University of Fribourg

Department of Economics

Flagging collusion using Machine Learning

Master Thesis

submitted by

Ana Paula Armendariz Pacheco

Principal Supervisor	Prof. Dr. Martin Huber
	Chair for Applied Econometrics & Policy Evaluation
Matriculation Number	21-215-793
Field of Study	Data Analytics & Economics
Contact Details	ana.armendarizpacheco@unifr.ch

Abstract

This study evaluated the effectiveness of an ensemble method in detecting collusive behavior across different countries and industries. The ensemble method combined six algorithms based on statistical screens of bid or price distributions. The method was applied to five different data sets from proven cartel cases across four countries (Brazil, Italy, Japan, and Switzerland) as well as on a combined data set. Later, a rotating scheme was implemented to examine the transferability of the method across different countries and industries by training the ensemble method on a mixed data set and testing its performance on unseen data. The ensemble method achieved overall correct classification rates ranging from 74% to 88% when applied on country and industry-level data sets, and from 74% to 96% when an additional screen was added. The estimation on the combined data set yielded an overall correct classification rate of 76%, and of 80% when applied on a perfectly balanced combined data set. However, the model struggled to capture the complexity of the data when trained on the combined data set and tested on unseen data from different countries and industries. This research underscores the potential of machine learning for detecting collusive behavior, while acknowledging the need for further research to develop models that can be applied more broadly across industries and countries.

Contents

1 Introduction

In markets, competition plays a vital role in stimulating innovation, promoting efficiency, and expanding consumer options. However, firms may have incentives to coordinate which can produce detrimental effects to all stakeholders, including consumers, firms, and the economy. Collusive agreements undermine competition, increase inefficiencies, create barriers to entry, stifle innovation, and adversely affect consumers by raising prices and reducing the quality of goods and services (Aghion et al., 2014; Bergman, 2019; OECD, 2020). Therefore, it is imperative to take steps to guard against collusion and promote fair competition in markets.

Despite being illegal, collusion continues to persist in markets around the world. In public procurement, it is the cause of significant financial damages to governments and taxpayers, with estimates suggesting losses amounting to billions of dollars annually. Moreover, these unethical practices erode public confidence in the bidding process, as well as the credibility of the authorities (Signor and Ika, 2020).

Given the significant harm caused by collusion in markets, there is a need to develop tools to detect and prevent collusive behaviour. For such reason, this research aims at contributing to the existing literature to develop a general model to flag collusion which can be used across industries and countries, similarly to the research done by Huber and Imhof (2019), Wallimann et al. (2022), Huber et al. (2022), García Rodríguez et al. (2022), Silveira et al. (2022), Adam et al. (2022), Fazekas et al. (2023). This study implements an ensemble method as machine learner based on the work of Huber and Imhof (2019), Wallimann et al. (2022) and Huber et al. (2022) which considers six algorithms that use as inputs, statistical screens of bid/price distributions. Such method is applied on five different data sets of proven cartel cases from four countries, namely, Brazil, Italy, Japan and Switzerland from various industries such as oil infrastructure, gasoline markets, road construction, and civil engineering.

The contributing factor of the present study is to determine how well the ensemble method can detect collusive instances across different countries and industries. For this, a rotating scheme was implemented where the ensemble method is trained on a combined data set and then tested on a new country data set that had not been used in the training phase. The performance of the ensemble method was evaluated based on the accuracy of its predictions in correctly classifying collusive and competitive tenders/markets. Accuracy was calculated by comparing the model's predictions to the actual classification of each tender, and it represents the proportion of correct predictions out of the total number of predictions made by the model.

Overall, this is an innovative approach that provides insights into the effectiveness and transferability of the method to different countries or industries and has significant implications for efforts to detect and prevent collusive behavior in global markets.

In this study the ensemble method was applied to different variations of the data available. Firstly, the ensemble method was applied on each country/industry data set, then it was estimated on a unified data set, and finally the rotating scheme was implemented. The estimations which used country or industry level data, achieved correct classification rates, ranging from 74% to 88% and from 74% to 96% when adding the number of bid/prices as an additional screen, indicating their effectiveness. However, there was variability in performance across different countries, suggesting that contextual factors may influence the models' effectiveness which is also reflected on the fact that no algorithm used in the ensemble method was consistently assigned more weight than others.

As for the results on the combined data set, the ensemble method achieved a 76% overall correct classification rate and of 80% when estimated on the perfectly balanced combined data set. Lastly, the results from the rotating scheme indicated that the model lacked the ability to perform well when it was trained on a mixed data set and later, tested on an unseen data set suggesting that the model is not transferable to different contexts.

Based on the results, the model remains to be a powerful tool to detect collusive behaviour when trained and tested on data from the same country, indicating its potential usefulness in specific contexts. Overall, this study contributes to the ongoing research on the use of machine learning to detect collusive behavior and highlights the need for further research to develop models that can be applied across different industries and countries.

This research begins by providing an introduction to collusion, including its definition, types, and detection methods. It then presents the literature on how machine learning can be employed to detect bid-rigging or collusion in markets. Next, the data sources used in the study are described, followed by an overview of the methodology. Subsequently, the findings are presented, and conclusions are drawn regarding the efficacy of the ensemble method in detecting collusion across industries and countries. Finally, the study concludes by exposing its limitations.

2 Context and Literature review

2.1 Competition and collusion

Competition is an essential element in markets as it promotes innovation and efficiency, as well as consumer satisfaction and prevents market abuse. Firstly, competition encourages innovation by motivating firms to invest in developing new and superior products in order to stay ahead of their competitors (Aghion et al., 2014). Furthermore, by increasing product diversity and lowering prices, competitiveness increases consumer choice and welfare (Bergman, 2019). Lastly, competition discourages market abuse by ensuring equal market access for suppliers and consumers (OECD, 2020).

In addition, competition holds significant importance in public procurement markets, which represents one of the most concrete expressions of public expenditure. Public procurement constitutes a significant proportion of GDP in most nations. Current estimates indicate that governments worldwide allocate approximately 13 trillion USD annually towards public contracts (WorldBank, 2022). In the European Union alone, public authorities spend 2 trillion EUR each year, representing about 14% of their GDP and remain important purchasers of various sectors such as education, health, energy and transport (OECD, 2020).

Effective procurement can promote economic growth, investment, and job creation, while fostering an innovative, energy-efficient, and socially inclusive economy. Improving public procurement can result in substantial savings, with even a 1% increase in efficiency yielding up to 20 billion EUR annually. Therefore, it is crucial to ensure that procurement processes are well-managed (European Commission, 2021, WorldBank, 2022).

Nevertheless, in both private and public markets, firms may have strong incentives to collude to increase profits. By colluding, firms can create barriers to entry, set higher prices, and limit production, among other strategies. For these reasons, consumers and contractors pay more for lower-quality goods and services. As a result, society suffers welfare losses and costly inefficiencies (Porter, 2005).

Collusion is specially damaging in public procurement, also commonly referred as bid-rigging, which is a recurring phenomenon that costs governments and tax payers billions of dollars each year. Collusion is estimated to increase the costs paid by public procurement agencies by 60% over competitive conditions. In summary, collusive arrangements undermine the benefits of a fair and transparent procurement market while lowering trust in public authorities and procurement process (European Commission, 2021, Signor and Ika, 2020).

2.2 Bid/price-rigging cartels and how to detect them?

There are multiple ways companies can collude. According to Toth et al. (2015), collusion can be classified into three groups: elementary collusion, rent sharing and the resulting market structure. Elementary collusion techniques ensure that contracts are won by the agreed-upon supplier which can be accomplished by withholding bids, submitting deliberately losing bids or joint bids. Moreover, companies can use the second method to allocate rent between them, distribute geographical markets, win contracts cyclically, subcontract or make secret payments to each other. Finally, collusion can lead to highly concentrated markets and companies can also coordinate to win cyclically (Adam et al., 2022). All these strategies constitute bid-rigging, hence, bid rigging is a serious violation of antitrust laws and can lead to higher prices for goods or services, reduced quality, and limited choices for buyers.

Detecting collusive behavior presents a significant challenge due to its highly secretive nature which often remains undetected during the procurement process and is only discovered after the contract has been awarded or completed. However, there are various methods that can be used to detect collusion before or after the presence of a collusive arrangement. One approach is the use of structural and behavioral methods to flag collusive behaviour. On one hand, a structural approach considers the factors that influence the likelihood of collusion while behavioral approaches examine the bidding or pricing behavior of competitors to identify markets where cartels may have already been established (OECD, 2022).

For example, the structural approach to collusion detection identified that cartels form in settings with fewer firms and where similar goods are sold. In contrast, behavioral approaches detect cartels based on their actions such as communication between cartel members or based on the patterns of their bids submitted in auctions (Signor and Ika, 2020, Harrington, 2006). In general, behavioural approaches are thought to be better to detect collusion (OECD, 2022).

One example of a behavioural method to detect collusive markets is the use of statistical screens. These screens are quantitative tools designed to identify markets where competition may be absent or insufficient, providing an indication of potential anti-competitive behavior that requires further investigation (Abrantes-Metz et al., 2012).

Statistical screens are effective tools to distinguish between collusive and competitive periods. This is because bid/price distributions tend to behave differently under competitive and collusive conditions. Under competition, companies act independently and compete with one another to secure a contract. As a result, with only a few outliers, their bids or prices are expected to fluctuate randomly around the market price. This implies that the distribution of bids or prices should be fairly even. On the contrary, in a collusive market, the distribution of bids or prices will be less spread out since firms coordinate to control prices. They may, for example, agree to maintain a certain range. In this sense, we can use statistical screens as a resource to detect unusual bid behaviour (Adam et al., 2022).

Therefore, statistical screens can be used to compare the distribution of bids or prices between periods. Specifically, variance-based statistical screens which are a type of statistical screen that can be particularly effective in detecting collusive behavior. Variance screens focus on various statistical measures of bids or prices, including but not limited to, normalized relative distance, percent difference, standard deviation, coefficient of variance, spread, skewness, and kurtosis (Adam et al., 2022). As mentioned earlier, the literature suggests that in the presence of collusion, bid or price variance reduces which is reflected in the statistical measures just mentioned (Huber and Imhof, 2019, OECD, 2022).

