Unemployment Assignment

DECEMBER 1

Authored by: Pablo Pérez, Juan Miguel Ramos and Andrés Canalejo

Contents

1. Introduction 4

2. Seasonal ARIMA Model 4

3. SARIMAX Model using an intervention variable 9

4. Model Comparison 10

5. Nonlinear Techniques 12

6. Unemployment Forecast November 2022 14

1. Introduction

In this report, the dataset we are going to use contains unemployment data for each month in Spain from January 2001 to October 2022. The evolution of unemployment during these years is shown in Figure 1.1, where an irregular behavior of the data is observed due to various factors, such as the outbreak of the economic crisis in 2008 and the COVID-19.

The main goal of this work is to analyze our collected data and apply different prediction models to finally predict the exact number of unemployed people for the month of November 2022.

Figure 1.1: Analysis of unemployment in Spain from 2001 to 2022

2. Seasonal ARIMA Model

To be able to predict the time series we must know what kind of process we are dealing with, for this, it is necessary to analyze if the series is stationary in the mean and variance. Therefore, the first step is to make a representation of the time series along with its ACF and PACF to examine the data collected, figure 2.1.

At first sight we can observe that the series may not be stationary on average as there are two clear trends along the series, first from 2008 to 2013, this is due to the crisis and therefore the unemployment

rate skyrocketed. Then, a downward trend from 2013 to 2020, where we started to recover from the crisis and therefore the low unemployment rate slowly decreased.
It also seems that this series is not stationary in variance since the range of values of the variable y for the values of x is changing with time.

We can observe that during the first 8 years there seems to be a seasonality, there is a pattern that repeats every year, this may be because normally the unemployment rate increases during the winter and decreases during the summer. In the other years we can also observe local minimums and maximums, this is surely due to seasonality, for example, it is normal that the unemployment rate at Christmas falls as we are in a period of high demand and many businesses require more staff and then, the unemployment rate increases as this demand falls and not so much staff is needed.

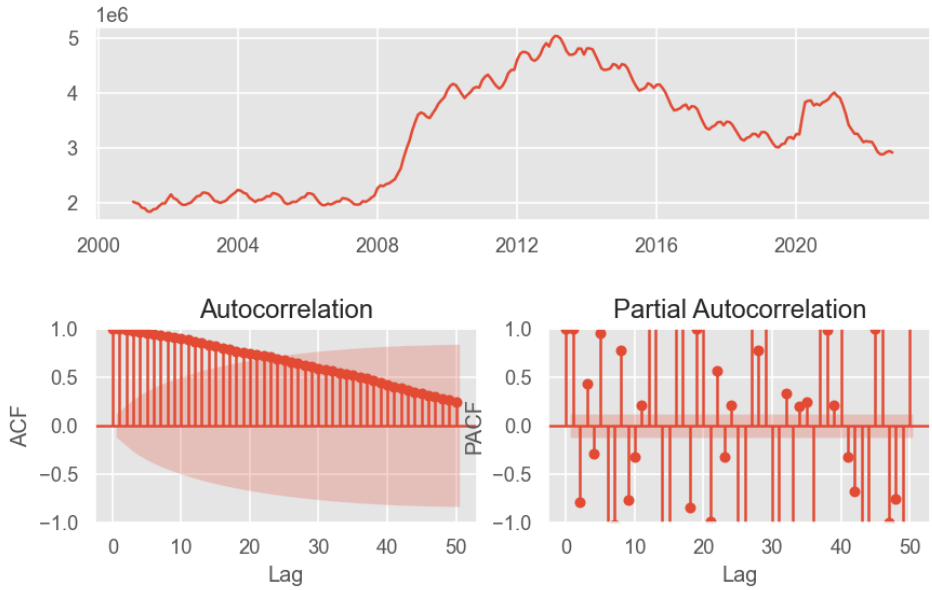


Figure 2.1: ACF and PACF of the time series

To verify that the variance is not stationary, the Box-Cox plot has been plotted as shown in Figure 2.2 confirming that the variance is not stationary, as the value of the logarithm of the moving average depends on the value of the logarithm of the standard deviation with a lambda of 0.1608. Then, the Box-Cox transformation has been applied, allowing us to obtain a stationary series in variance. As shown in Figure 2.3. the logarithm of the moving average no longer depends on the values taken by the logarithm of the standard deviation.

