Time Series Forecasting on Migration Data

Anand Kumar Shanmugam
shanmu01@ads.uni-passau.de

Chaitanya Gogineni
gogine01@ads.uni-passau.de

Nishanth Kandaswamy Subramanian
subram02@ads.uni-passau.de

Vishal Sowrirajan
sowrir01@ads.uni-passau.de

ABSTRACT

With the huge amount of data available, many sectors use various scientific methods and algorithms to retrieve valuable insights from these overabundant data. Based on reliable sources, the UNHCR agency has estimated the migration arrivals through the Western Balkan route on a daily basis. Time series forecasting is predicting the future events based on the past observations of a time series data. Therefore, the migration arrival dataset can provide valuable information to predict the upcoming migration pattern. This project focuses on developing three different time series models (ARIMA, SARIMAX, LSTM) on the migration dataset and forecasting the the migration arrivals of the different countries in the Western Balkan route. Using evaluation metrics such as Mean Absolute Error and Root Mean Square Error, we were able to compare the performance of the different models.

PHASE RESPONSIBILITY

Name	Phase
Nishanth Kandaswamy Subramanian	1
Anand Kumar Shanmugam	2
Vishal Sowrirajan	3
Chaitanya Gogineni	4

PHASE CONTRIBUTION

Phase	Contribution
Phase 1	Chaitanya, Nishanth, Vishal
Phase 2	Implementation : Chaitanya, Nishanth Report: Anand, Vishal
Phase 3	Implementation : Chaitanya, Nishanth Report: Anand, Vishal
Phase 4	Chaitanya, Nishanth

INTRODUCTION

In general, forecasting[1] can be defined as predicting the information of a future event or condition on the basis of currently available data. Time series forecasting, in particular, can be described as predicting the future based on the past observation of a time series [2]. Developing models that understand past data and predict the future outcomes has become a dynamic area of research which has gained attentions of the research community. Due to accurate predictions of time series models, it plays a vital role in several fields such as finance, science and engineering, etc[2].

The United Nations High Commissioner for Refugees (UNHCR) agency is part of the United Nations that functions to protect the refugees seeking help from different parts of the world and the primary purpose of the agency is to provide assistance to refugees in critical emergency with protection, shelter, and healthcare [13]. Since the beginning of the 21st century, the UNHCR has been involved with major refugee crisis in Africa, Asia and the Middle East [13].

The information in [3, 6] describes the routes and difficulties of migrants in reaching the EU countries. Before 2015, many migrants used different active routes to reach the countries of the European Union. Most of whom were forced to migrate, took immensely dangerous routes such as crossing the Mediterranean Sea to reach Italy and other countries of the European Union. In this process, many migrants lost their lives in shipwreck or were exploited by human smugglers [3, 6].

The introduction of the EU Agenda on Migration and Operation Sophia [6] in the mid-2015, re-routed the path migrants through the Western Balkan route. Even though the European Commission initially supported the cooperation between different EU states, it was later ruled by the Court of Justice of the European Union (CJEU) that it was not in accordance with the EU legislation which eventually led to the closure of the route [3, 6].

1 PROBLEM STATEMENT

Based on reliable information sources, the UNHCR has collected estimates on the number of migrants moving to each of the countries in the Western Balkan route on a daily basis [4]. Based on the available Balkan route dataset, this project aims to follow an univariate analysis to forecast the upcoming migration count of a particular country and compare the performance evaluation of different models. ARIMA, SARIMAX and LSTM model have been built on the given dataset and evaluated with different performance metrics. Mean Absolute Error (MAE) and Root Mean Squared Error are used to evaluate the performance of the models.

Section 1 gives a brief introduction to time series and a few time series models. Section 2 describes the various steps involved in the data preprocessing. Section 3 explains the implementation of different time series models. Finally, section 4 evaluated the different models and compares them based on the performance metrics.

1.1 Time Series

From [5], time series can be defined as a information recorded sequentially over a time period. Over the past few years, developing time series models that understand past data and predict the future

[1] https://www.merriam-webster.com/dictionary/forecast

Data Science Lab, 2019, Uni Passau
.

outcomes has become a dynamic area of research which has gained attentions of the research community [2]. With increasing research work for the development and enhancement of forecasting accuracy, the forecasting models have been evolved over time [2].

Based on different categories, the time series can be classified into different types. Considering the number of records, the time series can be univariate or multivariate [2]. The *univariate time series* refers to the set of observations recorded from a single variable over time and the *multivariate time series* refers to changing values taken from multiple variables over time [2]. Also, based on the statistical property (such as mean and variance), time series can be categorized as Stationary an Non-Stationary [8]. A time series is *stationary* if the statistical properties remains constant during the various time intervals, whereas a time series is *non-stationary* if the statistical properties changes over the time [2, 8].

