|  |  |  |
| --- | --- | --- |
| **A picture containing text  Description automatically generated** | Faculty of Computing, Engineering and Science |  |

**Assessment Cover Sheet 2022-23**

|  |  |  |  |
| --- | --- | --- | --- |
| Module Code:  MS4S09 | Module Title:  Data Mining and Statistical Modelling | | Module Team:  Rebecca Peters/Angelica Pachon |
| Assessment Title and Tasks:  Assessment | | | Assessment No.  1 |
| Date Set:  dd-mmm-yyyy. | | Submission Date:  **14-Feb-23** | Return Date:  dd-mmm-yyyy. |

**IT IS YOUR RESPONSIBILITY TO KEEP RECORDS OF ALL WORK SUBMITTED**

|  |
| --- |
| **Marking and Assessment** |
| This assignment will be marked out of 100%  This assignment contributes to 50% of the total module marks. |
| **Learning Outcomes to be assessed** (as specified in the validated module descriptor [https://icis.southwales.ac.uk/](https://icis.southwales.ac.uk/studentmodules/11368/studentmodulespecifications) ):  Learning Outcome 1:  To understand techniques for interrogating and evaluating complex datasets, and to design macros for extraction of patterns and relationships  Learning Outcome 2:  Critically analyse, interpret and evaluate the outputs of statistical modelling techniques to support useful insights from complex datasets. |
| **Marking Criteria/Marking Scheme**  Marking guidelines |
| **Feedback Method**  Via Grade Centre on Blackboard |
| *Provisional mark only: subject to change and / or confirmation by the Assessment Board* |

**INTRODUCTION**

In this study, we will examine the unemployment rate in the UK during a 22-year period, from 2000 to 2022, paying particular attention to the impact of the COVID-19 pandemic on the employment rate, and also forecast the future rates of the unemployment rate. The null hypothesis is that since the Covid-19 outbreak, the unemployment rate has increased.

We will be working on a time series dataset. Time series is a series of data points listed in the time sequence. Time series analysis is a statistical method of analyzing past data within a given time period to predict the future. It consists of a sequence of data orders at equal intervals.

The dataset is gotten from the below URL.

<https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/timeseries/mgsx/lms>

**EXPLORATORY DATA ANALYSIS**

Firstly, we convert the dataset into time series using the ts() function and define the start date, end date, and frequency, the data is collected and recorded every 12 months, so we set the frequency value to 12.





figure 1: converting dataset to time series.

Figure 2 shows the data’s summary statistics, giving a general idea of the distribution of the data. This is achieved using the summary() function



Figure 2: Summary of data

The plot below shown in figure 3 shows as the year goes on, the time series shows a substantial degree of fluctuation. It demonstrates a rising trend from 2009 to 2013 and a descending trend from 2013 to 2019. Although seasonality is difficult to detect, it is evident that the time series has a trend and is not stationary.

Chart, histogram

Description automatically generated

figure 3: Time series of Unemployment rate in the UK (2000 - 2022)

To get a clearer view of the trend shown in the above plot, the seasonal effect can be removed by aggregating the data to the annual level, which can be achieved by using the aggregate function shown in figure 4.

Chart, line chart

Description automatically generated

figure 4: aggregate()

To provide a clearer picture, the plot in figure 5 only displays a fraction of the period from 2008 to 2019. From the plot, we can see the ascending trend indicating that the unemployment rate climbed in 2008, a steady trend until 2011, then an increased variation in 2012, followed by a decreasing trend until 2018, indicating that the unemployment rate fell throughout this time.

Chart, line chart

Description automatically generated

figure 5: Plotting a portion of the time series

Figure 6 illustrates the trend, seasonal and random components of the time series separately in a graphical manner as above. The underlying time series is a non-stationary time series as it has a trend or a seasonal component and requires differencing to transform it into a stationary time series.

Chart

Description automatically generated

Figure 6: Decomposition of the Unemployment rate data

**MODEL FITTING**

To check for stationarity, the autocorrelation coefficients are plotted, this shows the autocorrelation function or ACF. This plot is also known as a correlogram, it is plotted in R using the acf function.

