Predicting Bill and Tax Passage: A Historical Data Analysis Approach

Mohammad Shahed Akhtar Mohammad Nizamoddin MS. Data Science ’23

Mentor: Dr. Christelle Scharff

Pace University, Seidenberg School of CSIS

Github

Abstract

In response to the escalating importance of predicting the passage of bills and tax proposals within legislative bodies, the demand for a robust predictive model has become increasingly apparent. This research addresses a noticeable gap in systematic studies in this domain by introducing an advanced predictive algorithm designed to forecast the outcomes of legislative decisions.

The study encompasses the exploration and evaluation of diverse machine learning models tailored for the specific context of legislative prediction. Models include Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, Support Vector Machines, Neural Networks, Naive Bayes, K-Nearest Neighbors, and Ensemble Methods such as Voting Classifier. Each model brings its unique strengths to the prediction task, catering to different aspects of legislative dynamics.

Additionally, the research delves into the identification of crucial variables essential for accurate predictions, employing model interpretation techniques such as Lime, SHAP, and Eli5. The models undergo rigorous evaluation using a dedicated dataset that incorporates legislative variables, historical data, and contextual factors. Performance metrics including accuracy score, confusion matrix, and classification report are employed, showcasing the exceptional performance of several models with accuracy values consistently exceeding 85%.

This study not only fills a significant void in the existing literature but also serves as a foundation for further exploration in this underexplored domain. The findings pave the way for advanced analyses, facilitating the development of even more accurate predictive models and fostering future avenues of research. In essence, this research represents a crucial step towards enhancing our understanding of legislative prediction, propelling the field towards new dimensions of inquiry.

Research Question

Que 1- Can Machine Learning models help us predict whether proposed laws and taxes will get approved, using different methods like decision trees and neural networks?

Que 2- What can we learn about how laws are made by figuring out which factors are most important, and how does this knowledge help make our predictions more accurate?

Dataset

We've gathered a dataset from the California State Treasurer's Office, specifically from the California Debt and Investment Advisory Commission. The dataset comprises 5,979 records and contains 13 columns capturing various aspects of legislative activities.

Key Points:

- Source: California State Treasurer's Office - California Debt and Investment Advisory Commission
- Size: 5,979 rows, 13 columns
- Purpose: Analyzing legislative data for predictive modeling in understanding and forecasting the passage of bills and tax proposals

Methodology

❑ **Data Collection and Preprocessing:**

- Compiled data from the California State Treasurer's Office and the California Debt and Investment Advisory Commission.
- Integrated datasets into a unified dataset for analysis.
- Applied data cleaning techniques to handle missing values and ensure data quality.

❑ **Exploratory Data Analysis (EDA):**

- Conducted in-depth data analysis using visualization libraries such as seaborn and matplotlib.
- Explored patterns and insights within the legislative data

❑ **Feature Engineering:**

- Utilized the Label Encoder library in Python to encode categorical variables.
- Assigned numerical values to labeled data to ensure compatibility with machine learning models.

❑ **Addressing Class Imbalance:**

- Mitigated class imbalance by applying the Synthetic Minority Over-sampling Technique (SMOTE) on the training dataset.
- Specifically addressed the overrepresentation of certain class labels

❑ **Interpretable Model Implementation:**

- Implemented Local Interpretable Model-Agnostic Explanations (LIME) to provide insights into individual predictions.
- Utilized SHAP (Shapley Additive Explanations) for global explanations of the models.

