Unemployment Rate

Nicholas Morris

Machine Learning

Wrangling → Feature Engineering → Modeling

Monthly Unemployment

Adding previous months of unemployment to the data by area.

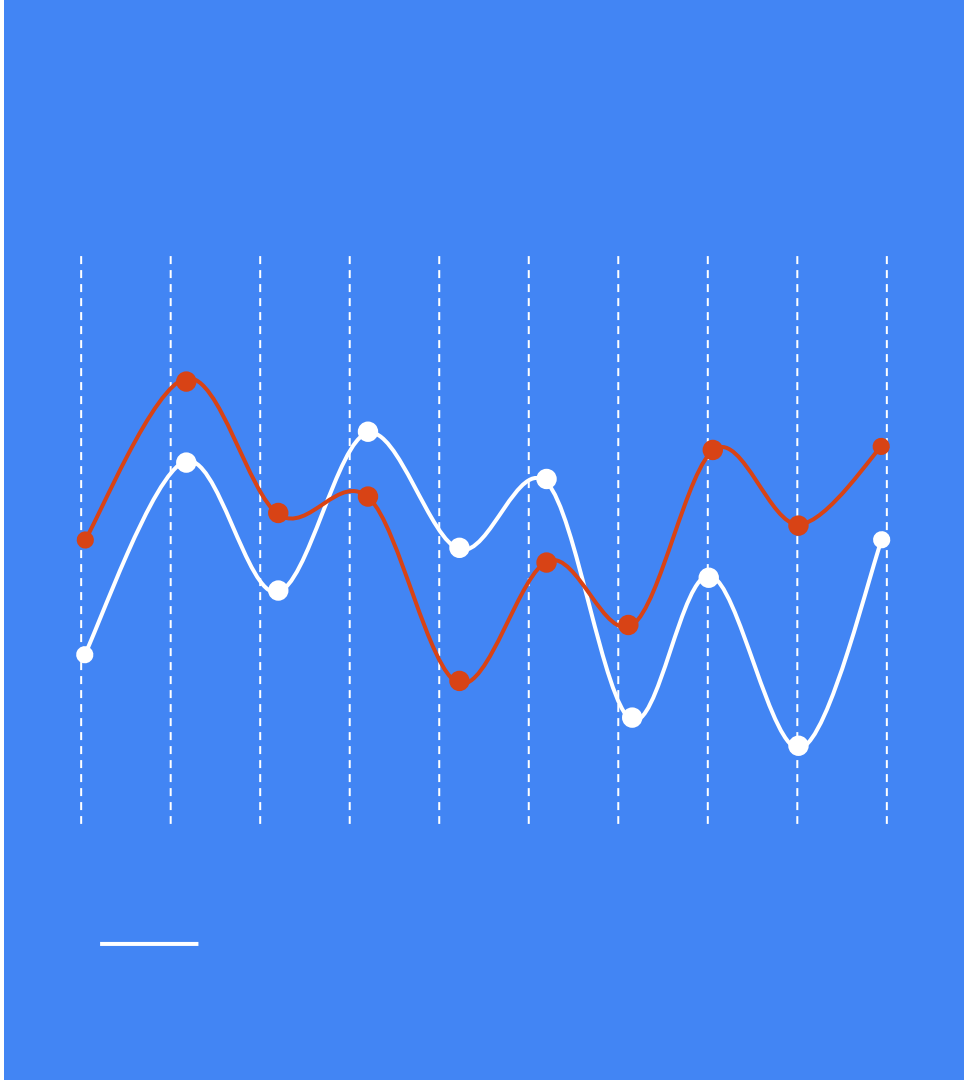
Additional Data

Adding economic trends such as unemployment and CPI. Converting timestamp components and area to binary data points. Attempting feature transformations.

Predictions

Training linear regression, XGBoost, and deep learning neural network models. Evaluating performance. Computing feature drift to signal retraining.

Wrangling



Dataset

Below is the first two entries in the data. There are 200,631 total entries. The target we are predicting is Unemployment Rate. [Link to the dataset]

Area Type	Area Name	Date	Year	Month	Seasonally Adjusted (Y/N)	Status (Preliminary / Final)	Labor Force	Employment	Unemployment	Unemployment Rate
State	California	01/01/1976	1976	January	N	Final	9672362	8668016	1004346	0.104
State	California	01/01/1976	1976	January	Y	Final	9774280	8875685	898595	0.092

