Traditional Gravity with Decision Trees

Exemplary prediction of bilateral aggregate trade flows

Adrienne Bohlmann

Workshop on Global Economic Studies 25.-26.05.2021, FernUni Hagen, Germany

background

machine learning vs „traditional“ econometric analysis

- econometric analysis: **coefficients**
    - for prediction
    - for evaluation of theoretical models
- machine learning: **out-of-sample prediction-performance**

so far: very few gravity analysis with machine learning
e.g. Gopinath et al. (2020), Jun et al. (2018)

actually few econometric analyses in general with machine learning
Mullainathan and Spiess (2017) consider these methods to have great potential

background

gravity analyses of international trade

Yotov et al (2016)

$$tradeFlow = \frac{GDP_{exp} * GDP_{imp}}{GDP_{world}} * \left(\frac{tradeCosts}{ML_{exp} * ML_{imp}}\right)^{1-elast}$$

- very successful econometric approach of trade analysis

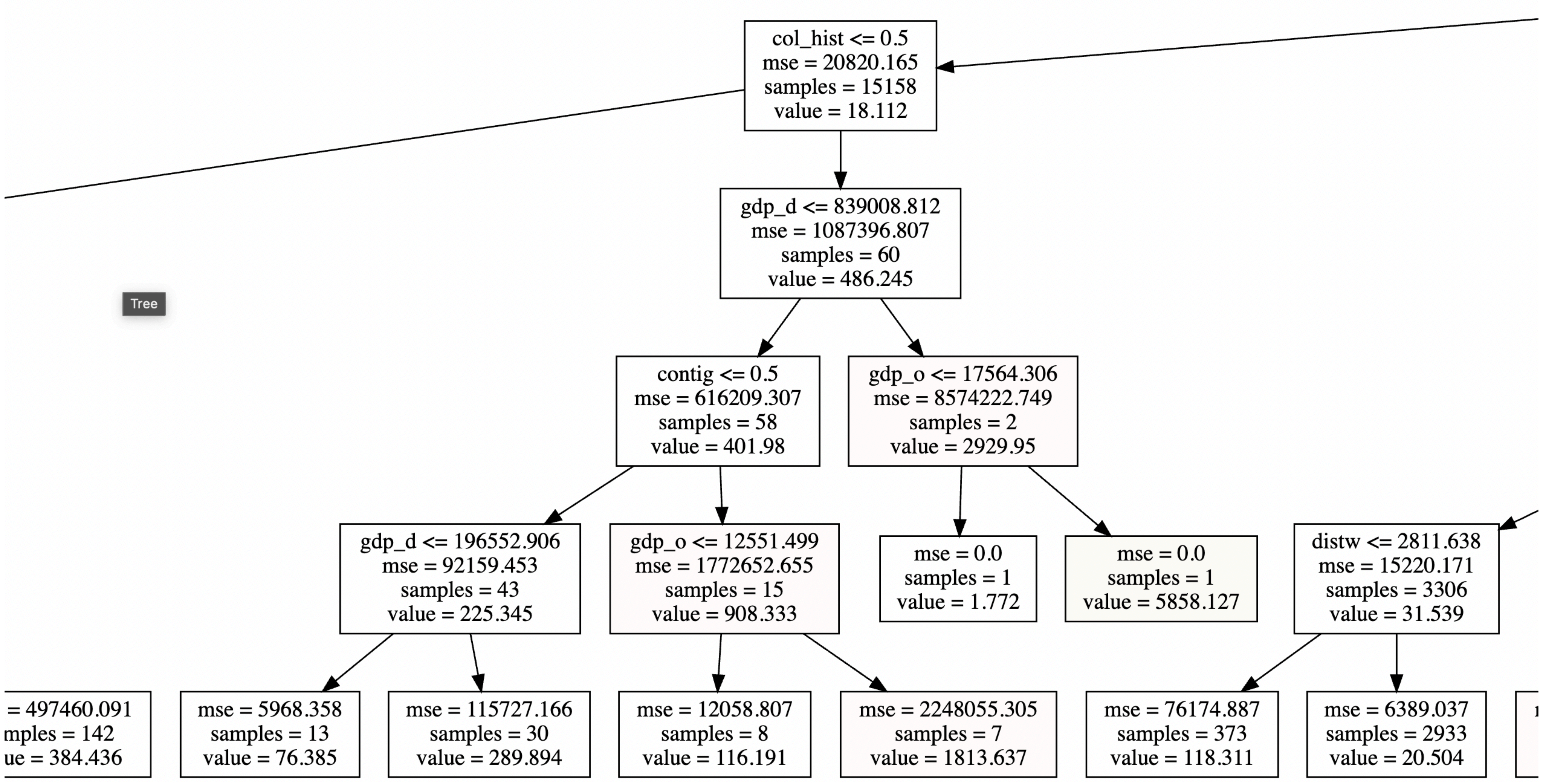
one of many challenges for gravity analyses: **Zeros in trade**
compromising logarithmic and also other transformation procedures, frequently used for handling non-linear relationships and skewed distributions

solutions in the literature e.g.

- dismissing or manipulating Zeros (theoretically inconsistent but widely used)
- two-step/-part analyses: Helpman et al (2008), Egger et al (2011)
- PPML (Poisson Pseudo Maximum Likelihood): Santos Silva and Tenreyro (2006), considered best practice to date

gravity with machine learning - why decision trees?

no transformation of data necessary
=> potential additional solution to Zero-trade problem



small part of simple gravity regression decision tree, restricted to max. depth = 7

very simple gravity setup & the data sources

```
flow ~ distw, gdp_o, gdp_d, contig, comlang_off, col_hist, isl_o, isl_d, lndl_o, lndl_d
```

- **Gravity Cookbook Website**, accompanying Head and Mayer (2014)
  https://sites.google.com/site/hiegravity/data-sources
  "lighter version“ = agg. bilateral international trade and some typical gravity variables, selection:
  - cross section analysis, selecting the year: 2000
  - `flow` (trade flow from origin to destination, Millions of current USD) = dependent variable,
  - `distw` (weighted geographical distance, population-weighted, in km),
  - `gdp_o`, `gdp_d` (GDP of origin and destination, Millions of current USD)
  - `contig` (Dummy for neighborhood status)
  - `comlang_off` (Dummy for common official language)
  - `col_hist` (Dummy for common colonial history)
- common in gravity analyses, data from **wikipedia**
  - `isl_o, isl_d` (Dummies for origin or destination being an island state)
  - `lndl_o, lndl_d` (Dummies for origin or destination being a landlocked state)

Decision Tree Analyses and Results

data preparation & validation of results

biggest problem after first runs:
lack of robustness (very high variation of R^2)
uncovered after repeated randomized train-test-split

=> **stratification with discrete representation of distribution of trade flows**

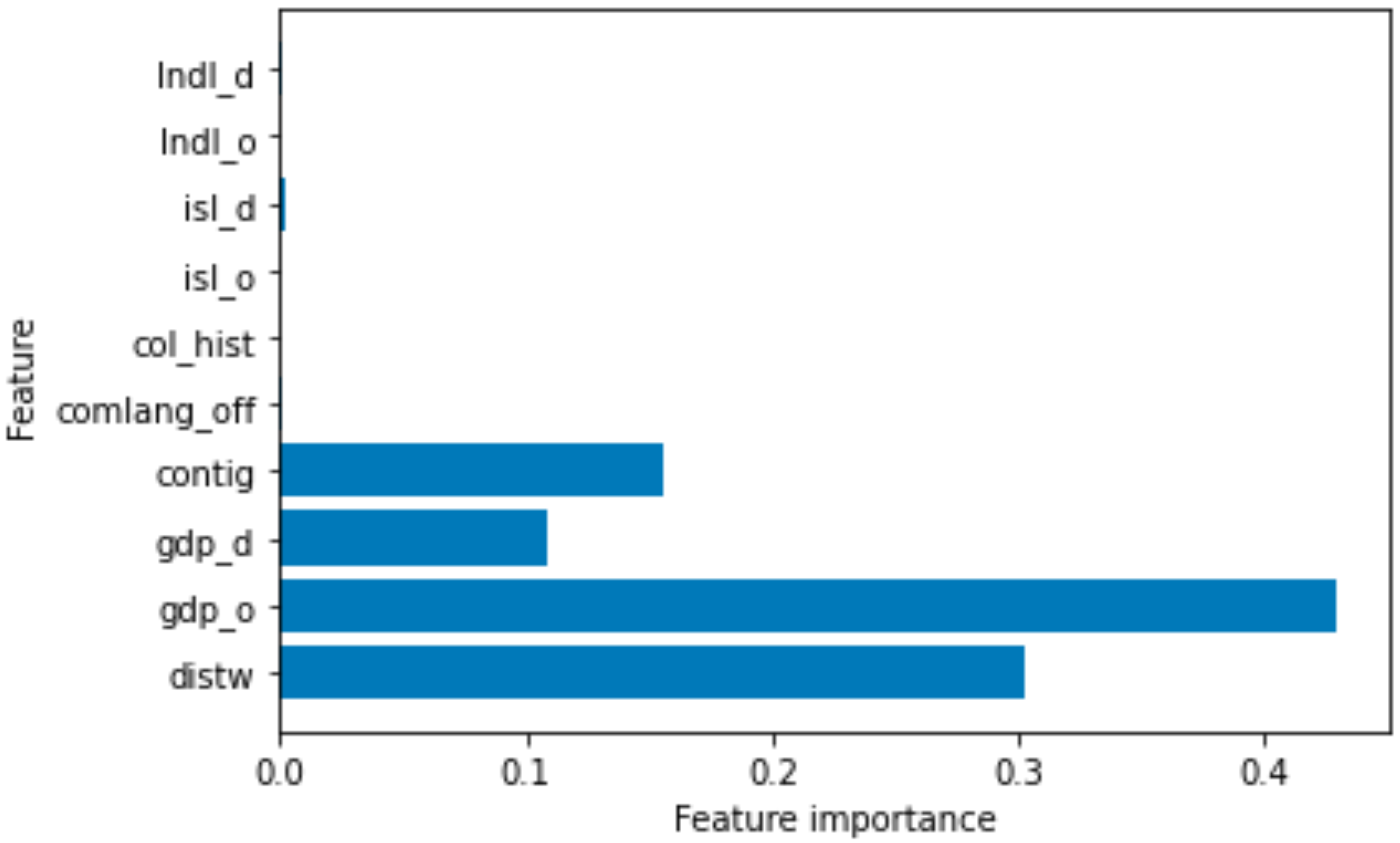
=> instead of cross validation: **random stratified validation (100 runs)**

