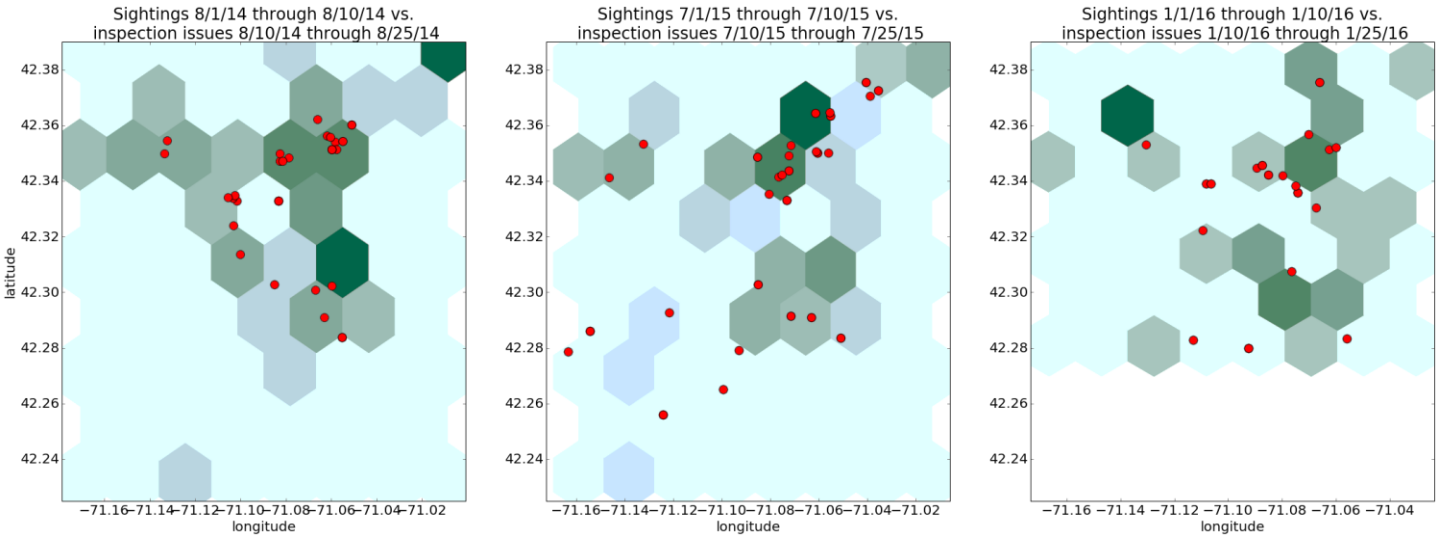
Our overall goal is to be able to predict the outcome (pass or fail) of a given health inspection, regardless of the specific underlying causes for a failure. However, our predictions are built upon environmental conditions near each inspected business, so it is worthwhile to consider some possible relationships between specific types of environmental issues and the reported causes of inspection failures. For instance, one aspect of food safety that often inspires public interest is the presence of rodents in food service establishments.

The dataset reveals that over 20% of inspected businesses have experienced inspection failures related to rodents. Inspections with rodent-related issues are an interesting subset of all inspections because we can investigate, relatively easily, whether citizens' reports of rodent sightings (through 311 service calls) are good predictors of subsequent rodent-related issues during health inspections of nearby food service establishments. Intriguingly, based on service call data, rodents are reported via 311 most commonly around June to August, and are detected in health inspections most commonly shortly thereafter, in July to September.



The pattern above shows an unmistakable correlation between the *timing* of rodent sightings and rodent-related inspection issues. With more difficulty, we can use visuals to explore whether there might be a correlation between the *locations* of sightings and inspection issues. For the graphs below, three different 10-day periods are selected (more or less at random) and service calls involving rodent activity during those periods are mapped. This reported rodent activity is indicated in green, with darker green corresponding to more activity. Then, rodent-related health inspections during an immediately subsequent

15-day period are shown in red. In each case, the distributions of sightings and rodent-related inspection results do appear to be roughly similar, although this conclusion is admittedly subjective.



