KPI-Driven Predictive ML Models Approach Towards Municipal Budgeting Optimization

Bo Shen Pradipta A. B. Hendri Kun Shao

{boshen, dip, kunshao}@stanford.edu

(CS229 Machine Learning Project Final Report)

***Abstract*— this project aims to use machine learning techniques to predict cities' Key Performance Indicators (KPI) based on municipal governmental budget components towards segments including education, public safety, and health. The project started with building prediction model for the City of Palo Alto and then generalized the model for a cluster of cities that contains the City of Palo Alto. The result shows that autoregressive linear models can predict Crime Index and Health Index reasonably well. By building satisfactory KPI prediction models, this project sets a path towards municipal budgeting optimization.**

I. Introduction

Key Performance Indicators (KPIs) of a city are a set of metrics used to evaluate factors that are crucial to the success of a city. Government annual budgets name and prioritize regional KPI. Through agency sub-budgets and programs, agency missions are turned into programs addressing single or multiple citizen concerns about economic development, education, energy, environment, health, housing, shared infrastructure, public safety and other factors.

Budgets reveal spending by a single program in isolation. Given the complexities of urban life, a single KPI (for instance, health) can be the function of multiple other elements (such as environmental and food pollutants, housing quality, transportation use, and education). Likewise, strategies of improving any given KPI can be ways of budget allocation across multiple agencies.

The project aims to use machine learning to predict KPI using municipal budget data and set a path to optimize a part of or entire budget towards certain KPIs.

II. Methodologies

As an overview, we began with accumulating budgeting and KPI data for different cities in the US. We started with building prediction model for the City of Palo Alto's Crime Index. We then applied unsupervised clustering on demographics data of these cities to identify similarity groups and attempted to generalize the prediction model. The cluster to which Palo Alto belongs was chosen as the group to which we generalized our analysis.

A. KPIs

Cities and other independent parties evaluate the performance using metrics such as the following.

1. Crime rate, as a measure of safety
2. Proportion of residents self-reporting better than fair or poor health, as a measure of health of city residents
3. Ratio of average household income to cost-of-living index (COLI), as a measure of affordability
4. Ethnicity diversity index, which is a normalized measure of ethnicity composition deviation index relative to statewide ethnicity composition

It is the interest of municipal government to develop the necessary infrastructure and environment to bring the maximum benefit for the inhabitants, and KPI-driven approach is a succinct method to summarize the performance of a city when it comes to numerical calculations within a data-driven decision support system.

B. Budget Components

Municipalities produce Comprehensive Annual Financial Report (CAFR) that they publish on online and printed media. The CAFR segments city financial budgeting as follows.

Category of CAFR components	
Revenues	**Expenses**
Sales and Use Tax Revenues Property Tax Revenues Other Tax Revenues Charges for Services Licenses and Fees Intergovernmental Grants Investment Earnings Miscellaneous Revenues Total Revenues	General Government Expenses Human Resources Expenses Public Safety Expenses Public Service Expenses Public Works Expenses Environmental Expenses Other Program Expenses Capital Outlay Principal Debt Service Interest Debt Service Total Operating Expenses

Category of CAFR components *(continued)*		
Sources & Uses	**Change in Fund Balances**	**Ratios**
NEGLECTED	*NEGLECTED*	*NEGLECTED*

These components serve as categories of focus areas in which the city invest for maintenance and future development. Bloomberg L.P. collected these financial data from numbers of cities in the United States, and we can access data through Bloomberg Professional service provided in special Bloomberg Terminal.

C. Data Set

For financial data, we obtained CAFR data from 380 cities in California for years between 2004 and 2014.

For KPI data, we obtained crime rates, cost of living index, health index, and median income from all cities in the United States annually, from 2004 up to 2014.

For demographics data, we obtained 2010 census population data of all census-registered cities in the United States.

To aggregate the data, we chose `geoid` primary index provided by United States Census Bureau to establish relationships in the rows of data obtained from different data sources.

D. Regression Analysis

First, we attempted to predict the crime index of the City of Palo Alto using its financial data. We explored five regression models:

1. Bayesian Linear Regression,
2. Neural Network Regression,
3. Boosted Decision Tree,
4. Linear Regression, and
5. Decision Forest Regression.

Among 20 financial features, we first cleaned the missing data by removing features with missing data; it left us with 14 features with complete dataset.

For each of the five regression models, we first trained the model using the all 14 features with complete dataset. Then, we trained the models using just one feature, the public safety expense, which intuitively, we think is most correlated to crime index. Lastly, we used filter-based selection methods to identify the features that are most predictive, and trained the models using the top two features. The five feature scoring method we tried are:

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

We apply this method to more cities as determined by the clustering result, and obtain a more generic model trained with data belonging to multiple cities instead of a single one.

