Analysis of Website Traffic Time Series Forecasting using ARIMA, Prophet, and LSTM RNN

Suraj Katwal[1], Rubim Shrestha[2], Dr. Gajendra Sharma[3]

[1]*MCA, Kantipur City College, Dept. of Science and Technology, Kathmandu 44600, Nepal*
[2]*Assistant Professor, Kantipur City College, Dept. of Science and Technology, Kathmandu 44600, Nepal*
[3]*Associate Professor, Kathmandu University, Dept. of Computer Science and Engineering School of Engineering, Dhulikhel 45200, Kavre, Nepal*

Abstract

The rapid growth of the Internet has led to a vast increase in website traffic data. Accurately forecasting website traffic is essential for informed decision-making and future planning. This study comprehensively analyzed four methods for forecasting website traffic time series data: Autoregressive Integrated Moving Average (Hereafter ARIMA), Prophet, Long Short-Term Memory (Hereafter LSTM), and Hybrid Long Short-Term Memory - Gated Recurrent Unit Recurrent Neural Network (Hereafter LSTM-GRU RNN). The study used Wikipedia Pageviews Dataset using API and Google Analytics data to train and test the forecasting models. The empirical analysis evaluated the accuracy and ability of each method to capture trends and seasonality. The results showed that the LSTM-GRU model with 50 epochs has the lowest MSE value of 0.0057022 and the lowest RMSE value of 0.075513. The LSTM model with 100 epochs has a low MSE value of 0.0057916 and a low RMSE value of 0.0761028, comparable to the LSTM-GRU model with 50 epochs.

Keywords: Web Traffic; ARIMA; Prophet; LSTM RNN; LSTM-GRU; Prediction; Time Series

1. INTRODUCTION

The increasing population has led to a greater familiarity with the internet worldwide. This has resulted in increased traffic almost everywhere. As a consequence, the company is experiencing a number of problems due to the increased traffic to its website. [4]. The company should invest additional resources to accommodate the increased traffic. This plan could include investing in more servers. The traffic report analyzed earlier can be used to determine if a website management technique plan needs to be implemented to reduce the risk of such mishaps [2,4]. One way to reduce or handle such problems in the future is by forecasting web traffic. Web traffic forecasting is the process of estimating the future size of a web traffic segment or group. This can be done by analyzing historical data, tracking online behavior, or using modeling techniques [7]. Web traffic time series forecasting uses historical web traffic data to predict future web traffic patterns [8]. This can help businesses decide about website design, capacity planning, and marketing campaigns. Web traffic forecasting can be used for various purposes, including estimating website traffic, understanding website trends, and forecasting future traffic volumes [9]. Several methods can be used to forecast web traffic, including time series and regression analysis. In order to determine the best forecast method, we must consider the type of data and the degree of accuracy we want [10]. Time-series analysis is a well-established approach for forecasting web traffic, as it can capture short-term patterns and long-term trends [11]. Regression analysis can identify relationships between different factors and web traffic [12]. In addition to creating more complex models, machine learning can also adapt to changing data over time.

2. RELATED WORK

2.1 Forecasting Website Traffic Using Prophet Time Series Model

This paper proposes a method for forecasting website traffic using the Prophet time series model. The prophet time series model is a tool that can be used to predict future events based on past data. The authors argue that the prophet model is well suited for forecasting website traffic because it can handle seasonality and trend changes. The authors collected data on website traffic from Google Analytics and it was used to train the prophet model. The model was then used to predict and forecast website traffic for the future. The authors found that the prophet model was able to accurately forecast website traffic. The model was able to capture seasonality and trend changes in the data. The authors discuss the implications of their findings. They argue that the prophet model can be used to effectively forecast website traffic and could be used to forecast other time-series data [1].

