Course: CS 109A Introduction to Data Science (Fall 2016)

Final Project: Predicting Food Inspection Outcomes in Chicago (Milestone #4/5)

Team Members: Calvin J Chiew, Angelo Kastroulis, Tim Hagmann

Teaching Fellow: Taylor Names

Baseline Model

Please view this ipython notebook on GitHub for our preliminary modelling: modeling.ipynb

The intermediate steps taken for this milestone are in these notebooks: Food_Inspections.ipynb, inspection_merge_climate.ipynb

GitHub URL: https://github.com/angelok1/cs109project

Summary Report

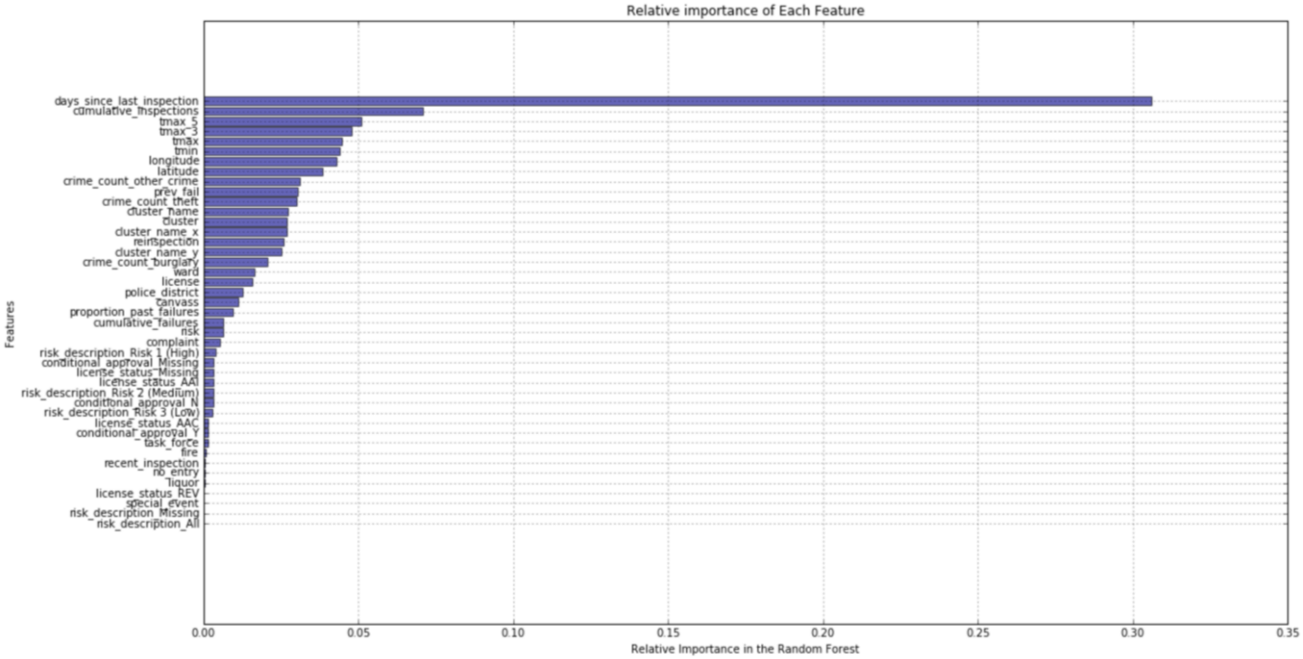
Since the last milestone (data exploration), we have cleaned up our datasets and generated new aggregate variables which we feel are potential predictors of inspection outcome. Specifically, for the main dataset, we created the following variables: proportion of past inspections failed, days since last inspection, and most recent inspection outcome. Inspection type and facility type were grouped into broad categories. For the climate dataset, we generated three-day and five-day average high temperature (rolling mean of max temperatures).

We then merged the main inspections dataset with climate data by date, with business licenses data by license ID number, and with crime and sanitation data by location. We dealt with missing data, applied one-hot encoding on categorical variables, and split our dataset into training and test sets (in 7:3 ratio).

Using a binary outcome (pass/fail), we created 2 baseline models as benchmarks, the first ("all-zero") model predicted all inspections to fail, and the second ("random label") model predicted failures randomly at a probability of the overall proportion of failures. We then modelled our data using various classifiers and used logistic regression as an ensemble method to "stack" all the models. The results, in terms of test accuracy, are as follows.

Model	***Overall Test Accuracy (%)***	***Accuracy for "Pass" class (%)***	***Accuracy for "Fail" class (%)***
All zero	22.57	0.00	100.00
Random label	34.58	22.16	77.21
Logistic regression (weighted)	70.59	64.20	92.52
Linear discriminant analysis (LDA)	77.63	96.45	13.08
Quadratic discriminant analysis (QDA)	77.56	99.34	2.81
Random forest (RF)	82.13	91.82	48.91
Support vector machine (SVM)	77.88	99.58	2.22
ADA boost (ADA)	80.79	91.85	42.85
Ensemble	82.13	91.82	48.91

The relative importance of features in the random forest model is shown below.



<u>Proposal of Future Work</u>

Moving forward, we will take these next steps for the project:

- Examine for collinear variables and explore other data transformations that make sense
- Tune the parameters of the various models to improve their predictive accuracy
- Consider if there is a more meaningful way of evaluating the models (other than test accuracy), such as simulated retrodiction to see which model discovers errant establishments earlier when used to prioritize inspections
- Select the final model based on results and overall objective
- Create a poster and website to communicate our findings with appropriate visualizations