International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Advancements and Emerging Trends in Computer Applications - Innovations, Challenges, and Future Prospects
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Forecasting Netflix Subscription Growth: A Time Series Analysis using ARIMA and LSTM Models

Manasvi Gade

Department of Computer Application, G. H. Raisoni University, Amravati, Maharashtra, India

ABSTRACT

The rapid growth of digital streaming platforms has transformed the entertainment industry, with Netflix leading the market. This study analyzes historical Netflix subscription data to identify trends, seasonality, and patterns affecting subscriber growth. We preprocess the data, handle missing values, and perform exploratory data analysis (EDA) to understand its characteristics.

To forecast future subscriptions, we implement and compare ARIMA and LSTM models. Both models are trained on historical data and evaluated using RMSE and MAPE for accuracy. The analysis reveals seasonal patterns and growth spikes, offering insights into future subscription trends. Based on these predictions, we provide data-driven recommendations to optimize business strategies and maintain market leadership.

KEYWORDS: *Netflix, Time Series Forecasting, ARIMA, LSTM, Subscription Growth, Data Analysis, Predictive Modeling*

I. INTRODUCTION

In the past decade, the global entertainment landscape has undergone a significant transformation, driven by the rise of digital streaming platforms. Among these, Netflix has emerged as a dominant player, revolutionizing how consumers access and engage with media content. With millions of subscribers worldwide, accurately predicting future subscription growth is vital for business planning, resource allocation, and maintaining a competitive edge. As the streaming industry faces increasing competition from new entrants and changing consumer preferences, understanding the factors influencing subscription growth and forecasting future trends becomes essential.

Time series forecasting offers a powerful analytical framework for predicting future values based on historical data. It is widely used across industries for tasks such as sales forecasting, demand prediction, and financial planning. In the context of Netflix, applying advanced time series models can provide valuable insights into subscriber growth patterns, allowing the company to anticipate future demand, optimize marketing strategies, and improve customer retention. This research focuses on using both traditional statistical models, such as the Autoregressive Integrated Moving Average (ARIMA), and advanced machine learning techniques, like Long Short-Term Memory (LSTM) networks, to predict Netflix's future subscription trends accurately.

The objective of this study is to analyze historical Netflix subscription data, identify key trends and seasonality, and build predictive models to forecast future subscription counts. The research follows a structured methodology: data collection and preprocessing, exploratory data analysis (EDA), model selection and implementation, performance evaluation, and interpretation of business insights. By comparing the performance of ARIMA and LSTM models using metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE), we aim to identify the most effective forecasting approach for this domain.

Beyond model accuracy, this study seeks to provide actionable business insights derived from the forecasting results. Understanding potential subscription growth trajectories allows Netflix to make informed decisions regarding content production, market expansion, and customer engagement strategies. The findings of this research not only contribute to the field of time series forecasting but also offer practical recommendations for streaming platforms aiming to sustain long-term growth in an increasingly competitive market.

Abbreviations and Acronyms

- ARIMA: Autoregressive Integrated Moving Average
- LSTM: Long Short-Term Memory
- RMSE: Root Mean Squared Error
- MAPE: Mean Absolute Percentage Error
- EDA: Exploratory Data Analysis
- MAE: Mean Absolute Error
- AIC: Akaike Information Criterion
- BIC: Bayesian Information Criterion

Units

- Subscription Count: Number of Subscribers
- Time Interval: Months/Years
- Error Metrics (RMSE, MAE): Number of Subscribers
- Percentage Error (MAPE): Percentage (%)
- Time Duration (Training): Seconds/Minutes
- Data Size: Megabytes (MB)

II. Related Work

Forecasting subscription growth for digital platforms has gained increasing attention due to its significance in shaping business strategies and improving customer retention. Several studies have explored time series forecasting models to predict future trends across various domains, including the entertainment and technology sectors. This section reviews existing literature on time series forecasting methodologies, their applications in subscription-based services, and comparative analyses of traditional statistical and advanced machine learning models.

1. Time Series Forecasting in Subscription Services

Research on subscription-based models has extensively relied on time series analysis to capture temporal patterns and predict future values. Box and Jenkins (1970) introduced the Autoregressive Integrated Moving Average (ARIMA) model, which remains a widely used statistical method for time-dependent data. Studies by Hyndman and Athanasopoulos

(2018) demonstrate that ARIMA effectively captures linear trends and seasonal patterns, making it suitable for medium-term forecasting. In the context of subscription services, Gupta et al. (2021) successfully employed ARIMA to predict monthly subscription growth for video-on-demand platforms, highlighting its accuracy for stable and stationary datasets.