E. Unsupervised Clustering

It is a possibility that there are multiple underlying models between financial budgeting and KPIs for cities depending on some hidden factors. In reality, cities dissimilar in scale and living standard require different policies of governance to make successful progress. To generalize the prediction model, we attempt to establish similarity clusters of cities in the dataset based on scale and living standard, with the following factors considered:

1. Population census in 2010
2. 10-year population growth from 2000 to 2010
3. Total revenue per capita in 2014
4. Total expense per capita in 2014
5. C2ER's Cost-of-living Index (COLI) in 2014

Each cluster will have its own regressive model, ultimately allowing higher degree of freedom to the big picture, to reduce generalization error. We explore multiple clustering algorithms, which are:

1. K-means clustering,
2. K-medoids clustering,
3. Agglomerative clustering, and
4. Gaussian mixture model.

We determine 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 $C(x_i) \neq C'(x_i)$

In addition, separately for each algorithm and cluster count we remove outlier points from data that induces the algorithm to create clusters containing low number of points. With this reduced data set, we run each clustering algorithm with 100 replicates per cluster count, choosing the result that maximizes mean Silhouette value. We then choose the model that maximizes the mean Silhouette value over different cluster counts. The pseudo-algorithm for each clustering algorithm is as follows:

```
def n_replicate = 100
for k = 2 to sqrt(size(data)):
  def data_k = data
  def C_low = {dummy point}
  while size(C_low) > 0:
    run clustering on data_k
    if any cluster has less than min_size points:
      C_low = join all clusters with size < min_size
      remove all points in C_low data from data_k
  end while
  for r = 1 to n_replicate:
    def index_kr = result of clustering on data_k
    def sil_kr = silhouette value of index_r
    if mean(sil_kr) > mean(sil_k):
      index_k = index_kr, sil_k = sil_kr
    end if
  end for
end for
def index = index_k that maximizes sil_k over k
```

III. Results

A. Baseline Model

First, we attempted to predict the crime index of the City of Palo Alto using its financial data.

The results of feature scoring in descending order is shown on Figure **1**. To evaluate the models, 10-fold cross validation was used. The estimated generalization errors of each model in terms of root mean squared error (RMSE) are presented in Figure 2. Each row represents errors of models using different sets of features. For example, the row, Public Safety Expenses, shows the errors from just using one feature, Public Safety Expenses; the row, Pearson Correlation, shows the errors models using the

top two features selected by Pearson Correlation Method. Among all the models we trained, the linear regression model using the public safety expenses alone was able to predict the crime index with the lowest error of 19.80.

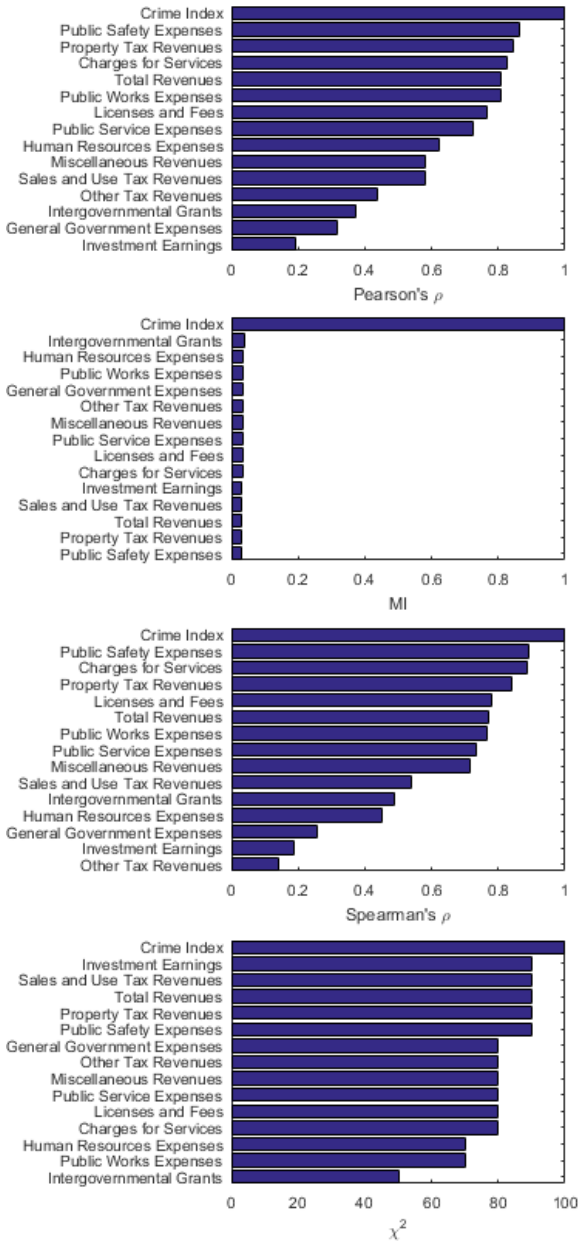

Figure 1—Feature scoring.