1.2 Time Series Models

ARIMA

The Autoregressive Integrated Moving Average (ARIMA) is a combination of differencing along with the autoregression model and the moving average model [5]. It consists of three different notations (p, q, d) representing the order of the autoregressive model, the level of difference and the order of moving average respectively [**?**]. An autoregression model forecasts values based on the linear combination of the past values of the given variable and a moving average model forecasts based on the past forecast errors [5]. Mathematically, an ARIMA model of a differenced time series can be represented as [5],

$$\begin{aligned} \mathrm{Y_t} = \mathrm{c}+\phi_1\mathrm{Y_{t-1}} + \phi_2\mathrm{Y_{t-2}} + \ldots + \phi_\mathrm{p}\mathrm{Y_{t-p}} + \mathrm{e_t} - \\ \theta_1\mathrm{e_{t-1}} - \theta_2\mathrm{e_{t-2}} \ldots - \theta_q\mathrm{e_{t-q}} \end{aligned} \tag{1}$$

SARIMAX

The SARIMAX model is an extension of Seasonal ARIMA model incorporated with an exogenous variable in order to improve the performance of the forecasting model [15]. In the migration dataset, the time period of the 'EU-Turkey cut-off' deal has been mentioned. Based on this, an additional column pertaining to the 'EU-Turkey cut-off' deal has been added to the dataset. The new column contains values '0' prior to the cut-off deal and '1' from the period when the deal came into existence. This information has been incorporated as the exogenous variable while developing the SARIMAX model.

LSTM

Long Short-Term Memory (LSTM) neural networks are a special type of RNN intended to address the problem of long term dependencies in which the architecture of LSTMs comprises of units called memory blocks, each of which contains an input gate, an output gate, and a forget gate [1]. The input gate controls the flow of input activations into the memory block, the output gate controls the flow of activations from the memory block to the rest of the network and the Forget gate estimates the internal state of the cell before adding it as an input to the memory block and performs forgetting or resetting the memory block [1].

Workflow

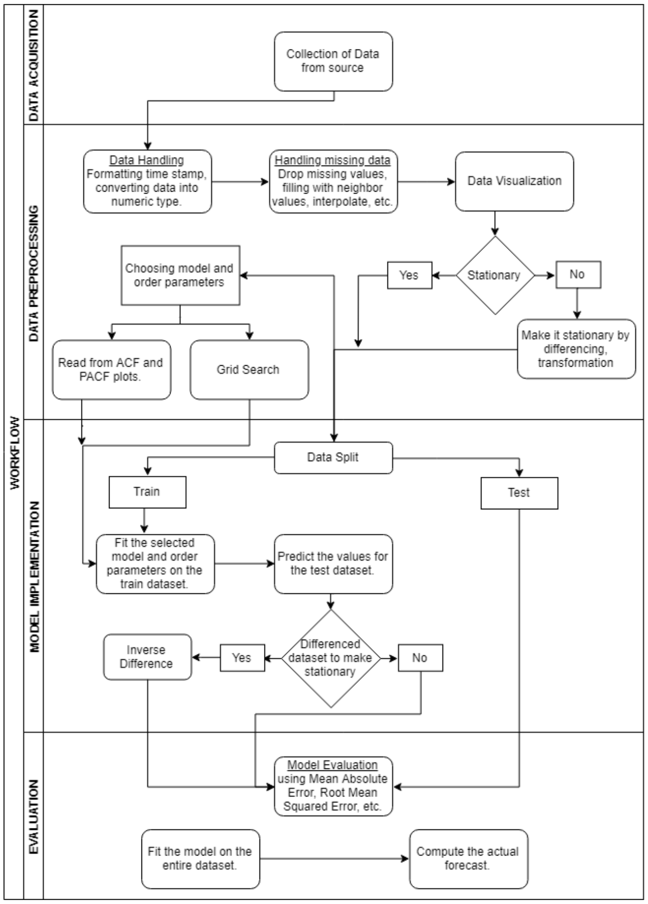
WORKFLOW
DATA ACQUISITION
DATA PREPROCESSING
MODEL IMPLEMENTATION
EVALUATION
Collection of Data from source
Data Handling
Formatting time stamp, converting data into numeric type.
Handling missing data
Drop missing values, filling with neighbor values, interpolate, etc.
Data Visualization
Stationary
Yes
No
Make it stationary by differencing, transformation
Choosing model and order parameters
Read from ACF and PACF plots.
Grid Search
Data Split
Train
Test
Fit the selected model and order parameters on the train dataset.
Predict the values for the test dataset.
Differenced dataset to make stationary
Yes
No
Inverse Difference
Model Evaluation
using Mean Absolute Error, Root Mean Squared Error, etc.
Fit the model on the entire dataset.
Compute the actual forecast.

