Functions are listed in the order they would be executed by the protocol.

Practical Function is tested on an AHU data set, detailed information can be taken from the “Evaluation” excel file, which lists all tests detailed.

More background information might be found in “Details for MainConcept” word file or on the website of Scikit-learn.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | In | Out | Theoretical function | Practical  Function |
| Resampling | Imported Data | Resampled Imported Data | Resampling to desired resolution (own) | As theoretical |
| (Initial custom feature selection) InitFeatureSelect | Output of previously conducted function | Same data minus features that haven´t been selected | Custom selection of features before tuning the data (own) | As theoretical |
| (Data cleaning)  NaN\_Dealing+impute | Output of previous function | Data without NaN´s | Free the data from NaN´s and infinite values (pandas and scikit) | As theoretical  Impute:  Not implemented (future work |
|  |  |  |  |  |
| Scaling and normalizing | Output of previously conducted function | Scaled and normalized Data | Most models need scaled and normalized data to properly weight features (scikit) | As theoretical +  Robust scaler achieved best results, while the choice of scaler has little influence if OwnLags are used |
| Time series plotting | Data without NaN´s + Scaled and normalized Data | Plots of each feature and signal over time. Unscaled and Scaled | Visualization of the signals time series for detecting extraordinary patterns or mistakes in the data (via matplotlib) | As theoretical |
| (Custom period selection)  ManSelect | Scaled and normalized Data | A specific period of that data | (Custom selection of a period for either selecting proper data or cutting data for faster computation)  Reduce computational costs. Extract important periods or neglect faulty periods (own) | As theoretical |
| Cross-, cloud- and autocorrelation plotting | Output of previously conducted function | Auto-correlation plot of signal & Cross- and cloud-correlation plots of each feature with the signal | Visualization of correlations for detecting influential lags and gather insight. (via matplotlib) | As theoretical |
| (Creation of differences)  DifferenceCreate | Output of previously conducted function | Same Data plus created “derivatives” of the selected features | (Creation of feature derivatives) of certain features for tuning performance. Especially meaningful if physical behavior is depending on the derivative of a parameter (own) |  |
| Polynomial feature creation | Output of previously conducted function | Same Data plus created polynomials | Creation of polynomials of all features (scikit) | Not implemented (future work |
| Custom feature creation | Output of previously conducted function | Same data plus transformed set of features | Any mathematical function, but only for transforming each feature without interaction with others (scikit) | Not implemented (future work |
| Automatic ownlag and featurelag construction through auto- and crossvalidation values | Output of previously conducted function + auto- and cross-validation values per lag per feature + scaled and normalized data of the whole period | Same Data plus created ownlags and featurelags | Automatic creation of meaningful lags without the need of a model (wrapper) | Not implemented (future work) |
| (Custom featurelag creation)  ManFeaturelagCreate | Output of previously conducted function + scaled and normalized data of the whole period | Same data plus manually chosen featurelags | Custom creation featurelags (own) | As theoretical |
| (Automatic featurelag creation)  AutoFeaturelagCreate | Output of previously conducted function | Same data plus automatically chosen featurelags | Wrapper for automatic creation of the best lag per feature within a custom lag range. Each lag is only created if beneficial. Based on the assumption that only one lag per feature has real predictive power.  Produces (if performing better) max one featurelag per feature. Lags are chosen between the stated boundaries. Through lagged features lagged behavior shall be introduced, e.g. the lagged influence of heater temperature to room temperature (own) |  |
| (Custom ownlag creation)  ManualOwnLagCreate | Output of previously conducted function + scaled and normalized data of the whole period | Same data plus manually chosen ownlags | Creation of custom ownlags.  Produce the ownlags chosen manually by the user(own) | As theoretical |
| PCA | Output of previously conducted function | Same data plus through PCA created features | Adding physical meaningless but statistical powerful features | Not implemented  (Not expected to work well on regression problems in HVAC systems) |
| ManualFeature-Combination or creation | Output of previously conducted function | Same data plus combinations, etc. of existing features | Manually adding physical and statistical powerful features (e.g. derivatives of features or voltage\*current = power) | Not implemented  (But expected to be powerful. Up to now implemented because this broad range of possible combination can more easily be done directly in the excel file) |
| Tsfresh |  |  | Automated feature construction for time series (creates more than 1000 features just out of the signal) (tsfresh) | Not implemented  (expected to very powerful if regular feature only give rare information; maybe very powerful in combination with an native recurrent(for pure timeseries forecastin)) |
| Low variance filter | Output of previously conducted function | Same data minus features that a lower than the stated variance | Delete features with low variance (scikit) | As theoretical |