Such patterns in both time and space lend some weak support to the idea that a restaurant that is near increased reported rodent activity might be at an increased risk of rodent-related issues during a health inspection. Soon, we will see firm statistical evidence that this is indeed the case.

Again, rodent-related issues are just one particularly interesting subset of public health challenges, serving here as a microcosm of the possible predictive relationship between service call data and inspection outcomes. Extending this beyond rodents, by combining other aspects of the available service call data, we construct a model for predicting the timing and location of a much broader variety of health inspection failures.

Extracting Features

In attempting to predict the outcome of a given restaurant inspection, we primarily look to conditions in the city near that restaurant in the recent past, as reflected in service calls. For each of the service call "reasons" discussed before, we count the number of times that particular issue has been reported near the inspected business, between 5 and 15 days before the inspection. Nearness is defined by the distance in latitude/longitude coordinates; more specifically, we count any reports that have occurred within about 3 miles of the inspected business. These choices for the time window and distance are arbitrary, to some extent. Systematic experimentation with other choices reveals that the accuracy of our predictive models is not very sensitive to these values, within reason.

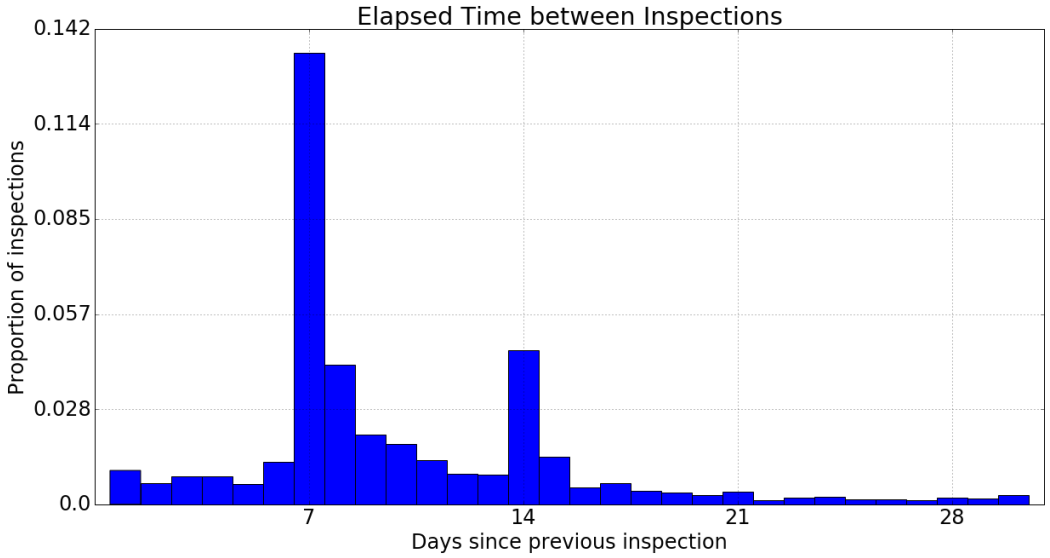
Because the features we will use are mostly derived from location information, for the remainder of this report, we will use only the condensed inspection dataset provided by Yelp, which provides latitude and longitude coordinates for every inspection. Another somewhat less important reason for using this dataset will also be seen later.

In addition to the features explained above, for each business inspected we also measure the length of time that has passed since its previous inspection, along with the results of the previous inspection, in the form of 1-star, 2-star, and 3-star violation counts.

