Twitter Volume Analysis

-- Amazon stock price

He Miao

Ayisha COK

Instructor:
Dr. Erfan Hoque
DASC 6510 – Time Series Analysis
Department of Mathematics and Statistics
Thompson Rivers University
2023-12-03

1. Abstract

This study aims to forecast the volume of Twitter mentions for Amazon (AMZN), utilizing a dataset from the realTweets collection within the Numenta Anomaly Benchmark (NAB) v1.1. The focus is on comparing the performance of two predictive models in time-series analysis: the Prophet model and the Neural Network Autoregressive (NNAR) model. The dataset provides a high-resolution view of AMZN mentions on Twitter, recorded every 5 minutes, posing a unique challenge for time-series forecasting. The Prophet model is renowned for its proficiency in handling seasonal time-series data, while the NNAR model offers a dynamic approach through its neural network-based autoregressive structure. The study evaluates these models using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE), with the objective of determining the more effective model for predicting social media trends.

Keywords: Prophet model, NNAR model, Evaluation Metrics

2. Introduction

The advent of social media platforms, particularly Twitter, has created new avenues for gauging public interest and sentiment towards corporations. This study leverages a high-frequency dataset from the realTweets collection of the Numenta Anomaly Benchmark (NAB) v1.1, focusing on the Twitter mentions of Amazon (AMZN). Recorded every 5 minutes, this dataset provides an intricate view of AMZN's presence in public discourse on Twitter, making it a prime candidate for time-series forecasting.

The research aims to compare two predictive models: the Prophet model, known for its effective management of time-series data with inherent seasonal trends, and the Neural Network Autoregressive (NNAR) model, which utilizes a neural network framework for autoregressive forecasting. The primary goal is to assess which model more accurately predicts the volume of Twitter mentions of AMZN. This will be achieved by evaluating the models against performance metrics such as MAE, MSE, and MAPE.

The study's findings are anticipated to offer valuable insights into the capabilities of these models in the context of time-series forecasting, particularly in social media trend analysis. It also aims to contribute to the broader understanding of applying advanced predictive techniques in the rapidly evolving field of social media analytics and market behavior prediction.

3. Data

The dataset in Numenta's NAB repository focuses on Twitter engagement pertaining to Amazon's stock (AMZN). It is located within the realTweets collection and is specifically named "Twitter_volume_AMZN.csv" and downloaded from(https://github.com/numenta/NAB/blob/master/data/realTweets/Twitter_volume_AMZN.csv).

The dataset comprises two main columns: 'timestamp', indicating the time at which tweet volumes were recorded, and 'value', representing the number of mentions of AMZN during each

5-minute interval. The 'timestamp' is recorded in a standard datetime format, making it suitable for time-series analysis without the need for extensive preprocessing.
To commence the analysis, the dataset underwent an initial loading process using the Python Pandas library. Subsequently, the data was modified to suit the specifications of the predictive models employed in this investigation. For the Prophet model, the column labels were adjusted to 'ds' (datestamp) and 'y' (value) to conform to the model's input format. Furthermore, the dataset was consolidated into hourly intervals to decrease granularity and emphasize overarching trends across time. This consolidation involved grouping the data by hour and computing the average tweet volume for each hourly segment.
This processed dataset forms the basis for the comparative analysis between the Prophet and Neural Network Autoregressive (NNAR) models. Its high-frequency nature poses unique challenges and opportunities for predictive modeling.

4. Methodology

To manage the computational complexity and optimize the training duration, the time-series data was aggregated to a granularity of one hour. This aggregation was performed to balance the detail and computational efficiency, providing a representative overview of trends without the excessive computational load of finer-grained data.

The model training and predictions were executed on Google Colab, utilizing its GPU capabilities (Tesla T4). This choice provided the necessary computational power for efficient processing of the time-series data, ensuring timely execution of the modeling tasks.

Prophet Model

Prophet model was introduced by Facebook (1). It is a non linear model of the form

$$Y_t = g(t) + s(t) + h(t) + \varepsilon_t$$

where

g(t) describes a piecewise-linear trend or growth term,

s(t) describes the various seasonal patterns,

h(t) captures the holiday effects, and

ε_t is a white noise error term

The prophet model generally manages outliers well and is resilient to missing data and changes in the trend.

Neural Network Autoregressive (NNAR) Model

A neural network (1,2) is a network of neurons which are organized in layers.