2.2 Web Traffic Time Series Forecasting Using LSTM Neural Networks with Distributed Asynchronous Training

This paper presents a method for forecasting web traffic on a web server using a supervised Long Short-Term Memory (LSTM) recurrent neural network model. The authors created an updated version of the Wikipedia page views dataset from a Kaggle competition for the years 2018-2020, which they used to train and test the model. The model was designed to perform distributed training using a data parallelism approach and the Downpour training strategy. The authors found that the model was able to accurately predict web traffic for the seven dominant languages in the dataset, with a mean absolute error of less than 30. The authors also found that the model's accuracy was improved by the distributed training approach and by extracting features and hidden patterns in the data before training the model. The paper concludes that the proposed model is a significant step forward in the field of time series prediction for web traffic forecasting, and is able to achieve accurate predictions despite having limited data in the dataset [2].

2.3 Time-Series Analysis with Smoothed Convolutional Neural Network

In the current study, Andika Dwiyanto et al. aims to improve the accuracy of time-series prediction by using a smoothed convolutional neural network (SCNN). This paper proposes a new approach for time-series analysis using Convolutional Neural Networks (CNNs). CNNs are a type of neural network that is well-suited for image processing tasks. However, they have not been widely used for time-series analysis. The authors propose using a CNN with a special type of layer known as a "smoothing layer" to improve the accuracy of predictions. The CNN is trained on a dataset of historical time-series data. The smoothing layer is used to "smooth out" the input data, making it easier for CNN to learn patterns. The CNN is then able to make predictions about future values in the time series. The authors evaluate their approach on several time-series datasets. They find that their approach outperforms traditional methods, also the use of Lucas Hidden Layer raises the performance of forecasting algorithms [3].

2.4 Web traffic time series forecasting using ARIMA and LSTM RNN

The paper explores the potential of using ARIMA and LSTM RNN for web traffic time series forecasting. The authors analyzed the sample datasets from Wikipedia’s pageview API for the project; the data was for ‘India’ only. Furthermore, the time series was divided into two sets, one for the training and another for the testing sets. The time series data were total hits per day. Here, the authors applied Discrete

Wavelet Transform (DWT) methodology to divide data into two sections. They separate data into linear and non-linear components, the detailed data are kept in linear components, and approximate data are kept in non-linear components. For non-linear components, they applied the LSTM RNN model to forecast whereas the ARIMA model for linear components and combined to give the final forecast. The authors compare the two methods and find that LSTM RNN brings more efficiency than ARIMA in terms of accuracy. They also suggest that the use of the DWT and a combination of the two methods may be more effective than either method alone [4].

2.5 A Novel Time Series Forecasting Model with Deep Learning

This paper describes a novel time series forecasting model called SeriesNet, which aims to improve the accuracy of traditional time series forecasting models. SeriesNet is composed of two networks: an LSTM network, which is used to learn holistic features and reduce the dimensionality of multi-conditional data, and a dilated causal convolution network, which is used to learn different time intervals. The model is designed to learn multi-range and multi-level features from time series data, and is able to achieve higher predictive accuracy than models using fixed time intervals. The model also utilizes residual learning and batch normalization to improve generalization. The authors conduct experiments on several typical time series with the same datasets and found that SeriesNet has higher forecasting accuracy and greater stability than other models [5].

2.6 Comparison of RNN, LSTM, and GRU Methods on Forecasting Website Visitors

The thesis examines the application of deep learning techniques in forecasting website visitor numbers. Specifically, the authors compare three recurrent neural networks (RNN) models: RNN, LSTM, and GRU. The study utilizes two types of data: first-time visits and unique visits. Tests are conducted using epoch parameters ranging from 1 to 500 across layers 1, 3, and 5, and the data includes both first-time visits and unique visits. The results indicate that the LSTM method outperforms the other models, demonstrating the lowest mean squared error (MSE) values for both first-time visits and unique visits [6].