| (Custom post-creation feature selection)  Manual feature Select | Output of previously conducted function | Same data minus features that haven´t been selected | Custom feature selection of created features  Chose manually which features should be used (own) | As theoretical |
| Independent component analysis | Output of previously conducted function | Data in the same shape as before but with through ICA changed values | Separating of superimposed signals (See Scikit-learn) (scikit) | Not known. |
| Variable ranking filter (univariate)  Filter\_univariate | Output of previously conducted function | Same data minus features that have been filtered out | Several search and rating strategies for univariate filter (scikit)  Delete meaningless features without the use of a model (embedded or wrapper) (scikit) | Not tested |
| Recursive embedded feature selection (Embedded) | Output of previously conducted function | Same data minus features that have been filtered out | Embedded variable subset selection, see (scikit) for information. Number of features can be found both automatically and custom.  Delete recursively meaningless features through using models with intrinsic feature weighting.  <http://scikit-learn.org/stable/modules/>  feature\_selection.html  #recursive-feature-elimination  (scikit) | As theoretical but it only increased accuracy for the model used for the RFS. ANN, SVR, GB and Lasso achieved for the investigated AHU better results with all features |
| (Embedded feature selection by threshold)  Embedded threshold feature selection | Output of previously conducted function | Same data minus features that have been filtered out | Variable ranking feature selection by a given threshold of importance. (scikit)  Delete meaningless features through using models with intrinsic feature weighting based on a threshold on their value of “Feature Importance”. <http://scikit-learn.org/stable/modules/>  generated/sklearn.feature\_selection.  SelectFromModel.html  (scikit) |  |
| Recursive wrapper feature selection  Wrapper (Recursive feature selection) | Output of previously conducted function | Same data minus features that have been filtered out | Variable subset selection wrapper which recursively deletes the worst feature as long as it improves the score. (own)  Delete meaningless features through comparing performance on test data sets using models  (own) | Inherent feature selection of the chosen models seem to perform so good, that the wrapper achieves similar results whether a feature is deleted or not. Retraining and statistical issues affect the result more, hence the wrapper deletes no or very few seemingly arbitrary features. |
| Wrapper with advanced search functions e.g. Bayesian Optimization or generic algorithms | Output of previously conducted function | Same data minus features that have been filtered out | Delete meaningless features through comparing performance on test data sets using models | Not implemented |
| Automated Period \_Ownlag Creation | Output of previously conducted function + scaled and normalized data of the whole period |  | Adding optimal period ownlags | Not implemented |
| Automated time series ownlag creation | Output of previously conducted function + scaled and normalized data of the whole period | Same data plus optimal number of TimeSeries-ownlags | Wrapper for creating the optimal number of time series ownlags. Ownlags are added as long as they improve the score, which is based on the assumption that the score is monotonically increasing with the number of ownlags till it reaches the global optimum. (own)  Find optimal number of timeseries-ownlags through adding ownlags (counting up) till accuracy is not increased anymore (own) | As theoretical |
| Automated Period- & TimeSeries- ownlag creation | Output of previously conducted function + scaled and normalized data of the whole period |  | Adding optimal period and timeseries ownlags | Not implemented |
| Below methods are in RighstideMainConcept |  |  |  |  |
| Individual Model “weekday & weekend” | Output of previously | Data sliced into week data and weekend data | Building one model for weekdays and one for weekends  Training one model only for weekdays, and one only for weekends (own) |  |
| Individual Model “hourly” | But only one | Data sliced into data sets per hour | Building one model per hour, e.g. one model for all days from 0 till 1 o´clock (own)  (own) |  |
| Individual Model “By Feature” | IndivModel can be selected | Data sliced into data sets with values of a feature above and below a certain threshold | Building one model for all samples with the feature´s value below and one for values above a certain threshold. The respective feature is user-defined.  E.g. for temperature, if with a certain temperature the control strategy changes, or for an on/off feature.  (own) |  |
| Shuffle | Output of previously | Data with randomly shuffled samples | Random shuffle of the train period.  (scikit)  to delete any bias while training the model (some models bias their training based on the samples already used for training)  (scikit) | As theoretical,  for the investigated AHU data shuffling increased accuracy |
| Model selection | Tuned Data | The error of the best model and a txt file stating the best model | Wrapper for selecting the best beneath all implemented models (own)  Selects automatically the model with the best accuracy (SVR; ANN; RF; GB; Lasso)  (own) | As theoretical |
| Recursive prediction |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |