Capstone Project

Predicting Inspection Results at Food Establishments

Introduction

As a destination that attracts millions of tourists every year, Southern Nevada, and particularly Las Vegas, offers many possibilities when it comes to dining options. With food offerings that range from small locales to Michelin star restaurants, the choices may seem endless, and even though this may be a great experience for visitors, it can become a challenging task for the Southern Nevada Health District, who must inspect every food establishment to ensure it meets food and safety requirements to protect public's health.

Since these are unannounced inspections that happen at least once a year, it's vital to prioritize visits to restaurants that are most likely to have food and safety violations to avoid foodborne outbreaks and maximize the effectiveness of inspections.

Hence, the goal of this analysis is to analyze publicly available data from restaurant inspection results and Yelp's data on businesses and consumers' ratings, to predict non-compliant results and help optimize and prioritize inspection visits. The hypothesis is that inspection results differ from different types of food establishments based on the type, date, and time of the inspection, establishment's cuisine type, business ratings, and the number of review it has received.

Thus, the null hypothesis is that holding everything else constant, the probability that a restaurant receives a passing grade is the same for all restaurants, independent of the month the inspection takes place, the location, restaurant's attributes, their Yelp's stars ratings, business ratings, review counts, type of food, or review counts. The alternative hypothesis is that there is a statistically significant difference between the attributes of a restaurant that have passing grades and those with downgrades. to use publicly available data

Approach

Using publicly available data from restaurant inspection results from the Southern Nevada Health District and Yelp's data on businesses and consumers' ratings, the goal is to identify the factors that contribute to downgrades and non compliance inspections results, using the inspection results variable as the predictor variable. The idea is to help narrow the factors considered when scheduling food and safety inspections. This will be accomplished by identifying patterns between business attributes, consumer ratings, and historic data on inspection results.

The results from this analysis can provide actionable insights to local officials to decide when to adjust the rank and frequency of inspections based on the factors that are most likely to influence low compliance rates.

Open Data

Inspections Data

The Southern Nevada Health District Restaurant Inspections is a publicly available dataset that provides inspection results of food establishments. Each record represents the inspection results of a single establishment, with the name, location, date of the inspection, and inspection grade, among the following attributes:

Serial Number	State	Employee ID
Permit Number	Zip Code	Inspection Type
Restaurant Name	Current Demerits	Inspection Demerits
Location Name	Current Grade	Inspection Grade
Category Name	Date Current	Permit Status
Address	Inspection Date	Inspection Result
City	Inspection Time	Violations

Record Updated

Target variable:

Once an inspection is conducted, the establishment receives an inspection grade based on the number and type of violations. Each violation is assigned a value, called demerit, that results in a grade. ‘A’ grades are given to establishments that are compliant (0 – 10 demerits), while ‘B’ (11 – 20) and ‘C’ (21-40) are downgrades and indicate critical or major violations. Hence, the objective is to predict that probability of downgrades or non compliance results.

Yelp Data

The Yelp Academic Dataset includes two files, the businesses description and the customer reviews. Each file is a JSON object file per line file. The business description contains:

business_id	state	review_count *
name	*postal_code*	is_open
neighborhood	*latitude*	attributes *
address	*longitude*	categories *
city	*stars**	hours

The reviews file contains customers’ reviews including:

review_id	*stars*	useful
user_id	date	funny
business_id	text	cool

How to Improve Decision Making

After obtaining the restaurant inspections results and the Yelp data, the datasets were merged to provide greater depth to the analysis and create new attributes that could give additional information. The idea is to give health inspectors a tool to schedule visits more efficiently by focusing on establishments that have a higher probability to have critical violations.

Data Wrangling

The inspections results were obtained from the Nevada Health District Restaurant Inspections webpage using their API and the Yelp academic files were downloaded directly from Yelp. All the JSON files were normalized to manipulate them using Pandas and some initial processing was done before merging them.

For the **restaurant inspections results**, which had over 160,000 records and 24 attributes, a subset was taken to include:
Address, category_name, city, current_demeritis[1], current_grade, inspection_date, inspection_time, inspection_demerits, inspection_grade, inspection_result, violations, inspection_type, location_coordiantes, location_name, permit_status, restaurant_name and zip code.

The main data cleaning steps included:

- removing trailing and leading spaces from column names and values in the category name and restaurant name attributes.
- removing special characters from the restaurant name values to merge with the Yelp data.
- splitting zip codes to leave only the first five digits (e.g. 89103-4004 became 89103)
- splitting the date to create a year and a month column to use as keys during merging and create a new season column
- creating a binary target variable called 'is_compliant', with 1, if the inspections results were 'compliant', 'A grade', 'approved', or 'no further action' and 0 otherwise.

[1] As noted in previous report ('Predicting Critical Health and Safety Violations at Food Establishments – Proposal'), each violation is assigned a value, demerit, that results in a grade. 'A's are given to establishments that are compliant (0 – 10 demerits). 'B's (11 – 20) and 'C's (21-40) are downgrades indicating critical or major violations.

For the Yelp data, only two files were used, the business data and reviews data. For the **business data file**, which had over 180,000 records, the main data cleaning steps included:

- filtering the data to include only records for the state of Nevada
- selecting business id, name, location, review count, categories, attributes, and zip code.
- performing the same first four steps done to the inspection results dataset.

For the **Yelp reviews data file,** which had over 5 million records, the main cleaning steps included:

- selecting only the date, rating, and business id attributes. The review text was not included for this analysis.
- grouping the data by date and business id and get the daily rating average for each business before merging it with the business data. This was done to mitigate 'unnecessary' data duplication when merging the Yelp file with the inspections file, since the text was not considered.

