Untitled

Angel Sarmiento

3/31/2020

Introduction

This investigation involved forecasting multiple steps ahead using a variety of method. The original dataset of nonfarm employment and the other 3 variables were updated to contain the latest values. In place of GSREG, a genetic algorithm (GA) was employed to fit the best possible model within twelve lags of each variable. This model was compared with 3 others, 2 being models also generated from the GA, and one being the best model from previous investigations.

After the best of these was chosen, this model will then be implemented using rolling windows to forecast the next 12 months (2019) and evaluated based on RMSE. The actual nonfarm employment will then be forecasted (instead of the log difference) and then displayed with a plot showing the interval and point forecast. This will then be compared with the empirical method for comparison between empirically generated confidence intervals and confidence intervals that assume normality.

Genetic Algorithm for best subset selection

In order to test a variety of possible models, a Genetic algorithm was employed automatically generate models. These models are created by the algorithm stochastically searching through the different possible combinations of parameters and parameter values to then minimize a metric based on the values. The metrics used here are the in-sample and out-of-sample Root Mean Squared Error (RMSE). In order to decrease the probability of overfitting, 5-fold cross-validation was implemented on top of 100 iterations of size 20 combinations to drive down the value for RMSE. In the interest of saving time, parallel processing was also implemented on 11 cores (threads). Parallel processing drastically decreased the time taken from many hours to under an hour. To visualize the relationship between RMSE and iterations, figure 1 was graphed. This shows both the in-sample and out-of-sample (OOS/external) RMSE changing over time.

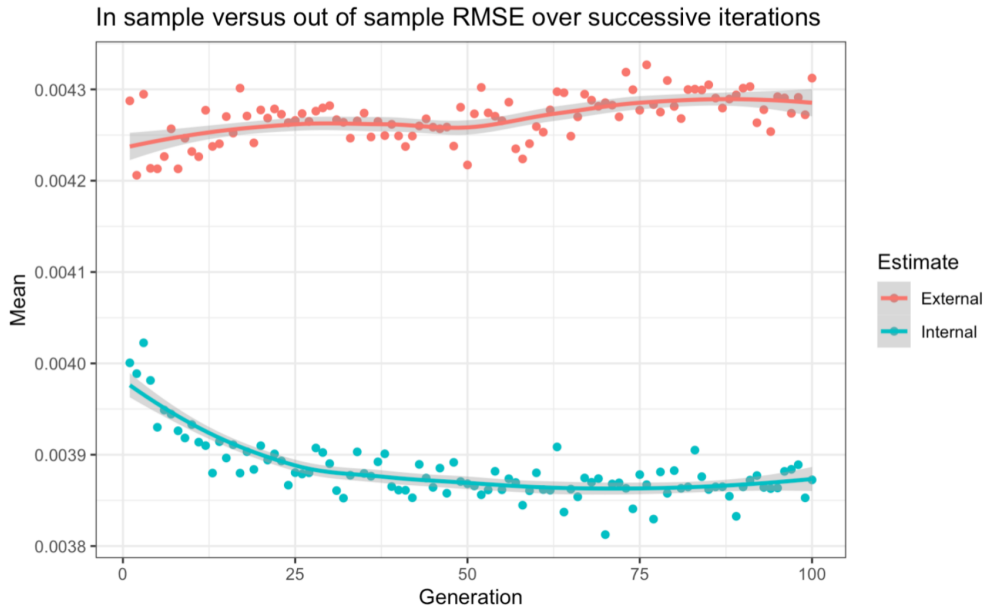


Figure 1: Genetic Algorithm iterations

Surprisingly, the model shows very little improvement over successive generations in external RMSE. This is a fairly significant demonstration that the model is still overfitting to the training data. Ideally, steps would be taken to improve this like increasing the number of folds or letting the algorithm run for more iterations with higher population sizes. This however would take much longer and it has been decided that the best model can be tuned further outside of the genetic algorithm for higher performance.