In summary. the use of statistical screens based on bid or price distribution can be a powerful tool for identifying collusive behavior. Nevertheless, their power can be limited by the complexity and variability of the data available. However, Machine Learning (ML) methods offer a way to overcome these limitations by combining multiple statistical screens and making use of large, complex data sets in a more accurate manner. ML techniques can help identify patterns and relationships in the data that may be difficult or impossible to detect through traditional statistical screens. Moreover, a variety of algorithms can be trained to recognize specific patterns of behavior associated with collusive conduct and can be continuously improved over time. Since collusion detection is a multi-stage process and screening is just one part of it, ML techniques can be used as a flagging tool that serve as an input for further investigation. As such, ML methods can be a valuable addition to the competition authorities' toolkit for detecting anti-competitive behavior in public and private markets (OECD, 2022).

2.3 Machine Learning to detect cartels

Recently, there has been a surge in the use of ML methods which combine multiple statistical screens based on bid or price distribution by both competition authorities and scholars (OECD, 2022). To begin, Machine Learning involves programming computers in a way that enables them to acquire knowledge from data without being explicitly programmed. Generally, ML can be supervised or unsupervised (Géron, 2017).

This project employs supervised learning where the algorithm is trained on labeled data, which means the data is already tagged with the correct output. In this case, each tender or market in the data set is labelled as collusive or competitive based on proven cases of collusion. Using the labelled data, the algorithm learns to map the input to the correct output. In this context, it learns to correctly classify collusive or competitive instances and later on, it can predict the class on unseen data. This is accomplished by minimizing the error between the predicted and actual output (Samuel, 1959, Géron, 2017).

Among the benefits of the use of Machine Learning techniques in collusion detection is that it enables to have data driven estimations and accuracy improvement. This is due to the fact that ML can combine several statistical screens to better predict collusion (Huber and Imhof, 2019, Huber et al., 2022 and OECD, 2022). Additionally, studies show that machine learning models frequently outperform standalone statistical indicators in terms of precision (The Danish Competition and Consumer Authority, 2022).

Adam et al. (2022), for example, implemented a supervised ML - based algorithm to predict the likelihood of collusion and applied to model to 13,640 contracts from France, Hungary, Latvia, Portugal, Spain and Sweden. The researchers used a random forest algorithm to build their models and estimated them on data from the six different countries which was collected before and after confirmed instances of cartels. The model used as inputs, ten collusion risk indicators, such as the number of bids, single bidding, and relative price, among others. Their results revealed that the random forest models were highly effective in identifying collusive activities. However, the authors also found that no single indicator could accurately detect all types of cartel strategies, highlighting the importance of employing a combination of indicators. As evidenced from the study, machine learning allows for the integration of various collusive risk indicators to improve detection accuracy.

Similarly, García Rodríguez et al. (2022) examined the performance of multiple ML algorithms to detect collusive instances using screens calculated from bid/prices values of various tenders and markets from Brazil, Italy, Japan, Switzerland and the United States. In total, eleven algorithms were tested which include Stochastic Gradient Descent, Ensemble Methods, Support Vector Machines, Nearest Neighbors, Neural Networks, Naive Bayes and Gaussian process. The authors found that Extra trees, Ada boost, Gradient Boosting (ensemble methods) and Random Forest were the top performing models which obtained an accuracy between 81% and 95%.

Another example is the study done by Silveira et al. (2022), in which the authors examined the case of a gasoline cartel in four regions in Brazil by studying the price distribution before and after the cartel was in place. They used a combination of algorithms which include random forest, lasso logistic, neural networks and ridge logistic. Concerning the screens, they found that the coefficient of variation and the standard deviation were the most powerful predictors as well as all statistical moments. Overall, the models correctly classified 87% of the collusive periods.

Moreover, Huber and Imhof (2019) investigated the case of a Swiss cartel which operated in the construction sector. They combined ML techniques such as lasso logit regression, ensemble methods and random forests. These models take the statistical screens from Imhof et al. (2018) as inputs. For the lasso and ensemble methods, the model achieved an accuracy of 84%. The researchers also discovered that the coefficient of variation and the normalized distance are the two most powerful screens. Lastly, they concluded that the there is no single screen or method that better detects collusion but instead it is a combination of screens and ML techniques that are useful for detecting collusion.

Furthermore, Huber et al. (2022) explore the degree to which statistical methods for detecting collusion, can be applied across nations and industries. The authors used data sets from Switzerland and Japan to train an ensemble method, which combined various algorithms to predict collusive instances. Specifically, the authors trained the ensemble method on the Swiss data and then tested it's performance using the Japanese data set and vice-versa. When doing so, the authors found that the imbalance of truly collusive or competitive tenders increases. Overall, when using an ensemble method, the model which was trained and tested on different country data had a correct classification rate between 85% and 90%.

Finally, Fazekas et al. (2023) used machine learning to develop a model for cartel detection which combined various cartel screens. One notable strength of their study is the extensive and diverse data set collected from seven countries, namely Bulgaria, France, Hungary, Latvia, Portugal, Spain and Sweden. The data sets were sourced from contract-level public contracting data from 2004 to 2021 in which the authors identified 68 cartel cases and were able to test 44 of them. Their best performing model was a random forest algorithm which achieved a 77-91% accuracy. In addition, the authors found that no single indicator or group of indicators could predict a wide set of cartel behaviors. The authors also used their model to predict the probability of cartel activity in over 3.3 million contracts.

Having reviewed the literature, there is evidence of the growing use various machine learning methods to detect collusion. As such, this study aims to contribute to the literature by examining the effectiveness of machine learning algorithms in detecting collusive activities across different industries in public and private markets. To achieve this objective, this research takes advantage of the availability of data on various proven cartel cases which will be described in the following section.

3 Data

3.1 Data sets

For this study, a total of five data sets on public procurement and product prices were gathered from publicly available sources from four countries. The raw data from cartels in Brazil and Italy came from research by García Rodríguez et al. (2022). Additionally, the data on the Brazilian gasoline cartel was compiled independently. However, the data search was motivated based on the study of Silveira et al. (2022). Finally, the data from the Japanese and Swiss cartels are from Huber et al. (2022). The data sets used in this analysis are sourced from verified instances of collusion which were investigated by competition authorities across the various countries. Hence, based on the evidence from these established cartel cases, it is possible to categorize each tender as either collusive or competitive and use this variable to train the supervised ML models.

Although the initial data included details such as tender date, pre-tender estimates, number of awardees, and company information, I only utilized the tender ID, submitted bids, and a binary indicator to determine whether a tender was collusive or competitive. In a next step, I transformed the data from long format to short format, where each tender now corresponds to a row and the columns are the submitted bids for a specific tender. Furthermore, a binary variable was assigned to each tender, taking the value of 1 if it was classified as collusive and 0 if it was competitive. The only exception is the Brazilian gasoline cartel data, which is the only data set derived from a public register of gasoline prices rather than from public procurement cases. Each row represents a month, and the column represents the weekly gasoline prices for that month. As a result, the unit of observation of all data sets, expect for the Brazilian gasoline case, are tenders. An overview of the data sets is presented in Table 1.

Table 1 shows the five data sets from Brazil, Italy, Japan, and Switzerland that were used in this project. To begin, the Brazilian data set was sourced from a bid-rigging cartel case by the Brazilian oil company 'Petrobras', involved in colluding in infrastructure projects from 2002 to 2013. The cartel was first examined by Signor and Ika (2020) and revealed that sixteen of the largest construction companies in Brazil colluded in many of the 'Petrobras' tenders. Overall, the data set has 101 observations of which 33% are classified as collusive and 67% are competitive.

Table 1 Original data sets

Name	Sector	Year	N	Collusive vs Competitive
Brazil	Oil infrastructure projects	2002-2013	101	33% vs 67%
Brazil	Gasoline prices	2004-2019	432	48% vs 52%
Italy	Road construction	2000–2003	278	51% vs 49%
Japan	Building construction and civil engineering	2003-2007	438	56% vs 44%
Switzerland	Road construction and civil engineering	1998-2010	2894	39% vs 61%

The second data set, also from Brazil, was obtained from a proven cartel case in the gasoline market, conducted by the Administrative Council for Economic Defense (CADE). The case demonstrated that gasoline resellers had colluded and fixed prices in four different municipalities in Brazil during 2004 to 2019 (Silveira et al., 2022). The actual gasoline price data was obtained from the Brazilian National Agency of Petroleum, Natural Gas and Bio fuels which is responsible for planning and maintaining the retail fuel price database in the country. From such database, I sourced the weekly collection of the retail gasoline prices where I obtained a sample of 432 observations which was quite balanced between collusive and competitive instances.

Moreover, the Italian data came from a proven cartel case in the road construction industry from the Municipality of Turin made up of seven companies involved in bid-rigging from 2000 to 2003. The data set consists of 278 tenders of which 51% are collusive and 49% are competitive. More information can be found in García Rodríguez et al. (2022) & Conley and Decarolis (2016).

In addition, the Japanese data set was sourced from a cartel case in construction and civil engineering contracts in the Okinawa Prefecture in Japan which found a significant number of companies guilty of bid-rigging which prevailed from 2003 to 2007. The data set contains 438 tenders, 56% were classified as collusive and 43% as competitive. Further details can be found in Ishii (2014) and Huber et al. (2022).

Finally, the data set from Switzerland is actually a pooled data set that covers cartels in road construction projects from the Canton of Ticino, Saint Gallen and Graubünden from 1998 to 2010. This data set is the largest, covering 2894 tenders of which 39% are collusive and 61% are competitive. Additional details can be accessed in the following studies: Imhof et al. (2018), Huber and Imhof (2019) and Huber et al. (2022).

In a further data cleaning step, for each data set, I removed the tenders for which only 1 bid was submitted given that the ML model requires more that 1 bid per tender to be estimated. Subsequently, as can be seen on Table 1, the Brazilian (oil infrastructure projects) was unbalanced in terms of the number of collusive and competitive observations and had to be re sampled. This is done since class imbalance can lead to biased predictions, low accuracy, over fitting and overall inadequate model evaluation. The modified data sets are described in Table 2.

Table 2 Modified / balanced data sets with additional columns

Name	Collusive	(%)	Competitive	(%)	N	Average bids per tender
Brazil	33	42%	45	58%	78	8
Brazil	207	48%	224	52%	431	4
Italy	143	51%	135	49%	278	73
Japan	245	56%	193	43%	438	13
Switzerland	1076	39%	1696	61%	2772	5

Table 2 presents the modified data sets for the four countries, offering insights into the share of collusive and non-collusive tenders, the total number of tenders, and the average number of bids per tender. The results highlight that now, most data sets have a roughly equal split between collusive and non-collusive tenders. Switzerland stands out as having a higher percentage competitive tenders while Japan has a greater number of collusive tenders. Additionally, Italy's procurement auctions show a higher average number of bids per tender compared to all other countries followed by Japan.