- Input Layer(predictors): The predictors form the bottom layer, the size of the input layer is determined by the chosen training window and coefficients attached to the inputs are called “weights”.
- First Fully Connected Layer (fc1): Transforms the input to a hidden layer of size 100.
- ReLU Activation Function: The ReLU activation function is a mathematical function that is used in artificial neural networks to introduce non-linearity and improve the performance of the model. It was first introduced in 2000 by Hahnloser et al.It is defined as follows:

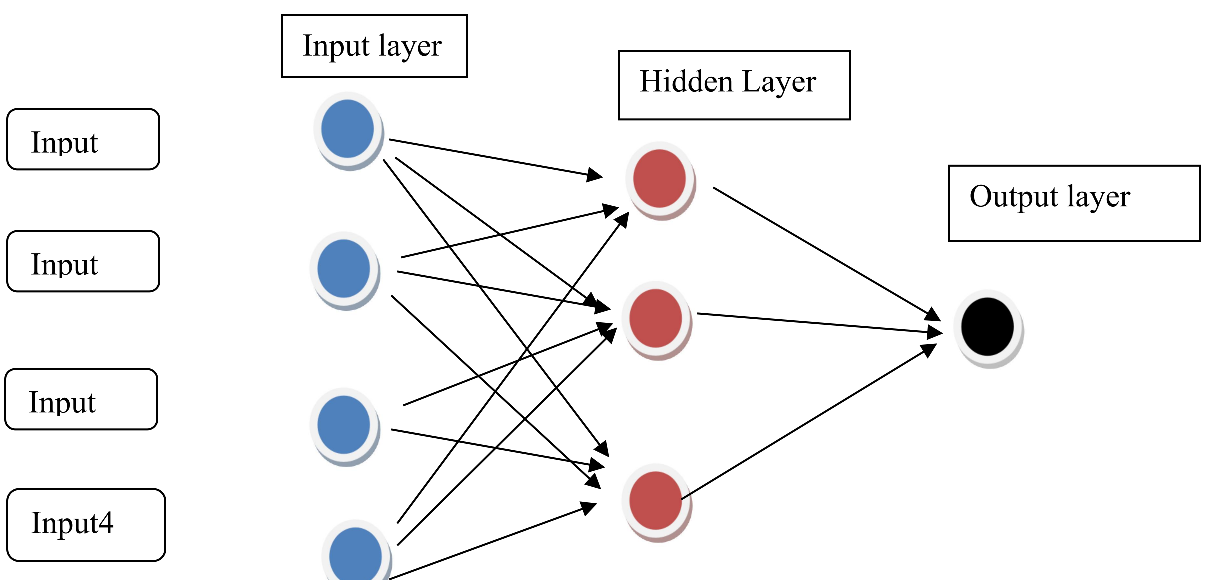
  $$\text{ReLU}(x)=\max(0,x)$$

  which means for any input x, the ReLU function will output x if it is positive, negative or zero. The ReLU activation function allows the network to learn sparse representations, because it sets the output of any negative input to zero

- Second Fully Connected Layer (fc2): Maps the output of the hidden layer to the final output size of 1, which corresponds to the predicted value.

NNAR is a special type of ANN which is used in time series forecasting. It stands for non linear Auto Regressive Neural Network with exogenous inputs. It has a recurrent architecture which allows to remember previous inputs and outputs and can capture non linear and dynamic patterns in time series data and generate accurate and robust forecasts.
A Neural Network AutoRegression(NNAR) Model consists of three components represented by NNAR(p,P,k),where

p represents the number of lagged values,
P represents the seasonal lagged values
k for number of hidden nodes



Neural Model

A neural network model is given by the combination of linear and non linear functions:

Inputs to hidden layer neuron j linearly combined :

$$U_j = b_j + \sum_{i=1}^{n} wij \ xi$$

This is then modified using the non linear function such as sigmoid

$$S(u)=\frac{1}{1+e^{-u}}$$

This helps to reduce the effect of extreme input values and hence the outliers (1)

Evaluation Metrics:

Model evaluation measures, such as the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (��2), are crucial instruments for assessing the predictive power of models. The tools used in this project for comparison are MAE, MSE, RMSE, MAPE.

Mean Absolute Error gives importance to the average bias between the model's predictions and actual results. It is given by

$$MAE = \frac{1}{n}\sum_{i=1}^{n} /yi - \hat{yi}\,/$$

Mean Square Error is the average difference between the predicted values and the actual values. It is given by

$$MSE=\frac{1}{n}\sum_{i=1}^{n} (yi - \hat{yi}\,)^2$$

Root Mean Square Error is the square root of MSE , which can evaluate the accurate accuracy of the model's prediction

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n} (yi - \hat{yi}\,)2}$$

Where

n is the total number of samples

yi is the actual value of the i^{th} sample

$\hat{yi}$ is the model's prediction

The Mean Absolute Percentage Error (MAPE) is another evaluation metric which is based on percentage of errors between the predicted and actual values. This gives an easy and straight forwad accuracy level. It is given by

$$\text{MAPE}=\frac{1}{n}\sum_{i=1}^{n}/\frac{yi-\hat{yi}}{yi}/$$

Github Repository Link
url: https://github.com/flistz/DASC-6510-Term-Project
member: flistz(He Miao), Ayisha Najeeha(Ayisha COK)