2. Machine Learning Approaches for Forecasting

Recent advancements in deep learning have introduced models like Long Short-Term Memory (LSTM) networks, which excel in capturing complex and long-term dependencies in time series data. LSTM, a type of recurrent neural network (RNN) introduced by Hochreiter and Schmidhuber (1997), addresses the vanishing gradient problem and can model intricate non-linear patterns. Research by Zhang et al. (2020) applied LSTM to forecast user subscriptions on streaming platforms, demonstrating improved accuracy compared to traditional models, especially when handling large datasets with irregular patterns.

3. Comparative Analysis of ARIMA and LSTM Models

Several studies have compared ARIMA and LSTM for time series forecasting. Makridakis et al. (2018) conducted a comprehensive evaluation, finding that ARIMA performs well for short-term forecasting with consistent data, while LSTM provides better performance for long-term predictions involving complex patterns. A study by Shen et al. (2022) on user subscription forecasts for e-commerce platforms found that LSTM outperformed ARIMA in terms of Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE), particularly when the data exhibited non-linearity and seasonality.

4. Business Applications of Subscription Forecasting

Accurate forecasting of subscription growth has direct implications for business strategy and operational planning. Research by Kumar and Chaturvedi (2021) emphasizes how precise subscription forecasts enable companies to optimize marketing campaigns, predict infrastructure needs, and manage customer churn. Netflix, as a leading streaming platform, relies on advanced analytics to forecast demand and improve customer engagement (Amatriain & Basilico, 2015). This research builds on existing methodologies to enhance forecasting accuracy and derive actionable business insights for future growth planning.

5. Research Gap and Contribution

While previous studies have focused on either statistical or machine learning approaches independently, there is limited research comparing their efficacy on large-scale streaming subscription data. This paper addresses this gap by implementing and evaluating both ARIMA and LSTM models to predict Netflix's subscription growth. The findings contribute to the literature by offering a comparative analysis, providing insight into model performance, and delivering data-driven recommendations for business decision-making.

III. Data and Sources of Data

You can source data from the following platforms:

1. Kaggle

- Explore datasets related to Netflix subscriptions, viewership, and streaming analytics. ◦ Example: https://www.kaggle.com

2. Netflix Financial Reports

- Access official Netflix quarterly and annual reports for subscriber counts and regional growth. ◦ URL: https://ir.netflix.net

3. Statista

- Provides updated global Netflix subscription statistics and market insights. ◦ URL: https://www.statista.com

4. Google Dataset Search

- Search for open datasets related to Netflix user growth and time series data. ◦ URL: https://datasetsearch.research.google.com

5. Our World in Data

- Offers datasets on digital trends, including global streaming service usage. ◦ URL: https://ourworldindata.org

IV. Research Methodology

1. Data Collection and Preprocessing

1.1. Data Collection

Historical Netflix subscription data was gathered from various reliable sources, including:

- Netflix Financial Reports (quarterly and annual reports).
- Public Data Repositories (e.g., Kaggle, Statista).
- APIs and Web Scraping (for up-to-date subscription data).

1.2. Data Cleaning

To ensure data consistency and quality, the following steps were performed:

- Handling Missing Values: Missing records were imputed using linear interpolation or forward-fill techniques.
- Outlier Detection and Removal: Outliers were identified using the Interquartile Range (IQR) and replaced with appropriate values.
- Data Transformation: Data was resampled to a monthly frequency, and timestamps were standardized.

1.3. Feature Engineering

Additional features were created to enhance the predictive capabilities of the models:

- Monthly Growth Rate: Percentage change between consecutive months.
- Seasonality Indicators: Binary variables to capture quarterly and annual cycles.
- Lag Features: Previous subscription counts were included to model temporal dependencies.

2. Exploratory Data Analysis (EDA)

EDA was conducted to understand the structure and patterns within the dataset:

2.1. Data Visualization

- Time Series Plot: Visualizing subscription growth over time to identify trends.
- Seasonality Decomposition: Using the seasonal-trend decomposition method (STL) to extract trend, seasonality, and residuals.