After the pre-processing, the datasets were merged to include only values that were present in both data sets. Since the inspection results didn't include a business id that matched the business id from Yelp, the datasets had to be merged using a combination of restaurant name, zip code, year, and month. Using the year and month ensured the ratings and business attributes were still relevant.

Feature engineering

- Three new attributes were derived from the inspection date: season, day of the week and inspection shift.
- similar levels of the 'category_names' and the 'categories' attributes were combined to later create separate binary attributes to identify cuisine type

(Latin, American, Fast food, etc.) and type of establishment (food truck, restaurants, etc.)

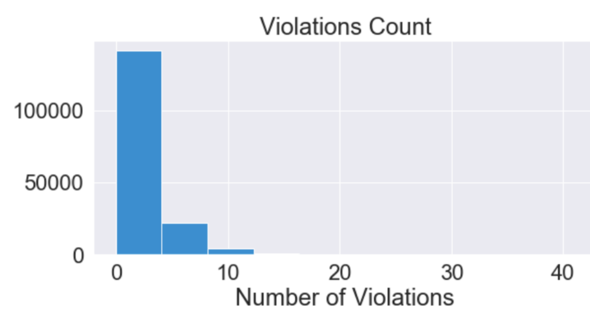
- A column with the number of days between inspection and customer star ratings.
- Used time weighting for star ratings that happened before the inspection date, with ratings closer to the inspection having higher weights.

EDA and Statistical Inference

Inspection Results Distribution

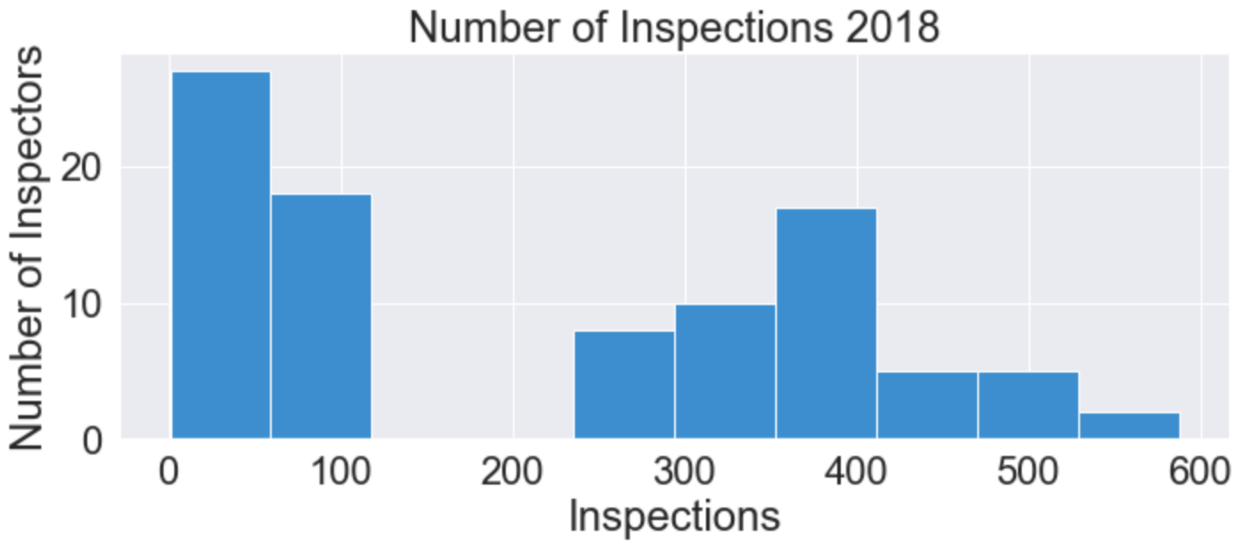
The first step was to look at the distribution of inspection results from the original data to compare the proportion of compliant vs. non-complaint establishments and see how well food establishments performed when it comes to safety and hygiene.

A visualization of the number of inspections by inspection results, shows that over 80% of inspections result in a passing grade with only less than 20% having downgrades or critical violations. In fact, a histogram of the total number of violations, shows that the most inspections have less than 10 violations.



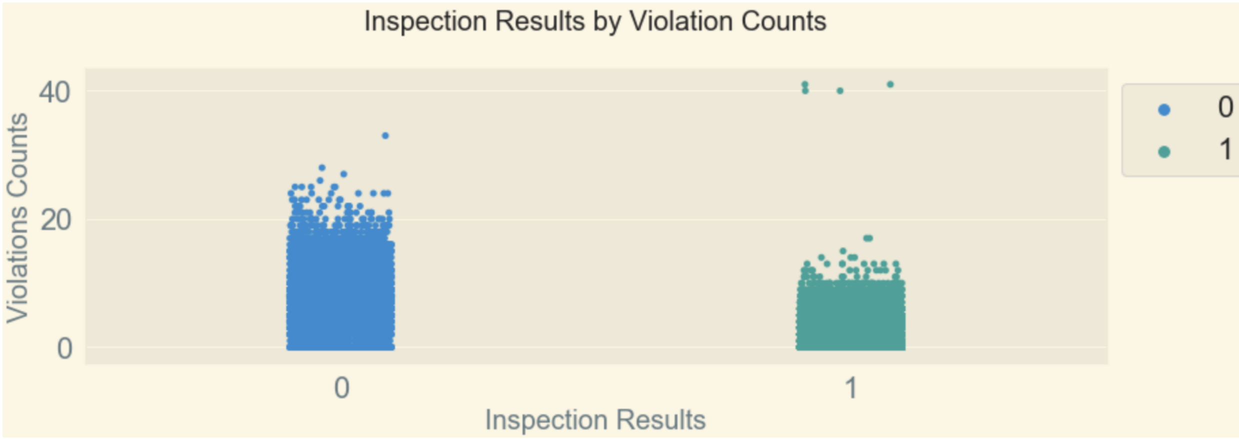
Inspectors and Inspections

Next, a close analysis to the number of visits done by employee ID, shows that in 2018, a few number of inspectors performed a large number of visits, with a single employee having over 550 recorded visits. This shows that the way schedules are being generated, few inspectors have to visit more establishments and perhaps spending less time at inspections due to the large workload.