Figure 1: Workflow representation

2 DATA ACQUISITION AND PREPROCESSING

2.1 Data Acquisition

The UNHCR has collected estimates on the number of migrants moving to each of the countries in the Western Balkan route on a daily basis, based on the information obtain from reliable sources such as UNHCR border teams, authorities and humanitarian partners [4, 14]. The western balkan route dataset[2] has been obtained from the Humanitarian Data Exchange[3]. The downloaded dataset contains the information about the daily arrivals of migrants to different countries in the Western Balkan route and the time period of 'EU Turkey cut-off' deal. The collected data is stored in a csv format and further modified such that suitable time series models can be developed.

2.2 Frameworks used

The following open source libraries and frameworks were used as part of the project implementation. The primary data manipulation

[2] https://data.humdata.org/dataset/daily-estimated-arrivals-through-western-balkans-route
[3] https://data.humdata.org/

and data analysis tasks were achieved by using the Pandas[4] and Numpy[5]. Seaborn[6] and MatplotLib[7] functions were used in the visualization of data. The time series models such as ARIMA, SARIMAX and LSTM were developed using the Statsmodels[8], pmdarima[9] and Keras[10]. The models were evaluated using the performance metrics from sklearn[11].

2.3 Data Pre-processing

As mentioned in the workflow diagram (figure 1), the preprocessing technique involves steps such as data handling (like format handling, missing data), data visualization, analyzing time series characteristics, etc. The 'Arrivals to Austria' time series data has been considered in order to explain each of these data preprocessing steps.

2.4 Data Handling

Initially, the dataset is loaded into a dataframe from the csv file. Using the date time formatting, the time stamp in the data is formatted and made as the index of the dataset. Also, it can be noted that the migration count of each country is of string datatype. With the help of regular expression and numeric functions, the data is converted into numeric datatype.

After analyzing the dataset, it can be found that some of the countries have missing or NaN values. As removing the missing data leads to loss of information, suitable methods to handle missing data can be applied. As the migration count of the countries show a particular trend at different instances of time, the interpolation method was used to handle the missing data. It retains the pattern of the time series as compared to methods such as ffill or bfill.

Data Visualization

Visual representation of data helps grab attention to detail and provide better understanding on information such as trends, outlier, pattern [12]. For instance in figure 2, the time series data pertaining to three different countries (namely Austria, Croatia and Italy) have been visualized using the line plot function. It can be noted that the migration count for Austria and Croatia gradually decreases over time, while Italy has a fluctuating migration for the same time period. Also, after the time period of 'EU Turkey cut-off deal' (mentioned in the dataset), the migration count in Austria remains almost constant and the migration count in Croatia becomes zero.

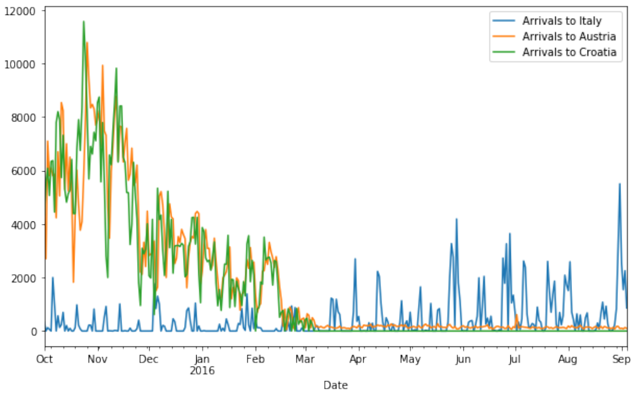

Figure 2: Line plot visualization

In order to better understand the characteristics of a given time series, they can be decomposed into several components [5, 10]. In figure 3, it can be seen that the 'Arrivals to Austria' data has been decomposed into trend, seasonal and residual components. While the trend component shows the overall trend pattern of the migration in Austria, the seasonal component explains the periodic pattern of people migrating to Austria that persists over a fixed time period. It can be noted that it follows a weekly pattern. Also, it can be observed that the time series has a varying statistic (such as variation in mean, variance and covariance), which brings to the conclusion that 'Arrivals to Austria' is non-stationary [8].

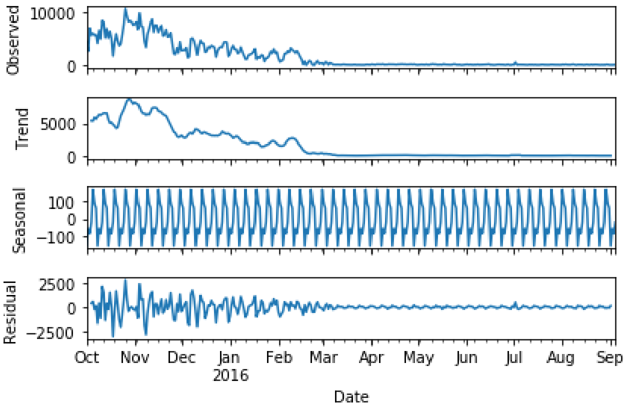

Figure 3: Decomposition of 'Arrivals to Austria' time series to understand the time series components

2.5 Stationarity Tests

Two statistical tests are used to check the stationarity of a time series. Augment Dickey Fuller test and Kwiatkowski-Phillips-Schmidt-Shim test are used to evaluate the stationarity of the migration time series.

Augmented Dickey Fuller Test

In the Augmented Dickey Fuller Test [10], the null hypothesis is that there exists a unit root and the alternate hypothesis is that there

[4] https://pandas.pydata.org/
[5] https://numpy.org/
[6] https://seaborn.pydata.org/
[7] https://matplotlib.org/
[8] https://www.statsmodels.org/stable/index.html
[9] https://www.alkaline-ml.com/pmdarima/
[10] https://keras.io
[11] https://scikit-learn.org/stable/

is no unit root. Based on [10], the ADF test has been formulated into the table 1 and the stationarity of a time series is determined.

Table 1: ADF Test Criteria

Test Scenario	Result	Inference
p-value < Critical value	Reject Null Hypothesis. There is no unit root.	Stationary
p-value > Critical value	Fail to reject Null Hypothesis. There exists a unit root.	Non-Stationary

Kwiatkowski-Phillips-Schmidt-Shim Test

In the KPSS Test [5, 10], the null hypothesis is that the time series is trend stationary. Based on [5, 10], the KPSS test has been formulated into the table 2 and the test is executed.

Table 2: KPSS Test Criteria

Test Scenario	Result	Inference
Test Statistics < Critical value	Fail to reject Null Hypothesis.	Trend Stationary
Test Statistic > Critical value	Reject Null Hypothesis.	Non-Stationary

Based on the adf and kpss tests, the stationarity of the different time series has been obtained as follows,

Table 3: Stationarity Test Result

Time Series	Stationarity
Arrivals to Italy	True
Arrivals to Greek Islands	False
Departures to mainland Greece	False
Arrivals to fyRoM	False
Arrivals to Serbia	False
Arrivals to Croatia	False
Arrivals to Hungary	True
Arrivals to Slovenia	False
Arrivals to Austria	False

2.6 Handling Non-Stationary Time Series

Based on the ADF and KPSS tests, the stationarity of the different time series has been identified. For instance, from the table 3, it can be found that 'Arrivals to Austria' is non-stationary. Using techniques such as differencing, log transformation these non-stationary time series can be made stationary [5]. By applying the differencing technique, the differences between the consecutive observation are computed to make a time series stationary [5]. Apart from visually analyzing the differenced data, the ADF and KPSS Tests can also be used to statistically verify if the differenced data is stationary or needs to be further differenced with a higher order.

2.7 ACF and PACF Plots

Based on the key properties of the ACF and PACF plots from [5, 8], the following insights about the time series have been obtained. While the ACF plot is slowly decaying indicating that the time series has a trend, the PACF plot displays a sharp cutoff. Even though the ACF plot slowly decays, it does not decay to zero. This emphasis that the Austria time series is non-stationary. A high number of lags have a positive correlation in ACF plot, verifies the need for differencing to make the time series stationary.

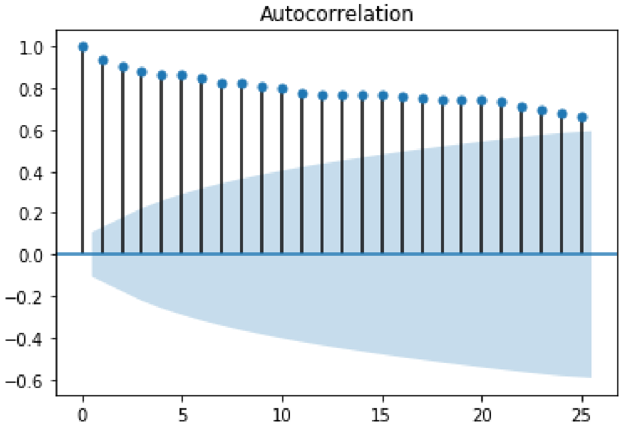

Figure 4: Auto correlation plot for 'Arrivals to Austria' with lag values from 0 to 25

These plots can also be used to determine the AR and MA parameters [5, 8]. The PACF plot has a sharp cutoff and eventually gets smaller, but does not decay to zero. Thus, the series requires a moving average parameter. The significant positive correlation at higher lags in the ACF plots indicate the need for autoregressive parameter.

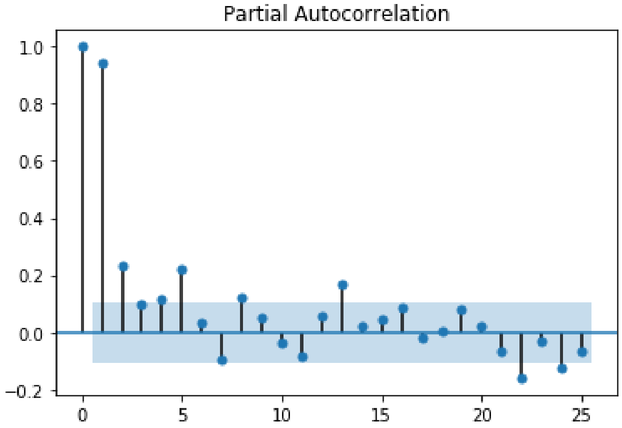

Figure 5: Partial auto correlation plot for 'Arrivals to Austria' with lag values from 0 to 25

3 MODEL IMPLEMENATION

In order to describe the various steps involved in building a forecasting model, the 'Arrivals to Austria' time series data has been considered.

3.1 Train and Test Split

An important part prior to developing a model is to separate the dataset into train and test data. From [7], it can be noted that, initially the model is built on the train data and it predicts for the given test data range. The comparison between predicted and test data helps to evaluate the performance of the models [7]. Ideally, the 'Arrivals to Austria' time series dataset has been split such that approximately 80% of the dataset is train data and 20% of the dataset is test data. Also, the dataset has been split in a chronological order.

3.2 Autoregressive Integrated Moving Average

In order to build an ARIMA model on the given dataset, the corresponding ARIMA parameters (p, d, q) needs to be obtained. To find the optimal parameters for the given time series, the auto_arima function (from pmdarima library) is used on the entire time series data. For 'Arrivals to Austria' time series, the auto_arima function returns the values (1,1,1) as the optimal values for (p, d, q) based on the Akaike Information Criterion [11].

Before making the actual forecast values, the model needs to be built on the train data and evaluated on the test data. Using the ARIMA function (from statsmodels library) and the obtained (p, d, q) parameter values are fitted on the train data to build the model. Then, the model forecasts the values in the range of the test data. Figure 6 shows the visual representation of the forecast of ARIMA model for the given test data range against the actual test data. This comparison helps to evaluate the model's performance with the help of evaluation metrics (which has been described in the Evaluation section).

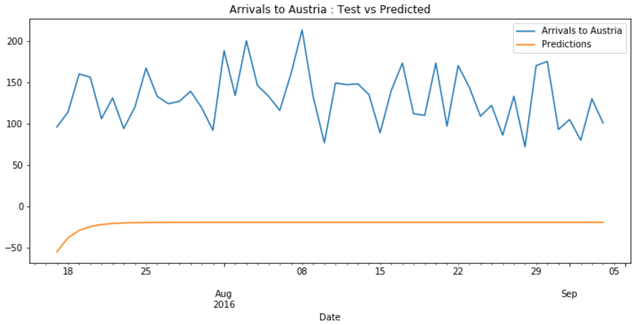

Figure 6: ARIMA : Test vs Predicted

3.3 SARIMA with Exogenous Variable

The SARIMAX model is an extension of Seasonal ARIMA model incorporated with an exogenous variable in order to improve the performance of the forecasting model [15]. With the available time period for 'EU Turkey cut-off' deal in the dataset, a new column (namely 'EU Turkey Cut-Off Deal') pertaining to this deal has been added to the dataset. Similar to the ARIMA function, the parameters of the SARIMAX needs to be identified. Also, both the 'Arrivals to Austria' time series and the exogenous variable data have been split into train and test data.