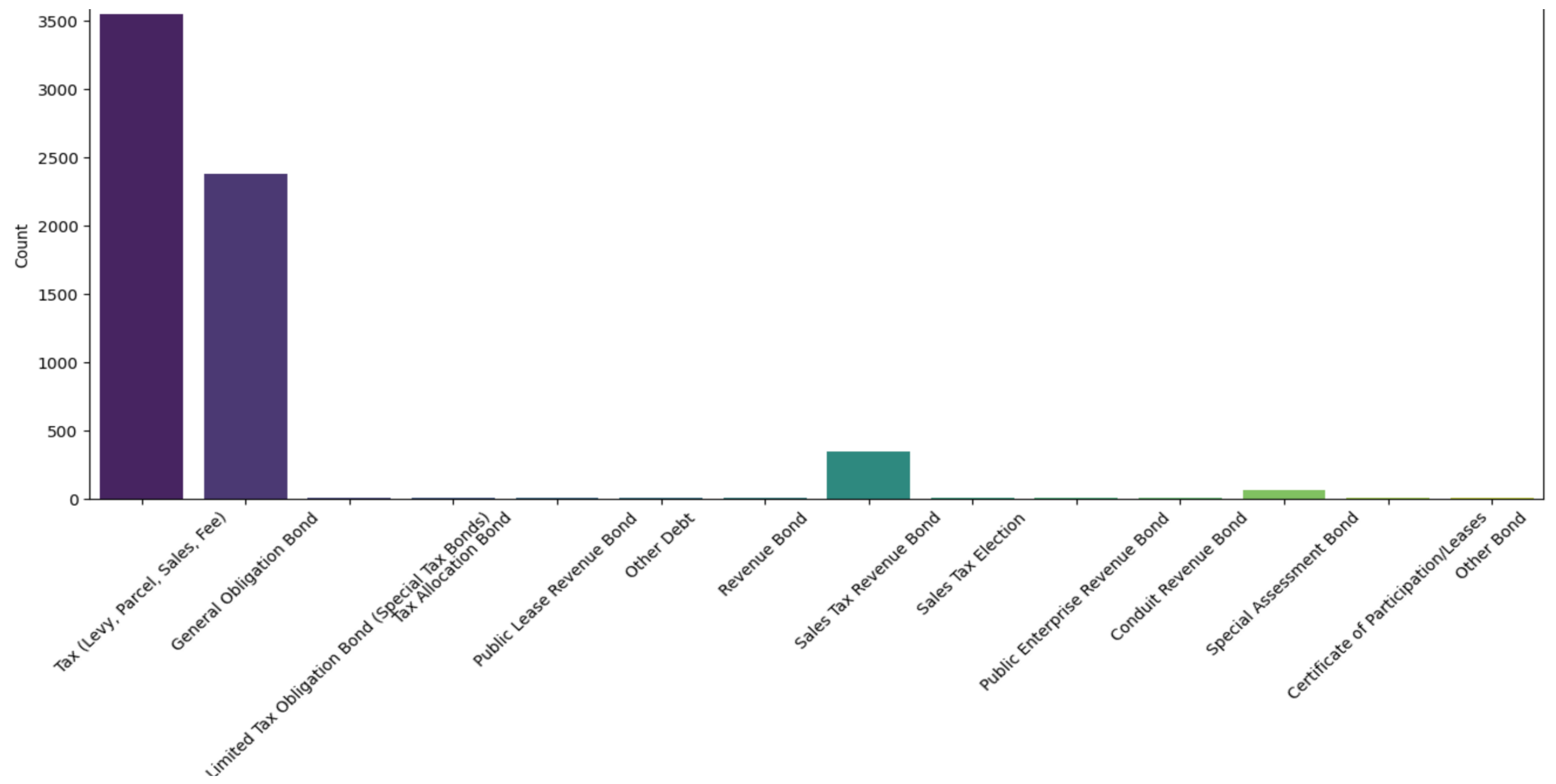
❑ **Performance Evaluation:**

- Assessed multiple models: Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, Support Vector Machines, Neural Networks, Naive Bayes, K-Nearest Neighbors, and Ensemble Methods.
- Key metrics: Accuracy, Confusion Matrix, and Classification Report.
- Evaluated overall correctness, classification details, discrimination ability, and precision/recall for legislative passage predictions.

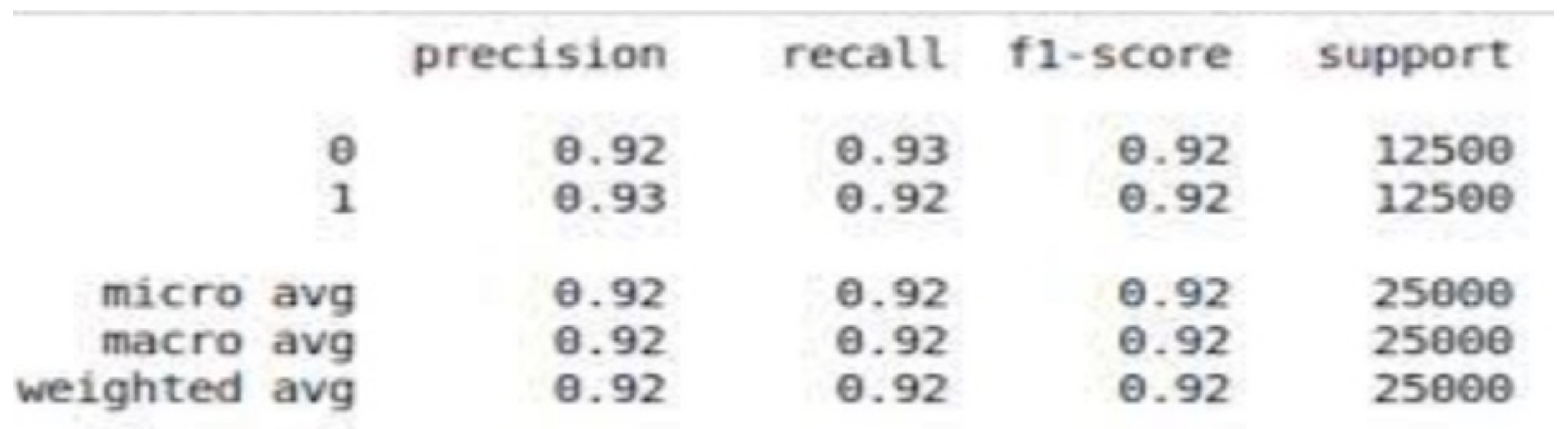
Note: These steps were executed to ensure robust model training, interpretability, and comprehensive evaluation of the classification models.