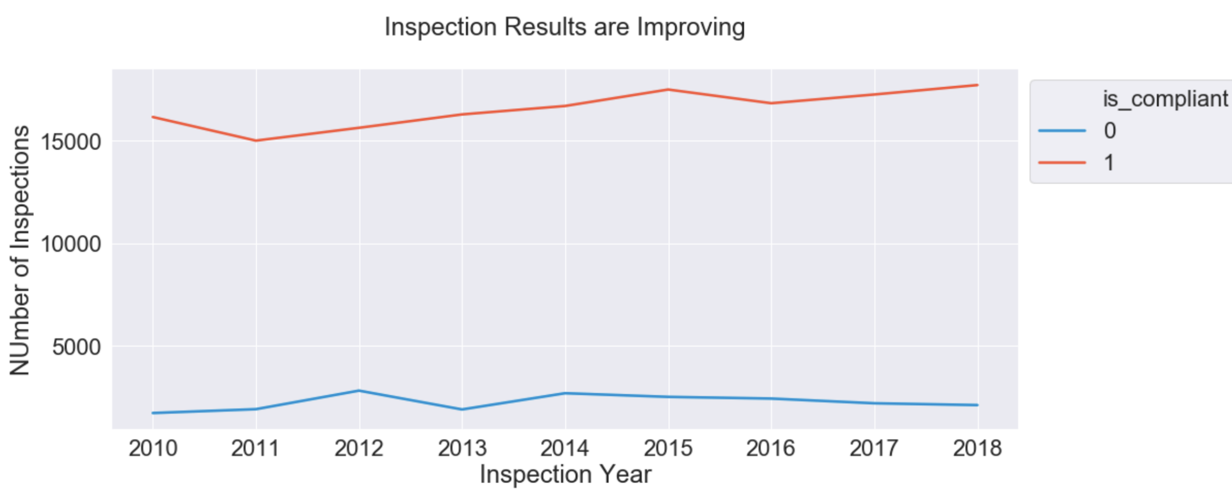


Inspection Results by Violation Counts

The Southern Nevada Health District manages 90 different violation codes for restaurant inspections with each violation corresponding to a value that results in a grade. ‘A’ grades are given to establishments that are compliant (0 – 10 demerits), while ‘B’ (11 – 20) and ‘C’ (21-40) are downgrades and indicate critical or major violations. An analysis of the violation counts by inspection shows that some restaurants with less than 17 violations can in fact received a non-compliance result. This shows that not all violations carry the same weight and even a restaurant with a single violation can be found non compliant, if the violation is an imminent public hazard. The same can be said for a few restaurants that had multiple violations, but received passing grades.



Looking at compliance rate overtime, it shows that In general, grades have improved over time. Since 2011 food establishments have constantly improved, with only a slight decline between 2015 and 2016 when grades were consolidated and inspectors started using A grades for compliance.



Compliance by Seasons

The general assumption for this project is that inspection results and passing rates are different across months, food categories, establishment's star ratings, business ratings, and review count.

Thus, one of the hypothesis is that seasonal changes in weather may affect food handling and preparation, increasing the likelihood of food violations. The assumption is that higher temperatures during the summer months may increase the growth of bacteria if food is not prepared or consumed promptly. A detailed analysis of inspection results by season, showed that although there are less inspections during the summer months, the number of non-compliance results is slightly higher compared to other months, particularly when compared to winter inspection results.

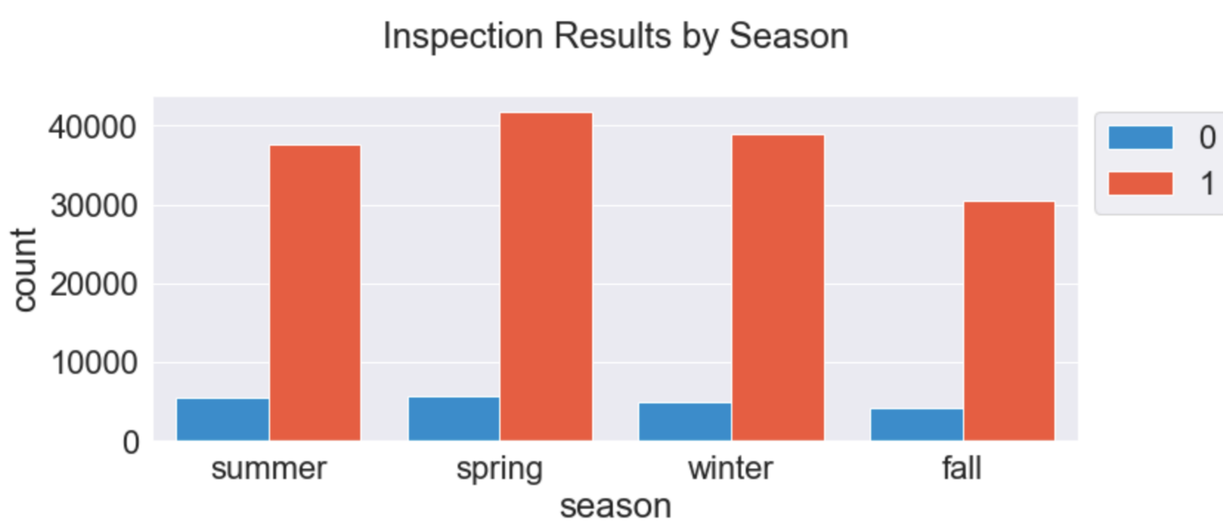
$H0$: Compliance rate in cooler months = Compliance rate in warmer months

HA: Compliance rate in warmer months ≠≠ compliance rate in cooler months

$\alpha = 0.05$

Based on all the inspection results, compliance in winter is slightly higher compared to other months with 88.7 %. Compliance during Spring is 87.93%, 87.89 % in the Fall and Summer has the lowest rate at 87.4%. Although compliance is 1.3 % lower during the Summer compared to winter.

