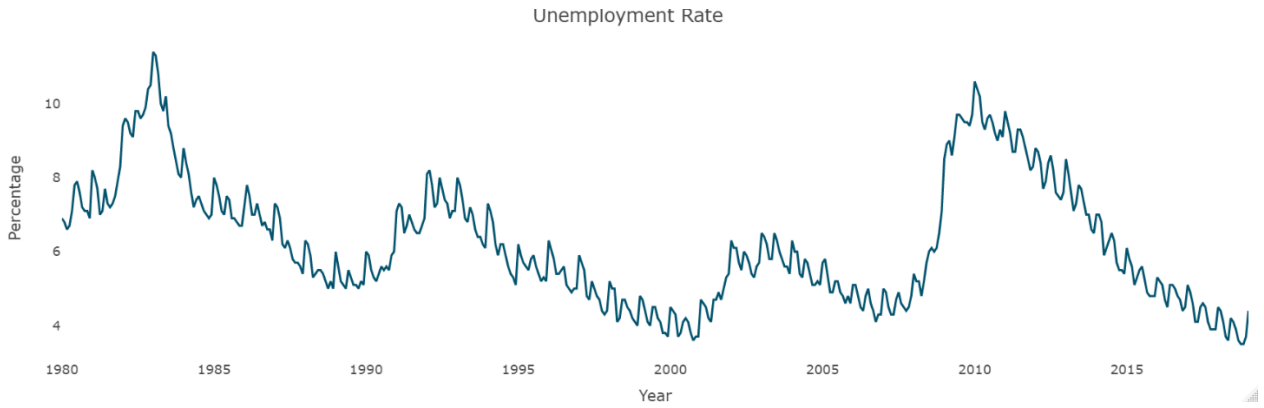
Random Forest (RF), Gradient Boosting Machine (GBM)