B. Unsupervised Clustering

Unsupervised clustering was applied against separately obtained scale and living standard data for 477 cities in California as described in Part E of Methodologies. Unfortunately, some values are found to be missing after merging the factors. Due to the nature of unsupervised clustering, it makes little sense to apply imputation or surrogate techniques, so we simply remove points having one or more missing factors and were left with 364 points. Clustering performance based on Silhouette value is shown in Figure 3. Additionally, another criterion is applied for a model with given parameter to be acceptable: all cluster within the model must have at least one point with Silhouette value higher than the model-wide mean Silhouette value. Points satisfying this criterion is marked with black circle in Figure 3.

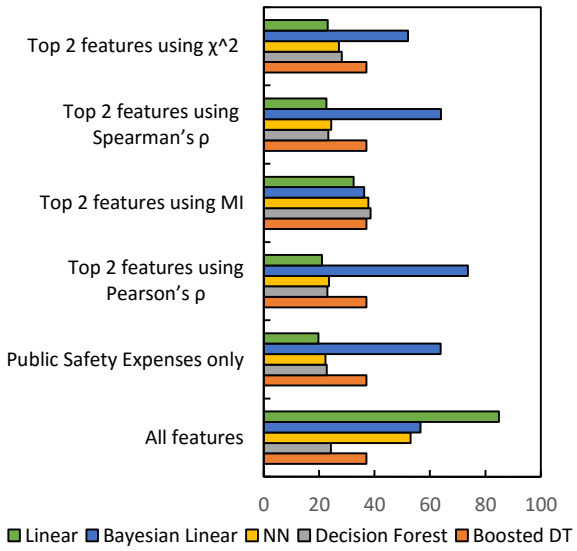

Figure 2—Cross-validation RMSE of baseline model.

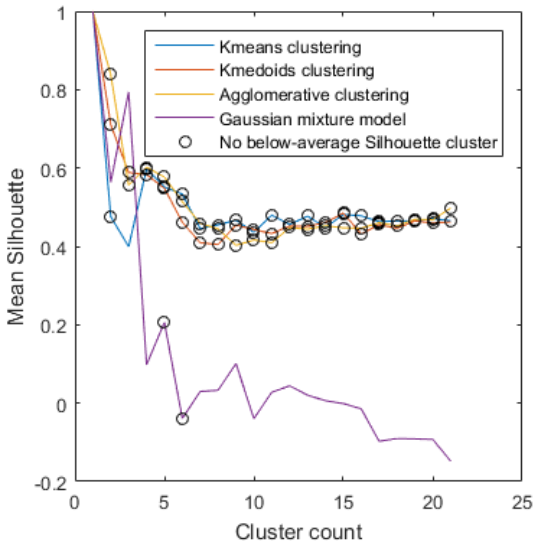

Figure 3—Clustering algorithms performance comparison.

From Figure 3 we can observe that Gaussian mixture model quickly degenerates with increasing cluster count, whereas the other 3 considered algorithms performs similarly. We note that Silhouette value by definition will be high for cluster count of 2 so long as the algorithm is allowed to converge, due to ℓ^2-norm being used for both clustering and Silhouette, as $C(x_i) \geq C'(x_i)$ always in this setting, keeping silhouette$(x_i) \geq 0$ for all points. With this consideration, as well as desire to lower the size of cluster to which Palo Alto belong, we pick the agglomerative clustering model pruned to 4 clusters because it maximizes the Silhouette value. Figure 4 illustrates how this model clusters the

points, using principal components. We found 225 cities similar to Palo Alto based on the factors considered.

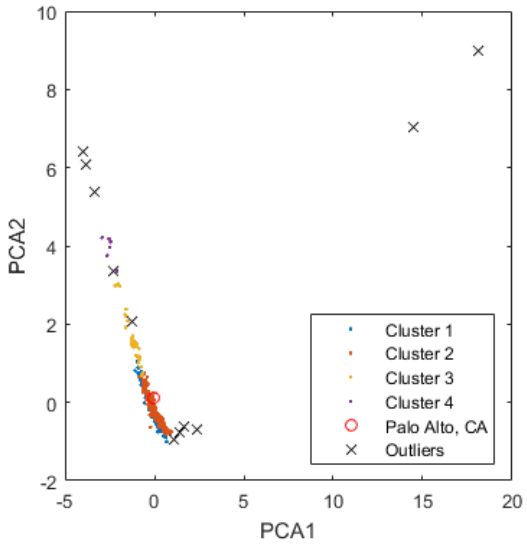

Figure 4—Agglomerative clustering on PCA, k=4.

