STATS 140XP Final Report

Team 13

Anna Heppelmann, Aurelia Santosa, Hyungjoo (HJ) Han,
Saul Magallon, Nicole Lee, Yiting Wang, Pavan Sah

Assessing Fiscal Risk and Financial Stability: Analyzing Trends and Predicting Risk Factors in California's State Budget

Introduction

The California State Auditor's Local High-Risk Program is an initiative designed to identify and address fiscal and operational risks in local governments across the state. The program assesses cities, counties, and special districts that may be experiencing financial distress or governance issues, ensuring accountability and transparency in local government operations. Data from the local high-risk dashboard enables the local government to make data-driven decisions, receive monitoring and specific recommendations, and take corrective actions in time when facing fiscal distress.

Literature Review

Fiscal stability in local governments is crucial for maintaining essential public services, ensuring economic growth, and fostering public trust. Stable finances provide consistent services such as education, public safety, and infrastructure maintenance, while also allowing for effective long-term planning and investment. Fiscal instability can lead to service disruptions, deferred maintenance, and a diminished quality of life for residents. According to the Urban Institute, in fiscal year 2021, California's combined state and local direct general expenditures totaled $574.8 billion. This significant expenditure underscores the extensive responsibilities shouldered by local governments in the state, bringing in significant challenges in wise operation. Additionally, due to high marginal tax rates, California is ranked 4th for the worst economic policy environment, as well as 3rd for the worst business tax climate.

Research Objectives & Hypotheses

Despite its importance, the California high risk dashboard has been discontinued since October 2023. Hoping to reinvestigate the importance of having a transparent risk analysis for local government and re-explore the potential of this data in detecting and managing fiscal distress, we will look at the audit of California local government fiscal health from 2019 to 2020.

Research Questions

In this study, we will perform exploratory data analysis on California Fiscal Health for the year 2019-2020. We will aim to solve the following questions:

1. What are the possible factors that contribute to a local government to be listed as "high risk"?
2. Can we predict fiscal risks in the future using past audit data?

Method:

Our research is derived from the California State Auditor's Local High-Risk Program for the year 2019-2020. Since the data is sourced from government audits and interactive dashboards, it is a secondary data source. This dataset includes the city names with variables like revenue, reserves, debt burdens, liquidity, pension obligations, other post-employment benefit (OPEB) obligations, and more. There are a total of 65 different variables which show the rank, points, risk factor, ratio, and balances of the different financial factors which have both quantitative and qualitative data.

Using this data, our goal is to analyze the financial data and identify key factors that contribute to California's local governments' fiscal risk. Our process of analysis will have four main steps:

1. Exploratory Data Analysis (EDA)

While we have briefly examined the data set, performing EDA will allow us to gain a better understanding of key trends, distributions, and even identify potential issues that may pose a problem to us in the future.

EDA Methods to perform:

a. Data Cleaning and Preprocessing: We check for missing values and identify any outliers within the dataset. If need be, we can either remove or impute missing values.

b. Summary Tables: Summary statistics are computed showing information like mean, median, standard deviation, count, etc. for all the variables.

c. Visualizations: i. Histograms: Generated to examine the distribution of each variable. ii. Scatter plots: Generated to show potential relationships between the variables and fiscal risk classifications.

