Identification of scams in Initial Coin Offerings with machine learning

Bedil Karimov and Piotr Wójcik

CONSOB – Sapienza Seminar on ICOs
ICOs, Blockchain e DLT, 15th July 2022

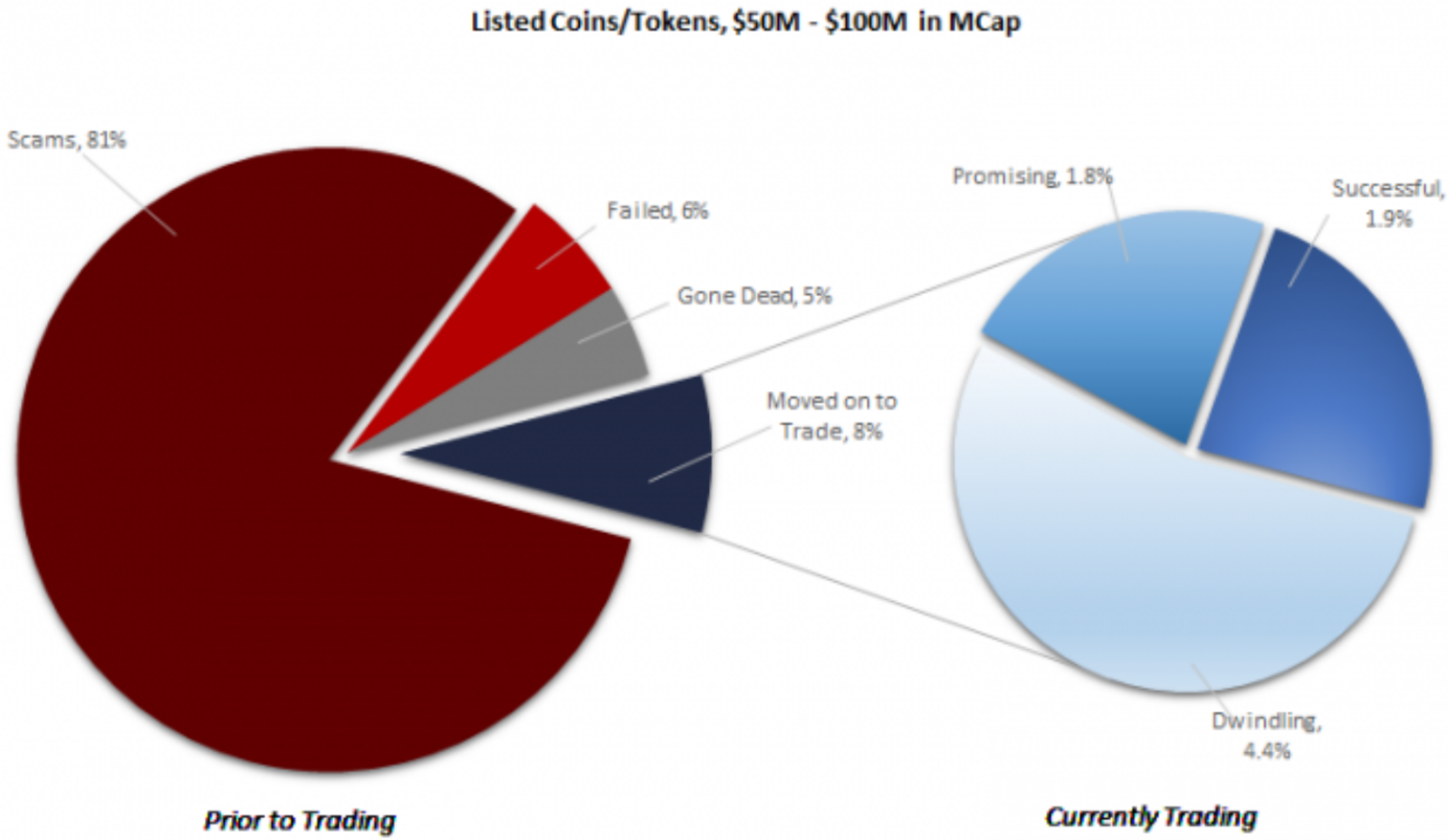
University of Warsaw
Faculty of Economic Sciences

Objective of research

- the **main objective** of the research is the **classification of ICOs** based on a wide range of features **known ex-ante**
- we assume that due to **non-linear relationships** between success factor and its predictors, using linear models might lead to **incorrect conclusions**
- non-linear models may **better reflect true relationships**, but **the shape on relationship is not known in advance**
- that is why we use selected **machine learning tools** that can **flexibly adjust** to data and **uncover unknown real relationships**
- published in Frontiers of Artificial Intelligence in 2021, https://doi.org/10.3389/frai.2021.718450

Motivation of research – popularity of crypto assets (e.g. ICOs, NFTs, DeFi)

Motivation of research – increase in scams (Dowlat, Hodapp 2018)



Novelty of our approach

- **Wider definition of a scam** and **updated information** not considered before
  - Definition of scam in Fahlenbrach and Frattarolli (2020): $1 million
  - Our updated definition: still active
- Use of **flexible machine learning** algorithms discovering complex nonlinear relationships and predict scams with greater accuracy
- The **application of XAI** tools to unhide the **black-box models**

Research hypotheses

1. the characteristics of ICOs **known ex-ante** help to predict that an ICO is a scam.
2. **nonlinear machine algorithms are more accurate** than traditional logistic regression and its regularized extensions (may capture potentially existing highly nonlinear relationships)
3. **Importance of technological background** of ventures – e.g. code availability, white paper availability, use of blockchain and decentralization

