Introducing Forecasting

FORECASTING · TIME SERIES DATA · MISSING DATA & RESAMPLING · ALGORITHMS · SAGEMAKER CANVAS · METRICS · AMAZON FORECAST

STUDY & REVIEW

01 MUST KNOW

if you remember five things, remember these

01 **Forecasting predicts the future from the past:** it estimates **future values from historical data** carrying a **time component**. Series are **univariate** (one variable) or **multivariate** (many), and most follow a **trend**, **seasonal**, **cyclical**, or **irregular** pattern. The time element adds information but makes the problem harder than ordinary prediction.

02 **Time series breaks regression's core assumption:** points are **captured in sequence** and tend to be **correlated** — yet regression assumes **independent** data points. That dependence (**autocorrelation**) is a form of noise that can *overstate* a model's accuracy, so time series needs special handling and algorithms that correct for it.

03 **Data preparation is most of the work:** fix **missing values** (forward fill, moving average, backward fill = *lookahead*, interpolation, or **zero fill** in retail), **resample** to align frequencies (**downsample** to a coarser grain by sum/average; **upsample** to a finer grain — hard), **smooth** outliers, model **seasonality**, and standardize timestamps to **UTC**.

04 **Forecasts are probabilistic — read the quantiles:** models predict at **P10, P50, P90** quantiles (the true value falls below the forecast that percent of the time). Accuracy metrics — **wQL, RMSE, MAPE, WAPE, MASE** — are all *lower-is-better*; choose the quantile by your cost of over- versus under-stocking.

05 **Two build paths, five algorithms: Amazon SageMaker Canvas** builds forecast models with **no code** (quick or standard build); **Amazon Forecast** is the dedicated service with **five algorithms** — **ARIMA, DeepAR+, ETS, NPTS, Prophet** — plus **AutoML**. Either way, a time series dataset needs a **timestamp**, a **target**, and an **item ID** column.

02 KEY TERMS

TERM	DEFINITION
Forecasting	Predicting future values based on historical data; an area of ML in which the data carries a time component
Univariate / multivariate	Time series with one variable versus a time series with more than one variable
Seasonality	Any repeating observation whose frequency is stable (hourly, daily, quarterly, yearly); a dataset can hold several at once
Stationarity	How stable a system is; a stationary series has constant statistical properties over time — low stability predicts the future poorly
Autocorrelation	How points in time are linearly related; observations are not independent, so it acts as noise that can overstate model accuracy
Lookahead	Using future values to calculate the past (e.g., backward fill) — invalid in forecasting and to be avoided
Downsampling	Moving to a less finely grained time (hourly → daily); values are combined by sum (sales) or average (temperature)
Upsampling	Moving to a more finely grained time (daily → hourly); difficult and usually needs extra data or domain knowledge
Smoothing	Removing outliers and anomalies that would distort a forecast (data preparation and cleaner visualization)
Quantile (P10 / P50 / P90)	Probabilistic forecast points; the true value is below the prediction 10, 50, or 90 percent of the time
Back testing	How Amazon Forecast measures accuracy — it splits the time series into training and test sets rather than a random holdout

TERM	DEFINITION
AutoML	Amazon Forecast feature that tries every supported algorithm and selects the best-performing predictor

03 CLEANING AND RESHAPING TIME SERIES DATA

the bulk of forecasting work

TECHNIQUE	WHAT IT DOES
Forward fill	Fills a gap with the **last known** value
Moving average	Fills a gap with the **average of recent known** values
Backward fill	Fills with the **next known** value — looks into the future (*lookahead*); **avoid** in forecasting
Interpolation	Uses an **equation** to estimate the missing value
Zero fill	Leaves the gap as **0** — correct in retail, where a missing sale means no order that day
Downsampling	Moves to a **coarser** frequency (hourly → daily); combine by **sum** (sales) or **average** (temperature)
Upsampling	Moves to a **finer** frequency (daily → hourly); hard — needs extra data or domain knowledge
Smoothing	Removes **outliers and anomalies** that a one-off event created

