Prediction of City KPIs Using Municipal Finance Data

Stanford University

Stanford ENGINEERING

Filter Feature Selection

- Pearson Correlation (Pearson's ρ),
- Mutual Information (MI),
- Spearman Correlation (Spearman's ρ), and
- Chi-squared (χ^2) Test.

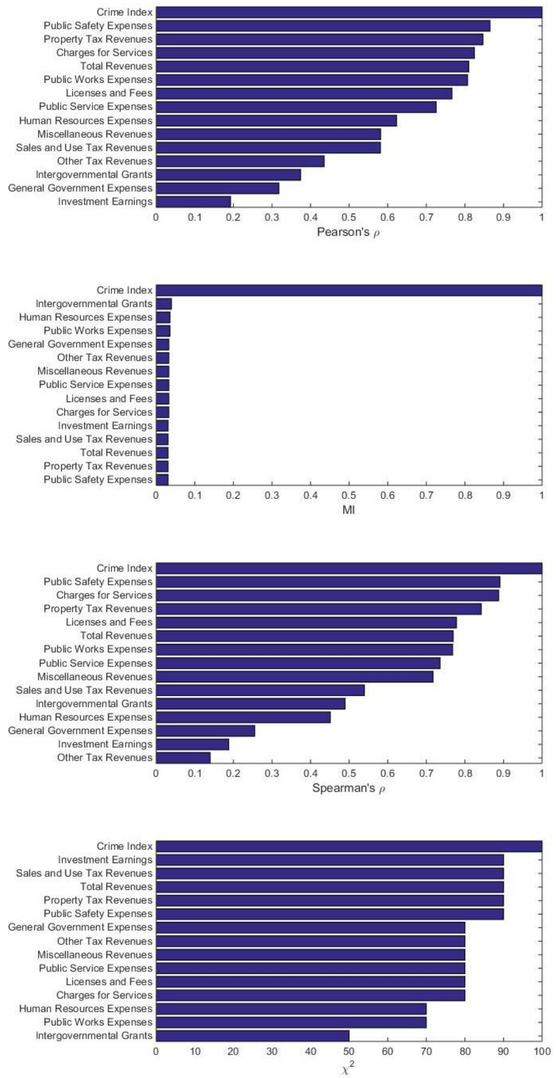


Figure 1—Feature scoring

Unsupervised Clustering

the cluster count using ℓ^2-norm Silhouette value as a measure of purity from dissimilarity (higher is better):

$$\text{silhouette}(x_i) = \frac{\|x_i - C'(x_i)\|_2 - \|x_i - C(x_i)\|_2}{\max\{\|x_i - C'(x_i)\|_2, \|x_i - C(x_i)\|_2\}} \in [-1,1]$$

where

$C(x_i)$: Centroid of cluster containing x_i

$C'(x_i)$: Centroid nearest to x_i satisfying $C(x_i) \neq C'(x_i)$

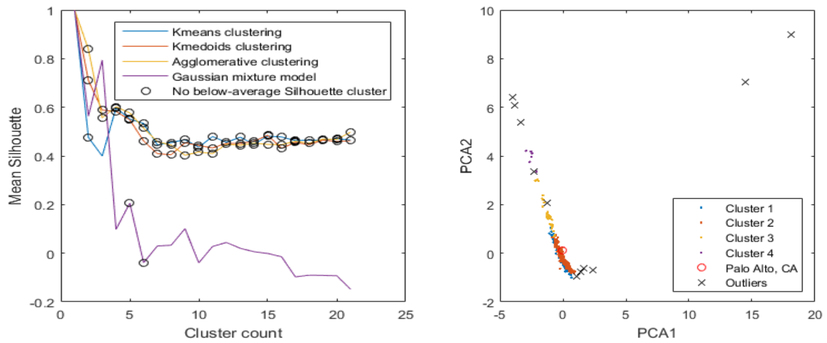


Figure 2—Clustering algorithms performance comparison and Agglomerative clustering, k=4

Data Overview

Figure 3—Data overview

Regression Model

- Linear Regression
- Bayesian Linear Regression
- Neural Network Regression
- Decision Forest Regression
- Boosted Decision Tree

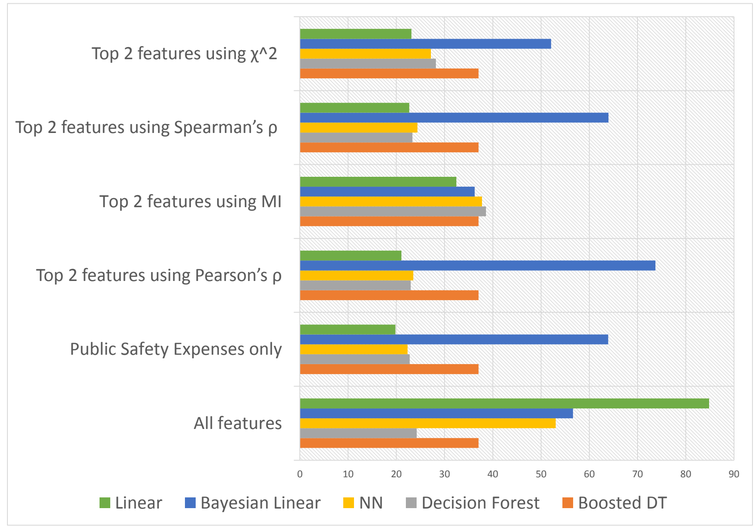
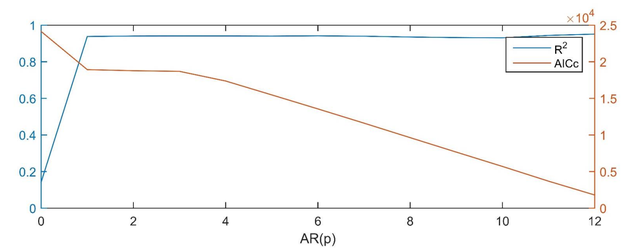


Figure 4—Cross-validation RMSE of baseline model

Autoregression Model



Acknowledgement:

We would like to thank Consulting Professor Bruce Cahan and Visiting Scholar Tomasz Golinski for providing us with city finance data.

Contact information:

Bo Shen | PhD Candidate in Civil Engineering | boshen@stanford.edu

Pradipta A. B. Hendri | SCPD Microsoft | dip@stanford.edu

Kun Shao| SCPD Microsoft | kunshao@stanford.edu