Time Series Forecasting

Forecasting is the process of predicting the future. Any time we make a prediction of a future event, we are forecasting.

A time series is a set of data points ordered in time. The data is equally spaced in time. In simpler terms, the data can be recorded either every minute/hour or it could be averaged over every month/year.

Common Interview Questions

1. What is the difference between stationery and non-stationary data? Why is this important?
2. What are some examples of time-series data sources?
3. Can you explain RNN and LSTM, and when you use each for TSA?
4. Can you explain Dynamic Time Warping?
5. What is the difference between ARMA and ARIMA?
6. How do you decide which forecasting model to use?
7. How do you handle seasonality in time series data?
8. How do you handle missing values in time series data?
9. How do you evaluate the accuracy of a time series forecast?
10. What methods do you use to detect anomalies in time series data?

Key Concepts

- Properties of time series
- Types of forecast
- Outliers
- Missing Values
- Accuracy of Forecast
- ARIMA
- SARIMA
- Error Metrics

Key Time Series Factors

Qualitative forecasting - This is based on expert decision.
Quantitative forecasting - repeating historical patterns in the data. E.g. Time series forecasting

Properties of Time Series

- **Level** - Average value (mean) of the series
- **Trend** - Gradual upward or downward movements of data over time
- **Seasonality** - Variation that repeats itself over time (Holidays, Promos)
- **Cycles** - Business Cycles, Economic Cycles, etc
- **Randomness** - Variation that cannot be explained by trend/seasonality/ caused by chance

Seasonality

- Seasonality is a variation that occurs at specific regular intervals of less than a year.
- Seasonality can occur on different time spans, such as daily, weekly, monthly, or yearly

Different Types of Forecasts

- **Long-term** - Forecasts support strategic planning and include decisions such as market trends, behavior and shifts in competitive markets, emerging competition, and long-term expansions
- Short-term - Forecasts support tactical decisions. They address questions such as how many stock-keeping units or SKUs of a product a retailer expects to sell
- **Promotions** - Seasonal based
- Impact of price, promotion on sales numbers

Stationary Tests

There are two tests to check if a time series is stationary:

- **Rolling Statistics** - Plot the moving avg or moving standard deviation to see if it varies with time. Its a visual technique.
- **ADCF Test** - Augmented Dickey–Fuller test is used to gives us various values that can help in identifying stationarity. The Null hypothesis says that a TS is non-stationary. It comprises of a Test Statistics & some critical values for some confidence levels. If the Test statistics is less than the critical values, we can reject the null hypothesis & say that the series is stationary. THE ADCF test also gives us a p-value. Acc to the null hypothesis, lower values of p is better.

How time series forecasting is different from other regression tasks

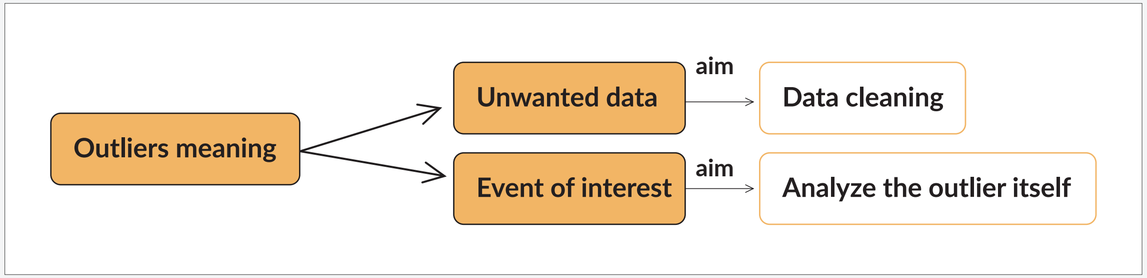
- The first concept to keep in mind is that time series have an order, and we cannot change that order when modeling
- In time series forecasting, we express future values as a function of past values
- Time series are indexed by time, and that order must be kept.
- Time series can be decomposed into three components: a trend, a seasonal component, and residuals.

Handling Outliers

Outlier - An observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism

Effects of Outliers

1. Skewing of mean and standard deviation
2. Could significantly affect significance readings when generating regression analysis
3. Can give us false readings on the magnitude of correlations



Outliers meaning

Unwanted data

aim

Data cleaning

Event of interest

aim

Analyze the outlier itself

Outlier Detection / Removal

1. Domain knowledge to identify data issues/min / max range
2. Inter Quartile Range to remove extreme values. Box Plot: The points lying on either side of the whiskers are considered to be outliers as shown in th image. The length of these whiskers is subjective and can be defined by you according to the problem.
3. Experiment with Outlier Detection techniques / Packages - Python Outlier Detection (PyOD), Python Streaming Anomaly Detection (PySAD)
4. Extreme value analysis: Remove the smallest and largest values in the dataset
5. Histogram: Simply plotting a histogram can also reveal the outliers - basically the extreme values with low frequencies visible in the plot

Time Series Forecast Models

Key Steps in Forecasting

- Define the problem
- Collect the data
- Analyze the data
- Build and evaluate the forecast model

Handling Missing Values

- **Mean Imputation:** Imputing the missing values with the overall mean of the data
- **Last observation carried forward:** We impute the missing values with its previous value in the data.
- **Linear interpolation:** You draw a straight line joining the next and previous points of the missing values in the data.
- **Seasonal + Linear interpolation:** This method is best applicable for the data with trend and seasonality. Here, the missing value is imputed with the average of the corresponding data point in the previous seasonal period and the next seasonal period of the missing value.

