Introduction

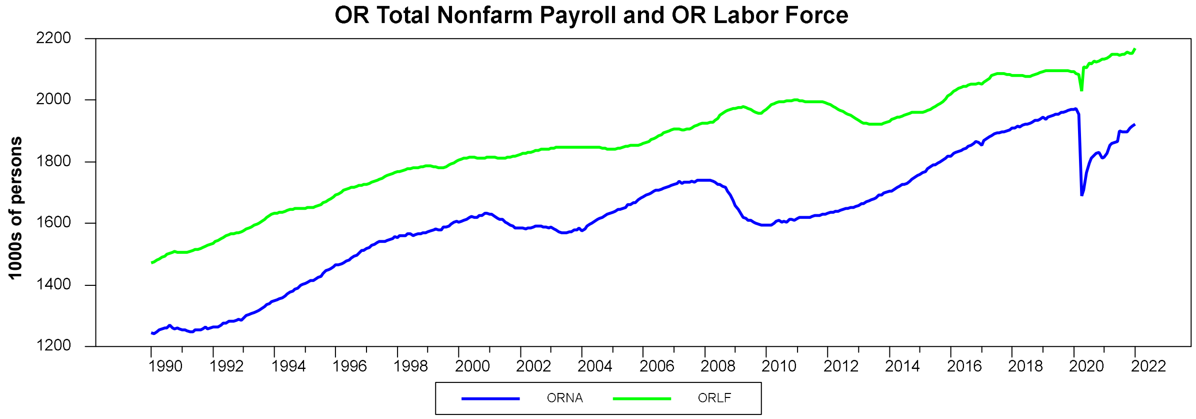
The purpose of this project is to forecast Oregon's total nonfarm payroll (FRED ID: ORNA) over the 2022-2023 period. Employment levels are important both as a general indicator of how the economy is performing and as input into tax revenue forecasts.

Methodology

In general Oregon's recovery from the COVID-19 recession has been robust; however, the pandemic continues to complicate predictions as decisions about how to train models around the economy's unprecedented response are largely unable to be guided by historical data. In addition to taking the standard forecasting approach of modeling employment levels using an ARIMA model, I tried several machine learning modeling techniques to see if non-linear modeling techniques were better able to deal with the large changes in input values due to the pandemic. After considering the alternative approaches I continue to believe that the ARIMA based forecast is the best for this situation. The body of this report presents my final methodology and forecast. Appendix A contains discussion of the machine learning techniques and results. Appendix B contains the RATS input used to create the final ARIMA models.

From the work documented in Appendix A it was found that one of the strongest predictors of Oregon's employment level (ORNA) besides the AR, I, and MA terms, was Oregon's labor force (ORLF). Figure 1 shows ORNA and ORLF on the same scale.

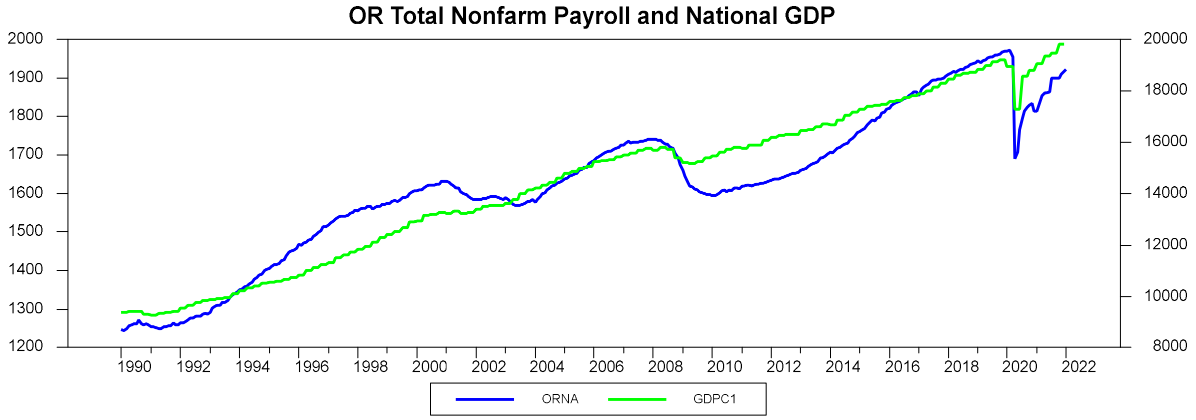
Figure 1



Both series show evidence for a single unit root when a Dickey Fuller test is used on the levels and first differences. The series were tested for evidence of cointegration and there is evidence that the two series are correlated but not cointegrated. The pandemic data was excluded from the cointegration test so that the pandemic spikes wouldn't wash out a longer-term relationship. Since no evidence for cointegration was found a vector autoregression (VARs) model was tested. With the pandemic data excluded RATS did not find evidence that ORLF granger-causes ORNA. Since there is not evidence of granger-causality a VAR model will not be a good choice and a different approach is needed.

