VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

Machine Learning

Forecasting

Jan Platoš

November 22, 2023

Department of Computer Science
Faculty of Electrical Engineering and Computer Science
VŠB - Technical University of Ostrava

Forecasting

Forecasting

What is forecasting?

- Procedure of creating a model able to predict future values of the target variable.
  - Usually provides prediction as a single number or an interval.
- Forecasting shares many aspects with regression analysis but with one major difference:
  - Regression analysis deals with prediction of **current value** based on the **current data**.
  - Forecasting deals with prediction of **future value** based on the **historical data**.

Forecasting applications

- Forecasting is not a niche area of data science and it can be more common discipline than you think.

- **Weather** - the most common example, everybody can think of that
- **Energy** - consumption of electricity, natural gas, etc; peak value forecasting
- **Informatics** - service load forecasts can be used for JIT container spinning
- **Sales** - consumer demand for certain products
- **Finance** - stock price prediction - don't try this at home :-)

- **What would you like to forecast?**

Forecasting methods limits

- Some things are easier to forecast than the others.
- We need to define if providing a forecast can be even done at first.
- E.g. Forecasting next hour temperature is no big deal but prediction of next lottery numbers is a whole different story.
- We need to take several factors into account:
  - Do we have enough data available?
  - Do we fully understand all factors that can have impact on the target variable?
  - Does our forecast affect the target variable?
  - How distant values are we forecasting?

Forecasting methods limits

- Take for example electricity consumption forecasting.
  - Electricity consumption is quite common variable to forecast so the data availability shouldn't be an issue.
  - Consumption is mainly driven by weather, calendar and economic conditions - these factors are easily understandable.
  - We usually do not affect target variable with our forecasts.
- Given these preconditions the forecasts could be very precise with the right model.
- However in terms of accuracy it still depends on the fact if we want to produce short term or long term forecasts.

Forecasting methods limits

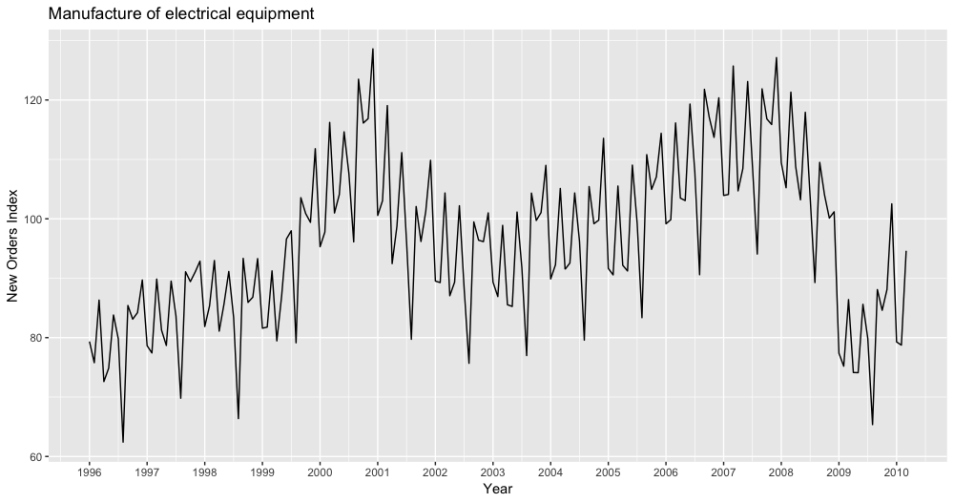
- Another good example is stock price forecasting.
  - We have plenty of data available as in the last case [1].
  - We have a limited understanding of external factors influencing stock prices.
  - Forecasted prices have a direct effect on the prices themselves.
  - Basically stock prices become their own forecasts - people will immediately adjust the price they are willing to pay.
- Given these preconditions the forecasting model abilities are very limited and we should be aware of it.

[1] You cant take a look at `https://finance.yahoo.com/`

Time series

- A time series is simply a series of data points ordered in time.
- **Each datapoint has timestamp assigned.**
- Essential feature is sampling frequency, e.g. hourly, weekly, monthly, ...
- We differ between two main types of time series data:
  - **Univariate** - only one variable is varying over time
    - One sensor which is measuring temperature in a room every minute.
  - **Multivariate** - multiple variables are varying over time
    - Two sensors which are measuring temperature and humidity in a room every minute.
- There are multiple ways of categorizing time series data, but this one is the most common.

