Countering Public Grant Fraud in Spain

Machine Learning for Assessing Risks and Targeting Control Activities

Mihály Fazekas (Central European University and Government Transparency Institute)

Regional Workshop
Using Technology to Prevent and Combat Corruption
Amman, Jordan, 15-16 June 2022

OECD
BETTER POLICIES FOR BETTER LIVES

Goals of the presentation today

- Introduce you to how machine learning can be used in an investigative context

- Highlight some ways of extending the data-driven risk assessment model

Rationale of data driven risk assessment for fraud detection

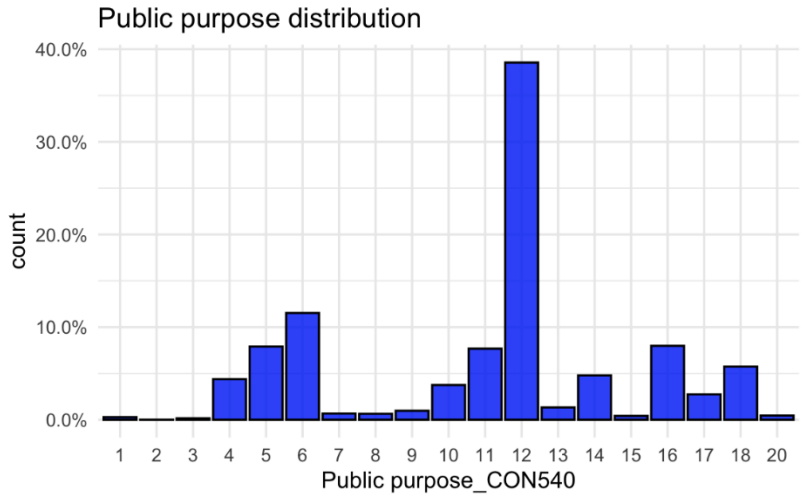
- There are hundreds of thousands of public grant awards each year
- Only a handfull of organisations and awards can be investigated
- As only few awards are fraudelent, a random/quota-based selection is unlikely to allocate scarce investigative resources efficiently

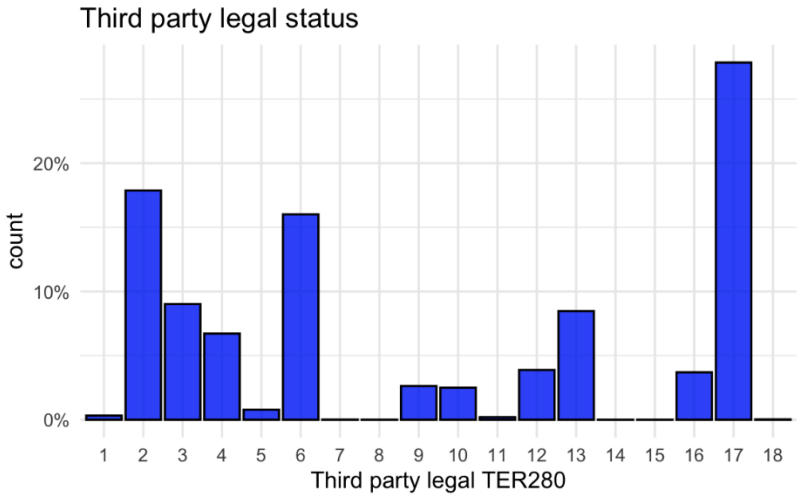
DATA & DATA PREPARATION

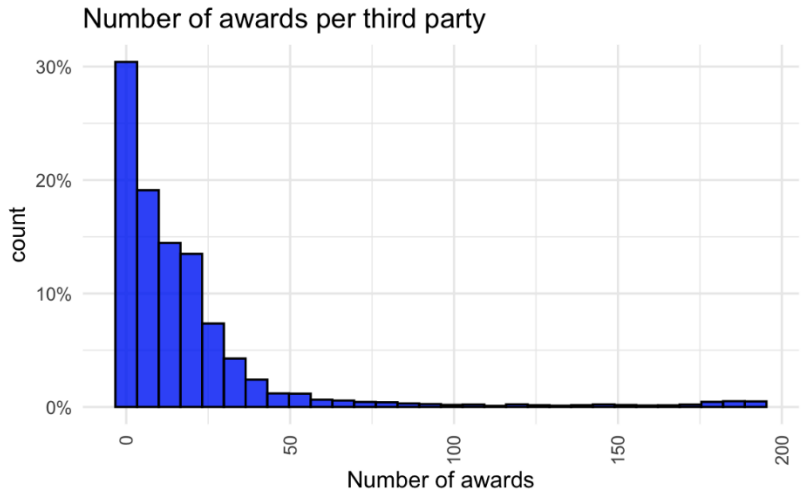
Data: final, merged dataset

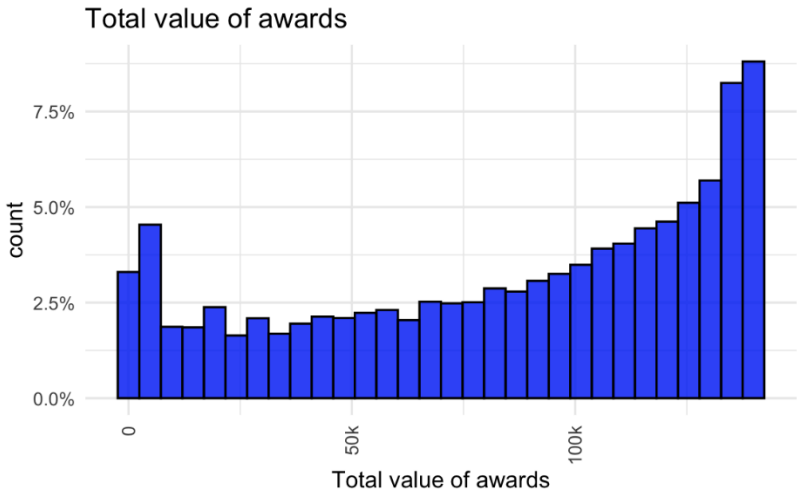
- 23 variables and 1,050,470 observations for years 2018 - 2020
- Sanctions dummy: marks if the third party was sanctioned for the corresponding award, as well as for all previous awards received by the same party
- Most of the variables are binary (regions, countries, types of grants and awards, types of third parties), 4 numeric (costs, payments), dates and IDs