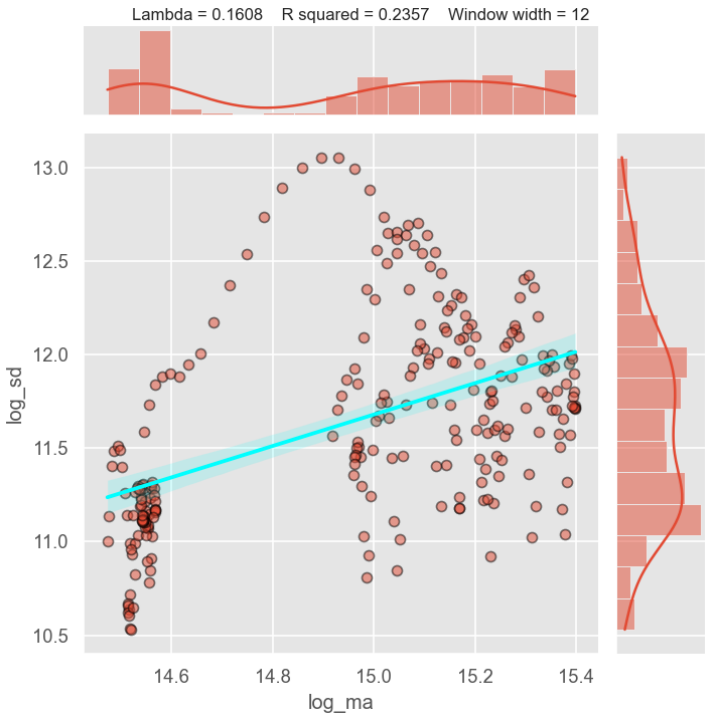


Figure 2.2: Box-Cox Transformation Nonstationary Variance

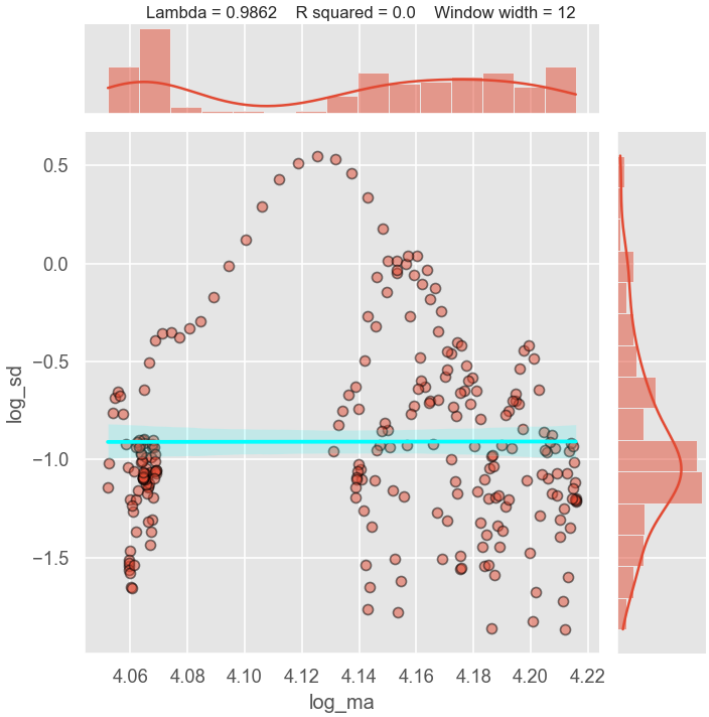


Figure 2.3: Box-Cox Transformation Stationary Variance

Once our variance is stationary, we must check if the series is stationary in the mean, for this we used the Dickey Fuller test. This test has returned a p-value of 0.4599, which means that it is not stationary in the mean, and it is necessary to differentiate the series because its p-value is much higher than 0.5.

Once the differentiation is done, the time series is represented again together with its ACF and PACF obtaining figure 2.4, where we can clearly see that now our series is stationary in the mean.