Below, for demonstration purposes, we display the extracted information for some of the inspections that occurred on July 22, 2014. With any one of these, we can illustrate the meaning of the features described above. For example, an inspection occurred at Penguin Pizza that day. This inspection was a failure, since some violations were found. Between 5 and 15 days prior to this inspection, there were 699 complaints related to "sanitation" in the area around Pengin Pizza, 455 complaints related to "highway maintenance," 180 related to "street cleaning," and so on. It had been 208 days since Penguin Pizza's last inspection. (For the sake of space, the results of the previous inspection are not shown here.)

violations			name	Sanitation	Highway Maintenance	Street Cleaning	...	Valet	Administration	Massport	Air Pollution Control	delay
*	**	***										
13	0	0	My Thai Vegan Cafe	587.0	382.0	137.0	...	1.0	0.0	0.0	0.0	204.0
8	0	1	Penguin Pizza	699.0	455.0	180.0	...	1.0	0.0	0.0	0.0	208.0
0	0	0	Rebecca's Cafe	521.0	318.0	114.0	...	1.0	0.0	0.0	0.0	194.0
22	2	12	New Saigon Sandwich	587.0	384.0	137.0	...	1.0	0.0	0.0	0.0	186.0
6	2	2	Anh Hong	669.0	359.0	210.0	...	0.0	0.0	0.0	0.0	166.0
3	0	0	Starbucks	628.0	408.0	148.0	...	1.0	0.0	0.0	0.0	148.0
9	2	1	Al Dente Restaurant	444.0	304.0	101.0	...	1.0	0.0	0.0	0.0	140.0
4	0	0	Chipotle Mexican Grill	656.0	415.0	133.0	...	1.0	0.0	0.0	0.0	273.0
2	1	0	Boloco	519.0	320.0	114.0	...	1.0	0.0	0.0	0.0	141.0
7	2	2	Benevento's	444.0	304.0	101.0	...	1.0	0.0	0.0	0.0	140.0

As mentioned before, there are systematic mistakes in Yelp’s dataset, with a kind of double-entry of health violations happening frequently throughout the dataset. While most inspections repeat after a period of about one year, we should be suspicious of any inspection results that supposedly took place within a few weeks of a previous inspection at the same business. With the “delay” feature of the data extracted (as seen above), we can take a closer look at such results. The graph below shows that quickly repeated inspections are overwhelmingly repeated after either exactly one week or exactly two weeks. Because of this, we will make one final correction to the data by dropping any inspection with a calculated delay of 14 or fewer days.



SOME INSIGHTS FROM INFERENTIAL STATISTICS

Once we have extracted a set of environmental features as explained above, we can attempt to construct a predictive model using modern machine learning techniques. But before moving to that stage, let's briefly investigate the relationship (if any) between our features and the outcomes of health inspections using elementary statistical methods.

Since the question we wish to answer – will a given establishment fail an unannounced health inspection or not? – is binary in nature, we will test the influence that our environmental features have on the proportion of failed inspections. For each kind of common issue (or "reason") that may be reported through service calls, we can use a permutation approach to test hypotheses of the following general form:

> **Null hypothesis:** Establishments where the issue has been reported frequently, recently, and nearby will fail health inspections at the *same* rate as establishments where the issue has not been reported frequently, recently, and nearby.
>
> **Alternative hypothesis:** Establishments where the issue has been reported frequently, recently, and nearby will fail health inspections at a *different* rate than establishments where the issue has not been reported frequently, recently, and nearby.

For example, we can test whether businesses that are near the sites of more than a few recent complaints related to "environmental services" are more likely than other businesses to fail a health inspection. Note that "environmental services" complaints largely consist of rodent sightings, and most inspected businesses are near at least 10 such recent complaints. Using the features we have extracted, a hypothesis test indicates a statistically significant difference in the inspection failure rates of businesses that are near at least 10 recent "environmental services" complaints, compared to other businesses (p-value less than 0.001 from a permutation test). In our dataset, these establishments fail inspections at a rate about 20% higher than others, which seems to be useful information in predicting inspection failures!

As another example, there is a significant difference in failure rates between businesses that are near very few (less than 3) recent "health" complaints and other businesses. The p-value from the permutation test is less than 0.001, and in the dataset, businesses that are near at least 3 recent "health" complaints have a

failure rate about 26% higher than others. Note that complaints with the reason "health" are mostly reports of "unsanitary conditions" in food establishments.