Variables in the analysis: selected examples of distributions









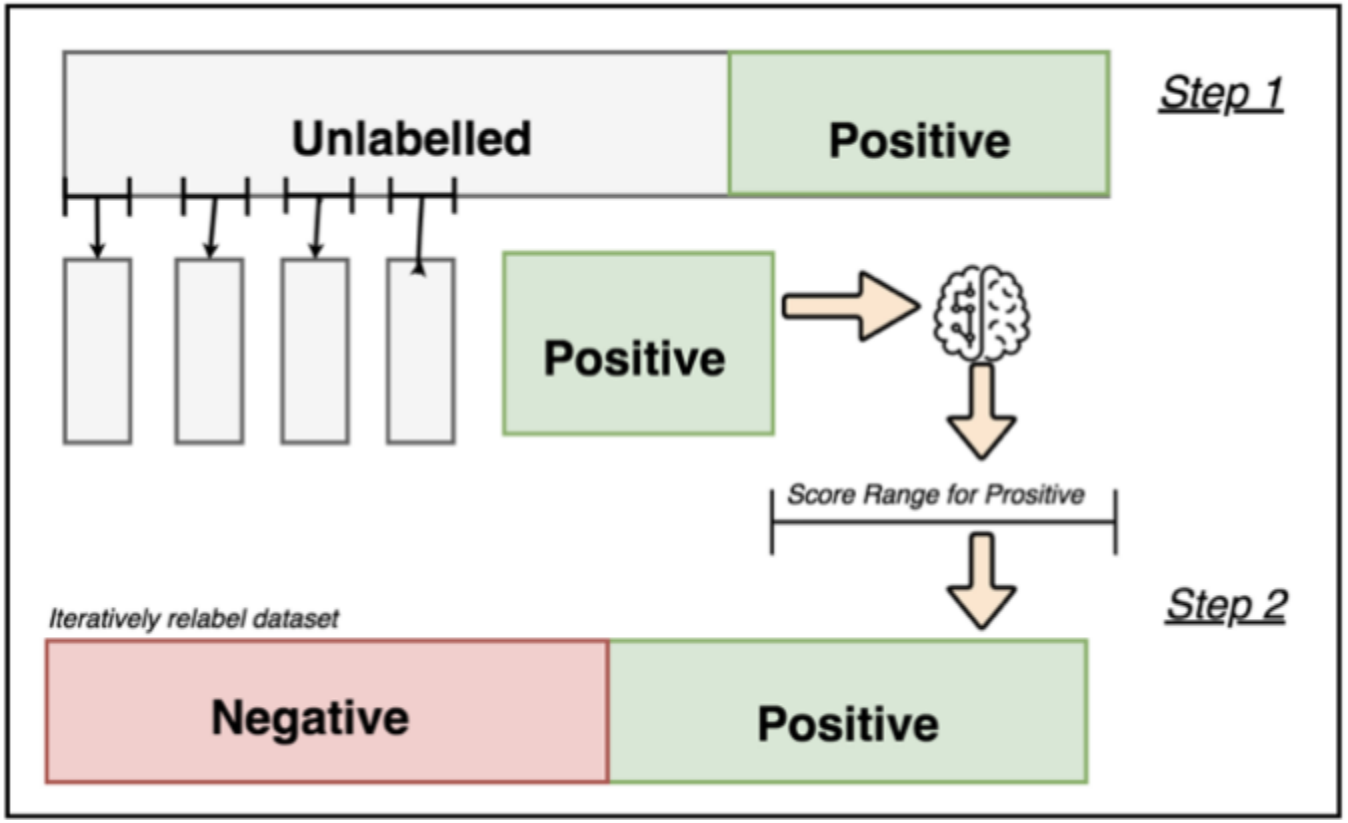
METHODOLOGY

Positive-Unlabelled learning I

- Problem: positive (sanctioned) cases are known, but negative cases are unclear
  - Some of the unsanctioned cases could have been sancitoned had they been investigated