Another predictor of ORNA, although much weaker, was national real GDP (GDPC1). Using national real GDP has another advantage in that a projected series (GDPPOT) is available from FRED so I can fill in future values of GDP with the projection and simplify my model. Figure 2 shows ORNA and GDPC1 on different scales where ORNA is in 1000s of persons and GDPC1 is in billions of chained 2012 dollars.

Figure 2

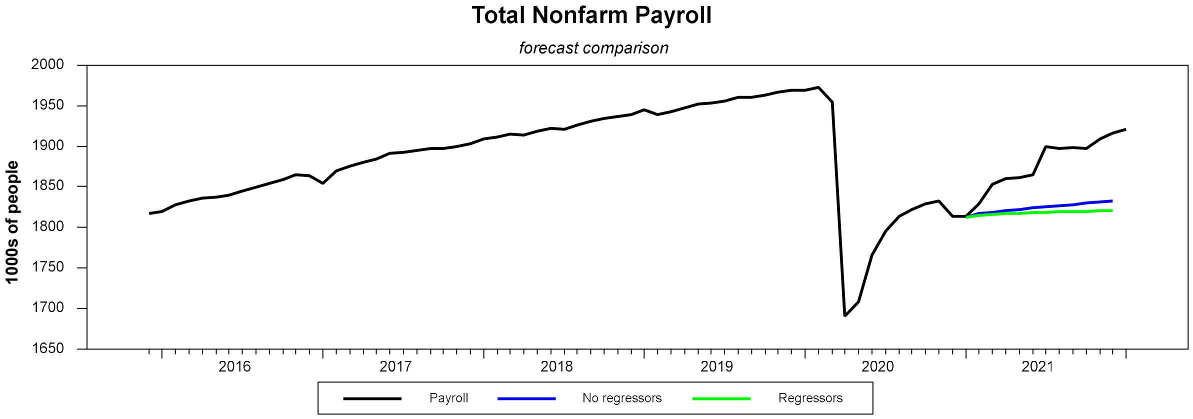


Again, both series show evidence for a single unit root when a Dickey Fuller test is used on the levels and first differences. The series were tested for evidence of cointegration and there is evidence that the two series are correlated but not cointegrated. The pandemic data was excluded from the cointegration test so that the pandemic spikes wouldn't wash out a longer-term relationship. A VAR model was used to test if there is evidence of granger-causality. With the pandemic excluded there is evidence that GDP granger-causes GDP and ORNA and that ORNA granger-causes itself. Rather than use a VARs model to forecast both series I will use an ARIMA model with the regressor option to include GDP (including projected values) to create a forecast for ORNA.

The model with GDP as a regressor and the pandemic data included would not converge so GDP was transformed by taking the natural log and all the steps were repeated. The conclusions stated above remain valid and the model regressed on the natural log of GDP did converge.

Training on the pre 2021 data and forecasting onto 2021 as a validation set suggests that the ARIMA model with no regressors provides a better forecast than the ARIMA model that includes the natural log of GDP as a regressor. This can be seen visual in Figure 3 and a @dmariano test was used to evaluate the two forecasts. Figure 3 shows the two forecasts (with and without regressors trained on pre-2021 data along with the actual 2021 values of ORNA.

Figure 3. Using 2021 as a validation set



Results

The next step was to train the models on all the available data and forecast through 2023. The results are shown in Figures 4 through 6.