Methodology – research framework

- **Feature engineering** – omitting non-informative and correlated variables, transformation on training data to avoid information leakage
- **Final dataset**:
  - 53 categorical and 23 numerical features (labeled as **"all"**)
  - Selected variables after filtering based on realtionship with success factor: 18 categorical features (with Cramer's $V > 0.1$) and 10 numeric predictors (statistically significant at 10% level in one-way ANOVA) labeled as **"selected"**
- **Performance measures**: area under ROC curve, accuracy, sensitivity (recall), specificity, precision, F1 and balanced accuracy (average of sensitivity and specificity)

Methodology – research framework

- **Models** used: logistic regression, LASSO, ridge, SVM (linear & polynomial kernels), random forest, XGBoost, Catboost, LightGBM
- **Training and test split**: 70% and 30%
- Estimation on two sets of variables (**all** and **selected**)
- **Tuning of hyperparameters** with respect to **F1**
- Application of Explainable Artificial Intelligence (**XAI**)
- Model agnostic methods:
  - Permutation-based feature importance (**PFI**)
  - Partial dependence profiles (**PDP**)

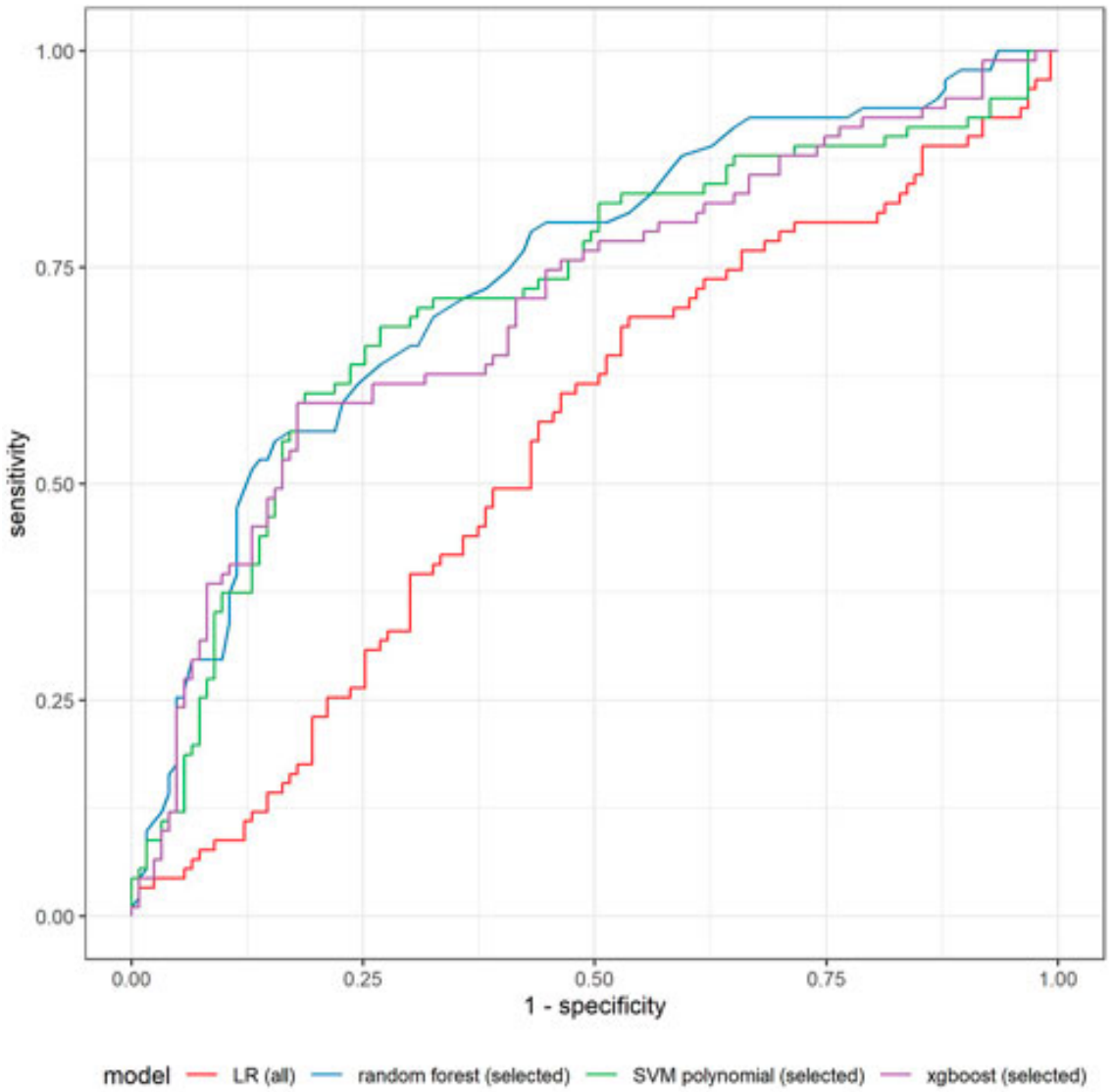
Data

- The same data that is used in paper "ICO investors" by Fahlenbrach and Frattarolli (2020)
- **305 completed ICOs with 136 features**, that took place between January 2016 and March 2018
- The information about ICOs were collected from **multiple sources**: websites, white papers, social media and Github pages with over 4700 contributors (Fahlenbrach, Frattaroli., 2020)
- icorating.com, smithandcrown.com, icowatchlist.com and coinschedule.com