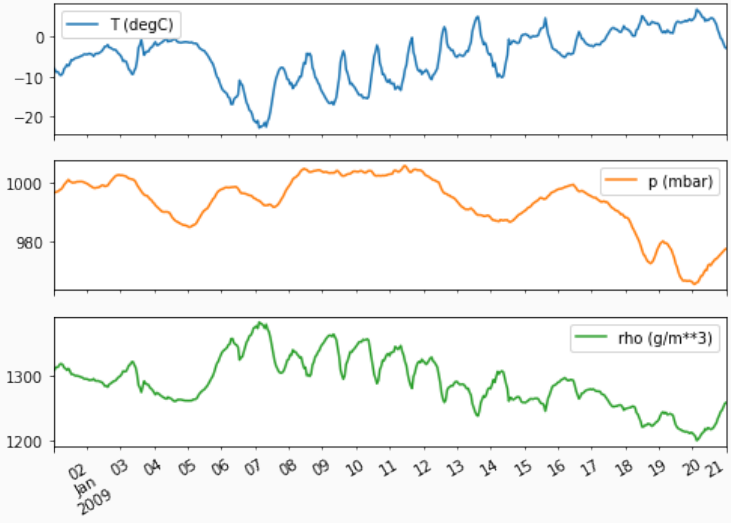
Time series example



Multivariate time series

- Number of a covariate time series in the dataset has significant impact on the method selection.
- Covariate time series have to be aligned to each other and have the same sampling frequency
  - Imagine you have multiple independent sensors.
  - Each sensor started measuring in different point of time - may lead to disalignment - may be fixed by shifting.
  - Each sensor can have its own sampling frequency - may be fixed by resampling.
  - This circumstances make analysis of raw data very hard however both issue could be fixed during pre-process phase.

Multivariate time series example



Time dependant variables

- There are two types of variables in case of multivariate time series:
  - Endogenous
    - A variable that depends on other variables in a model.
    - Value changes because there are changes to its relationships with other variables in the same model.
    - i.e. the variable you forecast.
  - Exogenous
    - A variable that depends on external factors outside of the model.
    - Changes in values of these variables influence the endogenous variables.
    - Not every exogenous variable is important because it can have strong correlation with other exogenous variables.

Forecast horizon and forecast types

- Forecast horizon is very important parameter in the model definition phase.
- It is length of time into the future for which forecasts are produced by the model.
- The forecasting could be either short-term or long-term. The precise definition of short-term or long-term forecast interval depends heavily on the sampling frequency and the dataset domain.
  - E.g. For hourly sampled electricity consumption could be next 24 hours forecast horizon considered short-term although for weekly sampled cosmetic product sales we could consider next 4 weeks forecasts short-term as well.
- More reasonable is to treat the forecast horizon as number of time steps of the sampling frequency into the future.

Forecast horizon and forecast types

- The simplest type of forecast is plain next value forecast, e.g. stock price of Tesla tomorrow.
- Forecast horizon length is one in this case, because we forecasted only single value.
- With longer forecast horizons we distinguish between two approaches:
  - Direct - make the predictions all at once with no relationship among forecasted values.
  - Cumulative - make one prediction at a time and feed the output back to the model.

Cumulative forecasting

- The cumulative approach takes the forecasts of the previous values into account and uses them as additional input variables for the model.
- Many traditional methods (simple autoregressive models) are built upon this methodology.
- The advantage of this approach is the need for only one model.
- The final forecast of the whole forecast horizon can be ensembled from the partial next-value forecasts.
- The main disadvantage of this approach is the accumulation of forecast error through the forecast horizon.
- This can lead to increasing error during the period being forecasted.

Direct forecasting

- The direct approach treats forecasts as independent variables.
- There is no error accumulation.
- Disadvantage of this approach is its complexity.
- Multiple models or heavy preprocessing of the data are usually needed.
- Complexity depends to a certain extent on the specific model.
  - Models can be able to provide either scalar (e.g. regression tree) or vector output (e.g. neural network).
  - There is no need for multiple models if the model is capable of vector output.

Forecasting methods

- There is a wide variety of methods available.
- You can divide them basically into three main groups.
  - **Traditional statistical methods**
  - **Machine learning and deep learning**
  - **Hybrid models**

Traditional statistical methods

- Many different methods and their modifications.
- Based on some sort of linear combination of past values.
- Exponential smoothing, autoregression, moving average and many modifications, e.g. (S)ARIMA(X)
- Usually useful for smaller datasets.
- Multi-step forecasting uses cumulative approach.
- Many essential properties are still used in modern approaches (autoregression, differencing, etc).
- `forecast` [2] package in R or `statsmodels` [3] in Python

[2] `https://www.rdocumentation.org/packages/forecast`
[3] `https://www.statsmodels.org/`

Short look at simple traditional approaches

Exponential smoothing

- The most basic variant is plain moving average.
  - All future forecasts are equal to a simple average of the observed data.
  - The average method assumes that all observations are of equal importance.
- We usually want something less extreme. It is sensible to attach larger weights to more recent observations than to observations from the distant past.

$$y_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \ldots$$

- $\alpha \in (0, 1)$ is the smoothing parameter.