Figure 4. No Regressors

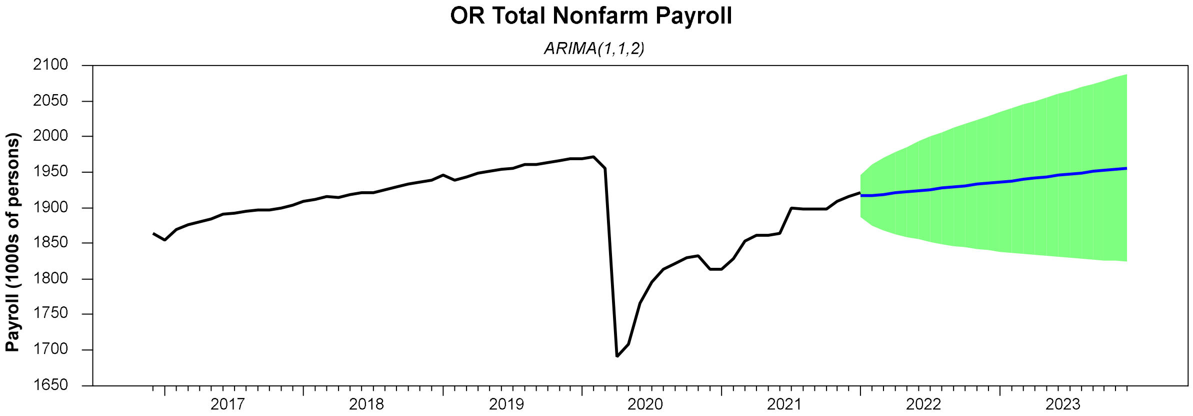


Figure 5. Including Ln of GDP as a regressor

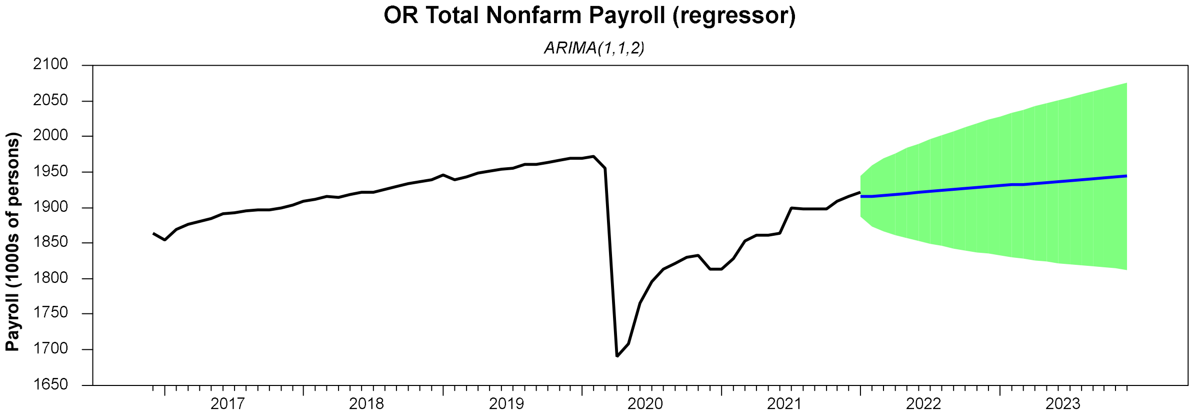
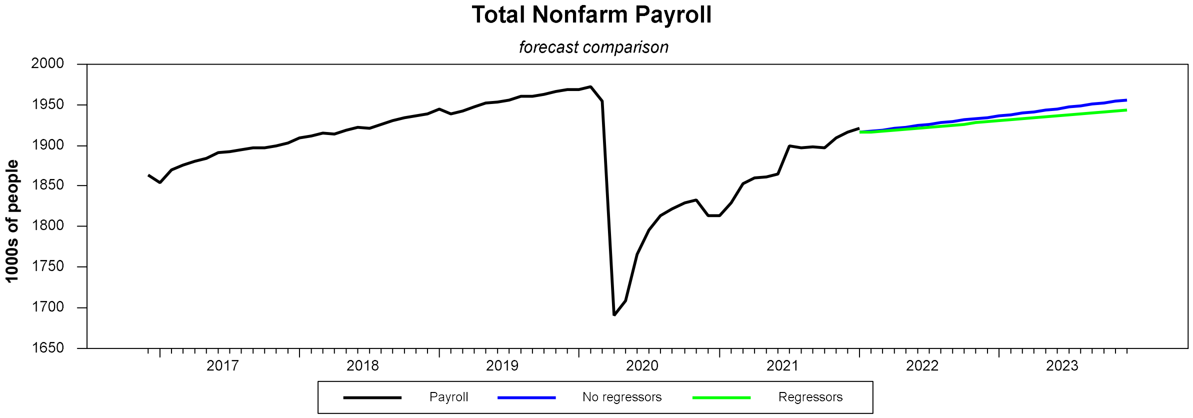


Figure 6. Including Ln of GDP as a regressor



Now the question become did the 2021 validation set without the regressor perform better because it is a better model or because 2021 was not a representative year. I think 2021 was not a representative year and I recommend the forecast with GDP as a regressor (the green line in Figure 6) and the forecast shown with 95% confidence intervals in Figure 5. The risks to this forecast are that GDP does not grow at the projected rate.

Conclusion

I predict that Oregon's nonfarm payroll will continue to grow steadily as long as we don't experience another recession. I predict that growth will be at a somewhat slower rate compared to 2021.

References

Oregon Office of Economic Analysis, "Oregon Economic and Revenue Forecast", March 2022, https://www.oregon.gov/das/oea/pages/forecastecorev.aspx

Appendix A: Machine Learning

I was interested in how the standard ARIMA model approach to forecasting would compare with machine learning techniques I was introduced to in EC 524. I printed the no regressor forecasts from RATS to an excel file and read that file into R for comparison with several machine learning techniques. I trained all my models on pre-2021 data and used 2021 as a validation set. Each training model was chosen using 5-fold cross validation. In other words, the models were trained 5 times each with 20% of the pre-2021 data held out for use in evaluating hyperparameter tuning and model accuracy. The metric for evaluating model accuracy was minimizing the root mean squared error (RMSE). As with the ARIMA model, the FRED series ORNA was the dependent variable. Table A.1 shows the additional FRED series fed to the models for use as predictors.