Empirical results – validation set

Model (variables set)	ROC	Accuracy	Sensitivity	Specificity	Precision	F1	Balanced accuracy
LR (all)	0.5592	0.5468	0.5859	0.4944	0.4711	0.4756	0.5402
LR (selected)	0.7044	0.6630	0.7071	0.6056	0.6251	0.6061	0.6563
LR + backward (all)	0.5914	0.5643	0.6006	0.5167	0.5086	0.4997	0.5587
LR + backward (selected)	0.6887	0.6394	0.6987	0.5600	0.5964	0.5705	0.6294
LASSO (all)	0.6805	0.6907	0.7955	0.5489	0.6693	0.5976	0.6722
LASSO (selected)	0.7270	0.6768	0.7308	0.6044	0.6401	0.6154	0.6676
ridge (all)	0.6501	0.6260	0.7147	0.5067	0.5767	0.5309	0.6107
ridge (selected)	0.7345	0.6911	0.7641	0.5933	0.6716	0.6221	0.6787
SVM linear (all)	0.6488	0.6108	0.6885	0.5056	0.5589	0.5246	0.5970
SVM linear (selected)	0.7204	0.7000	0.7795	0.5922	0.6730	0.6245	0.6859
SVM polynomial (all)	0.6488	0.6342	0.7045	0.5389	0.5981	0.5581	0.6217
SVM polynomial (selected)	0.7277	0.7190	0.8038	0.6044	0.7091	0.6443	0.7041
random forest (all)	0.7200	0.7050	0.8115	0.5611	0.7109	0.6212	0.6863
random forest (selected)	0.7458	0.7143	0.8276	0.5611	0.7238	0.6235	0.6943
xgboost (all)	0.6750	0.6775	0.7231	0.6178	0.6264	0.6158	0.6704
xgboost (selected)	0.7128	0.7093	0.7712	0.6256	0.6783	0.6467	0.6984
catboost (all)	0.7277	0.7195	0.8365	0.5600	0.7194	0.6243	0.6983
catboost (selected)	0.7335	0.7015	0.7808	0.5933	0.6652	0.6212	0.6871
lightgbm (all)	0.6788	0.6634	0.7333	0.5711	0.6065	0.5746	0.6522
lightgbm (selected)	0.7009	0.6686	0.7256	0.5911	0.6375	0.6036	0.6584

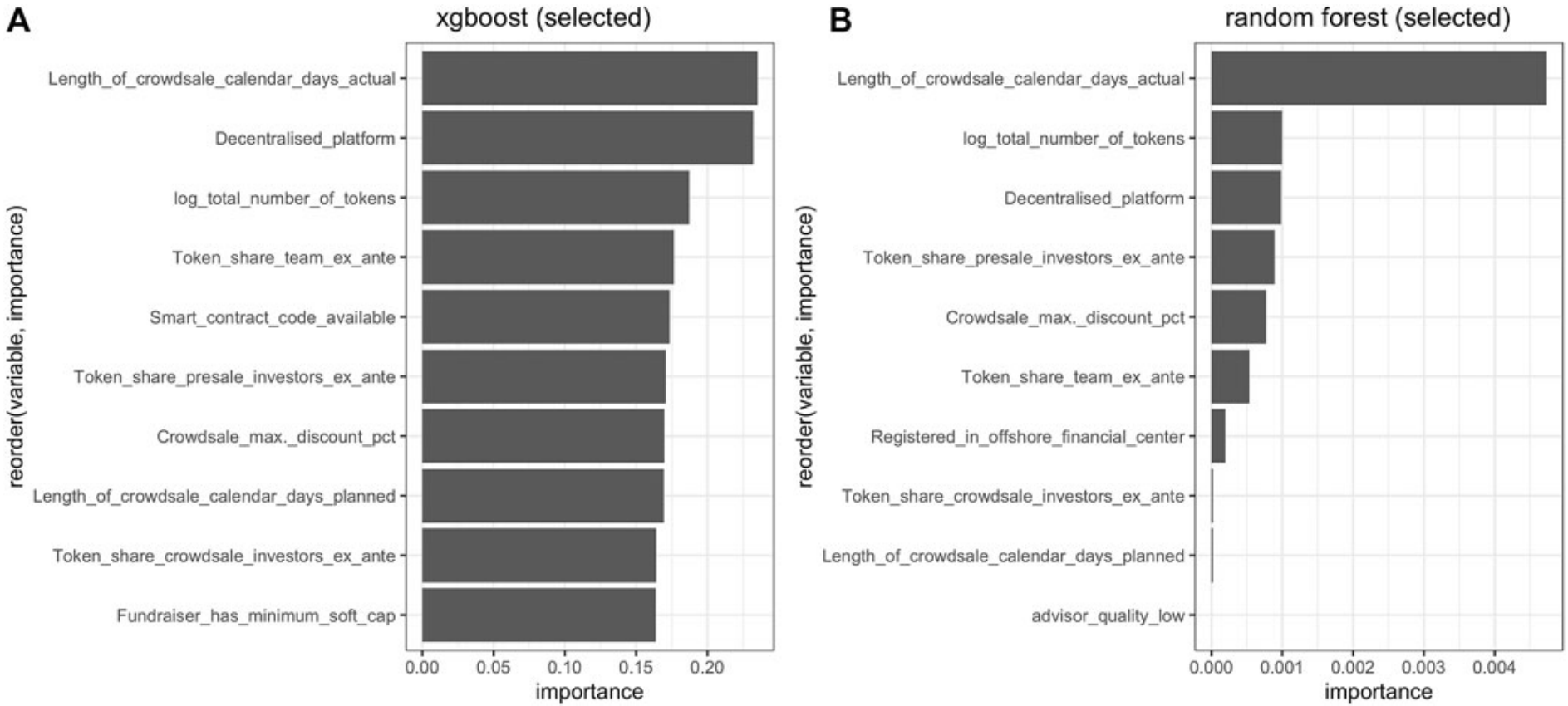
Empirical results – validation set ROC curve