$$SARIMA\ (1,1,0)\times(0,1,1)_{12}$$

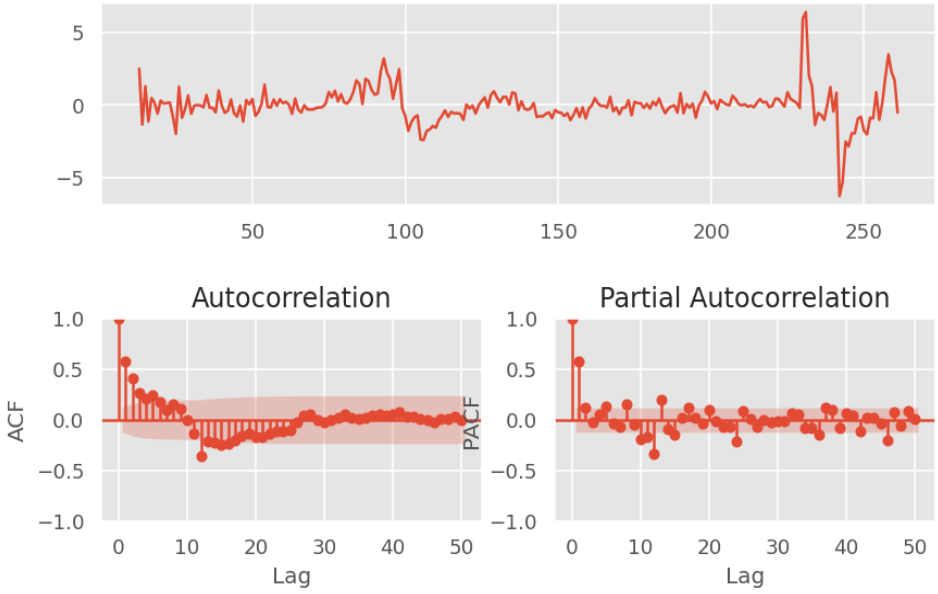


Figure 2.4: ACF and PACF of the time series differentiated

The next step consists on fitting the SARIMA model and testing the significance of the coefficients. By analyzing figure 2.4, we can observe that for the regular part we have two possible options, AR(1) or MA(3), we will choose AR(1) for its simplicity. While for the seasonal part we have an AR(1) or an MA(1), where we choose an MA(1).

Once the model has been fitted and the significance of its coefficients checked, the residual error is plotted in Figure 2.5 in order to analyze the residuals of the model.

Analyzing them, we can observe how they are almost at zero, and they are below the threshold, so we can conclude that the random residuals are white noise, therefore, they are not predictable and the chosen model is correct.

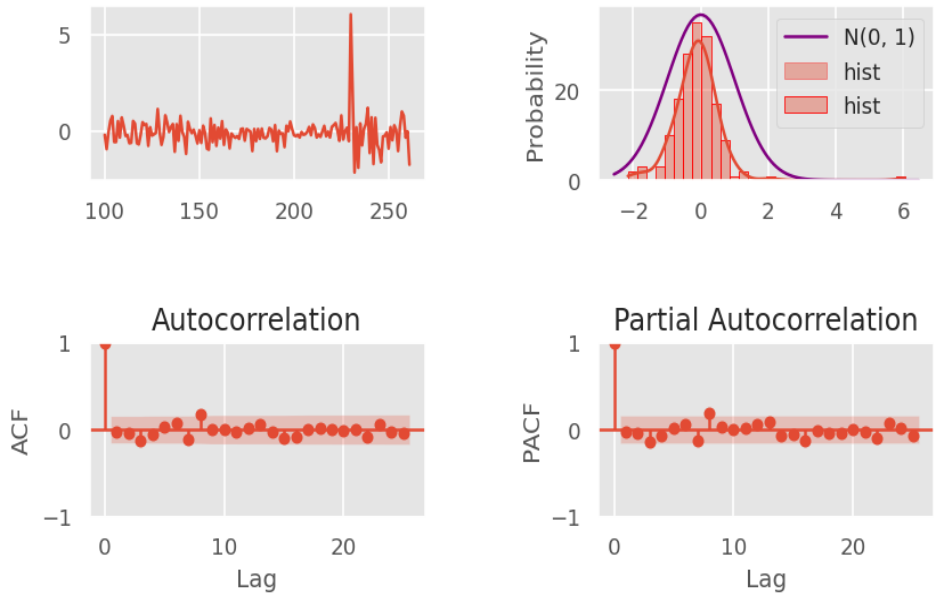


Figure 2.5: Analyzing model residuals

Finally the last step has been to perform a prediction for 20 future instances, as plotted in Figure 2.6, where the actual time series and the predicted values are shown.

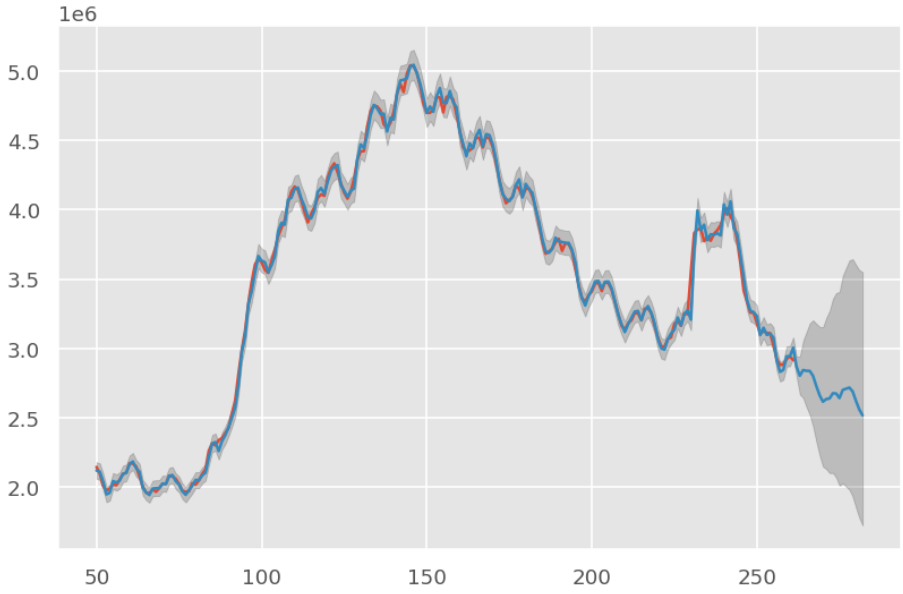


Figure 2.6: Forecasts for in-sample and out-of-sample

3. SARIMAX Model using an intervention variable

This type of model takes into account intervention variables which may have affected the variable to be predicted. In this case, we have used a spike variable to model the behavior of unemployment during the Covid pandemic.