Autoregressive models

- The term autoregression indicates that it is a regression of the variable against itself.
- We forecast the variable of interest using a linear combination of past values of the variable.
- We can view the model as an extension to Exponential smoothing, because the core idea is similar however we use multiple ”smoothing” parameters.
- You can often see the notation $AR(p)$, where p denotes the number of used past values.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + ... + \phi_p y_{t-p} + \epsilon_t$$

- ϵ_t is normally distributed white noise with mean zero and variance one

Machine learning in time-series forecasting domain

- We saw on the previous slides that traditional approaches were designed specifically for the time-dependent data and forecasting.
- Machine learning algorithms do not provide time series forecasting support out of the box, because they were not designed for it in the first place.
- They do not provide an implicit way to capture the interdependence between observations like e.g. AR(p) models.
- Every feature vector is treated independently, and thus the time-related part of the data is ignored.

Machine learning in time-series forecasting domain

- Information on the time-dependency of the individual vectors has to be captured through engineered features.
- Many popular algorithms are based on decision trees.
- There comes another complication - tree-based machine learning algorithms are unable to extrapolate trends.
- The trend extrapolation issue is caused by the fact that the forecast values provided by the regression tree-based learners are averaged target variable values of samples which belong to the same leaf in the particular decision tree.
- Thus you can’t predict higher value than you already saw in the training set.

Machine learning in time-series forecasting domain

- The good news is that every mentioned issue can be overcame by precise data pre-processing and feature engineering.
- Their versatility in the area of feature engineering is an indisputable advantage.
- Machine learning is not limited to only using endogenous variable like other statistical approaches but it is no problem to include many exogenous variables to the model as well.
- You can even work with data with complex seasonal pattern using careful pre-processing which would be very problematic for traditional approaches.
- **Machine learning algorithms are able to outperform statistical approaches if the dataset is big enough, moreover with lower computational complexity.**

Machine learning - capturing interdependence

- We have no autoregressive part in the model by default.
- If we want to include past values into the model, we need to engineer additional features.
- Common technique is adding so-called lagged values of the original variables as new features.
- This can be done not only for the target variable but for exogenous variables as well.
- We are not limited to including only raw lagged values, but common practice is including some summary statistics of past fixed length periods, e.g. mean of values for the past 24 hours.
- Usual practice is including calendar features as well, e.g. day of the week, month, is the current day holiday?, etc.

Machine learning - capturing interdependence - example

Time	Consumption	Temperature
7 a.m.	7000	12
8 a.m.	8000	13
9 a.m.	8500	13
10 a.m.	8700	14
11 a.m.	9000	15
12 p.m.	9100	16
13 p.m.	9400	17
14 p.m.	9900	17

Table 1: Example of raw data. Consumption is the forecasted variable.

Machine learning - capturing interdependence - example

Time	Consumption	Temperature	$Consumption_{t-1}$	$Temperature_{t-1}$
7 a.m.	7000	12	NaN	NaN
8 a.m.	8000	13	7000	12
9 a.m.	8500	13	8000	13
10 a.m.	8700	14	8500	13
11 a.m.	9000	15	8700	14
12 p.m.	9100	16	9000	15
13 p.m.	9400	17	9100	16
14 p.m.	9900	17	9400	17

Table 2: Example of lagged data. Consumption is the forecasted variable.

Minimum lag number

Beware that the length of the forecast horizon is a minimal lag number.

Examples

- We are doing short-term forecast of electricity consumption for the next 24 hour with hourly sampled data. The minimum lag of a variable is 24 because of this. If we include values with shorter lag, our forecast would be based on currently unknown values.
- Imagine that now is midnight and you forecast consumption at noon tomorrow. If you include consumption with lag 1 as a feature (i.e. consumption in the last hour) you would need to know real consumption at 11 a.m. tomorrow.
- *This is obviously not true because you would need an oracle for this thus no forecasting model would be needed :-)*

Machine learning - capturing interdependence

- We can add as many lagged values as we need using this approach.
- It is worth to mention that the first instances in the datasets must be dropped before training the model because we do not know the past (lagged) values for them.

Common properties of time-series

- You will usually come across these four terms describing time series properties in the literature:
  - Autocorrelation
  - Trend
  - Seasonality
  - Stationarity
- Each of these properties is important and influence the model and preprocessing steps selection.