Table A.1. Predictors

Oregon-Specific Predictors	
ORPHCI	Coincident Economic Activity, index 2007=100. The Coincident Economic Activity Index includes four indicators: nonfarm payroll employment, the unemployment rate, average hours worked in manufacturing and wages and salaries. The trend for each state's index is set to match the trend for gross state product
LBSSA41	Labor Force Participation rate, %
ORLF	Civilian Labor Force, persons
ORMFG	all employees manufacturing, 1000s of persons
ORCONS	all employees construction, 1000s of persons
ORGOVT	all employees gov, 1000s of persons
SMS41000005000000001	all employees information, 1000s of persons
ORFIRE	all employees financial activities, 1000s of persons
ORSRVO	all employees other services, 1000s of persons
ORLEIH	all employees leisure and hospitality, 1000s of persons
ORNRMN	all employees mining and logging, 1000s of persons
ORPBSV	all employees professional and business services, 1000s of persons
OREDUH	all employees education and health services, 1000s of persons
ORTRAD	all employees trade, transportation, and utilities, 1000s of persons
ORUR	Unemployment rate, %
ORPOP	Resident population, 1000s of persons
National Average Predictors	
GDPC1	National Real GDP, quarterly, billions of chained US dollars
LNS11300002	National average of the labor force participation rate for women
LNS11300001	National average of the labor force participation rate for men
TEMPHELPS	National average of Temporary Help Services, 1000s of persons

Oregon-specific predictors were used as much as possible, but several national average predictors were used when OR-specific information was difficult to find (LNS11300002, LNS11300001, and TEMPHELPS) or thought to be flawed (GDPC1). Because I'm interested in forecasting future values, I offset each of the predictors by one year. Inputs from 1990 were used to predict 1991 values; 1991 inputs were used to predict 1992 values etc. I also gave the model the one-year-ago value of ORNA to use as a predictor to mimic the AR term in an ARIMA model and a one-year-ago month-over-month change in ORNA to mimic the first difference term.

I was hoping the machine learning algorithms would be better than me at “learning” around the anomalous pandemic data but instead many of my models just predict that something crazy happens in March and April. As we collect data into the future it’s probable that this effect will disappear. I can already see improvement comparing my 2021 forecast (based on 2020 data) to my 2022 forecast (based on 2021 data). I trained one linear (elastic net) and two non-linear algorithms (random forest and support vector machine). The results of the 2021 forecasts compared to the actual 2021 ORNA values are shown in Figures A.1 – A.3.

Figure A.1. Elastic Net 2021 Predicted Values

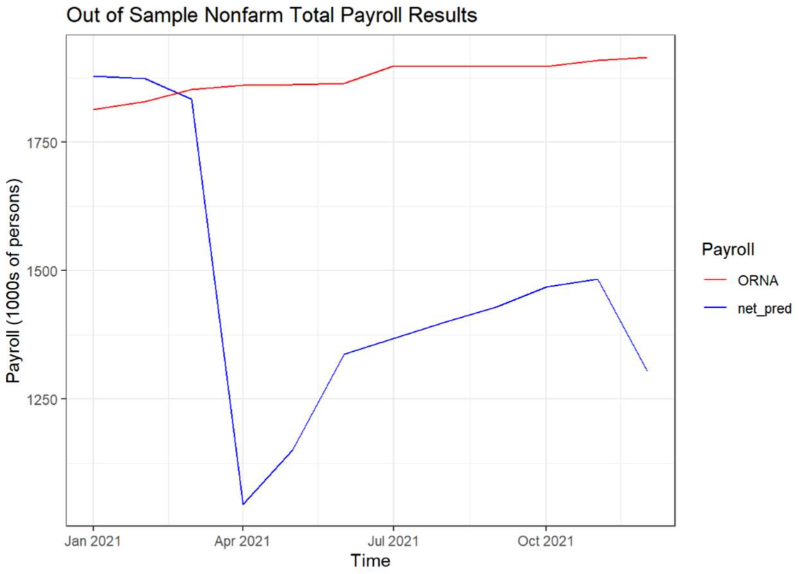


Figure A.2. Random Forest 2021 Predicted Values

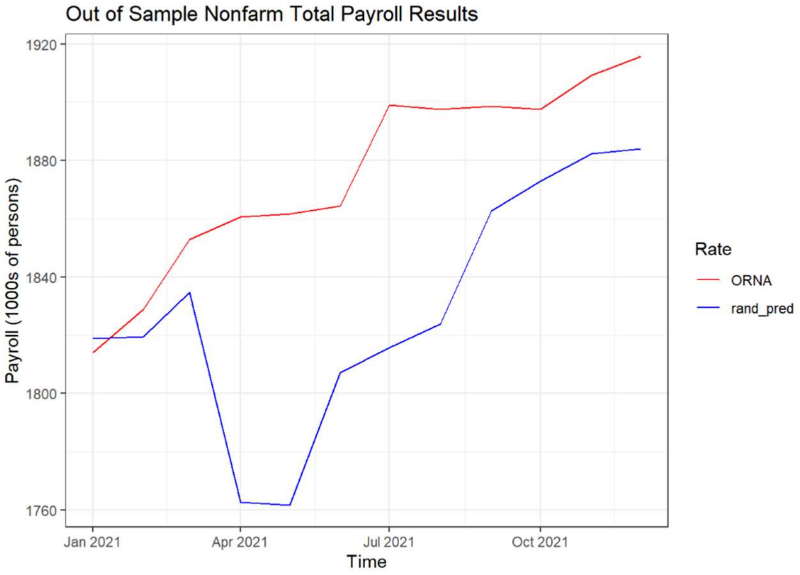


Figure A.3. Support Vector Machine (SVM) 2021 Predicted Values

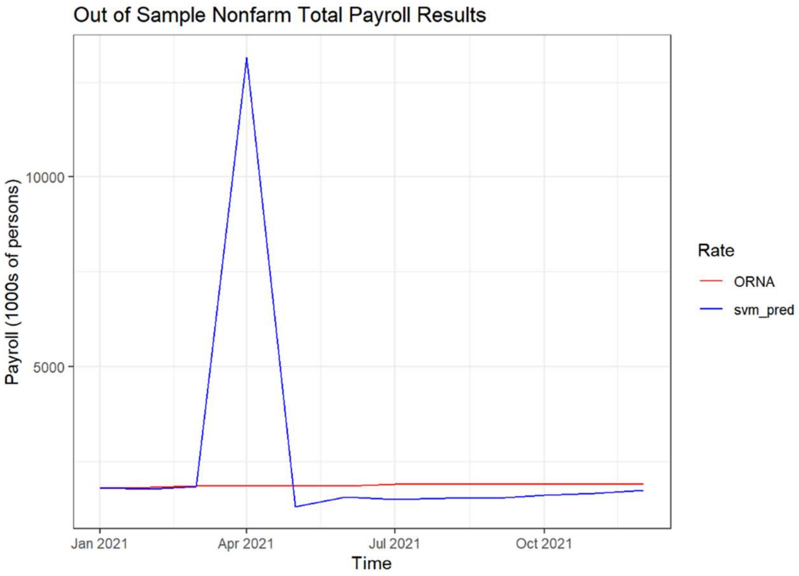
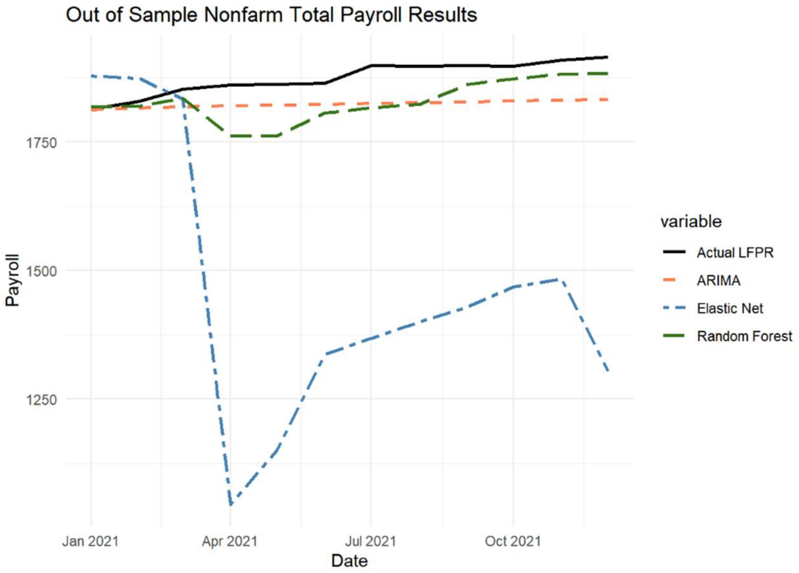
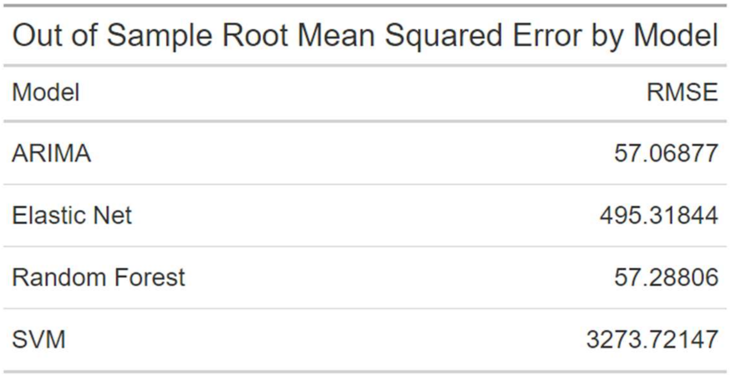


Figure A.4 shows all the 2021 predictions excluding SVM so that the ARIMA forecast can be compared to the machine learning forecasts. The random forest approach appears to have some value as a forecasting method.