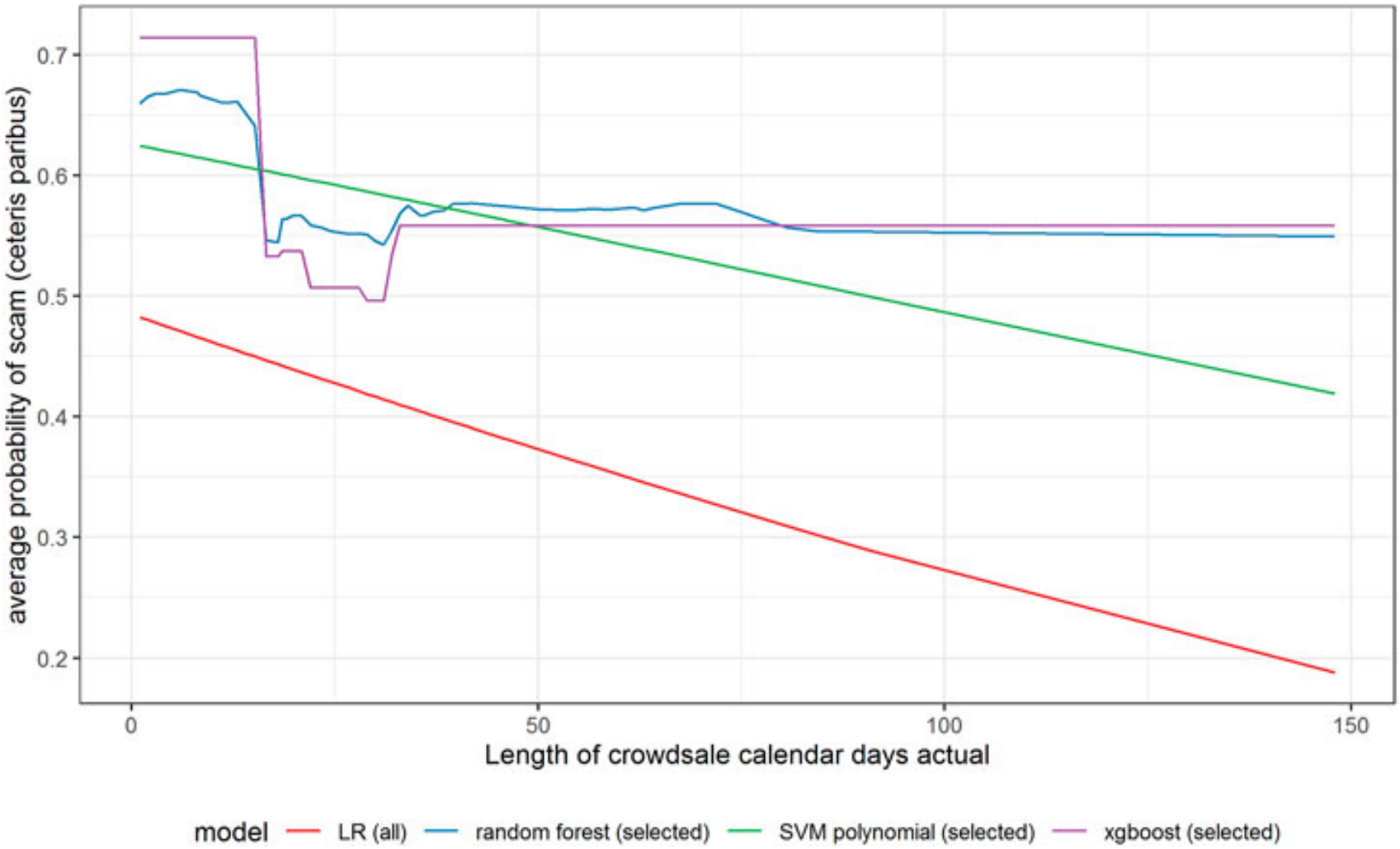
Empirical results – test set

Model (variables set)	ROC	Accuracy	Sensitivity	Specificity	Precision	F1	Balanced accuracy
LR (all)	0.6575	0.6196	0.6792	0.5385	0.5526	0.5455	0.6089
LR (selected)	0.6971	0.6304	0.6415	0.6154	0.5581	0.5854	0.6284
LR + backward (all)	0.7199	0.6304	0.6604	0.5897	0.5610	0.5750	0.6251
LR + backward (selected)	0.7068	0.6630	0.7170	0.5897	0.6053	0.5974	0.6534
LASSO (all)	0.7131	0.6196	0.7358	0.4615	0.5625	0.5070	0.5987
LASSO (selected)	0.7059	0.6739	0.7547	0.5641	0.6286	0.5946	0.6594
ridge (all)	0.7068	0.6196	0.7170	0.4872	0.5588	0.5205	0.6021
ridge (selected)	0.7010	0.6739	0.7547	0.5641	0.6286	0.5946	0.6594
SVM linear (all)	0.6991	0.6739	0.7547	0.5641	0.6286	0.5946	0.6594
SVM linear (selected)	0.6831	0.6413	0.6981	0.5641	0.5789	0.5714	0.6311
SVM polynomial (all)	0.6986	0.6739	0.7547	0.5641	0.6286	0.5946	0.6594
SVM polynomial (selected)	0.6918	0.6522	0.7358	0.5385	0.6000	0.5676	0.6372
random forest (all)	0.6986	0.6522	0.7170	0.5641	0.5946	0.5789	0.6405
random forest (selected)	0.6606	0.6196	0.6604	0.5641	0.5500	0.5570	0.6122
xgboost (all)	0.6652	0.6087	0.6038	0.6154	0.5333	0.5714	0.6096
xgboost (selected)	0.6478	0.6413	0.6792	0.5897	0.5750	0.5823	0.6345
catboost (all)	0.6497	0.6196	0.6792	0.5385	0.5526	0.5455	0.6089
catboost (selected)	0.6023	0.5543	0.6226	0.4615	0.4737	0.4675	0.5421
lightgbm (all)	0.6584	0.6413	0.6415	0.6410	0.5682	0.6024	0.6413
lightgbm (selected)	0.6512	0.6522	0.6604	0.6422	0.5814	0.6098	0.6507