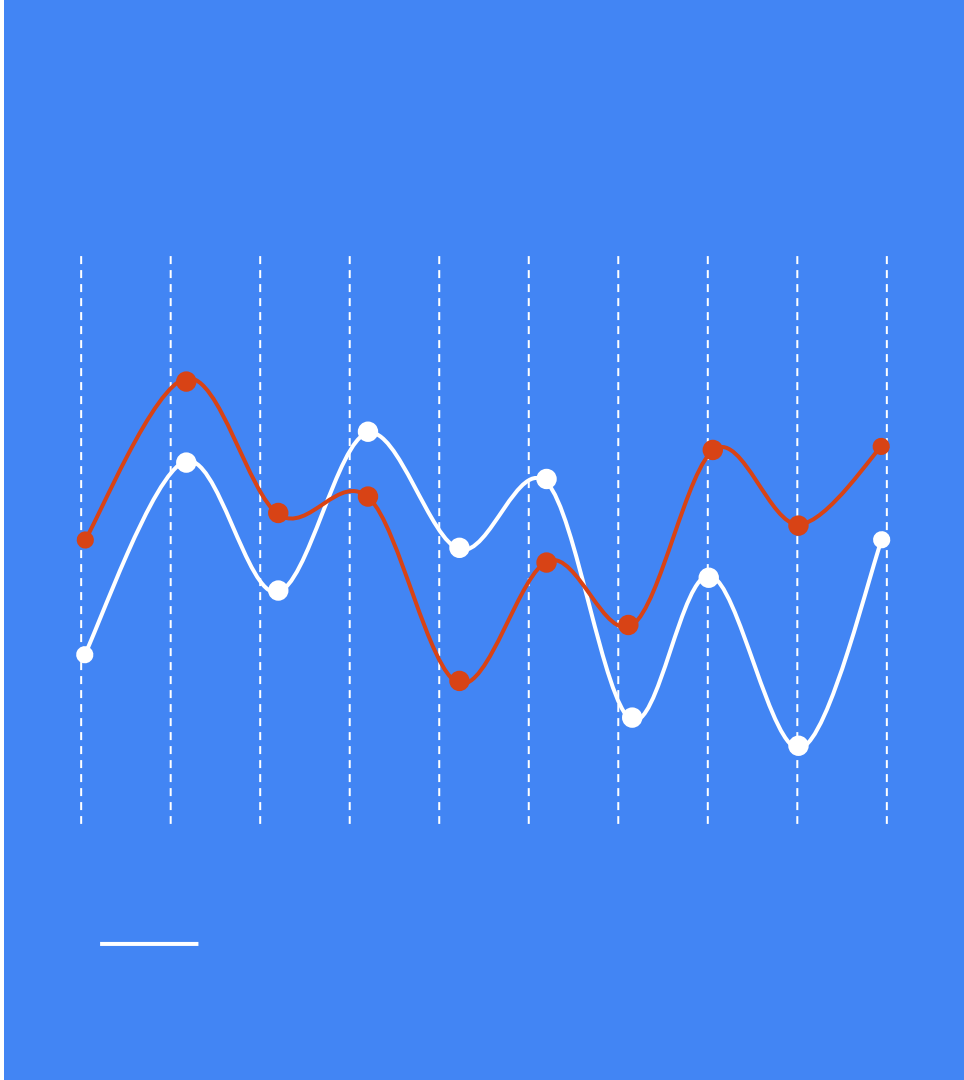
Previous Months Of Unemployment

The last four months of unemployment for each area was added to the data.

Removing Data

Employment volume and unemployment volume were removed from the data because they leak the unemployment rate.

Feature Engineering



Binary Data

Area Type, Area Name, timestamp components, Seasonally Adjusted, and Status were all converted to binary data points.

Atwood Numbers

An Atwood Number is a calculation that shows the relative change between two variables. The formula for two variables x and y is:
$(x - y) / (x + y)$

This calculation was done on all pairs of non-binary variables; but did not improve model performance, so, it was left out of the final model.

Binning

Binning is when a non-binary variable is grouped into histogram bins, and represented as binary variables.

Binning did not improve model performance, so, it was left out of the final model.

Reciprocals

A reciprocal is when a non-binary variable x is calculated as 1 / x.

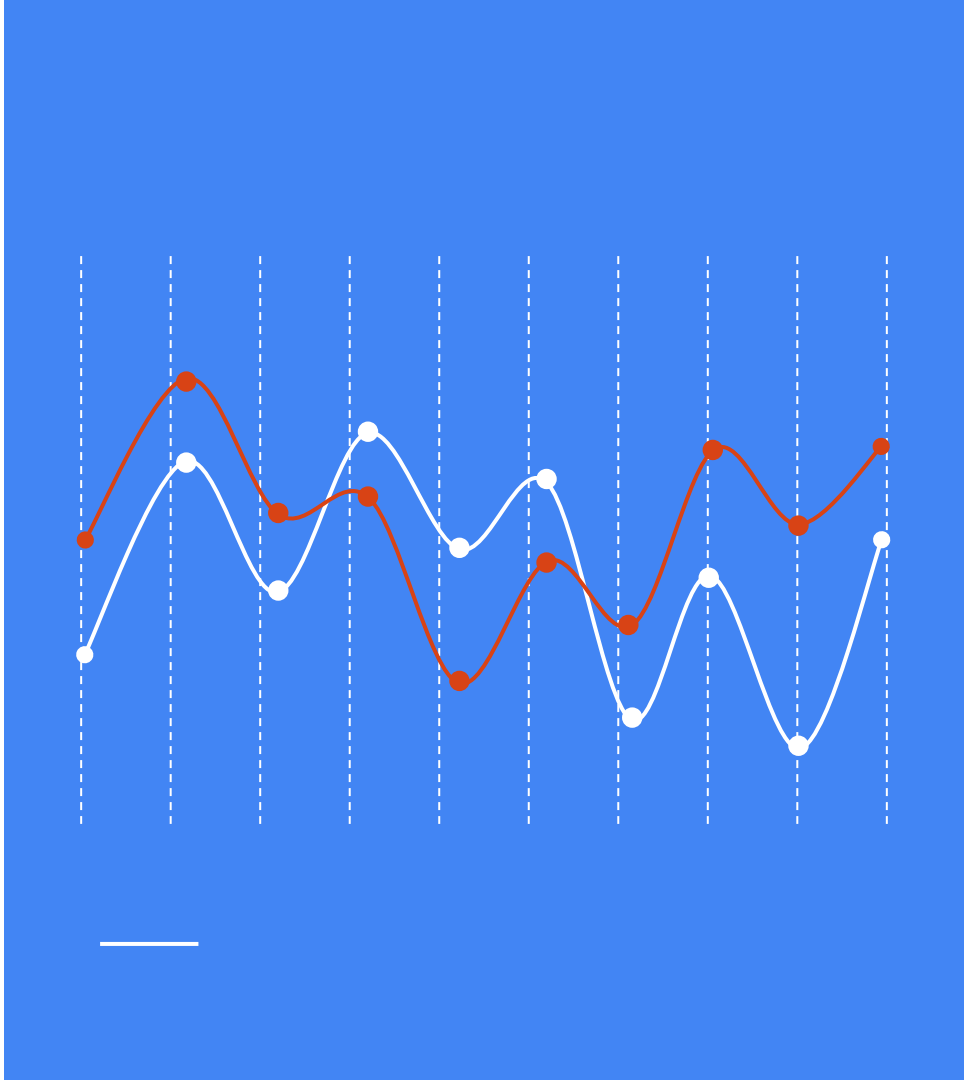
Reciprocals did not improve model performance, so, it was left out of the final model.