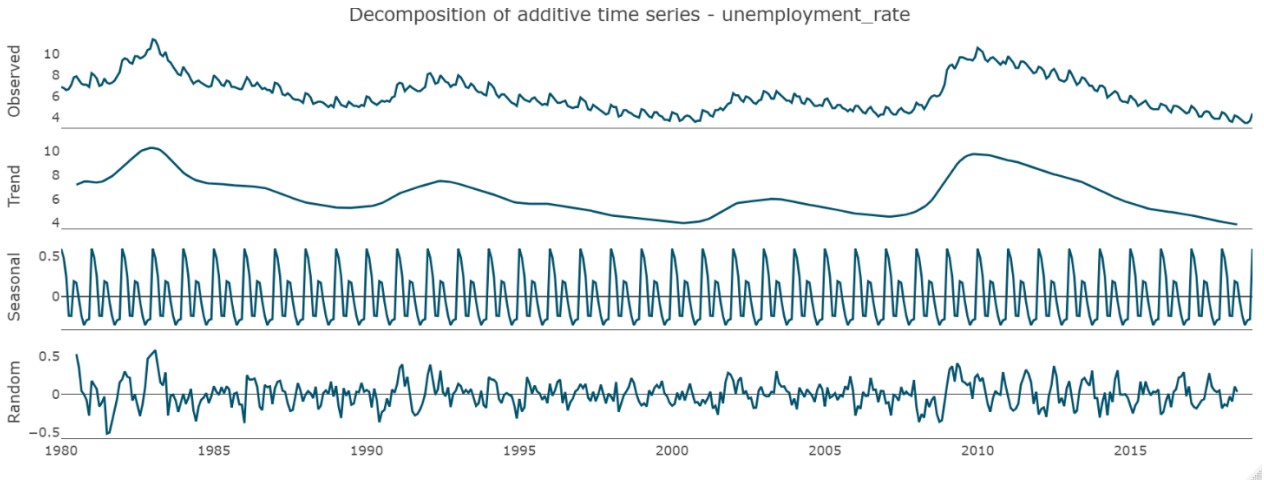
Name: Ellen Wei

Feature engineering

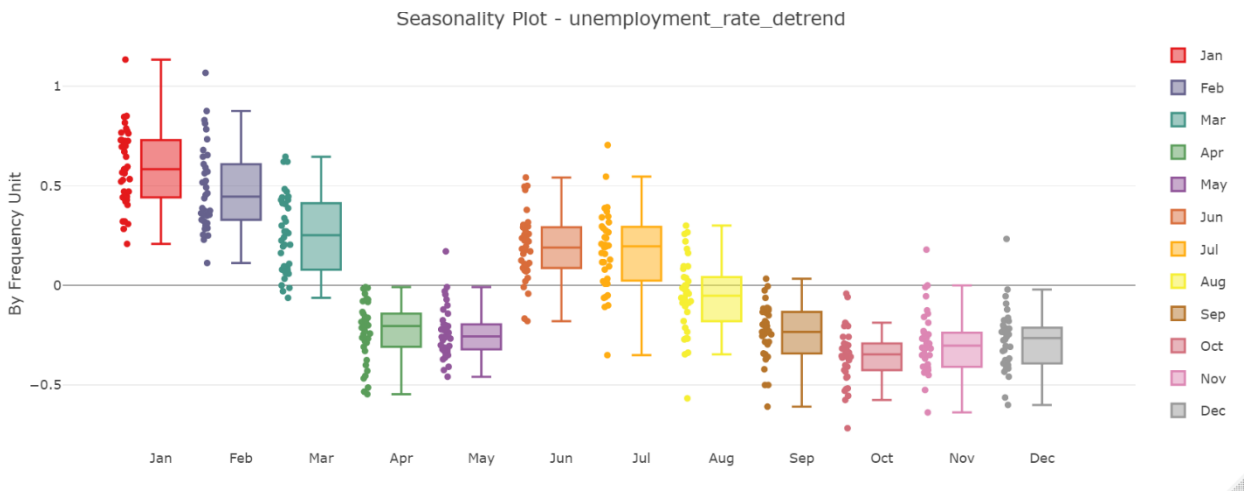
The features lag12, month, and trend were used to create the benchmark regression model, the Random Forest model and the GBM model. After an initial plot of the data, there was a trend observed as seen in Plot 1. From Plot 2, we confirm that there appears to be a trend component and seasonality component which is why lag12 and trend were included as variables for the models. Plot 4 includes plots of the lags at 12, 24, 36 with the plot of lag12 displaying the strongest trend and least variance. After constructing a boxplot based on months and observing the variance across the different months as seen in Plot 3, month was included as a feature.



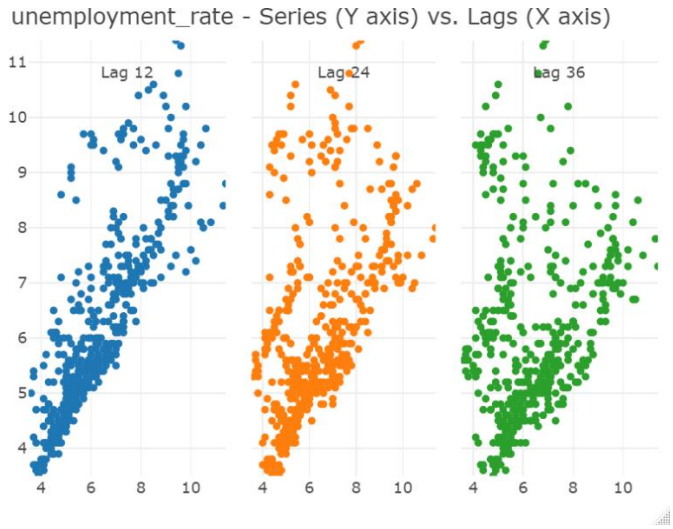
Plot 1: Unemployment Rate over Time



Plot 2: Additive Decomposition of Unemployment Rate



Plot 3: Seasonality plot (months)