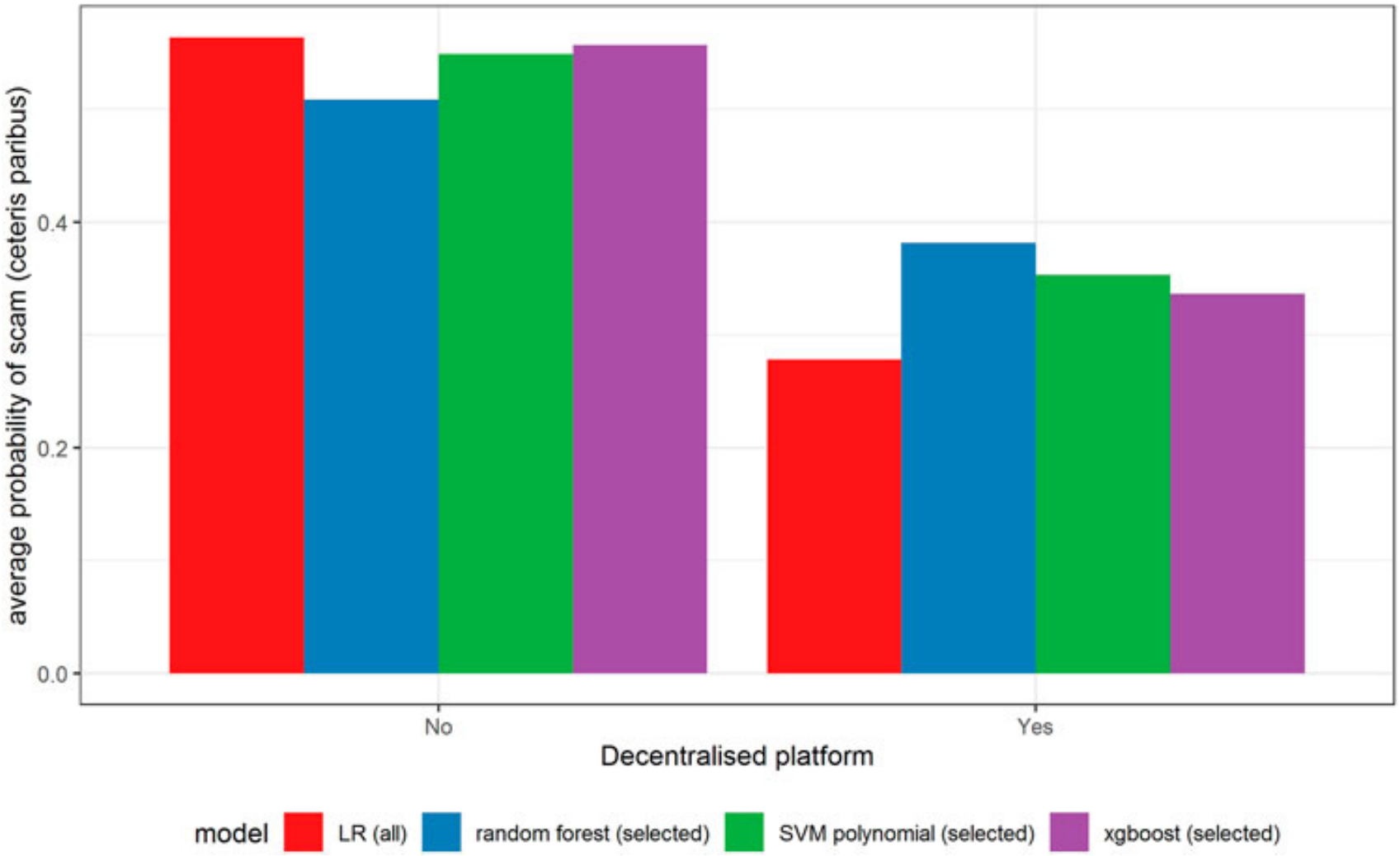
Variable importance



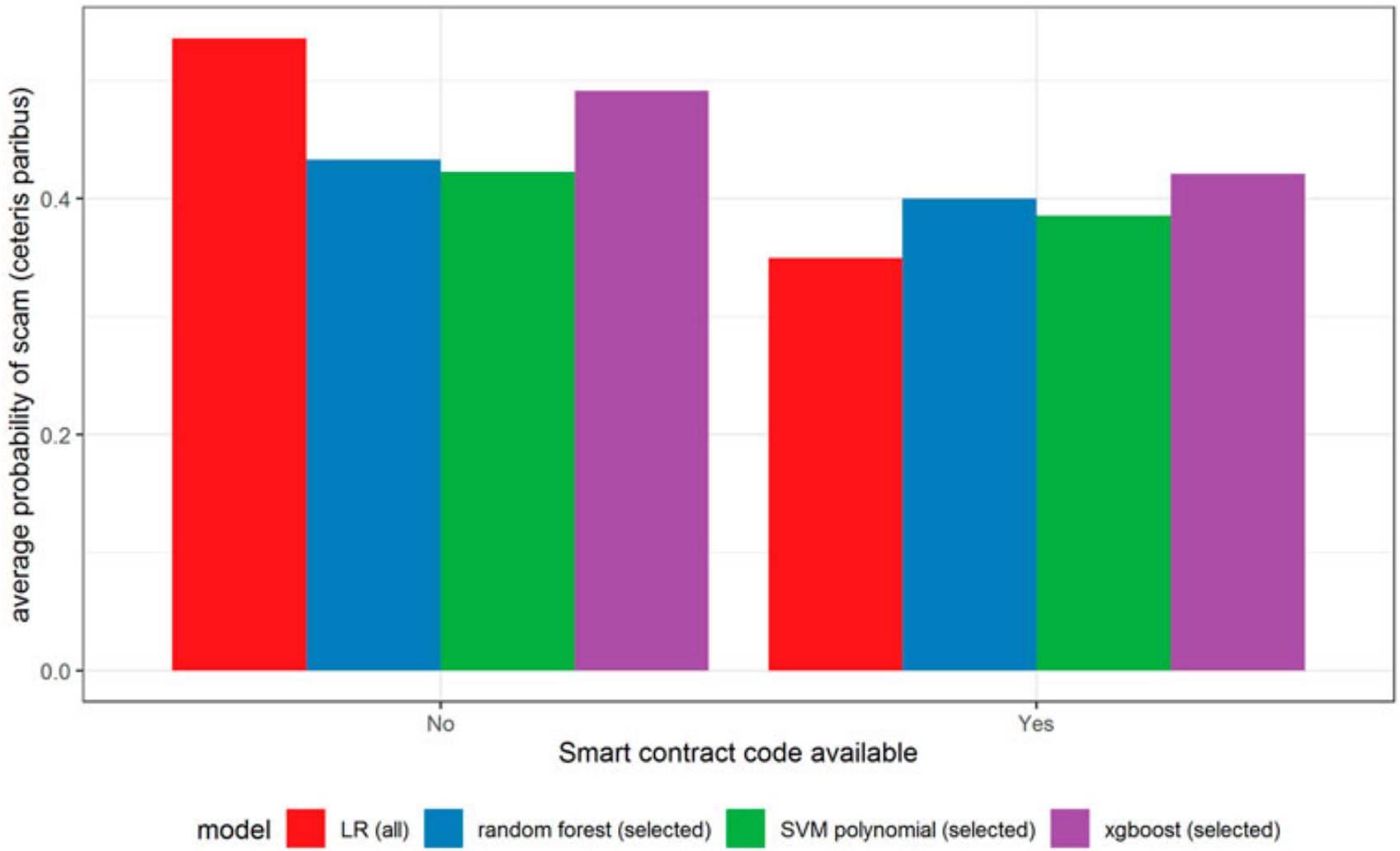