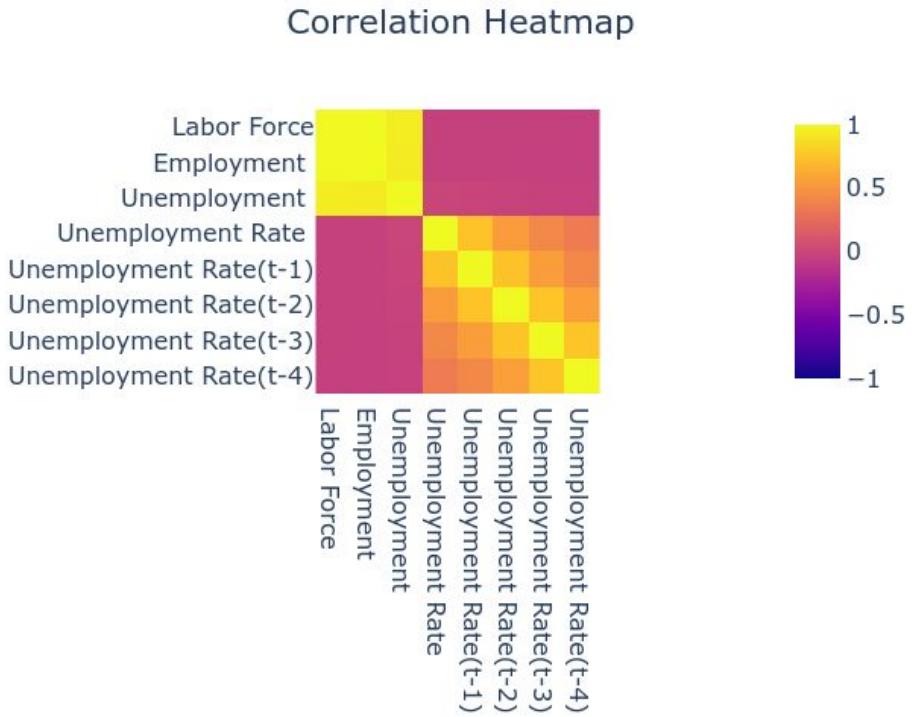
Interactions

An interaction is when two variables x and y are calculated as x * y. Reciprocals were fed into this calculation to generate x / y as well.

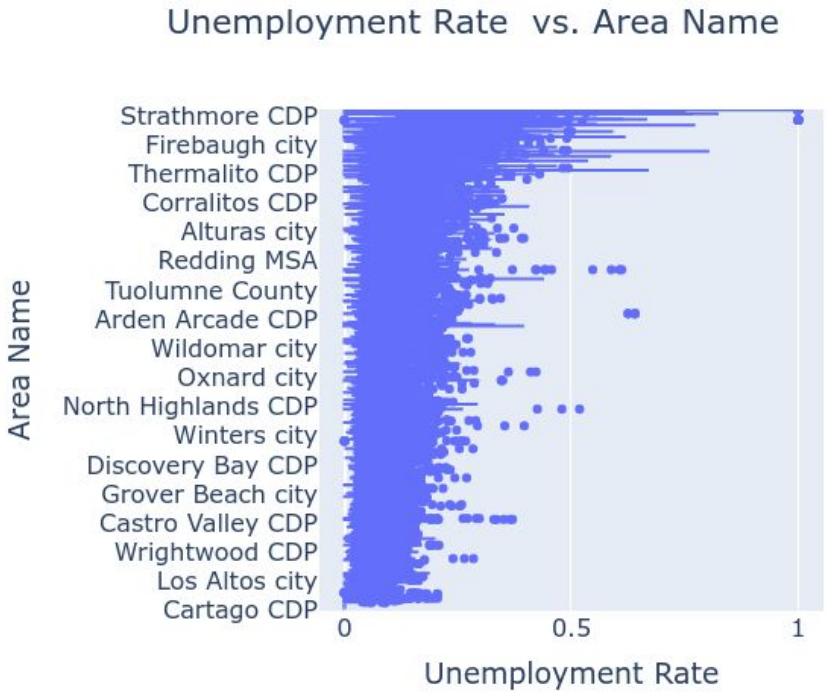
Interactions did not improve model performance, so, it was left out of the final model.