Figure A.4. Summary of 2021 Predicted Values



The root mean squared errors for the various forecasts are summarized on the following page. This summary provides further evidence that the performance of the ARIMA model and the random forest were similar.

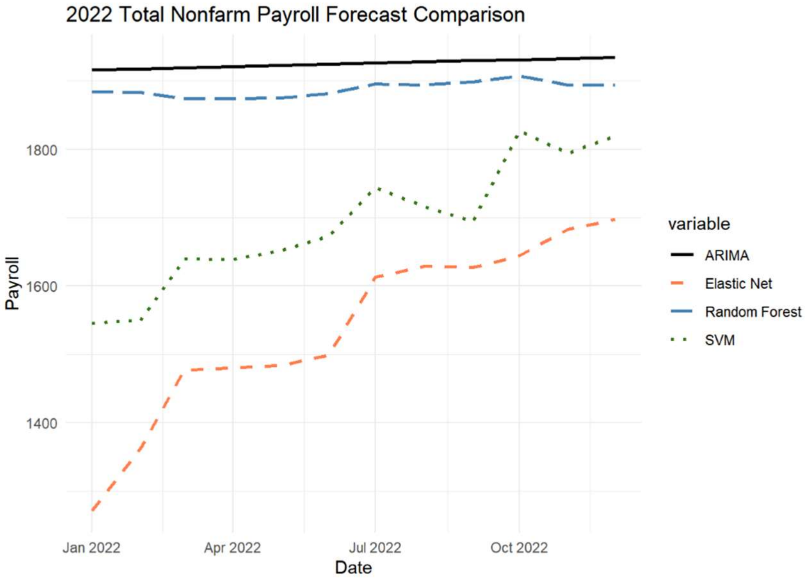


Out of Sample Root Mean Squared Error by Model

Model	RMSE
ARIMA	57.06877
Elastic Net	495.31844
Random Forest	57.28806
SVM	3273.72147

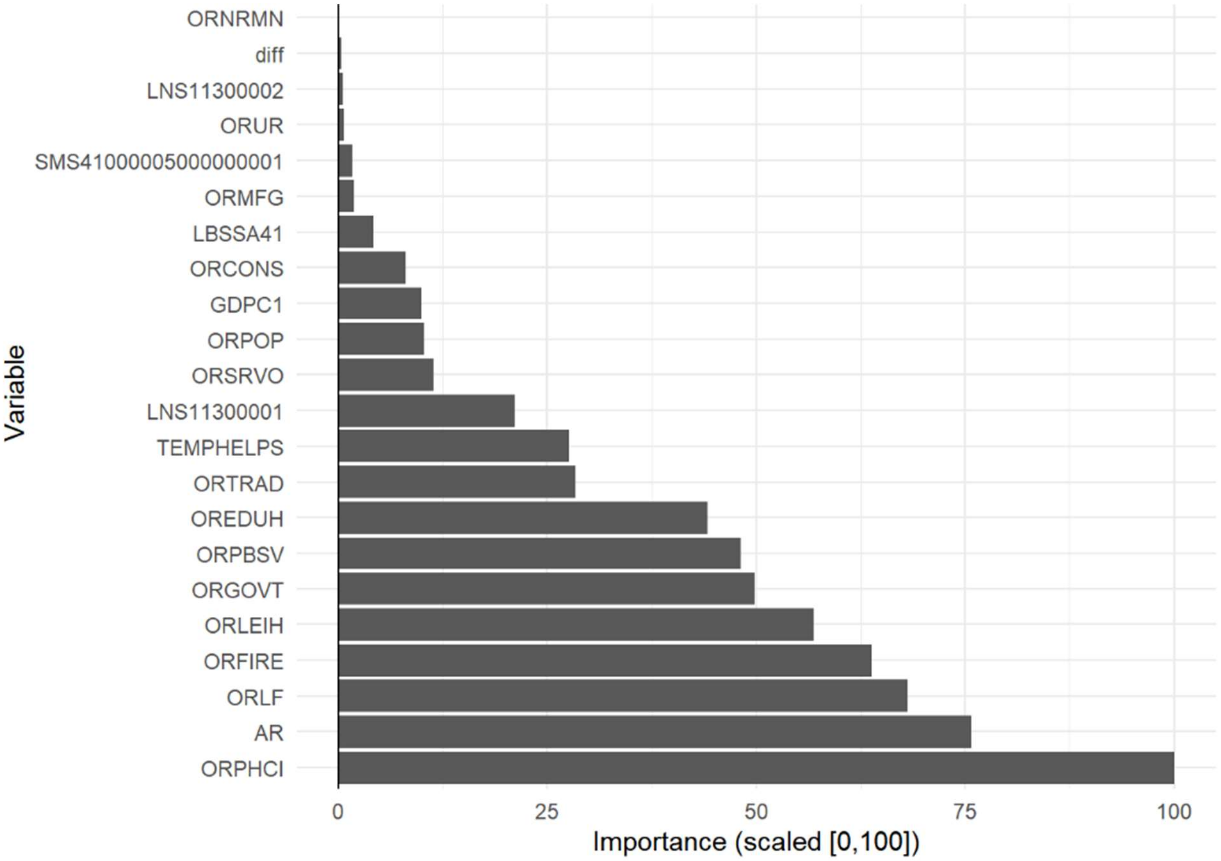
I used each of the trained models to predict ORNA out into 2022, shown in Figure A.5. The ARIMA model (no regressors from RATS) and the random forest predictions continue to be very similar. It will be interesting to see how ORNA progresses over the year.