For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly. Also, the ACF plots can be used to observe if the data has a trend and seasonal component. If there is a trend or seasonal component in the time series, then it cannot be stationary.

When data have a trend, the autocorrelations for small lags tend to be large and positive because observations nearby in time are also nearby in size. So, the ACF of trended time series tend to have positive values that slowly decrease as the lags increase.

When data are seasonal, the autocorrelations will be larger for the seasonal lags (at multiples of the seasonal frequency) than for other lags.

When data are both trended and seasonal, you see a combination of these effects.

The plot in figure 7 shows that the autocorrelation slowly decreases as the number of lags increases, this exhibits the trait of non-stationary data.

Chart, bar chart

Description automatically generated

Figure 7: Correlogram of the unemployment rate

Conducting a statistical test is one approach to determine whether a time series is stationary. The augmented Dickey-Fuller (ADF) test is one of the statistical tests that check for stationarity in a time series. It is the t-statistic of the estimated coefficient of a from the method of least squares regression.

ADF tests the null hypothesis that a unit root is present in an autoregressive time series model.

(H0): If accepted, it suggests the time series has a unit root, meaning it is non-stationary. It has some time-dependent structure.

(H1): The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary.

p-value > 0.05: Accept H0, the data has a unit root and is non-stationary

p-value ≤ 0.05: Reject H0. the data does not have a unit root and is stationary.

Figure 7, shows the result of the ADF test, the p-value(0.761) which is not less than the significance level of 0.05. Then we accept the null hypothesis that the time series is non-stationary.

Text

Description automatically generated

Figure 7: ADF test

Another test is the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, In this test, the null hypothesis is that the data are stationary. Figure 8, shows the result of the KPSS test with a p-value of 0.01 which means that we accept the null hypothesis that our time series is stationary.

Graphical user interface, text, application, email

Description automatically generated

Figure 8: Kpss test

We will use differencing to try and make our non-stationary series into stationary series. The first difference was applied to the time series; figure 9 displays the first difference's time plot. As can be seen, both the upward and downward trends have been erased, which is a good thing as it looks better than the previous plot, but the seasonality is still present this can also be seen in figure 10.

Chart

Description automatically generated

Figure 9: Time plot of the first difference

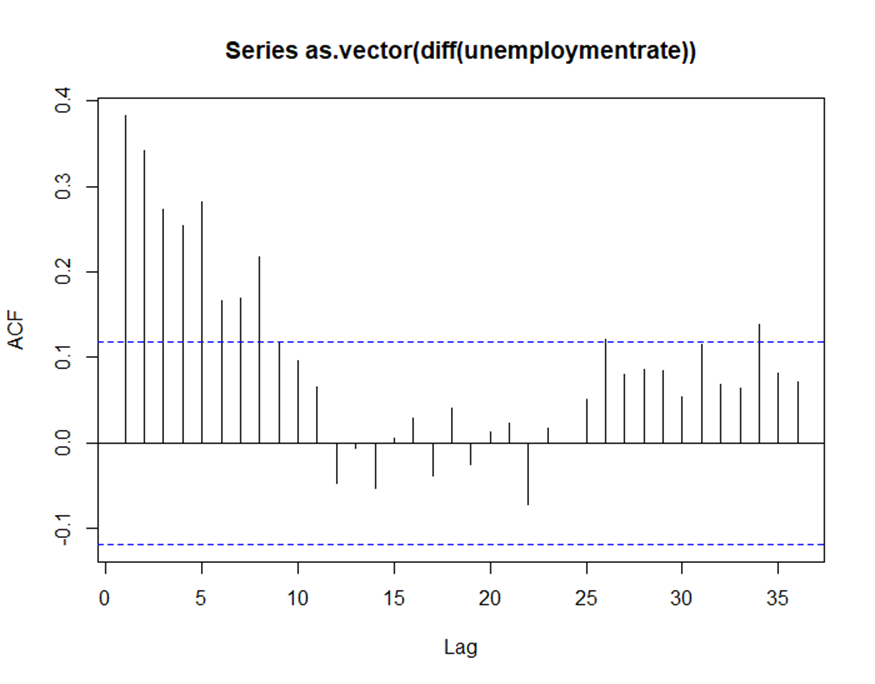


Figure 10: Plot of ACF of the First Differences

Next, we take both a first difference and a seasonal difference, figure 11 shows the time series plot of the first difference and seasonal difference. The seasonality seems to be gone for the most part, if not completely.