Therefore, a dummy variable is created with value 0 during the non-Covid season and 1 during the Covid season. According to the values taken by the unemployment rate in the time series, we establish that our intervention variable should take value 1 from March 2020 to August 2021 (inclusive).

To run the model we use the same parameters used in the Seasonal ARIMA model including the intervention variable.

$$SARIMAX\ (1,1,0)\times(0,1,1)_{12}$$

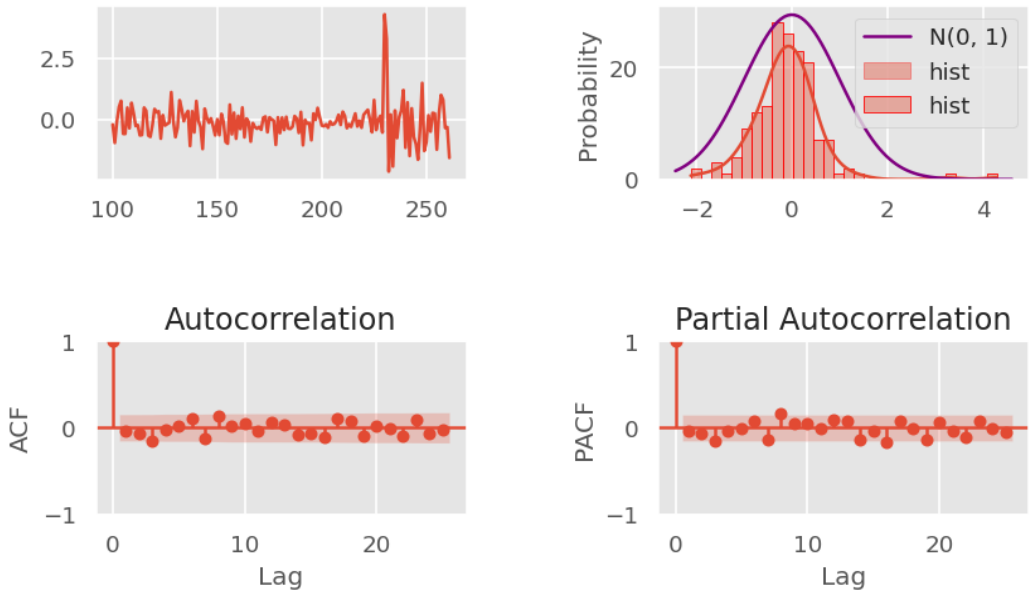


Figure 3.1: ACF and PACF of the SARIMAX model

Analyzing the residuals of the model (Figure 3.1), we can observe that they are around value 0, and below the threshold, therefore, we can conclude that the residuals seem random (white noise). They are not predictable, therefore, the chosen model seems to be working correctly.

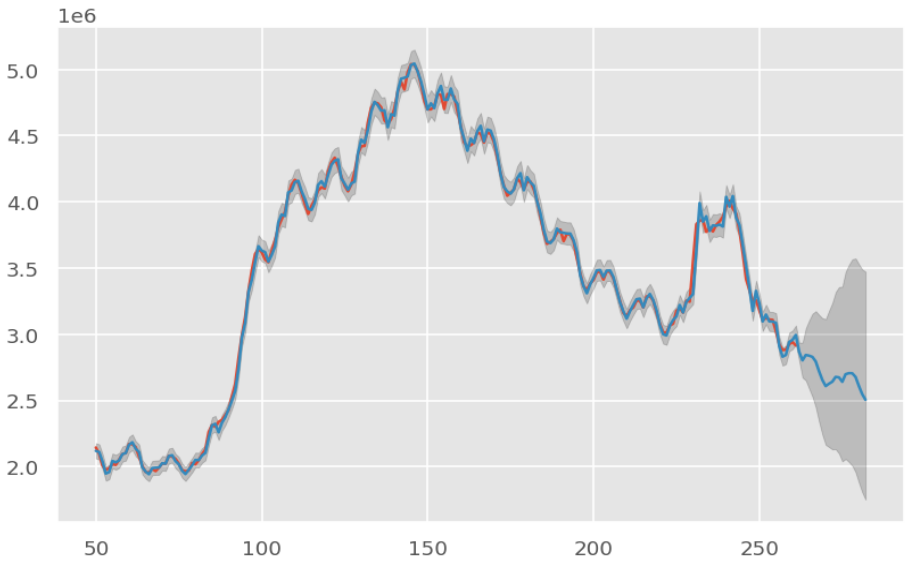


Figure 3.2: Forecasts for in-sample and out-of-sample

Figure 3.2 states the predicted time series (blue) against the real time series (red) and the predicted values for the future (20 moments more) using the SARIMAX model proposed.