5. Results

The inference from table 1 is that the Prophet model has a better performance than the NNAR model in terms of the four evaluation metrics. The table shows the values of MAE, MSE, RMSE, and MAPE for the two models, which are measures of the prediction error rates and model performance in regression analysis. The lower the values of these metrics, the better the model fits the data and the lower the error. You can see that the Prophet model has lower values than the NNAR model for all four metrics, which means that it has a smaller prediction error and a higher accuracy. dataset yielded the following performance metrics:

	MAE	MSE	RMSE	MAPE
Prophet	7.2061	103.9179	10.1940	13.8728
NNAR	7.5691	104.4294	10.2190	15.3612

Table 1- Evaluation Metrics

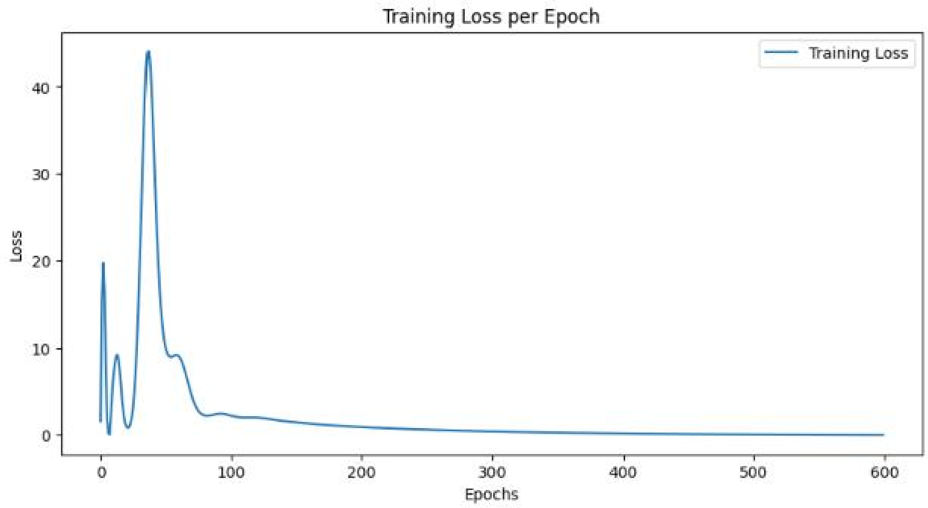


Figure 1 - Training loss per Epoch

Figure 1 illustrates the training loss of the Neural Network Autoregressive (NNAR) model over 600 epochs. The training loss is a measure of how well the model fits the training data. The lower the loss, the better the model. The graph shows that the model is learning and improving over time, as the loss decreases as the number of epochs increases. The graph also shows that the loss has some fluctuations in the beginning, but becomes smoother as the model converges to a stable state.