PDPs – actual crowdsale calendar days



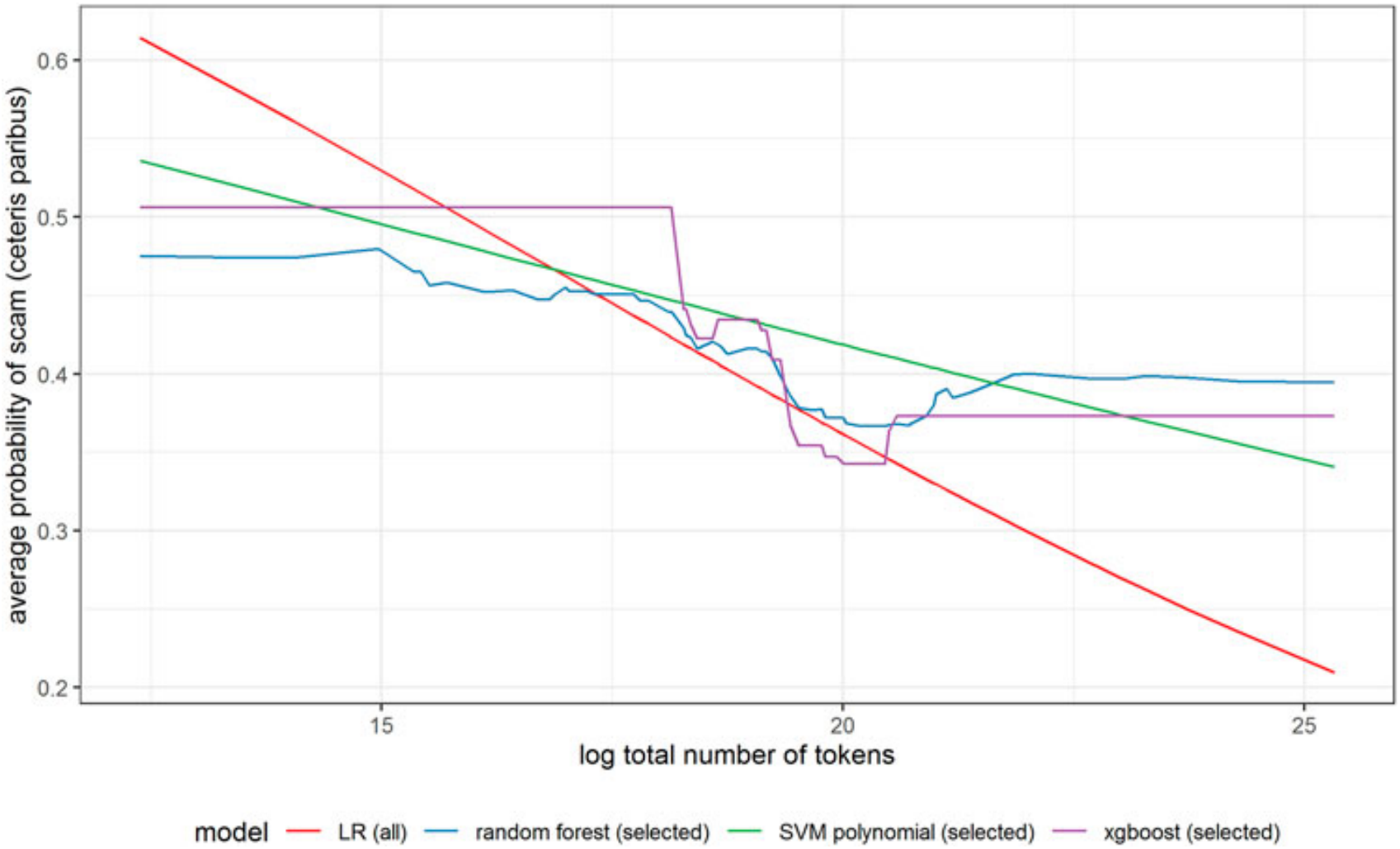
PDPs – decentralized platform