Correlation Analysis

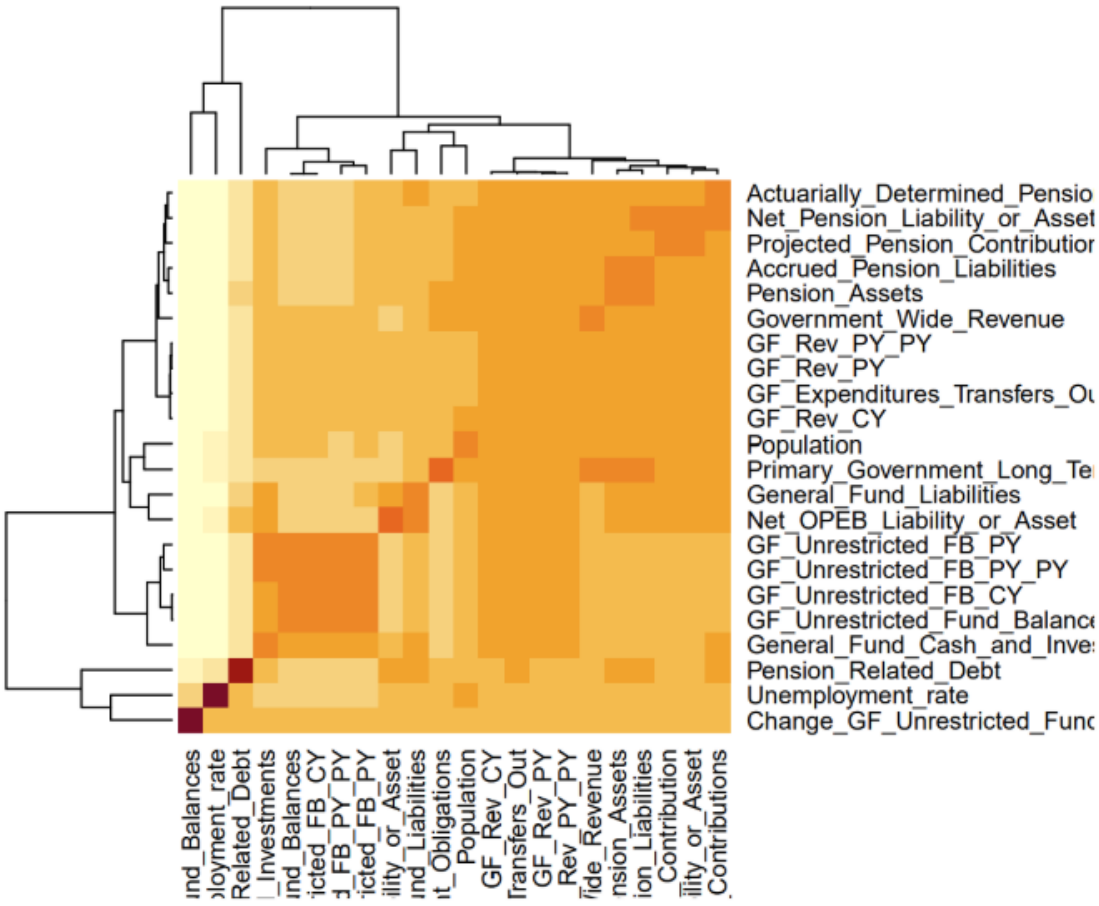

Figure 1: image

Strong correlations found between pension-related ratios and debt metrics. Since pension obligations are a long-term liability, they significantly impact a city's financial health. Revenue-related ratios show moderate correlations with liquidity and reserves.

Visualizations: Possible Factors Leading to High Risk

Figure 2: image

All three of these graphs demonstrate that the less points you have in terms of general fund reserve points, burden points, and liquidity points, the more likely you are to be at high risk for that respective category.

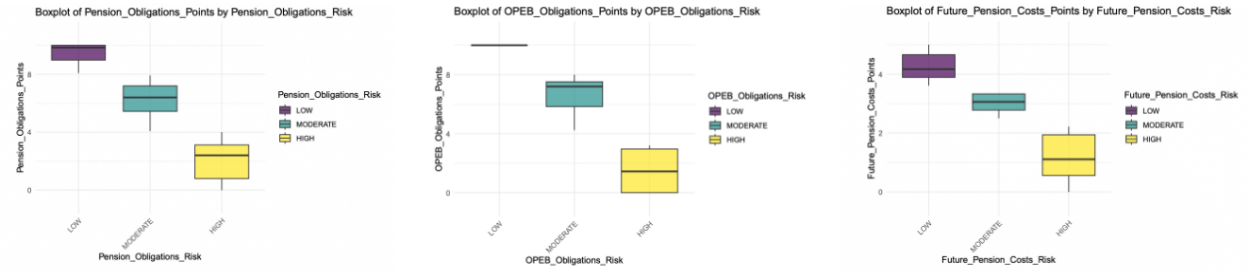

Figure 3: image

All three of these graphs demonstrate that the less points you have in terms of obligation points, OPEB obligation points, and Future Pension Costs points, the more likely you are to be at high risk for that respective category.