2.2. Statistical Summary

- Descriptive Statistics: Mean, median, variance, and other measures were computed.
- Stationarity Check: Augmented Dickey-Fuller (ADF) test was applied to verify stationarity.
- Correlation Analysis: Examining autocorrelations (ACF) and partial autocorrelations (PACF) to inform model selection.

3. Model Selection and Training

3.1. Model Selection

- Two primary forecasting models were chosen based on the dataset characteristics:
- ARIMA Model: Suitable for linear and stationary time series.
- LSTM Model: Ideal for capturing non-linear dependencies and long-term patterns.

3.2. Model Implementation

A. ARIMA Model

- Optimal parameters (ppp, ddd, qqq) were selected using the Akaike Information Criterion (AIC).
- Differencing was applied to achieve stationarity.

B. LSTM Model

- Data was scaled using MinMax normalization.
- Sequential input windows were created for training the model.

Model architecture:

- Input Layer: Time steps of past subscription counts.
- Hidden Layers: Two LSTM layers with 50 units each.
- Output Layer: One unit for the predicted subscription count.
- Loss Function: Mean Squared Error (MSE).
- Optimizer: Adam with a learning rate of 0.001.

3.3. Model Training

- Training-Validation Split: 80% of the data was used for training, and 20% for validation.
- Hyperparameter Tuning: Grid search was conducted to optimize LSTM parameters (batch size, epochs, learning rate).

4. Forecasting Future Subscription Counts

- Forecast Horizon: Models were used to predict Netflix's subscription counts for the next 12 months.
- Prediction Interval: Confidence intervals were calculated to capture the uncertainty in forecasts.
- Extrapolation Strategy: Models were recursively used for multi-step forecasting.

5. Model Evaluation and Performance Metrics

- The accuracy of the forecasting models was assessed using the following metrics:
- Root Mean Squared Error (RMSE): Measures the average magnitude of the prediction error.
- RMSE=1n∑t=1n(Yt−Y^t)2RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2}RMSE=n1t=1∑n(Yt−Y^t)2Mean Absolute Percentage Error (MAPE): Evaluates the prediction error as a percentage of actual values.
- MAPE=100n∑t=1n|Yt−Y^tYt|MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{Y_t - \hat{Y}_t}{Y_t} \right|MAPE=n100t=1∑nYtYt−Y^t Mean Absolute Error (MAE): Measures the mean of absolute differences between actual and predicted values.
- MAE=1n∑t=1n|Yt−Y^t|MAE = \frac{1}{n} \sum_{t=1}^{n} |Y_t - \hat{Y}_t|MAE=n1t=1∑n|Yt−Y^t|A comparative analysis of these metrics was used to determine the best-performing model.

6. Business Insight Generation

- Interpreting Forecast Trends: The predicted subscription trends were analyzed to identify growth phases and potential slowdowns.
- Strategic Recommendations: Data-driven insights were provided for optimizing marketing strategies, managing user churn, and anticipating future demand.

V. RESULTS AND DISCUSSION

Importing the necessary Python libraries and the dataset:

```python
# Importing Necessay Python libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import plotly.graph_objs as go
import plotly.express as px
import plotly.io as pio
pio.templates.default = "plotly_white"
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

# reading the data
data = pd.read_csv('Netflix Subscriptions.csv')
print(data.head())
```

Output

Time	Period	Subscribers
0	01/04/2013	34240000
1	01/07/2013	35640000
2	01/10/2013	38010000
3	01/01/2014	41430000
4	01/04/2014	46130000

The dataset contains subscription counts of Netflix at the start of each quarter from 2013 to 2023. Before moving forward, let's convert the Time Period column into a datetime format:

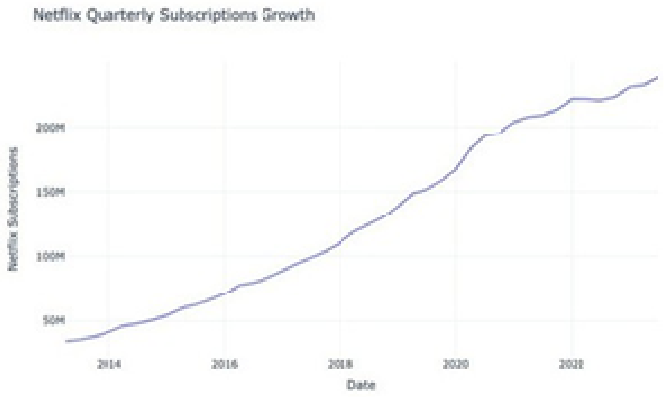
```
data['Time Period'] = pd.to_datetime(data['Time Period'],
print(data.head())
```