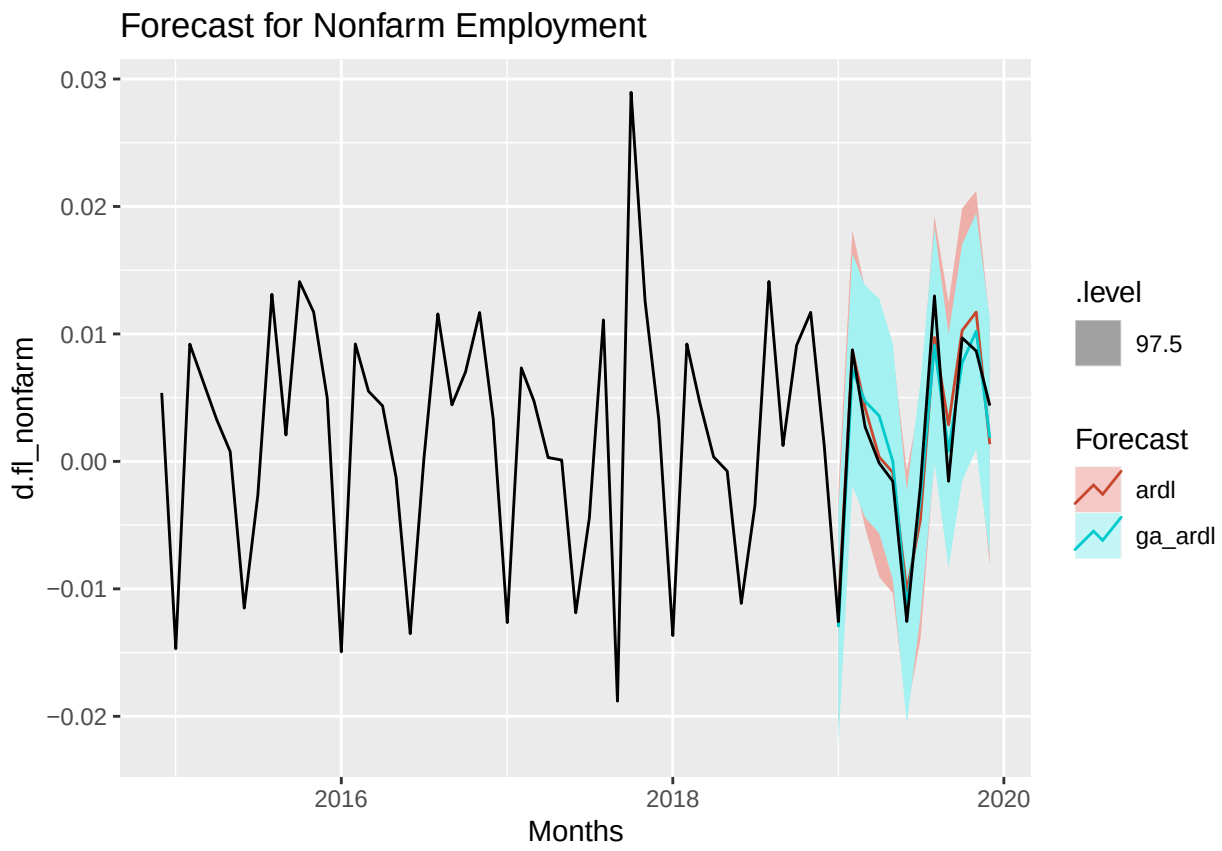
The following table shows 4 different model performances. The first three models are models generated by the GA that seemed to perform well with model three being the model selected as the best by the GA and tweaked outside of the GA. The tweaking involved running on different seeds and training with more folds to alleviate the overfitting situation. The fourth model in this table is the model generated from multiple previous investigations into the modelling and forecasting of nonfarm employment.