3.2 Statistical screens

Building upon the comprehensive data presented in the previous section, I will now introduce the statistical screens that will be employed to detect collusion in procurement auctions and markets. These statistical screens are based on the modified and balanced data described in the previous section which will serve as inputs for the subsequent machine learning models.

The statistical screens used in this study are based on Imhof et al. (2018), Huber and Imhof (2019) and Huber et al. (2022), which are the following:

1. Coefficient of variation which is an statistical measure of the dispersion of data points around the mean:

$$CV_t = \frac{SD_t}{\mu_t} \tag{3.1}$$

In this equation, CV represents the coefficient of variation, SD represents the standard deviation, and μ represents the mean for a tender/market t. During cartel periods, it is anticipated that the variance of bids/prices will decrease. The reason is that often, cartels members agree to place bids/prices which are higher than the lowest bid in a competitive scenario. As a result, the distribution of the bid/prices becomes more concentrated.

2. Spread is a measure in statistics that is calculated as the difference between the maximum and minimum values of a variable. In this case, it is also divided by the minimum value:

$$SPDt = \frac{b_{max,t} - b_{min,t}}{b_{min,t}} \tag{3.2}$$

3. Kurtosis is a statistical measure that indicates the degree to which a probability distribution is peaked or flat compared to a normal distribution. It is calculated by comparing the distribution's fourth moment to the square of its variance. Positive values of kurtosis indicate a peaked distribution, while negative values indicate a flat distribution.

$$KURTO_t = \frac{n_t(n_t+1)}{(n_t-1)(n_t-2)(n_t-3)} \sum_{i=1}^{n_t} \left(\frac{b_{i,t} - \bar{b_t}}{s_t} \right)^4 - \frac{3(n_t-1)^3}{(n_t-2)(n_t-3)} \tag{3.3}$$

Where: n_t: is the number of bids in the tender/market t. $b_{i,t}$: is the i-th bid of the variable of tender/market t. $\bar{b_t}$: is the mean of the bids in tender/market t. s_t: the sample standard deviation of the bids/prices.

Kurtosis is a useful tool for detecting collusive behavior among firms because in collusive situations, firms may designate a winner and submit bids/prices that are lower than the established winning bid/price. This can result in a convergence of bids and create peaked distributions, which can be detected through higher kurtosis values.

4. A useful measure related to the symmetry of the distributions of the bids/prices is the difference between the two lowest bids/prices.

$$DIFFPt = \frac{b_{2,t} - b_{1,t}}{b_{1,t}} \tag{3.4}$$

Where $b_{2,t}$ is the second lowest bid and $b_{1,t}$ is the lowest bid/price for tender t. There is a prevailing notion that the existence of substantial gaps between the two lowest bids/prices in tenders is inconsistent with a competitive environment.

5. Another symmetry measure of the bid/prices distribution is the relative distance between the bids/prices measured by the following equation:

$$RD_t = \frac{b_{2,t} - b_{1,t}}{s_{d,t}} \tag{3.5}$$

In the formula above, $b_{1,t}$ is the lowest bid/price in the tender at time t, $b_{2,t}$ is the second lowest bid/price in the tender at time t. $s_{d,t}$ is the standard deviation of the losing bids/prices in the tender at time t.

6. A variation of the relative distance measure is the normalization of the difference. It consists of using the mean difference between all adjacent bids/prices in a tender or market as a denominator.

$$RDNOR_t = \frac{b_{2,t} - b_{1,t}}{\sum_{i=1}^{n_t-1}(b_{i+1,t} - b_{i,t})/(n_t - 1)} \tag{3.6}$$

Where n_t is the number of bids/prices in the tender at time t. The denominator of the equation is the average difference between bids/prices in a tender.

7. Another approach is to use an alternative distance measure that excludes the lowest bid/price from the calculation of the mean difference between losing bids/prices, which is then used as the denominator.

$$ALTRD_t = \frac{b_{2,t} - b_{1,t}}{\sum_{i=2}^{n_t-1}(b_{i+1,t} - b_{i,t})/(n_t - 2)} \tag{3.7}$$

8. Skewness is a statistical measure that indicates the level of asymmetry of a probability distribution. Positive skewness implies a longer tail on the right side, while negative skewness indicates a longer tail on the left side.

$$SKEW_t = \frac{n_t \sum_{i=1}^{n_t} \left(\frac{b_{it} - b_t}{s_t}\right)^3}{(n_t - 1)(n_t - 2)} \tag{3.8}$$

In the equation, the variables are represented as n_t which denotes the number of the bids/prices. Also, $b_{i,t}$ is the i_{th} bid/price for tender t. Lastly, s_t is the standard deviation of the bids/prices and $\bar{b_t}$ is the mean of the bids/prices for tender t.

9. The last statistic used was the nonparametric Kolmogorov-Smirnov statistic which measures whether the bids/prices follow a uniform distribution. A small KS statistic indicates a good fit between the sample data and the uniform distribution, while a large KS statistic indicates a poor fit.

$$D_t^+ = \max_i \left[\frac{x_{it} - i}{n_t + 1} \right], D_t^- = \max_i \left[\frac{i}{n_t + 1} - x_{it} \right], KS_t = \max(D_t^+, D_t^-) \quad (3.9)$$

In the context of tenders, the variables n_t, i_t, and x_{it} represent the number, the rank, and the standardized bids/prices for the i_{th} rank in tender/market t, respectively. The standardized bids/prices are obtained by dividing the bid/prices by their standard deviation within the tender/market. Such step enables a fair comparison of tenders/markets with different contract values. Based on Wallimann et al. (2022), it is expected that collusion makes the bids/prices less uniform which can be detected by the KS statistic.

To have a concise overview of the screens previously presented, the tables below provide a summary statistics of two main statistical screens, namely the coefficient of variation and the normalized distance. The analysis is done for each country and based on whether the tender/market was categorized as collusive or competitive. I focused on the coefficient of variation and the normalized distance based on Huber and Imhof (2019) given that in their study, the authors found that these two screens were the most powerful [1]. As previously stated, the coefficient of variation is expected to be lower during cartel periods compared to non-cartel periods. On the contrary, during cartel periods, the normalized distance is expected to be higher than in non-cartel periods.

Table 3 Descriptive statistics of the Coefficient of Variation

Country	Cartel Period			Non-Cartel Period		
	Mean	Median	SD	Mean	Median	SD
Brazil (oil infrastructure)	0.07	0.06	0.03	0.18	0.17	0.1
Brazil (gasoline)	0.01	0.00	0.01	0.01	0.01	0.01
Italy	0.04	0.04	0.02	0.03	0.05	0.01
Japan	0.01	0.0	0.02	0.05	0.05	0.03
Switzerland	0.05	0.04	0.04	0.09	0.08	0.08

Table 3 provides a comprehensive overview of some descriptive statistics, including mean, median, and standard deviation (SD), for the coefficient of variation of

[1] The means for the rest of the screens are displayed in Tables 18-22 in the Appendix.

tenders/markets of the all cartel cases examined. In the case of the Brazilian (oil infrastructure), Japan, and Switzerland, the mean is lower in cartel periods, suggesting decreased competition. On the contrary, in the case of Italy, the coefficient of variation is higher during cartel periods which contradicts the expected trend or in the case of Brazil (gasoline), it remains constant. This could be due to outliers driving the mean down. Additionally, the median increased in all cases in non-cartel periods indicating an increase in the variation of the bids in the post-cartel periods. Finally, the SD shows little variation between periods, expect for Brazil (oil infrastructure), where it increased from 0.03 to 0.1.

Table 4 Descriptive statistics of the Normalized Distance

Country	Cartel Period			Non-Cartel Period		
	Mean	Median	SD	Mean	Median	SD
Brazil (oil infrastructure)	1.34	1.37	0.52	1.03	1.07	0.63
Brazil (gasoline)	0.03	0.01	0.06	0.06	0.03	0.08
Italy	6	0.04	9.17	3.67	0.05	1.49
Japan	3.11	2.99	1.9	1.67	0.86	2.1
Switzerland	1.32	1.12	1.02	0.90	1	0.72

Table 4 presents the descriptive statistics of the Normalized Distance for four countries during cartel and non-cartel periods. The Normalized Distance is an indicator of market competition, where a higher value indicates a less competitive market. The results show that during cartel periods, three out of four countries experienced an increase in the mean Normalized Distance, indicating a decrease in market competition. In contrast, during non-cartel periods, all countries, except for Brazil (gasoline), had a lower mean Normalized Distance, indicating a more competitive market. The Brazilian gasoline cartel is the only exception, as it had a lower mean Normalized Distance during the non-cartel period. From the results, the Brazilian gasoline cartel screens appears to contradict the expected behaviour.

The results of the data analysis suggest that the coefficient of variation and normalized distance can be useful screening tools to detect potential cartel behavior in bidding data. However, the findings are not consistent across all countries, like the case of the Brazilian gasoline cartel. This could be due to several factors, such as small sample size or the presence of outliers.

In addition to the descriptive statistics presented in Tables 3 and 4, visual representations of the screens can provide further insights into potential cartel behavior which are not apparent from descriptive statistics alone. Therefore, in a next step, scatter plots of the coefficient of variation during cartel and non-cartel periods are presented or each cartel case.

Figure 1 exhibits, for each country/industry data set, a scatter plot which displays the coefficient of variation for each tender over time. The vertical red line indicates the end of the cartel period. For all cases except the Brazilian (gasoline) and Italian data sets, there is a noticeable increase in the coefficient of variation after the cartel's dissolution. This evidence suggests that competition increased after the end of the cartels and led to more varied tender prices. On the other hand, for the Brazilian (gasoline prices), and Italian cartels, the graphical analysis of the coefficient of variation does not show a significant difference before and after the cartel period. There are multiple explanations for such result, including the possibility that collusion continued even after the end of the cartel.