Graphical user interface

Description automatically generated with low confidence

Figure 11: Time Series Plot of First and Seasonal Differences of the Unemployment rate

The following ACF, PACF, and EACF plots have been analyzed to determine the parameters initially. The ACF plot in figure 12 demonstrates that approximately 95% of the lags are within the interval of confidence. However, very few substantial autocorrelations are outside of the interval of confidence. Lag 1 and 22 are significantly different from zero, but the remaining lags are all zero or very close to zero. This would lead us to believe that the time series is stationary, the ACF plot has been used to determine the moving average (MA) component of the model specification. Thus, ACF depicts both a seasonal moving average component (q) and a non-seasonal moving average component (Q). Therefore, it could be assumed that q is either 1 or 2, and Q is 1. The first difference (d) was set as 1.

Chart, box and whisker chart

Description automatically generated

Figure 12: ACF

The PACF plot shown in figure 13, has a few significant autocorrelations outside the interval of confidence. However, we can assume that 95% of the data falls within the confidence interval. The remaining lags are all zero or almost zero, except for lags 1, 2, 3, 4, and 7, which are significantly different from zero. Hence, it could assume that the time series is stationary. PACF plot illustrates the autoregressive (AR) component of the model specification. The time series in this analysis consists of both non-seasonal autoregressive components(p) and seasonal autoregressive components (Q). Significant peaks are at the beginning of the lags. Lag 25 is not considered as it may lead to an increase in the errors of the model).

Chart

Description automatically generated

Figure 13: PACF

The EACF plot is shown in figure 14, the result is not so clear, and it is difficult to determine an appropriate model.

Calendar

Description automatically generated

Figure 14: EACF Model

**CHOOSING THE BEST MODEL**

It would be suitable to use the seasonal ARIMA(p, d, q)(P, D, Q)s model for the time series in this study that has nonseasonal orders (p, d, q), seasonal orders (P, D, and Q), and seasonal period S. ACF, PACF plots and AIC criteria have been used in the process of model specification & parameter estimation of the model. After numerous tests, the model that produces the lowest value for AIC(-528.18) is (1,1,1)(0,0,1)[12] so, we will take this model to be the best model.

A picture containing table

Description automatically generated

**RESIDUAL ANALYSIS**

We first have a look at the time series plot of the residuals to validate the estimated ARIMA (1,1,1) X (0,0,1)12 model. The standardized residuals plot is shown in figure 15. This plot does not indicate any significant model abnormalities, aside from some odd behavior in the middle of the period, though we may need to dig at the model more thoroughly for outliers because the 2009 standardized residual looks suspicious.

A picture containing bar chart

Description automatically generated

Figure 15: Residuals from the ARIMA(1,1,1) X (0,0,1)12 Model

The ACF of the residuals shows no significant autocorrelations. Although lag 22 appears to be close to the significant line, 95% of the lags are inside the interval of confidence. This plot is shown in figure 16.

Chart

Description automatically generated

Figure 16: ACF of Residuals from the ARIMA(1,1,1) X (0,0,1)12 Model

The p-value for the Ljung-Box Q test is 0.7614, which is well above 0.05, so we have no evidence to reject the null hypothesis that the error terms are uncorrelated.

Letter

Description automatically generated with low confidence

In the histogram of the residuals, the shape is somewhat “bell-shaped” but certainly not ideal. This is shown in Figure 17.

Chart, histogram

Description automatically generated

Figure 17: Residuals from the ARIMA(1,1,1) x (0,0,1)12 Model

Q-Q plot: It doesn't seem too bad, but once more, the upper tail contains the one outlier. The plot is shown in figure 18.

Chart, line chart

Description automatically generated

Figure 18: Residuals: ARIMA(1,1,1) x (0,0,1)12 Model

The Kolmogorov-Smirnov test of normality has a test statistic of D = 0.13442, leading a p-value of 0.1, while the Shapiro Wilk has a test statistics W = 0.99242, leading to a p value of 0.1756, and normality is not rejected at any of the usual significance levels for the KS test.