3. METHODOLOGY

The research involved a framework for predicting website traffic using ARIMA, Prophet, and LSTM RNN and creating a hybrid LSTM-GRU model. The process included data collection, pre-processing, cleaning, splitting, model training, evaluation, and comparison. The website traffic data was collected from collegesnepal.com over 5-year period and tracked by Google Analytics. The data was cleaned and divided into a training, validation, and test set in the ratio of 80:10:10. The ARIMA, Prophet, LSTM, and Hybrid LSTM-GRU models were trained and evaluated on the data, and their performance was compared. The performance of the models was evaluated and compared using metrics such as MSE and RMSE. Overall, this framework provided a structured approach for forecasting website traffic using ARIMA, Prophet, LSTM, and Hybrid RNN models and allowed for a thorough evaluation and comparison of the performance of these methods.

3.1 Workflow

We propose a hybrid model for forecasting web traffic time series. The model integrates two common recurrent neural network architectures, namely LSTM and GRU. We optimize the model's hyperparameters to achieve the best performance. We compare the performance of our proposed model with other established approaches as in Fig. 1 such as ARIMA, Prophet, and LSTM methods. Our goal is to determine the most efficient model for forecasting web traffic in time series analysis. We follow a specific methodology to construct the Hybrid LSTM-GRU Stack model as in Fig - 2. This methodology outlines the step-by-step process employed in this study to develop and evaluate our proposed forecasting model.

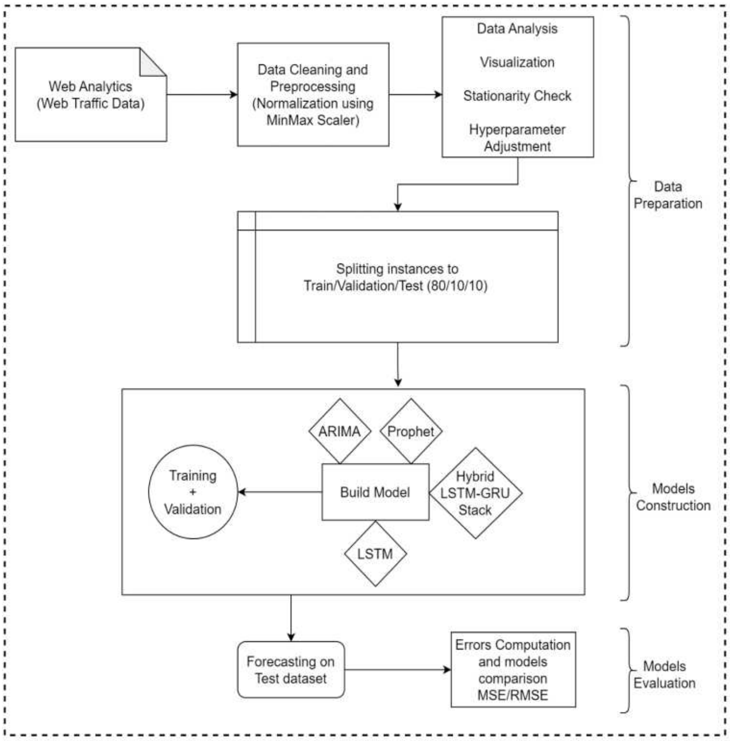

Fig. 1: Proposed Methodology for Time Series Web Traffic Forecasting

3.2 Proposed Hybrid LSTM-GRU Stack Model Architecture

The hybrid LSTM-GRU stack model architecture is a new approach to web traffic forecasting that combines the strengths of LSTM and GRU models. This architecture addresses the vanishing gradient problem and can effectively capture complex patterns in the data.