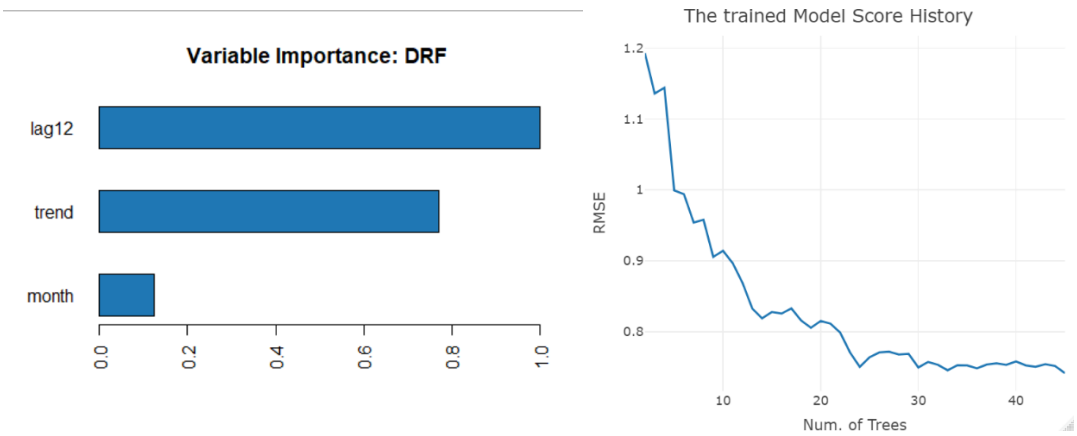


Plot 4: Lag Plots at 12,24,36

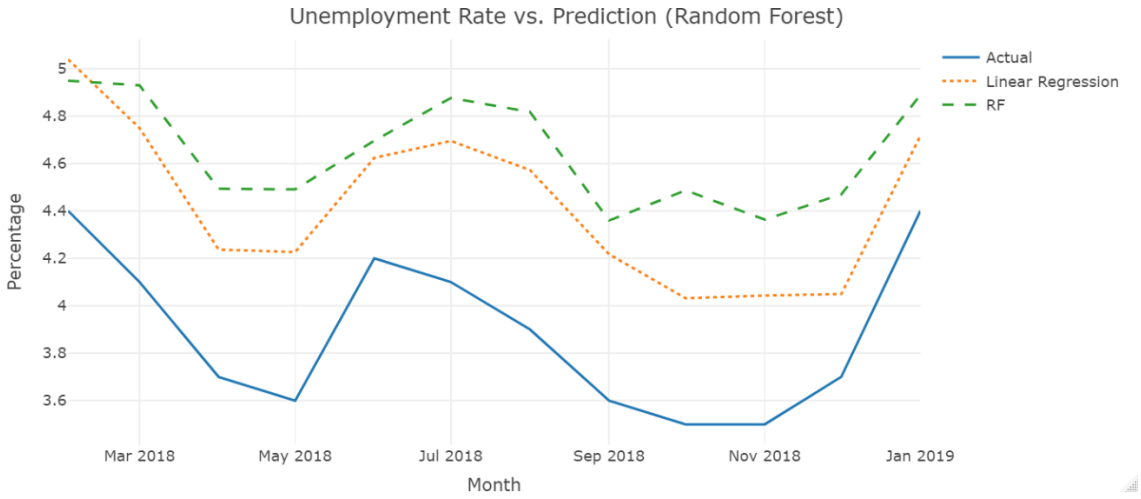
Random Forest and GBM

To evaluate the forecasts generated from the Random Forest and GBM models, we establish a benchmark regression model with month, trend, and lag12 as features. The MAPE from this benchmark model was 0.1404531 and the RMSE was 0.5533557.

A random forest model (RF) with a maximum of 500 trees was trained on the training subset data from the unemployment rate data using the same features month, trend, and lag12. Plot 5 shows the contribution of the mode inputs with lag12 being the most important variable assessed by the random forest algorithm. Ultimately, 44 trees were used as the stopping criteria was met as seen in Plot 5. The MAPE measured against the testing data was 0.1991295 and the RMSE was 0.775971. Plot 6 compares the actual values, benchmark regression model, and the random forest model.