Results

- Highest legislative activity observed in 2022 compared to the previous years.
- California State stands out with the highest legislative data entries in our dataset.
- Legislative activities primarily clustered around sectors like Finance, Infrastructure, and Education.
- Top legislative contributors identified include major committees such as Budget, Revenue, and Appropriations.
- Legislative proposals show varying attention across different age groups and geographic regions.
- Random Forest model identified specific legislative features as crucial in predicting bill and tax proposal outcomes



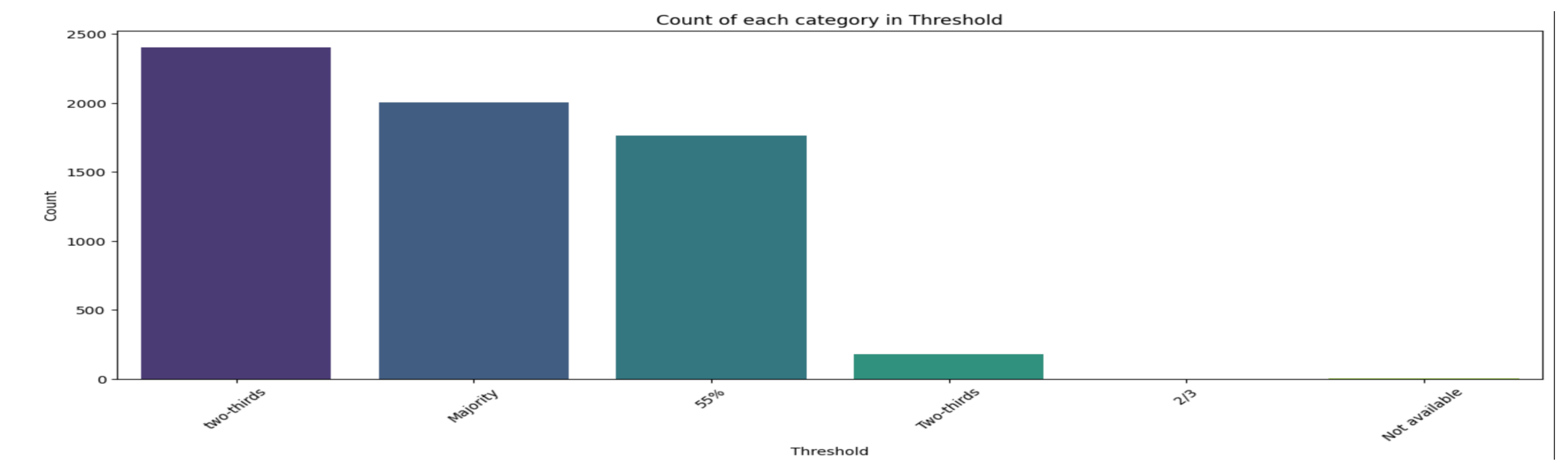
- We achieved an impressive F1 score of 92 for Logistic Regression and it demonstrated higher precision scores across both labels compared to the Random Forest model. Below are the detailed classification reports for the Logistic Regression.



```
              precision    recall  f1-score   support

           0       0.92      0.93      0.92     12500
           1       0.93      0.92      0.92     12500

   micro avg       0.92      0.92      0.92     25000
   macro avg       0.92      0.92      0.92     25000
weighted avg       0.92      0.92      0.92     25000
```



Limitations

- Unforeseen alterations in tax policies or laws at the state or federal level can significantly affect the accuracy of predicted tax outcomes.
- Economic uncertainties, regional variations, and unforeseen events may challenge the model's ability to accurately predict tax bill outcomes for the diverse regions of California.

Conclusions & Future work

Conclusion:
The tax bill outcome project offers valuable predictive insights for the California State Department. However, the model's reliability is contingent on addressing challenges like legislative changes and economic uncertainties. Interpretation caution is crucial given the dynamic landscape.

Future Work:
Enhancements should include real-time legislative monitoring, refined economic modeling, localized predictions, dynamic data integration, consideration of external factors, stakeholder collaboration, and user-friendly interfaces. By refining these aspects, the model can evolve into a more robust tool, adapting to real-world complexities and fostering effective decision-making within the dynamic economic and legislative context of California.

References

- Sentiment Analysis on Twitter using Neural Network: Indonesian Presidential Election 2019 Dataset . Ahmad Fathan Hidayatullah *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1077** 012001
- Forecasting stock market movement direction with support vector machine. Institutions (2)30 Sep 2005-(Elsevier Science Ltd.)-Vol. 32, Iss: 10, pp 2513-2522. link : https://typeset.io/papers/forecasting-stock-market-movement-direction-with-support-3ls6z5dgj2