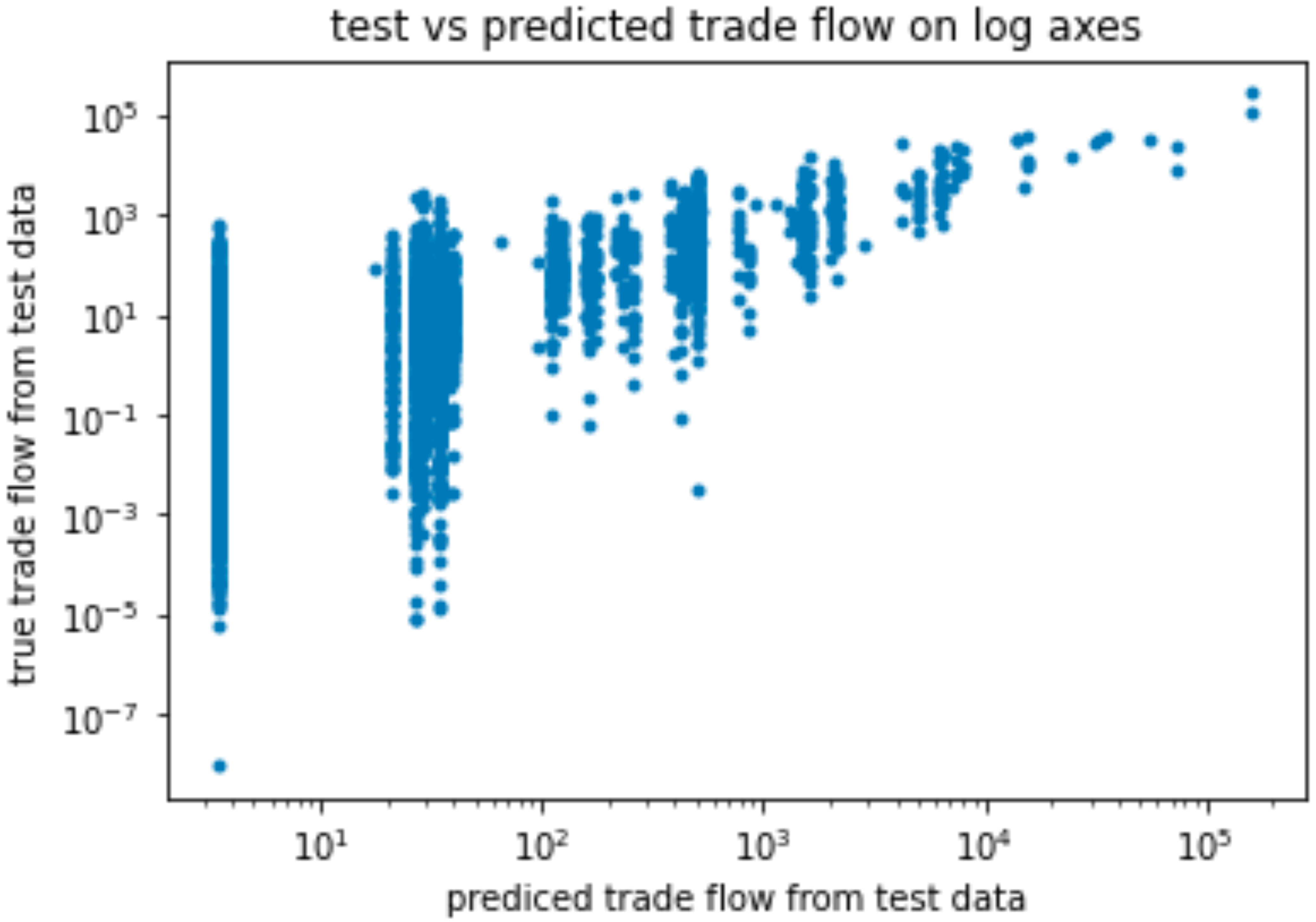
Decision Tree Analyses and Results

maximum depth of 7

single decision tree, exemplary R^2: 74%, mean R^2: 38% (restricted)

(fixed random states) (100 random runs)