Autocorrelation

- Correlation measures the extent of a linear relationship between two variables.
- Autocorrelation measures the linear relationship between current and lagged values of a time series.
- High autocorrelation of a time series is a positive phenomenon because the current value is dependent on the past values thus this dependency can be included in the model.

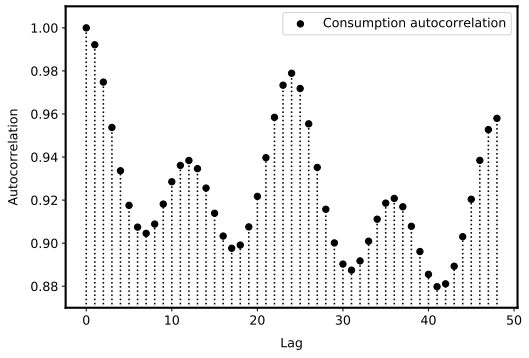


Figure 1: Consumption autocorrelation function values for lags from 0 to 48.

Trend

- Trend is a long-term increase or decrease in the data.
- It does not have to be linear, can be exponential, polynomic, etc.
- Trend can change in the long run, it is referred as “changing direction,” when it might go from an increasing trend to a decreasing trend.

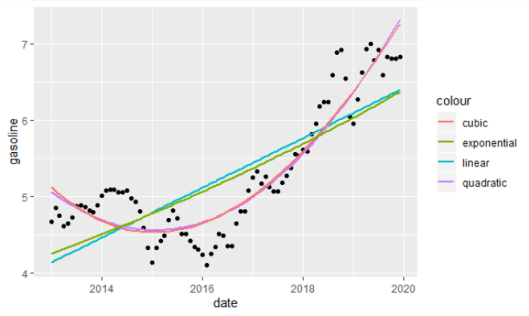


Figure 2: Example of different trend types.

Seasonality

- A seasonal pattern occurs when a time series is affected by seasonal factors.
- It can be a month of the year or a day of the week for example.
- Seasonality is always of a fixed and known frequency.

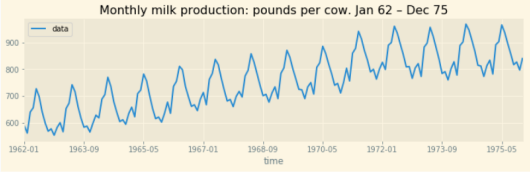


Figure 3: Example of seasonality.

Common properties of time-series

Cycle vs. Seasonality

- Many people confuse cyclic behaviour with seasonal behaviour.
- If the fluctuations are not of a fixed frequency then they are cyclic.
- If the frequency is unchanging and associated with some aspect of the calendar, then the pattern is seasonal.
- Cyclic fluctuations can be associated for example with economic conditions (business cycle).

Common properties of time-series

Stationarity

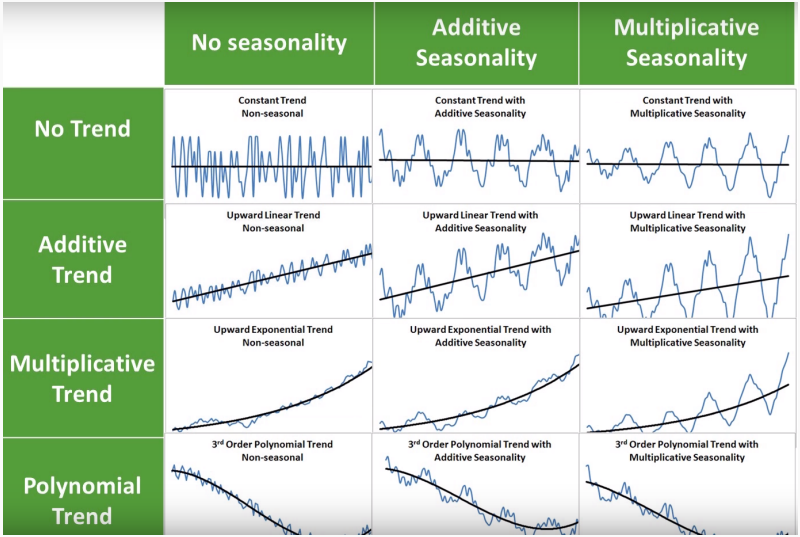
- A stationary time series doesn't depend on the time at which the series is observed.
- Most intuitive definition would be that the time series has the same mean and variance over time.
- Time series with trends, or with seasonality, are not stationary — the trend and seasonality will affect the value of the time series at different times.
- If there is a cycle present time series still could be stationary on the other hand - cycles are aperiodic.