Based on our baseline model findings in Figure 2, we determine that the project should be scoped to linear models. To improve the performance of models, we augment stepwise linear model construction with autoregressive terms. The significance of autoregressive terms is supported by assumption that pre-existing quality of life in the city affects the extent to which budgeting decision for the year could influence the outcome of the development which is incremental in nature.

C. Public Safety KPI

We applied linear model augmented with autoregressive terms to the public safety KPI, which is the crime index of cities inside cluster containing Palo Alto. The chronological extent of our dataset allows for autoregression up to 12th order, and we used 10-fold cross-validated R^2 as goodness-of-fit statistic of models. The result is shown in Figure 5, along with p-values of the coefficients involved in the linear model. To evaluate the significance of augmented autoregressive terms, we consider only the maximum of p-values associated with these terms.

Shown also in Figure 5 is the Akaike Information Criteria (AIC) of models, corrected to account for finite sample sizes. Although AIC typically provides sensible measure to compare models having varying number of coefficients, in this application we observe that the AIC decreases due to the likelihood component diminishing in value. This may be caused by the number of training samples accounted for in AIC calculation decreases as we expend the temporal limit of our dataset into providing values used for autoregressive terms. Therefore, we simply use the cross-validated R^2 to choose the model.

In Figure 5 we see that the cross-validated R^2 receive little to no improvement past the first order of autoregression, so in the interest of simplicity and avoiding overfitting, we opt for the first order model:

$$I_S(t) = 0.0956 \cdot E_{PS}(t) + 0.9410 \cdot I_S(t-1) + 4.4689$$

where:

$I_S(t)$: Crime index of the year

$I_S(t-1)$: Crime index of preceding year (autoregressive)

$E_{PS}(t)$: Public safety expense of the year, in million US$

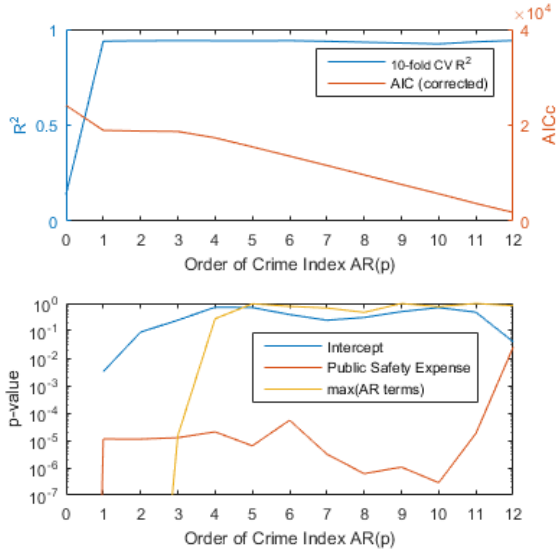

Figure 5—Performance of autoregressive linear models for public safety KPI.

D. Health KPI

We noted from Figure 2 that using all CAFR components increased the validation error of model. For health KPI model, we apply stepwise construction of linear model in addition to the autoregressive terms similar to public safety KPI model. Stepwise selection will prevent overfitting due to including all CAFR components.

During model construction, we encountered a lot of missing data for CAFR components for a given year. To maximize training dataset utilization, since the model is linear, we impute missing data with zeros. Our temporal dataset enables us to construct up to 9th order autoregression.

We found, as can be seen in Figure 6, that the autoregressive models perform poorly for models lower than 8th order. The AR(8) model, however, does not include any of CAFR components. Therefore, we will pick the AR(9) model:

$$I_H(t) = 0.0020 \cdot E_{HR} + \sum_{p=1}^{9} \{\alpha_p I_H(t-p)\} + 0.1060$$

where:

$I_H(t)$: Health index of the year

$I_S(t-p)$: Health index of p years ago (autoregressive)