Potentially Important Factors:

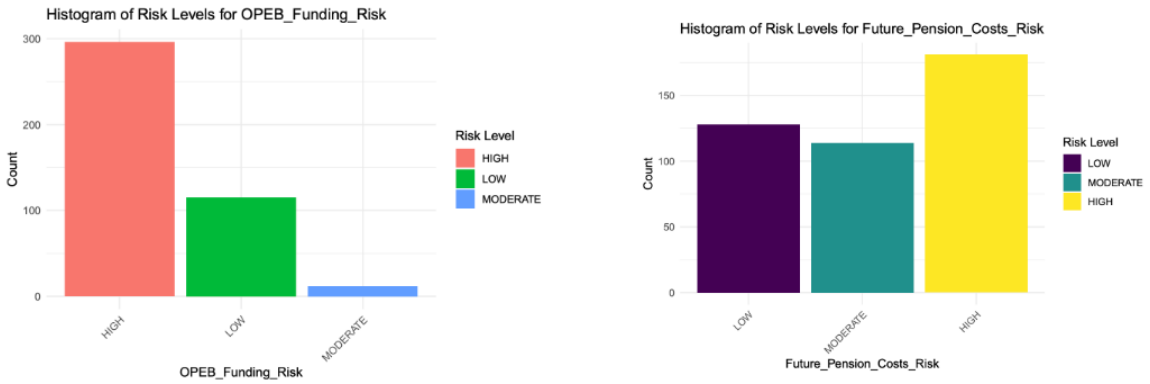

Figure 4: image

The factors of OPEB Funding Risk and Future Pension Costs Risk both have high counts of high risk, exemplified in the histograms above.

2. Statistical Modeling:

Once EDA is performed, we create statistical models to see which variables contribute to a local city's fiscal risk status.

a. Logistic regression: Used to identify which variables increase the probability of a city being high-risk (classification).
b. Linear regression: Used to identify which variables contribute to a city's funds (numerical values like revenue, debt, funding, etc).
c. Variance Inflation Factor (VIF): We check for multicollinearity between the variables and identify whether we should remove/transform them.

Ordinal Logistic Regression:

```
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## General_Fund_Reserves_Ratio -0.24111    0.08069  -2.988  0.00281 **
## Debt_Burden_Ratio            0.58626    0.21870   2.681  0.00735 **
## Liquidity_Ratio             -0.06083    0.01540  -3.950  7.8e-05 ***
## Revenue_Trends_Ratio        -2.20938    1.71059  -1.292  0.19650
## Pension_Obligations_Ratio    1.16436    0.56056   2.077  0.03779 *
## Pension_Funding_Ratio        1.74199    1.57109   1.109  0.26753
## Pension_Costs_Ratio         -5.52329    6.08241  -0.908  0.36384
## Future_Pension_Costs_Ratio  14.62631    5.26400   2.779  0.00546 **
## OPEB_Obligations_Ratio      -0.02841    0.70131  -0.041  0.96769
## OPEB_Funding_Ratio          -1.03841    0.36949  -2.810  0.00495 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                Estimate Std. Error z value
## LOW|MODERATE    -2.319      1.425  -1.627
## MODERATE|HIGH    2.458      1.380   1.781
```

Figure 5: image

The ordinal logistic regression demonstrates the following factors are statistically significant: General_Fund_Reserves_Ratio, Debt_Burden_Ratio, Liquidity_Ratio, Pension_Obligations_Ratio, Future_Pension_Costs_Ratio, and OPEB_Funding_Ratio

A positive coefficient indicates an increase in the predictor raises the chances of being in a higher-risk category and vice versa for a negative coefficient.

Linear Regression and VIF

For linear regression, the R-squared of about 0.0565 means that the four predictors (General_Fund_Reserves_Ratio, Debt_Burden_Ratio, Liquidity_Ratio, Revenue_Trends_Ratio) used tell us about 5.65% of the variation in government-wide revenue (which is pretty low). Debt_Burden_Ratio is the only statistically significant predictor, with its p-value close to 0, and tells us that a one-unit increase in Debt_Burden_Ratio is linked to an increase of about 474 million in Government_Wide_Revenue. Cities that have higher debt burdens seem to have higher government-wide revenues. For the VIF, all values circle around 1.0, telling us there is no significant multicollinearity among the four variables.

3. Predictive Modeling:

In order to predict whether a city will be low or high-risk, we need to build a predictive model. These are some of the different models we will test to predict risk status:

a. Logistic regression: The simplest model for predictive modeling.
b. Decision trees/Random Forests, Gradient Boosting (i.e. XGBoost): Gradient Boosting is the most complex predictive model which improves based on previous errors. Random Forests build independent trees but are much easier to interpret.

4. Model Selection and Evaluation:

From all the different models we have created so far, in order to choose the best predictive model, we look at the accuracy rate for the classification models and generate a feature importance plot to see which variables

contributed the most to predicting risk status.