Graphical user interface, text, application

Description automatically generated

Checking the stationarity of the residuals, we can see that the p-value (0.01) is less than the significance level of 0.05 and therefore we can reject the null hypothesis and accept stationarity.

Text

Description automatically generated

Stationarity of residuals for SARIMA model

Next, we consider fitting the ARIMA(1,1,1) x (0,0,1)12 model with the results shown below

A picture containing table

Description automatically generated

**FORECASTING**

In this module, we use the SARIMA model for a predictive goal for a full cycle of the future unemployment rate in the UK. Figure 19 shows the forecasts and 95% forecast limits for a lead time of two years for the ARIMA(1, 1, 1) X (0, 0, 1)12 model that we fit. The forecasts mimic the stochastic periodicity in the data quite well, and the forecast limits give a good feeling for the precision of the forecasts. From the plot, we can see that the unemployment rate in the future maintained a steady trend as there was no increase or decrease in the employment rate.

Chart

Description automatically generated

Figure 19: Forecasts and Forecast Limits for the Unemployment ARIMA(1,1,1) x (0,0,1)12

Figure 20 displays the last year of observed data and forecasts for four years. At this lead time, it is easy to see that the forecast limits are getting wider, as there is more uncertainty in the forecasts.

Chart

Description automatically generated

Figure 20: Long-Term Forecasts for the Unemployment ARIMA(1,1,1) x (0,0,1)12

Figure 21 below shows the comparison of predictions for the year 2022, where we used the model to predict the values for 2022 and compared it to the test data(actual values). We can see that the predicted values are closer to the actual values.

Table

Description automatically generated with low confidence

Figure 21: Comparison of Predictions for the year 2022

**FORECAST ERRORS AND FORECAST ACCURACY**

A forecast error is a difference between an observed value and its forecast. Here error does not mean a mistake, it means the unpredictable part of an observation. It can be written as Note that forecast errors are different from residuals in two ways. First, residuals are calculated on the training set while forecast errors are calculated on the test set. Second, residuals are based on one-step forecasts while forecast errors can involve multi-step forecasts.

Figure 22 shows the result from the forecast accuracy.



Figure 22: Evaluation metric for selected models

ME (Mean Error): A value close to zero indicates that the model is correctly predicting the mean of the test data. In this case, the ME is -0.19, which is close to zero, indicating a good model fit.

RMSE (Root Mean Squared Error): This metric measures the average magnitude of the error. The lower the value of RMSE, the better the model fit. In this case, the RMSE is 0.21, which is low, indicating a good fit.

MAE (Mean Absolute Error): This metric measures the average magnitude of the error. The lower the value of MAE, the better the model fit. In this case, the MAE is 0.19, which is low, indicating a good fit.

MPE (Mean Percentage Error): This metric measures the average percentage error. A value close to zero indicates that the model is accurately predicting the test data. In this case, the MPE is -5.22, which is low, indicating a good fit.

MAPE (Mean Absolute Percentage Error): This metric measures the average magnitude of the percentage error. The lower the value of MAPE, the better the model fit. In this case, the MAPE is 5.22, which is low, indicating a good fit.

ACF1 (Autocorrelation of the residuals at lag 1): This metric measures the residual autocorrelation. A value close to zero indicates that the residuals are uncorrelated, meaning that the model has captured all the relevant information in the data. In this case, the ACF1 is 0.38, which is close to zero, indicating a good fit.

Theil's U: This metric measures the ratio of the prediction mean square error to the prediction mean square error of a naïve model that predicts the mean of the dependent variable for all observations. The lower the value of Theil's U, the better the model fit. In this case, Theil's U is 2.20, which is low, indicating a good fit.

**CONCLUSION**:

In this study, the time series for the UK's unemployment rate were examined. The time series was made stationary, a list of potential models was created, and model fitting and residual analysis were carried out. Finally, to forecast the time series, Seasonal ARIMA(1,1,1) x (0,0,1)12 was fitted. We may infer that this is the best model for estimating the rate of unemployment by examining the forecasting error and forecast accuracy results. We may also observe how covid affected the UK's unemployment rate and projected rates for the years to come.