Each of these models were selected for different reasons. Model 1 has less overall external RMSE performance but does so with a better selection of variables that is more parsimonious and explains an amount of variance that is competitively similar to model 3. Model 2 was chosen because the external RMSE and internal RMSE are fairly close together. This, to me, shows that this model is relatively robust to overfitting. This model is also slightly more parsimonious than model 3.

	RMSE	OOS_RMSE	Rsquared	N_of_Vars
Model 1	0.00374	0.005380	0.8620	22
Model 2	0.00396	0.003969	0.8411	32
Model 3	0.00364	0.002206	0.8690	34
Prev_Best Model	0.00484	0.002333	0.7704	26

Table 1. 4 Model comparison between the previous best and the models generated by the GA

As one can see, the “best” model from the GA slightly overperforms over the others. Because of this, and despite the reasoning above, this model will be selected as the best model. Higher out of sample RMSE is much more indicative of forecasting ability.

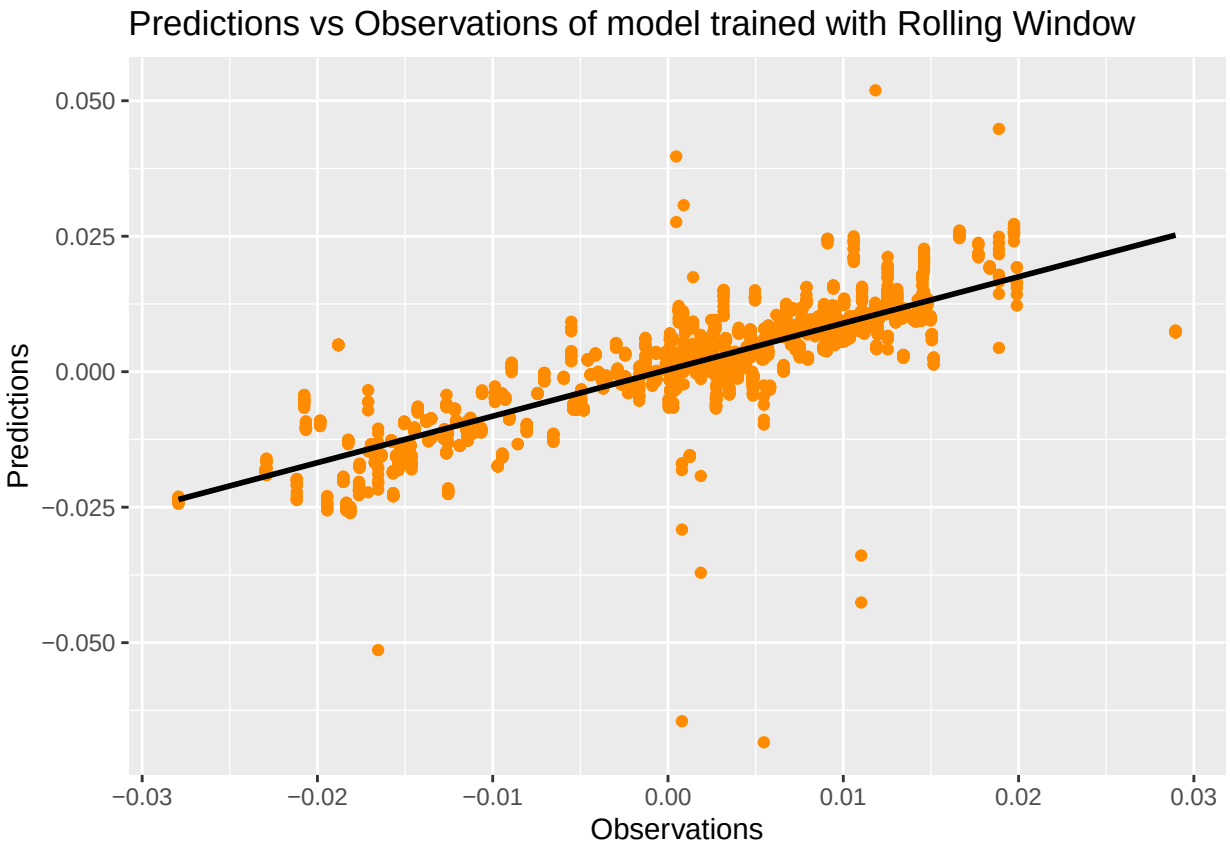


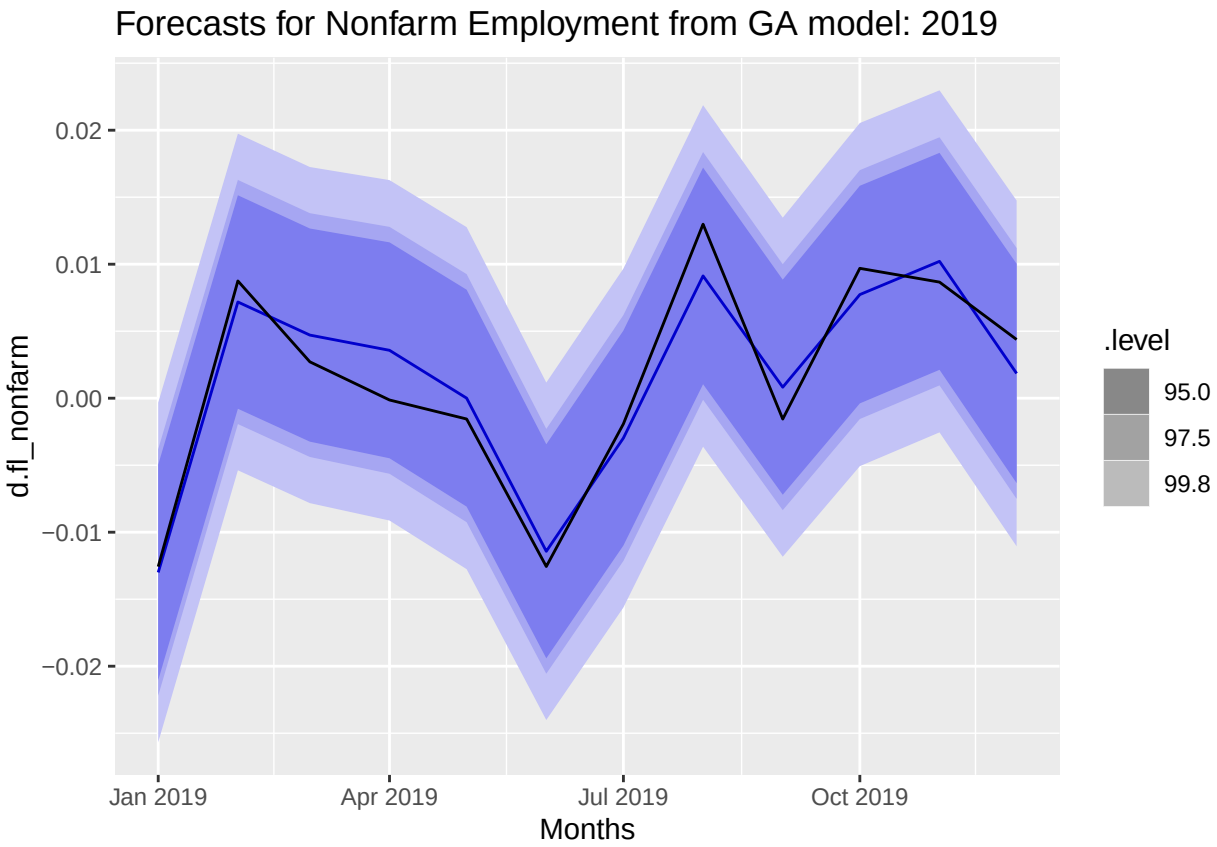
Explain that these are without rolling window