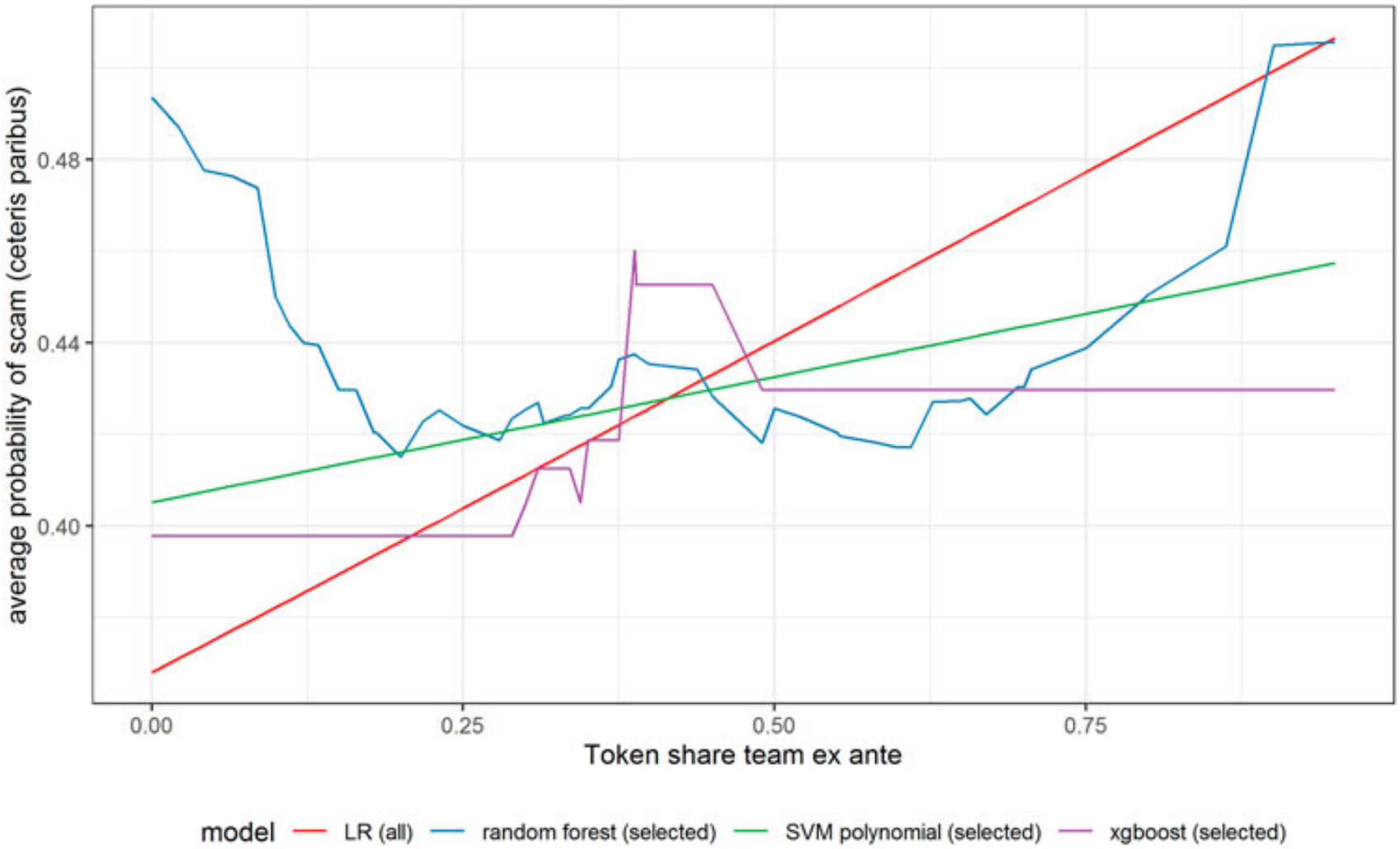
PDPs – availability of smart contract code