Common properties of time-series

Stationarity

- White noise series is stationary — it does not matter when you observe it, it should look much the same at any point in time.
- A stationary time series will have no predictable patterns in the long-term.
- It may seem that making time series stationary would make it unpredictable at first glance. Opposite is true, if you are able to extract trend and seasonal patterns from the series, you can forecast them separately easily and the last amount of variance in the stationary series can be explained often by past values or exogenous variables.

Trend and seasonality combined



Preprocessing the data

As we could see on the previous slide - there is a wide variety of trend, seasonality and variance combinations.

Our goal is usually to remove these factors from the time series in the pre-processing phase so is become stationary (ideally).

We can use these steps for doing so:

- Transformation
- Differencing
- Decomposition

Transformation

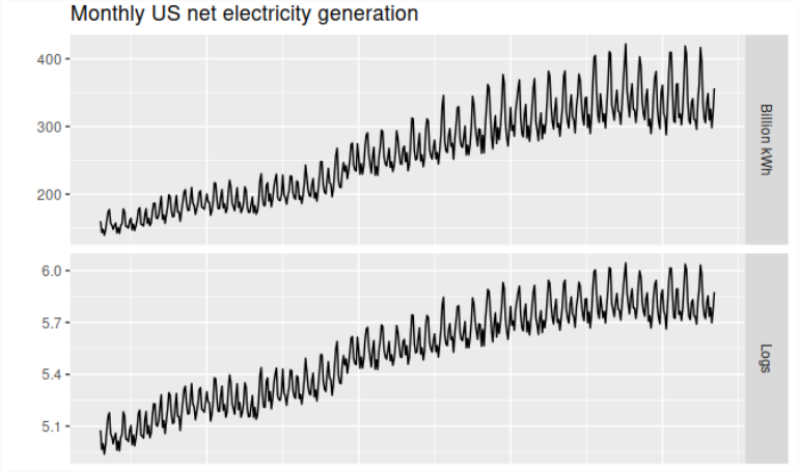
- Our goal is stabilizing the variance over time.
- The most common is logarithmic or Box-Cox [4] transformation.
- Beware that the logarithmic transformation works only for positive values, zeroes can be fixed by adding small constant.
- If the numbers are negative, you can use Box-Cox instead.

$$y_t^{log} = log_a\left(y_t\right)$$

$$y_t^{original} = a^{(y_t^{log})}$$

[4] https://www.statisticshowto.com/box-cox-transformation/

Transformation



Preprocessing the data

Differencing

- Computing the differences between consecutive observations.
- Differencing can help stabilizing the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality.
- Differencing can have several orders or be based on seasonality. The most common is first-order difference. Order tells you how many differencing operations were performed on the time series.

$$\textit{1st order: } y'_t = y_t - y_{t-1}$$

$$\textit{2nd order: } y''_t = y'_t - y'_{t-1}$$

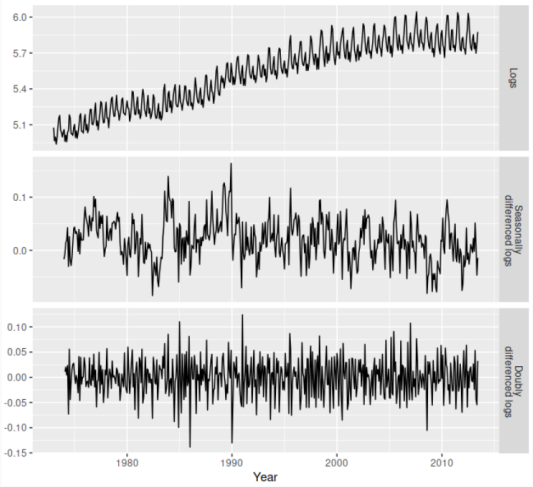
Differencing

- A seasonal difference is the difference between an observation and the previous observation from the same season. These are also called “lag-m differences,” as we subtract the observation after a lag of m periods.

$$y'_t = y_t - y_{t-m}$$

Preprocessing the data

Differencing



Decomposition

- Time series data can exhibit a variety of patterns, and it is often helpful to split a time series into several components, each representing an underlying pattern category.
- Decomposed time series usually have these three components:
  - Trend component
  - Seasonal component
  - Residual component
- Decomposition can be either additive or multiplicative. Multiplicative decomp. is less frequent, but can be used for time series with non-constant variance. Usual practice is that the time series is transformed in the first step thus has constant variance and we can employ additive approach.

