NBER WORKING PAPER SERIES

IS MODEL ACCURACY ENOUGH?
A FIELD EVALUATION OF A MACHINE LEARNING MODEL
TO CATCH BOGUS FIRMS

Taha Barwahwala
Aprajit Mahajan
Shekhar Mittal
Ofir Reich

Working Paper 32705
http://www.nber.org/papers/w32705

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2024

J-PAL South Asia has been instrumental in getting this study off the ground, and we are grateful for their support in navigating the tax department partnership. Aparna Krishnan, Sharanya Chandran, Vaishnavi Mulay, Piyali Chatterjee and the rest of the policy team championed the project with the state leadership at every turn. Mihir Bhaskar and Suddhasatwa Bhattacharya were exceptional and tireless in managing the multitude of stakeholder relationships and administration matters. We would like to thank the efforts of the leadership at the taxation department who supported the study especially through a challenging time during the pandemic. This project would not have been possible without the generosity of tax inspectors at the department headquarters as well as various district offices for their plentiful time in helping us develop an understanding of the legal, procedural and technical aspects of revenue recovery. We are especially grateful to the technical services team at the department for providing data and helping us navigate the technical infrastructure of the Goods and Services Tax Network. This study was supported by funding from the taxation department and the International Center for Tax & Development through grant 22/573 GV/18012/4H/2W. This work began prior to author SM joining Amazon Inc. The views expressed herein are those of the author(s) and can not be attributed to Amazon Inc., its Executive Boards, or management teams, or the National Bureau of Economic Research.

Is Model Accuracy Enough? A Field Evaluation Of A Machine Learning Model To Catch Bogus Firms
Taha Barwahwala, Aprajit Mahajan, Shekhar Mittal, and Ofir Reich
NBER Working Paper No. 32705
July 2024
JEL No. H0,O10

ABSTRACT

We investigate the use of a machine learning (ML) algorithm to identify fraudulent non-existent firms. Using a rich dataset of tax returns over several years in an Indian state, we train an ML-based model to predict fraudulent firms. We then use the model predictions to carry out field inspections of firms identified as suspicious by the ML tool. We find that the ML model is accurate in both simulated and field settings in identifying non-existent firms. Withholding a randomly selected group of firms from inspection, we estimate the causal impact of ML driven inspections. Despite its strong predictive and field performance, the model driven inspections do not yield a significant increase in enforcement as measured by the cancellation of fraudulent firm registrations and tax recovery. We provide two rationales for this discrepancy based on a close analysis of the tax department's operating protocols: selection bias, and institutional friction in integrating the model into existing administrative systems. Our study serves as a cautionary tale for the application of machine learning in public policy contexts and of relying solely on test set performance as an effectiveness indicator. Field evaluations are critical in assessing the real-world impact of predictive models.

Taha Barwahwala
Columbia University
taha17.kb@gmail.com

Aprajit Mahajan
Dept. of Agricultural & Resource Economics
University of California, Berkeley
219 Giannini Hall
Berkeley, CA 94720-3310
and NBER
aprajit@gmail.com

Shekhar Mittal
Amazon Inc.
shekhar.mittal@gmail.com

Ofir Reich
Independent Data Scientist
e.scribbler@gmail.com

1 Introduction

Improving the state's ability to tax effectively is increasingly seen as central to the development process. In 2017, the federal and state governments of India together launched the Goods and Service Tax (GST), a unified Value Added Tax (VAT) regime, to improve their ability to raise revenues.[1] However, the rollout of the GST appears to have been accompanied by a number of technical and design problems. There has been a proliferation of firms claiming fraudulent tax credits against "ghost" purchases. Such purchases are facilitated by "bogus" firms – firms that exist only on paper with no physical presence. Although various changes to the process of claiming tax credits have been implemented by the government, the problem of bogus firms remains widespread.[2] Consequently, identifying bogus firms is an important task for tax authorities both in India and elsewhere.