Similar results can be found for many of the other environmental features, which suggests that these features will indeed be directly useful in developing a model to fulfill our stated purpose.

DEVELOPING A PREDICTIVE MODEL

OVERVIEW

We aim to devise a means for predicting the outcomes of future health inspections – will a given establishment fail an unannounced health inspection, or not? To this end, we have a matrix of 9641 data points and 60 features. Recall that of these features, 54 measure the recent and nearby occurrences of various kinds of city service complaints, while 4 of the features relate to the outcome of the restaurant's previous inspection, and 2 of the features are simply the latitude and longitude of the restaurant. We also have a vector of 9641 inspection outcomes, classifying each inspection as failed or passed.

To develop a model for classifying inspections as failing or passing, we will take the following general steps:

- Partition the data into a training set and a test set. We will randomly assign 80% of the data to the training set. The assignment is "stratified," meaning that the relative frequencies of failed and passed inspections are preserved in the training set.
- Use 3-fold cross-validation on the training set to search for hyperparameter values to maximize accuracy. More specifically, the training set is split into three parts. For each set of candidate hyperparameter values, each part of the training set in turn is held back while a model is created from the other two parts, then checked on the third. After all hyperparameter values are tested this way, the best-performing set of values are finally used on the full training date to create the final model.
- Use the previously untouched test set to check the model's accuracy, along with other more graphical measures of performance:

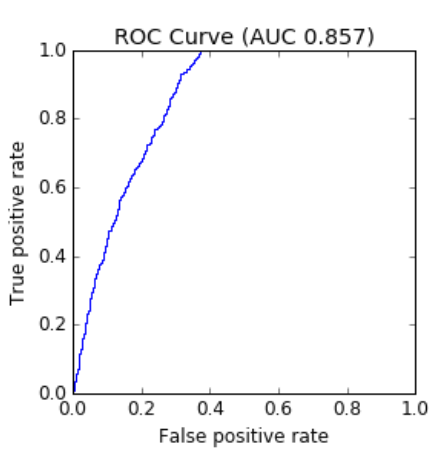
[DISCUSSION OF OTHER MEASURES WILL GO HERE]

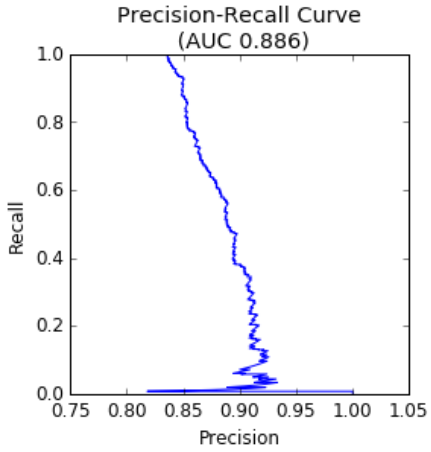
These steps have been applied to create a variety of predictive models for restaurant inspections, using the concepts of logistic regression, decision trees, support vector machines, and ensemble techniques. The performance of each type of model has been fairly comparable, but we will now detail two of the most successful.