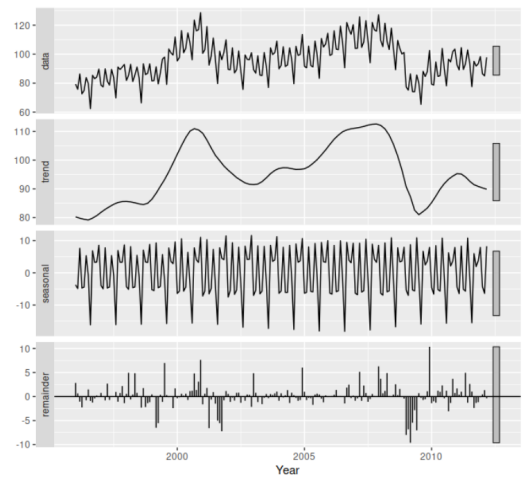
Decomposition

- If we assume an additive decomposition, then we can write:

$$y_t = T_t + S_t + R_t$$

- T_t is trend component, S_t is the seasonal component and R_t is the the remainder (residual) component.
- Seasonal component is periodic, thus it is trivial to predict. Trend can be extrapolated if there is any. Forecasting model usually predicts only a residual component values and uses trend and seasonal component as an another exogenous variable.
- All three components are added together after the forecast phase and preprocessing steps are reversed thus you obtain raw value of endogenous variable in the end.

Decomposition using STL



Assessing Model Effectiveness

All models are wrong, but some are useful. - George E. P. Box

- Every model you create is able to provide you with some number as a forecast.
- Models are not equal and it is always important to evaluate them.
- There are several metrics which focuses on amount of errors in the forecasts from different perspectives.
- We will list the most common ones, but note that there exist some other metrics or variants of them specific for different domains.
- Usually we use multiple metrics for model accuracy assessment.

- **Mean Absolute Error (MAE)** - is the average of the absolute difference between the predicted and actual value. It is highly affected by outliers.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - g(\overline{X_i})|$$

Forecasting- Assessing Model Effectiveness

- **Mean Squared Error (MSE)** - is the average of the squared difference between the predicted and actual value. It is differentiable and may be used for optimization.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - g(\overline{X_i})\right)^2$$

- **Root Mean Squared Error (RMSE)** - is the square root of the average of the squared difference of the predicted and actual value. The root mean is able penalize large errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - g\left(\overline{X_i}\right)\right)^2}$$

Forecasting- Assessing Model Effectiveness

- The effectiveness of the linear regression models can be evaluated with a measure known as **R^2-*statistics*** or *coefficient of determination*.
- The standard Sum of Squared Error is defined for a model $g(\overline{X})$ as:

$$SSE = \sum_{i=1}^{n} \left(y_i - g(\overline{X_i})\right)^2$$

- The Squared Error of the response variable about its mean is defined as:

$$SST = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} \frac{y_j}{n} \right)^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Forecasting- Assessing Model Effectiveness

- The R^2-*statistics* is then defined as:

$$R^2 = 1 - \frac{SSE}{SST}$$

- The value is always between 0 and 1 and higher are more desirable.
- For high dimension data, **adjusted** version is more accurate:

$$R^2 = 1 - \frac{(n-d)SSE}{(n-1)SST}$$

- The R^2-*statistics* is not applicable on the nonlinear models.
- The nonlinear regression may be evaluated using pure SSE [5].

[5] `https://statisticsbyjim.com/regression/r-squared-invalid-nonlinear-regression/`

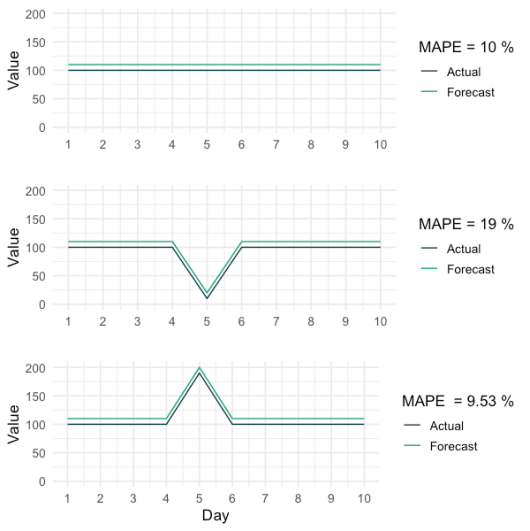
Forecasting- Assessing Model Effectiveness

- **Mean Average Percentage Error (MAPE)** - is the average percentage error between the predicted and actual value. It is scale invariant.