The increased use of bogus firms to originate and pass-through tax credit has been noted in the popular press and is a pressing policy concern.[3] The current approach towards identifying bogus firms is time-consuming and takes place in a context of strong capacity constraints. First, a ward-level tax official expends considerable time and effort to manually filter information from a variety of sources to identify firms for physical verification. Second, the rapidly expanding taxpayer base[4] and the simultaneous increase in tax department duties[5] has further increased the burden on tax officials' limited time. In such a setting, developing approaches to improve the detection rate for bogus firms is evidently of first-order importance.

In this paper we present evidence from field inspections on the use of machine learning methods to detect bogus firms. In particular, we ask whether ML methods can improve upon existing methods on this task and also identify more bogus firms (that are missed by tax officials). We begin by building a machine learning tool adapted from existing research on detecting bogus firms based on administrative data. We next evaluate the predictions of the ML tool, first on the labeled set, and then by carrying out physical field inspections of firms with a high model-predicted likelihood of being bogus, (henceforth "suspicious firms"). The inspections found 53% of the firms identified by the ML tool as suspicious to be non-existent at their registered address (i.e. bogus). This rate compares favorably to two natural benchmarks which we estimated using physical inspections: it is substantially higher than both the hit rate of 4% from randomly inspecting firms and the hit rate of 38% for inspections based on a rule-of-thumb based approach typically employed by the tax department.

We next assess the causal effect of these physical inspections on the formal registration status of suspicious firms. We view this as the first stage in assessing the effect of the tool on the tax authority's enforcement activities. In particular, we identify the extent to which the use of the ML tool increased firm cancellations i.e. led to the removal of bogus firms from the tax authority's database of firms eligible to file returns. We use a randomized holdout sample to make this comparison. The holdout sample is a random sample of firms from the ML tool's list of suspicious firms that were not provided to the inspec-

[1] See IMF (2018)

[2] See Bureau (2024)

[3] See PTI (2021), Behl (2021a) and Behl (2021b)

[4] Total registered firms and active return filers from GSTN (2024). The number of registered firms increased by ∼42% in our state from July-2017 to April-2020

[5] See Figure A.12 in the Appendix

tors. The cancellation rate of firms in this holdout sample can credibly be compared to the cancellation rate of firms whose names were provided to inspectors. We find that the two cancellation rates are not significantly different from each other. Thus, we conclude that the ML tool based physical inspections and identification exercise did not lead to increased or earlier removal of firms from the authority's list of legitimate firms.

Following identification, tax authorities are interested in revenue recovery. However, revenue recovery from bogus firms is not straightforward. Bogus firms are hard to track down as by definition they have no physical presence. The key targets for revenue recovery are those firms that claim illegitimate input tax credits, i.e. by claiming purchases from the bogus firms. We refer to these firms as beneficiary firms. We find that these beneficiary firms are often registered in a different tax jurisdiction from that of their partner bogus firms. The recovery process therefore depends critically on tax officials sharing information about beneficiary firms with concerned officials in the relevant jurisdictions and efficiently coordinating across jurisdictions (see Appendix B.5).

Given the lack of effect on cancellations, it is perhaps unsurprising that the use of the ML tool for targeting physical inspections did not have any impact on beneficiary firms i.e. we do not observe any changes in enforcement against firms that transacted with the firms on the ML tool list provided for inspections. This appears to be, at least in part, because the tax authority does not have a clearly articulated process for tracking such firms. Based on this experience we suggest simple technology upgrades to improve tax department performance on this front by improving information sharing and coordination across jurisdictions.

Finally, we speculate on the extent to which technology more broadly can enable revenue recovery after the tax department identifies the beneficiaries from a detected bogus firm. We suggest a number of measures that should improve monitoring by targeting information to officials in the relevant jurisdictions.

This paper contributes most directly to a growing literature on the use of technology to improve tax collections in developing countries. Mittal and Mahajan (2017) find that an improvement in the tax authority's technology to automatically detect and investigate broad-based evasion only improved collections among a subset of firms receiving increased regulatory attention. This last finding is consistent with our conclusion here that the use of the ML tool in improving tax department efficacy requires that the tool be viewed as a complement to, rather than a substitute for, tax authorities' ongoing efforts. In earlier work (under a different tax regime) Mittal et al. (2018) propose the use of ML to improve identification of bogus firms and evaluate its potential through a cross-validation exercise. In this paper, we adapt their work to a new data-set and obtain a field-based evaluation of ML based inspections.[6] Our work is also closely related to Battaglini et al. (2022) who develop an ML tool to improve audit targeting in Italy and estimate that implementing their tool in the field would improve detection rates by 38%. More generally, there is a body of work in the ML literature developing algorithms for detecting tax evasion (Wu et al., 2020; Zheng et al., 2019) although without any field validation the tool's performance is evaluated on a hold-out sample rather than in actual practice. Our work goes a step further and eval-

[6] See e.g. "Using AI and machine learning to reduce government fraud" (Brookings Institution, 2021) for a general overview of the use of ML to reduce fraud in government settings.

uates the ML tool in a setting that is closer to where it will eventually be applied. This added step brings into focus constraints (i.e. that the tax authority was not able to cancel the identified firms) that would impact the effectiveness of the ML tool in improving primary outcomes (i.e. going beyond detection) of interest to policy-makers.

Our work is also linked to the fledgling literature in economics examining the use of technology more broadly in improving tax collections. Fan et al. (2020) document a significant improvement in VAT collections in China following a policy reform that computerized VAT returns. Dzansi et al. (2022) document an improvement in property tax collectors in Ghana from the use of tablets with a geo-coded address database that allows inspectors to improve targeting. Bellon et al. (2022) examine the roll-out of an e-invoicing policy in Peru and find improvements in reported sales and purchases though limited effects on tax collections. This strand of work is well summarized in Okunogbe and Santoro (2022) who emphasize the lack of institutional and regulatory support as a key constraint in the increased adoption of technology in improving tax collections. Our work is also related to Carrillo et al. (2022) who examine an innovative scheme aimed at reducing tax evasion by bogus firms in Ecuador. Our work is also related to Bachas et al. (2022) who compare alternative approaches for targeting firm audits and find that discretionary targeting outperforms risk-score based targeting in terms of fines levied (in Senegal). Our work complements this broad line of research by examining the actual field performance of a widely discussed technology in the context of a tax authority with limited capacity. The field-performance of an ML tool also links us to the literature on “on-policy" evaluation, see Athey et al. (2023).

The paper is structured as follows. Section 2 describes the context and provides a brief overview of the tax regime in place in our state. Section 3 describes the methodology we use to build the ML tool as well as a brief description of the field inspections. Section 4 presents the results from our analysis and Section 5 provides a discussion of the results and finally Section 6 concludes.

2 Context

We work in an Indian state and use data on tax returns over the period 2017-2021. This period coincides with the introduction of the Goods and Services Tax (GST) regime. The GST is a nation-wide value added tax (VAT) system that replaced the patchwork of state-specific VAT systems in 2017.

The GST is intended to improve upon the previous state-level VAT systems in at least two important ways. First, by unifying the state-level VAT systems as well as other tax systems nationally and in particular by making inter-state transactions visible to tax officials across states, the GST system was intended to remove tax loopholes that stemmed from the differential treatment of inter-state transactions. Second, the GST was intended to improve upon the third-party verification systems currently in place in many states (see Mahajan and Mittal, 2017).

In practice, technological challenges in the initial years made automated third party verification unreliable (see Shah, 2023). Firms faced severe cash flow problems due to unavailable, unverified tax credits (see Dave, 2021; Prabhakaran, 2022). As a stopgap solution, firms were allowed to self-declare the tax credits due to them and matching against counterparty sales declarations was not automatically enforced, except in cases where an audit was undertaken. In such a setting, firms have strong incentives to overreport available credits and this is widely viewed to have occurred (see Dhasmana, 2021).

This led to the rapid proliferation of firms that existed only on paper (“bogus" firms) whose only role was to provide fraudulent credits via fictitious sales. In extreme cases, firms could offset their entire tax liability through such fraudulent credits from bogus firms. All the bogus firm had to do was file returns with no exchange of goods or services or even a physical presence (a filed return would increase scrutiny costs for officials, see Mittal and Mahajan (2017) and Mittal et al. (2018)).

A key feature of the GST system is the division of responsibilities between the central government (the “center") and state governments. The GST rules mandate that firm registration be randomly assigned to one of the two jurisdictions. That is, when firms above a turnover threshold (₹15 million) apply for registration, they are randomly chosen to be registered at the center or at the state level with equal probability.[7] As we show below, bogus detection rates vary systematically depending upon whether a firm is state or center registered (see Appendix B.2). Moreover, to increase tracking frictions, it is common for tax evading firms to “buy" fake input credits from bogus firms that are registered under a different authority (see Table A.11 in the Appendix). These differentials will have implications for our policy recommendations.

A second key feature of the system is the lengthy procedure between the detection of a bogus firm and its subsequent cancellation (i.e revocation of status as a registered firm) in the tax authority’s system. Since beneficiary firms (i.e. those firms that trade with the bogus firm) can only be investigated after a formal cancellation of the bogus firm, physical inspection and detection by itself may not have a significant impact on the tax behavior of beneficiary firms.

3 Research Questions, Data and Methodology

We answer the following research questions (RQ):

1. RQ 1: Is the ML model effective in detecting bogus firms?
2. RQ 2: Does ML based detection improve enforcement?

In this section we first describe our data and then detail the work that went into model construction. The framework for the model construction and evaluation is similar to that described in Mittal et al. (2018). Finally, we describe the different components of the field evaluation methodology.

3.1 Tax returns data

We have an unbalanced panel of three years of monthly tax returns for all firms ($\sim$448,000) registered in a state in India (as of March 2020).[8] These are consolidated returns containing monthly sales, purchases, tax liabilities and credits claimed. The data also includes sale invoices furnished by the firms which allow

[7] Below this threshold, 90% of firms are allocated to the state authorities.

[8] After sharing our field inspection lists, we received 2 years of additional data (from April 2021 to March 2022). We use this data for some of our impact evaluation exercises.

us to link sellers to buyers. A key difference from the setup in Mittal et al. (2018) is that firms are not required to report disaggregated (i.e. firm-by-firm) purchases like most VAT systems. As a result, we do not have self-reported purchases. Instead, we use sales declarations from sellers to impute purchases to buyers—this is similar to the GST authority's own matching mechanism. For example, if firm A declares they have sold goods worth ₹X to firm B we use this information to impute features for Firm B i.e. we record that Firm B has purchased ₹X worth of goods from supplier A.

In addition to tax returns, we also have some firm-level information such as the date of registration, the revenue ward (geographical zone), the nature of business, and jurisdiction (Federal or State).[9] Whether a firm is under Federal or State jurisdiction turns out to be an important classification criterion, as each jurisdiction has separate enforcement mechanisms. Our partnership is with the state's tax department. We have registration snapshots through December 2022. This registration data (available for 21 months after the return data) enables us to track whether the potentially bogus firms and their trading partners are active and eligible to file returns in any given month.

A unique identifying number for each firm allows us to track a firm's records over time as well as match its records to counterparty sales declarations. For confidentiality reasons, this number is separate from a firm's public GST identification. The state department removed the public GST identification number along with all other personally identifiable information for the firm before making the data available to us. As a result, we cannot link this data to any other publicly available information on the firms.

3.2 Machine learning model construction

3.2.1 Labels

A key challenge was building an initial list of bogus firms for the training data. The department did not have such a list. Based on extensive conversations with tax officials, we labeled a firm as bogus if it had been canceled and the cancellation was made effective from its date of registration. The retroactive aspect meant that none of the firms' reported sales were legitimate and buyers could not claim any input tax credits from sales reported by such firms. Of the 448,000 registered firms we used to train the model, 4,837 (1.08%) were marked as bogus by this criterion.

This led to the same challenges as described in Mittal et al. (2018). Specifically, first, we have a biased labeled set. The bogus firms were explicitly chosen by the tax officials based on certain criteria so are not random. Second, the labels are one-sided. We do not have results from inspections when tax officials investigated a firm that fit their criteria but turned out to be legitimate. Therefore, while we have a list of bogus firms we do not have a truly labeled training set. We define classes for all our observations in the following way. We label firms that the department found to be bogus as 1 (“bogus") and the rest of the firms as 0 (“probably legitimate"). We are interested in predictions on the “probably legitimate" firms as we expect some of them to be bogus.