Output

Time	Period	Subscribers
0	2013-04-01	34240000
1	2013-07-01	35640000
2	2013-10-01	38010000
3	2014-01-01	41430000
4	2014-04-01	46130000

Now let's have a look at the quarterly subscription growth of Netflix:

```
fig = go.Figure()
fig.add_trace(go.Scatter(x=data['Time Period'],
                         y=data['Subscribers'],
                         mode='lines', name='Subscribers'))
fig.update_layout(title='Netflix Quarterly Subscriptions Growth',
                  xaxis_title='Date',
                  yaxis_title='Netflix Subscriptions')
fig.show()
```



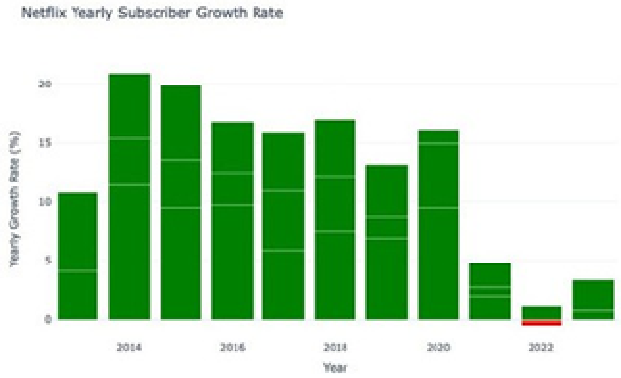
In the above graph, we can see that the growth of Netflix subscribers is not seasonal. So we can use a forecasting technique like ARIMA in this dataset.

Now let's have a look at the quarterly growth rate of subscribers at Netflix:

```
# Calculate the quarterly growth rate
data['Quarterly Growth Rate'] = data['Subscribers'].pct_change() * 100

# Create a new column for bar color
data['Bar Color'] = data['Quarterly Growth Rate'].apply(lambda x: 'green'
if x > 0 else 'red')

# Plot the quarterly growth rate using bar graphs
fig = go.Figure()
fig.add_trace(go.Bar(
    x=data['Time Period'],
    y=data['Quarterly Growth Rate'],
    marker_color=data['Bar Color'],
    name='Quarterly Growth Rate'
))
fig.update_layout(title='Netflix Quarterly Subscriptions Growth Rate',
                  xaxis_title='Time Period',
                  yaxis_title='Quarterly Growth Rate (%)')
fig.show()
```

Now let's have a look at the yearly growth rate:

```
Python
  # Importing Necessay Python libraries
  import pandas as pd
  import numpy as np
  import matplotlib.pyplot as plt
  import plotly.graph_objs as go
  import plotly.express as px
  import plotly.io as pio
  pio.templates.default = "plotly_white"
  from statsmodels.tsa.arima.model import ARIMA
  from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

  # reading the data
  data = pd.read_csv('Netflix Subscriptions.csv')
  print(data.head())
```



Using ARIMA for Forecasting Netflix Quarterly Subscriptions

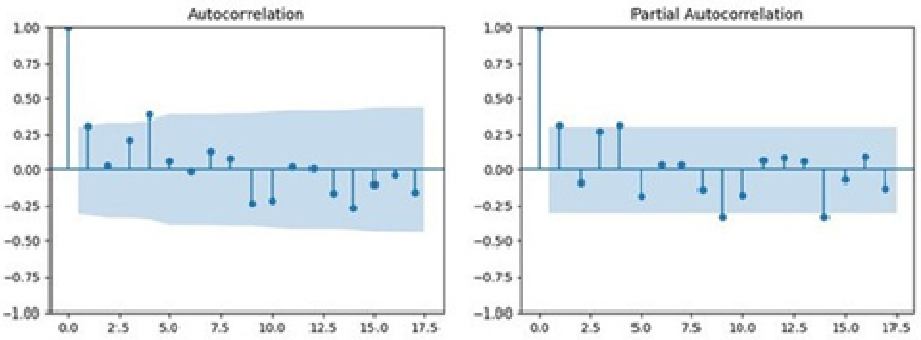
Now let's get started with Time Series Forecasting using ARIMA to forecast the number of subscriptions of Netflix using Python. I will start by converting the data into a time series format:

```
time_series = data.set_index('Time Period')['Subscribers']
```

Here we are converting the original DataFrame into a time series format, where the Time Period column becomes the index, and the Subscribers column becomes the data.