Positive-Unlabelled learning II

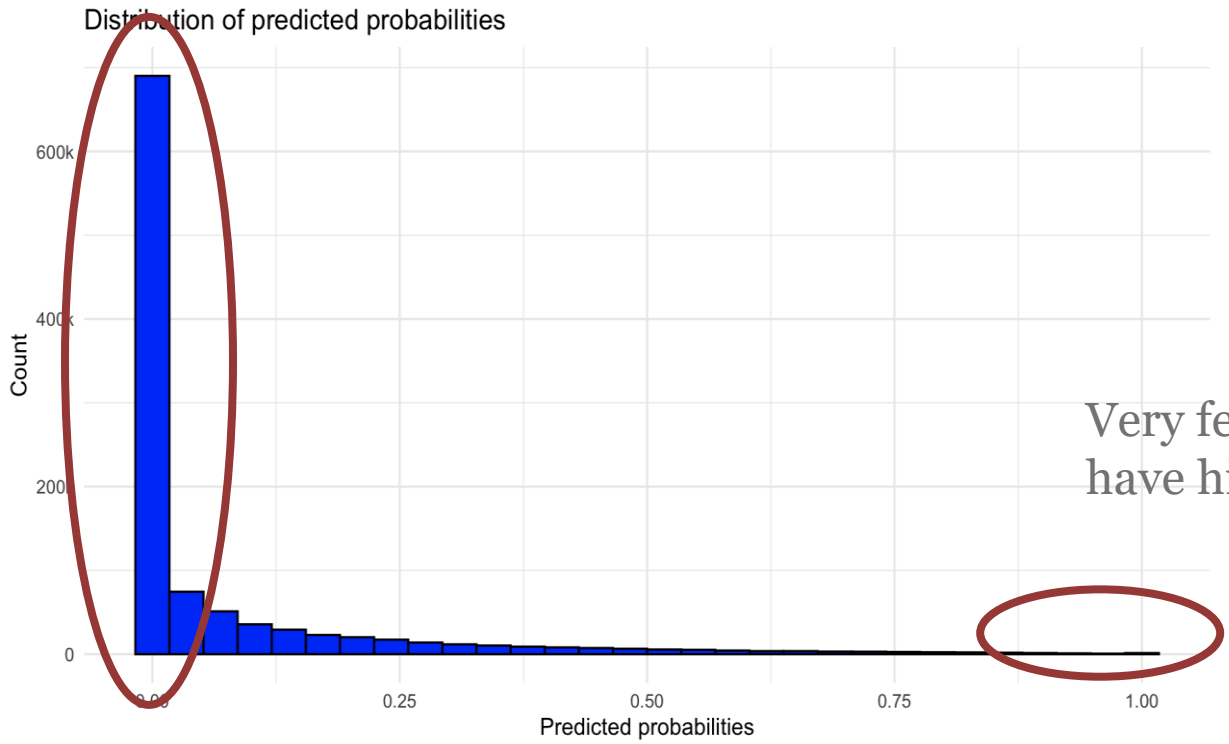
- Solution: Positive-Unlabelled learning using Random Forest (aka PU bagging)
  - Sequentially relabel unknown cases as negatives based on known risks



RESULTS

Distribution of predicted fraud risks: Tree-based Machine Learning approach

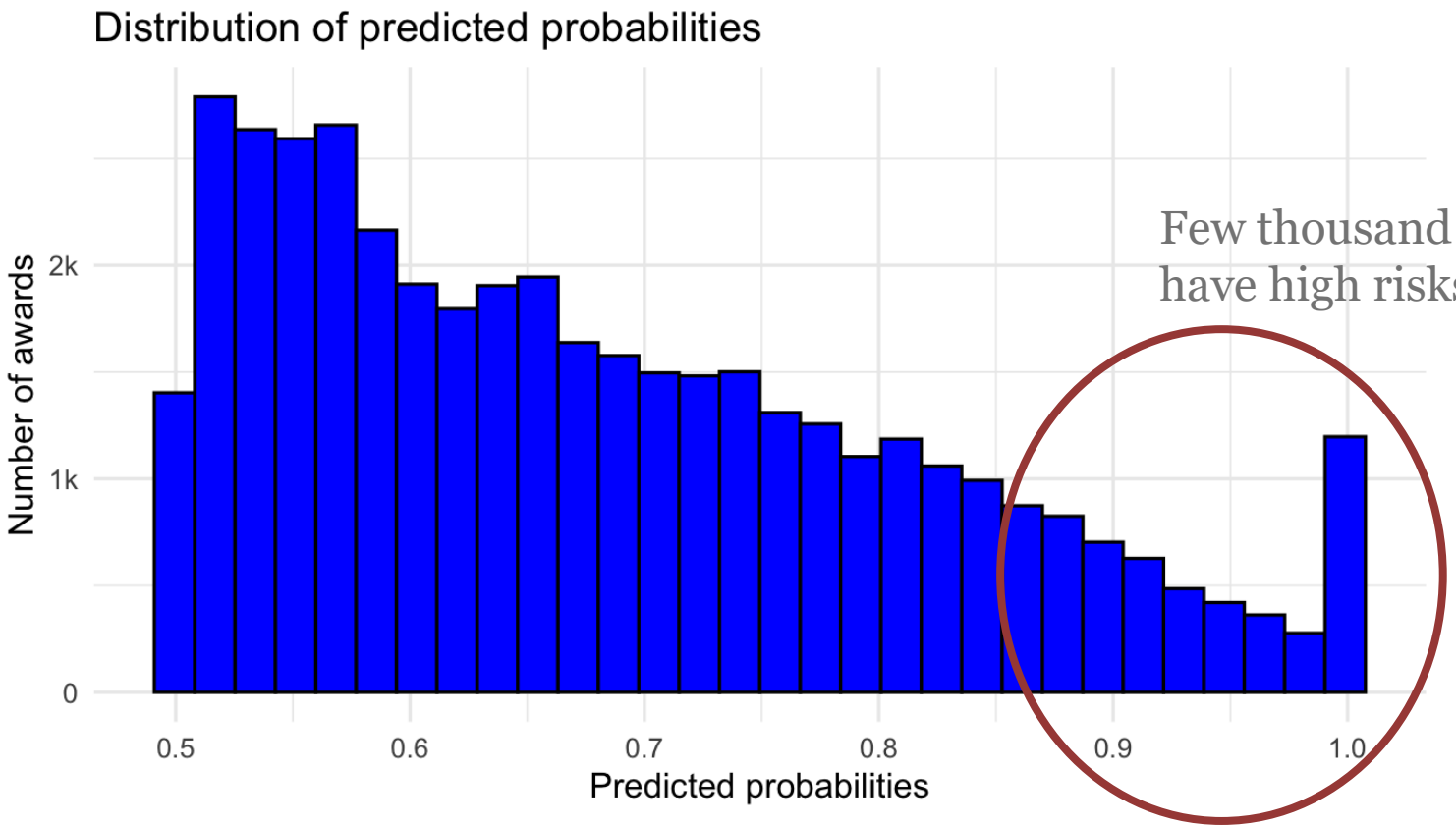
Most awards have virtually no risk



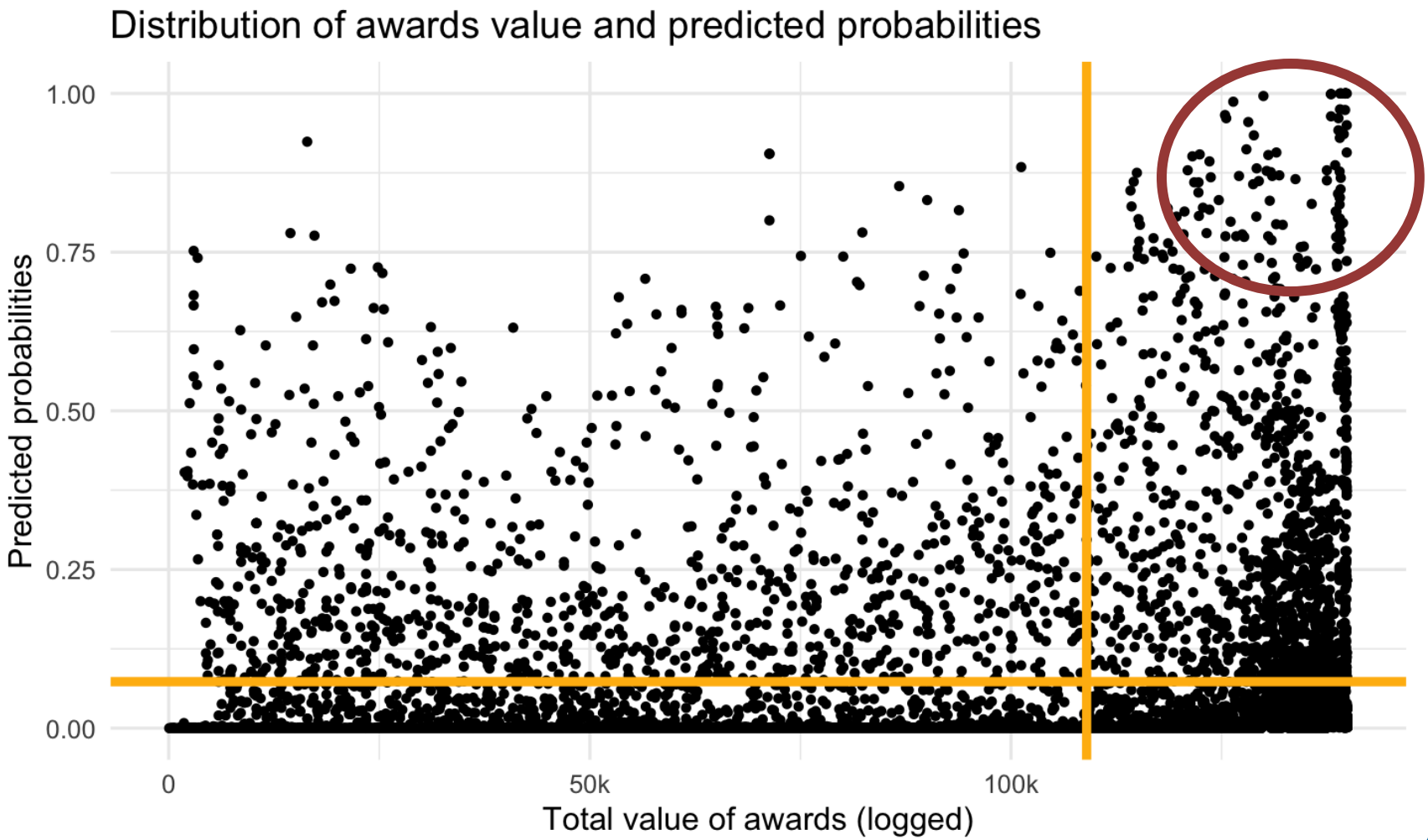
Very few awards have high risks

1,050,470 awards from 2018-2020 risk scored based on their similarity to observed sanctions

Distribution of predicted fraud risks: Zooming in on >50% risk



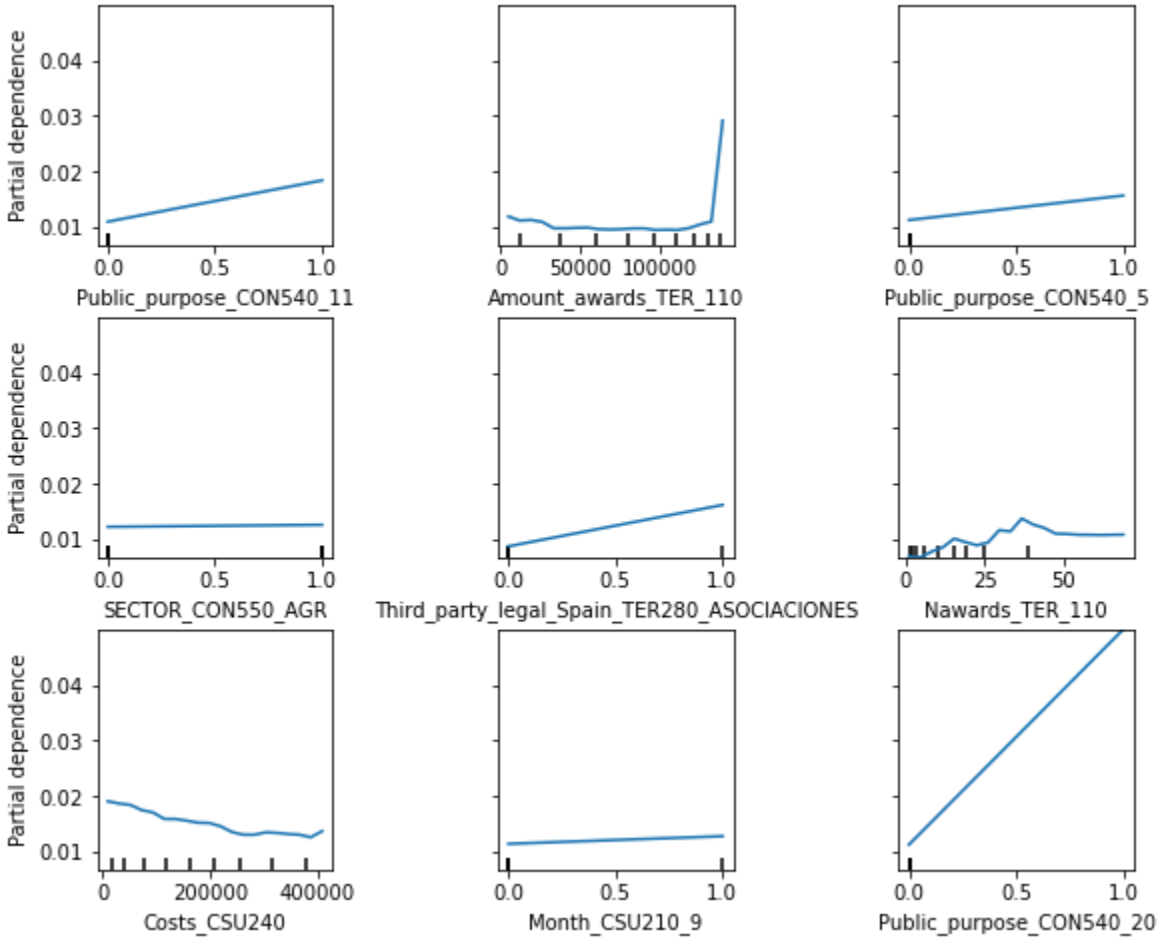
Potential uses of the results: combining risk and grant value