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - g(\overline{X_i})}{y_i} \right|$$

- It fails if some of the actual values are equal to zero.
- If any true values are very close to zero, the corresponding absolute percentage errors will be extremely high and therefore bias the MAPE.

Forecasting- Assessing Model Effectiveness



- **Symmetric Mean Average Percentage Error (SMAPE)** - is the symmetric average percentage error between the predicted and actual value.

$$SMAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{|y_i - g(\overline{X}_i)|}{\frac{|y_i| + |g(\overline{X}_i)|}{2}}$$

- A limitation to SMAPE is that if the actual value or forecast value is 0, the value of error will boom up to the upper-limit of error - 200 %.

Model evaluation methods

- We know several methods for evaluation models for vector data, e.g. k-fold CV, Leave-one out or traditional train/test split.
- These methods are not directly applicable in the time series domain because they are based on random (with or without stratification) division of the **unordered** data to training and testing set.
- We must be very careful when we are splitting the time series data into groups because of the temporal dependencies in them.
- We must utilize methods specific for the time series area to prevent data leakage.

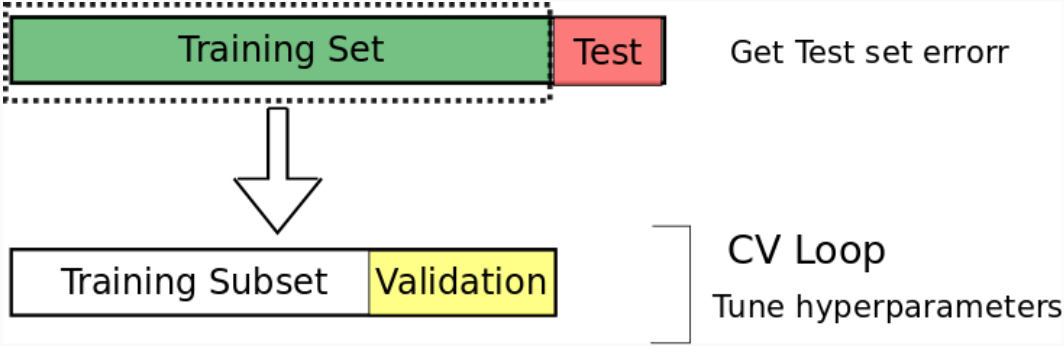
Model evaluation methods

Train/test set split

- This is the simplest method of the mentioned.
- The whole point is that you take part of the data to the specific timestamp in the chronological order as training set.
- Test set consists of the data following the specific timestamp.
- This way you make sure that there will be no information leakage because ”future” data are strictly separated from the training set.
- You can of course employ validation set as well, validation data must be in between train and test set split timestamps.
- Disadvantage of this method is that arbitrary chosen test set may lead to the bias (either too good or too bad results).

Model evaluation methods

Train/test set split



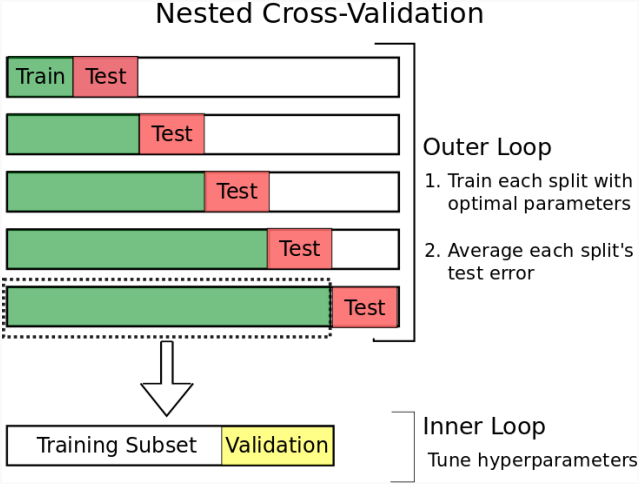
Cross validation

- This method is very similar to a classical k-fold CV.
- It splits the dataset into multiple different training and test sets as well.
- The error on each split is averaged in order to compute a robust estimate of model error.
- The split is chronological as in the previous case, but we are utilizing so-called time window.
- Time window could be either fixed of expanding (sometimes called nested).