This also potentially leads to bias in our set of labeled bogus firms. Specifically, $\sim$89% of bogus firms

[9] This information is provided at registration and firms are required to update it as needed.

fall under the state jurisdiction (whereas only $\sim$69% of all firms fall under the state jurisdiction, see Table A.6 in the Appendix). We hypothesize that this is driven by the state jurisdiction putting more effort in canceling firms that were still filing or active (see Figure A.2 in the Appendix).

3.2.2 Features

We used 593 features to build the first iteration of the model. We began with discussions with officials to identify existing indicators used by the department and built features to replicate some of these indicators.[10] Second, we incorporated ideas from the literature on similar classification problems in financial settings.[11] Third, features that we defined were based on the patterns that we observed in the data.[12]

A strict data-sharing protocol limited our ability to link other firm-level, publicly available data. Since we had access to transaction-level information, all possible identifying characteristics (registration numbers, names, physical addresses) were encrypted at the department's end. Nevertheless, the unique encrypted values still allowed us to identify when multiple registered firms shared one of these characteristics. We use this information, for example, to flag if a large number (1000+) of firms seem to be registered with the same street address - a common tactic used by shell firms elsewhere.

During the field evaluation, we learnt that bogus firms are likely to operate as part of a trading network, consisting of other non-existent firms which also interact with each other before engaging with the ultimate beneficiary. This makes investigations difficult. Based on this and other patterns, we upgraded the model by adding 70 new features.[13] Of the top-1000 riskiest firms identified from the full sample by the new model 640 were labeled as bogus compared to 472 in the original model. The additional data led to compute resource constraints so we pruned the features which led to a positive impact on model performance.[14] However, we were unable to send out predictions from the new model for field evaluation.

3.2.3 Classifier

Mittal et al. (2018) find that the Random Forest (RF) classifier is best on this class of problems. This makes sense because we have enough labels to use supervised learning, but not enough to use more sophisticated models like Deep Neural Networks. We selected Random Forest also because it can capture nonlinear relationships between features without needing them to be explicitly specified and works well with both categorical and continuous variables. We trained a RF model with n_trees=128 and

[10] For example, rank by firm size in terms of credits claimed, percentage mismatch between cumulative sales invoices and self-reported turnover, number of days for registration approval (registration application is auto-approved after a set number of days, so this is a proxy for auto-approval), whether a firm is trading in "high risk" commodities.

[11] For example, whether a firm's listed address is unique, whether a firm uses a public vs. private email domain, total B2B turnover.

[12] For example, filing delay and filing regularity for sales invoices and consolidated returns, number of firms registered on the same street address (busy street), earliest date of other (VAT, CST) tax registration, network features etc.

[13] This also includes mismatch features, a few risk flags generated by the central authority, features from imputed purchases, and from bank account information.

[14] On a 10% sample, the model with the full set of features correctly identified 299 firms labeled as bogus in the top-1000 firms by model score compared to 316 firms correctly identified with a pruned set of features.

max_tree_depth=256 in scikit-learn.[15] Our subsequent methodology is similar to the one described in Mittal et al. (2018). First, we wish to verify the status, through physical inspections, of the in-sample firms currently classified as probably legitimate but predicted to be bogus. Therefore, we carry out cross-validated holdout predictions with 8 folds (where different returns by the same firm are always assigned to the same fold). Second, we need one prediction per firm. Firms file a return every month so we have multiple observations for each firm and their bogus classification should not change over time. We aggregate multiple firm predictions by averaging them to give a single prediction score per firm. This approach increases the number of data points available for training, but neglects certain between-return relationships and indications of data availability such as irregular filing or last filing date. We then verify the efficacy of our model by physically inspecting the riskiest firms amongst those that are currently classified as 'probably legitimate'.

3.3 Generating firm lists for physical inspection

We use the trained model to calculate the probability of being bogus for each firm registered in the state. We then weight the predicted probabilities by transaction volume (proxied by square root of total input credits claimed). The tax authority's inspection capacity is limited and they are primarily interested in bogus firms with a high value of fraudulent transactions. Weighting by the input credits claimed adjusts for this constraint. In addition, input credits are fat tailed. Therefore, we weight by the square root of input credits to reduce the skew towards very large firms. We then share these sorted lists with the state tax department for inspection over multiple rounds. In this section, we describe the selection criteria for these lists (details are included in Appendix A).