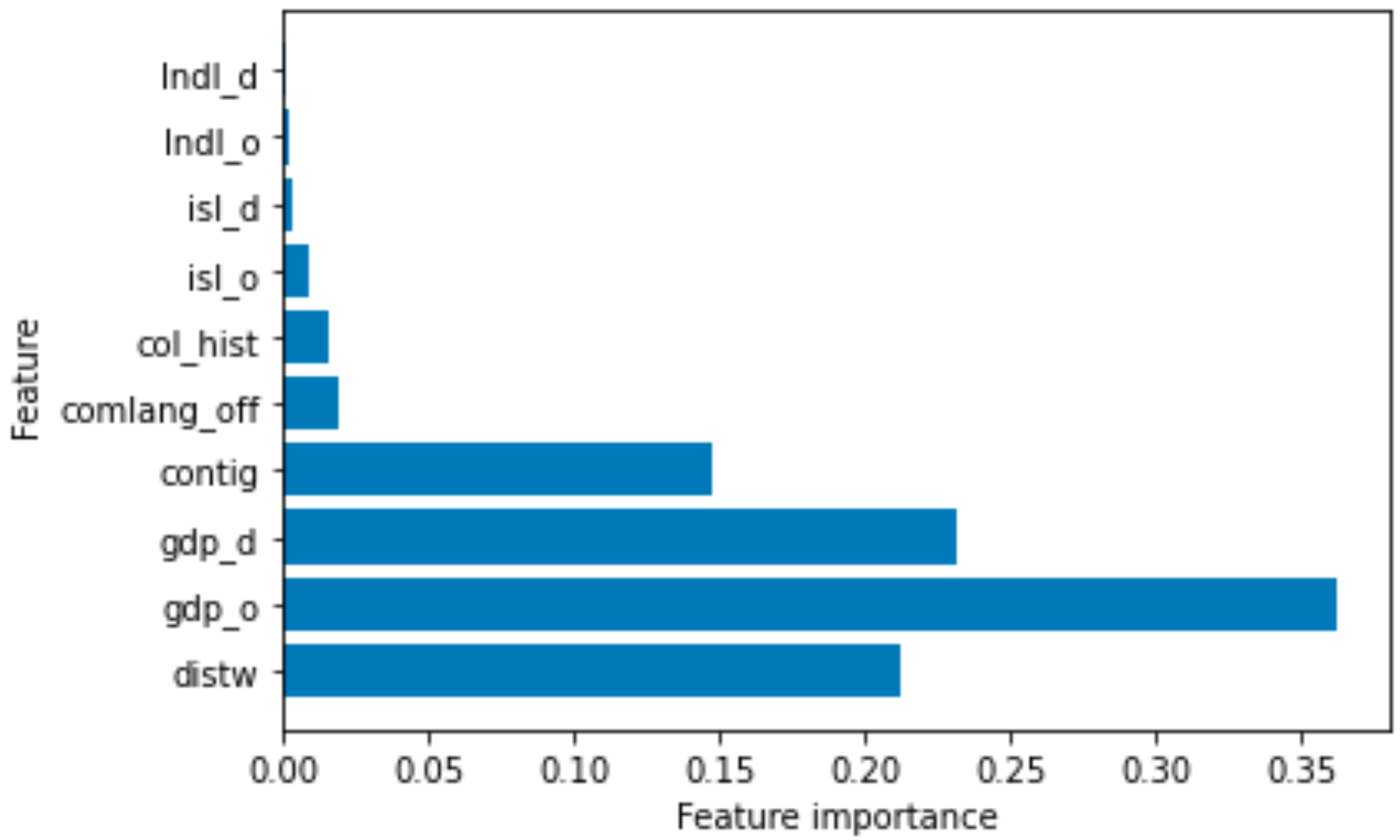
Decision Tree Analyses and Results

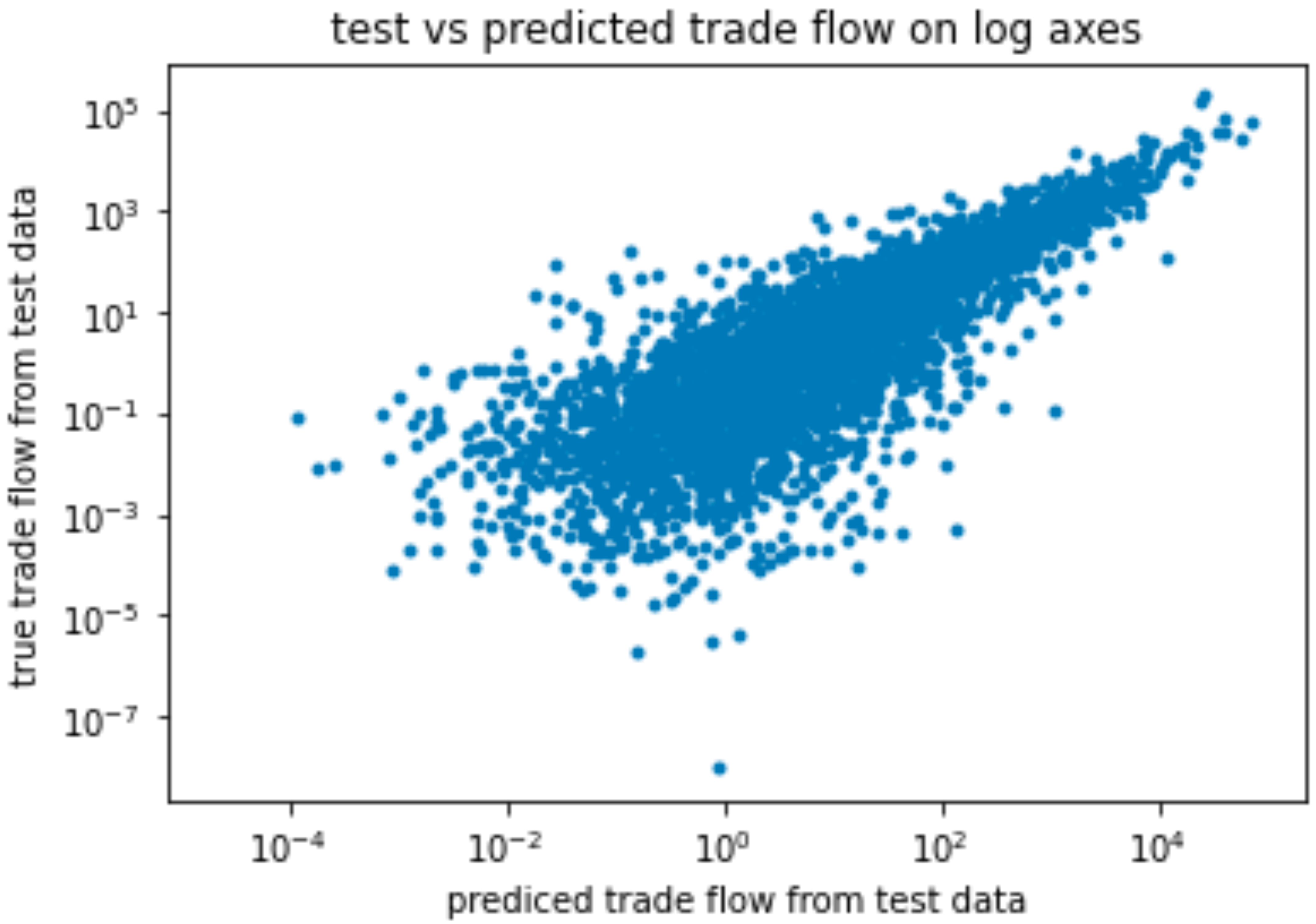
average results of 377 trees of max depth of 13

Random Forest, exemplary R^2 68%, mean R^2 62% (optimized)

(fixed random states)

(100 random runs)