4. Model Comparison

When checking the results (Table 4.1) of the SARIMA model we can observe that all of the coefficients are significant. The same occurs with the SARIMAX model (Table 4.2). However, the AIC is lower in the second one as well as the BIC and HQIC. These results show that the second model seems to perform better than the first one. Both models perform poorly in the assumptions tests (Ljung-Box, Heteroskedasticity and Jarque-Bera) so the residuals may not be white noise even though the residuals plots seemed fine. They present some skewness and a large kurtosis both reduced in the second model.

SARIMAX Results

Dep. Variable:	TOTAL	No. Observations:	262
Model:	SARIMAX(1, 1, 0)x(0, 1, [1], 12)	Log Likelihood	129.159
Date:	Wed, 30 Nov 2022	AIC	-252.318
Time:	17:26:19	BIC	-241.926
Sample:	0	HQIC	-248.129
	- 262		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6787	0.028	23.951	0.000	0.623	0.734
ma.S.L12	-0.8464	0.043	-19.569	0.000	-0.931	-0.762
sigma2	0.0189	0.001	33.577	0.000	0.018	0.020

Ljung-Box (L1) (Q):	0.27	Jarque-Bera (JB):	3864.65
Prob(Q):	0.61	Prob(JB):	0.00
Heteroskedasticity (H):	2.15	Skew:	2.36
Prob(H) (two-sided):	0.00	Kurtosis:	22.26

Table 4.1: SARIMA results

SARIMAX Results

Dep. Variable:	TOTAL	No. Observations:	262
Model:	SARIMAX(1, 1, 0)x(0, 1, [1], 12)	Log Likelihood	138.161
Date:	Wed, 30 Nov 2022	AIC	-268.322
Time:	17:26:22	BIC	-254.466
Sample:	0	HQIC	-262.737
	- 262		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
COVID	0.3311	0.023	14.156	0.000	0.285	0.377
ar.L1	0.6706	0.027	24.584	0.000	0.617	0.724
ma.S.L12	-0.8376	0.051	-16.484	0.000	-0.937	-0.738
sigma2	0.0175	0.001	17.298	0.000	0.016	0.020

Ljung-Box (L1) (Q):	0.25	Jarque-Bera (JB):	551.98
Prob(Q):	0.62	Prob(JB):	0.00
Heteroskedasticity (H):	1.87	Skew:	1.28
Prob(H) (two-sided):	0.01	Kurtosis:	10.04