Data Exploration

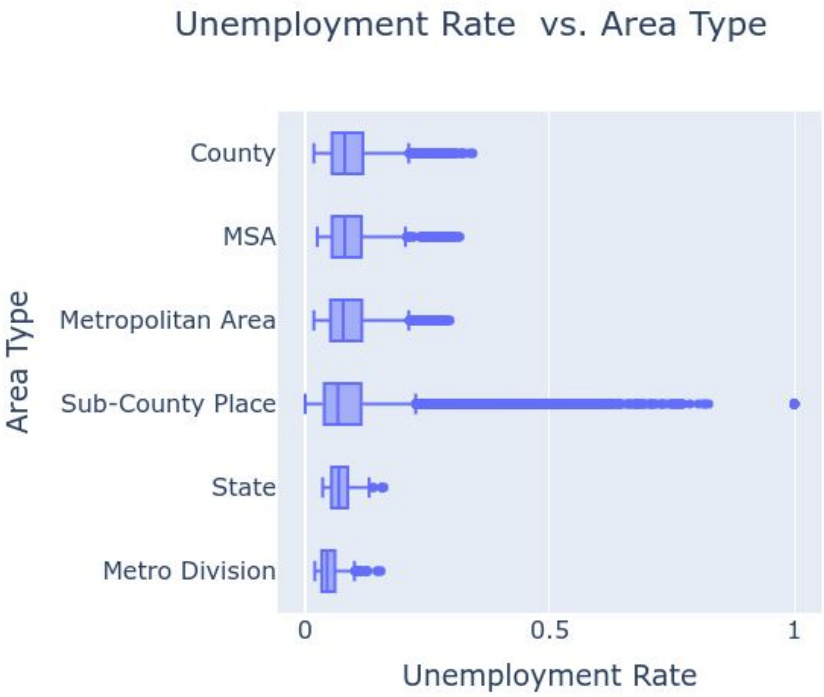




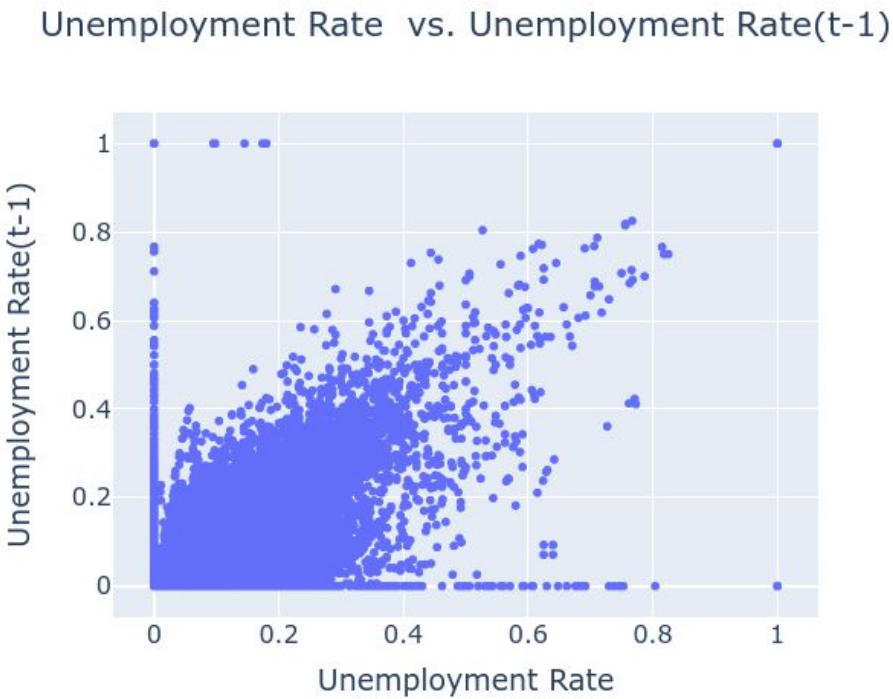
There are two zones of correlations. The first is between the labor force and the number of employed and unemployed people. The second is between the unemployment rate and the previous months of unemployment.



We can see that the unemployment rate varies by area with Strathmore CDP having the highest average rate and Cartago CDP having the lowest.

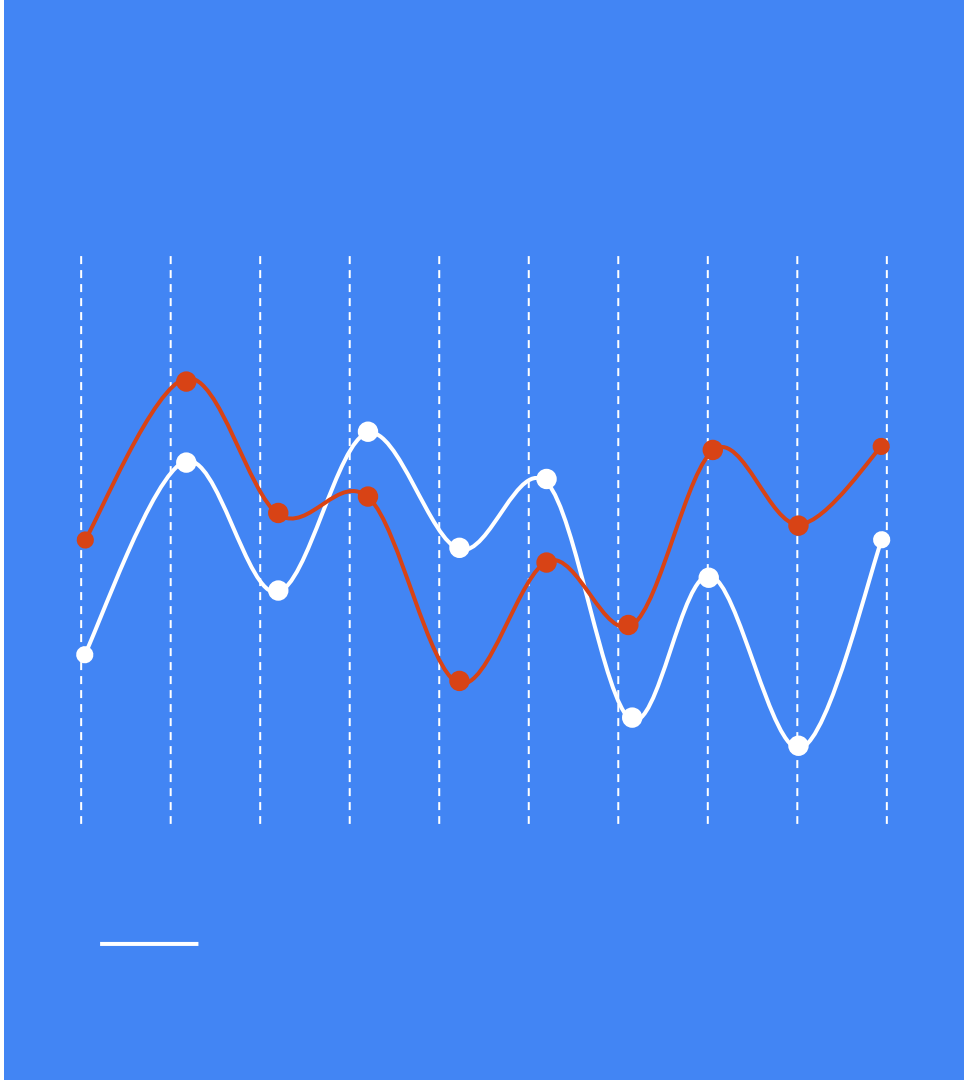


The unemployment rate is highest in County with a median of 0.081 and lowest in the Metro Division with a median of 0.045.



There is some positive trend showing that the unemployment rate follows the previous month of unemployment; but there are many outliers on the edges.

Modeling



Model Parameters

Linear Regression

Library: scikit-learn
Length Of Path: 1e-9
Number Of Alphas: 16
Cross Validation Folds: 3
Tolerance: 1e-4
Max Iterations: 500

XGBoost

Library: xgboost
Boosting Rounds: 100
Learning Rate:
0.001, 0.01, 0.1
Max Depth:
5, 7, 10, 14, 18
Min Child Weight: 1
Column Sampling: 0.8
Row Sampling: 0.8
Cross Validation Folds: 3

Neural Network

Library: Tensorflow
Epochs: 500
Learning Rate:
0.0001, 0.001, 0.01
Batch Size: 16
Layers: 10
Nodes Per Layer:
32, 64, 128, 256, 512
Solver: Adam
Cross Validation Folds: 3

Model Comparison

Linear Regression

R2: 0.67
RMSE: 0.04
In Control: 95.04%

Model Indicators:
1. Unemployment_Rate(t-1)
2. Prattville_CDP
3. Chilcoot_Vinton_CDP
4. Richgrove_CDP
5. Unemployment_Rate(t-4)