For a similiar approach see: European Investment Bank: Prior Integrity Reviews (OECD, 2019)

Results: Partial dependence plots
Most influential predictors

- Partial dependence plots



Most of these are proxies, rahter than directly pointing at wrongdoing

Strenghts and weaknesses

- Strengths
  - Efficient, well-tested methodology
  - Well-defined outcome (yes/no)
  - Precisely replicating past sanctions: 93% accuracy
- Weaknesses
  - Limited data
  - Past investigations may not have uncovered all major types of fraud

EXTENDING THE DATA & MODELING

Promissing data sources

Four groups of data:

- organisational data on the parties of the granting process
- connections and conflict of interest
- organisational reliability and violation of rules
- other funds and contracts

Dataset name	Dataset group	Unit of measurement	Number of observations	ID to match on to IGAE main dataset	Priority for IGAE follow-up work
National Company registry	i, ii	Organization	>5000000	NIF of beneficiaries, names of organizations	high
BO registry	i, ii	Organization	>5000000	NIF of beneficiaries	high
Database of Spanish senior positions and secretariats	ii	Institutions and State Bodies	~100000	Name of organizations	high
CINCO net	iii	Organizations	should be accessed by official body	NIF of organizations	high
Public procurement data	iv	Tender	1391558	NIF of organizations	high
Public Bankruptcy Registry	iii	Organizations	website does not allow to search	NIF of organizations	medium
Spanish Association of Foundations (AEF)	iv	Foundation	15840	Location and type of beneficiary	medium
State Tax Administration Agency-AEAT	iii	Organizations	not in public access	NIF of organizations	medium
European Union aid	iv	Grant or contract	40567	Name of beneficiary, vat number	medium
National Register of Associations of the Ministry of Interior	i, iii	Accredited NGO	44	CIF of organization	low
Fundación Lealtad	i, ii, iii	Accredited NGO	191	Name of organization	low

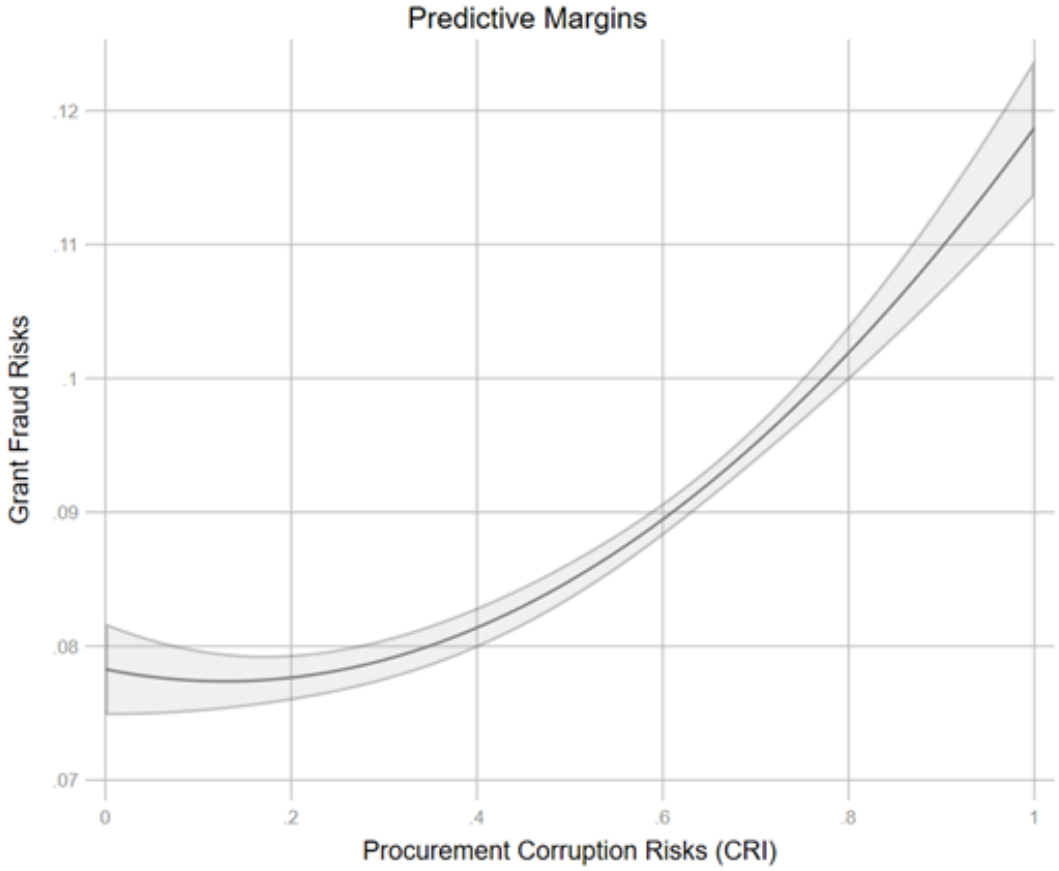
Public procurement data: external risk scores