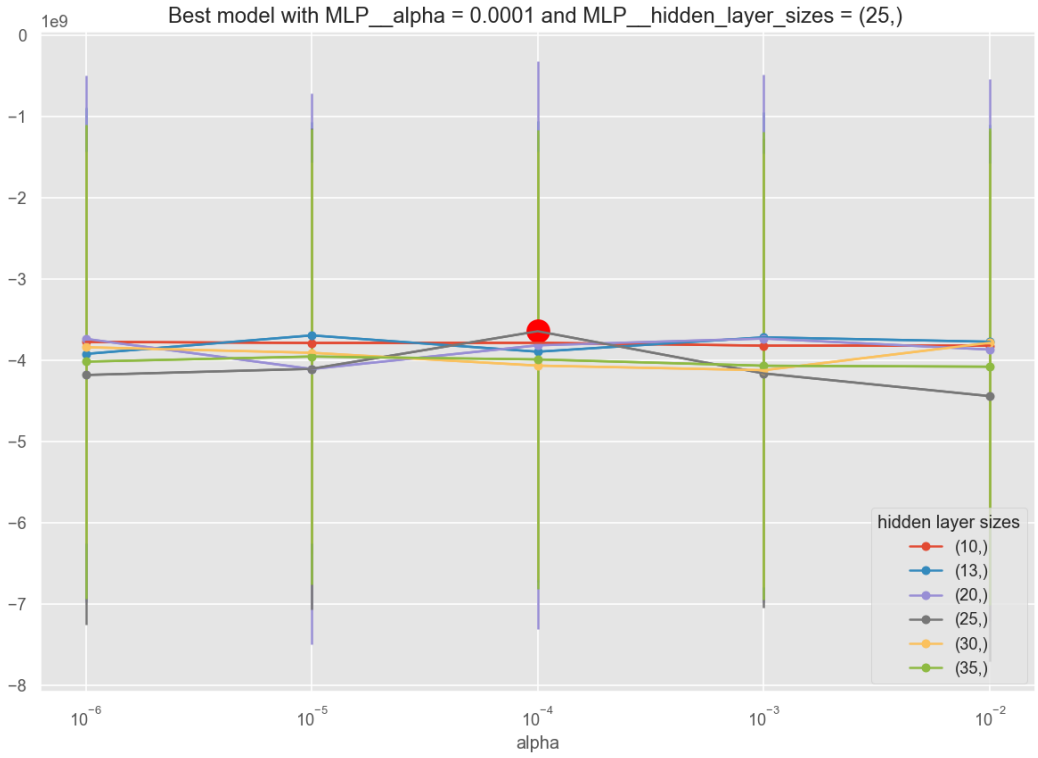
Table 4.2: SARIMAX results

5. Nonlinear Techniques

We are going to use a Multilayer Perceptron (MLP) to model our univariate series. In order to achieve this, we must first transform our variable into multiple variables from which the model can learn. To do so, six lag variables have been added to the model (lags 1, 2, 3, 4, 5, and 6) to predict the total unemployment.

The hyperparameters (Figure 5.1) that best fit our data are:

$$\alpha = 0.0001$$
$$Hidden\ Layers = 25$$



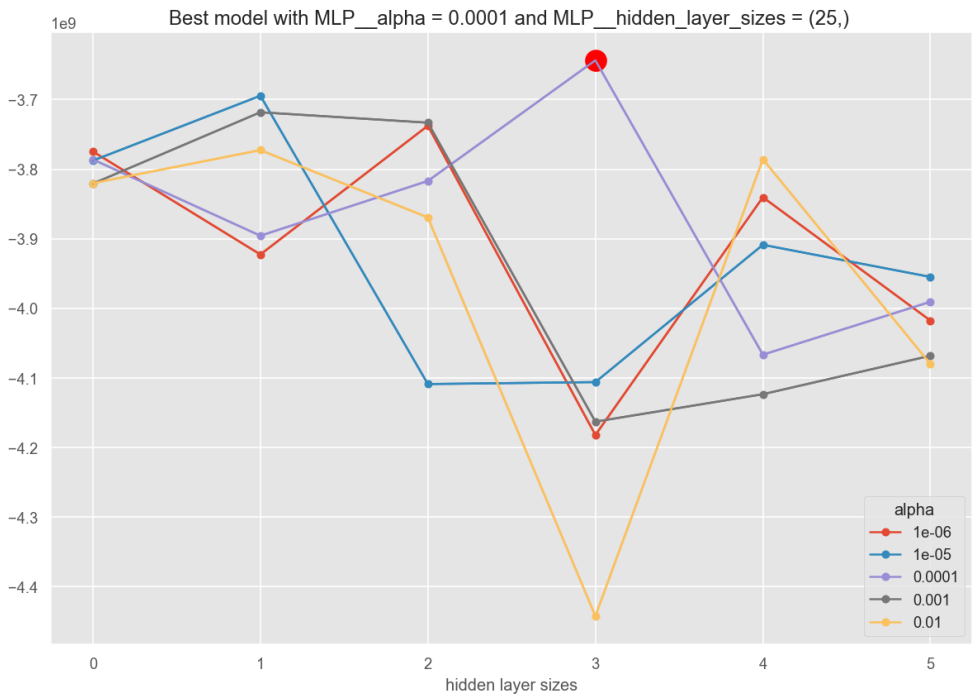


Figure 5.1: MLP hyperparameter selection

The variable that has more importance in our model is lag 1 followed by 4, 2, 5, 3, 6 as shown in Figure 5.2.

