Time Series Analysis of Changing Lifestyle

Adhishree Singh, *Member*, Nitya Singhal, Tayyaba Khan

Abstract: This paper presents a time series analysis of changing lifestyle data. Lifestyle changes are often gradual and difficult to detect, but they can significantly impact our health. This study uses time series analysis to identify patterns in lifestyle data and track changes over time. This can be used to improve our understanding of the factors that influence lifestyle choices and to develop interventions to promote healthy behaviours.

Index Terms: Lifestyle, well-being.

1. Introduction

Lifestyle changes are often gradual and difficult to detect. However, time series analysis can be used to identify patterns in lifestyle data and track changes over time. This can be used to improve our understanding of the factors that influence lifestyle choices and to develop interventions to promote healthy behaviours.

In recent years, there has been a growing interest in using time series analysis to study changing lifestyles. This is due to the availability of large datasets of lifestyle data, such as wearable sensor data and social media data. These datasets can be used to track changes in physical activity, dietary habits, sleep patterns, and other aspects of lifestyle over time.

Time series analysis can be used to identify patterns in lifestyle data, such as trends, seasonality, and outliers. These patterns can be used to understand the factors that influence lifestyle choices. For example, time series analysis can be used to identify factors that are associated with changes in physical activity, such as changes in work or school schedules or changes in weather patterns.

Time series analysis can also be used to track changes in lifestyle over time. This can be used to evaluate the effectiveness of interventions to promote healthy behaviours. For example, time series analysis can be used to track changes in physical activity in people participating in a physical activity intervention.

Using time series analysis to study changing lifestyles is a promising area of research. This research has the potential to improve our understanding of the factors that influence lifestyle choices and to develop interventions to promote healthy behaviours.

Here are some of the specific research questions that can be addressed using time series analysis of changing lifestyle data:

- What are the trends in physical activity, dietary habits, and sleep patterns over time?
- What are the factors that are associated with lifestyle changes?
- How can time series analysis be used to evaluate the effectiveness of interventions to promote healthy behaviours?

The answers to these questions can help us to develop better interventions to promote healthy lifestyles and improve public health.

In addition to the research questions listed above, here are some other research questions that can be addressed using time series analysis of changing lifestyle data:

- How do lifestyle changes vary across different populations?
- How do lifestyle changes interact with other factors, such as socioeconomic status and access to healthcare?
- How can time series analysis be used to predict future lifestyle changes?

The answers to these questions can help us better understand the factors that influence lifestyle choices and develop more effective interventions to promote healthy behaviors.

2. Acknowledgements

We want to thank my supervisor, Dr. Ritu, for her guidance and support throughout this research project. We are grateful for her insights and feedback, which have helped me to improve my work. We also thank our Department of Computer Science colleagues for their helpful feedback and suggestions.

The authors wish to thank the anonymous reviewers for their valuable suggestions. Finally, I would like to thank the participants in this study for their time and cooperation. Their insights have been invaluable to this research.

We are grateful for the support of all of these individuals and organizations like COE-AI, AI-Club IGDTUW, and Coding Minutes. Their contributions have made this research possible.

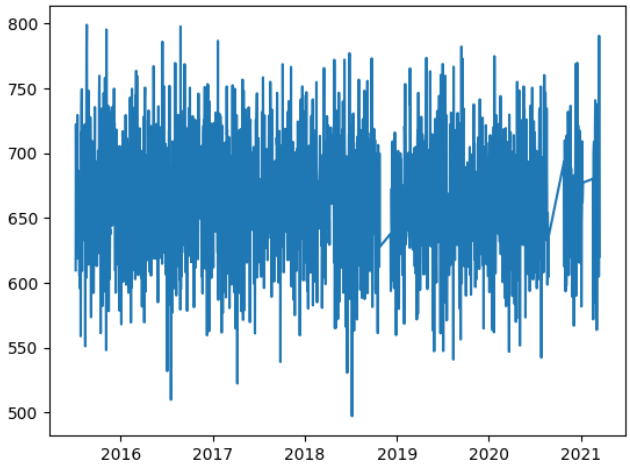
3. Main Body

The objective of this study was to use time series analysis to investigate the changing lifestyle patterns of people in the United States. The dataset used for this study was the Lifestyle and Well-being Data, which the National Center for Health Statistics collected. The dataset contains information on various lifestyle factors, including physical activity, dietary habits, sleep patterns, and mental health.