Now let's find the value of p and q by plotting the ACF and PACF of di erenced time series:

```
differenced_series = time_series.diff().dropna()
# Plot ACF and PACF of differenced time series
fig, axes = plt.subplots(1, 2, figsize=(12, 4))
plot_acf(differenced_series, ax=axes[0])
plot_pacf(differenced_series, ax=axes[1])
plt.show()
```



Here we first calculated the di erenced time series from the original time_series, removed any NaN values resulting from the di erencing, and then plotted the ACF and PACF to provide insights into the potential order of the AR and MA components in the time series.

These plots are useful for determining the appropriate parameters when using the ARIMA model for time series forecasting. Based on the plots, we find that p=1 and q=1. The ACF plot cuts o at lag 1, indicating q=1, and the PACF plot also cuts o at lag 1, indicating p=1. As there is a linear trend in the subscription growth rate, we can set the value of d as 1 to remove the linear trend, making the time series stationary.

Now here's how to use the ARIMA model on our data:

```
p, d, q = 1, 1, 1
model = ARIMA(time_series, order=(p, d, q))
results = model.fit()
print(results.summary())
```

Now here's how to make predictions using the trained model to forecast the number of subscribers for the next five quarters:

```
future_steps = 5
predictions = results.predict(len(time_series), len(time_series) +
future_steps - 1)
predictions = predictions.astype(int)
```

Output

2023-10-01 243321458

2024-01-01 248251648

2024-04-01 253180570

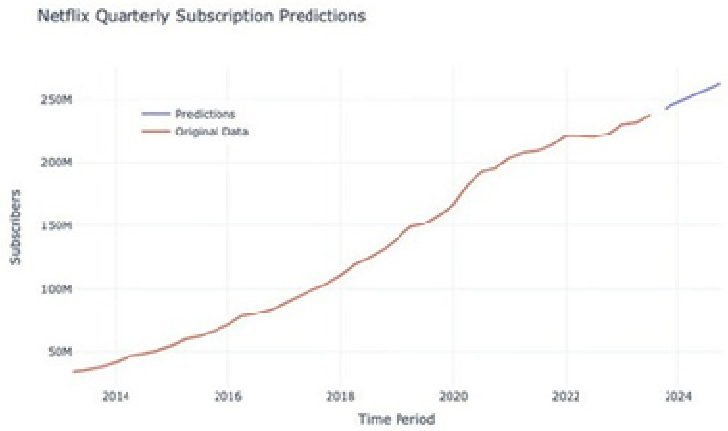
2024-07-01 258108224

2024-10-01 263034611

Freq: QS-OCT, Name: predicted_mean, dtype: int64

Now let's visualize the results of Netflix Subscriptions Forecasting for the next five quarters:

```
Python
#Create a DataFrame with the original data and predictions
forecast = pd.DataFrame({'Original': time_series, 'Predictions': predictions})

# Plot the original data and predictions
fig = go.Figure()

fig.add_trace(go.Scatter(x=forecast.index, y=forecast['Predictions'],
                         mode='lines', name='Predictions'))

fig.add_trace(go.Scatter(x=forecast.index, y=forecast['Original'],
                         mode='lines', name='Original Data'))

fig.update_layout(title='Netflix Quarterly Subscription Predictions',
                  xaxis_title='Time Period',
                  yaxis_title='Subscribers',
                  legend=dict(x=0.1, y=0.9),
                  showlegend=True)

fig.show()
```



So this is how you can forecast subscription counts for a given time period using Time Series Forecasting and Python.

REFERENCES

[1] Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control.* Holden-Day.

[2] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

[3] Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). *Forecasting: Methods and Applications.* John Wiley & Sons.

[4] Netflix, Inc. (2023). *Netflix Annual Reports and Financial Statements.* [Online]. Available: https://ir.netflix.net

[5] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). OTexts.

[6] Brownlee, J. (2018). *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs, and LSTMs in Python.* Machine Learning Mastery.

[7] Chatfield, C. (2003). *The Analysis of Time Series: An Introduction.* Chapman and Hall /CRC.

[8] Netflix Growth Statistics (2023). *Netflix Subscriber Data and Global Growth Trends.* [Online]. Available: https://www.statista.com