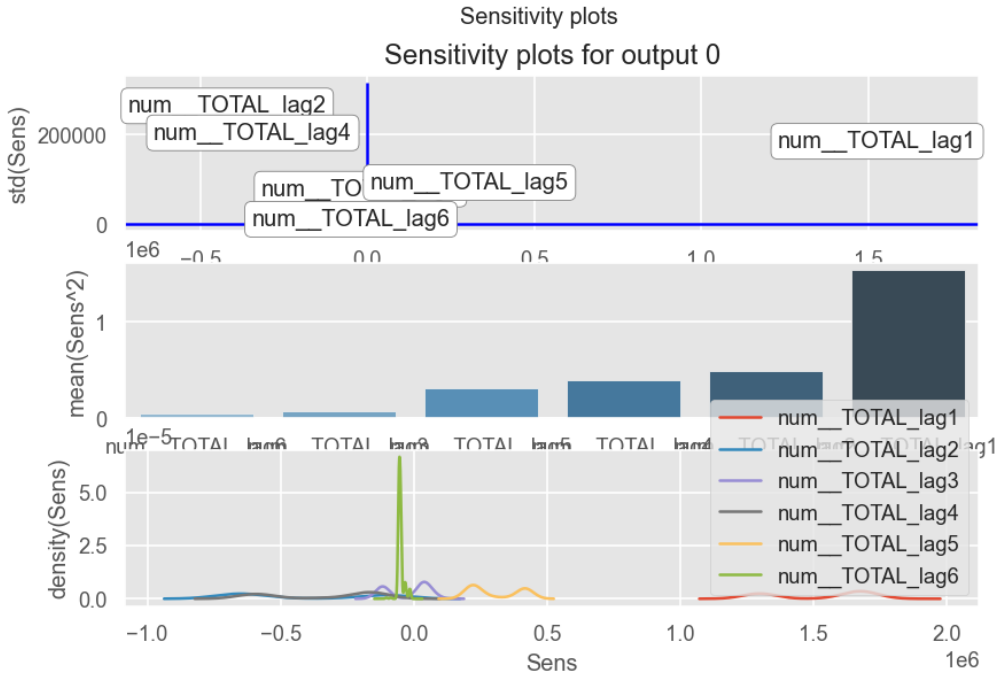


Figure 5.2: MLP sensitivity plots

The model results (Table 5.1) reveal a good fit as well as the residual plots (Figure 5.3).

MLP(0.01, 25)	MAE	RMSE	R2
Training	35801.10	48982.60	0.99
Test	61201.93	79512.87	0.95