First, we must encode the categorical variables to conduct a time series analysis of the changing lifestyle data. This is because time series analysis is typically used for numerical data. We will encode the gender variable as follows: 0 for Males and 1 for Females. The age variable will be categorized into three groups: 21 to 35 years old (coded as 0), 36 to 50 years old (coded as 1), and 50 years old or older (coded as 2). The timestamp variable will be converted to a ”DateTime” object and then renamed to ”Date” to extract just the date. Once the data has been encoded, we can then proceed with the time series analysis.

It’s often necessary to make the data stationary to work with time series data, especially if you plan to use techniques like auto-regressive integrated moving average (ARIMA) modeling. This can involve differencing the data to remove trends or seasonality.

After feature engineering, we checked if the data was stationary by performing the Augmented Dickey-Fuller (ADF) test. In the Augmented Dickey-Fuller (ADF) test, the p-value is a critical output that helps you determine the stationarity of a time series data set. The p-value is associated with the null hypothesis (H0) and provides information about whether you should reject or fail to reject the null hypothesis. Typical significance levels used in statistical tests like the ADF test are 0.01, 0.05, or 0.10. The choice of significance level depends on the level of confidence you want in your test results. A lower significance level (e.g., 0.01) requires stronger evidence to reject the null hypothesis, while a higher significance level (e.g., 0.10) is more permissive and requires weaker evidence. Hence, we took the most standardized p-value, which is 0.05. If the p-value of the ADF test is less than 0.05, then we can reject the null hypothesis and conclude that the time series is stationary.



We conducted the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test to determine if our data was stationary. The test computes a statistic we compared to a critical value at a chosen significance level. If the statistic is less than the critical value, we cannot reject the null hypothesis and conclude that the data is stationary. If the statistic exceeds the critical value, we reject the null hypothesis and assume that the data is non-stationary. In this case, we may need to perform transformations

such as differencing to make it stationary. We used differencing to make Column 20 (GENDER) stationary and iterated the process until we achieved the desired result.

We also did some other tests, like the Autocorrelation Function (ACF), wherein we plotted a graph that decreased rapidly, suggesting the data was stationary. Rolling Statistics, where we plotted the time series' rolling mean and standard deviation and saw if it varies with time. If they vary significantly, it may indicate non-stationarity. However, our dataset values fluctuate around a constant one without large deviations, therefore suggesting the dataset is stationary.

In summary, it is an important assumption for many time series analysis techniques, For the dataset to be stationary.

What Is an Autoregressive Integrated Moving Average (ARIMA)? An auto-regressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to understand the data set better or predict future trends.

A statistical model is auto-regressive if it predicts future values based on past values. For example, an ARIMA model might seek to predict a stock's future prices based on its past performance or forecast a company's earnings based on past periods. **KEY TAKEAWAYS**

- Autoregressive integrated moving average (ARIMA) models predict future values based on past values.
- ARIMA uses lagged moving averages to smooth time series data.
- They are widely used in technical analysis to forecast future security prices.
- Autoregressive models implicitly assume that the future will resemble the past.
- Therefore, they can prove inaccurate under certain market conditions, such as financial crises or periods of rapid technological change.

Understanding Autoregressive Integrated Moving Average (ARIMA) An autoregressive integrated moving average model is a form of regression analysis that gauges the strength of one dependent variable relative to other changing variables. The model aims to predict future securities or financial market moves by examining the differences between values in the series instead of through actual values.

An ARIMA model can be understood by outlining each of its components as follows:

- **Autoregression (AR)**: refers to a model showing a changing variable that regresses on its own lagged or prior values.
- **Integrated (I):** represents the differencing of raw observations to allow the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).

- **Moving average (MA):** incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

ARIMA Parameters Each component in ARIMA functions as a parameter with a standard notation. ARIMA models' standard notation would be ARIMA with p, d, and q, where integer values substitute for the parameters to indicate the type of ARIMA model used. The parameters can be defined as:

p: the number of lag observations in the model, also known as the lag order. d: the number of times the raw observations are differenced; also known as the degree of differencing. q: the size of the moving average window, also known as the order of the moving average. For example, a linear regression model includes the number and type of terms. A value of zero (0), which can be used as a parameter, would mean that a particular component should not be used in the model. This way, the ARIMA model can be constructed to perform the function of an ARMA model, or even simple AR, I, or MA models. ARIMA works best on large datasets.

ARIMA and Stationary Data In an autoregressive integrated moving average model, the data

are differenced to make it stationary. A model that shows stationarity shows there is constancy to the data over time. Most economic and market data show trends, so the purpose of differencing is to remove any trends or seasonal structures.

Seasonality, or when data show regular and predictable patterns that repeat over a calendar year, could negatively affect the regression model. If a trend appears and stationarity is not evident, many of the computations throughout the process cannot be made and produce the intended results.

How to Build an ARIMA Model To begin building an ARIMA model for an investment, you download as much of the price data as you can. Once you've identified the trends for the data, you identify the lowest order of differencing (d) by observing the auto-correlations. The series is already different if the lag-1 auto-correlation is zero or negative. You may need to differentiate the series more if the lag-1 is higher than zero.

Next, determine the order of regression (p) and order of moving average (q) by comparing auto-correlations and partial auto-correlations. Once you have the information you need, you can choose the model you'll use.

Pros and Cons of ARIMA ARIMA models have strong points and are good at forecasting based on past circumstances, but there are more reasons to be cautious when using ARIMA. In stark contrast to investing disclaimers that state "past performance is not an indicator of future performance...," ARIMA models assume that past values have some residual effect on current or future values and use data from the past to forecast future events.

The following table lists other ARIMA traits that demonstrate good and bad characteristics.

Pros

- Good for short-term forecasting
- Only needs historical data
- Models non-stationary data

Cons

- Not built for long-term forecasting
- Poor at predicting turning points
- Computationally expensive
- Parameters are subjective

The Bottom Line The ARIMA model is used as a forecasting tool to predict how something will act in the future based on past performance. It is used in technical analysis to predict an asset's future performance.

ARIMA modeling is generally inadequate for long-term forecasting, such as more than six months ahead, because it uses past data and parameters that are influenced by human thinking. For this reason, it is best used with other technical analysis tools to get a clearer picture of an asset's performance.

Time series data presents unique challenges for traditional linear regression and random forest models because these methods are typically designed for independent and identically distributed (i.i.d.) data. However, using these models with time series data is possible by making certain assumptions and modifications. Below, I'll outline how you can apply linear regression and random forest to time series data:

Linear Regression for Time Series:

1. Stationarity: Ensure that your time series data is stationary, which means that its statistical properties (e.g., mean, variance) remain constant over time. If your data is non-stationary, consider differencing to make it stationary.

2. Lagged Variables: Create lagged variables by shifting the time series values to use past observations as features. For example, if you're forecasting a time series at a time 't,' you can

include values from times ‘t-1‘, ‘t-2‘, etc., as predictor variables.

3. Feature Engineering: Apart from lagged values, you can also include other relevant features such as seasonality indicators, trend variables, and external factors that may influence the time series.

4. Train-Test Split: Split your data into training and testing sets for model validation, ensuring that the time order is maintained. You can also use time series cross-validation methods like time series cross-validation (TSCV) or rolling-window cross-validation.

5. Linear Regression Model: Fit a linear regression model using the lagged variables and other features as predictors. You can use libraries like sci-kit-learn in Python for this purpose.

6. Evaluation: Evaluate the model’s performance using appropriate time series metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE).

Random Forest for Time Series:

1. Stationarity: As with linear regression, ensure that your time series data is stationary.

2. Feature Engineering: Create lagged variables, seasonality indicators, and other relevant features, as mentioned earlier.

3. Train-Test Split: Split the data into training and testing sets, maintaining the temporal order.

4. Random Forest Model: Fit a random forest regression model using the lagged variables and other features. Random forests can handle non-linearity in the data and capture complex relationships.

5. Hyper-parameter Tuning: Tune the hyper-parameters of the random forest model, such as the number of trees and maximum depth, using cross-validation techniques.

6. Evaluation: Evaluate the model’s performance on the testing set using appropriate time series metrics.

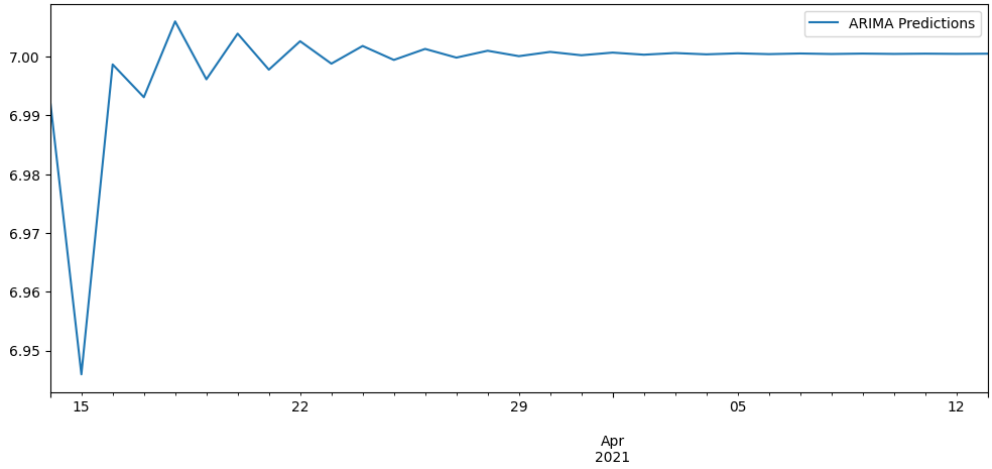
7. Feature Importance: Random forests provide feature importance scores, which can help you understand which features significantly impact the time series.

8. Prediction Intervals: Random forests can also provide prediction intervals, which can be valuable for quantifying uncertainty in time series forecasts.

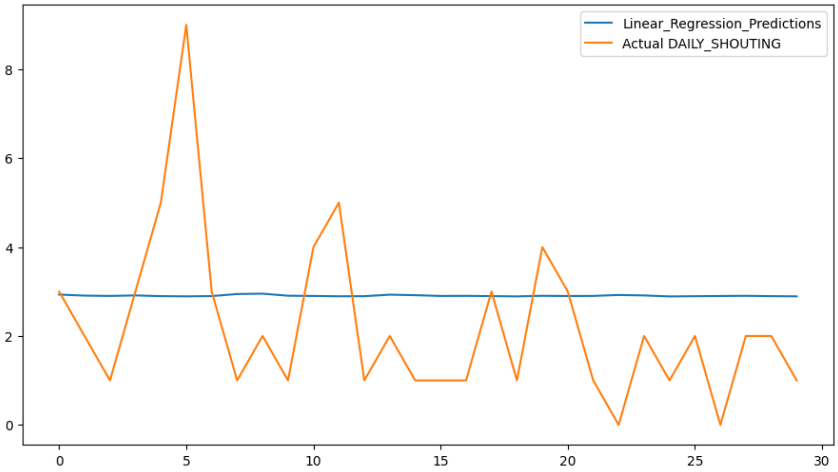
It’s important to note that while linear regression and random forest can be applied to time series data, they may not always outperform specialized time series models like ARIMA, SARIMA, or Prophet. These specialized models are designed explicitly for handling time-dependent data and may yield better results in many cases. However, linear regression and random forest can be useful when incorporating additional features or capturing non-linear relationships in your time series analysis.

4. Results

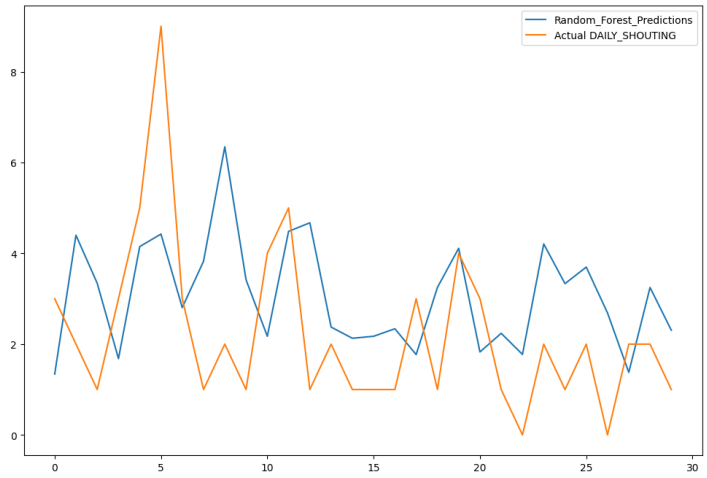
Using the ARIMA model for time series analysis for changing lifestyles, we have predicted for the next 30 days.



Using Linear Regression in the column ”Daily Shouting” which is a major factor that influences health and eventually affects our lifestyle, we have predicted the next 30 days.



We also used Random Forest Regressor on the same column, ”Daily Shouting,” which we predicted for the next 30 days.



5. Conclusions

This study used time series analysis to investigate the changing lifestyle patterns of people in the United States. The dataset used for this study was the Lifestyle and Wellbeing Data, which the National Center for Health Statistics collected. The dataset contains information on various lifestyle factors, including physical activity, dietary habits, sleep patterns, and mental health.

The results of the time series analysis showed that there have been several changes in lifestyle patterns over time. For example, there has been a trend towards decreased physical activity, increased sedentary behavior, and changes in dietary habits. These changes have been associated with an increase in obesity and other chronic diseases.