PUBLIC HEALTH AWARENESS

Phase 2:innovation

introduction

Public health, the art and science of preventing disease, prolonging life, and promoting physical and mental health, sanitation, personal hygiene, control of infectious diseases, and organization of health services.

Food Inspections Evaluation:

This is our model for predicting which food establishments are at most risk for the types of violations most likely to spread food-borne illness. Chicago Department of Public Health staff use these predictions to prioritize inspections. During a two month pilot period, we found that that using these predictions meant that inspectors found critical violations much faster.

You can help improve the health of our city by improving this model. This repository contains a training and test set, along with the data used in the current model.

Feel free to clone, fork, send pull requests and to file bugs. Please note that we will need you to agree to our Contributor License Agreement (CLA) in order to be able to use any pull requests.

Original Analysis and Reports

In an effort to reduce the public’s exposure to foodborne illness the City of Chicago partnered with Allstate’s Quantitative Research & Analytics department to develop a predictive model to help prioritize the city's food inspection staff. This Github project is a complete working evaluation of the model including the data that was used in the model, the code that was used to produce the statistical results, the evaluation of the validity of the results, and documentation of our methodology.

The model evaluation calculates individualized risk scores for more than ten thousand Chicagoland food establishments using publically available data, most of which is updated nightly on Chicago’s data portal. The sole exception is information about the inspectors.

The evaluation compares two months of Chicago’s Department of Public Health inspections to an alternative data driven approach based on the model. The two month evaluation period is a completely out of sample evaluation based on a model created using test and training data sets from prior time periods.

The reports may be reproduced compiling the knitr documents present in ./REPORTS.

REQUIREMENTS

All of the code in this project uses the open source statistical application, R. We advise that you use R version >= 3.1 for best results.

Ubuntu users may need to install libssl-dev, libcurl4-gnutls-dev, and libxml2-dev. This can be accomplished by typing the following command at the command line: sudo apt-get install libssl-dev libcurl4-gnutls-dev libxml2-dev

The code makes extensive usage of the data.table package. If you are not familiar with the package, you might want to consult the data.table [FAQ available on CRAN] (http://cran.r-project.org/web/packages/data.table/vignettes/datatable-faq.pdf).

FILE LAYOUT

The following directory structure is used:

DIRECTORY DESCRIPTION

. Project files such as README and LICENSE

./CODE/ Sequential scripts used to develop model

./CODE/functions/ General function definitions, which could be used in any script

./DATA/ Data files created by scripts in ./CODE/, or static

./REPORTS/ Reports and other output are located in

We have included all of the steps used to develop the model, evaluate the results, and document the results in the above directory structure.

The scripts located in the ./CODE/ folder are organized sequentially, meaning that the numeric prefix indicates the order in which the script was / should be run in order to reproduce our results.

Although we include all the necessary steps to download and transform the data used in the model, we also have stored a snapshot of the data in the repository. So, to run the model as it stands, it is only necessary to download the repository, install the dependencies, and step through the code in CODE/30\_glmnet\_model.R. If you do not already have them, the dependencies can be installed using the startup script CODE/00\_Startup.R.

DATA

Data used to develop the model is stored in the ./DATA directory. Chicago’s Open Data Portal. The following datasets were used in the building the analysis-ready dataset.

Business Licenses

Food Inspections

Crime

Garbage Cart Complaints

Sanitation Complaints

Weather

Sanitarian Information

The data sources are joined to create a tabular dataset that paints a statistical picture of a ‘business license’- The primary modelling unit / unit of observation in this project.

The data sources are joined (in SQLesque manner) on appropriate composite keys. These keys include Inspection ID, Business License, and Geography expressed as a Latitude / Longitude combination among others.

Acknowledgements

This research was conducted by the City of Chicago with support from the Civic Consulting Alliance, and Allstate Insurance. The City would especially like to thank Stephen Collins, Gavin Smart, Ben Albright, and David Crippin for their efforts in developing the predictive model. We also appreciate the help of Kelsey Burr, Christian Hines, and Kiran Pookote in coordinating this research project. We owe a special thanks to our volunteers from Allstate who put in a tremendous effort to develop the predictive model and allowing their team to volunteer for projects to change their city. This project was partially funded by an award from the Bloomberg Philanthropies' Mayors Challenge.