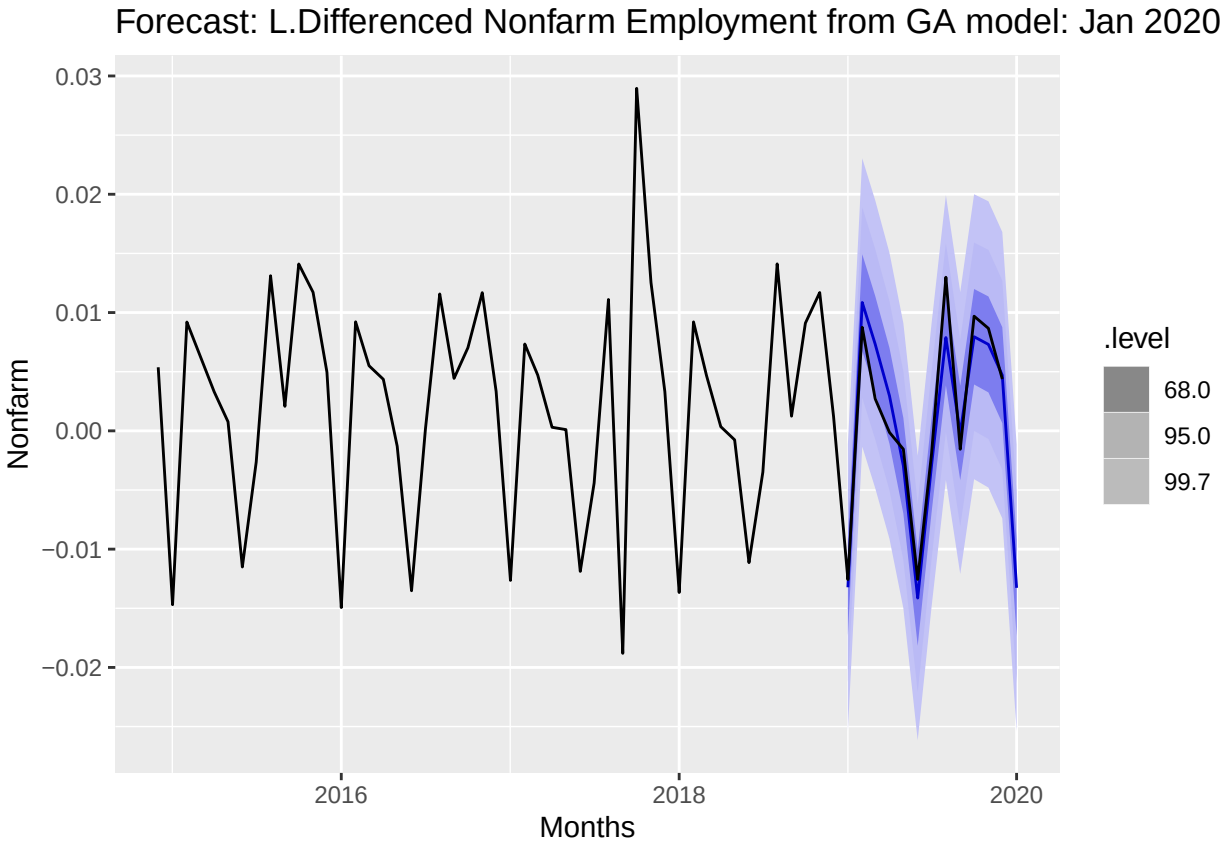
#These are also without rolling window

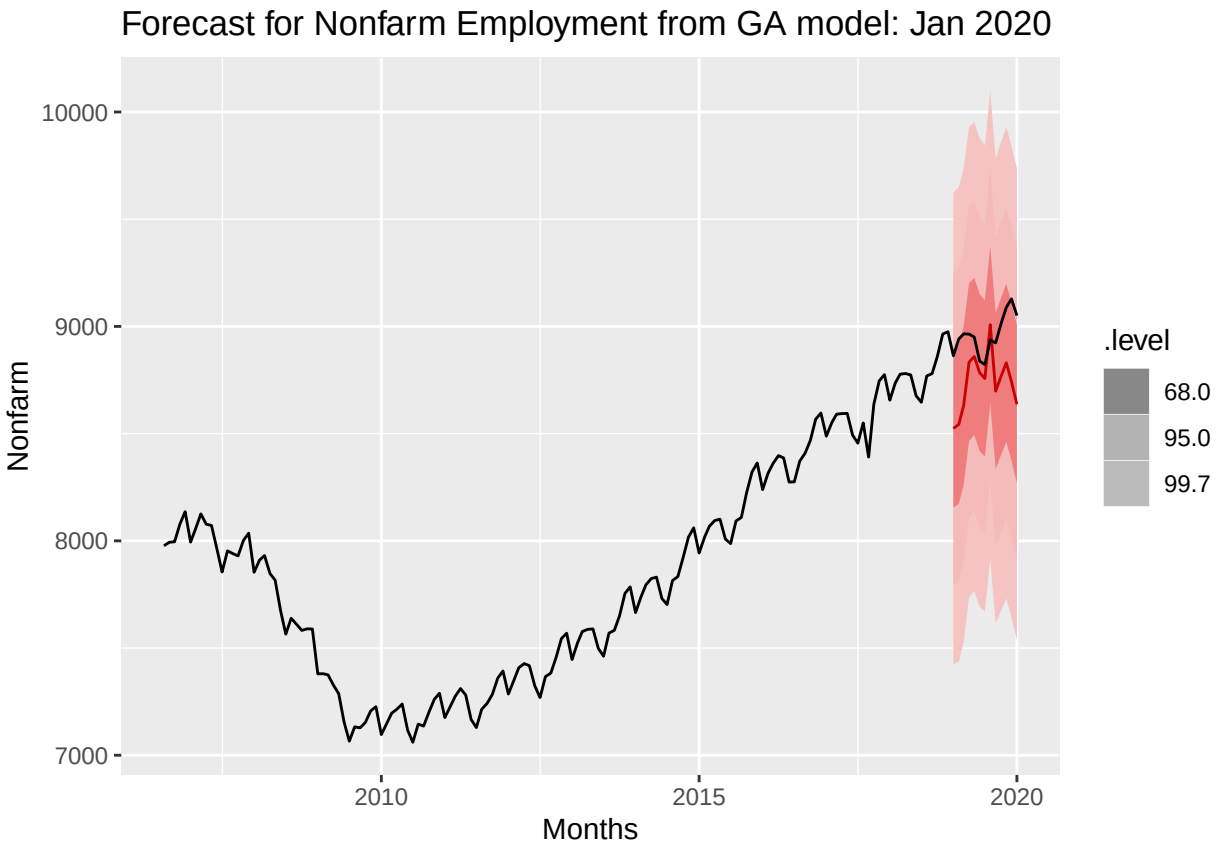
Rolling Window

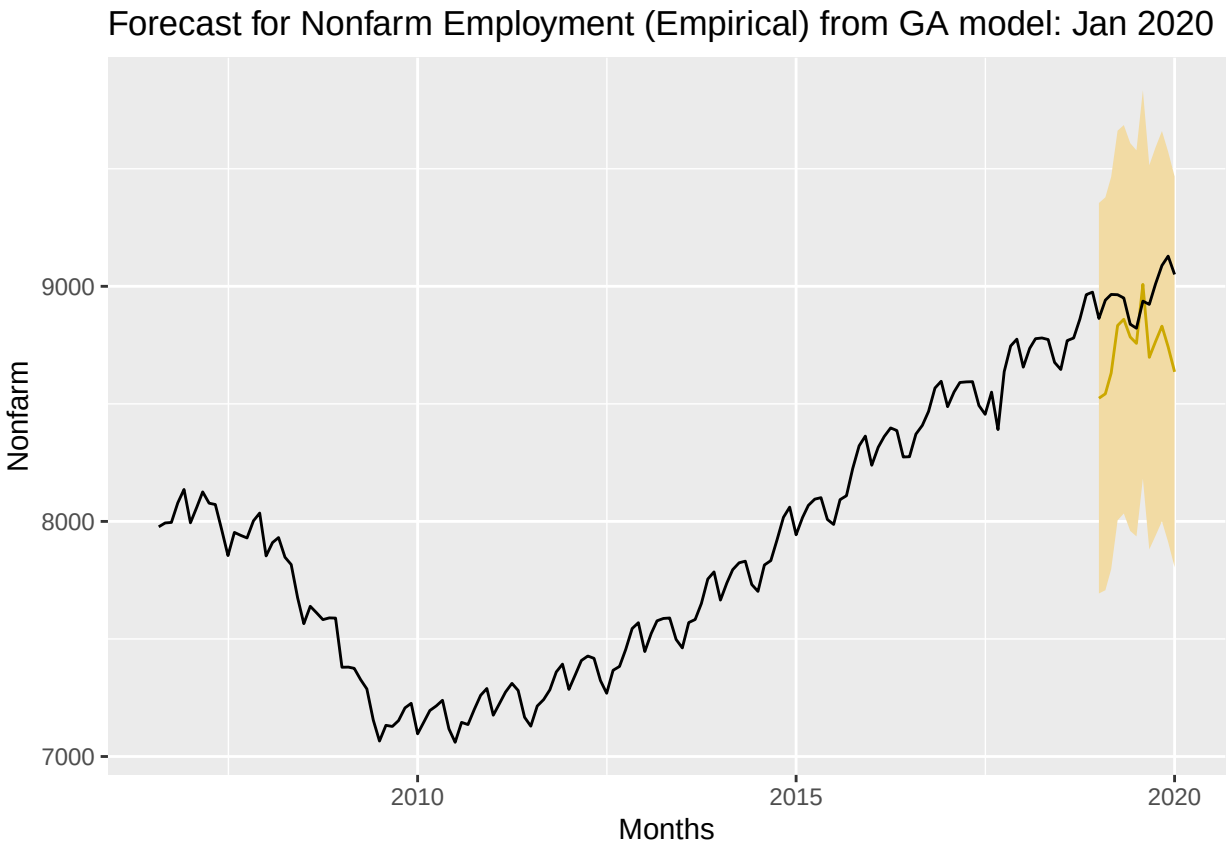
```
Linear Regression

362 samples
  4 predictor

No pre-processing
Resampling: Rolling Forecasting Origin Resampling (12 held-out with no fixed window)
Summary of sample sizes: 36, 37, 38, 39, 40, 41, ...
Resampling results:

  RMSE         Rsquared   MAE
  0.004528818  0.8343578  0.003535837

Tuning parameter 'intercept' was held constant at a value of TRUE
```


Predictions vs Observations of model trained with Rolling Window
Predictions
0.050
0.025
0.000
−0.025
−0.050
−0.03
−0.02
−0.01
0.00
0.01
0.02
0.03
Observations