Since the Summer months tend to be specially busy with a growing number of tourists visiting the city, being able to predict and narrow the search of possible critical health violations, can minimize the risks of foodborne illness outbreaks.

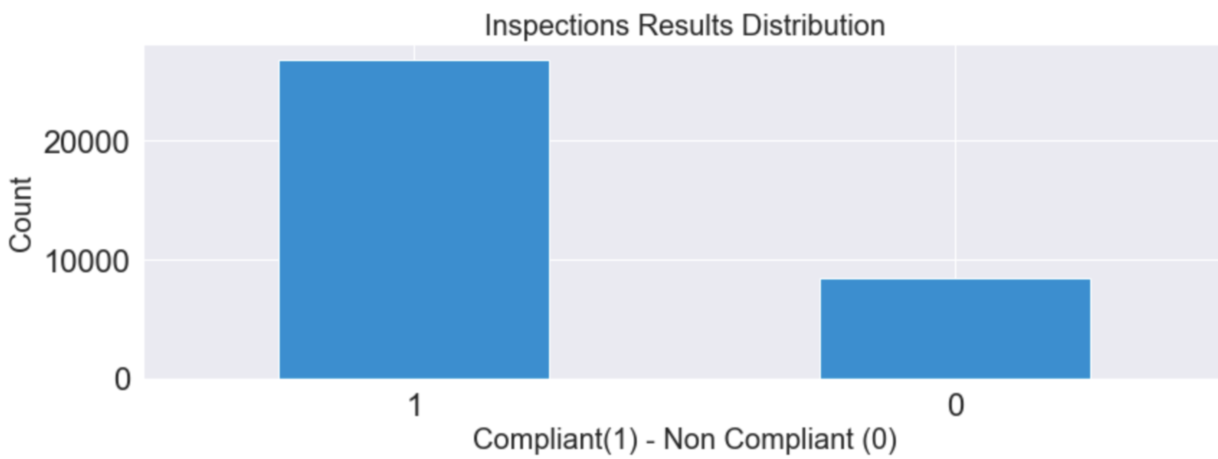


In order to see the likelihood of getting these inspection results again, the data for winter and summer were randomly sampled, ignoring the season 'labels' to see if there is a statistically significant difference between the compliance results obtained during each season. Using a t-test, the results show that after 10000 simulations, the difference is statistically significant at the 0.05 level, as none of the simulations reached a difference at least as high as the observed difference of 1.3%.

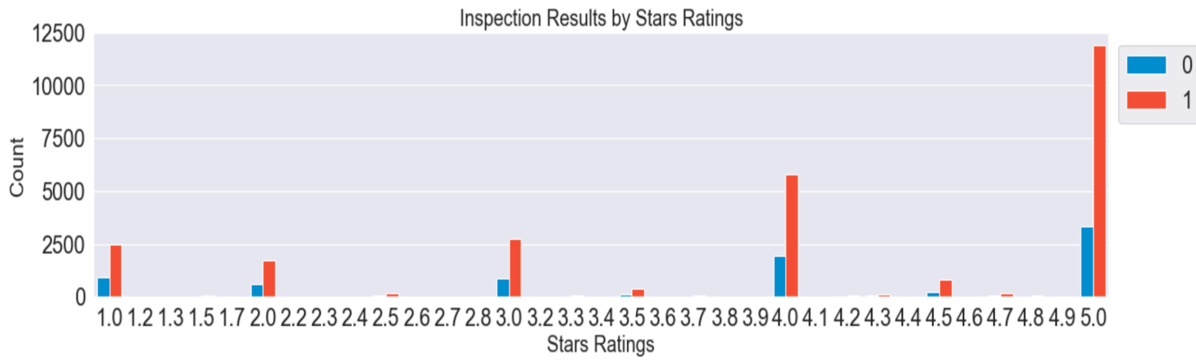
Merged Data

Distribution of Target Variable

Once the inspections and Yelp datasets were merged, the first step was to review the distribution of the target variable to see if it was unbalanced. The plot showed that the proportion of passing grades is 76%, which is a fairly balanced distribution.



A comparison of Yelp's stars rating and the inspection results, showed that the vast majority of restaurants with daily four and five average stars rating, have positive compliance results.



Feature Engineering Merged Data

Since the data were very sparse across business attributes, similar business and food type categories were grouped, rename, and changed into binary features. Thus, there were 41 attributes for food and business types.

Food Type and Compliance

Another general assumption was that the establishment type and the food type may impact inspection results. For instance, establishments with buffet options and/or raw food menus may have higher food violations due to stricter standards of temperature and food handling.