Random Forest Model and Feature Importance

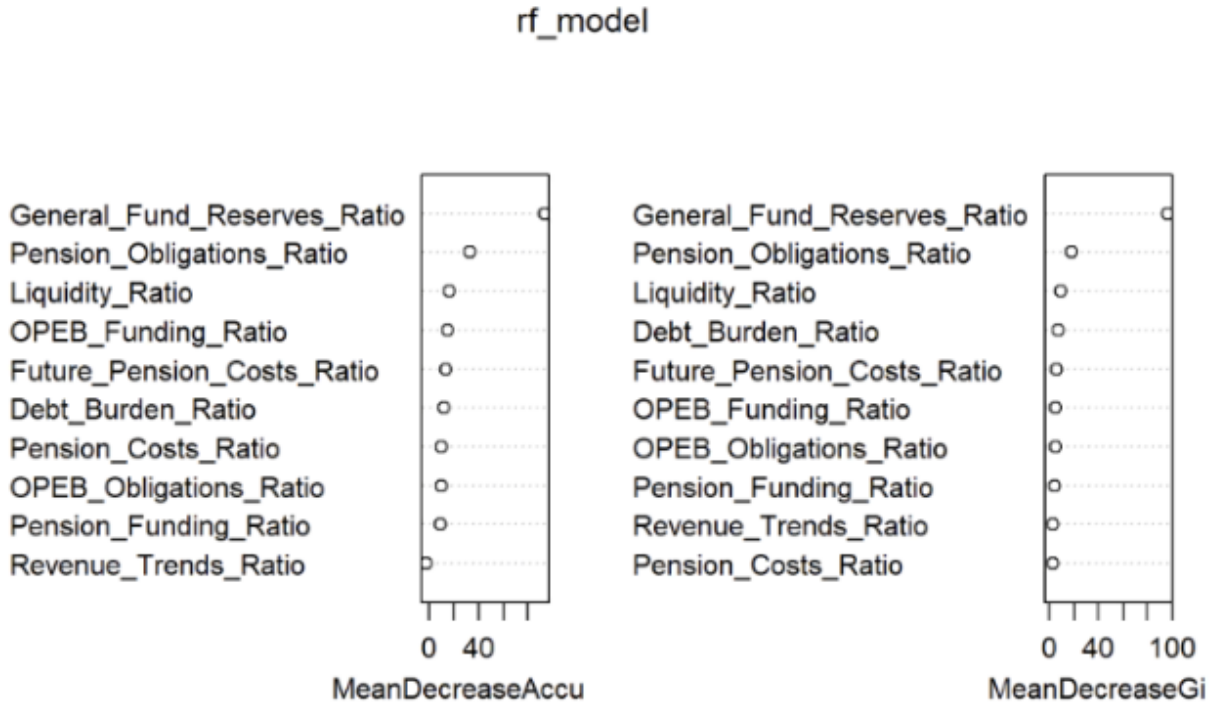

Figure 6: image

Using 10 splits (all predictors), the model misclassified about 12.42% total observations (OOB error rate). While the OOB error rate is relatively low, the class.error for LOW is extremely high. Variable importance was determined through the Mean Decrease in Accuracy, and the Mean Decrease in Gini Impurity. In both these charts, general fund reserves has the greatest value, meaning it is a significant contributor to predicting the overall risk.