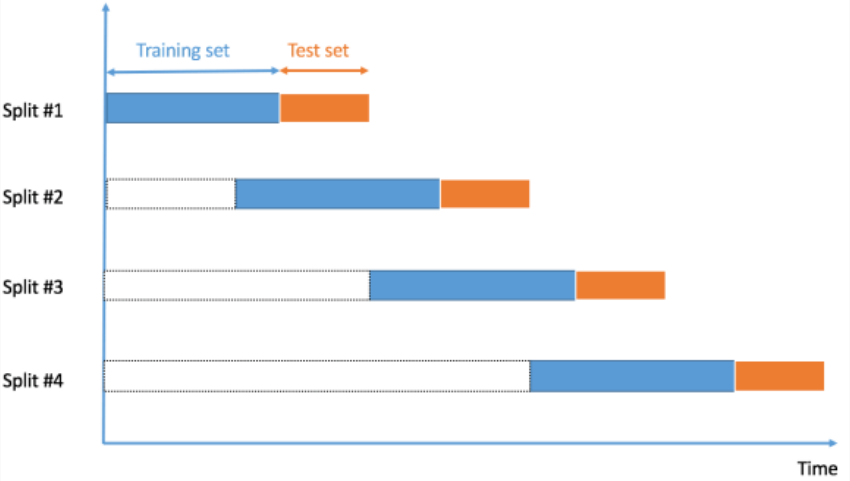
Cross validation

- Fixed time window use the same timespan in each split, thus it uses training and testing set of the same size each time. The fixed window is shifted by the test set length in each split.
- Expanding time window uses only the same size of a test set. Training set starts small and is expanded by the test set from the previous split. Test set is shifted in each split.
- Fixed time window approach is more fair to the model performance among the splits because each model has the same amount of data available in the training phase.
- Expanding time window approach is on the other hand more realistic because you would expect that the model won't be trained only once but you would want to re-train the model after some period of time, when you have

Cross validation - expanding window



Cross validation - fixed window



Visual evaluation of the model

- Mentioned metrics give you information about the amount of a error in the model in global.
- Sometime it is desirable to focus on the specific aspects of the model.
  - You may want to take a look how the errors are distributed according to the calendar features for example.
- It is very useful to analyze residual errors of the model, e.g. if the residual errors are stationary or normally distributed.
- Used visualization techniques depends on the specific case, but scatter plots, histograms or box plots are commonly employed.

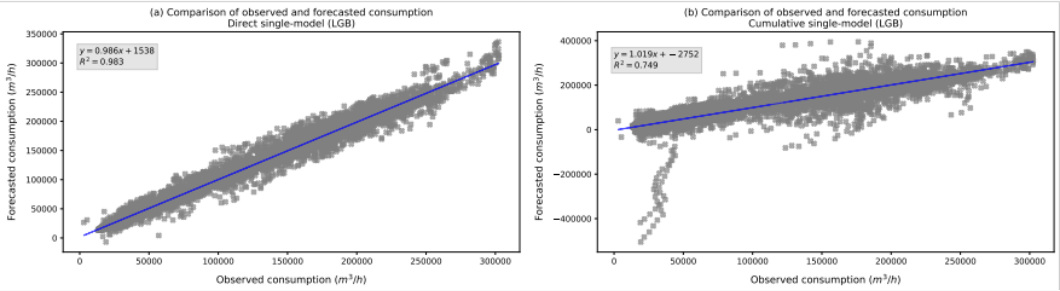
Visual evaluation of the model

Residual auto-correlation
If there are correlations between residuals, then there is information left in the residuals which should be used in computing forecasts.

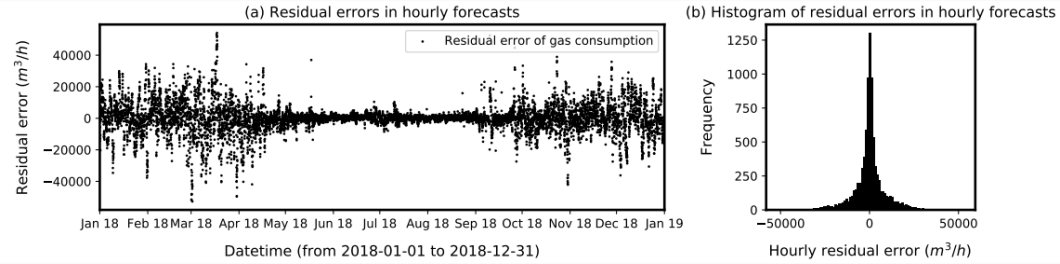
Zero mean
If the residuals have a mean other than zero, then the forecasts are biased.

- Note: Residual errors in a time series model are what is left over after fitting a model.
- I.e. the difference between the observations and the corresponding fitted values ($e_t = y_t - y_f$).
- You can use histogram, Q-Q plot or ACF plot for this task.
- It can be also useful to plot the true and forecasted values in the scatter plot. It is mostly done for the linear model, majority of the points should be centered around the regression line.

Visual evaluation of the model - examples



Visual evaluation of the model - examples



Questions?

Questions?