Apart from the (p, d, q) parameters of an ARIMA model, the SARIMAX additionally has (P, D, Q, s) parameters. With seasonal value set as True and seasonal differencing (m) set as 7, the auto_arima function is applied on the 'Arrivals to Austria' time series along with exogenous variable data. For this dataset, the optimal parameters have been identified as (4,1,3) for (p, d, q) and (0,0,1,7) for (P, D, Q, s) by the auto_arima function using the Akaike Information Criterion [11]. To build the model, the obtained parameter values are used in SARIMAX function on the train data (containing both the time series and exogenous variable data). Then, by passing the exogenous variable values of the test data, the model forecasts the values for the test data range. Figure 7 shows the visual representation of SARIMAX forecasts for the given test data range against the actual test data.

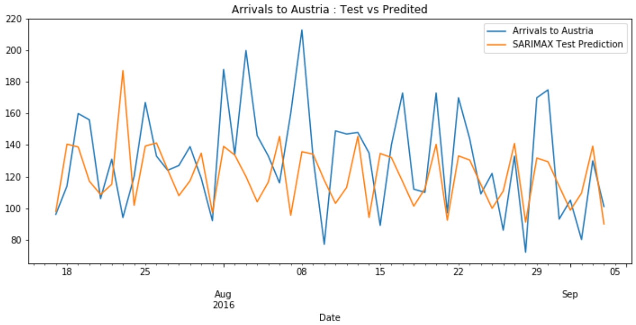

Figure 7: SARIMAX : Test vs Predicted

3.4 Long Short Term Memory

The foremost step in the implementation of the LSTM model for time series is that, the dataset is converted into a supervised learning problem using an user-defined method. Then, the time series is checked for the stationarity. In case the data is not stationary, the time series is made stationary by differencing the data and later the inversed while forecasting.Followed by this, the model uses an activation function (tanh) and transforms the dataset in the scale -1 to +1. Finally, the model is fit on the train data and forecasts for the test data range (figure 8).

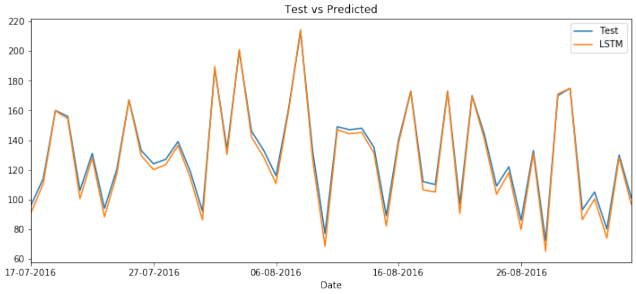

Figure 8: LSTM : Test vs Predicted

3.5 Models with sliding window

In order to better understand the model's performance on the dataset, the models are built and evaluated on a sliding window dataset. Initially, the first 100 records of the time series (say 'Arrivals to Austria') are taken as the train data and the next 30 records as the test data. The model is built on this train data set and it forecasts for the test data range. The forecasted and the test data are compared to evaluate the model's performance. Followed by this, the first 130 records are taken as the train data and the next 30 as the test data. And the same performance evaluation procedure is repeated. Likewise, the train dataset is increased sequentially with additional 30 records in every iteration until the entire dataset becomes the train data. This allows to compare the model's performance at the different phases of the time series (discussed in evaluation section). The complete procedure is evaluated on all the time series with the ARIMA, SARIMAX and LSTM models.

4 EVALUATION

The performance of a model can be evaluated by comparing the predicted data and the actual test data [7]. As part of the project, the evaluation or performance metrics (Mean Absolute Error and Root Mean Square Error) are used to evaluate the model's performance.

Mathematically, the mean absolute error can be defined as [9],

$$MAE = \frac{\sum_{t=1}^{n} |Y_\mathrm{t} - \hat{Y}_\mathrm{t}|}{n} \tag{1}$$

where $\hat{Y}_t$ is the predicted value and $\mathrm{Y_t}$ is the true value. Also, based on [9] the mean square error can be defined as,

$$MSE = \frac{\sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2}{n} \tag{2}$$

where $\hat{Y}_t$ is the predicted value and $\mathrm{Y_t}$ is the true value. The root mean square error is the square root of the mean squared error.

With the use of evaluation metrics, the performance of each models on the different time series is evaluated. The mean of corresponding test dataset is calculated, which when compared to the results of the evaluation metrics, gives a clear picture on the performance of the different models. Also, this can be used to compare the models amongst each other and across all the different time series.

Table 4, 5, 6 describes the performance metrics on the different countries for ARIMA, SARIMAX and LSTM models respectively.