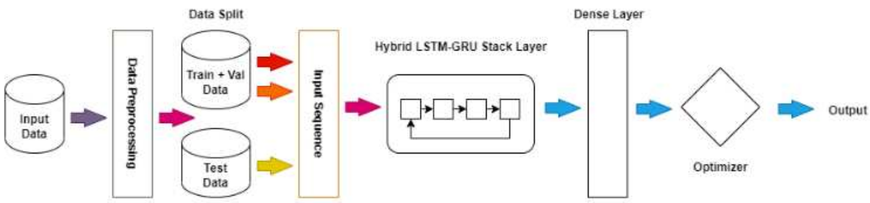

Fig. 2: Proposed Overview of Hybrid LSTM-GRU Stack Model Architecture

3.3 Data Collection

The analysis utilized website traffic data from collegesnepal.com over 5-year period (2017 to 2022) with date and daily traffic stats as a univariate dataset. Google Analytics was employed for data tracking. To enhance the analysis, traffic data from Wikipedia Pageviews covering 7 years (2015 to 2022) was obtained via an API for model validation and analysis improvement. The raw data was transformed into a suitable format for modelling.

3.4 Data Pre-Processing and Cleaning

The website traffic data was pre-processed and cleaned to remove any missing or invalid values. Necessary transformations were also applied to the data, such as scaling or normalization. Min Max normalization was used to normalize all models to fit the model.

3.5 Data Splitting

The website traffic data was split into a training set (80%), a validation set (10%), and a test set (10%) to train and evaluate the ARIMA, Prophet, and LSTM RNN models. The training set data helped to extract knowledge and data patterns; a 10% validation set was used to ensure the model captured the underlying patterns accurately and extracted knowledge was then tested on the test data.

3.6 Model Training

The ARIMA, Prophet, LSTM, and Hybrid LSTM-GRU models were trained on the training set using appropriate algorithms and hyperparameters. Any necessary data resampling techniques were also applied to the training data.

3.7 Model Evaluation

The performance of the ARIMA, Prophet, LSTM and Hybrid LSTM-GRU models was evaluated using the test set. Mean Square Error (MSE) and Root mean square error (RMSE) were used as an error or loss metric.

4. RESULT

4.1 Result Analysis Training Data

Tables 1 and 2 present the training data results for various models, including ARIMA, Prophet, LSTM, and LSTM-GRU. The smallest MSE and RMSE value in the table is 0.0014493 and 0.0380697 respectively, which is for the Hybrid LSTM-GRU model with 200 epochs and is the best-performing model.

Table 1: Training Data for All Models with (p,d,q)/epochs with MSE metrics

Model	(p,d,q)/Epochs	MSE
ARIMA	(1,1,1)	0.0018209
ARIMA	(2,1,4)	0.0017051
ARIMA	(3,1,2)	0.0018167

ARIMA	(5,1,3)	0.0017497
Prophet	-	0.005301
LSTM	5	0.002200587
LSTM	50	0.001657788
LSTM	100	0.001660538
LSTM	200	0.001664046
LSTM-GRU	5	0.0027758
LSTM-GRU	50	0.0020562
LSTM-GRU	100	0.001704
LSTM-GRU	200	0.0014493

Table 2: Training Data for All Models with (p,d,q)/epochs with RMSE metrics

Model	(p,d,q)/Epochs	RMSE
ARIMA	(1,1,1)	0.0426721
ARIMA	(2,1,4)	0.0412924
ARIMA	(3,1,2)	0.0426225
ARIMA	(5,1,3)	0.0418288
Prophet	-	0.072813
LSTM	5	0.046910413
LSTM	50	0.04071594
LSTM	100	0.040749703
LSTM	200	0.040792715
LSTM-GRU	5	0.0526861
LSTM-GRU	50	0.0453458
LSTM-GRU	100	0.04128
LSTM-GRU	200	0.0380697