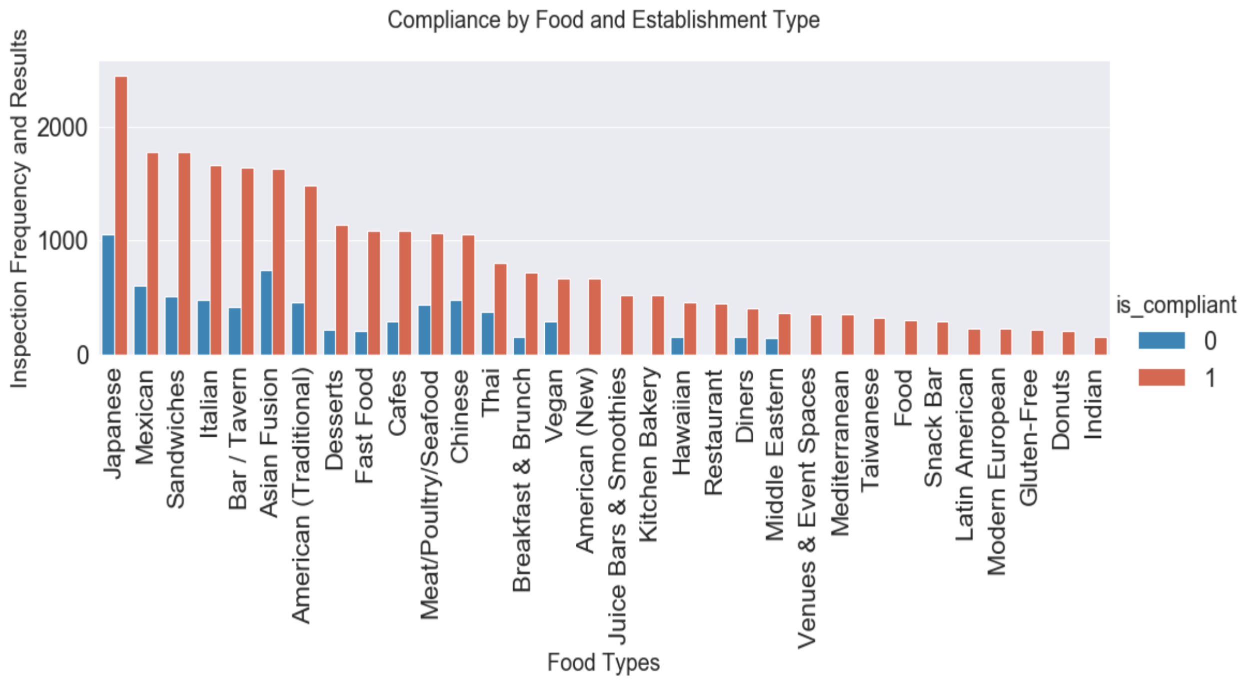
The hypotheses were:

H_0: There is no significant difference in compliance rates by food type.

H_A: compliance rate among establishments varies between food type.

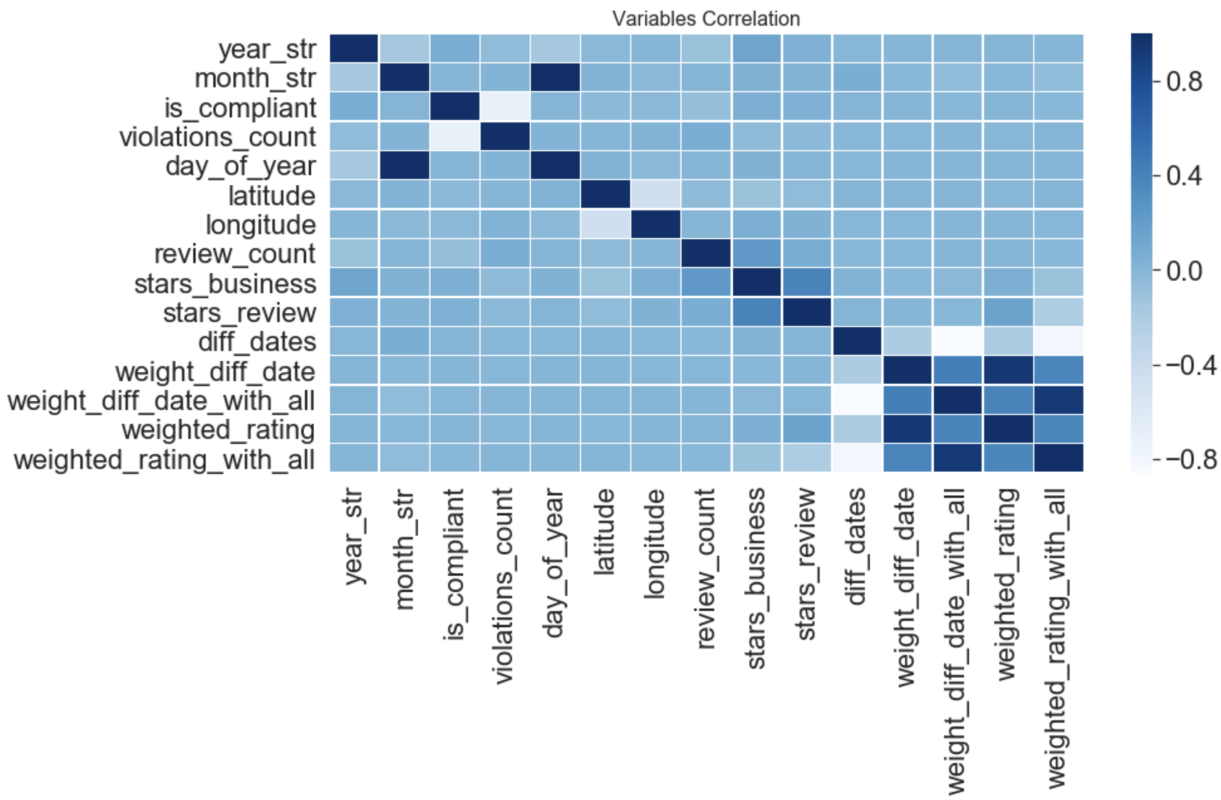
$$\alpha = 0.05$$

Looking at the distribution of compliance rates across different establishment types, Yelp data showed that establishments categorized as 'Japanese' had a higher number of 'non compliant' results followed by Asian Fusion, Mexican, Sandwiches, and Italian.