04 FORECASTING ALGORITHMS (AMAZON FORECAST'S FIVE)

dataset features drive the choice

ALGORITHM	BEST SUITED FOR
ARIMA	Removes **autocorrelations** that can distort the pattern of observations
DeepAR+	**Supervised** algorithm for one-dimensional series; trains a **recurrent neural network** across multiple series
ETS (Exponential Smoothing)	Datasets with **seasonality**; a weighted average of all observations, weights decreasing over time
NPTS (Non-Parametric Time Series)	Predicts by **sampling from past observations**; specialized for seasonal and climatological data
Prophet	A **Bayesian** model; long time spans, missing data, or large outliers
AutoML	Tries **all** algorithms and picks the best predictor

05 FORECAST ACCURACY METRICS — ALL LOWER IS BETTER

probabilistic, not a single number

METRIC	WHAT IT MEASURES	REMEMBER
wQL	Average weighted quantile loss across P10/P50/P90	Best when errors vary a lot
RMSE	Standard deviation of the prediction errors	Best when errors are similar in size
MAPE	Percent error averaged over all time points	0 = a perfect model
WAPE	Absolute error normalized by the absolute target	0 = a perfect model
MASE	Forecast error scaled against a simple baseline	< 1 beats the baseline; > 1 is worse
Column impact	Relative influence of each input attribute	Larger value ⇒ larger impact

06 AMAZON FORECAST WORKFLOW

the ML pipeline applied to time series

`01` Import as much historical + related data as you have → `02` Evaluate & feature-engineer the data → `03` Train a predictor — pick an algorithm or **AutoML**, and a **domain** → `04` Generate the forecast from an input dataset group → `05` Query it, or export to an **Amazon S3** bucket

07 EXAM TRAPS

where the knowledge check gets you

- × "Backward fill is a safe way to patch gaps." — It fills the past with **future** values (*lookahead*) and leaks information. Forward fill, moving average, and interpolation are safe; backward fill is not.
- × "Every missing sales value should be filled in." — In retail a missing value usually means **no orders that day**; a **zero fill** (not an average) is right — zero is sometimes the perfect fill value.
- × "You can upsample as easily as you downsample." — Downsampling to a coarser grain is straightforward; **upsampling** to a finer grain is difficult and usually impossible without another data source or domain knowledge.
- × "A strong correlation between two series proves one causes the other." — **Correlation is not causation.** Shared **trends** inflate correlation — give two random series the same slope and they correlate strongly.
- × "Regression handles time series out of the box." — Regression assumes **independent** points, but time series data is **correlated** (autocorrelation); untreated, it overstates the model's accuracy.
- × "A higher P90 forecast is a more accurate forecast." — P10/P50/P90 are **quantiles of uncertainty**, not accuracy grades. Pick the quantile by your risk; judge accuracy by **lower** wQL/RMSE/MAPE/WAPE/MASE.
- × "A quick build is too rough to use." — In SageMaker Canvas a **quick build** (2–15 min) is *reasonably accurate*; the **standard build** (2–4 hours) is the most accurate — you trade time for accuracy.
- × "DeepAR+ is unsupervised because it uses a neural network." — **DeepAR+** is a **supervised** learning algorithm that trains a **recurrent neural network** over multiple time series.

08 SELF-QUIZ

answers are upside-down below

Q1 What are the two broad categories of a time series based on how many variables it has?

Q2 Name the four common patterns a time series can follow.

Q3 Regression assumes points are independent — what property of time series data breaks that assumption?

Q4 Which missing-value fill uses future values to estimate the past, and what is that dangerous practice called?

Q5 Moving from an hourly dataset to a daily one is called what?

Q6 In a P90 forecast, what fraction of the time will the true value fall below the predicted value?

Q7 List the five algorithms Amazon Forecast supports.

Q8 Which Amazon Forecast feature tries every algorithm and selects the best predictor?

Q9 What three columns must a SageMaker Canvas time series dataset contain?

Q10 How does Amazon Forecast split data to compute predictor accuracy metrics instead of a random holdout?

ANSWER KEY (rotate the page)

Q1 univariate (one variable) and multivariate (more than one variable) **Q2** trend, seasonal, cyclical, irregular **Q3** correlation between points (autocorrelation) **Q4** backward fill; the practice is lookahead **Q5** downsampling **Q6** 90 percent of the time **Q7** ARIMA, DeepAR+, ETS, NPTS, Prophet **Q8** AutoML **Q9** a timestamp column, a target column, and an item ID column **Q10** back testing (a time-ordered train/test split)