Over a period of 15 months from March 2021 to July 2022, we shared a total of 4 lists containing 1,879 firms. Besides the firms predicted to be risky by the model we also included firms that satisfied other criteria. First, to compare model effectiveness with proxy department procedures, we included a set of firms using the rule-of-thumb risk criteria that tax officials claimed to use while deciding which firms to inspect. Second, we included firms linked to known bogus firms to better understand the behavior of the trading network of bogus firms.

We received full inspection results for the first 3 lists (covering 822 firms) and received partial inspection results of our final list (containing 1,057 firms). As a result, when we analyze verified non-existent firms, we only describe results from the first 3 lists. By contrast, when we compare model predicted risky firms, we use firms from all 4 lists. For analysis, we sometimes group all 4 lists together, which we term the "Combined List". Moreover, for the first three lists, the state performed inspections for both center jurisdiction and state jurisdiction firms. In principle, it should cancel all verified bogus firms under its jurisdiction and send a request for cancellation to the center for firms that were under center jurisdiction. For our fourth list, we know that the state department did not inspect center jurisdiction firms, so we should not expect any cancellation of those firms.

For physical inspections, we also filtered out firms that were inactive at the time of list sharing. We followed this rule because it seemed unnecessary (and a waste of effort) for the tax department to

[15] For relevant documentation, see scikit-learn developers (2021)

Group	Group Name	Description	N. Firms shared	N. Inspection results received
A	**Model Score**	Firms that would actually be inspected based on our model prediction. Firms with a high model score and a high input tax credit (ITC).	1,205	938
B	**Random firms**	Randomly selected firms, only meeting a certain minimal transaction volume threshold.	100	49
C	**Rule-of-thumb**	An approximation of the department's current methods. Firms predicted to be bogus using a criterion common in tax departments across India (a high ITC to turnover (or tax paid) ratio).	50	50
D	**Trading partners**	Trading partners of identified bogus firms.	90	90
E	**ML with limited data**	Based on the ML predictions, but using fewer returns than those available in order to assess how the tool's performance changed when using fewer returns.	330	139
F	**Shared identifiers**	Firms that had the same key variables (e.g., PAN numbers, phone numbers) as identified bogus firms.	104	34

Table 1: Selection Criteria for Physical Inspection of Suspicious Firms by the Department. *Notes: We did not receive the inspection results for the complete set of suspicious firms shared with the department. To check if this introduces a bias, we confirm that the model scores do not systematically differ for the sets of firms where the inspection results were received and not received, stratified by group (results available upon request).*

inspect already canceled firms. In the analysis we sometimes include these canceled firms to ensure that our main and comparison lists are balanced. All list criteria are summarized in Table 1.

3.4 Shadow lists (comparison groups) from holdout set

We retained 10% of the firms as a holdout set. Specifically, we did not use data from these firms to train or improve the model or share the predictions from this sample with the tax authorities for inspection. We use this holdout set to construct four comparison groups using the same criteria used on the four main lists to identify firms that were sent for inspection. For the main lists, the cross-validated model used to predict a firm's probability of being bogus is one where its training fold did not include that firm. In the holdout group, we randomly select one of the possible eight models to calculate this probability. This approach mimics the random division of the main group into folds.[16] These comparison groups, that we refer to as shadow lists, help us evaluate whether the use of the ML model improves state enforcement (Section 3.6). We take care to ensure that the risk profiles of the comparison groups are identical to the risk profile of the firms that we shared with the authorities for field inspection.

[16] We use the same seed to randomize holdout firms between models as we do to randomize main firms between folds. The predicted probabilities for a holdout firm calculated using different cross-validated models are highly correlated, so the random assignment is not expected to have significant bearing on the calculated outcome.