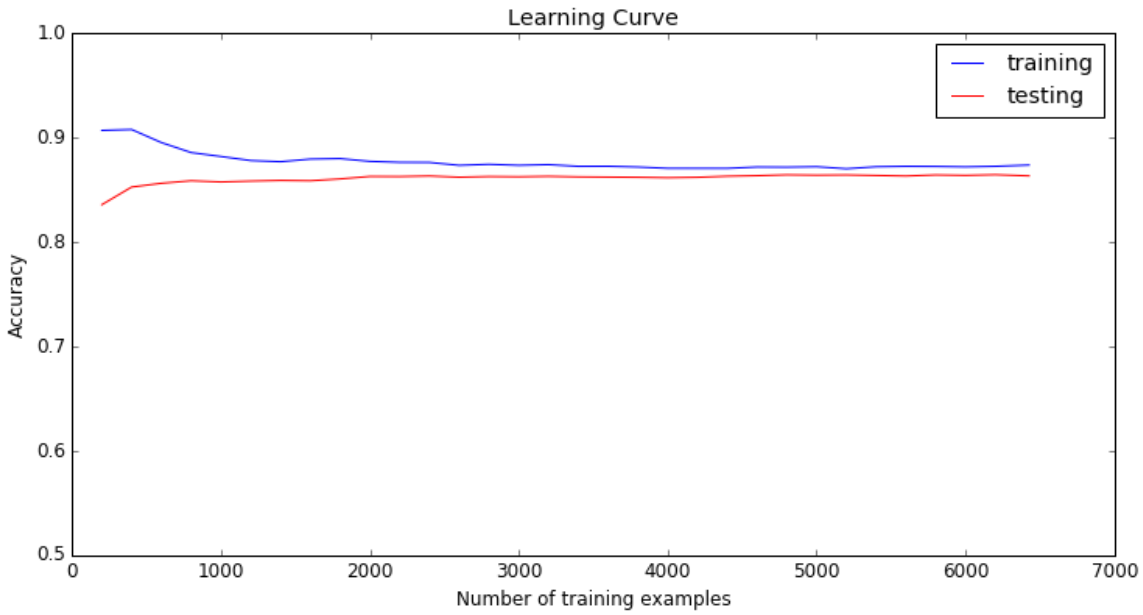
SUPPORT VECTOR MACHINES

Using the features and general steps described above, a support vector machine classifier can be constructed. We first standardize the feature values, then search for optimal values of the regularization hyperparameter, kernel type, and kernel coefficient. Once the model is created, its accuracy on our test set for this problem is typically 0.86 or 0.87.

Below we have the receiver operating characteristic (ROC) curve and precision-recall curve for a typical SVM classifier. The learning curve is also shown.





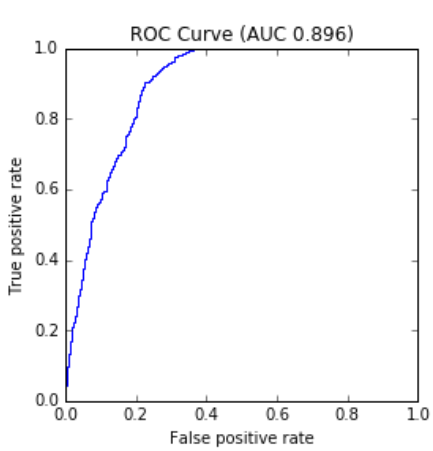


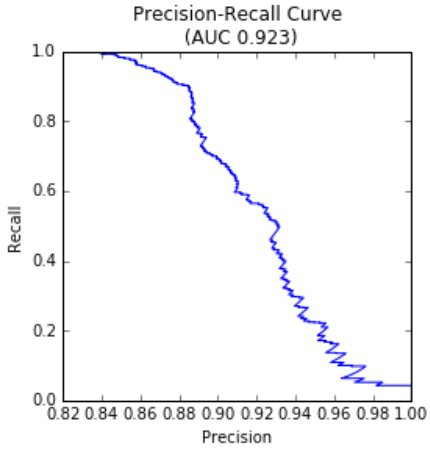
[MORE DISCUSSION HERE]