Correlations

A heatmap of the correlation coefficients showed that violation counts are highly correlated to the inspection results, which makes sense, since the inspection grades are directly based on the violation points received. Other variables that were highly correlated with each other were those derived from the same attribute, such as the weighted star ranking and the days difference between inspection results and star rating. These were removed since they were highly correlated and did not provide new information.



Machine Learning

The majority of the predictive stage was spent feature engineering and converting categorical variables into binary attributes since the data were very sparse.

One of the variables that was derived from the Yelp data was the weighted star rating. Since one establishments would have multiple daily ratings, the first step was to get the average daily rating. Then a new feature was created based on the inspection date and the rating date to find the time between rating and inspections and being able to give more weight to those ratings that happened shortly before an inspection. Hence, the same rating would have a different value based on the inspection date and review date.

$$Weight = 1/\sigma(star\ ratings)$$

$$Weighted\ difference = -1/(Weight * (inspection\ date - inspection\ review))$$

*Weighted rating = Weighted difference * star rating*

Modeling

The final dataset had 35242 records and 92 attributes. Since this is a classification problem with a binary target variable, two different classification algorithms, logistic regression and Random Forest models, were tried with different data splits and slightly different predictors to compare the predictive power of each model.

Features

The target variable, 'is_complaint', indicates whether an establishment received a passing grade or not.

Input variables:

- Binary variables for food type attributes. Whether the restaurant served Japanese, Asian, Mexican, Mediterranean, among others.
- Business review counts, weighted star rating, business rating.
- Binary variables for business attributes. Food truck, restaurant, venue, or other.
- Binary variables for month of inspection (Logistic Regression and one Random Forest)
- Binary variables for seasons
- Binary variables for weekdays
- Binary variable for inspection shift (morning and afternoon)
- Binary variables for establishment's grade at the time of inspection
- Demerits at the time of the inspections
- Type of inspections (Routine, re-inspection, survey)

Models

Logistic Regression

Two logistic regression models were trained after splitting the data into training and testing sets with 75% of the data to fit the model and 25% to test the model's accuracy on unseen data.

Using 76 attributes (see Appendix) to train the model, the accuracy of the model to predict inspection results on the test set was 76%. However, the precision and recall for non-compliant results was zero. The model was not able to accurately identify any of the non-compliant results.

The following classification report provides the results of the logistic model.

Classification report:				
	precision	recall	f1-score	support
0	0.00	0.00	0.00	2125
1	0.76	1	0.86	6686
micro avg	0.76	0.76	0.76	8811
macro avg	0.38	0.5	0.43	8811
weighted avg	0.58	0.76	0.65	8811

Logistic Regression 2

A second logistic regression model was tried setting the class_weight parameter to 'balanced', to see if the model will improve. Based on the LogisticRegression documentation:
"The "balanced" mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as

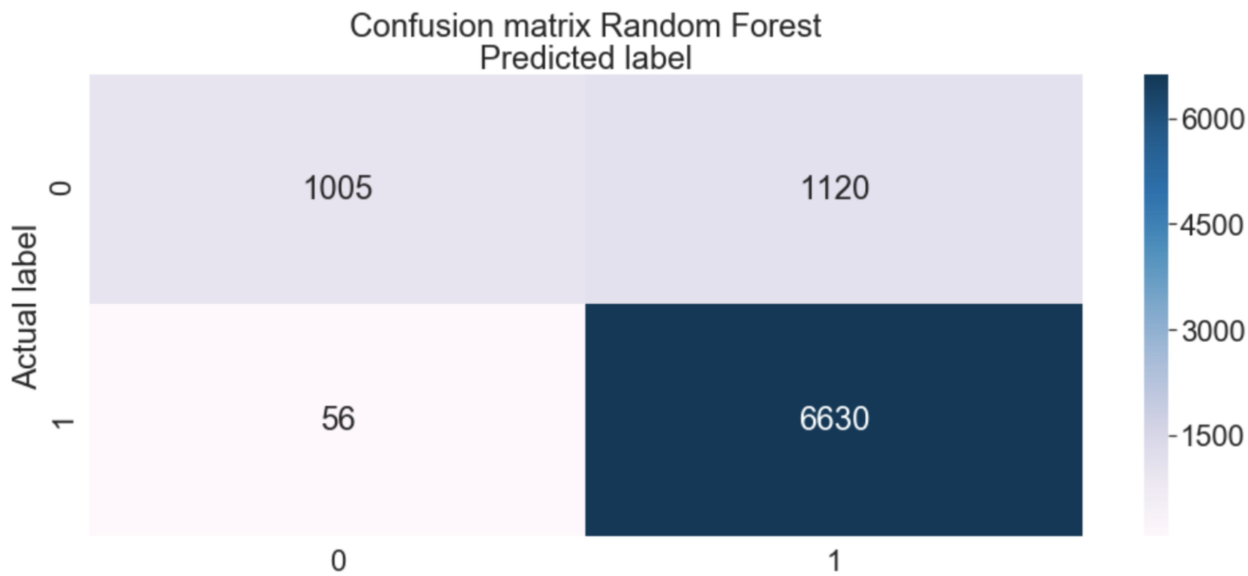
n_samples / (n_classes * np.bincount(y)).”

The overall performance of the model decreased to just 58.2% and although it was able to accurately identify 81% of non-compliance results, it did not perform as well with passing grades, identifying only 51%.