$E_{HR}(t)$: Human resources expense of the year, in million US$

α_p: Coefficient of p-th autoregressive term

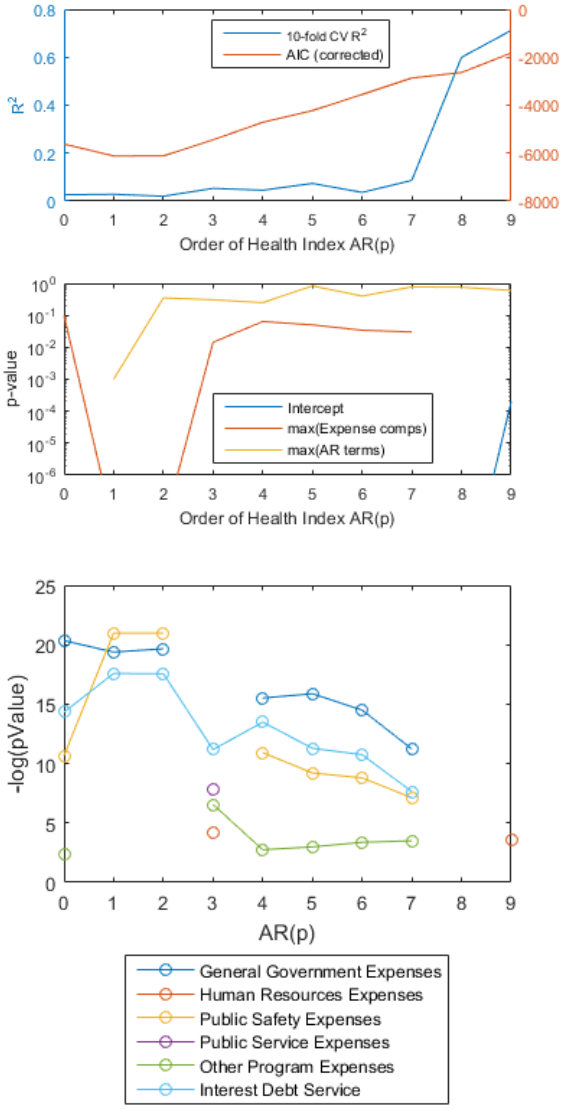

Figure 6—Performance of autoregressive linear models for health KPI.

E. Affordability KPI

Affordability KPI data was limited to a single year, which puts autoregression out of reach. The resulting model is weak:

	Estimate	p-value
(Intercept)	675.1712	$8.5665 \cdot 10^{-48}$
Public Safety Expenses	−1.7796	0.0319

$n = 223$ $R^2 = 0.0207$ 10-fold CV $R^2 = 0.0020$

However, we also see weak performance in non-autoregressive linear models for public safety and health KPI. The failure to obtain a meaningful model for this KPI does not present evidence that would invalidate our proposed methodologies for the project. A meaningful model may be obtained with more temporal data.

F. Diversity KPI

Diversity KPI data, as with that of affordability KPI, was limited to a single year. The resulting model is again weak without any autoregressive terms:

	Estimate	p-value
(Intercept)	0.5380	$8.6080 \cdot 10^{-82}$
Public Safety Expenses	0.0015	0.0002
Interest Debt Services	−0.0928	0.0341

$n = 223$ $R^2 = 0.0691$ 10-fold CV $R^2 = 0.0047$

For this results we also maintain that our proposed methodologies may perform well given more temporal data.

IV. Conclusion

This project showed promising outcome for public safety KPI (crime rate) and health KPI (proportion of residents self-reporting better than fair or poor health) as shown by the goodness of fit of the autoregressive linear models for these KPI. The optimization of municipal budget is achievable using the outcome of this project, for public safety and health KPI for a cluster of 225 cities.

Further works deriving from this project should consider obtaining more training data. It is observed from the goodness of fit of models for public safety KPI and health KPI that linear models without autoregressive terms can be dramatically improved by adding autoregressive index values. We observed that affordability KPI and population diversity KPI models have poor performance using only the latest publicly available KPI values. Obtaining affordability KPI would require us to purchase the Cost-of-living Index report for US$750 to enable the level of analysis equivalent to that of public safety and health, which was cost-prohibitive. Nevertheless, based on the evidence available to us and given the importance of affordability to the quality of life, we recommend further works to acquire temporally extensive data points for affordability KPI.

Further works should also consider expanding geographical extent of the analysis. Due to Bloomberg Terminal usage limits imposed on us, our analysis could only consider financial budgeting data from the state of California.

Lastly, further works should consider augmenting other nonlinear regressive models with autoregressive terms for the final models. Due to time and resource limitation, we only considered autoregressive linear models.

V. Acknowledgements

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

VI. References

Bloomberg L.P. (2015) City Credit Report. Retrieved Nov. 11, 2015 from Bloomberg database.

City-data.com, 'City-Data.com - Stats about all US cities', 2015. [Online]. Available: http://www.city-data.com/. [Accessed: Oct-2015].