**Project Outline:**

Our project focused on analyzing and predicting employment and unemployment trends in California, with a population nearing 40 million. Employment trends have significant implications for California's economy, lawmakers, and residents considering their future in the state. This analysis aimed to predict unemployment rates using the features in the dataset provided while considering whether unemployment trends exhibit self-exciting behavior, potentially benefiting from models like Recurrent Neural Networks (RNNs).

**Data Information:**

The dataset showed that unemployment rates had a highly skewed distribution, with most values concentrated at the lower end and a long tail for higher unemployment rates, indicating an imbalance where most areas experienced lower unemployment. Employment rates were also skewed, with values concentrated close to 1, reflecting high employment and low variability, which could pose challenges for predictive modeling. The labor force distribution was similarly skewed, with a concentration near the lower end but containing outliers representing areas with extremely large labor forces that required special handling. Most data points belonged to "Sub-County Place," creating a potential imbalance by over-representing sub-county characteristics while under-representing counties and metropolitan areas. Additionally, there was an imbalance in seasonality and status labels, with most data labeled as "Final" and "Non-seasonally adjusted," potentially introducing bias into the model.

We also viewed the data in terms of heatmap visualizations which showed strong positive correlations between “Labor Force”, “Employment”, and “Unemployment”. There was also a strong negative correlation between “Unemployment Rate” and “Employment Rate”. These correlations highlighted potential multicollinearity among variables, which made it extremely important we performed hyperparameter testing and used different optimizers throughout the process.

**Model Selection and Testing:**

We start by utilizing our baseline model which was a Linear Regression model. We used it in an attempt to predict the unemployment rate on past data and replicate those predictions. During preprocessing, non-numeric entries were converted, and missing values were replaced with zeros. StandardScaler was applied to normalize features, as this improves model performance. The dataset was split into an 80-20 ratio for training and evaluation. The model achieved an MSE of 1.79 and an R-squared value of 0.97, meaning it explained 97% of the variance in the unemployment rate. While this provided a reasonable starting point, the low R-squared value indicated the need for more complex models to capture the trends and variability in the data. More importantly, it indicated that there was a need for a more advanced model to capture the intricacies of the data especially if one wanted to make predictions. Multiple runs comparing Adam and SGD optimizers revealed that Adam consistently achieved better accuracy and lower test losses, solidifying it as the preferred optimization method for further modeling. Given the hypothesis that unemployment trends might be self-exciting, RNNs were considered for capturing sequential dependencies and dynamic patterns in unemployment rates. Initial results suggested that RNNs could better model these trends than static models like linear regression.

**Challenges and Observations:**

The skewed distributions and dominance of "Sub-County Place" and "Non-seasonally adjusted" data required thoughtful preprocessing and careful modeling to avoid biases. Large labor force values presented challenges, necessitating special handling or transformations to ensure they didn’t disproportionately impact model training. High employment rates and limited variability posed challenges for creating robust predictive models, particularly for linear regression. In addition, there were a lot of data points which consistently caused us to revise the training to train within a reasonable time frame to make sure the process was replicable.

**Key Takeaways:**

Ultimately, we observed that an RNN was clearly the correct tool to solve the problem as it accurately captured the dynamism of the problem and its associated features. The baseline model was able to achieve moderate performance with linear regression, establishing a starting point for further experimentation, but clearly RNNs were necessary to be able to perform predictions and perform proper analysis of the data. Optimizers like Adam and SGD offered better potential for capturing complex patterns in the data. Feature selection and scaling as well as constant hyperparameter tuning were critical to managing multicollinearity and improving model performance. This was a very key takeaway for us because it showed the importance of optimally tuning each neural network parameter to optimize training and optimize accuracy as best as possible.

**Recommendations for Future Work:**

Future work should explore RNNs, GRUs, or LSTMs to fully leverage sequential dependencies in unemployment trends. Addressing data imbalances using techniques like oversampling, undersampling, or reweighting can help correct biases in area types and seasonally adjusted data. Handling outliers through transformations (e.g., log scaling) will reduce their disproportionate influence. Adding external features such as economic indicators or policy changes could improve predictive power, and fine-tuning preprocessing techniques like PCA can address multicollinearity. Moreover, we could always go back and do a super deep assessment of each individual feature to perfectly optimize our prediction.

**Conclusion:**

This project successfully established a baseline for predicting unemployment rates in California and highlighted the complexities of employment data. The work provides a strong foundation for future research and development of more sophisticated models to support California’s economic decision-making and policy planning.