**FORECASTING ASSIGNMENT**

**PART 1**

Juan Jose Medina

**PROBLEM**

For the following assignment we are going to analyze the air passengers by Belgium and EU partner countries. To determine if we are able to make an accurate forecast on the year 2020 using the previous data, and the best model based on the different aspects and factors from the time series.

As soon as we plot the data we can see a huge difference in the timeline, but we are going to omit this for our forecast.

**DATA EXPLORATION**

Time Series Plot

Chart

Description automatically generated

As our first impression we ca see that there seems to be a high seasonality based on the fluctuations from the graph and a trend going upwards.

Chart, line chart

Description automatically generatedChart

Description automatically generated

Seeing the seasonal plot we can see a clear spike ion the month of July in every single year, and then starts to descend up to the end of the year.

On the seasonal subseries plot we can again confirm the seasonality on, that is exemplified by the month of July, and also see the trend that is going upward on every month.

Chart, histogram

Description automatically generatedChart

Description automatically generated

Next on our ACF graph we determine that there are peaks at lag 12,24 and 36 indicating a seasonality of length 12, meaning that its monthly based, and the slow decay indicates a that there exists a trend on the timeline.

On the PACF graph we can see a significant correlation at lag 1, following by a correlation on lag 2, but also on lag 3, this means we should use an an AR(3) model

**TRANSFORMATION**

After analyzing if the time series needs transformation, based on the lambda value that we got of -0.03592379, we decide to apply a Box Cox transformation, since the value is so close to zero the transformation will work similar to a logarithmic transformation.

Since we are using mainly the forecast function, we can apply the transformation on every forecasting using the lambda parameter, this will allow us to transform the data using the Box Cox transformation and then reverse it to its original scale without extra steps.

**FORECASTING**

As our first model to run we run a Seasonal Naïve method forecast

Chart, histogram

Description automatically generatedGraphical user interface, chart, histogram

Description automatically generated

We can clearly appreciate that the prediction considers the seasonality of the timeline.

And after analyzing the residuals we see that they have a close to normal distribution, but the mean is not on zero rater its to the rights side. Also on the ACF we can see that there is a lot of information on different lags that there is information on the residuals that could be used in the forecasting. To confirm this we realized a Ljung – Box Test.

Text

Description automatically generated

Based on this test we can see that on lag 1 to 10 all the P values are below 0.05 meaning that we reject the null hypothesis that the residuals are independently distributed, in other words they show a correlation.

Text

Description automatically generated

Finally we check the accuracy based on RMSE, MAE, MAPE and MASE and we see that the model has a high error compared to the test data, meaning that the model is best fitted to predict the future rather than the past

**DECOMPOSITION**

Since the data is seasonal, we want to try to use the STL decomposition to adjust the forecast and see if we can get better results

Diagram

Description automatically generated

Based on the decomposition we can clearly see the seasonality factor and the trend; in this way we can use non seasonal methods to predict and then apply these factors again to the model to make the forecast seasonal.

For the next experiment we used the naïve, rwdrift, ets and arima methods and compare them.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

After making the four different forecasts we check the accuracy on all of them

Text

Description automatically generated

We can see that all models performed worse than the regular naïve and comparing these 4 we can conclude that the first model that is the naive is the best performing one having the lowest MAE, MAPE and MASE, RMSE.

To further compare the models we checked the residuals and see that on the first model the Seasonal Naïve with STL

Graphical user interface

Description automatically generated with medium confidenceChart

Description automatically generatedChart

Description automatically generatedHistogram

Description automatically generated with medium confidence

Naïve RWDRIFT ETS ARIMA

There is information that is not taking into account at lag 1 ,2, 12 and 24, for the Random walk with drift + STL we have similar results as the previous, for ETS we have a significance of the residuals at lag 12 and a only by a small margin on lag 2 and 24, finally the ARIMA model have a significance on lag 12 and lag 24. After comparing the residuals, we see that the 3rd model, ETS, is the one that fits the data the best using the decomposition.

**ETS**

For the next step in the experiment, we use the Exponential smoothing methods, here we selected the AAA, an error additive, additive tend with additive seasonal, the MAM, multiplicative error, additive trend and multiplicative season model ,and a AAA, Additive damped method. With bias adjusted Additionally we selected the auto ets model that will try to find the best model automatically.

Text

Description automatically generated

Based on the accuracy we can see that the that the ZZZ model is the best one based on the RMSE, MAE, and MASE, and second place on the MAPE, the first model based on the MASE tells us that this model is not good at predicting out of sample forecasts

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart

Description automatically generatedChart

Description automatically generated

AAA MAM AAdA ZZZ

On ly on the first model we see that there is information wthat is not white noise on lag 6 , 12,18,25, while the rest every single lag is not significant, meaning that the residuals can be interpreted as simply white noise

For the last model we can see that the parameters estimated was ETS(A,Ad,A)

Chart, histogram

Description automatically generated

**AUTO ARIMA**

As the last model we want to compare we applied an Auto.ARIMA

Chart

Description automatically generatedA picture containing graphical user interface

Description automatically generated

On this model we can see that there is only a small loss of information at lag 36, while the remain stays significant.



The errors of the auto arima are higher than the rest that we have been using and based on its MASE its not good at making out of sample predictions.

Text

Description automatically generated

The parameters for the Auto Arima where the following ARIMA(1,0,2)(2,1,0)[12] with drift

**COMPARISON**

Finally we compare the 4 main models, the Seasonal Naïve, the STL, ETS, AUTO.ARIMA

Text

Description automatically generated

Based on the accuracy test and MASE we can see that the SNAIVE, STL and ETS models are able to predict well out of sample, having Snaive with the lowest errors, followed by ETS

Graphical user interface, chart, histogram

Description automatically generatedChart

Description automatically generated with low confidenceChart

Description automatically generatedChart

Description automatically generated

SNAIVE STL ETS ARIMA

After seeing the residuals, we see that only the third model, the ETS that was recognized as the most stable model is the only one that all the residuals can be seen as white noise, meaning that it includes all the information for every lag. Even if the Snaive have better results, the ETS model contains white noise on the residuals

Therefore the best model for this data set is the ETS model with the ETS(A,Ad,A) parameters.

Chart

Description automatically generated

**OUT OF SAMPLE**

After making a forecasting with the data from January 2003 up to February 2020, we can see that the forecast seems to follow the seasonality and the trend of the previous data

Chart, histogram

Description automatically generated

**COVID**

Chart

Description automatically generated

But due to unforeseen circumstances that occurred in the world, with the global pandemic Covid-19, the forecast was unable to predict a total drop on the number of flights since the economy went to a total stop for several months, this allows to know that a forecast model can only work if the information is correct and if the variables remain untouched. Since the pandemic is an unforeseeable disaster that has no seasonality, nor trend nor cycle, there is no way that any model can be able to predict its impact on the flight industry.