Accuracy score Linear Regression Model:0.582				
Best parameter: {'C': 0.01}				
Classification Report Logistic Regression				
	precision	recall	f1-score	support
0	0.35	0.81	0.48	2125
1	0.9	0.51	0.65	6686
micro avg	0.58	0.58	0.58	8811
macro avg	0.62	0.66	0.57	8811
weighted avg	0.76	0.58	0.61	8811

Random Forest (min. Sample split = 50)

A random forest was run using the same random split for the training and the test set used for the logistic regression model, 75% - 25% and the same input variables. The minimum number of samples for each decision tree split was set to 50 and the number of decision trees classifiers was set to 200. The accuracy of the model on unseen data was 86.65%, however the recall for non-compliant results was only 47%. This means that the model was able to predict only 47% of inspection resulting on non-compliance.



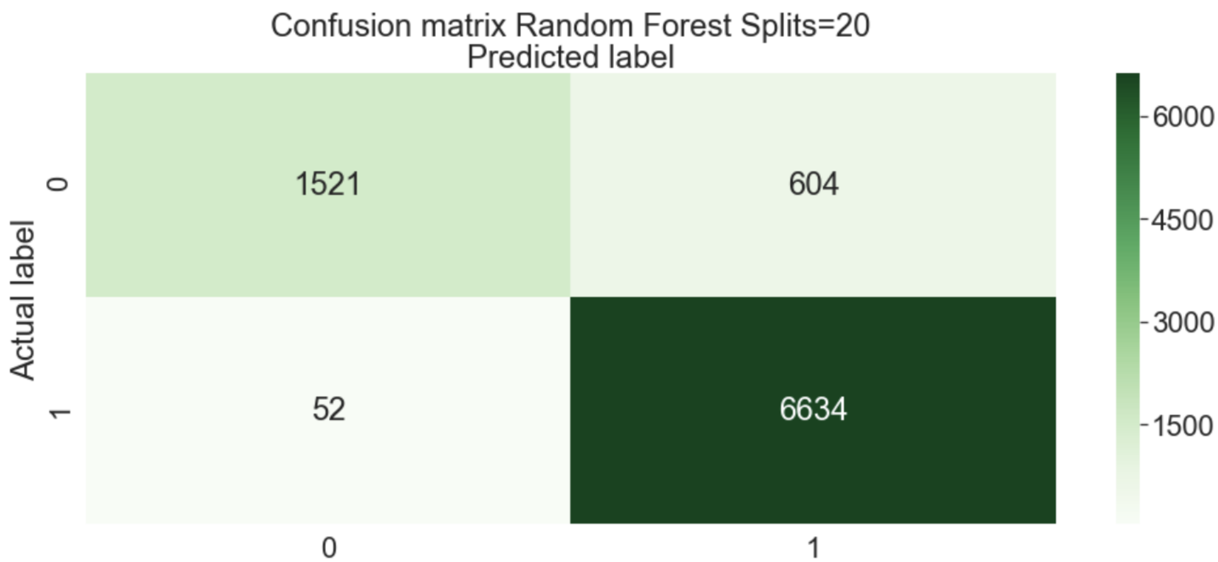
Accuracy score Random Forest (splits=50): 0.8665304732720464

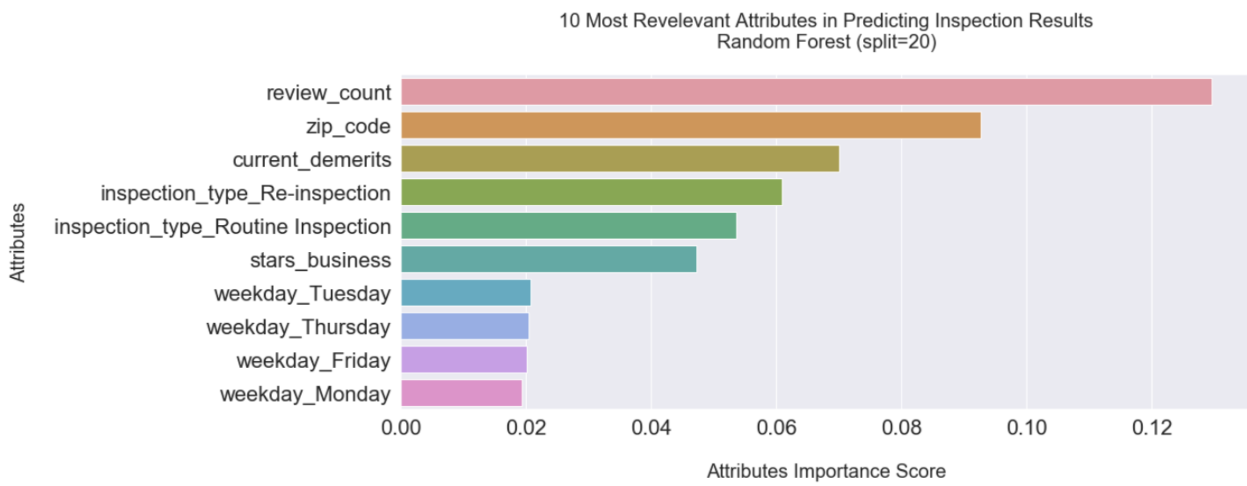
Classification Report Random Forest				
	precision	recall	f1-score	support
0	0.95	0.47	0.63	2125
1	0.86	0.99	0.92	6686
micro avg	0.87	0.87	0.87	8811
macro avg	0.9	0.73	0.77	8811
weighted avg	0.88	0.87	0.85	8811

Random Forest (min. Sample split = 20)

Since the random forest model performed better than the logistic regression model, one more random forest was run, using 80 input variables (see Appendix), 80% of the data to train the model, and 100 decision trees classifiers. Using these parameters, the predictive accuracy of the model went up to 92.5% on unseen data with a 0.72 recall and 0.97 precision on non-compliance results. Thus, this model was able to accurately predict 72% of non-compliance results.