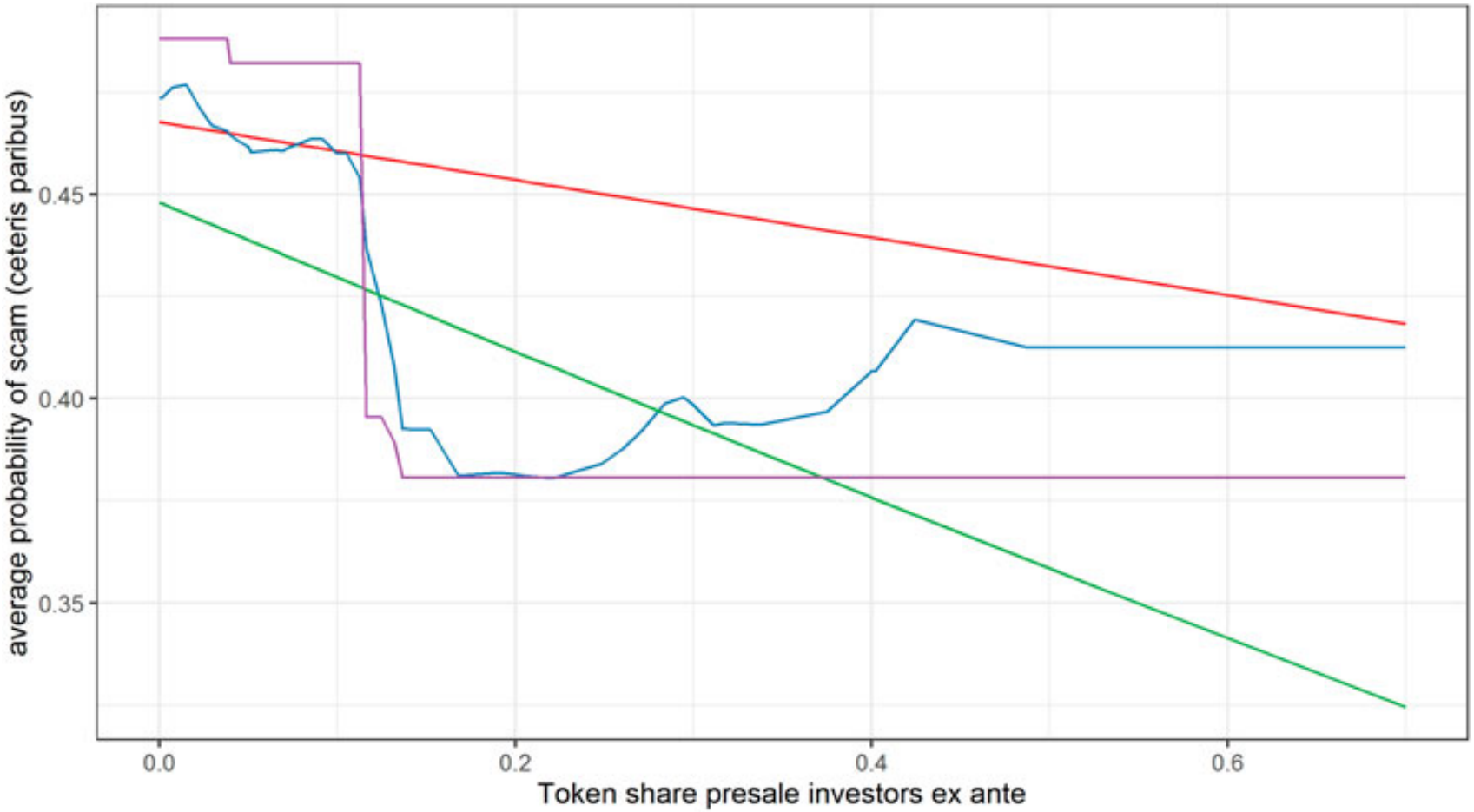
PDPs – total number of tokens (log)



PDPs – share of tokens for the team (ex-ante)



PDPs – share of tokens for the investors in presales (ex-ante)



Conclusions

- all three research hypotheses were confirmed
- ex-ante characteristics of ICOs allow us to distinguish between scams and non-scams with a relatively high probability
- we confirmed the **superiority of nonlinear machine learning models**
- however, this superiority was **mainly visible in the validation sample** and much weaker in the test data
- Last but not least, we notice the **importance of the technological capabilities** of these ventures, which are a crucial factor in the future success of a project
- although it was not obvious that all technical aspects are important, we uncovered a **positive relationship with the availability of smart contract code and being decentralized**
- the patterns revealed may **help investors to identify reliable ICO projects** and to make more rational decisions

Thank you !