Table 4: Performance Evaluation : ARIMA

Country	**Error Values**		
	Mean	**MAE**	**RMSE**
Arrivals to Italy	843.56	846.08	1356.96
Arrivals to Greek Islands	100.18	119.26	140.69
Departures to mainland Greece	38.12	61.92	71.97
Arrivals to fyRoM	0.0	22.00	24.78
Arrivals to Serbia	244	260.66	265.71
Arrivals to Croatia	0	24.26	27.17
Arrivals to Hungary	38.64	51.16	52.26
Arrivals to Slovenia	0	30.47	34.71
Arrivals to Austria	130.8	151.23	154.59

Table 5: Performance Evaluation : SARIMAX

Country	**Error Values**		
	Mean	**MAE**	**RMSE**
Arrivals to Italy	843.56	734.83	1089.26
Arrivals to Greek Islands	100.18	120.74	146.50
Departures to mainland Greece	38.12	141.66	187.67
Arrivals to fyRoM	0.0	69.08	82.17
Arrivals to Serbia	244	44.87	46.34
Arrivals to Croatia	0	73.87	94.87
Arrivals to Hungary	38.64	59.53	62.14
Arrivals to Slovenia	0	464.33	498.95
Arrivals to Austria	130.8	25.86	33.78

Table 6: Performance Evaluation : LSTM

Country	**Error Values**		
	Mean	**MAE**	**RMSE**
Arrivals to Italy	843.56	622.86	865.35
Arrivals to Greek Islands	100.18	64.78	94.86
Departures to mainland Greece	38.12	21.26	25.60
Arrivals to fyRoM	0.0	6.32	6.32
Arrivals to Serbia	244	20.95	25.75
Arrivals to Croatia	0	2.69	2.69
Arrivals to Hungary	38.64	11.17	15.56
Arrivals to Slovenia	0	2.84	2.84
Arrivals to Austria	130.8	39.36	47.88

After the evaluation of the model's performance, the actual future forecast needs to be estimated. Instead of the train/test data, the obtained optimal parameters are used to fit the model on the entire time series data and the actual unknown future values are predicted. For instance, considering the ARIMA model for the 'Arrivals to Austria' time series, the optimal values (1,1,1) for the parameters (p, d, q) is used to fit the model on the entire time series. With this, the actual future values are predicted as shown in figure 9.

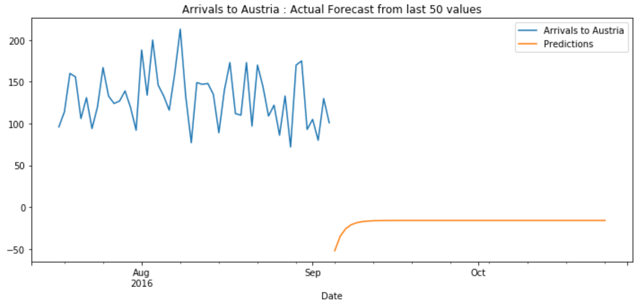

Figure 9: ARIMA : Actual Forecast

Similarly, for SARIMAX and LSTM models, the models are fitted on the entire dataset and the actual future values are forecasted. Figures 10 represents the actual future values predicted for 'Arrivals to Austria' time series using SARIMAX model.

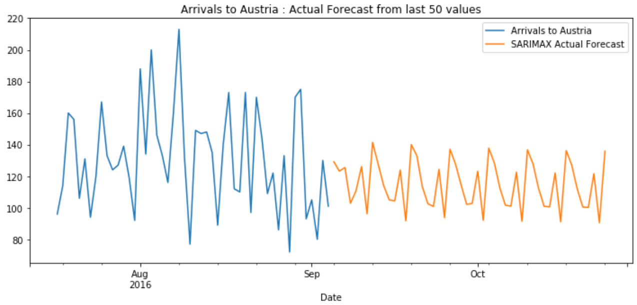

Figure 10: SARIMAX : Actual Forecast

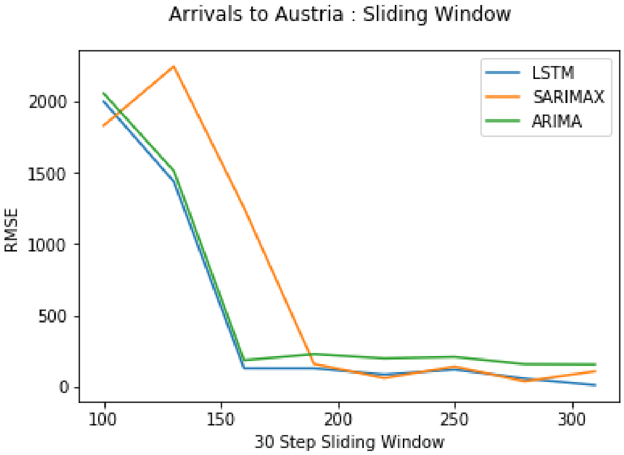

Figure 11: Sliding Window : ARIMA

Figure 11 represents the model's performance on the different phases of the 'Arrivals to Austria' dataset. The train dataset is increased sequentially with additional 30 records in every iteration until the entire dataset becomes the train data. This allows to compare the model's performance at the different phases of the time series. It can be noticed that the rmse value decreases along iteration and the LSTM performs better compared to other models for 'Arrivals to Austria'.