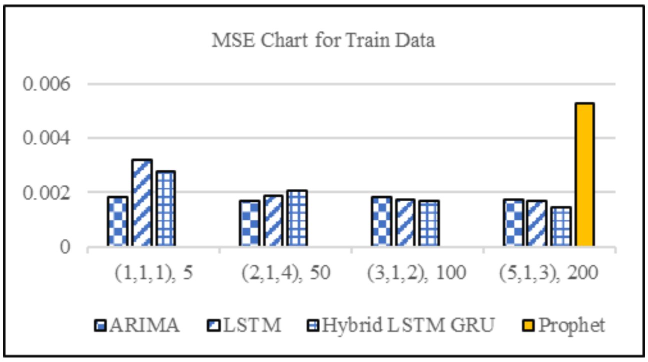

Fig. 3**:** MSE Chart for Training Data

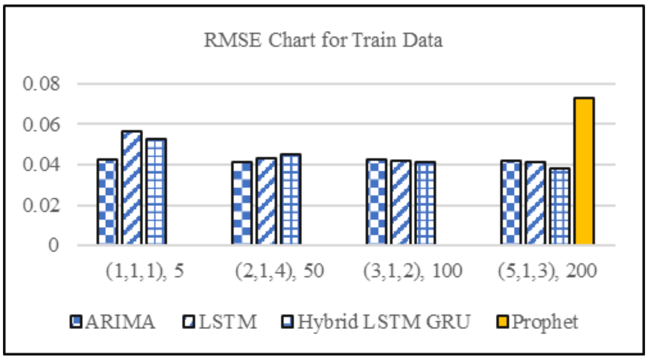

Fig. 4**:** RMSE Chart for Training Data

4.2 Result Analysis Test Data

Tables 3 and 4 present the test data results for various models, including ARIMA, Prophet, LSTM, and LSTM-GRU. The smallest MSE and RMSE values in the table are 0.0057022 and 0.075513 respectively, which is for the Hybrid LSTM-GRU model with 50 epochs and is the best-performing model.

Table 3: Test Dataset for All Models with (p,d,q)/epochs and MSE metrics

Model	(p,d,q)/Epochs	MSE
ARIMA	(1,1,1)	0.0319291
ARIMA	(2,1,4)	0.0301552
ARIMA	(3,1,2)	0.0319156
ARIMA	(5,1,3)	0.0305497
Prophet	-	0.0223472
LSTM	5	0.008509

LSTM	50	0.0058864
LSTM	100	0.0057916
LSTM	200	0.0060289
LSTM-GRU	5	0.0081866
LSTM-GRU	50	0.0057022
LSTM-GRU	100	0.0059244
LSTM-GRU	200	0.006205

Table - 4: Test Dataset for All Models with (p,d,q)/epochs and RMSE metrics

Model	**(p,d,q)/Epochs**	**RMSE**
ARIMA	(1,1,1)	0.1786873
ARIMA	(2,1,4)	0.1736525
ARIMA	(3,1,2)	0.1786494
ARIMA	(5,1,3)	0.1747846
Prophet	-	0.1494899
LSTM	5	0.0922444
LSTM	50	0.0767231
LSTM	100	0.0761028
LSTM	200	0.077646
LSTM-GRU	5	0.0904798
LSTM-GRU	50	0.075513
LSTM-GRU	100	0.0769704
LSTM-GRU	200	0.0787718