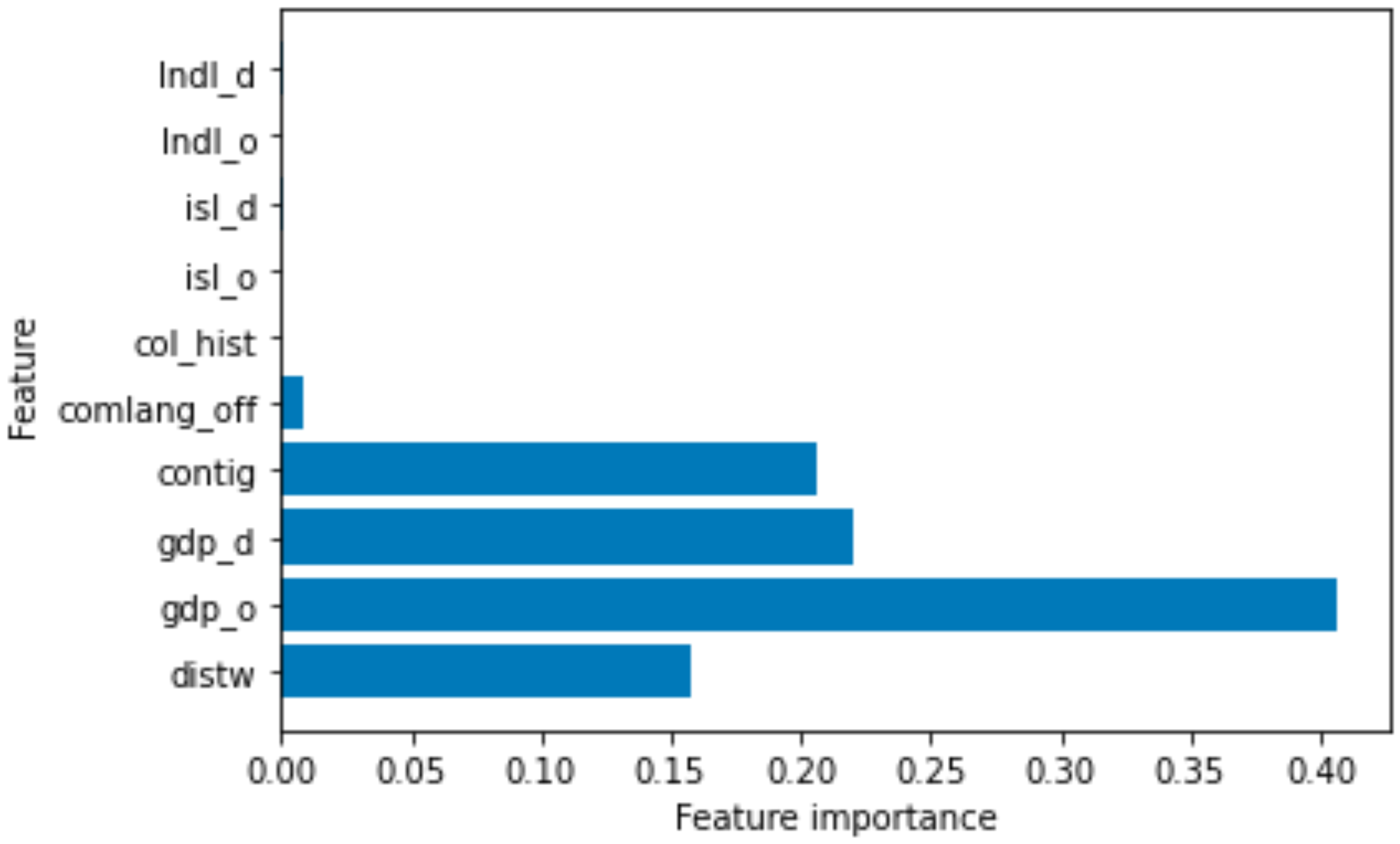


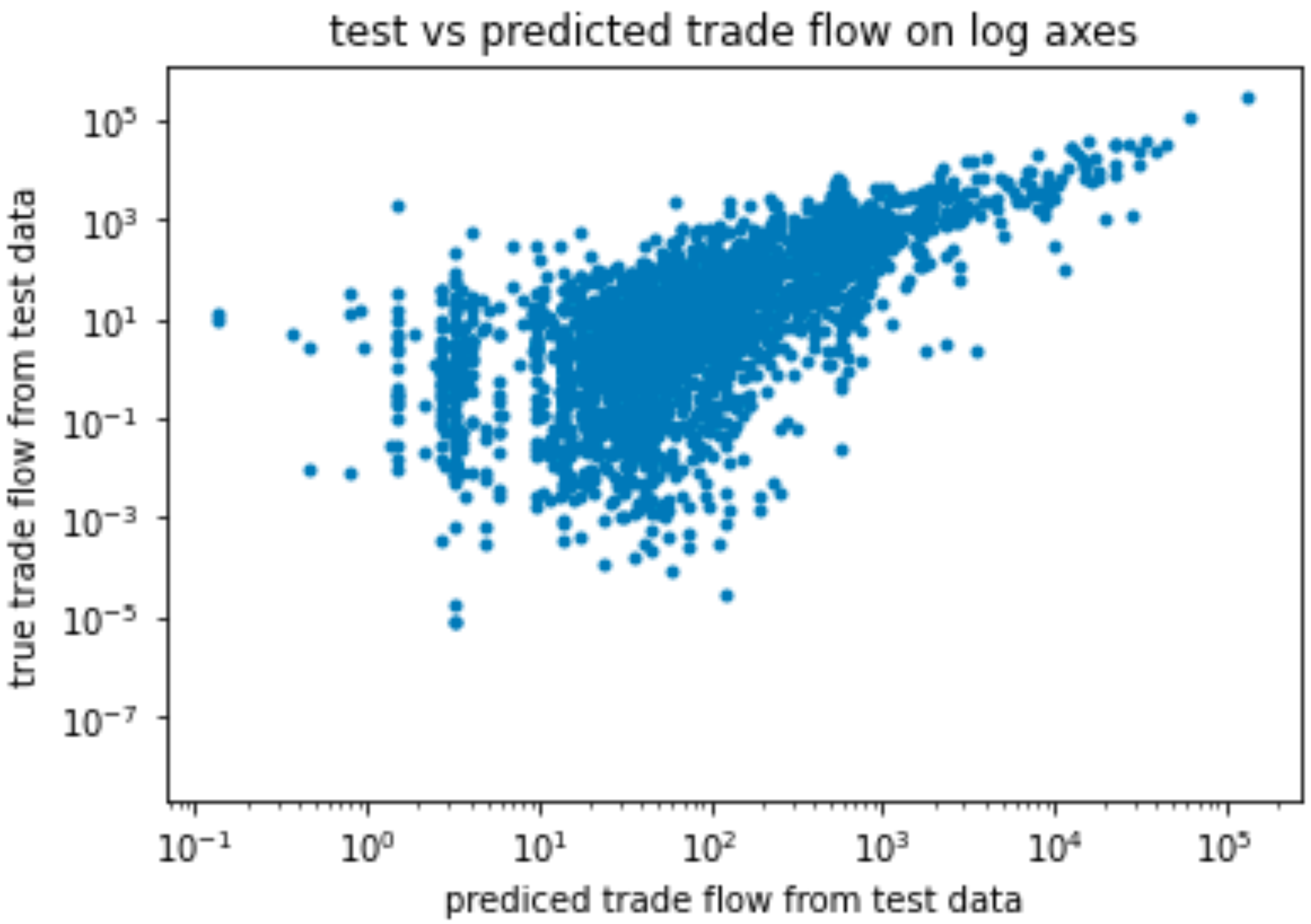
Decision Tree Analyses and Results

200 trees with max depth of 3, learning from another in sequence

Gradient Boosting, exemplary R^2 74%, mean R^2 69% (optimized)

(fixed random states) (100 random runs)



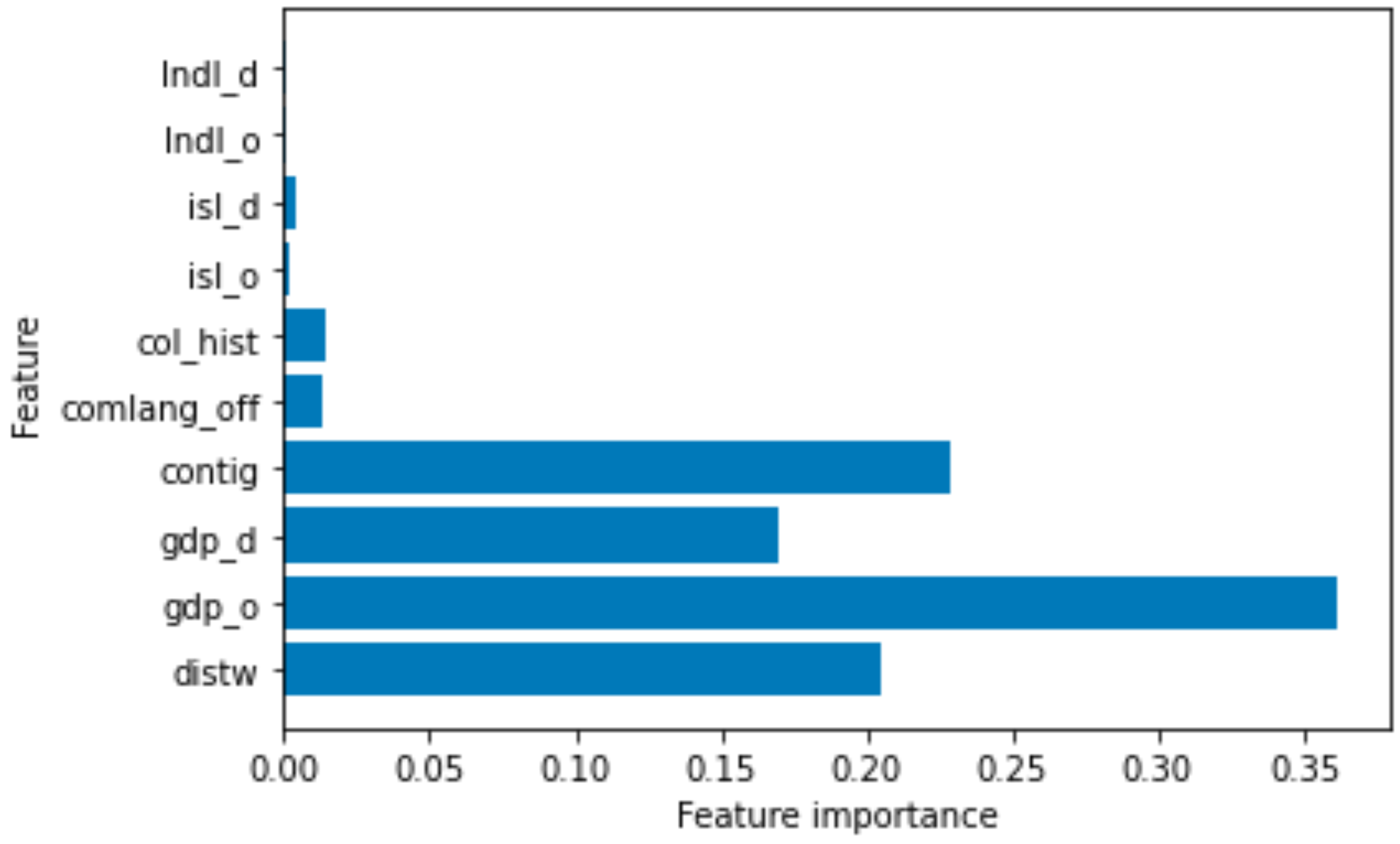


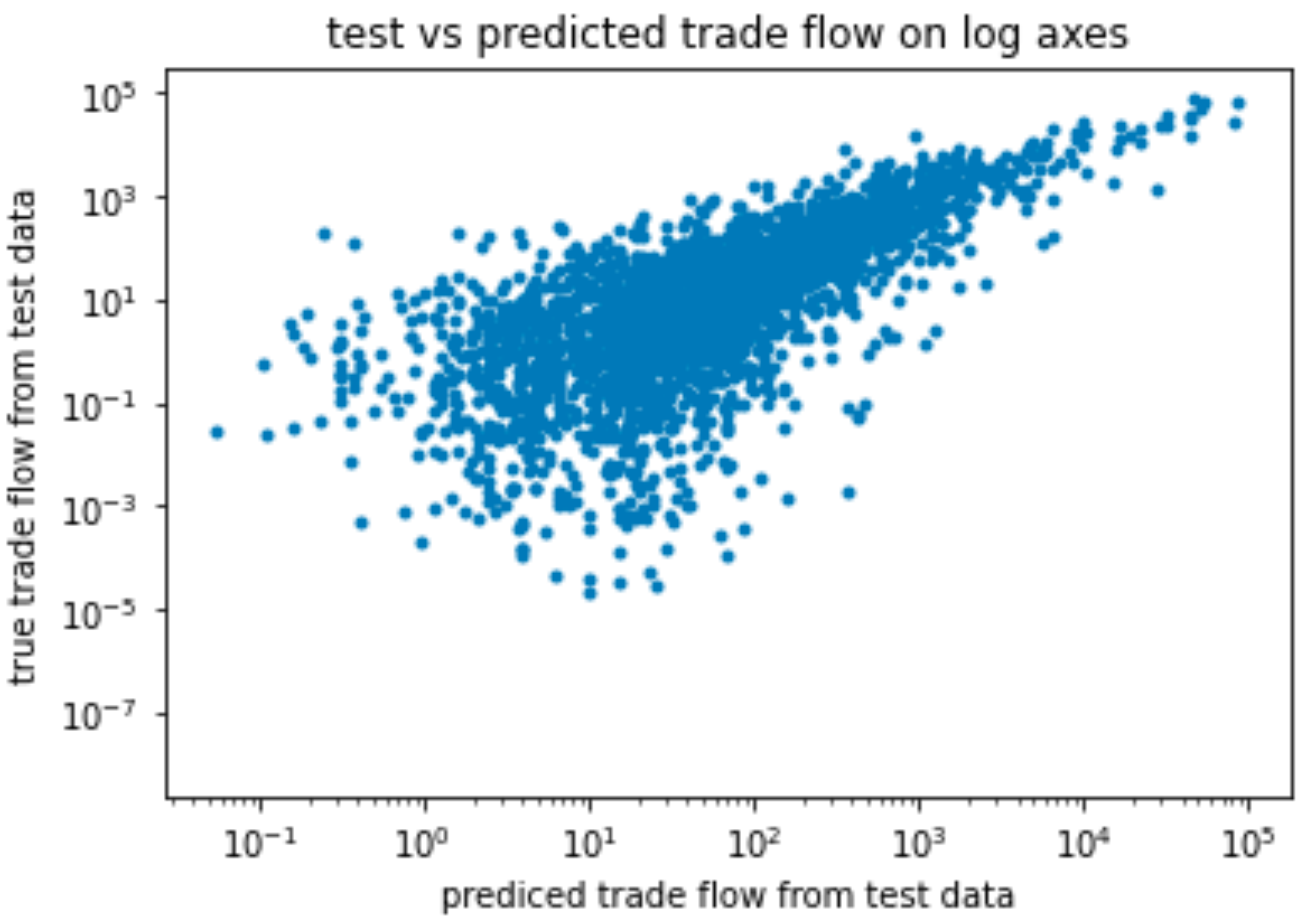
Decision Tree Analyses and Results

200 trees with max depth of 5, learning from another in sequence

Stochastic Gradient Boosting, exemplary R^2 85%, mean R^2 73% (optimized)

