Predicting Restaurant Inspections

Duc Vu, Metis Chicago, Summer 2019

Chicago is home to thousands of restaurants.

Can the inspectors use the historical data on the past inspections and reviews that citizens generate to get a better view of active risks to public health?

Client - City of Chicago

- Narrow the search for health code violations

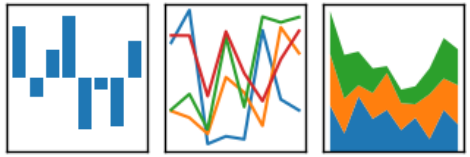
Objective

- Improve the City’s inspection efforts
- Which features are predictive?
- Create the apps to test the individual restaurant for the potential risk

Tools

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



plotly

+ableau

amazon web services™ | EC2

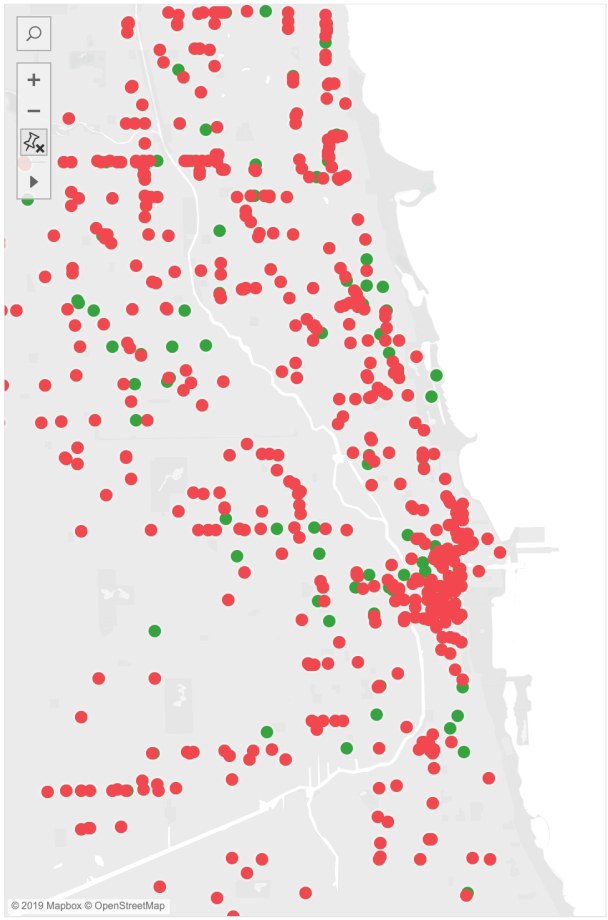
Flask

scikit learn

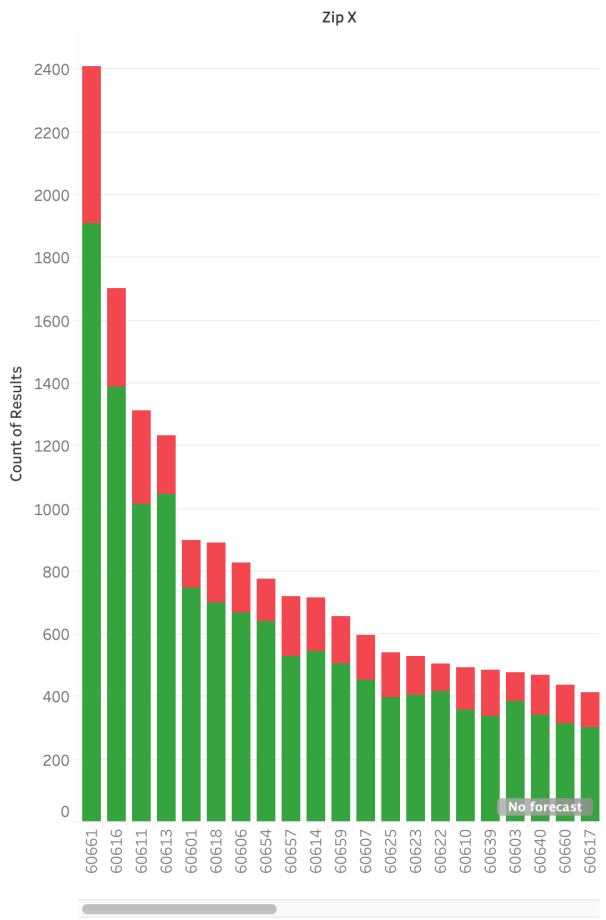
Data source

Open data from City of Chicago

map



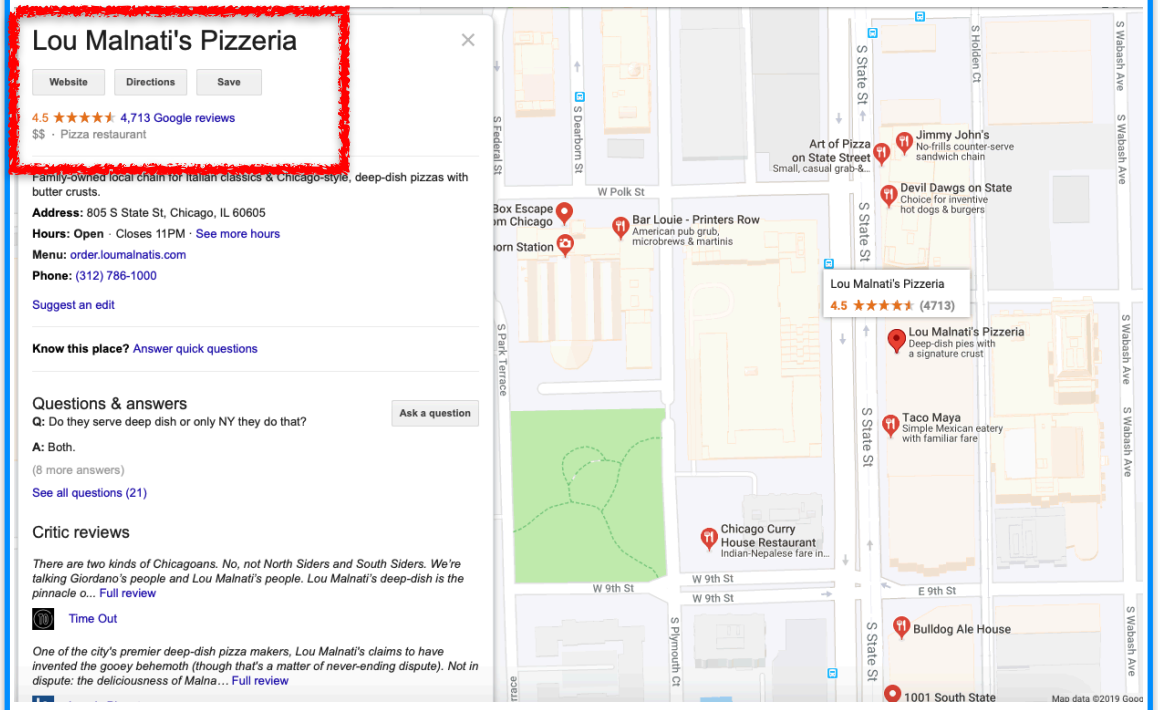
Histogram



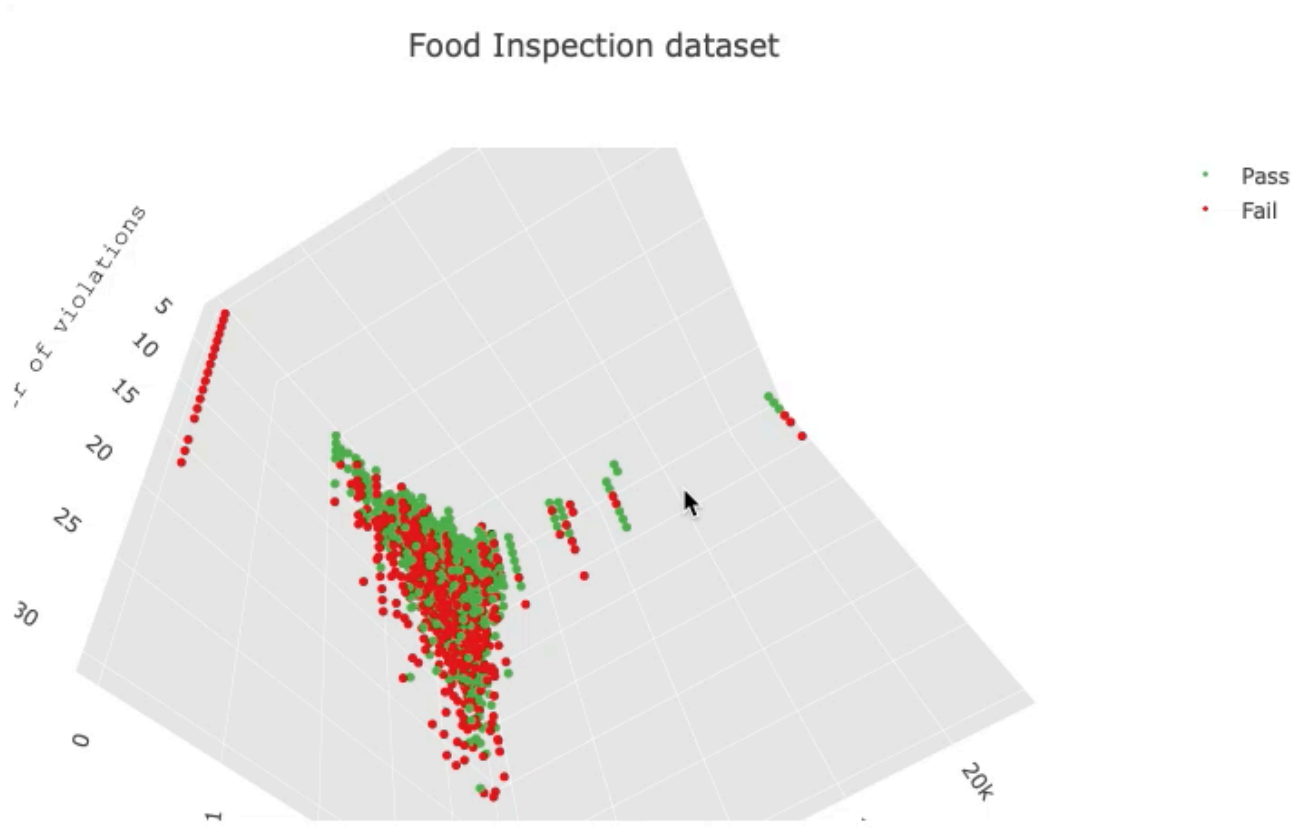
Collect data

- Extract rating, price level and total review from Google Place Detail

Google Place API



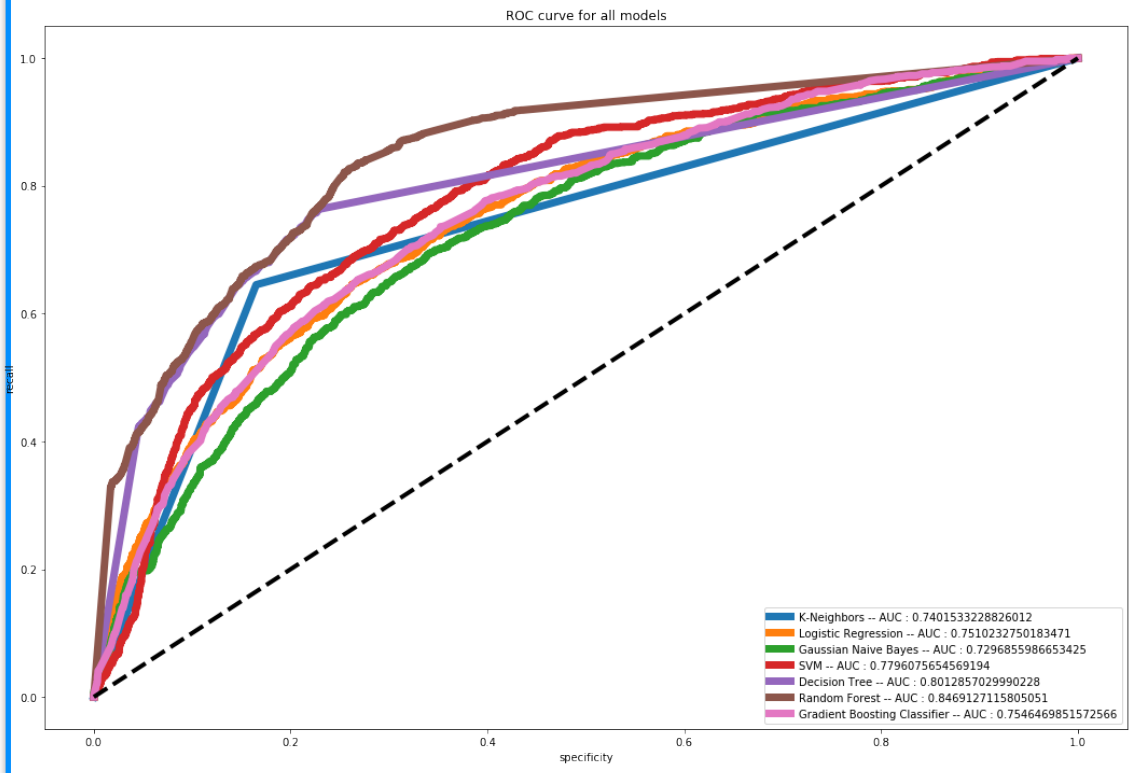
EDA


Food Inspection dataset
Pass
Fail