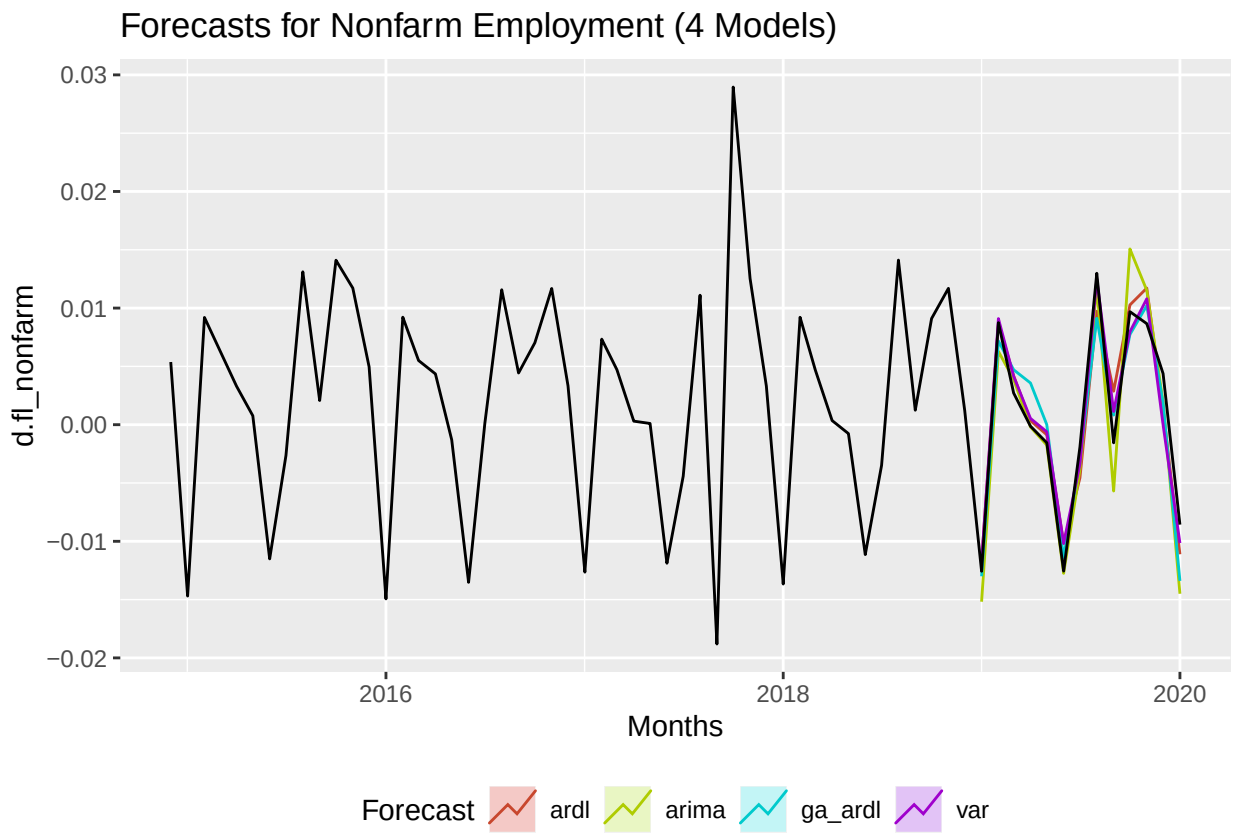

Forecasts for Nonfarm Employment from GA model: 2019
d.fl_nonfarm
0.02
0.01
0.00
−0.01
−0.02
Jan 2019
Apr 2019
Jul 2019
Oct 2019
Months
.level
95.0
97.5
99.8


Forecast: L.Differenced Nonfarm Employment from GA model: Jan 2020
Nonfarm
0.03
0.02
0.01
0.00
−0.01
−0.02
2016
2018
2020
Months
.level
68.0
95.0
99.7


Forecast for Nonfarm Employment from GA model: Jan 2020
Nonfarm
10000
9000
8000
7000
2010
2015
2020
Months
.level
68.0
95.0
99.7


Forecast for Nonfarm Employment (Empirical) from GA model: Jan 2020
Nonfarm
9000
8000
7000
2010
2015
2020
Months

Which is better?



.model	RMSE	MAE	MASE	Winkler	CRPS
var	0.0019615	0.0016202	NaN	0.0164736	0.0013302
ga_ardl	0.0022064	0.0019778	NaN	0.0161018	0.0014120
ardl	0.0023330	0.0019053	NaN	0.0165546	0.0014734
arima	0.0025753	0.0020503	NaN	0.0153963	0.0015334