Figure A.5. Summary of 2022 Predicted Values



In addition to providing predictions, I interrogated my machine learning models to find which series had the most predictive power in the hopes that I could use that knowledge to improve my ARIMA forecast. Figure A.6 shows the relative importance of the predictors, keeping in mind that these values are offset by one year.

Figure A.6. Relative Importance of Predictors



Appendix B: RATS Input

```
* Winter 2022 EC522 Final Project

cal(m) 1900:1
data(format=fred) * * usrec ORNA GDPC1 GDPPOT
all 2025:12

* ORNA: Oregon Total Nonfarm Payroll
* GDPC1: real national GDP
* GDPPOT: projected real GDP

compute startdate = 1990:1
compute enddate   = 2020:2

* log transformation of GDP
set LnGDP = log(GDPC1)

******************* Look at Data
* plot data, duel axis
graph( header="OR Total Nonfarm Payroll and National GDP",
overlay=line,ovcount=1,key=below) 2
# ORNA startdate * 2
# GDPC1 startdate * 3

******************************************************************

* test for unit roots
@dfunit(det=trend,method=aic) ORNA startdate enddate
@dfunit(det=trend,method=aic) LnGDP startdate enddate
** there is evidence of unit roots in each series

* first differences
set diffpay = ORNA - ORNA{1}
set diffgdp = LnGDP - LnGDP{1}

* test for 2nd unit roots
@dfunit(det=constant,method=aic) diffpay startdate enddate
@dfunit(det=constant,method=aic) diffgdp startdate enddate
** there isn't evidence of a 2nd unit root


************** Test for Co-integration *****************

* create cointegrated regression and test ORNA and GDP
linreg(define=coint) ORNA startdate enddate resids
#constant LnGDP
* the coefficient on LnGDP is statistically significant so correlation exists

graph(header="residuals of cointegrationg relationship") 1
# resids startdate enddate
```

```
@egtestresids(maxlags=10,method=aic,det=none) resids startdate enddate
** there is not evidence of cointegration


***********  VARs Models, Pandemic Excluded

*** VAR modeling LN GDP
* let RATS suggest a model
@varlagselect(lags=24, crit=aic)  startdate enddate
# LnGDP ORNA

* Vector AR model
system(model=emp)
variables LnGDP ORNA
lags 1 to 13
det constant
end(system)
estimate(outsigma=omega) startdate enddate
** When the pandemic is excluded
** there is evidence of ORNA and LNGDP causing ORNA
** there is evidence of LNGDP causing LNGDP
** instead of projecting both ORNA and GDP I will use
** the projected GDP available on FRED


*********************************************************************

* test the 1st diff term of the prepandemic ORNA series to get an idea of ARIMA model
@bjautofit(pmax=5, qmax=5, crit=sbc) diffpay startdate enddate

boxjenk(constant, ar=1, ma=2, diffs=1, define=mod) ORNA startdate enddate resids
@autocorr resids startdate enddate
@histogram2(stats,counts) resids startdate enddate
* also tried 1,1,1 and 2,1,2


boxjenk(constant, ar=1, ma=2, diffs=1, define=modr,regressors) ORNA startdate enddate residr
#LnGDP
@autocorr residr startdate enddate
@histogram2(stats,counts) residr startdate enddate
* model converges, nonnormal resids, no more time dep


************* train on pre-2021 data and validate on 2021

* reset enddate to train on all pre2021 data
compute enddate   = 2020:12
compute forend    = 2021:12
```

```
********* ARIMA model, no regressors

boxjenk(constant, ar=1, ma=2, diffs=1, define=mod1) ORNA startdate enddate resids1
@autocorr resids1 startdate enddate
@histogram2(stats,counts) resids1 startdate enddate
* model did converge
* no more evidence of time dependence but nonnormal resids

uforecast(equation=mod1,nostatic,stderrs=errors1, from=enddate+1,to=forend) fore1
set uerror1 enddate+1 forend = fore1+1.96*errors1
set lerror1 enddate+1 forend = fore1-1.96*errors1

graph(style=line,vlabel="Payroll (1000s of persons)",header="OR Total Nonfarm Payroll", $
