Notes

3: **Methodology**

General approach here

To estimate a continuous value such as the normalized total expenditures (cTOTEXn) for Distribution System Operators (DSOs), we applied various predictive regression models. For all models a two-stage model training approach is implemented, to ensure stable performance evaluation and robustness of feature selection. This is achieved by training and testing across five different seeds and assessing how often different features were selected across models and their respective dataset. The most robust features in different seeds are selected based on a model-specific frequency threshold to retrain the best performing models and obtain stable evaluation and models.

3.1 **Models**

For each of the models, hyperparameter tuning was conducted using cross-validation. Lasso regression was employed due to its intrinsic feature selection and regularization capabilities, which reduce model complexity by shrinking irrelevant coefficients to zero. After training, coefficients smaller than zero are manually set to zero to adhere to monotonicity constraints.

Linear Regression is applied after pre-selecting features using lasso regression to obtain unbiased estimates, offering clearer interpretations of variable relationships with cTOTEXn and statistical inference using p-values. Both Lasso and Linear Regression models required standardization, and coefficients were unstandardized for interpretation.

To capture non-linearities, we utilized Decision Tree Regression, which allowed exploration of complex interactions, with partial dependency plots used for interpretation.

Random Forest Regression, known for its robustness, averaged predictions across multi-ple trees, reducing overfitting risks, and feature importance was assessed using Variable Importance Plots (VIP).

In addition to these individual models, we implemented a cluster-based modeling approach to handle heterogeneity among DSOs. Using DBSCAN, we identified clusters within the training data based on technical block variables—features selected by expert engineers for their relevance to cost factors. Lasso and Random Forest models are then trained within these clusters and the best combination of models is evaluated.

3.2 **Evaluation**

The performance of the predictive models was evaluated using metrics such as Root

Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Per-

centage Error (MAPE), with MAPE prioritized due to its ability to provide relative error

measures that are more interpretable, especially in the context of log-transformed out-

comes. These metrics provided a comprehensive view of each model’s predictive ability.

We also evaluate adherence to monotonicity constraints and identify variable contribution specific to the models. For Lasso and Linear Regression, unstandardized coefficients are analyzed on their sign and strength and for Random Forest and Decision Tree the partial dependency plot gives insight into monotonicity.

1. **Results**

We found that cluster-based models exhibited high variability across different random seeds, leading us to limit further exploration of this approach due to concerns about robustness. Figure 1 shows the best performing models after training and evaluating over all different datasets. Lasso and Linear Regression performed considerably better than the tree-based models with a Testing MAPE of 0.14 and 0.138 respectively. For both regression models and Random Forest, log-transforming the outcome worked well in conjunction with only aggregate variables. When using the whole dataset, log-transforming the features in addition to outcome also performed well for regression models. In Figure X all of the different model and dataset runs are presented.

Ein Bild, das Text, Schrift, Zahl, Screenshot enthält.

Automatisch generierte Beschreibung

Figure 1

Retraining the models based on the robust frequently selected features (See Table XY, table that contains in a list format e.g. yEnergy.losses.tot (5), yInstalledPower.other.tot (5)… Ein Bild, das Text, Screenshot, Zahl, Schrift enthält.

Automatisch generierte Beschreibung

listing all the respective stable variables ), results in following final evaluation of best models:Ein Bild, das Text, Schrift, Zahl, Reihe enthält.

Automatisch generierte Beschreibung

* The best models are trained with log-transformed features and outcome, withLinear regression lower MAPE 0.126 but **lasso** actually even lower RMSE 9.5 mil and test MAE 4.99 mil while very similar MAPE so lasso is probably best model
  + Lasso also adheres to monotonicity constraints and uses rather stable features that appear with a threshold of 80%+
  + Most of predictions are smaller than 15% with one exception around 40%
  + Describe those final best coeficients of lasso Ein Bild, das Text, Screenshot, Schrift, Zahl enthält.

    Automatisch generierte Beschreibung
    - log y log x so like 1% increase in factor is coefficient% increase in y, so e.g. coefficient 0.1 that means 1% increase in x means 0.1% increase in y
* From the linear regression statistically significant are following features

Ein Bild, das Text, Screenshot, Schrift, Zahl enthält.

Automatisch generierte Beschreibung

* Mostly aggregate features are statistically significant
* tree based models tend to overfit do not work as well as the linear models, possibly because they are worse at catching the mostly linear relationships, and also have worse interpretability so our recommendation
  + if we look at regulatory org that wants to have interpretable and monotonous effects with certain cost factors such ensemble models are not suitable in their common form
  + RF partial dependency plot shows that it almost always adheres to monotonicity ocnstraints, which one can see with the flat and increasing parts of the line – show that in appendix

3.3 hier feature selection and variable importance wie du schon hattest

* 1. Limitations

This study presents several limitations that must be acknowledged.

In modeling DSO expenditures, maintaining monotonicity is essential, as certain vari-

ables like energy losses and installed capacity should logically only increase costs. To

enforce this in our Lasso Regression models, we applied a post-hoc adjustment by setting

negative coefficients to zero. However, as this method is applied post-hoc and might limit the optimality of the models, future work could benefit from integrated approaches like non-negative least squares (NNLS) in both Lasso and Linear Regression to enforce monotonicity during model training.

The cluster-based approach demonstrated high variability in performance across clusters, raising concerns about the robustness and reliability of predictions. This instability led to the decision not to pursue cluster-based

models further. Approaches for feature selection were mostly data-driven and could benefit from more domain knowledge. Lastly, due to computational limitations, nested cross-validation

was not employed for model selection and hyperparameter tuning, potentially resulting

in slightly optimistic performance estimates.

This study applied multiple predictive modeling approaches to estimate normalized total expenditures (cTOTEXn) for Distribution System Operators (DSOs), focusing on methods that ensure both accuracy and interpretability. Among the models tested, Lasso Regression, particularly with logarithmic transformations, demonstrated superior performance in terms of MAPE, RMSE and MAE, confirming its robustness and suitability for regulatory applications. The enforcement of monotonicity constraints in Lasso models ensured that the predictions adhered to logical cost behaviors, further enhancing model reliability.

* These were the 2 most important and influential variables : yEnergy.losses.tot and yInstalledPower.other.tot
* We identified specific variable sets that performed best for each model (table xy oben)
* We recommend linear models especially lasso and further exploration in that direction
* In the future using domain knowledge and more compiutational power for exhaustive feature selection using forward or RFE and for better hyperparmater tuning