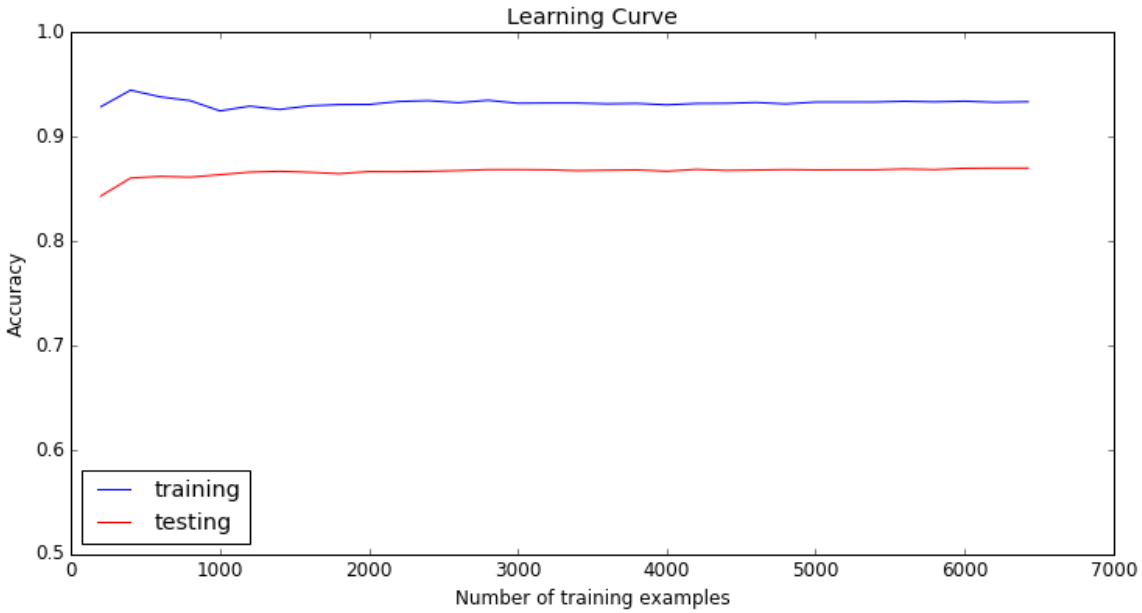
RANDOM FORESTS

Using the same features and general steps again, we construct a random forest classifier. We attempt to optimize for the following hyperparameters: the maximum depth of trees in the forest, the number of features to consider when splitting, the minimum number of samples allowed in a new leaf, and the minimum number of samples required when splitting a node. Once created, a random forest model typically achieves an accuracy of around 0.87 or 0.88 on our test set for this problem.

Below, we have the ROC curve, precision-recall curve, and learning curve for a typical random forest classifier.

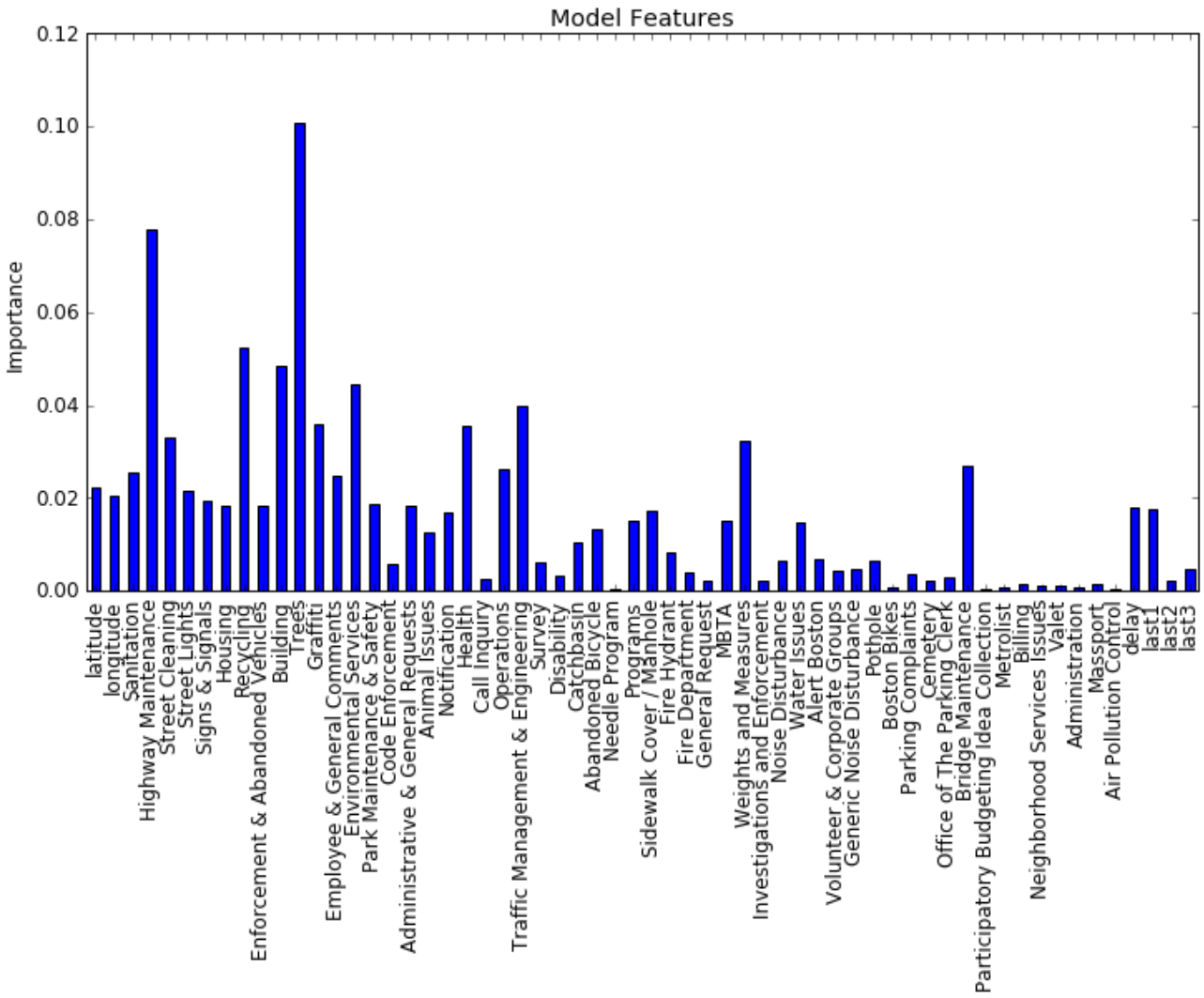






The use of a tree-based classifier provides a straightforward measure of the importance of the model's features. For the random forest with the results shown above, the most important features were the service call reasons of "Trees," "Highway Maintenance," "Recycling," "Building," and "Environmental Services." The features that are ranked as the most important in one random forest model may vary somewhat when a new forest is constructed – even if the same data and hyperparameters are used – but in the present situation, the set of most important features is fairly consistent from one forest to another.

In the graph below, we visualize the importance of all 60 features used, as revealed via the Random Forest model discussed above. Note that other than the first 2 features on the left, and the last 4 features on the right, the rest of the features are related to service requests. Note also that these features are listed in order of decreasing frequency (just as they were listed in a previous graph). Unsurprisingly, we can see that the more prevalent service call reasons tend to have higher importance than the less prevalent ones.



COMPARISON TO PREVIOUS CONTEST RESULTS

As mentioned before, the Yelp dataset that has been used here, primarily as a target for classifiers, was released as part of the "Keeping it Fresh: Predict Restaurant Inspections" contest at DrivenData.org in 2015. The goal of that contest was to be able to predict the number of one-star, two-star, and three-star violations found at each health inspection during a six-week period after the closing of contest submissions. So, in contrast to our present goal of simple classification, the DrivenData contest involved a problem of regression. Contestants' predictions were judged using the following weighted root mean square log error formula:

$$\text{WRMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[\log(y_i \cdot W + 1) - \log(\hat{y}_i \cdot W + 1)]^2}$$

In the notation used for this scoring function, there are N inspections, y_i represents the results (a vector of three integers) of the i^{th} inspection, W is the weighting vector $(1, 2, 5)$†, and $\hat{y}_i$ represents the predicted

† In plainer terms, contestants' predictions for the number of one-star, two-star, and three-star violations at each inspection were weighted so that the two-star prediction was worth twice as much as the one-star prediction, while the three-star prediction was worth five times as much.

results of the i^{th} inspection. The winning contestant's error score, as reported at DrivenData.org, was 0.8901.

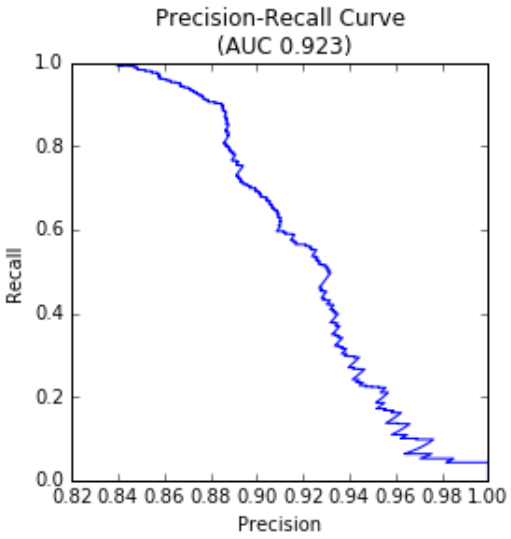
Contestants extracted features mainly from the text of customer reviews and other information provided through Yelp, in addition to details from past inspection outcomes. In the present project, we have taken a very different approach, using features related to reported environmental conditions near upcoming restaurant inspections. It would be interesting to compare the predictive ability of our features with those used in the contest. To make a somewhat fair comparison, using the same set of 60 features upon which we built our previous classification models, we take the following steps:

- Set aside the last 6 weeks of available data as a test set, with the rest of the data as a training set.
- Use cross-validation on the training set to search for hyperparameter values to minimize the WRMSLE score, and use those best values to create a regression model.
- Use the previously untouched test set to calculate the regression model's WRMSLE score.

A random forest regression model created in this way typically achieves a WRMSLE score of around 0.85 on the test data – a level of error quite comparable to the leading results of the DrivenData competition.

RECOMMENDATIONS

Given our successful development of an accurate predictor for health inspection failures, the City of Boston is advised to make use of such a predictor in planning upcoming inspections, in order to make more efficient use of inspectors' time.



While a randomly-selected restaurant has around a 64% chance of failing an unexpected inspection, a restaurant identified by our random forest predictor will have an 84% or higher chance of failure, as seen in the precision-recall curve reproduced again here. If desired, this precision can be increased, at the cost of decreasing recall. For example, if the City is willing to accept only 50% recall (i.e. only 50% of failed inspections are foreseen), then the precision of predictions could be increased to about 93%. This trade-off should be calibrated in consultation with the Inspectional Services Department of Boston.

The models developed in this project have been based on a limited record of past health inspection results, for the principal reason that the City of Boston's live records lack latitude and longitude information for many inspections. As such, our models can be seen as a proof of concept; a fully operational model for ongoing use would need incorporate the City's live records. Therefore the City is advised to require all health inspections to include location data. Alternatively, missing location data for past and future inspections may be generated rather inexpensively through the use of the Google Maps Geocoding API. The code included with this report already gives some indication as to how this would be done.

POSSIBLE DIRECTIONS FOR FUTURE STUDY

The environmental features we have extracted provide enough information for a successful classifier, but there is likely room for improvement. We have used the same time window and nearness threshold in the calculation of each of those features. Perhaps some variation in these choices from feature to feature would lead to a better-performing model, since some environmental factors might have influence at different distances, or on different timescales, than others.

When the City of Boston's more detailed inspection records are used, we may incorporate new features based on the fine details of past inspections, such as the exact reasons for past failures. After all, it seems plausible that some health issues are more prone to repeat offense than others.

Lastly, given the success of the DrivenData contestants in using customer reviews to predict inspection outcomes, our model might benefit from an incorporation of features derived from textual and other data provided by Yelp and Google restaurant reviews.