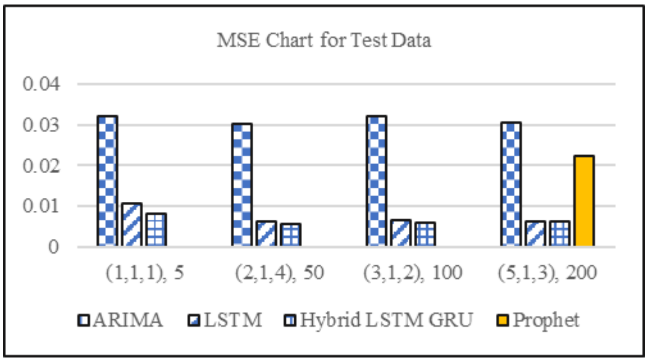

Fig. 5: MSE Chart for Test Data

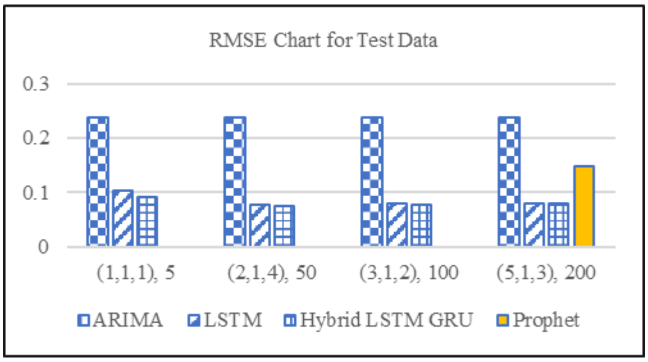

Fig. 6: RMSE Chart for Test Data

4.3 Summary of the Results

The mean squared error (MSE) and root mean squared error (RMSE) values were lower for the Hybrid LSTM-GRU Stack model than for the other models on the test data. This indicates that the Hybrid LSTM-GRU Stack model is better suited for this type of analysis and can generalize well to unseen data. Overall, the Hybrid LSTM-GRU Stack model performed better than the other models on the test data. The results suggest that the Hybrid LSTM-GRU Stack model can be a useful tool for analyzing website traffic.

5. CONCLUSIONS

The models created in this study can be leveraged to further develop the system for practical applications by utilizing highly effective models. It also contributes in the field of time series forecasting by demonstrating the effectiveness of deep learning models, specifically hybrid LSTM-GRU Stack, in forecasting web traffic. The results of this study can help practitioners and researchers to make informed decisions when selecting a model for web traffic forecasting tasks.
In conclusion, the study compared four approaches for web traffic forecasting: ARIMA, Prophet, LSTM, and Hybrid LSTM-GRU. Among them, the LSTM-GRU model emerged as the best performer in terms of prediction accuracy based on the test dataset. With the lowest MSE value of 0.0057022 and the lowest RMSE value of 0.075513, the LSTM-GRU model with 50 epochs outperformed the other models. The LSTM model also showed good results, with the best RMSE of 0.0767231 achieved with 50 epochs. These results suggest that Hybrid LSTM-GRU is a suitable model for web traffic forecasting that outperforms ARIMA, Prophet, and LSTM in terms of prediction accuracy based on test data. This research contributes to the field of time series forecasting by demonstrating the effectiveness of deep learning models, specifically Hybrid LSTM-GRU, in forecasting web traffic patterns.

ACKNOWLEDGEMENTS

I express my sincere gratitude to Mr. Rubim Shrestha, supervisor, for his continuous guidance, support, and constructive feedback throughout this research process. His expertise and insights were invaluable. I thank the faculty members of Kantipur City College, especially Assoc. Prof. Dr. Gajendra Sharma, Mr. Raju Kattel, Mr. Bhanu Bhakta Niroula, and Mr. Ramesh Parajuli, for their support and advice. I'm also grateful to Shailendra Basnet and Srijan Shah from Colleges Nepal for providing the dataset essential to my research. I appreciate the unwavering support and understanding of my family and friends during this journey. Finally, I acknowledge all the authors whose works have been referenced in my thesis.

REFERENCES

[1] A. Subashini, S. K, S. Saranya, and U. Harsha, “Forecasting Website Traffic Using Prophet Time Series Model,” International Research Journal of Multidisciplinary Technovation, pp. 56–63, Jan. 2019, doi: 10.34256/irjmt1917.

[2] R. Casado-Vara, A. Martin del Rey, D. Pérez-Palau, L. de-la-Fuente-Valentín, and J. M. Corchado, “Web Traffic Time Series Forecasting Using LSTM Neural Networks with Distributed Asynchronous Training,” Mathematics, vol. 9, no. 4, p. 421, Feb. 2021, doi: 10.3390/math9040421.

[3] A. P. Wibawa, A. B. P. Utama, H. Elmunsyah, U. Pujianto, F. A. Dwiyanto, and L. Hernandez, “Time-series analysis with smoothed Convolutional Neural Network,” Journal of Big Data, vol. 9, no. 1, Apr. 2022, doi: 10.1186/s40537-022-00599-y.