(fixed random states) (100 random runs)





Gravity Best Practice: PPML, Santos Silva and Tenreyro (2006)

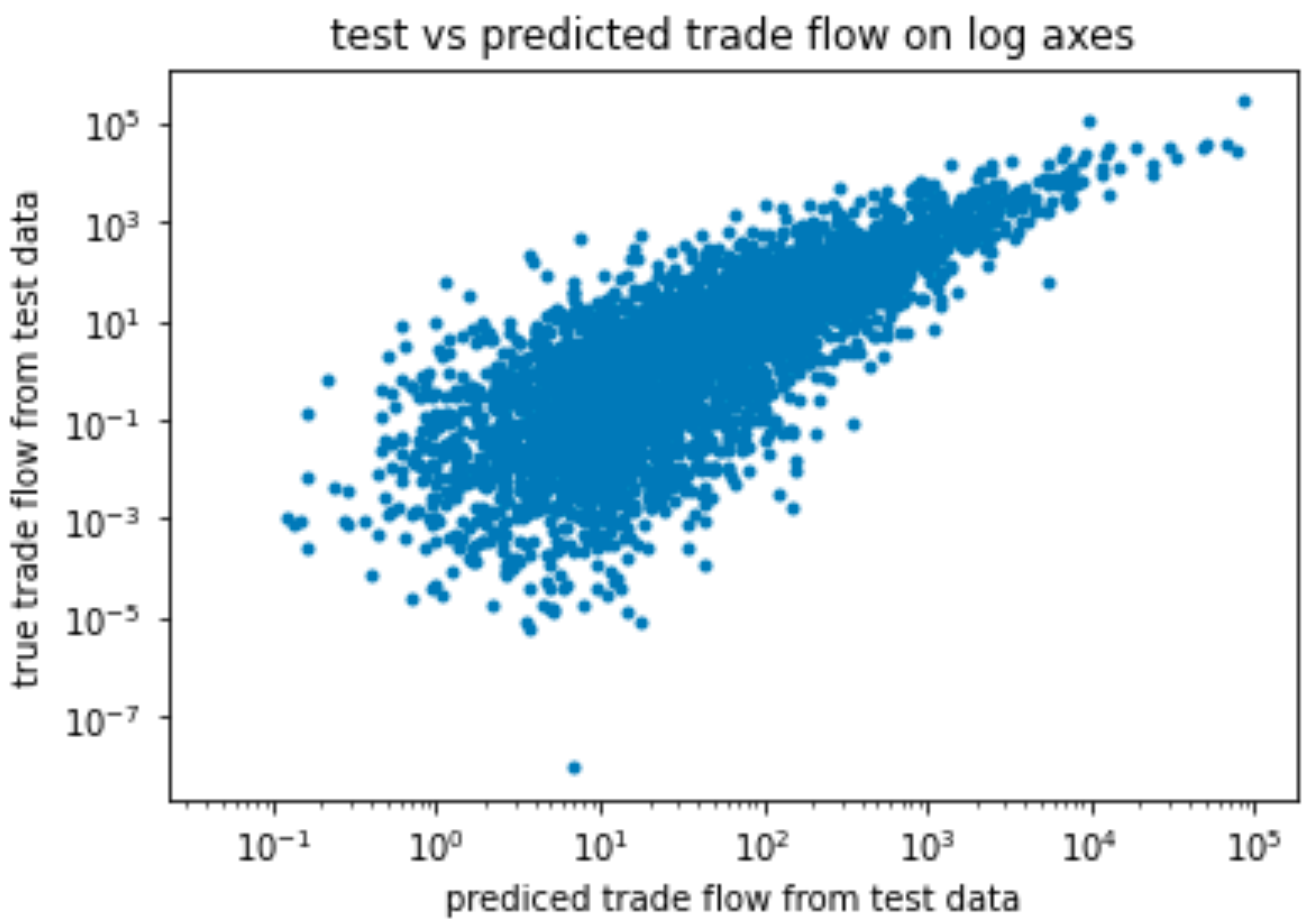
PPML, exemplary R^2 52%, mean R^2 55% (not optimized)

(fixed random states) (100 random runs)

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	21573
Model:	GLM	Df Residuals:	21563
Model Family:	Poisson	Df Model:	9
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3.1565e+06
Date:	Sun, 16 May 2021	Deviance:	6.2600e+06
Time:	16:31:30	Pearson chi2:	1.03e+07
No. Iterations:	7		
Covariance Type:	HC1		

		coef	std err	z	P>|z|	[0.025	0.975]
x1	ln distw	-1.0123	0.025	-40.667	0.000	-1.061	-0.963
x2	ln gdp_o	0.6068	0.021	29.523	0.000	0.566	0.647
x3	ln gdp_d	0.6300	0.021	29.791	0.000	0.589	0.671
x4	contig	0.3409	0.191	1.783	0.075	-0.034	0.716
x5	comlang_off	0.4507	0.167	2.696	0.007	0.123	0.778
x6	col_hist	-0.2000	0.154	-1.301	0.193	-0.501	0.101
x7	isl_o	0.2064	0.149	1.387	0.165	-0.085	0.498
x8	isl_d	0.1646	0.102	1.611	0.107	-0.036	0.365
x9	lndl_o	-0.8997	0.124	-7.249	0.000	-1.143	-0.656
x10	lndl_d	-0.9404	0.118	-7.957	0.000	-1.172	-0.709



results for traditional gravity from Santas Silva & Tenreyro (2006)

TABLE 3.—THE TRADITIONAL GRAVITY EQUATION

Estimator: Dependent Variable:	OLS $\ln(T_{ij})$	OLS $\ln(1 + T_{ij})$	Tobit $\ln(a + T_{ij})$	NLS T_{ij}	PPML $T_{ij} > 0$	PPML T_{ij}
Log exporter's GDP	0.938**	1.128**	1.058**	0.738**	0.721**	0.733**
	(0.012)	(0.011)	(0.012)	(0.038)	(0.027)	(0.027)
Log importer's GDP	0.798**	0.866**	0.847**	0.862**	0.732**	0.741**
	(0.012)	(0.012)	(0.011)	(0.041)	(0.028)	(0.027)
Log exporter's GDP per capita	0.207**	0.277**	0.227**	0.396**	0.154**	0.157**
	(0.017)	(0.018)	(0.015)	(0.116)	(0.053)	(0.053)
Log importer's GDP per capita	0.106**	0.217**	0.178**	−0.033	0.133**	0.135**
	(0.018)	(0.018)	(0.015)	(0.062)	(0.044)	(0.045)
Log distance	−1.166**	−1.151**	−1.160**	−0.924**	−0.776**	−0.784**
	(0.034)	(0.040)	(0.034)	(0.072)	(0.055)	(0.055)
Contiguity dummy	0.314*	−0.241	−0.225	−0.081	0.202	0.193
	(0.127)	(0.201)	(0.152)	(0.100)	(0.105)	(0.104)
Common-language dummy	0.678**	0.742**	0.759**	0.689**	0.752**	0.746**
	(0.067)	(0.067)	(0.060)	(0.085)	(0.134)	(0.135)
Colonial-tie dummy	0.397**	0.392**	0.416**	0.036	0.019	0.024
	(0.070)	(0.070)	(0.063)	(0.125)	(0.150)	(0.150)
Landlocked-exporter dummy	−0.062	0.106*	−0.038	−1.367**	−0.873**	−0.864**
	(0.062)	(0.054)	(0.052)	(0.202)	(0.157)	(0.157)
Landlocked-importer dummy	−0.665**	−0.278**	−0.479**	−0.471**	−0.704**	−0.697**
	(0.060)	(0.055)	(0.051)	(0.184)	(0.141)	(0.141)
Exporter's remoteness	0.467**	0.526**	0.563**	1.188**	0.647**	0.660**
	(0.079)	(0.087)	(0.068)	(0.182)	(0.135)	(0.134)
Importer's remoteness	−0.205*	−0.109	−0.032	1.010**	0.549**	0.561**
	(0.085)	(0.091)	(0.073)	(0.154)	(0.120)	(0.118)
Free-trade agreement dummy	0.491**	1.289**	0.729**	0.443**	0.179*	0.181*
	(0.097)	(0.124)	(0.103)	(0.109)	(0.090)	(0.088)
Openness	−0.170**	0.739**	0.310**	0.928**	−0.139	−0.107
	(0.053)	(0.050)	(0.045)	(0.191)	(0.133)	(0.131)
Observations	9613	18360	18360	18360	9613	18360
RESET test p-values	0.000	0.000	0.204	0.000	0.941	0.331

summary and preliminary results

various machine learning techniques based on decision trees were used for a very simple traditional gravity analysis with **untransformed data**:

```
flow ~ distw, gdp_o, gdp_d, contig, comlang_off, col_hist, isl_o, isl_d, lndl_o, lndl_d
```

simple decision tree, random forest, gradient boosting, stochastic gradient boosting

PPML (best practice for gravity to date) for comparison
(continuous explanatory variables log-transformed)

the **mean of random validation (100x)**
out-of-sample performance (test R^2) indicates:
gradient boosting tree ensembles perform better than all other

PPML more robust

Thank you for the attention!

data and code:

https://github.com/adriennebohlmann/DecisionTrees_on_Gravity

Feedback very welcome!