Additionally, Figure 2 displays scatter plots of the normalized distance for each tender plotted over time. Here also, a vertical red line marks the end of each cartel period. Upon examination, it is evident that the normalized distance is higher during collusive periods for Japanese and Swiss cartels. Conversely, for the remaining cartels, there is no noticeable disparity between the cartel and non-cartel periods, as demonstrated by the scatter plots. This lack of a significant difference in the normalized distance between periods could be attributed to several factors, such as ineffective collusion, partial compliance with the cartel agreement, or the influence of other market forces on tender prices.

In a final step, I conducted two statistical tests - the Mann-Whitney and the Kolmogorov-Smirnov test on the five data sets to investigate the differences between the collusive and competitive periods within samples. The tests were done on the two screens previously presented, the coefficient of variation and the normalized distance. The p values of both statistical tests are displayed in Table 5 and 6.

Table 5 Summary of Statistical Test Results on Coefficient of Variation

Test	p-value (MW)	p-value (KS)
Brazil (oil infrastructure)	0.00	0.00
Brazil (gasoline)	0.00	0.00
Italy	0.00	0.00
Japan	0.00	0.00
Switzerland	0.00	0.00

According to the results presented in Table 5, the null hypothesis of no difference between cartel and non-cartel periods was rejected for all cases, at any level of sig-

Figure 1 Coefficient of variation for each tender over time

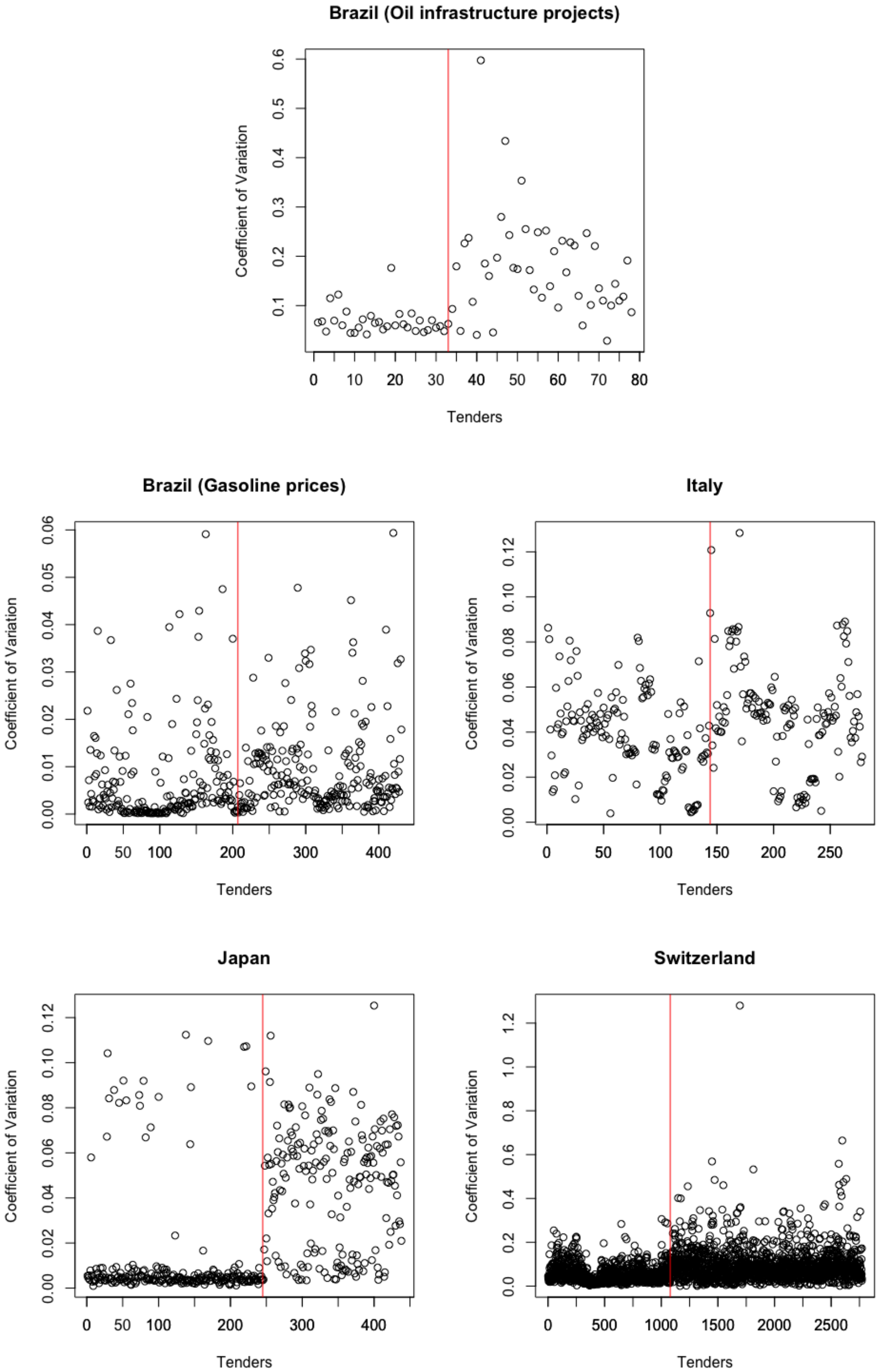
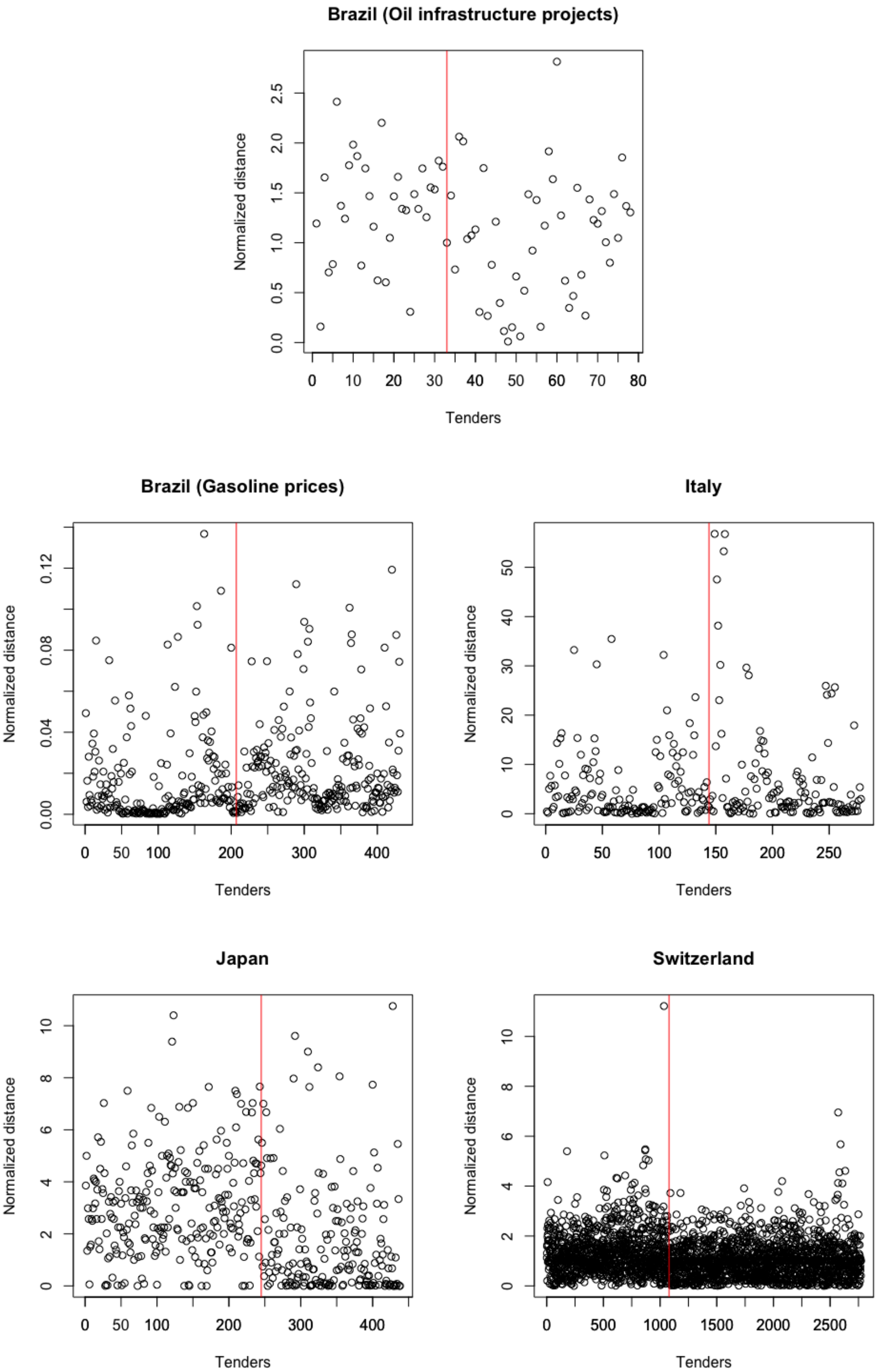


Figure 2 Normalized distance for each tender over time



nificance. For the results of the Mann Whitney test, this means that the samples in collusive and competitive periods are significantly different from each other. Furthermore, the Kolmogorov-Smirnov test also confirm this as the results suggests that the two samples come from different distributions with no overlap between them.

Table 6 Summary of Statistical Test Results on the Normalized Distance

Test	p-value (MW)	p-value (KS)
Brazil (oil infrastructure)	0.01	0.06
Brazil (gasoline)	0.00	0.00
Italy	0.80	0.60
Japan	0.00	0.00
Switzerland	0.00	0.00

Additionally, in Table 6, the p values obtained from both statistical tests on the normalized distance are displayed. The analysis showed that the null hypothesis of no difference between periods can be rejected at a 10% significance level except for Italy. This implies that there is a significant difference between the normalized distance between periods in all data sets except for the Italian case. The lack of statistical significance in the Italian data set suggests that the observed distances are likely to be similar to the expected distances. Overall, these results suggest that the normalized distance can be a useful tool to detect potential cartel behavior in bidding data for most cases.

In summary, the aim of this section was to give an overview of the data used in this project, including the original data and processed input which in this case are the screens per tender/market, that will be used for the estimation of the ensemble method. In addition, the differences in the data sets between cartel and non cartel periods were examined using descriptive statistics, graphs and statistical tests to examine if there are significant differences across periods. The following section focuses on the methods used to flag collusive tenders or markets.

4 Methodology

4.1 Machine learning model to detect bid-rigging cartels

As previously mentioned, the method used in this project is based on the work of (Huber and Imhof, 2019, Huber and Imhof, 2019 and Huber et al., 2022), which aims at correctly predicting whether a tender or market is collusive or competitive, by making use of statistical screens of bid distributions. In order to achieve this, the authors developed an ensemble method as machine learner which designs a weighted average of six algorithms including random forest, bagged decision trees, bayesian additive regression trees, lasso regression, support vector machines and neural nets. The authors take advantage of the fact that ensemble methods are a ML technique that can combine multiple models in order to improve predictive accuracy and reduce over fitting. Instead of relying on a single model to make predictions, ensemble methods use multiple models and then aggregate their predictions to arrive at a final prediction (Hastie et al., 2009, Zhou, 2012).

In order to evaluate the performance of the model, the data needs to be split into a training set and a test set, with a default split of 75% for training and 25% for testing. The model learns on the training set and is subsequently evaluated on the test set. The model's performance is evaluated based on the correct classification of collusive and competitive tenders or markets, achieved by comparing the predictions to the actual classification of each tender. This performance measure is often referred as accuracy which measures the proportion of correct predictions made by the model out of the total number of predictions (Khadka, 2023).

The authors used cross-validation in the training sample to select the best performing model, a common technique in machine learning to evaluate a model's performance by partitioning the data into subsets and using each subset for training and testing (Hastie et al., 2009). Specifically, cross validation was used to identify the optimal weights for each algorithm in the ensemble method, resulting in the selection of the model with the highest performance. Once identified, the optimal weights were assigned to each algorithm, ensuring that the ensemble method would achieve the best possible accuracy.

After selecting the best performing model using cross-validation on the training sample, it was then applied to the test sample to predict whether a tender or market is collusive or not. At first, the model predicts the probability of collusion, and subsequently, a tender is classified as collusive if the collusion probability is greater than or equal to 0.5. By doing so, the correct classification rate can be calculated in the test set, where the predictions match the actual outcomes.

Lastly, the whole procedure of randomly splitting the data into train and test and subsequently evaluating the performance of the optimal model found during training, is repeated 100 times. The correct classification rate of each split is recorded, and afterwards, averaged to finally be used as the ultimate performance measure. This approach provides a more reliable estimate of the model's accuracy and helps to account for any variability that may arise from different random splits of the data.

4.2 Machine learning algorithms used in ensemble method

Next, a brief overview of each ML algorithm which is used in the ensemble method is presented. Firstly, bagged decision trees and random forests are both ensemble methods that use decision trees as their base models. Bagged decision trees involves training multiple decision trees on different subsets of the training data, using sampling with replacement, and then combining their predictions by averaging them to obtain a final prediction. The goal of bagging is to reduce variance in the predictions, which can occur when using a single decision tree. In contrast, random forests use a similar approach to bagging, while on the tree level all features are available, at each split, a new subset of features are randomly sampled, using sampling without replacement. This additional level of randomness can help reduce correlation between trees and improve performance. The final prediction in random forests is made by taking the mode of the predictions from all the decision trees (Breiman, 1996, Breiman, 2001, Hastie et al., 2009, James et al., 2013).

The third model that was included in our ensemble method is based on bayesian additive regression trees (BART) which is a machine learning method that combines decision trees with bayesian modeling. BART model the relationship between the target variable and the input features by building a sum of decision trees, with each tree weighted by its contribution to the prediction (Chipman et al., 2010).

Moving on to the fourth algorithm, Lasso regression is a type of linear regression that adds a penalty term to the objective function. The penalty term helps to shrink the coefficients of some of the features towards zero. This results in a simpler model which includes only the most important features. The strength of the penalty term is controlled by a parameter called lambda, which determines the degree of shrinkage applied to the coefficients (Tibshirani, 1996).

The fifth algorithm is the Support vector machines (SVMs) which is a type of machine learning algorithm used for classification and regression analysis. The goal of SVMs is to find a hyperplane in a high-dimensional space that separates the data into different classes (in this case, collusive or non-collusive) with the largest possible margin between the two classes (Hastie et al., 2009).

Finally, the last algorithm used in the study are neural nets which aim is to fit a system of nonlinear regression functions that flexibly capture relationships. The predictors are used as inputs for intermediate functions known as hidden nodes, which are associated with the outcome. The model's flexibility is determined by the number of hidden nodes and layers, with an increase in parameters leading to reduced bias but increased variance (Ripley, 1996, Hastie et al., 2009).

In summary, this project uses a combination of statistical screens and algorithms to predict whether a tender or market is collusive or not. Through this process, it is possible to identify the optimal ensemble method that provides the highest performance. In a next step, the ensemble method as a machine learner will be estimated for each country/industry data sets which were previously presented.

5 Results

5.1 Model estimation by country or industry

Table 7 displays the correct classification rates overall, under collusion and under competition for all country/industry level data sets. The results were obtained by training the ensemble method on the share of training data and afterwards, evaluating it's performance on previously unseen test data. The results below were obtained by averaging the correct classification rates from 100 simulations of training and testing, which provides a comprehensive evaluation of the model's accuracy and ability to generalize to new data.

Table 7 Correct classifications rates for all data sets

Model	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
Brazil (oil infrastructure)	0.85	0.84	0.86
Brazil (gasoline)	0.88	0.90	0.86
Italy	0.77	0.79	0.76
Japan	0.88	0.89	0.87
Switzerland	0.74	0.63	0.81

Table 7 displays the model's performance on different country/industry data sets [2]. The Brazilian (gasoline) and Japanese cases achieved the highest overall accuracy rates. The model on the Brazilian gasoline cartel obtained an overall accuracy rate of 88%, with correct prediction rates of 90% for collusion and 86% for competition. Similarly, the model on the Japanese cartel achieved an overall accuracy rate of 88%, with correct prediction rates of 89% under collusion and 87% under competition. These results indicate that both models are effective in predicting collusion, although slightly less effective in identifying competitive cases. Also, the overall accuracy for these cartel cases is balanced between correct predictions under collusion and competition. This suggests that the models are not over fitting in one category

[2] A graphical representation of the results is displayed in Figure 3 in the Appendix

and are able to distinguish between collusive and competitive behavior in a effective manner [3].

In third place comes the model that used the Brazilian (oil infrastructure) data set which achieved an overall correct prediction rate of 85% while the model on the Italian data set performed less effectively, with a 77% overall correct prediction rate. For these two cases the accuracy is also balanced between classes which can be explained by the fact that data sets have similar number of collusive and competitive bids/prices. In contrast, the Swiss data set had a lower proportion of collusive cases, resulting in a lower correct prediction rate under collusion (63%) but a higher rate under competition (81%). These findings suggest that the effectiveness of collusive behavior detection models can be influenced by the balance of collusive and competitive bids/prices in the data set.

To summarize, the models performed reasonably well, with correct prediction rates ranging from 74% to 88%. The models on the Brazilian gasoline and Japanese data set had the highest overall correct prediction rates, both achieving a 88%, while the model on the Swiss data set had the lowest overall correct prediction rate of 74%. In terms of identifying collusion, the models perform well, with correct prediction rates starting from 63% in the case of the Swiss cartel to 90% for the Brazilian gasoline cartel. Overall, the results suggest that the models are effective in predicting cartel cases. However, there is some variability in performance across different countries/industries, indicating that factors such as legal frameworks, industry characteristics, and other contextual factors may play a role in the effectiveness of the models

Additionally, as a further performance measure, for each data set, the mean squared error (MSE) of the six machine learning algorithms used in the ensemble method was estimated and are shown in Table 8. The MSE values indicate the average squared difference between the predicted and actual values, with lower values indicating better performance. From Table 8, we can see that the performance of the machine learning algorithms varies across data sets. Firstly, for both Brazilian data sets, Lasso regression performed better than other algorithms as it exhibited the lowest MSE values while neural nets have the highest. Regarding the Italian data set, bagged decision

[3] The underlying model can be modified by changing the threshold used to classify a tender or market as collusive or competitive. This can help to avoid false positives or false negatives, depending on the cost that is associated with each. Such metrics are important in this context given that false positives incorrectly identify a tender or market as collusive when it is actually a competitive transaction. The findings can result in unnecessary investigations, legal actions, or other negative consequences for the parties involved. On the other hand, a false negative occurs when the model fails to identify a collusive tender or market when it actually is collusive. As a result, anti competitive behavior gets unnoticed.

Table 8 Mean squared error of the machine learning algorithms

Data set	Random forest	Bayesian additive regression trees	Bagged decision trees	Lasso regression	Support vector machines	Neural nets
Brazil	0.13	0.14	0.13	0.12	0.15	0.26
Brazil (gasoline)	0.19	0.20	0.16	0.09	0.21	0.22
Italy	0.19	0.20	0.19	0.23	0.18	0.25
Japan	0.03	0.18	0.25	0.38	0.17	0.00
Switzerland	0.17	0.17	0.18	0.18	0.19	0.24

trees and support vector machines had the lowest MSE while Lasso regression has the highest. For the Japanese data set, the random forest algorithm had the lowest MSE value, while Lasso regression had the highest and notably, the neural nets algorithm had an MSE value of 0. Finally, for the Swiss data set, most algorithms had roughly the same level of MSE except for the neural nets algorithm which was higher. The results suggests that depending on the data set, different ML algorithms may perform better or worse.

Moreover, to inspect the importance of the different ML algorithms used in the ensemble method, For each country/industry data set, I examine the average weight assigned to each machine learning algorithm.

Table 9 Average weight of the machine learning algorithms

Data set	Random forest	Bayesian additive regression trees	Bagged decision trees	Lasso regression	Support vector machines	Neural nets
Brazil	0.06	0.00	0.76	0.00	0.18	0.00
Brazil (gasoline)	0.00	0.00	0.03	0.91	0.00	0.06
Italy	0.12	0.00	0.40	0.00	0.47	0.00
Japan	0.00	0.09	0.09	0.07	0.08	0.07
Switzerland	0.54	0.32	0.00	0.13	0.10	0.00

Table 9 shows that the weights assigned to each machine learning algorithms vary greatly depending on the data set being analyzed. For instance, in the first Brazilian data set, bagged decision trees were given the highest weight, while in the second Brazilian data set, lasso regression was found to be most effective and therefore, was given a significant weight. In the case of Italy, support vector machines had the highest weight, followed by bagged decision trees. In the Japanese case, the weights were distributed relatively evenly across the different algorithms, with the exception of random forest, which received no weight. Lastly, for the Swiss case, random forest and Bayesian additives regression trees received the most importance. In summary, Table 9 suggests that different machine learning algorithms may be better suited for different data sets and contexts. Also, it highlights that bagged decision trees were consistently given some weight while neural nets were often given less weight or excluded.