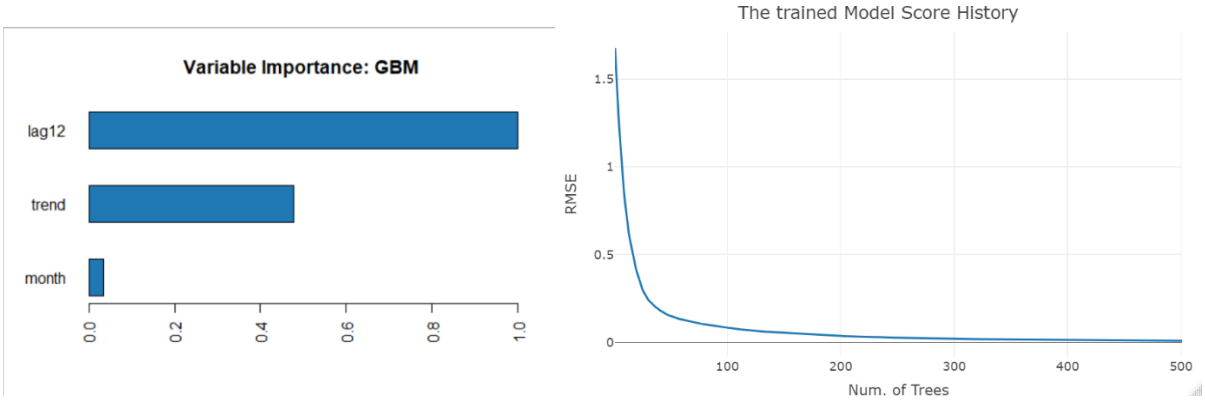


Plot 5: Variable Importance and Model Score on Number of Trees in Random Forest

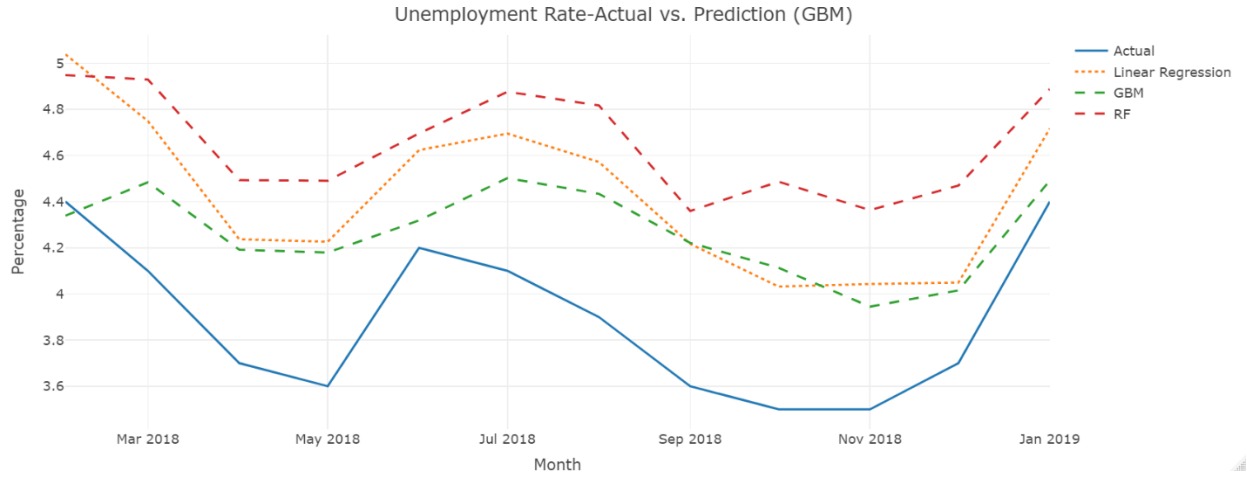


Plot 6: Random Forest Forecast

Subsequently, a GBM was trained with a maximum of 500 trees. Plot 7 shows the contribution of the mode inputs with lag12 being the most important variable assessed by the random forest algorithm. Ultimately, 500 trees were used as the stopping criteria was met as seen in Plot 7, but it seems the most improvement in RMSE score were achieved in the first 100 trees. The MAPE measured against the testing data was 0.1037769 and the RMSE was 0.4336706. Plot 8 compares the actual values, benchmark regression model, random forest model, and GBM model.



Plot 7: Variance importance, Model score on number of trees of GBM



Plot 8: GBM Forecast

Prophet was not included in this report as there were problems with the package on my machine.