XGBoost

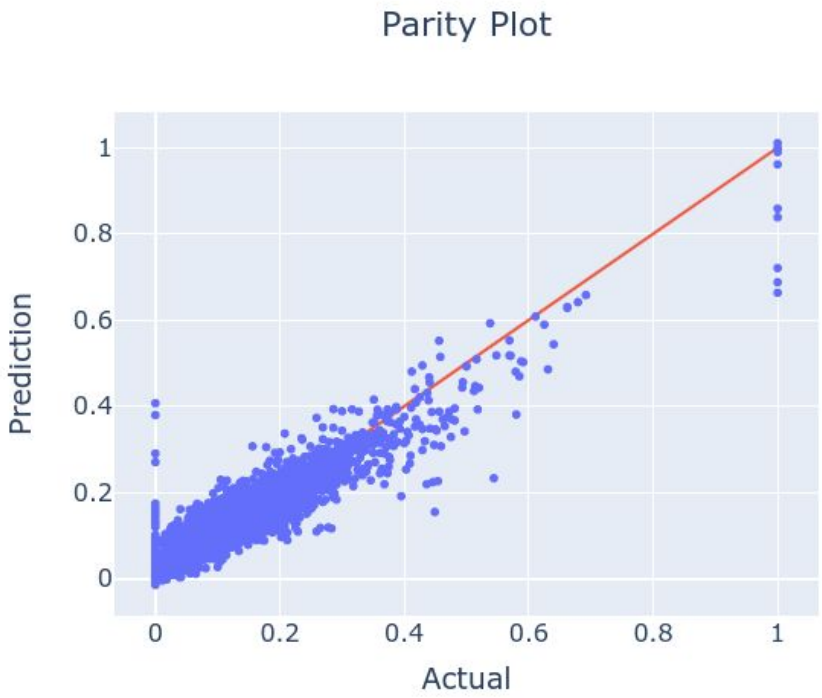
R2: 0.89
RMSE: 0.02
In Control: 94.53%

Model Indicators:
1. Date_year_2010
2. Prattville_CDP
3. Chilcoot_Vinton_CDP
4. Unemployment_Rate(t-1)
5. Canyondam_CDP

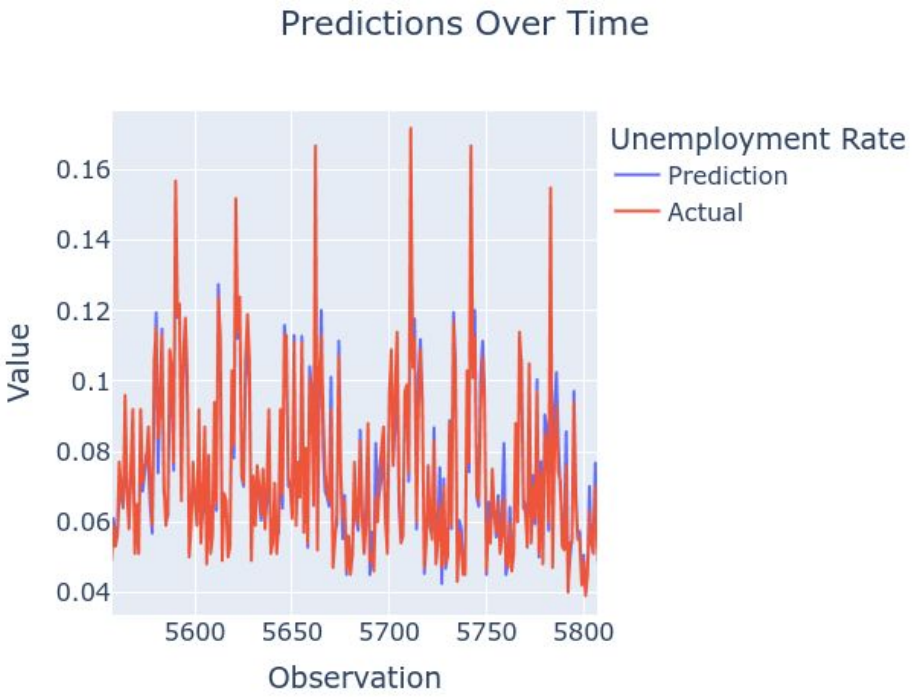
Neural Network

R2: DNF
RMSE: DNF
In Control: DNF

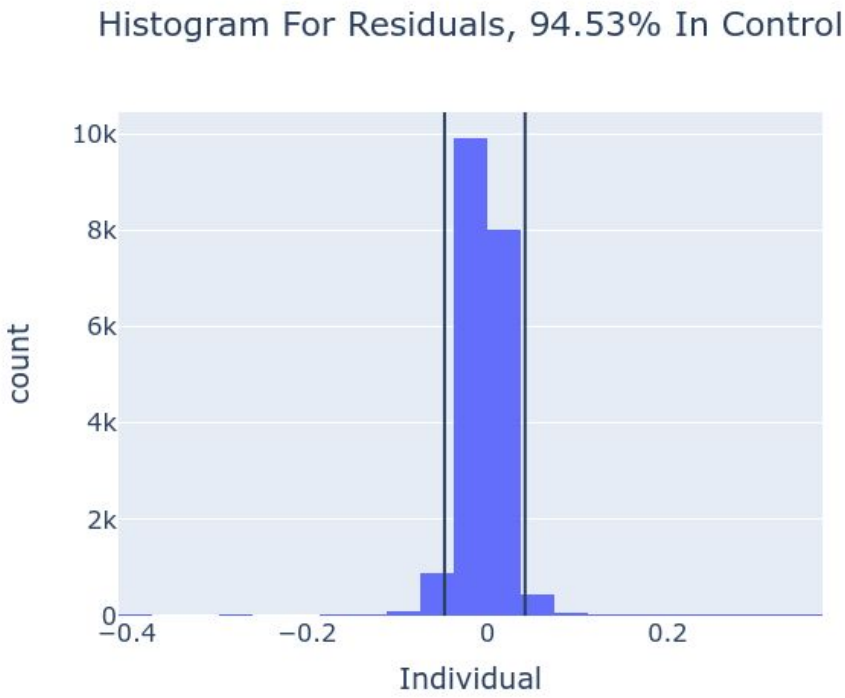
Model Indicators:
DNF



These predictions come from the XGBoost model. These predictions are done on 20% of the data that the model did not see during training. The predictions are centered on the red line (perfect predictions). There is a tendency to under-predict the unemployment rate. There's also some outliers at both ends of the plot where a rate of 0 is over-predicted and a rate of 1 is under-predicted.

