BOSTON HEALTH INSPECTION PROJECTION

SI 670 FALL 2020
MCCOY DOHERTY
NICOLAS ORTEGA

An exploration into the effectiveness of machine learning methods in predicting the health-inspection outcomes of Boston-area food service entities based on the city's public-inspection data regarding health-inspections and building-permits. Our work entails a binary prediction model that predicts whether a given health inspection will pass or fail.

MISSION

As we're more aware of the hygiene in places we frequent, food establishment health codes are essential. Our goal is to build a classifier to predict whether a restaurant will pass or fail an inspection, and approximate the severity of the violation if committed.

DATA

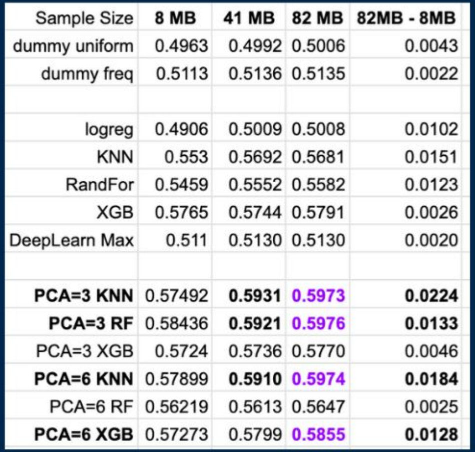
Data available through Boston's Department of Innovation and Technology open-sourced data. Two main sources:

- 625k records detailing outcomes of food-service entities in the Boston area from 2006 to present (Inspections)
- Data detailing regulation-required fixes for building permit eligibility.
- 69 features when joined, 8.8GB filesize

METHODS

- Established performance baseline with static-rule (or "dummy") classifiers
- Experimented with various modeling approaches (LSVC, KNN, RF, XGB, NN)
- Optimized performance and runtime via PCA
- GridSearched best candidates with principal components for model accuracy maximization

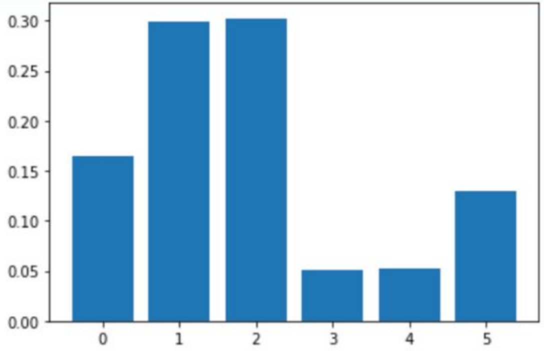
MODEL ACCURACY COMPARISON



Sample Size	**8 MB**	**41 MB**	**82 MB**	**82MB - 8MB**
dummy uniform	0.4963	0.4992	0.5006	0.0043
dummy freq	0.5113	0.5136	0.5135	0.0022
logreg	0.4906	0.5009	0.5008	0.0102
KNN	0.553	0.5692	0.5681	0.0151
RandFor	0.5459	0.5552	0.5582	0.0123
XGB	0.5765	0.5744	0.5791	0.0026
DeepLearn Max	0.511	0.5130	0.5130	0.0020
PCA=3 KNN	0.57492	**0.5931**	**0.5973**	**0.0224**
PCA=3 RF	0.58436	**0.5921**	**0.5976**	**0.0133**
PCA=3 XGB	0.5724	0.5736	0.5770	0.0046
PCA=6 KNN	0.57899	**0.5910**	**0.5974**	**0.0184**
PCA=6 RF	0.56219	0.5613	0.5647	0.0025
PCA=6 XGB	0.57273	0.5799	**0.5855**	**0.0128**

RESULTS

6-COMPONENT FEATURE IMPORTANCE



- Highest-Accuracy Classifier: Gridsearch-tuned XGBoost with 6-PC-transformed data
- Modest improvement from baseline classifier (~10%)
- Unable to achieve better results by utilizing Deep Learning methods