[4] T. Shelatkar, S. Tondale, S. Yadav, and S. Ahir, “Web Traffic Time Series Forecasting using ARIMA and LSTM RNN,” ITM Web of Conferences, vol. 32, p. 03017, 2020, doi: 10.1051/itmconf/20203203017.

[5] Z. Shen, Y. Zhang, J. Lu, J. Xu, and G. Xiao, “A novel time series forecasting model with deep learning,” Neurocomputing, Apr. 2019, doi: 10.1016/j.neucom.2018.12.084.

[6] I. N. G. A. Astawa, I. P. B. A. Pradnyana, and I. K. Suwintana, “Comparison of RNN, LSTM, and GRU Methods on Forecasting Website Visitors,” Journal of Computer Science and Technology Studies, vol. 4, no. 2, pp. 11–18, Aug. 2022, doi: 10.32996/jcsts.2022.4.2.3.

[7] Tableau, “Time Series Analysis: Definition, Types, Techniques, and When It’s Used,” Tableau, 2022. https://www.tableau.com/learn/articles/time-series-analysis

[8] V. Tambe, A. Golait, S. Pardeshi, R. Javheri, and Prof. G. Arsalwad, “Web Traffic Time Series Forecasting Using ARIMA Model,” International Journal for Research in Applied Science and Engineering Technology, vol. 10, no. 5, pp. 2447–2453, May 2022, doi: https://doi.org/10.22214/ijraset.2022.42800.

[9] K. Kumari, “Guide to Web Traffic Forecasting Using Deep Learning,” Analytics Vidhya, Sep. 09, 2021. https://www.analyticsvidhya.com/blog/2021/09/web-traffic-forecasting-using-deep-learning/

[10] Frost, J. (2023) Mean squared error (MSE), Statistics By Jim. Available at: https://statisticsbyjim.com/regression/mean-squared-error-mse/

[11] A. Tripathi, “Time Series Analysis & Forecasting Guide ‖ AnalytixLabs,” Blogs & Updates on Data Science, Business Analytics, AI Machine Learning, Nov. 21, 2022. https://www.analytixlabs.co.in/blog/time-series-analysis/

[12] X. Hub, “Regression Analysis: The Ultimate Guide - Qualtrics,” Qualtrics AU, Mar. 23, 2022. https://www.qualtrics.com/au/experience-management/research/regression-analysis/

BIBLIOGRAPHY

[1] J. Brownlee, "How to Setup Your Python Environment for Machine Learning with Anaconda," Machine Learning Mastery, Sep. 16, 2020. [Online]. Available: https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/.

[2] J. Brownlee, "The Promise of Recurrent Neural Networks for Time Series Forecasting," Machine Learning Mastery, Aug. 5, 2019. [Online]. Available: https://machinelearningmastery.com/promise-recurrent-neural-networks-time-series-forecasting/.

[3] "Export and share reports - Analytics Help," Analytics Help. [Online]. Available: https://support.google.com/analytics/answer/1038573.

[4] J. Brownlee, "What is time series forecasting?" Machine Learning Mastery, Dec. 2016. [Online]. Available: https://machinelearningmastery.com/time-series-forecasting/.

[5] J. Brownlee, "How to create an Arima model for time series forecasting in Python," MachineLearningMastery.com, Dec. 9, 2020. [Online]. Available: https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/.

[6] J. Brownlee, "Time Series Forecasting With Prophet in Python," MachineLearningMastery.com, Aug. 26, 2020. [Online]. Available: https://machinelearningmastery.com/time-series-forecasting-with-prophet-in-python/.

[7] J. Brownlee, "Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras," MachineLearningMastery.com, Jul. 21, 2016. [Online]. Available: https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/.

[8] A. Choudhary, "Generate Quick and Accurate Time Series Forecasts using Facebook's Prophet (with Python & R codes)," Analytics

Vidhya, May 10, 2018. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/05/generate-accurate-forecasts-facebook-prophet-python-r/.