Table 5.1: Results table

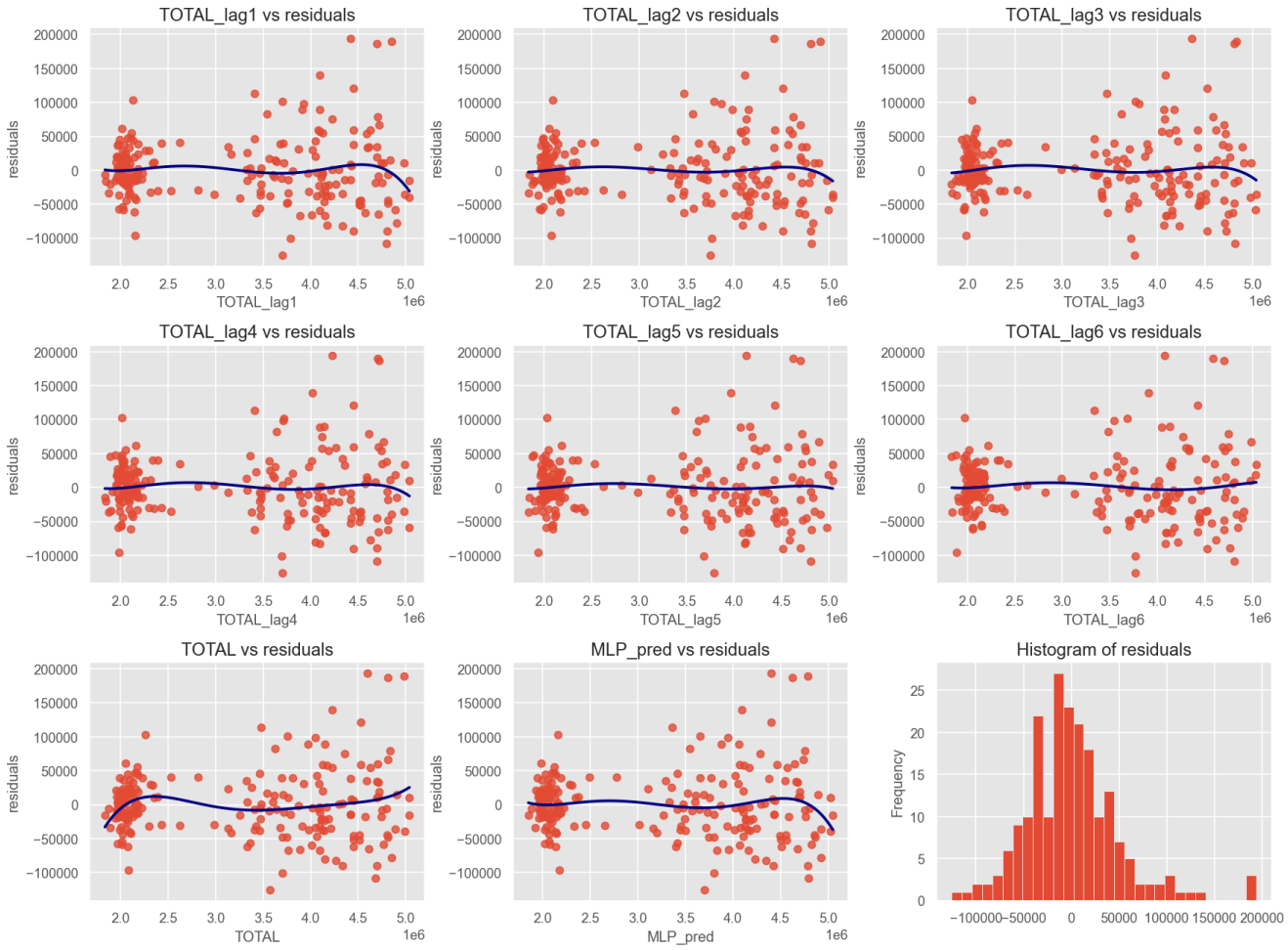


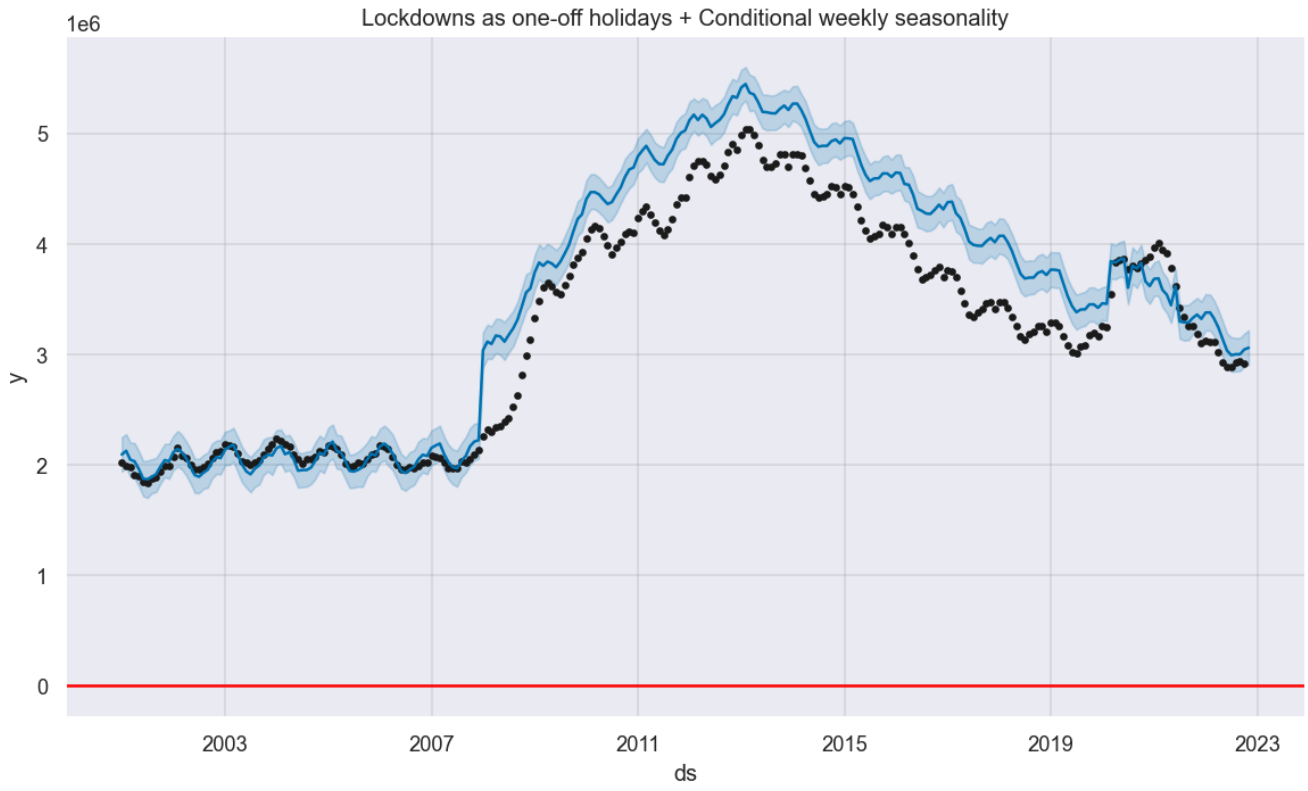
Figure 5.3: MLP residual plots

6. Unemployment Forecast November 2022

In order to forecast the unemployment rate we are going to compare the results between two different models: Prophet and SARIMAX . Prophet is a kind of model that implements a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly,

weekly, and daily seasonality, plus holiday effects. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

In order to get an accurate prediction using the Prophet Model we have introduced the different periods of lockdown as holidays and we have taken into account the pre and post periods of crisis and covid. To implement this we have introduced 4 different types of seasonalities, one for each of the seasonalities.



In addition, we have established just 1 month of prediction. The Unemployment Rate forecasted for November 2022 was 3.058.783.
Using SARIMAX's Model we have obtained an Unemployment Rate of 2.868.528.

To conclude, we have chosen as the best prediction the one obtained with the SARIMAX model, i.e. our forecast for the Unemployment Rate of November 2022 is 2.868.528.