XGBoost Model and Feature Importance

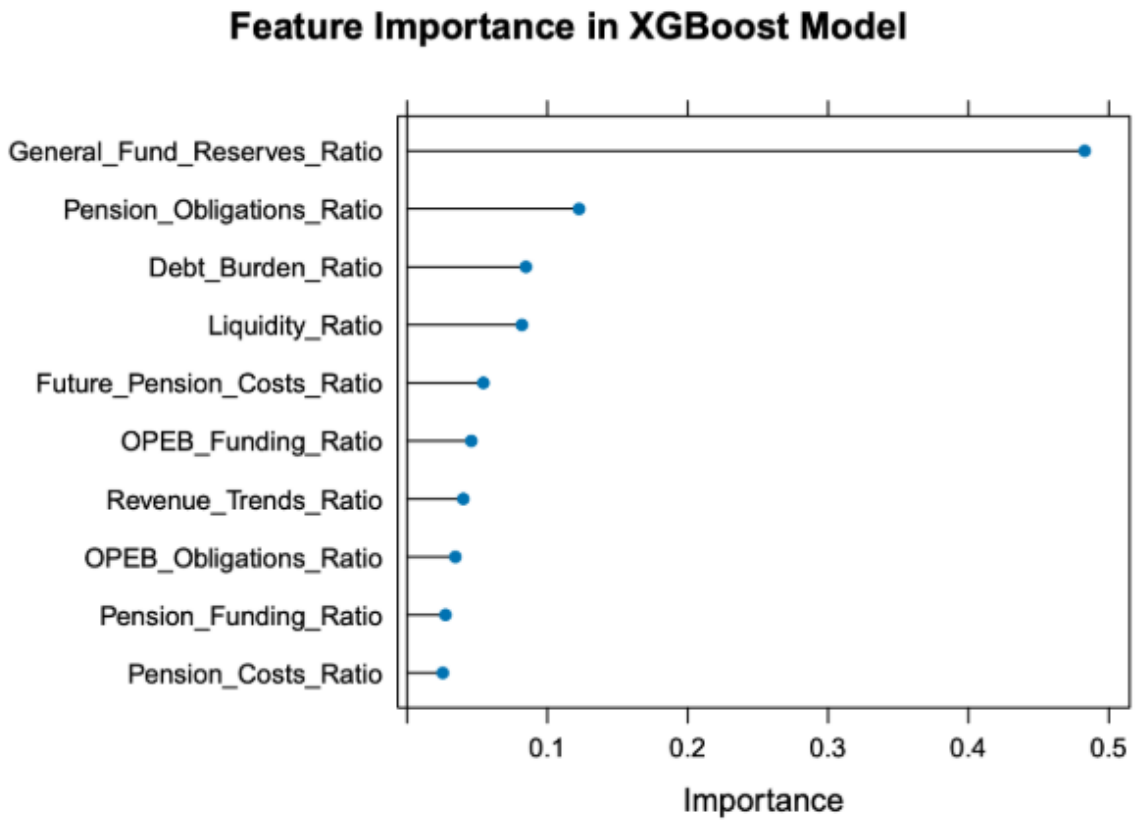

Figure 7: image

The XGBoost model was trained using 10-fold cross validation and applied to the training dataset. The model was tuned using hyperparameters nrounds = 150, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1 and subsample = 0.5 . Similar to the Random Forest model, the XGBoost model also performed relatively well achieving a prediction accuracy of 92%. Furthermore, we see similar results of general fund reserves being a key contributor for risk classification.

Can We Predict Fiscal Risks in the Future Using Past Audit Data?

Random Forest Model (87.58% accuracy):
-Misclassified 12.42% of observations.
-Struggled with identifying low-risk cities.

XGBoost Model (92% accuracy):
-Best performing model.
-Reinforced that General Fund Reserves is the strongest predictor.
-More robust against overfitting compared to Random Forest.

Yes, but to an extent. Machine learning models provide reasonable accuracy (up to 92%), but there are limitations to the predictions.

Limitations

While the models may perform exceptionally well, there are certain limitations that must be acknowledged. This dataset covers only 2019-2020, meaning we only have yearly data and time series analysis cannot be done to analyze long-term trends/changes over the years. Furthermore, 2019-2020 was during the COVID-19

pandemic. Due to the pandemic, government funds may have branched off from normal trends that may have been observed over the years. This would make it difficult for us to generalize our findings and predictive models to future/previous years. Additionally, the dataset does not provide enough information to perform deeper analysis for risks (e.g. expenditures breakdown or sources of revenue).

Potential Impact:

Our research can help in identifying key issues that come with fiscal distress in local governments. By being able to identify such variables/trends, we can contribute to developing efficient strategies, laws, systems, etc. that can help local governments from facing fiscal risk. For instance, using our predictive model, auditors can watch out for warning signs (e.g. high debt, low reserves, low revenue, etc) to recommend early interventions like policy changes or budgeting adjustments. With an effective predictive model, local governments can recover at a much faster rate and even completely prevent fiscal distress if proper measures are taken to watch out for warning signs.

References

[1] Davenport, Andrew. "California's High-Risk Dashboard Is Gone without a Trace but Should Not Be Forgotten: California Policy Center." California Policy Center, 25 Oct. 2024, californiapolicycenter.org/californias-high-risk-dashboard-is-gone-without-a-trace-but-should-not-be-forgotten/ .

[2] Winegarden, Wayne. "Californians Have Little to Show for All That Government Spending." Pacific Research Institute, 28 Jan. 2025, www.pacificresearch.org/californians-have-little-to-show-for-all-that-government-spending/?utm_source=chatgpt.com#_ftn3.