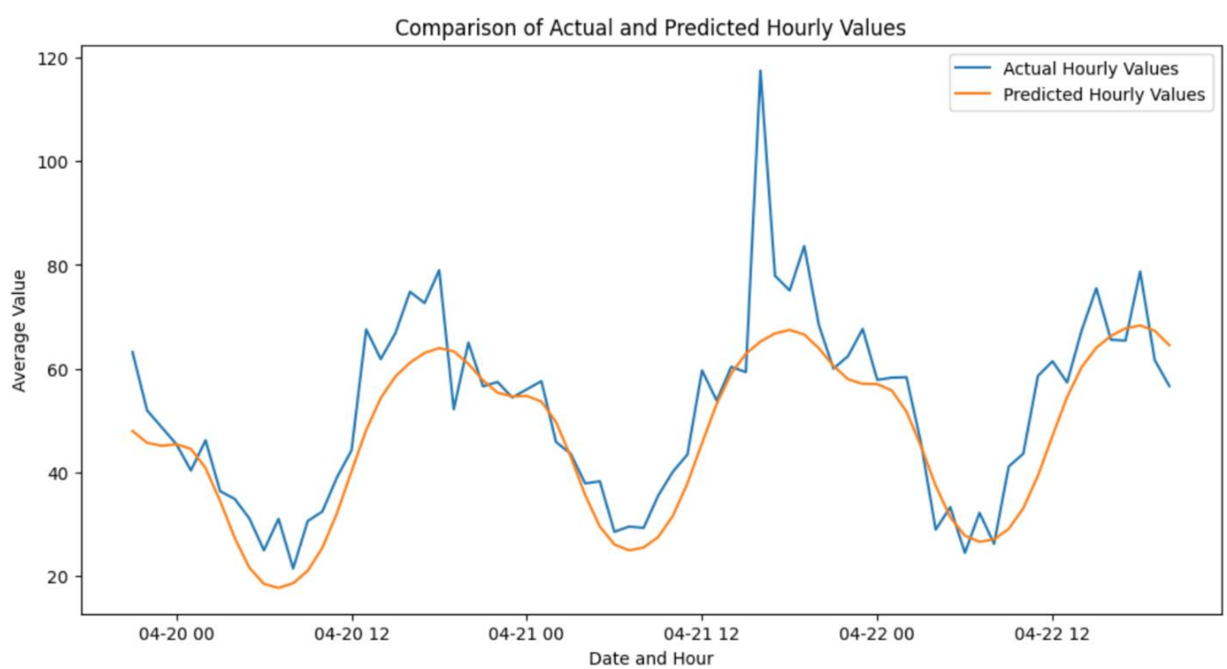


Figure 2:Prophet Model Predictions vs. Actual Values

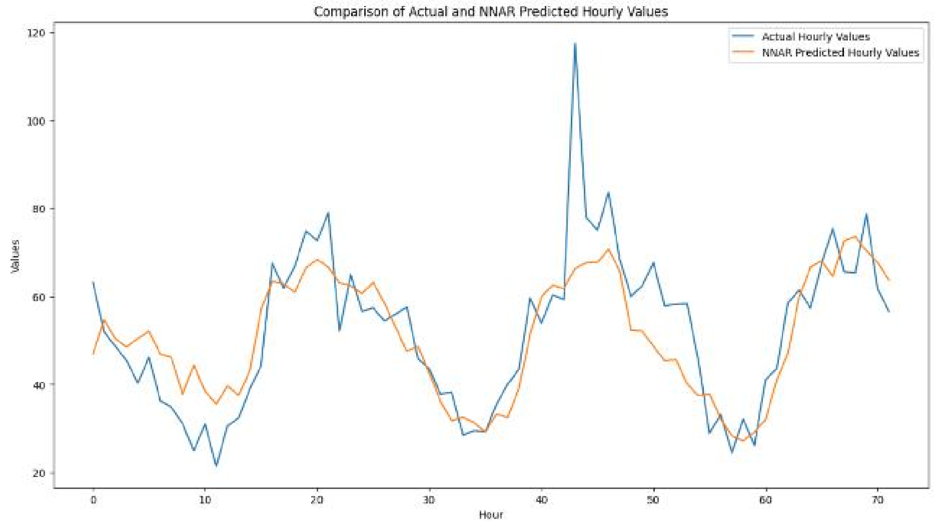


Figure 3: NNAR Model Predictions vs. Actual Values

Figure 2 compares the actual hourly Twitter volume values for Amazon (AMZN) against the predicted values generated by the Prophet model.
Figure 3 displays the predicted hourly values by the NNAR model in comparison to the actual data.

6. Conclusions

The results of the study reveal that the Prophet model marginally outperforms the NNAR model in predicting the hourly Twitter volume for Amazon (AMZN). The lower MAE, MSE, RMSE, and MAPE values indicate that the Prophet model is slightly more accurate in capturing the trend and fluctuations in the Twitter data. Notably, the NNAR model, despite its slightly higher error metrics, showcased a strong performance, particularly highlighted by its R-squared (R^2) value of 0.6606, which is competitively close to the Prophet model's 0.6621. This measure of determination suggests that the NNAR model is nearly as effective as the Prophet model in explaining the variance in the Twitter volume data. The commendable R^2 value, alongside the model's satisfactory loss curve, points to the NNAR model's potential for further development and optimization.

However, the results also highlight the potential of the Neural Network Autoregressive (NNAR) model. The training loss curve for the NNAR model (Figure 1) is particularly noteworthy. It showcases a swift reduction in loss during the initial epochs and stabilizes as the training progresses, which is a hallmark of effective learning and model convergence.

In conclusion, while the Prophet model is recommended for immediate use in similar time-series forecasting tasks, the NNAR model remains a viable contender. Its potential, especially with the advantage of running on powerful GPU hardware like the Tesla T4, should not be overlooked in future research. Further investigations could aim to refine the NNAR model, explore alternative neural network architectures, and even combine forecasts from multiple models to improve accuracy.

7. References

1. Hyndman RJ AG. Forecasting: Principles and practice. 2014th ed. first, editor.: otexts.
2. Taylor SJ LB. Forecasting at scale. The American Statistician. 2018 Jan 2 Jan 2; 37-45: 72.
3. Kożuch A CDAK. A Comparison of Artificial Neural Network and Time Series Models for Timber Price Forecasting. Forests. 2023; 177: 14.
4. Zhang L, Wang R, Li Z, Li J, Ge Y, Wa S, et al. Time-Series Neural Network: A High-Accuracy Time-Series Forecasting Method Based on Kernel Filter and Time Attention. Information. 2023; 500: 14.

8. Github Repository Link

url: https://github.com/flistz/DASC-6510-Term-Project

member: flistz(He Miao), Ayisha Najeeha(Ayisha COK)