Public procurement data can point at risk features of both grators, grantees and third parties

Variables	Description	Type of variable
Supplier ID	Unique ID of supplier	Text
Buyer ID	Unique ID of buyer	Text
Name of supplier	Name of supplier winning the contract	Text
Name of buyer	Name of buyer providing tender call	Text
Number of bids	How many bids were made per tender	Numeric
Procedure type	Is the procedure type open or restricted	Categorical
Public call	Was the call for tender available to public	Categorical
Length of bid submission	What is the length between start and end date of bid submission	Numeric
Length of decision period	What is the length between end date of bid submission and decision	Numeric
Connections	Are there recorded connections between supplier and procurement authority	Categorical

Public procurement corruption risks and grant fraud risks

- Matching based on beneficiary NIF
- Corruption risk indices such as single bidding or non-publication of the call for tenders co-vary with grants fraud risks of our model
- R2 for non-linear regression model is 0.17



SUGGESTIONS AND QUESTIONS?


OECD

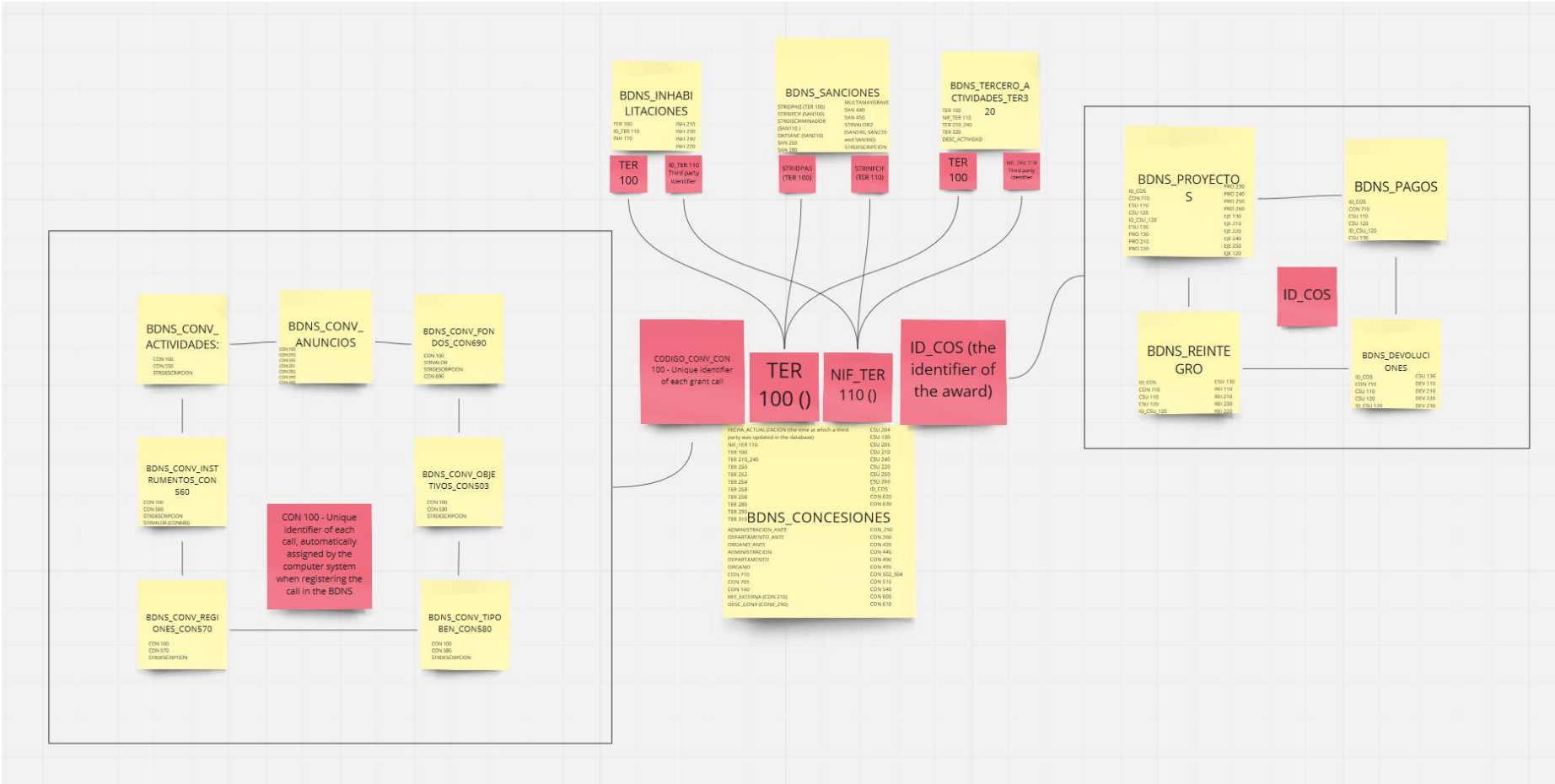
ANNEX

Data

Three stages of data processing:

1. Merging: 17 datasets with different levels of observations (call level, award level, third party level) => aligning to award-level
2. Anonymisation (replacing identifiers)
3. Cleaning: dropping variables with low variation OR high missing rate (>50%)
4. Preparing for analysis: dropping text variables, keeping only complete observations