Moreover, to ensure the validity and reliability of the results obtained above, two robustness checks were conducted. The first check involved using demeaned screens as inputs in the ML models, the results are displayed in Table 10.

Table 10 Correct classifications rates with demeaned screens

Model	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
Brazil (oil infrastructure)	0.80	0.57	0.92
Brazil (gasoline)	0.61	0.42	0.79
Italy	0.60	0.62	0.56
Japan	0.84	0.87	0.82
Switzerland	0.73	0.62	0.81

From Table 10, it can be seen that for all cases, the models performed significantly worse. For the Brazilian oil infrastructure case, the correct prediction overall reduced from 85% to 80% while in the case of the Brazilian gasoline cartel it decreased from 88% to 61%, exhibiting a significant reduction in performance. In addition, for the Italian case, the model's performance reduced from 77% to 60% which is also a significant reduction. Moreover, for the Japanese model, it's performance reduced from 88% to 84%. Lastly, the model on the Swiss cartel data reduced it's accuracy from 74% to 73%. These findings suggest that demeaning the screens does not lead to any improvements in the model's predictions and that the original screens may already be providing a clear and accurate representation of the data.

The second robustness check involved adding as an additional screen, precisely, the number of bids/prices per tender/market. The aim was to investigate whether this additional screen had a significant influence on the accuracy of the models. The results of this check are provided in Table 11.

The results from Table 11 suggest that there is evidence that for the Brazilian (oil infrastructure), Italian, and Japanese data sets, the inclusion of the number of bids

Table 11 Correct classifications rates with number of bids per tender/market as an additional screen

Model	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
Brazil (oil infrastructure)	0.88	0.91	0.84
Brazil (gasoline)	0.88	0.90	0.87
Italy	0.81	0.82	0.81
Japan	0.96	0.96	0.94
Switzerland	0.74	0.65	0.80

per tender/market significantly improved the model's performance [4]. This suggests that the number of bids per tender is an important factor that contributes significantly to the detection of collusive behavior in these markets. Moreover, in the case of the Swiss and Brazilian gasoline cartel, the inclusion of the new screen also led to slight improvements in the models' performance, although the effect was not as significant as in the other data sets. This may be because the number of bids per tender was not as strongly correlated with the presence of collusive behavior in these markets.

In conclusion, the robustness checks performed provide evidence for the sensibility of some data sets to different specifications. On one hand, when using the demeaned screens as inputs, the models performed worse [5]. On the other hand, adding the number of bids/prices, most models led to an improvement in performance or remained stable. This suggests that the number of bids or prices may be informative in predicting cartel activity. Overall, these findings highlight the importance of exploring innovative screens when applying machine learning techniques to detect cartels.

Finally, for a last estimation step, all the data sets were merged into a unified data set which was later used to train and test the ensemble method. Such combined data set had 3997 tenders/markets in total of which 1705 were collusive and 2292 were competitive. For this estimation, the number of price/bids was added as an additional screen and the share of trained data was increased from 75% to 85%. The results of the analysis are presented below.

The findings from Table 12 suggests that the ensemble method performed reasonably well on the combined data set, achieving an overall correct classification rate of 76%.

[4] A graphical representation of the results is displayed in Figure 3 in the Appendix

[5] A further robustness check was implemented where the statistical screens were demeaned and standardized by dividing them by their respective standard deviation. However the results showed no improvement in accuracy across any of the data sets.

Table 12 Correct classifications rates on combined data

Model	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
Combined data	0.76	0.71	0.79

Additionally, the model obtained a lower correct classification rate for collusive instances (71%) than for competitive ones (79%). One explanation for this issue is that the combined data set has a larger proportion of competitive tender/markets and therefore the models learns how to identify competitive instances better. Additionally, to evaluate the model's performance, the MSE and the weight of each algorithm are presented below.

Table 13 Mean squared error of the machine learning algorithms

Data set	Random forest	Bayesian additive regression trees	Bagged decision trees	Lasso regression	Support vector machines	Neural nets
Combined data	0.17	0.18	0.19	0.21	0.20	0.25

From Table 13, the random forest algorithm had the lowest MSE of 0.17, followed closely by Bayesian additive regression trees. Bagged decision trees had an MSE of 0.19, while lasso regression and support vector machines have MSEs of 0.21 and 0.20, respectively. Neural nets had the highest MSE of 0.25, indicating the poorest performance among the six algorithms. Overall, the results suggest that random forest and Bayesian additive regression trees are the best performing algorithms on this mixed data set, while Neural nets may not be the most suitable algorithm for the unified data.

Table 14 Average weight of the machine learning algorithms

Data set	Random forest	Bayesian additive regression trees	Bagged decision trees	Lasso regression	Support vector machines	Neural nets
Combined data	0.99	0.00	0.00	0.00	0.00	0.01

Based on Table 14, it can be seen that the random forest algorithm received a nearly unanimous importance weight, indicating that it was the algorithm that contributed the most to the model's performance. On the contrary, the neural nets algorithm was assigned a much lower weight. All other algorithms received a weight of 0, suggesting that they did not significantly contribute to the model. Additionally, two robustness checks of the ensemble method on the combined data set were implemented. On one hand, incorporating the demeaned screens did not result in any improvement in performance. However, when the number of bids was included in the model as an additional screen, the performance improved, as previously observed with other models [6].

Lastly, a final robustness check was performed by modifying each country/industry level data sets to achieve a perfect balance of 50% collusive and 50% competitive tenders/prices and subsequently combining all data sets. This was done to ensure that the results were not driven by any imbalance in the original data sets. The new combined data set consisted of 3634 observations with an equal number of collusive and competitive observations instances, specifically 1817 each. The results obtained are presented in Table 15.

Table 15 Correct classifications rates on perfectly balanced combined data

Model	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
Combined data	0.80	0.80	0.79

The results from the perfectly balanced data set from Table 15 show that the correct prediction rate has increased. Comparing Table 12 and Table 15, it is clear that the balanced data set in Table 15 resulted in better performance for the model. Table 15 shows that the model was able to achieve a higher overall correct classification rate (80%) than the model in Table 12 (76%). Additionally, the model in Table 15 performed well on both collusive and competitive cases, with correct prediction rates of 80% and 79%, respectively. In contrast, the model in Table 12 had a lower correct prediction rate for collusive cases (71%) than competitive cases (79%). These findings suggest that class imbalance can have a negative impact on model performance, and balancing the data set can improve the accuracy of the model's predictions [7].

In summary, this section has provided an in-depth analysis of the performance of the ensemble method at predicting collusive and competitive instances for each coun-

[6] The results obtained from implementing two robustness checks are shown in Table 23 and 24 in the Appendix.

[7] The findings from demeaning and standardizing the screens did not improve the models performance when estimated on the perfectly balanced combined data set.

try/industry case as well as on the combined data set. The country level estimations demonstrated correct classification rates, ranging from 74% to 88% and from 74% to 96% when adding an additional screen. However, there was some variability in performance across different countries, suggesting that contextual factors such as legal frameworks and industry characteristics may influence the models' effectiveness. Additionally, the findings highlighted that the choice of machine learning algorithms may vary depending on the data set and context. Nevertheless, bagged decision trees consistently received some weight across the different country data sets.

Furthermore, the robustness checks conducted in this study yielded interesting results. Demeaning the screens did not improve the model's predictions, implying that the original screens already provided a clear and accurate representation of the data. On the other hand, the inclusion of the number of bids per tender as a screen significantly contributed to the detection of collusive behavior in the markets that were analyzed.

In the case on the combined data set, the ensemble method achieved a 76% overall correct classification rate and of 80% when trained and tested on a perfectly balanced data set. The model performed worse at predicting collusive than competitive instances in the combined data set (71% vs 79%) which can be related to the class imbalance in the data set. However, when the ensemble method was trained on the perfectly balanced data set, the model achieved a much higher accuracy when predicting collusive instances (80% vs 79%). Overall, the ensemble method demonstrated promising performance in predicting collusion and competition in the combined data set, with a significantly higher accuracy achieved when trained and tested on a perfectly balanced data set.

Regarding the relative importance of the ML algorithms, for the case of the combined estimation, the random forest algorithm has the lowest MSE of 0.17 and received a significant weight, therefore it was the algorithm that contributed the most to the model's performance. Lastly, demeaning the screens did not improve performance while including the number of bids increased accuracy, as was previously observed with other models. Ultimately, these results offer valuable insights into the effectiveness of machine learning techniques across industries and countries and underscore the importance of addressing class imbalance in the data sets to improve model performance.

The subsequent section delves deeper into the transferability of the ensemble method to different contexts by employing a rotating scheme. In this scheme, each country data set is temporarily excluded and later used to test the model which was trained on the remaining data sets. This is a the most challenging test given that the data sets come from different industries, countries and therefore, contexts.

5.2 Model estimation: Rotating scheme

The objective of this section is to evaluate the transferability of machine learning models across different industries and countries. Initially, a unified data set is created by combining all the available data sets, with only tenders/markets having more than 4 bids/prices being included [8]. Next, one country data set is withheld for testing purposes. Subsequently, an ensemble method as machine learner is applied on the unified data set. Such trained model is then used to predict the nature of tenders (collusive or competitive) on the previously unseen test data set. In a last step, the model performance is evaluated by comparing the actual and predicted outcomes overall, under collusion and under competition. Through this approach, the adaptability and robustness of the machine learning models in different settings can be assessed. Note that the number of bids/prices was included as an additional screen in all the following models [9].

Table 16 Correct classifications rates: training on mixed data and testing on unseen data

Test data set	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
Brazil (oil infrastructure)	0.66	0.00	1.00
Brazil (gasoline)	0.48	0.99	0.01
Italy	0.50	0.02	1.00
Japan	0.68	0.92	0.38
Switzerland	0.56	0.02	0.99

The results in Table 16 suggest that the ensemble method used as a machine learner did not perform well when trained on a combined data set and tested on previously unseen data. The first row of results indicate that the model always predicts competition for the Brazilian oil infrastructure data set. This shows that the model cannot capture the complexity of the data which may be due to the imbalances in classes however, there was not a significant difference between the number of collusive and competitive instances in the data set.

[8] Now the data set contains a total of 3202 observations, of which 47% are classified are collusive and 53% as competitive

[9] The models that were trained and tested on the combined data sets (composed of N-1 data sets) as well as on the perfectly balanced data sets were also evaluated, their results and performance measures are shown in Tables 25-28 in the Appendix.

On the other hand, in the case of the Brazilian gasoline market, the model always predicts the collusive class and very rarely the competitive class, suggesting poor performance. For the Italian case also, the model always predicts competition which can also be explained by the class imbalance in the data set. In the case of Japan, the models struggled to predict competitive instances while achieving a high accuracy in prediction collusive tenders. Finally, for the Swiss case, the model consistently predicted competitive cases and therefore, performed poorly in predicting collusive tenders.

In general, the model appears to lack the ability to perform well when trained on a combined data set and tested on unseen data. The model appears to struggle with capturing the complexity of the data and may have been affected by class imbalances in the data sets. To address the issue of class imbalance, I applied the ensemble method on the perfectly balanced data sets used earlier for the combined data set estimations. However, the results indicated that the model's accuracy did not improve significantly in most cases, except for the Swiss data set. For such case, the overall correct classification rate was more balanced, but still relatively low, with values of 62%, 40% and 84% for collusive, competitive tenders/markets and overall, respectively [10].

To ensure the reliability of the results and examine the impact of new specifications on the performance of the model, two additional robustness checks were conducted. The first involved demeaning the screens while the second involved both demeaning and normalizing them. Demeaning the screens did not lead to any increase in accuracy [11]. However, when the screens were demeaned and standardized, the model improved, Table 16 presents the results obtained from the second robustness check.

Table 17 Correct classifications rates (demeaned & standardized screens): training on mixed data and testing on unseen data

Test data set	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
Brazil (oil infrastructure)	0.75	0.37	0.95
Brazil (gasoline)	0.49	0.92	0.08
Italy	0.55	0.48	0.62
Japan	0.67	0.67	0.67
Switzerland	0.57	0.53	0.59

10 The complete results from the estimations on the perfectly balanced data are shown in Table 29 in the Appendix.

11 The results from demeaning the screens are shown in Table 30 in the Appendix.

The results presented in Table 17 demonstrate that when the screens are demeaned and standardized, the model performs better overall and has a far more balanced correct prediction rate between collusion and competition. These findings are in contrast to the results obtained from Table 16. For instance, for the Brazilian (oil infrastructure) test data set, the results in Table 16 report an overall correct prediction rate of 0.66, while the results from Table 17 report a higher rate of 0.75 which is also more balanced between class predictions. In summary, the results suggest that demeaning and standardizing the screens enhances the model's accuracy and improves the balance of predictions across different classes [12].

[12] As a further robustness check, the ensemble method with demeaned and standardized screens was applied using the perfectly balanced data sets. However, the accuracy only improved in the case of the Brazilian gasoline case. The detailed results are shown in Table 31 in the Appendix

6 Conclusion

This study aimed to assess the correct classification rate of an ensemble method at predicting collusive behavior across different countries and industries. The results showed that the ensemble method performed well for country level estimations, achieving an overall correct classification rate of 74% to 88%, and up to 96% when an additional screen was added. However, there was some variability in performance across different countries, indicating that contextual factors may influence the model's effectiveness.

Moreover, the research found that no single algorithm consistently received a significant weight, and the choice of machine learning algorithms may vary depending on the data set and context. Despite this, the ensemble method based on statistical screens of bid distributions remains a powerful tool for detecting collusive behavior when trained and tested on data from the same country, demonstrating its usefulness in specific contexts.

From the results of the rotating scheme, the model did not perform well when trained on a combined data set and tested on unseen data. The model appeared to struggle with capturing the complexity of the data, indicating that the method may not be transferable to different contexts. However, it was evidenced that demeaning and standardizing the screens enhanced the model's accuracy and improved the balance of predictions across different classes. Additionally, when applying the ensemble method on different combinations of the data sets after leaving out one country's data, the random forest algorithm consistently received the highest weight, indicating that such algorithm may be particularly useful in collusion detection.

In addition, there are certain practical limitations that need to be taken into account when using the ensemble method to detect collusive behavior. For instance, the lack of data containing the winning and losing bids/prices for tenders in proven cartel cases can hinder the calculation of statistical screens. This issue is compounded by the fact that the ensemble method requires access to a large historical data set to effectively learn and predict accurately. However, despite these challenges, the method examined in this study is one of the most parsimonious in the literature on collusion detection given that it only requires the submitted losing and winning bids/prices and no additional information about the tender or market. Another strength of the model is that the method is data driven which directly learn from the patterns in the data in order to identify the accurate combination of individual indicators.

Also, it is important to acknowledge that the data sets used by the machine learner are derived from proven cartel cases, which may introduce a potential bias towards certain types of cartels and their behavior. Therefore, the ensemble method may not be as effective in detecting other types of cartels, and the results may not be representative to all collusive behaviors.

In summary, this study provides valuable insights into the effectiveness and transferability of machine learning methods in detecting collusive behavior in specific contexts. Further research is needed to explore the transferability of the ensemble method to different industries and countries, taking into account the heterogeneity of collusive behavior.

A Appendix

Table 18 Means of screens by collusion status from Brazil

Variable	Cartel period	Non cartel period
Variance	2.34e+16	1.27e+16
Spread	0.17	0.69
Kurtosis	1.78	2.39
DiffFlowEst	1.25e+08	4.60e+07
DiffPercent	0.07	0.11
RelDist	7.53	0.89
RelDistAlt	5.79	1.21
Skewness	-0.19	0.28
KolSmir	16.48	8.78

Table 19 Means of screens by collusion status from Brazil (gasoline)

Variable	Cartel period	Non cartel period
Variance	0	0
Spread	0.02	0.02
Kurtosis	1.86	1.9
DiffFlowEst	0.01	0.03
DiffPercent	0.01	0.01
RelDist	0.03	0.06
RelDistAlt	0.03	0.06
Skewness	-0.1	-0.03
KolSmir	1086.69	298.67

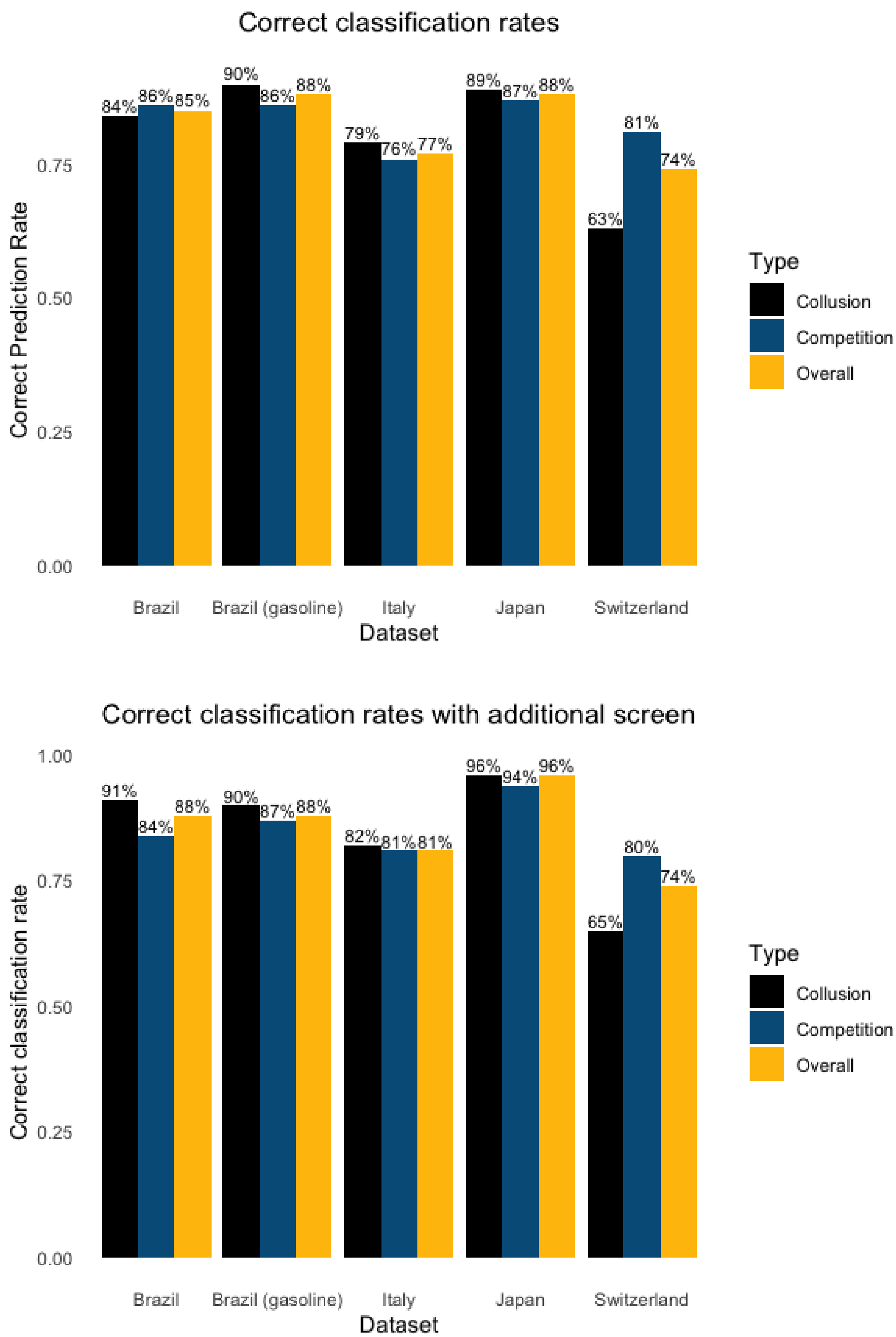


Figure 3 Correct classification rates

Table 20 Means of screens by collusion status from Italy

Variable	Cartel period	Non cartel period
Variance	6.31e+12	9.90e+12
Spread	0.18	0.25
Kurtosis	5.63	5.55
DiffFlowEst	668784.1	737954.5
DiffPercent	0.02	0.02
RelDist	0.76	0.44
RelDistAlt	7.78	8.5
Skewness	0.60	0.84
KolSmir	42.29	33.71

Table 21 Means of screens by collusion status from Japan

Variable	Cartel period	Non cartel period
Variance	6.61e+12	3.41e+13
Spread	0.04	0.17
Kurtosis	3.24	3.89
DiffFlowEst	4.71e+05	1.18e+06
DiffPercent	0.01	0.02
RelDist	1.63	0.75
RelDistAlt	5.10	2.58
Skewness	-0.56	0.47
KolSmir	243.46	49.67

Table 22 Means of screens by collusion status from Switzerland

Variable	Cartel period	Non cartel period
Variance	1,214,538,490	29,688,402,098
Spread	93.06	1.66
Kurtosis	2.08	1.86
DiffFlowEst	11,900.43	42,674.07
DiffPercent	80.19	1.16
RelDist	1.75	6.52
RelDistAlt	2.22	5.21
Skewness	0.00	0.12
KolSmir	31.8	21.70

Table 23 Correct classifications rates with demeaned screens on combined data

Model	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
Combined data	0.70	0.60	0.78

Table 24 Correct classifications rates with number of bids per tender/market as screen on combined data

Model	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
Combined data	0.75	0.64	0.84

Table 25 Correct classifications rates when training on N-1 data sets

Data set	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
All - Brazil oil infrastructure	0.72	0.63	0.80
All - Brazil gasoline	0.79	0.73	0.84
All- Italy	0.80	0.74	0.85
All - Japan	0.71	0.69	0.73
All - Switzerland	0.81	0.81	0.80

Table 26 Correct classifications rates when training on N-1 data sets on perfectly balanced data sets

Data set	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
All - Brazil oil infrastructure	0.72	0.63	0.80
All - Brazil gasoline	0.79	0.73	0.84
All- Italy	0.80	0.74	0.85
All - Japan	0.71	0.69	0.73
All - Switzerland	0.80	0.79	0.81

Table 27 Mean squared error of the machine learning algorithms when training on N-1 data sets

Data set	Random forest	Bayesian additive regression trees	Bagged decision trees	Lasso regression	Support vector machines	Neural nets
All - Brazil	0.16	0.17	0.19	0.21	0.18	0.25
All - Brazil	0.17	0.17	0.17	0.20	0.18	0.25
All - Italy	0.16	0.17	0.18	0.20	0.18	0.25
All - Japan	0.18	0.19	0.19	0.22	0.28	0.25
All - Switzerland	0.15	0.16	0.13	0.20	0.18	0.25

Table 28 Average weight of the machine learning algorithms when training on N-1 data sets

Data set	Random forest	Bayesian additive regression trees	Bagged decision trees	Lasso regression	Support vector machines	Neural nets
All - Brazil	0.75	0.20	0.00	0.00	0.05	0.00
All - Brazil	0.46	0.48	0.00	0.00	0.06	0.00
All - Italy	0.94	0.06	0.00	0.00	0.00	0.00
All - Japan	1.00	0.00	0.00	0.00	0.00	0.00
All - Switzerland	0.00	0.00	1.00	0.00	0.00	0.00

Table 29 Correct classifications rates: training on perfectly balanced mixed data and testing on perfectly balanced unseen country/industry data

Test data set	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
Brazil (oil infrastructure)	0.50	0.00	1.00
Brazil (gasoline)	0.50	0.99	0.01
Italy	0.50	0.01	1.00
Japan	0.63	0.91	0.34
Switzerland	0.62	0.40	0.84

Table 30 Correct classifications rates: training on mixed data and testing on unseen country/industry data with demeaned screens

Test data set	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
Brazil (oil infrastructure)	0.66	0.00	1.00
Brazil (gasoline)	0.48	0.96	0.04
Italy	0.51	0.01	1.00
Japan	0.68	0.93	0.37
Switzerland	0.54	0.23	0.79

Table 31 Correct classifications rates (demeaned & standardized screens): training on perfectly balanced mixed data and testing on perfectly balanced unseen country/industry data

Test data set	Correct Prediction Rate (Overall)	Correct Prediction Rate (Collusion)	Correct Prediction Rate (Competition)
Brazil (oil infrastructure)	0.61	0.42	0.79
Brazil (gasoline)	0.50	0.97	0.03
Italy	0.55	0.30	0.80
Japan	0.63	0.60	0.67
Switzerland	0.56	0.58	0.54

Bibliography

Abrantes-Metz, R. M., Kraten, M., Metz, A. D., & Seow, G. S. (2012). Libor manipulation? *Journal of Banking Finance*, *36*(1), 136–150. https://doi.org/https://doi.org/10.1016/j.jbankfin.2011.06.014

Adam, I., Fazekas, M., Kazmina, Y., Teremy, Z., Tóth, B., Villamil, I. R., & Wachs, J. (2022). *Public procurement cartels: A systematic testing of old and new screens* (Working Paper Series GTI-WP/2022:01). Government Transparency Institute. Budapest, Hungary. https://www.govtransparency.eu/wp-content/uploads/2022/03/GTI-WP-Cartel_20220907.pdf

Aghion, P., Bechtold, S., Cassar, L., & Herz, H. (2014). The causal effects of competition on innovation: Experimental evidence. *Working Paper*. https://scholar.harvard.edu/files/aghion/files/causal_effects_of_competition.pdf

Bergman, M. A. (2019). How and when consumer choice drives efficient competition in publicly funded services – Note by Mats. A Bergman.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *Annals of Applied Statistics*, *4*(1), 266–298.

Conley, T. G., & Decarolis, F. (2016). Detecting bidders groups in collusive auctions. *American Economic Journal: Microeconomics*, *8*(2), 1–38. https://doi.org/10.1257/mic.20130254

European Commission. (2021). Notice on tools to fight collusion in public procurement and on guidance on how to apply the related exclusion ground. *Official Journal of the European Union*, *C 91*(2), 1–2. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52021XC0318(01)&from=EN

Fazekas, M., Tóth, B., & Wachs, J. (2023). *Public procurement cartels: A large-sample testing of screens using machine learning* (Working Paper GTI-WP/2023:02). Government Transparency Institute. Budapest, Hungary. https://gti.org.hu/wp-content/uploads/2023/04/GTI-WP_2023_02.pdf

García Rodríguez, M. J., Rodríguez-Montequín, V., Ballesteros-Pérez, P., Love, P. E., & Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction*, *133*, 104047. https://doi.org/https://doi.org/10.1016/j.autcon.2021.104047

Géron, A. (2017). *Hands-on machine learning with scikit-learn and tensorflow: Concepts, tools, and techniques to build intelligent systems* (1st ed.). O'Reilly Media.

Harrington, J., Joseph E. (2006). Behavioral screening and the detection of cartels [To be published in Claus-Dieter Ehlermann and Isabela Atanasiu (eds.), European Competition Law Annual 2006: Enforcement of Prohibition of Cartels]. *EU Competition Law and Policy Workshop/Proceedings.*

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd). Springer Series in Statistics.

Huber, M., & Imhof, D. (2019). Machine learning with screens for detecting bid-rigging cartels. *International Journal of Industrial Organization*, *65*, 277–301. https://doi.org/10.1016/j.ijindorg.2019.04.002

Huber, M., Imhof, D., & Ishii, R. (2022). Transnational Machine Learning with Screens for Flagging Bid-Rigging Cartels. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(3), 1074–1114. https://doi.org/10.1111/rssa.12811

Imhof, D., Karagök, Y., & Rutz, S. (2018). Screening for bid rigging - does it work? *Journal of Competition Law & Economics*, *14*(2), 235–261. https://doi.org/10.1093/joclec/nhy006

Ishii, R. (2014). Bid roundness under collusion in japanese procurement auctions. *Review of Industrial Organization*, *44*, 241–254. https://doi.org/10.1007/s11151-013-9408-6

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r* (Vol. 112). Springer Science+Business Media. https://doi.org/10.1007/978-1-4614-7138-7

Khadka, N. (2023). Adaboost algorithm. https://dataaspirant.com/adaboost-algorithm/

OECD. (2020). Abuse of dominance in digital markets. https://www.oecd.org/daf/competition/abuse-of-dominance-in-digital-markets-2020.pdf

OECD. (2022). Data screening tools in competition investigations. www.oecd.org/daf/competition/data-screening-tools-in-competition-investigations-2022.pdf

Porter, R. H. (2005). Detecting collusion. *Review of Industrial Organization*, *26*(2), 147–167. https://doi.org/10.1007/s11151-004-7294-7

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge University Press.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Signor, R., & Ika, L. (2020). White collar crime: Unearthing collusion in the procurement of infrastructure projects. *IEEE Transactions on Engineering Management*, *PP*, 1–12. https://doi.org/10.1109/TEM.2020.2994636

Silveira, D., Vasconcelos, S., Resende, M., & Cajueiro, D. O. (2022). Won't get fooled again: A supervised machine learning approach for screening gasoline cartels. *Energy Economics*, *105*, 105711. https://doi.org/https://doi.org/10.1016/j.eneco.2021.105711

The Danish Competition and Consumer Authority. (2022). Collusion detection in public procurement using computational methods. (56). https://www.kfst.dk/en/news/2022/collusion-detection-in-public-procurement-using-computational-methods/

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Toth, B., Fazekas, M., Czibik, A., & Toth, I. J. (2015). *Toolkit for detecting collusive bidding in public procurement: With examples from hungary* (Working Paper). Corruption Research Center Budapest. Budapest, Hungary. http://www.crcb.eu/

Wallimann, H., Imhof, D., & Huber, M. (2022). A machine learning approach for flagging incomplete bid-rigging cartels. *Computational Economics*. https://doi.org/10.1007/s10614-022-10315-w

WorldBank. (2022). The hidden $1 trillion: Halting waste in public procurement. *World Bank*. https://blogs.worldbank.org/voices/hidden-1-trillion-halting-waste-public-procurement

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. CRC Press.

Declaration of Authorship

I hereby declare that I wrote this thesis on my own and followed the principles of scientific integrity.

I acknowledge that otherwise the department has, according to a decision of the Faculty Council of November 11th, 2004, the right to withdraw the title that I was conferred based on this thesis.

I confirm that this work or parts thereof have not been submitted in this form elsewhere for an examination, according to a decision of the Faculty Council of November 18th, 2013.

Fribourg, 09.05.2023

Ana Paula Armendariz Pacheco