# **Group 21 - Fifer**

# **California Unemployment Rate Prediction**

### Team Members:

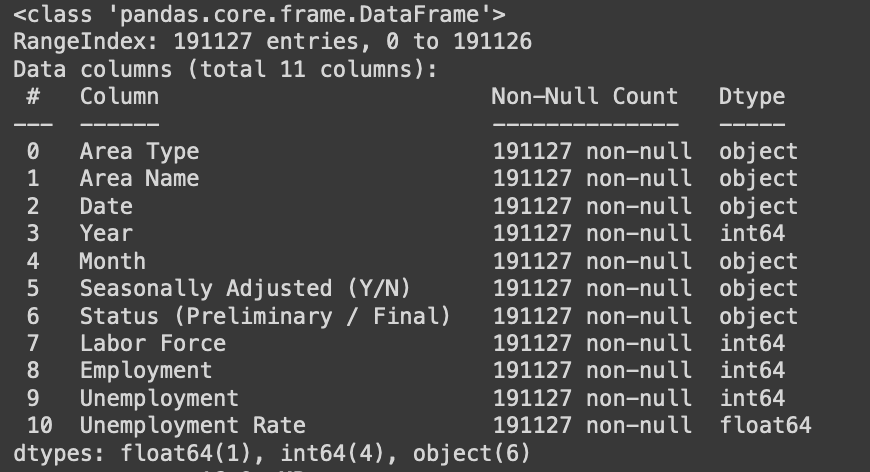
| [Pritham Sriram Govindaraj](mailto:psg4@illinois.edu) | psg4 |
| --- | --- |
| [Suraj Vashista Bettadapura Krishna](mailto:sb70@illinois.edu) | sb70 |
| [Saiharshith Karuneeger Ramesh](mailto:sk120@illinois.edu) | sk120 |
| [Maahi Patel](mailto:maahidp2@illinois.edu) | maahidp2 |

### **Data info:**

We are considering the California unemployment dataset from the California open data portal which can be found in the link below:

<https://data.ca.gov/dataset/local-area-unemployment-statistics-laus/resource/b4bc4656-7866-420f-8d87-4eda4c9996ed>

The given dataset contains 11 columns and 191,127 rows.



Understanding more about each of the columns:

* **Area Name:** It contains the names of areas in California like Los Angeles, Long Beach and so on. It has 1050 unique area details.
* **Area type:** Describes the category of the area based on size it can be categorized into 6 types 'State', 'County', 'Metropolitan Area', 'Sub-County Place', 'Metro Division', 'MSA’
* **Date, Month, Year:** The date in which the data was recorded. It ranges from 01/01/1970 to 08/01/2022
* **Seasonally Adjusted:** it holds ‘Y’ or ‘N’ indicating if the data is seasonally adjusted.
* **Labor Force:** Used to describe the No. of labor at that given time.
* **Employment:** The number of people employed out of the labor force
* **Employment rate:** rate of employment out of labor force
* **Unemployment:** The number of people unemployed out of the labor force
* **Unemployment rate:** rate of unemployment out of labor force

### **Data Preprocessing:**

* The data provided in the data set did not contain any null values
* The date columns of the given dataset were converted into pandas timestamps for easier processing.
* We excluded rows that had a labor force equal to zero.
* Unnecessary spaces in column names were trimmed and columns that had zero values were removed.
* Finally, we save two versions of the dataset. One, complete data as a pickle file and second, small enough to test code with.

### Exploratory data analysis:

Upon performing exploratory analysis on the preprocessed data, we made the following observations:

* The unemployment rate showed a general decreasing trend over the passage of time, other than a couple of outliers of which 1 was near 2008 (which might have been due to the 2008 recession).
* The Metropolitan Area has the highest labor force and it was lowest for metro division and subcounty places.
* The unemployment rate was particularly high during July and October of each year.
* Sub county places had the highest unemployment rate along with the lowest labor force.
* The seasonally adjusted column had very less “N” value unevenly distributed and does not add any significance for the data. Hence, this was removed.
* Finally, we conclude that the columns : Year, Area Name, Area Type, labor force and employment have higher importance in making an informed decision.

### Linear Regression:

Inorder to get a brief idea and set benchmarks, we perform a linear regression on the dataset. Based on the correlation values among each column, we used Year, Labor Force and employment as input of the training data. Unemployment rate if our output.

We find that the data is overfitting with these columns and we get an overall mean squared error of 0.04

### Our Approach:

In our approach, we plan to use the LSTM time series forcasting to check how the unemployment rate has varied over time. We have experimented with the dataset 3 optimizers Namely RMSprop, Adam, Adagrad.

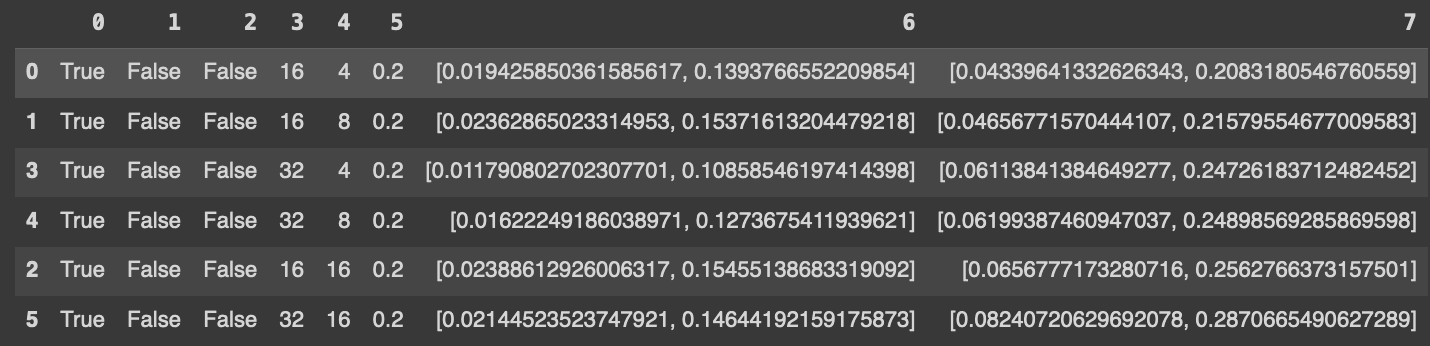
We performed hyperparameter tuning by experimenting with combinations of (16, 4), (16, 8), (16, 16) and (32, 4) neurons and batch size and with a dropout rate of 0.2.

## **Adam Optimizer**

## 

## **Adagrad optimizer**

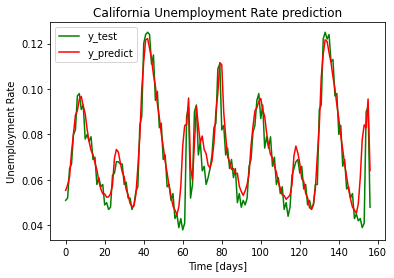
## **RMSprop optimizer**



### Conclusion

Upon experimenting different hyperparameters and optimizers, we found that the model obtained with the combination of adam optimizer with a batch size of 16 run over 45 epochs produces a loss of 0.0078

Comparing the predicted and the actual value on the test data for the unemployment rate:



The green line represents the ground truth and the red line represents the actual value.

The RMSE of the test data for the final model was 0.00964