References

Egger, P., Larch, M., Staub, K., and Winkelmann, R. (2011) The Trade Effects of Endogenous Preferential Trade Agreements. American Economic Journal: Economic Policy 3[3], 113-143.

Gopinath, Munisamy; Batarseh, Feras A.; Beckman, Jayson (2020): Machine Learning in Gravity Models: An Application to Agricultural Trade. NBER WORKING PAPER 27151. DOI: 10.3386/w27151

Head, Keith; Mayer, Thierry (2014): Gravity Equations: Workhorse,Toolkit, and Cookbook. Handbook of International Economics, Vol. 4, pp. 131–191. DOI: 10.1016/B978-0-444-54314-1.00003-3

Helpman, E., Melitz, M., and Rubinstein, Y. (2008) Trading Partners and Trading Volumes. Quarterly Journal of Economics 123[2], 441-487.

Jun, Jingwen; Suo, Yuan; Park, Seeha; Xu, Tianze; Liu, Yizhu; Wang, Weiqi (2018): Analysis of Bilateral Trade Flow and Machine Learning Algorithms for GDP Forecasting. Engineering, Technology & Applied Science Research, Vol. 8(5): 3432–3438. DOI: 10.48084/etasr.2311

Mullainathan, Sendhil; Spiess, Jann (2017): Machine Learning: An Applied Econometric Approach. Journal of Economic Perspectives. Vol. 31(2): 87–106. DOI: 10.1257/jep.31.2.87

Santos Silva, J. M. C. and Tenreyro, S. (2006) The Log of Gravity. The Review of Economics and Statistics 88[4], 641-658.

Yotov, Yoto; Piermartini, Roberta; Monteiro, José-Antonio; Larch, Mario (2016): An Advanced Guide to Trade Policy Analysis: The Structural Gravity Model. UNCTAD/WTO.

References

Island Dummies
https://en.wikipedia.org/wiki/List_of_island_countries

Landlocked Dummies
https://en.wikipedia.org/wiki/Landlocked_country

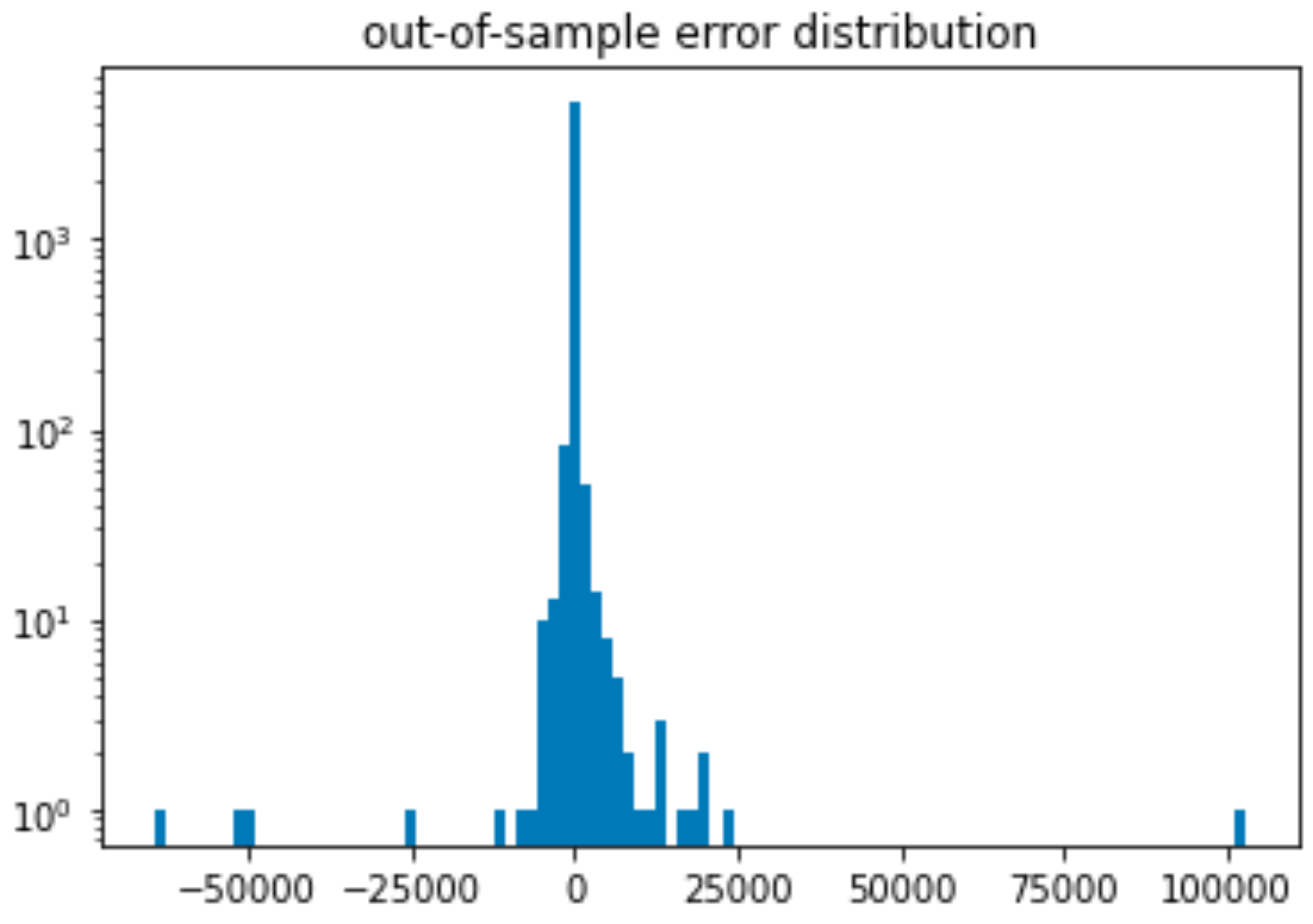
„The Log of Gravity“ website with very helpful information:
https://personal.lse.ac.uk/tenreyro/LGW.html?

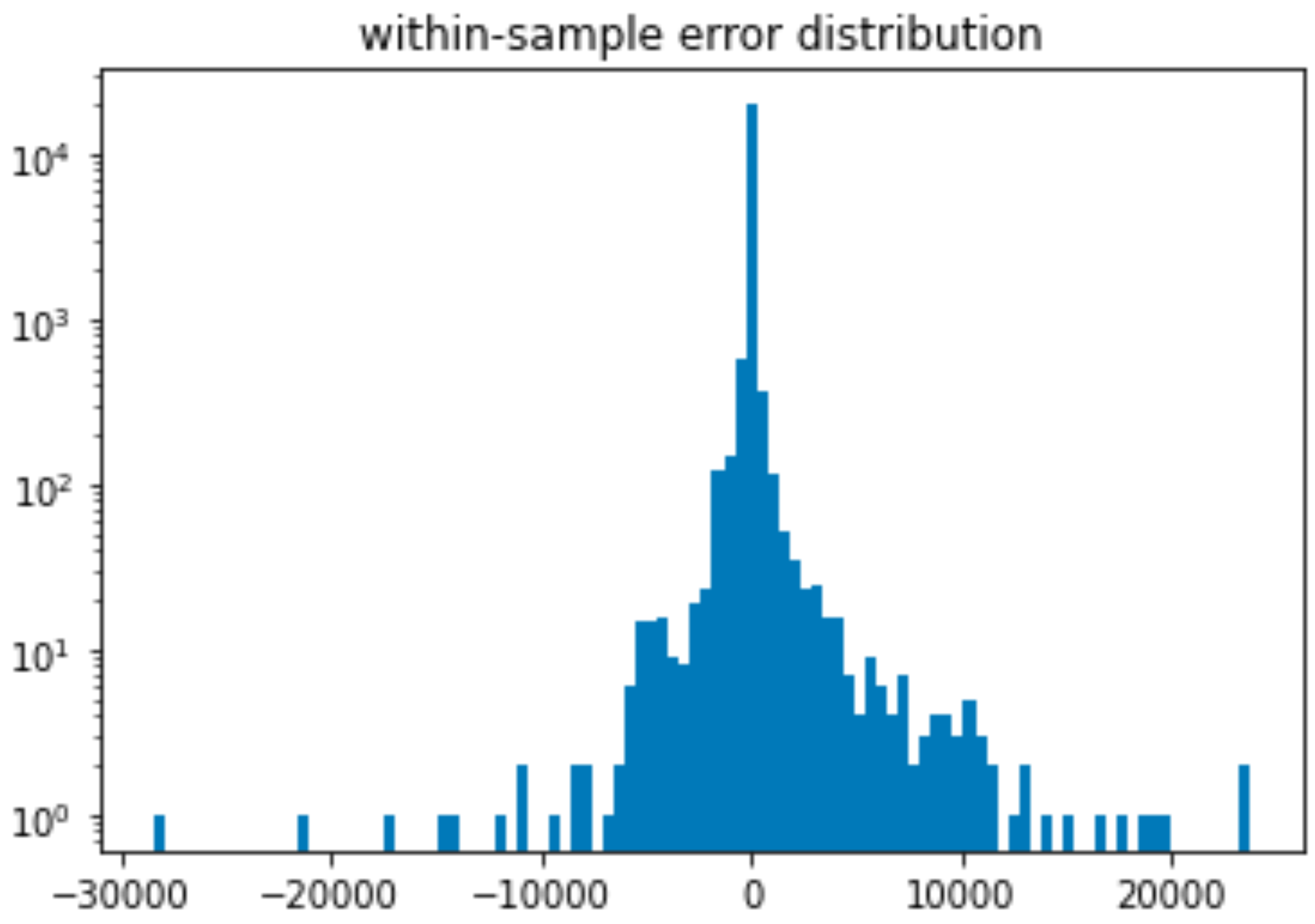
appendix: a glance at the variation of R^2