ARIMA(Auto Regressive Integrated Moving Average)

- ARIMA(Auto Regressive Integrated Moving Average) is a combination of 2 models AR(Auto Regressive) & MA(Moving Average). It has 3 hyperparameters –P(auto regressive lags) - the number of past values included in the AR model
- P(auto regressive lags) - the number of past values included in the AR model
- Q(moving avg.) - the size of the moving average window

How to choose values of p, d and q?

- Draw a partial autocorrelation graph(ACF) of the data.This will help us in finding the value of p because the cut-off point to the PACF is p.
- Draw an autocorrelation graph(ACF) of the data. This will help us in finding the value of q because the cut-off point to the ACF is q.

Time Series Forecast Models

SARIMA builds upon the concept of ARIMA but extends it to model the seasonal elements in your data. Key params are

- P: Seasonal autoregressive order.
- D: Seasonal differencing order.
- Q: Seasonal moving average order
- S: Length of the seasonal cycle.

Holt-Winters Forecasting and Exponential Smoothing

- Holt-Winters is a way to model three aspects of the time series: a typical value (average), a slope (trend) over time, and a cyclical repeating pattern (seasonality)
- Exponential smoothing refers to the use of an exponentially weighted moving average (EWMA) to “smooth” a time series.
- Single, Double, and Tripe exponential smoothing options are available.
- The additive method is preferred when the seasonal variations are roughly constant through the series, while the multiplicative method is preferred when the seasonal variations are changing proportional to the level of the series

Forecast Algo Summary

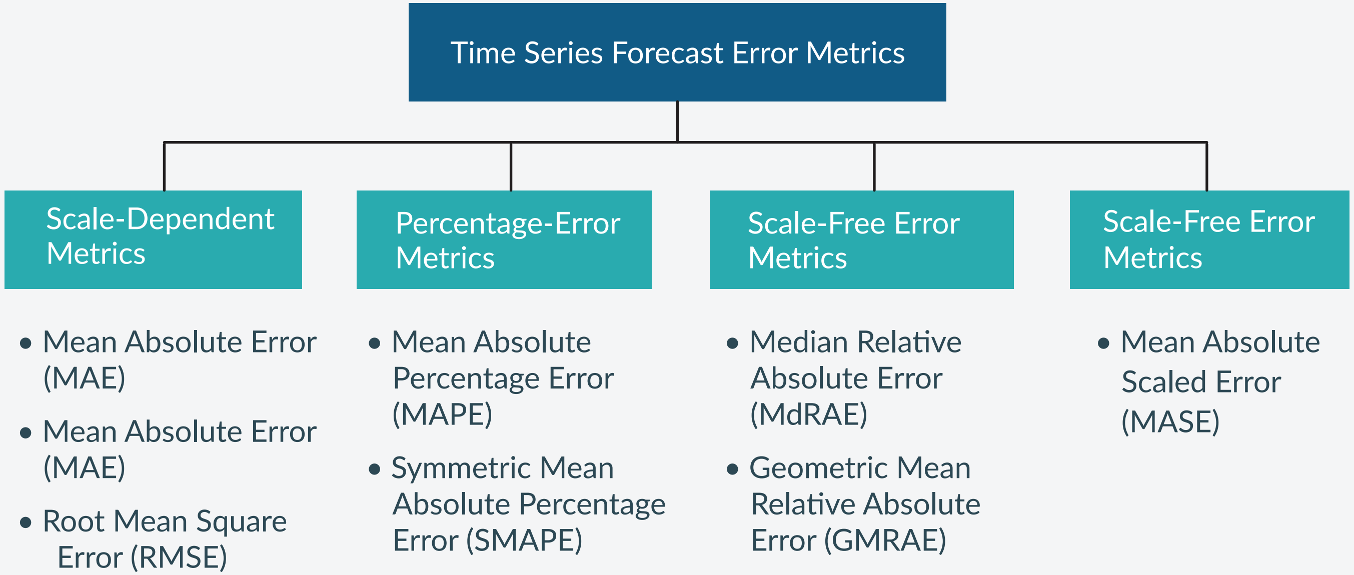
- Baseline Algos - Moving Average, Weighted Moving Average, ARIMA, SARIMA, Exponential Smoothing
- Regression-Based - Linear Regression, Random Forest Regressor, SVM Regressor, Gradient Regressor
- Deep Learning based - CNN, RNN, DART, TFT, GluonTS

Feature Engineering / Metrics

Feature Variables for Time series data

Date time features - hour, month, and day of week for each observation. Daylight savings or not, Public holiday or not, Quarter of the year, Hour of day, Season of the year.

- Lag features and window features - Business day, Quarter start, Weekly frequency
- Rolling window statistics - moving average
- Expanding window statistics - minimum, mean, and maximum values
- Domain-specific features - additional research into each feature and find out domain-specific information beyond what is provided in the dataset description
- Additional factors, such as trends, seasonality, holidays, and external economic variables.



Time Series Forecast Error Metrics

Scale-Dependent Metrics

- Mean Absolute Error (MAE)
- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)

Percentage-Error Metrics

- Mean Absolute Percentage Error (MAPE)
- Symmetric Mean Absolute Percentage Error (SMAPE)

Scale-Free Error Metrics

- Median Relative Absolute Error (MdRAE)
- Geometric Mean Relative Absolute Error (GMRAE)

Scale-Free Error Metrics

- Mean Absolute Scaled Error (MASE)

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_l|$$

Root of Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_l|}$$