  subheader="ARIMA(1,1,2)",overaly=fan, ovcount=2, ovsame) 4
# ORNA enddate-60 forend
# fore1 enddate+1 forend
# uerror1 enddate+1 forend 3
# lerror1 enddate+1 forend 3


********* ARIMA model, with regressors

boxjenk(constant, ar=1, ma=2, diffs=1, define=mod2,regressors,iters=500) ORNA startdate
enddate resids2
#LnGDP
@autocorr resids2 startdate enddate
@histogram2(stats,counts) resids2 startdate enddate
* model did converge
* no more evidence of time dependence but nonnormal resids

uforecast(equation=mod2,nostatic,stderrs=errors2, from=enddate+1,to=forend) fore2
set uerror2 enddate+1 forend = fore2+1.96*errors2
set lerror2 enddate+1 forend = fore2-1.96*errors2

graph(style=line,vlabel="Payroll (1000s of persons)",header="OR Total Nonfarm Payroll
(regressor)", $
  subheader="ARIMA(1,1,2)",overaly=fan, ovcount=2, ovsame) 4
# ORNA enddate-60 forend
# fore2 enddate+1 forend
# uerror2 enddate+1 forend 3
# lerror2 enddate+1 forend 3


**** forecast comparison ***
*A small P value indicates that the forecast on the line will be rejected in favor of the other.
* note @dmariano not for nested models two models are nested if one model contains
* all the terms of the other, and at least one additional term.
@dmariano(lags=6,lwindow=newey) ORNA  fore1 fore2
* results indicate that fore1 is better than fore2
```

```
************************ View forecasts together ************************
graph(style=line,vlabel="1000s of people", key=below,header="Total Nonfarm Payroll", $
        klabel=||"Payroll"|"No regressors"|"Regressors"||,subheader="forecast comparison") 3
# ORNA enddate-60 *
# fore1 * *
# fore2 * *


************* train on 2022 data and forecast through 2023

compute enddate   = 2021:12
compute forend    = 2023:12

* use FRED's projected GDP
set GDPC1 enddate+1 forend = GDPPOT
set LnGDP = log(GDPC1)

********* ARIMA model, no regressors

boxjenk(constant, ar=1, ma=2, diffs=1, define=mod3) ORNA startdate enddate resids3
@autocorr resids3 startdate enddate
@histogram2(stats,counts) resids3 startdate enddate
* model did converge
* no more evidence of time dependence but nonnormal resids

uforecast(equation=mod3,nostatic,stderrs=errors3, from=enddate+1,to=forend) fore3
set uerror3 enddate+1 forend = fore3+1.96*errors3
set lerror3 enddate+1 forend = fore3-1.96*errors3

graph(style=line,vlabel="Payroll (1000s of persons)",header="OR Total Nonfarm Payroll", $
  subheader="ARIMA(1,1,2)",overaly=fan, ovcount=2, ovsame) 4
# ORNA enddate-60 forend
# fore3 enddate+1 forend
# uerror3 enddate+1 forend 3
# lerror3 enddate+1 forend 3


********* ARIMA model, with regressors

boxjenk(constant, ar=1, ma=2, diffs=1, define=mod4,regressors) ORNA startdate enddate
resids4
#LNGDP
@autocorr resids4 startdate enddate
@histogram2(stats,counts) resids4 startdate enddate
* model did converge
* no more evidence of time dependence but nonnormal resids

uforecast(equation=mod4,nostatic,stderrs=errors4, from=enddate+1,to=forend) fore4
set uerror4 enddate+1 forend = fore4+1.96*errors4
set lerror4 enddate+1 forend = fore4-1.96*errors4
```

```
graph(style=line,vlabel="Payroll (1000s of persons)",header="OR Total Nonfarm Payroll
(regressor)", $
  subheader="ARIMA(1,1,2)",overaly=fan, ovcount=2, ovsame) 4
# ORNA enddate-60 forend
# fore4 enddate+1 forend
# uerror4 enddate+1 forend 3
# lerror4 enddate+1 forend 3


*********************** View forecasts together ***********************
graph(style=line,vlabel="1000s of people", key=below,header="Total Nonfarm Payroll", $
        klabel=||"Payroll"|"No regressors"|"Regressors"||,subheader="forecast comparison") 3
# ORNA enddate-60 *
# fore3 enddate+1 forend
# fore4 enddate+1 forend


************** create output file
* create an xls file holding output data
open copy ORNAforedata.xls
copy(dates,format=xls,org=columns) * * fore1 fore3
```