The following graphs show the classification matrix and the 10 most relevant features for predicting inspection results.





Classification Report Random Forest				
	precision	recall	f1-score	support
0	0.97	0.72	0.82	2125
1	0.92	0.99	0.95	6686
micro avg	0.93	0.93	0.93	8811
macro avg	0.94	0.85	0.89	8811
weighted avg	0.93	0.93	0.92	8811

Conclusions

Using publicly available data on food inspection results and information on business features and reviews from Yelp, four predictive models were trained to identify the attributes that can help predict food inspection results.

Although all models had over 70% overall predictive accuracy, the latest random forest provided a robust model able to predict 92.5% of inspection results. Using a variety of explanatory variables, including the business review count, business location (zip code) and the number of demerits at the time of the inspection (demerits received in previous inspections), the model accurately identify 72% of non-compliance results.

These results can be used to help narrow and prioritize inspections to food establishments by comparing current scheduled visits with the parameters that are most likely to result in a non-compliance result.

Further Analysis

Although most of the work was done during the feature engineering, there is room for improvement in tuning and combining features to find even more relevant variables that help explain inspection results.

Additional features might also improve the predictive power of the model, such as date of last inspection or how long the establishment has been opened, as these can show patterns of compliance or suboptimal scheduling.

Appendix

Features used for Models

Features / **Models**	**Logistic Regressions** *Test size=0.25*	**Random Forest** (splits = 50) *Test size=0.25*	**Random Forest** (splits=20) *Test size=0.25*
is_compliant'	TARGET	TARGET	TARGET
current_demerits'	✓	✓	✓
year_str'			
zip_code'	✓	✓	✓
inspection_demerits'			
violations_count'			
day_of_year'			
latitude'			
longitude'			
review_count'	✓	✓	✓
stars_business'	✓	✓	✓
stars_review'			
diff_dates'			
weight_diff_date'			
weight_diff_date_with_all'			
weighted_rating'			
weighted_rating_with_all'			
inspection_type_Epidemiological Investigation'	✓	✓	✓
inspection_type_Re-inspection'	✓	✓	✓

inspection_type_Routine Inspection'	✓	✓	✓
inspection_type_Survey'	✓	✓	✓
food_type_American (New)'	✓	✓	✓
food_type_American (Traditional)'	✓	✓	✓
food_type_Asian Fusion'	✓	✓	✓
food_type_Bar / Tavern'	✓	✓	✓
food_type_Breakfast & Brunch'	✓	✓	✓
food_type_Buffet'	✓	✓	✓
food_type_Cafes'	✓	✓	✓
food_type_Caterer'			✓
food_type_Chinese'	✓	✓	✓
food_type_Delis'	✓	✓	✓
food_type_Desserts'	✓	✓	✓
food_type_Diners'	✓	✓	✓
food_type_Donuts'	✓	✓	✓
food_type_Ethnic Food'	✓	✓	✓
food_type_Fast Food'	✓	✓	✓
food_type_Food'	✓	✓	✓
food_type_Food Trucks / Mobile Vendor'	✓	✓	✓
food_type_Gluten-Free'	✓	✓	✓

food_type_Hawaiian'	✓	✓	✓
food_type_Health Markets'	✓	✓	✓
food_type_Indian'	✓	✓	✓
food_type_Italian'	✓	✓	✓
food_type_Japanese'	✓	✓	✓
food_type_Juice Bars & Smoothies'	✓	✓	✓
food_type_Kitchen Bakery'	✓	✓	✓
food_type_Latin American'	✓	✓	✓
food_type_Meat/Poultry/Seafood	✓	✓	✓
food_type_Mediterranean'	✓	✓	✓
food_type_Mexican'	✓	✓	✓
food_type_Middle Eastern'	✓	✓	✓
food_type_Modern European'	✓	✓	✓
food_type_Public Services & Government'			
food_type_Restaurant'	✓	✓	✓
food_type_Sandwiches'	✓	✓	✓
food_type_Snack Bar'	✓	✓	✓
food_type_Spanish'	✓	✓	✓
food_type_Special Kitchen'	✓	✓	✓
food_type_Steakhouses'	✓	✓	✓

food_type_Taiwanese'	✓	✓	✓
food_type_Thai'	✓	✓	✓
food_type_Vegan'	✓	✓	✓
food_type_Venues & Event Spaces'	✓	✓	✓
season_fall'			✓
season_spring'			✓
season_summer'			✓
season_winter'			✓
month_str_1'	✓	✓	✓
month_str_2'	✓	✓	✓
month_str_3'	✓	✓	✓
month_str_4'	✓	✓	✓
month_str_5'	✓	✓	✓
month_str_6'	✓	✓	✓
month_str_7'	✓	✓	✓
month_str_8'	✓	✓	✓
month_str_9'	✓	✓	✓
month_str_10'	✓	✓	✓
month_str_11'	✓	✓	✓
month_str_12'	✓	✓	✓

weekday_Friday'	✓	✓	✓
weekday_Monday'	✓	✓	✓
weekday_Saturday'	✓	✓	✓
weekday_Sunday'	✓	✓	✓
weekday_Thursday'	✓	✓	✓
weekday_Tuesday'	✓	✓	✓
weekday_Wednesday'	✓	✓	✓
inspection_shift_afternoon'	✓	✓	✓
inspection_shift_morning'	✓	✓	✓
current_grade_A'	✓	✓	✓
current_grade_B'	✓	✓	✓
current_grade_C'	✓	✓	✓
current_grade_O'	✓	✓	✓
current_grade_X'	✓	✓	✓