The detailed results of the models on the different countries can be found here [12].

5 CONCLUSION

In this project, different forecasting methods such as ARIMA, SARIMAX, LSTM were developed on the given migration dataset to forecast the future arrival estimates of migrants on a daily basis. The performance of the models were evaluated on different performance metrics such as Mean Absolute Error and Root Mean Sqaure Error. Also, using sliding window technique on the dataset, the model's performance on the different phases of the dataset were calculated. Based on the results of the performance metrics, it can be found that for the given dataset, the ARIMA and SARIMAX models provided good prediction in most of the time series, but comparatively the LSTM model performed better in most of the time series.

[12] https://github.com/ChaitanyaGogineni/DataScienceLab

REFERENCES

[1] Guy Pujolle Abdelhadi Azzouni. 2017. A Long Short-Term Memory Recurrent Neural Network Framework for Network Traffic Matrix Prediction. (2017). https://arxiv.org/abs/1705.05690

[2] Ratnadip Adhikari and R. Agrawal. 2013. *An Introductory Study on Time series Modeling and Forecasting*. https://doi.org/10.13140/2.1.2771.8084

[3] Frontex European Border and Coast Guard Agency. [n. d.]. Migratory Routes. ([n. d.]). Retrieved May 19, 2019 from https://frontex.europa.eu/along-eu-borders/migratory-routes/western-balkan-route/

[4] The Humanitarian Data Exchange. [n. d.]. Daily Estimated Arrivals through Western Balkans Route. ([n. d.]). Retrieved May 19, 2019 from https://data2.unhcr.org/en/documents/details/47375

[5] Rob J Hyndman and George Athanasopoulos. 2018. *Forecasting: principles and practice* (2nd. ed.). OTexts, Melbourne, Australia. Retrieved May 19, 2019 from http://OTexts.com/fpp2/.

[6] Green European Journal. [n. d.]. The Western Balkan Route: A New Form of Forced Migration Governance in Europe? ([n. d.]). Retrieved May 19, 2019 from https://www.greeneuropeanjournal.eu/the-western-balkan-route-a-new-form-of-forced-migration-governance-in-europe/

[7] R. Medar, V. S. Rajpurohit, and B. Rashmi. 2017. Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning. In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*. 1–6. https://doi.org/10.1109/ICCUBEA.2017.8463779

[8] Robert Nau. 2019. Statistical forecasting: notes on regression and time series analysis. (2019). Retrieved September 14, 2019 from https://people.duke.edu/~rnau/411home.htm This web site contains notes and materials for an advanced elective course on statistical forecasting that is taught at the Fuqua School of Business, Duke University.

[9] D. Purwanto, C. Eswaran, and R. Logeswaran. 2010. A Comparison of ARIMA, Neural Network and Linear Regression Models for the Prediction of Infant Mortality Rate. In *2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation*. 34–39. https://doi.org/10.1109/AMS.2010.20

[10] Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

[11] Taylor G. Smith et al. 2017–. pmdarima: ARIMA estimators for Python. (2017–). Retrieved September 14, 2019 from http://www.alkaline-ml.com/pmdarima

[12] Tableau. [n. d.]. Data visualization beginner's guide: a definition, examples, and learning resources. ([n. d.]). Retrieved September 14, 2019 from https://www.tableau.com/learn/articles/data-visualization

[13] UNHCR. [n. d.]. UNHCR, The UN Refugee Agency. ([n. d.]). Retrieved May 19, 2019 from https://www.unhcr.org/

[14] UNHCR. 2013. Balkan Route dataset. (Jan. 2013). Retrieved June 09, 2019 from https://www.unhcr.org/statistics/STATISTICS/45c06c662.html#DATA_COLLECTION_METHODS

[15] S. I. Vagropoulos, G. I. Chouliaras, E. G. Kardakos, C. K. Simoglou, and A. G. Bakirtzis. 2016. Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting. In *2016 IEEE International Energy Conference (ENERGYCON)*. 1–6. https://doi.org/10.1109/ENERGYCON.2016.7514029