Dataset merging: overview



- Merging used different IDs depending on the level of observation in each dataset
- Lower levels of observations were aggregated to match to award level

Variables in the analysis

Variable	**Short description**	**Variable description**	Type
ABIERTO_CON420	Open admission period	Indicates if the call keeps the application admission period open permanently	factor
AUDAESTADO_CON490	Condition of State Aid	Indicates if the aid of the call should be classified as ADE	factor
FINALIDAD_CON540	Purpose	Public utility or social interest or promotion of a public purpose pursued with the granting of the subsidy	factor
NOMINATIVA_CON610	Nominative grant	Nominative grant condition	factor
PUBLICABLE_CON620	Publication	Condition of publicity of the concessions	factor
IMPACTOGENERO_CON630	Gender impact	It rates the expected results in relation to the elimination of inequalities between women and men and the fulfillment of the equality policy objectives	factor
PAIS_TER100	Third party country	Country that generates the identification of the third party	factor
PAISDOM_TER250	Country of domicile		factor
NATURALEZA_TER280	Legal nature of the third party		factor
TIPOBEN_TER290	Third party type	Cataloging of third parties based on their legal nature and economic activity	factor
COSTE_ACT_CSU240	Costs	Amount of the fundable budget of the activity to which the grant award applies	numeric
IMPORTE_PAG220	Amount paid (grant)		numeric

RETENCION_PAG230	Retention	Condition of tax withholding carried out	factor
CON560	Help instrument	One or more of the legal or economic figures on the basis of which the subsidies and aid are awarded	factor
CON580	Types of beneficiary	One or more of the types of beneficiary foreseen in the call	factor
SAN_dum	Sanctions	If the award was sanctioned	factor
Month_CSU210	Month of award	Month of the date when the grant was awarded	factor
Nawards_TER_110	Number of awards	Number of awards received by the same third party	numeric
Amount_awards_TER110	Amount of awards	Overall amount of awards received by the same third party	numeric
NATIONAL_CSU260 REGIONAL_CSU260 MUNICIAPAL_CSU260	Level of award	If the grant was awarded by national, regional or municipal body	factor
NATIONAL_TER310 REGIONAL_TER310 MUNICIAPAL_TER310	Level of third party location	If the third party is located at national, regional, municipal level	factor
LOCAL_IMPL	Local implementation	If the location of third party is the same as location of granting body	factor
SECTOR_CON550_AGR...EXTRATER	Sector of economy	Sectors of the economy foreseen in the call	factor

Methodology

Positive-Unlabelled learning using Random Forest (aka PU bagging)

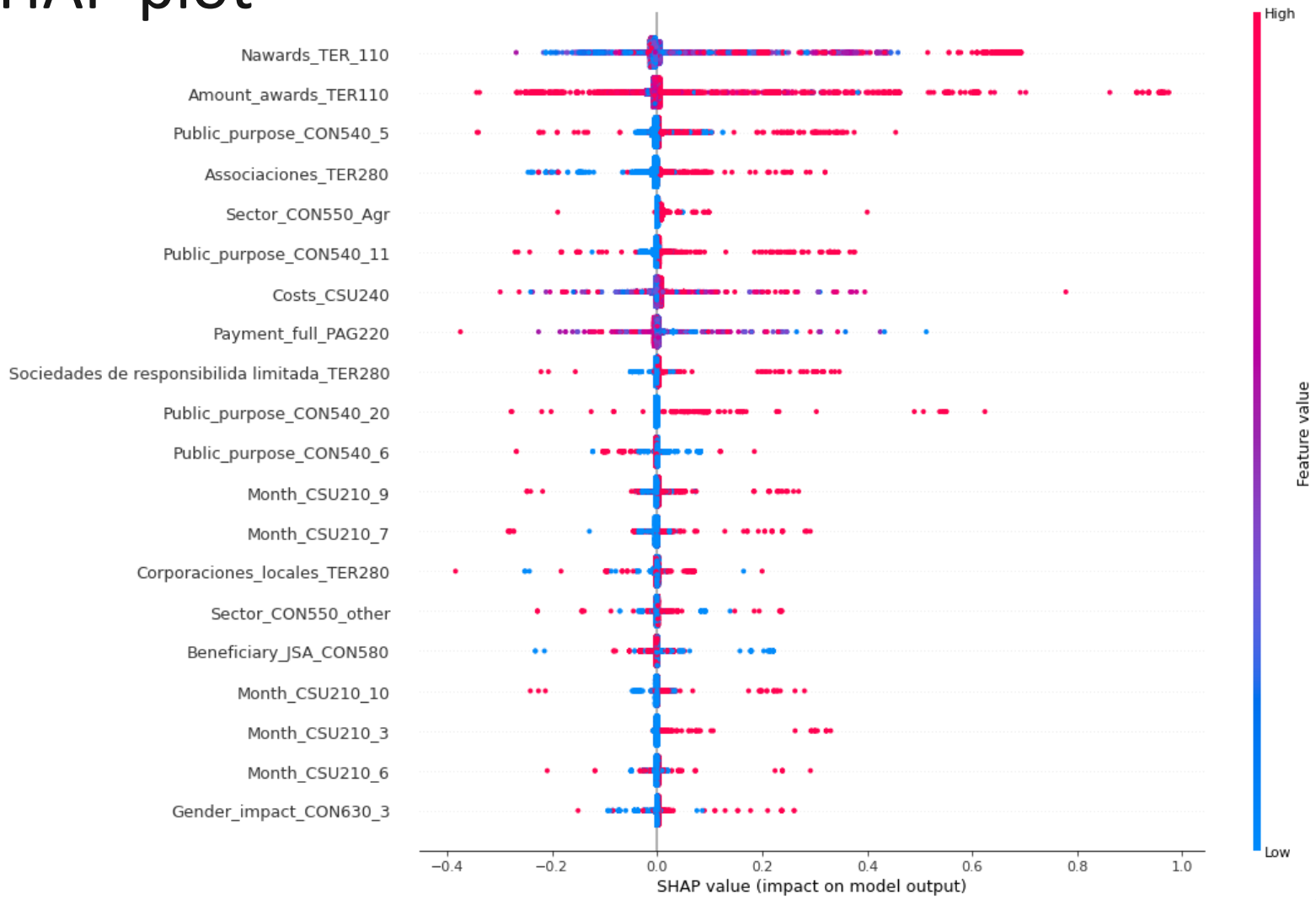
1. Run several standard Random Forests
2. Relabel unknown cases as negatives if they get a very low risk score compared to proven cases
3. Training a Random Forest model on relabelled sample: positive vs likely negative cases
4. Model quality assessment