Long-term and Short-term performance comparison

Date	Raw Data Values	Exponential Smoothing	Polynomial Regression	Multiple Regression	Diff. Raw Data
2018, 2	4.4	4.441674	3.939979	6.247967	0.1
2018, 3	4.1	4.249191	3.589400	6.278818	0
2018, 4	3.7	3.728389	2.977189	6.263476	0.1
2018, 5	3.6	3.779068	2.851970	6.263235	-0.1
2018, 6	4.2	4.163007	3.148521	6.234664	0.2
2018, 7	4.1	4.137926	3.003566	6.210484	-0.2
2018, 8	3.9	3.781335	2.635782	6.184804	-0.1
2018, 9	3.6	3.320676	2.325321	6.172834	0.1
2018, 10	3.5	3.155865	2.078094	6.141373	0.1
2018, 11	3.5	3.111732	1.999822	6.144801	0.0
2018, 12	3.7	3.230751	1.871225	6.091753	0.2
2019, 1	4.4	3.963826	2.602641	6.161996	0.1
	RMSE (long term)	**0.263217**	**1.222444**	**2.328397**	
	RMSE (short term)	**0.1095324**	**0.485969**	**2.020177**	

Date	VAR	ARIMA	Average Forecast	Benchmark	Random Forest	GBM
2018, 2	-0.01856	4.69013	4.441674	5.038384	4.948561	4.339016
2018, 3	-0.0221	4.58632	4.249191	4.749052	4.929292	4.483886
2018, 4	0.03820	4.18789	3.728389	4.236735	4.493171	4.191636
2018, 5	-0.0056	4.37285	3.779068	4.226481	4.490608	4.179157

2018, 6	0.0274	4.93323	4.163007	4.623387	4.695871	4.318728
2018, 7	0.00535	5.12979	4.137926	4.694976	4.875974	4.501810
2018, 8	0.00179	5.05041	3.781335	4.571970	4.817123	4.434040
2018, 9	0.01552	4.88785	3.320676	4.217094	4.359416	4.221325
2018, 10	-0.0047	4.8918	3.155865	4.031660	4.486536	4.113262
2018, 11	-0.0007	5.00000	3.111732	4.042802	4.363063	3.944921
2018, 12	0.00705	5.16746	3.230751	4.049321	4.469576	4.015434
2019, 1	-0.0011	5.99869	3.963826	4.716452	4.888258	4.491061
RMSE (long term)	**0.41848**	**0.5637**	**0.263217**	**0.5533557**	**0.775971**	**0.4336706**
RMSE (short term)	**0.08527**	**0.40042**	**0.109532**	**0.6437401**	**0.7030805**	**0.2748522**

Table 1: Actual Values, Forecasted Values, and respective long-term and short-term RMSE

Conclusion

Referring to Table 1, we can make a comparison between model performance based on RMSE. The best model for long-term forecasting based on RMSE is the exponential smoothing model. It has an RMSE for long-term forecasting of **0.263217.** The worst model for long-term forecasting is the multiple regression model with an RMSE of **2.328397.** The best model for short-term forecasting is VAR with an RMSE of **0.08527.** However, this model was based on the differenced data, making it difficult to compute the actual forecasted value. Excluding the VAR model, the best for short-term forecasting is the exponential smoothing model, RMSE of **0.1095324**. The worst model for short-term forecasting is the multiple regression model, RMSE of **2.020177.**

Holt-Winters exponential smoothing was the best model for our dependent variable (unemployment data) as it had the lowest short-term and long-term RMSE. It was able to adequately capture the trend and seasonality trends. Polynomial regression had better short-term performance compared to its long-term performance, but was not a good fit for our data as it was only able to fit the trend. Furthermore, it was a high order polynomial that would have probably not been suitable for forecasts longer than our period of 12 months. Multiple regression performed the worst out of all the models. It couldn't fit the trend at all, although it tried to use an ARMA model to fit the residuals. The VAR model performed well but it had to be conducted on differenced data. It effectively used the independent variables in its forecast. The ARIMA model also performed well, probably due to the fact that it was constructed while taking into

account the trend, seasonals, and random part of the unemployment data. The average forecast of exponential smoothing, polynomial regression, multiple regression, and ARIMA also performed well as it used an ensemble method that ensured that its predictions would not be extreme. The benchmark model refers to a regression model constructed with the features lag12, trend, and month. It performed better than the random forest model, which was surprising. The random forest model performed the worst out of the machine learning models- the data seemed like it was not fit for a tree approach. However, the gradient boosting model performed better than the benchmark regression model. This is probably because it is a sequential model where each subsequent tree is modeled on the residuals from the former tree, similar to time series algorithms that were explored earlier.