of violations
0
5
10
15
20
25
30
1
20k

Model Selection

- Random Forest (1st)
- Decision Tree
- Logistic Regression
- KNN
- Naive Bayes

ROC-AUC plot



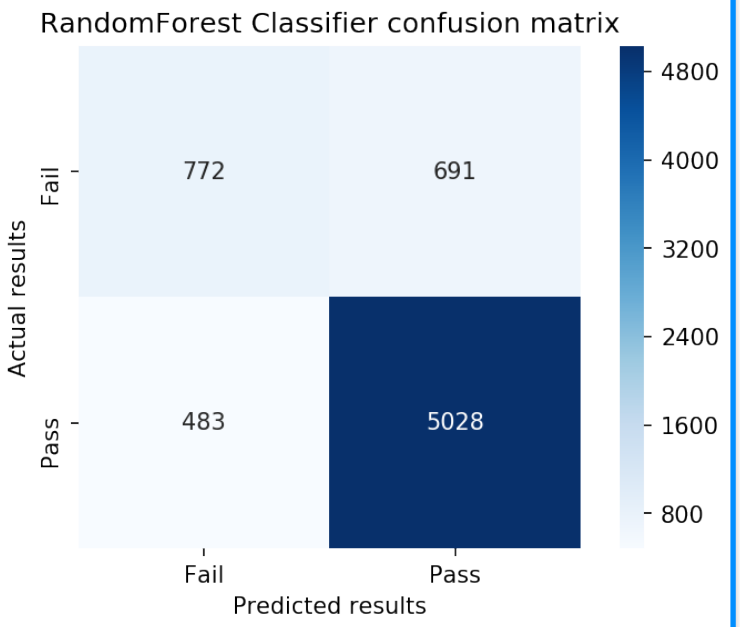
Features

- latitude
- longitude
- number violations
- price_level
- rating
- user_ratings_total
- business activity
- license description

Best Model

- Random Forest
- Hyperparameters :
  - # estimators : 20
  - Tree depth : 10
  - min_samples_split: 5
  - Which feature is best ? : number violations

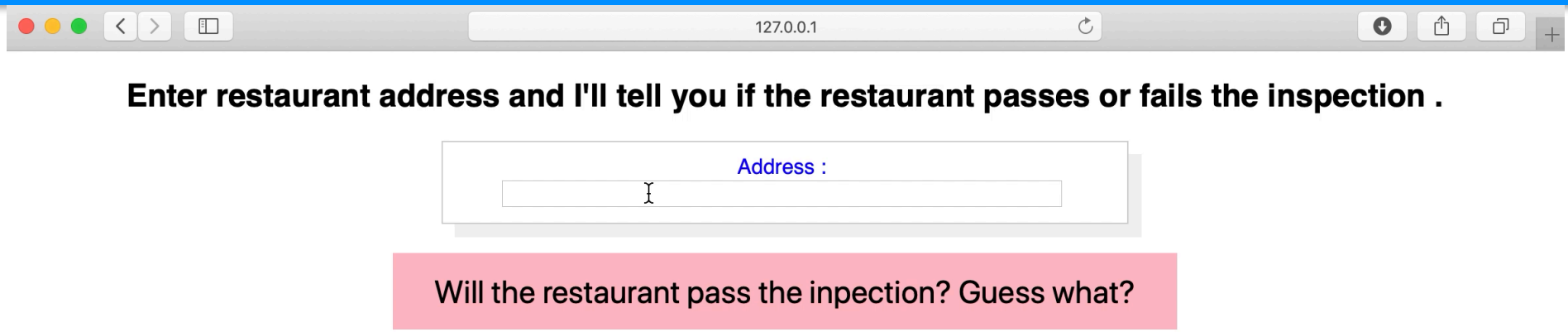
Confusion Matrix



Scores

- Accuracy = 0.8274
- Precision = 0.879175
- Recall = 0.912357
- F1 = 0.895459
- Log loss = 0.4004 (< -log(0.5) = 0.69)

Flask Apps


127.0.0.1
Enter restaurant address and I'll tell you if the restaurant passes or fails the inspection .
Address :
Will the restaurant pass the inpection? Guess what?

Actionable Insight

- Low-rated restaurants are likely to fail the inspection
- Work with other features such as text data from Yelp review, # of available parking, etc
- Use app data/feedback to pinpoint where the model struggle

Questions