Details of our methodology as implemented in the IGAE data

- Bagging model with nr.trees = 1000
- Highly unbalanced data: 1031 sanctioned awards vs 1,049,439 non-sanctioned awards => relable negative (not sanctioned) cases using PU bagging
  - Relabel unlabelled cases
  - Full RF model

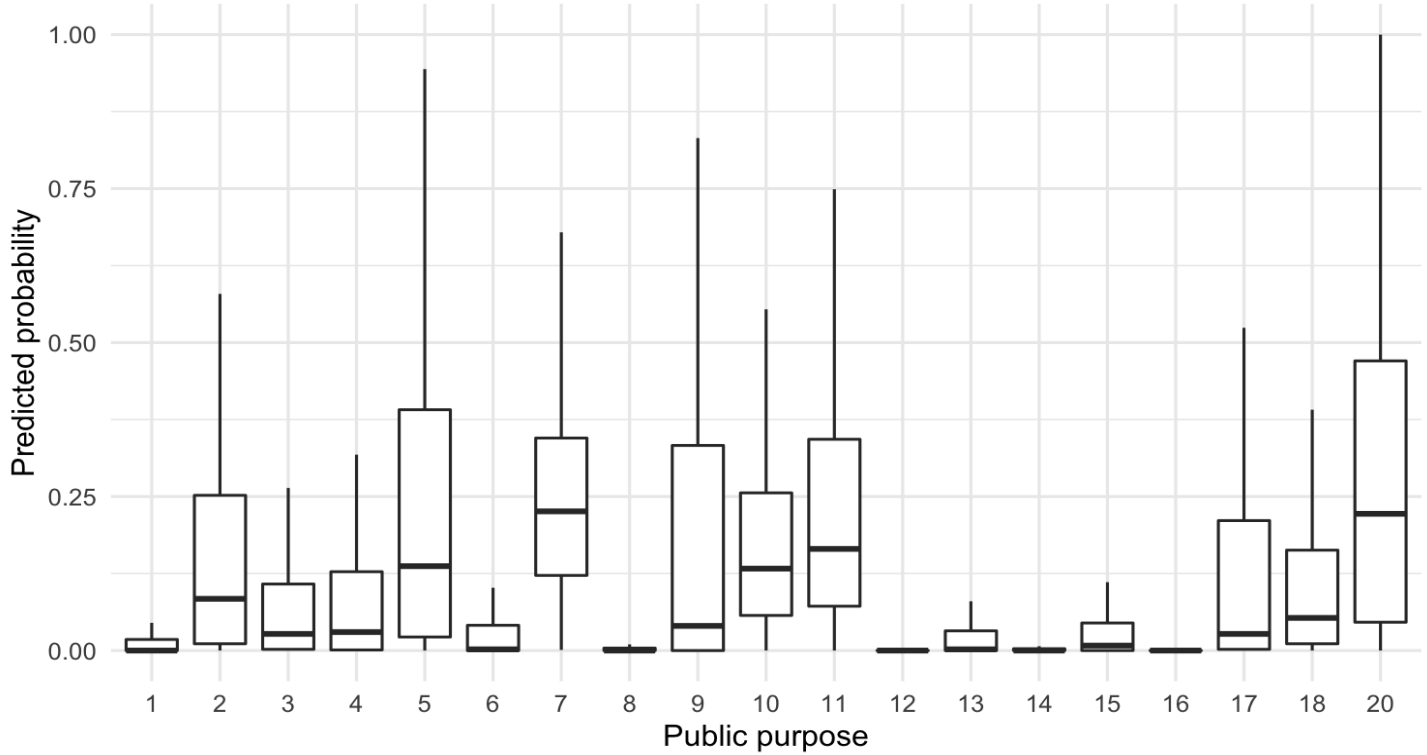
Results: influential predictors

SHAP plot



Results: Predicted probabilities by public purpose of the call

Two categories show the most extended risk of sanctions: social services (5) and international cooperation for development and culture (20).



Recommendations: Data

1. Data quality
   - Filling gaps in the variables with high missing rate
2. Behavioral indicators
   - More focus on risk indicators rather than background variables: company data
3. Efficient data pipeline
   - Developing techniques of data aggregation and flattening for merging purposes

Highest priority dataset for matching: National company registry and financial data

Can be matched to the main BDNS dataset by the company's NIF number

Plenty of potential risk indicators and red flags

Variables	Description	Type of the variable
Name	What is the name of the company	Text
NIF	What is the NIF number of the company	Text
Date of incorporation	When the company was incorporated	Date
Company address	Where the company is registered	Text
Sector of economic activity	In which economic sector the company operates (NACE)	Categorical
Legal form	Official legal form of the company (national forms)	Categorical
Company status	If company is active and operational	Categorical
Company's assets	Total value of items benefiting the company economically	Numeric
Company's liabilities	Total value of company's obligations	Numeric
Company's income	Total amount of income generated annually	Numeric
Company's expenditures	Total amount of spendings list per year	Numeric
Changes in equity	If there were any changes in equity for the past year	Binary + text
Cash flows	Increase or decrease in the amount of money	List
Members	Includes the name of all members of the current organic representation	Text
Beneficial owners	List of names of final owners of the company	Text