	random forest	gradient boosting	SGD boosting	PPML
R2 test	**0.580**	**0.455**	**0.664**	**0.513**
R2 train	0.951	0.949	0.996	0.721
R2 test	**0.754**	**0.828**	**0.632**	**0.496**
R2 train	0.941	0.967	0.996	0.688
R2 test	**0.629**	**0.775**	**0.419**	**0.463**
R2 train	0.949	0.960	0.996	0.494
R2 test	**0.561**	**0.741**	**0.710**	**0.519**
R2 train	0.959	0.961	0.996	0.654
R2 test	**0.636**	**0.644**	**0.553**	**0.470**
R2 train	0.946	0.972	0.995	0.613
R2 test	**0.747**	**0.261**	**0.626**	**0.457**
R2 train	0.956	0.972	0.996	0.709
R2 test	**0.736**	**0.583**	**0.769**	**0.497**
R2 train	0.950	0.971	0.995	0.501
R2 test	**0.524**	**0.596**	**0.406**	**0.511**
R2 train	0.944	0.974	0.994	0.612

appendix

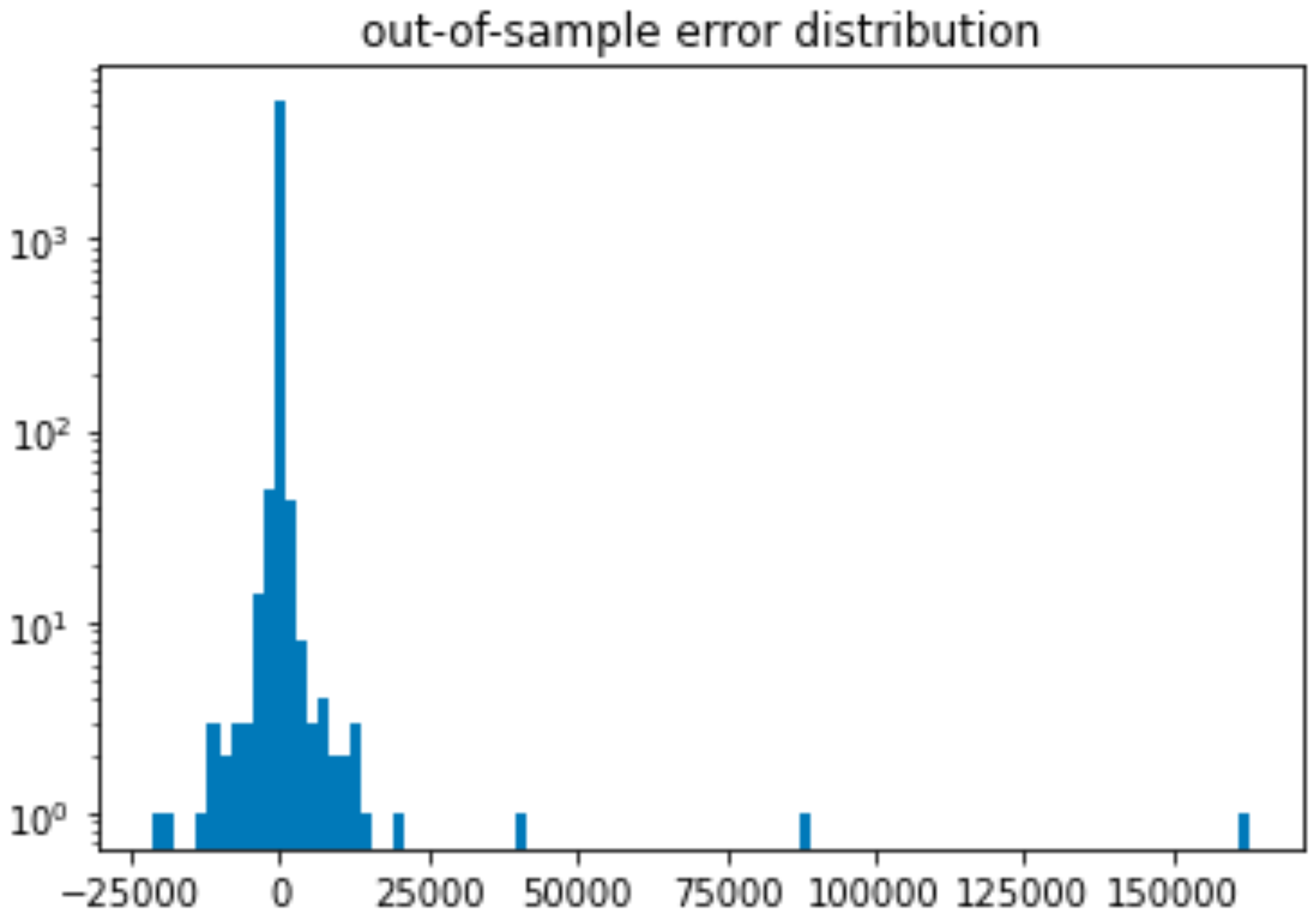
errors from single decision tree

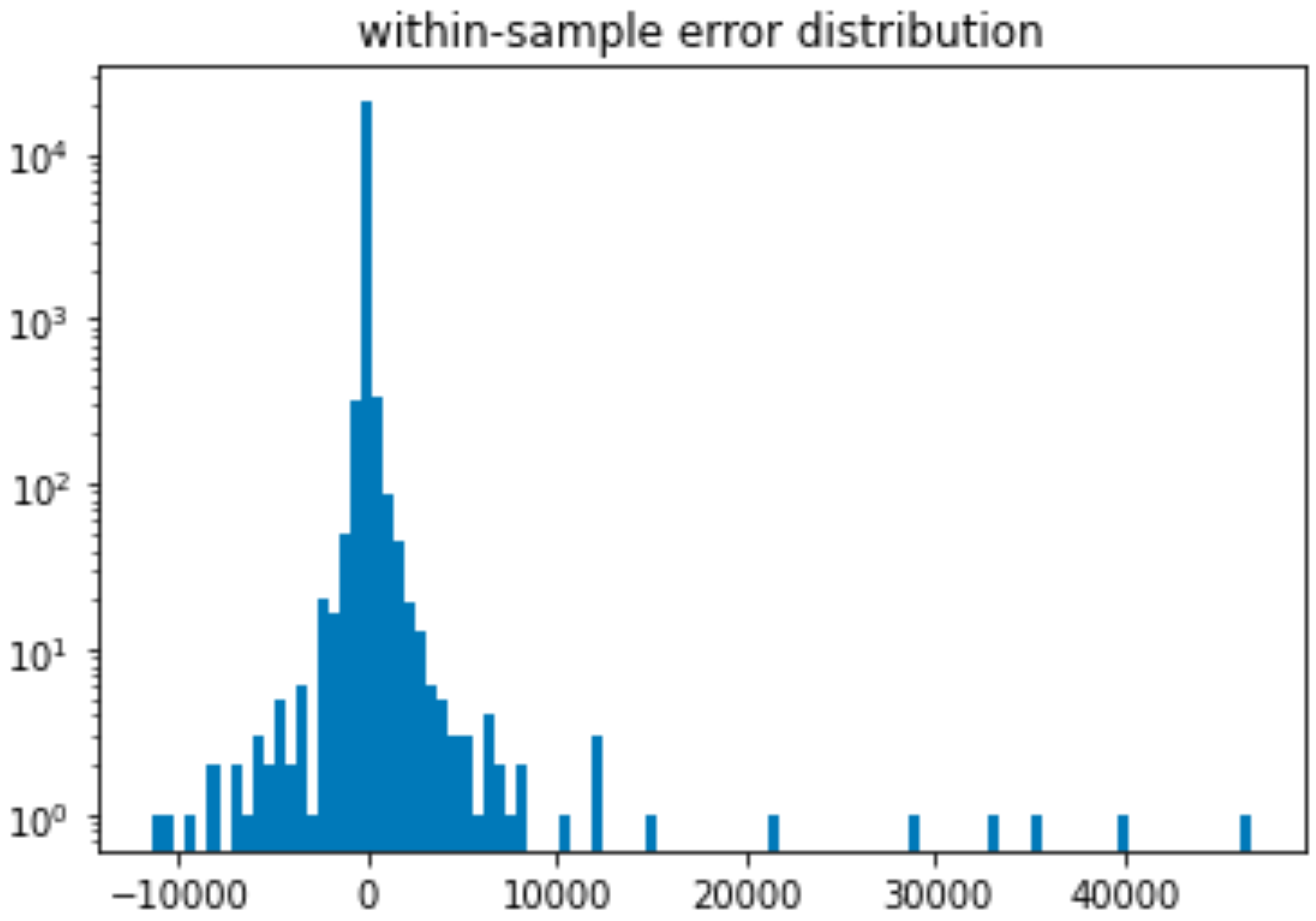




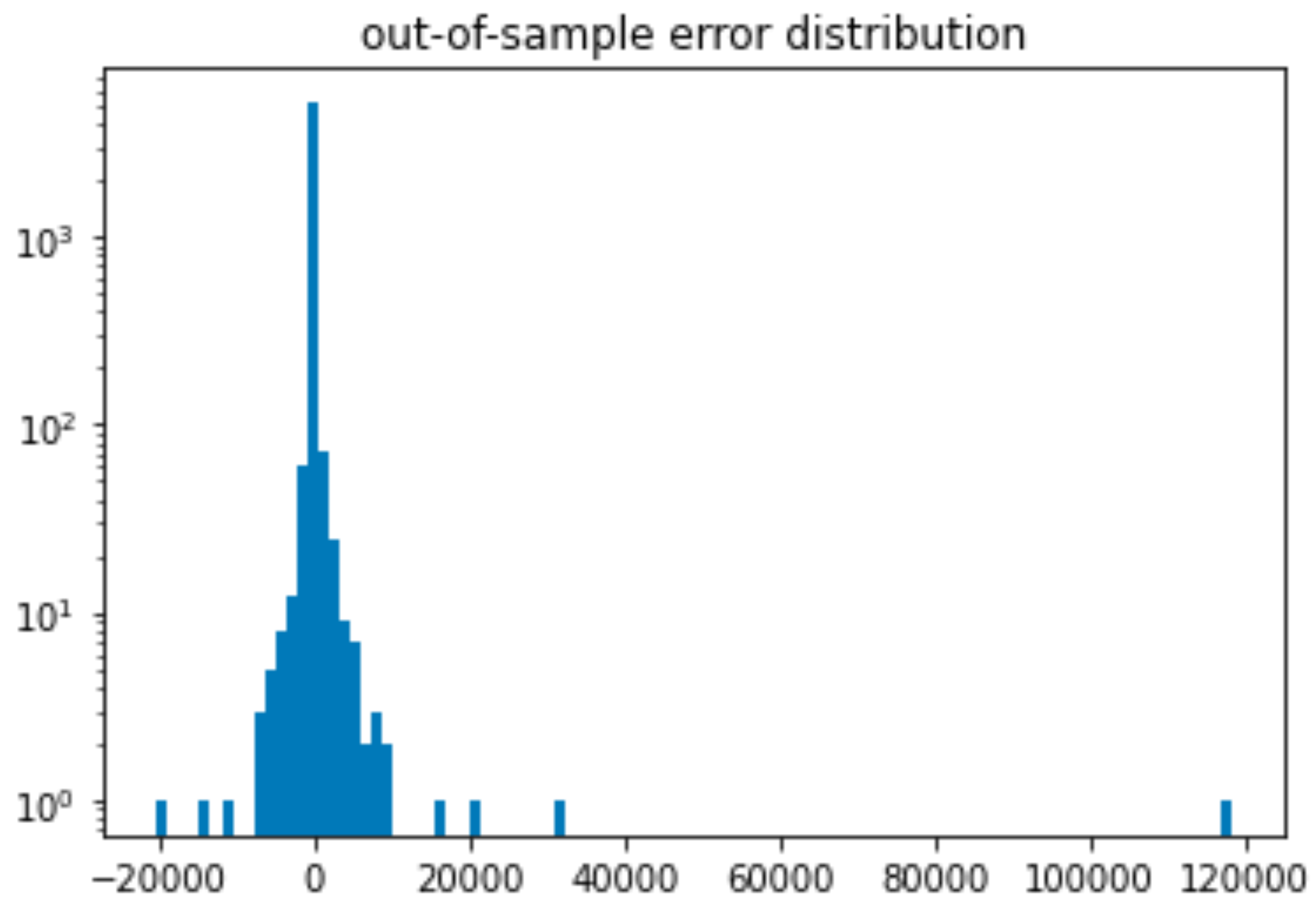
appendix

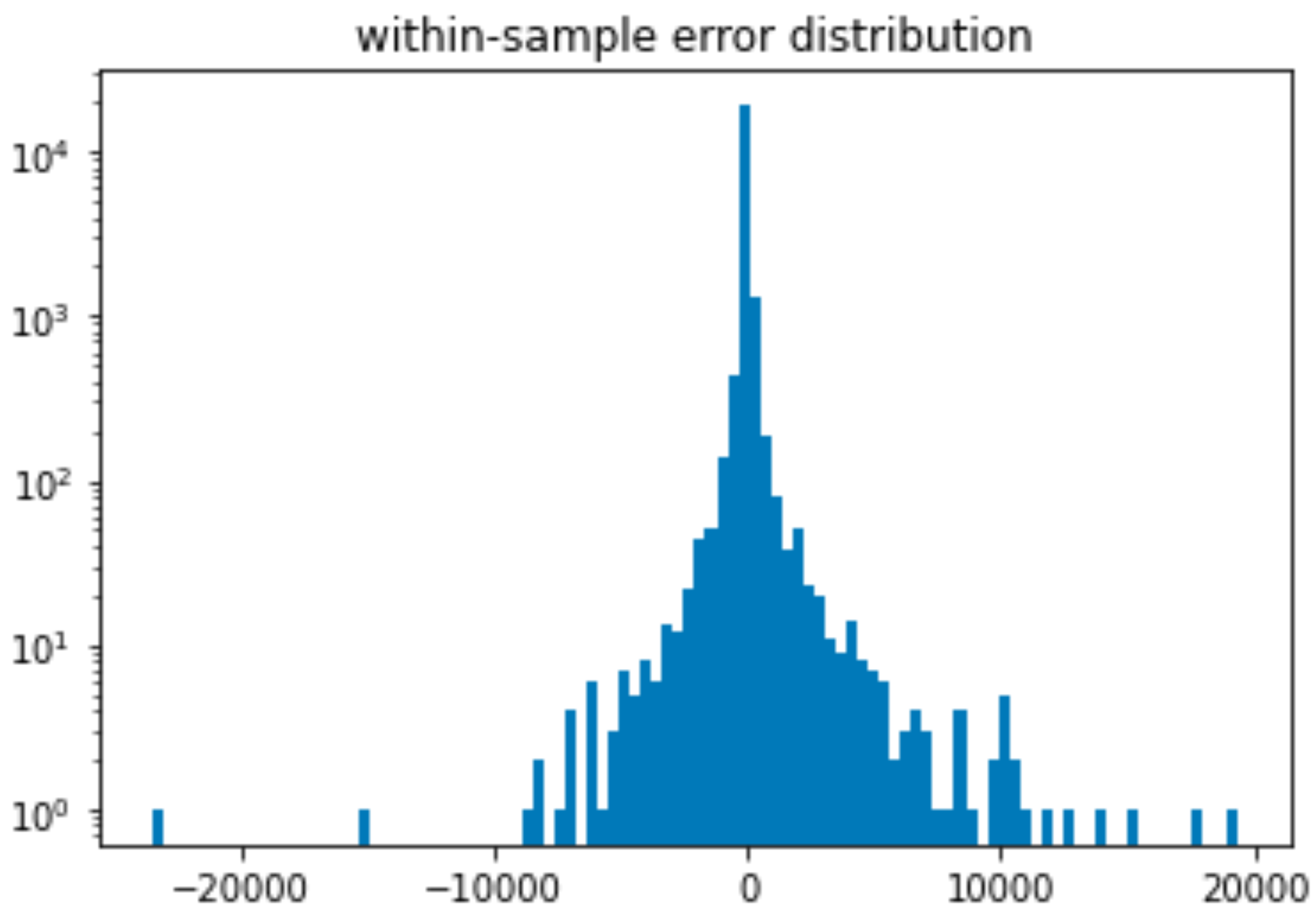
errors from single optimized random forest





errors from single optimized gradient boosting





appendix

errors from single PPML