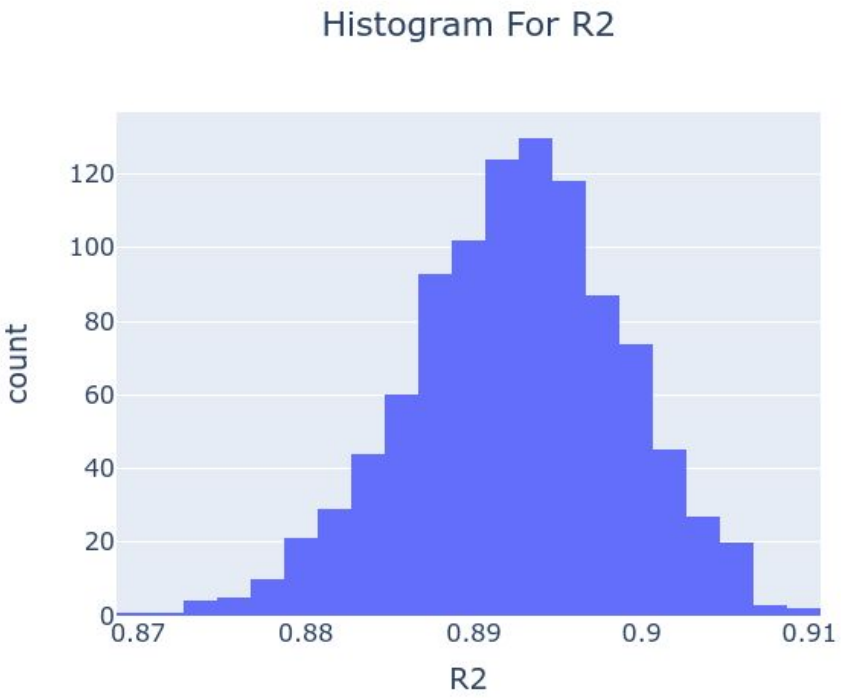


Here’s a snapshot of the predictions over time. We can see the the blue predictions follow the actual values well. There is some extreme values where the red line jumps up well past the predictions. We can see there is some seasonality in the data as well.



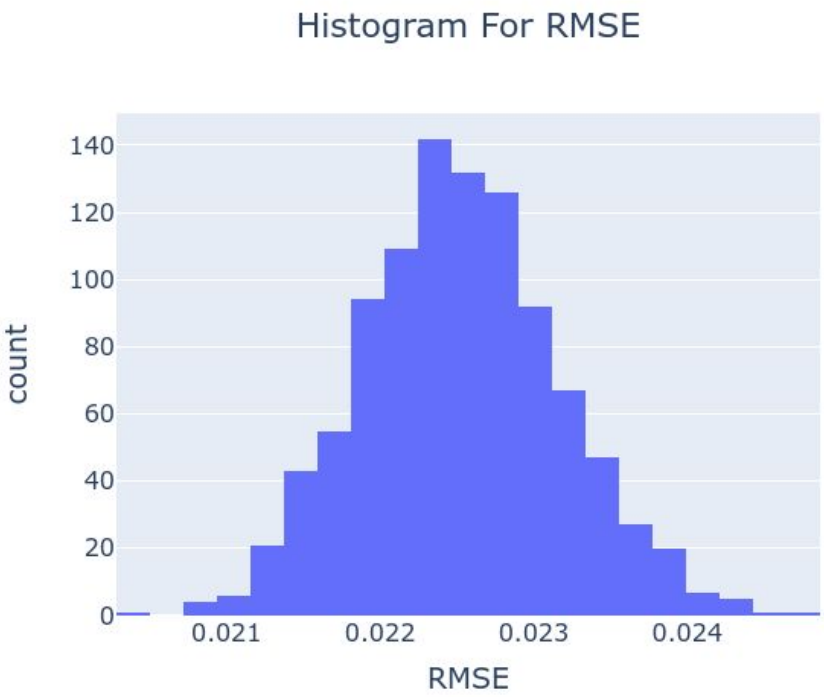
The residuals are prediction error = actual - predicted.

The residuals have a tight bell shape, which is good, and they are centered on zero. Control limits were computed on the residuals and we can see that the prediction error is mostly under control. We see can long tails which means the model has a small tendency to over and under-predict unemployment by a large margin.



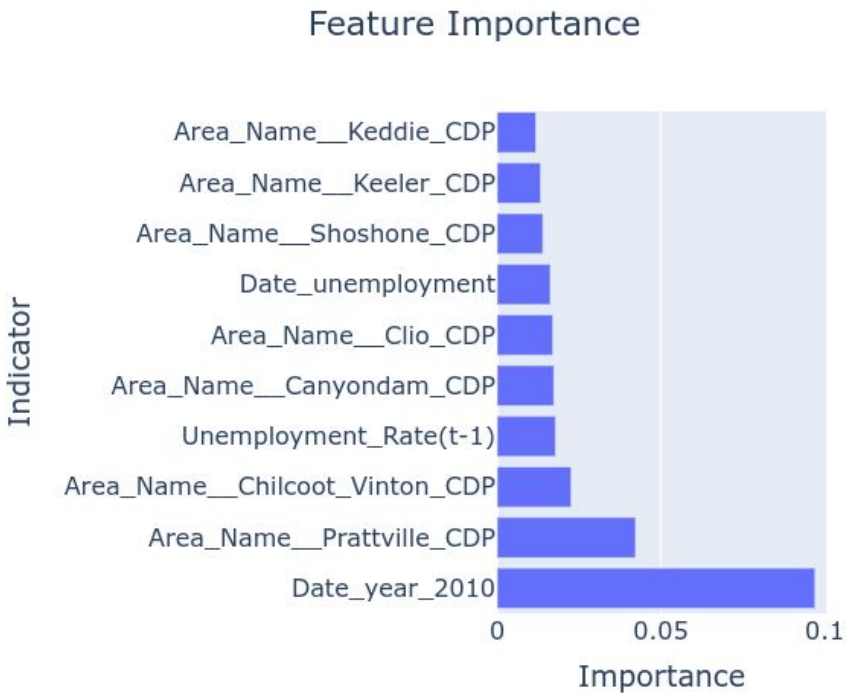
The prediction error was resampled 1000 times at a 50% sampling rate with replacement. Then R2 was computed on each sample to get a distribution.

R2 has a tight range between 0.87 and 0.91, which is good. R2 has a bell shape, which is good.

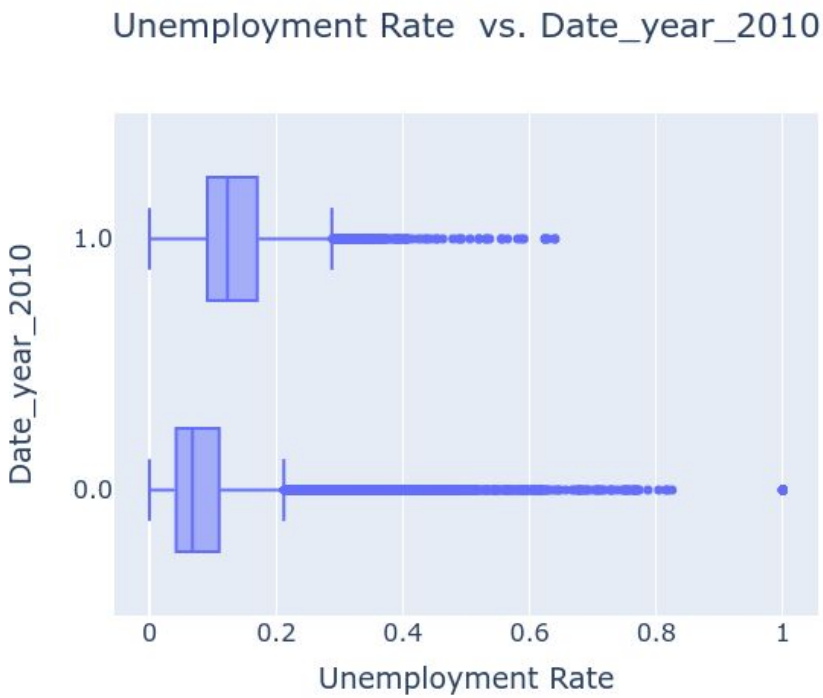


The prediction error was resampled as previously mentioned to get a distribution for RMSE.

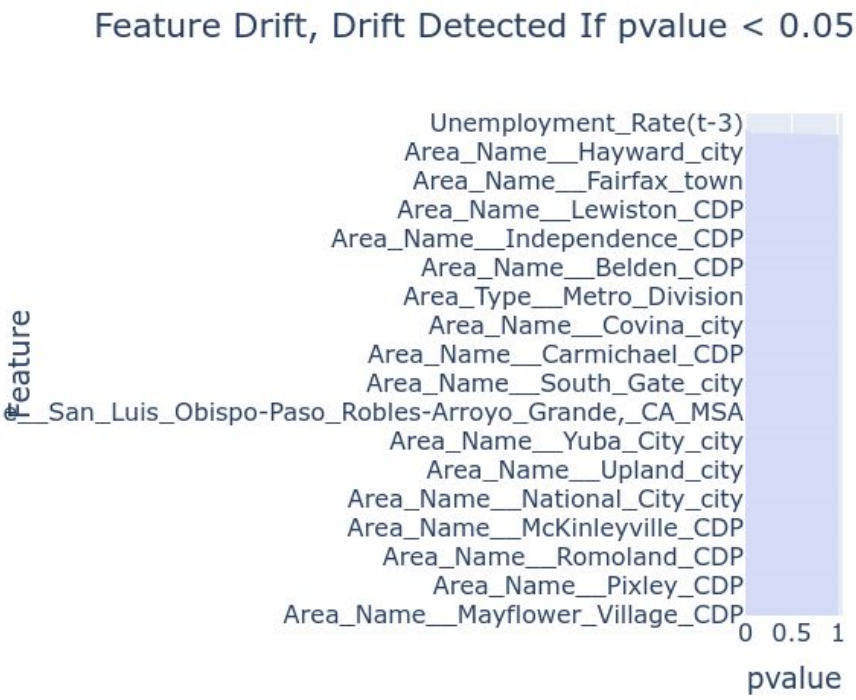
On average, the predictions are off by 0.021 to 0.024, which is a tight range. The RMSE also has a bell shape, which is good.



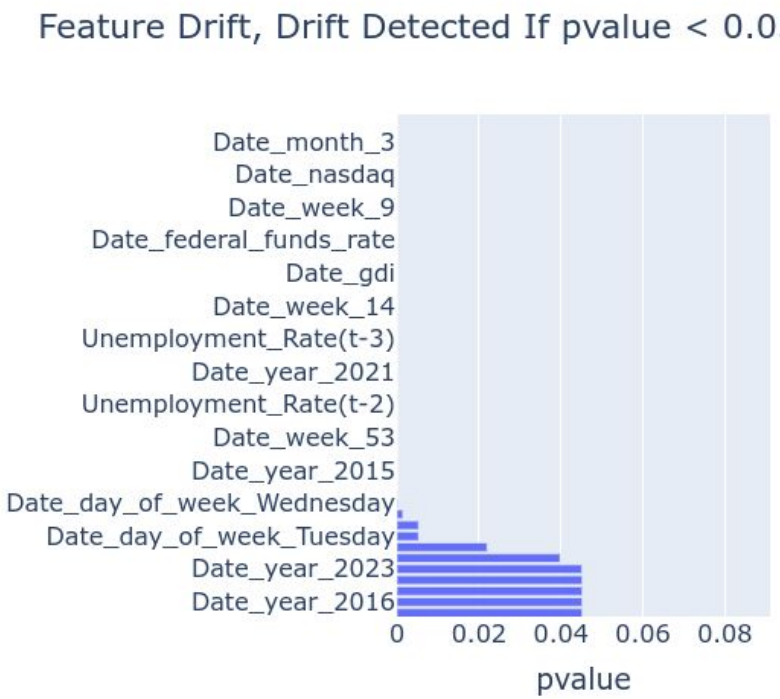
Here's the top ten most important indicators of unemployment rate. We can see that the year 2010 is much more important than the other features, followed by Prattville_CDP.



We can see the unemployment rate is much higher in 2010 than in other years. In 2010, the median unemployment rate was 0.123 and in other years it was 0.068.

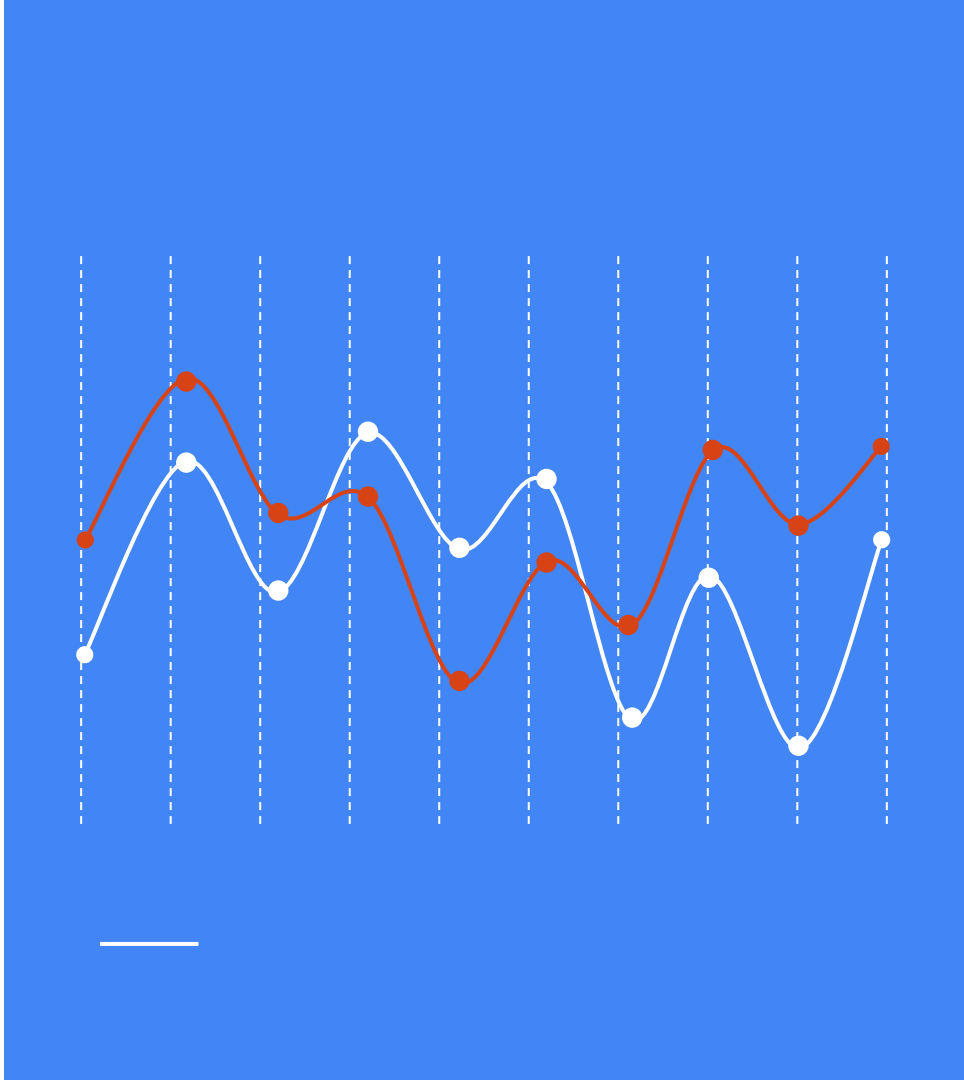


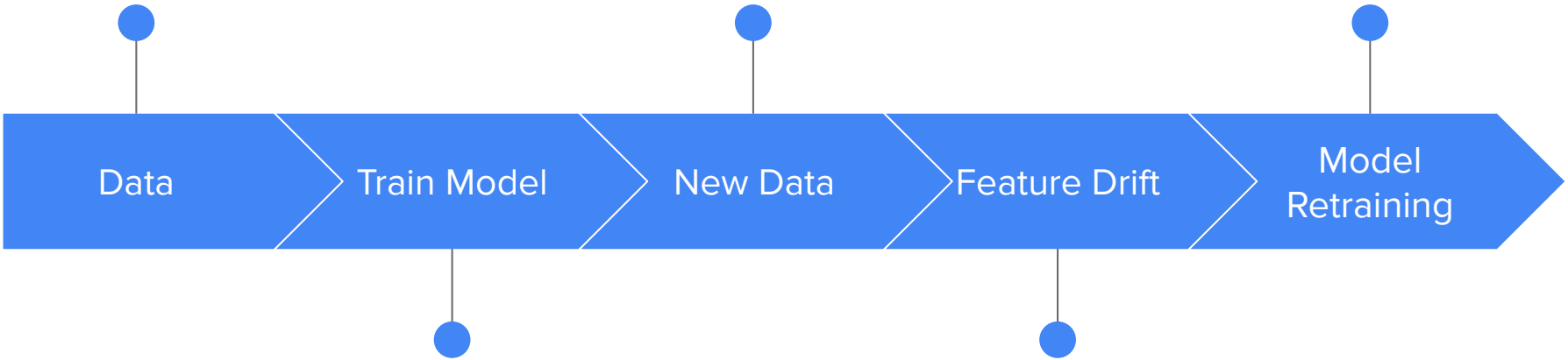
A Kolmogorov-Smirnov test was performed for each column in the data to see if the distribution of the testing data is the same as the training data. If the testing data does not share the same distribution as the training data, then there is a drift, which signals for model retraining. Most of the columns do not experience a drift, which is good.



The features that experience a drift are the timestamp components, unemployment rate, and economic indicators. This is because time is moving forward, unemployment shifts from year to year, and economic indicators have a trend over time.

Deployment




The data we start with.
The latest data we want predictions for.
Retrain the model on the initial data and new data.
Data
Train Model
New Data
Feature Drift
Model Retraining
Data wrangling, feature engineering, model training.
See if the distribution of the new data is significantly different than the initial data.

Thank You