Firm Attribute	List-Combined		Shadow-Combined		Difference (Main - Shadow)	
	Mean	St. Dev	Mean	St. Dev	Coef.	p-value
Model Score	0.05	0.04	0.04	0.03	0.00	0.17
Registration Month	19.36	9.80	18.18	9.75	1.17	0.16
If registered with State authorities	32.68%	46.92%	35.03%	47.86%	-2.35%	0.60
Turnover, till FY20	218.0M	898.1M	253.8M	1,327.6M	35.8M	0.72
If Turnover (till FY20) is not available	1.67%	12.82%	2.55%	15.81%	-0.88%	0.50
Total Tax Paid (Cash), till FY20	0.7M	3.6M	1.4M	7.3M	0.7M	0.05
Total ITC Claimed, till FY20	27.5M	71.0M	33.6M	153.7M	6.1M	0.37
Tax-to-Turnover Ratio, till FY20	0.58%	3.98%	6.52%	75.96%	-0.06	0.06
log(Turnover, till FY20)	18.03	1.45	18.00	1.43	0.02	0.82
log(Total Tax Paid (Cash), till FY20)	8.67	4.39	8.71	4.72	-0.04	0.91
log(Total ITC Claimed, till FY20)	16.06	1.39	16.01	1.44	0.04	0.72
log(Tax-to-Turnover Ratio, till FY20)	0.01	0.03	0.02	0.19	-0.01	0.06
No. of Monthly Sales Statements Filed, till FY20	10.27	7.67	11.02	7.52	-0.75	0.21
No. of Consolidated Returns Filed, till FY20	11.87	7.92	12.99	7.69	-1.12	0.10
If registered as Composition firm	0.44%	6.59%	1.27%	11.25%	-0.84%	0.19
If migrated from VAT to GST	9.22%	28.95%	11.46%	31.96%	-2.24%	0.39
If registered as proprietorship, partnership or Pvt Ltd firm	96.73%	17.79%	95.54%	20.71%	1.19%	0.47
N. Observations	Combined_main	1377	Combined_shadow	157		

Table 2: Balance Tests for Randomization of Firms. *Notes: We verify that shadow list firms do not systematically differ from main list firms prior to treatment. We use several firm-level attributes: the model-generated risk score, State vs. Center registration, total Input Tax Credit (ITC) claimed, the ratio of Tax to total revenue, and other attributes. Amounts are in nominal Rs., and their log is often included as well due to the fat tailed distributions. Note that we do not explicitly adjust for multiple hypotheses, so due to the large number of attributes some have a p-value less than 0.05, but when adjusting this significance would vanish. The two sets are balanced, implying that any observed variations in subsequent enforcement outcomes can be causally attributed to the treatment itself, rather than to pre-existing disparities between the two sets.*

3.4.1 Balance tests between shadow lists and main lists

We verified that the randomization functioned as expected and that there are no systematic differences between the main and the holdout set (prior to sharing the lists with the department). We check for balance on various firm attributes between the main lists and shadow lists. For convenience, we present in Table 2 the balance table for the Combined List (aggregation of all four lists). For list-wise tables, see Table A.3 in the Appendix.

3.5 RQ1: Is the ML model effective in identifying bogus firms?

We are constrained by the number of firms that tax authorities can physically inspect. This constraint is relevant for both model evaluation as well as eventual deployment. The model's performance on firms that are unlikely to be inspected has limited implications. An ideal algorithm to target bogus firms should deliver a high accuracy rate with low investigation cost (in terms of time and effort). Therefore, we evaluate only the top firms ordered by riskiness and not the list of all firms with risk scores. Specifically, we take a few realistic numbers of top recommendations and check our model's success on those i.e. of these top N firms predicted to be bogus by our model, how many are known to be bogus or are found to be bogus in inspections.

We perform four comparisons to assess ML model performance. First, we compare the model performance based solely on cross-validated predictions (as carried out in Mittal et al. (2018)) with the actual field inspection results. The field inspection results tell us whether the firm was found operating at its stated location. This comparison evaluates the effectiveness of the ML model in real world resource constrained environments, rather than by typical evaluation metrics described in the ML literature. Second, we compare our field inspection results with a random selection of firms (Group B in Table 1) i.e., if firms were randomly selected for physical verification, what percentage of firms inspected would be non-existent. The efficacy of the random inspection approach depends on the true prevalence rate of bogus firms. If a large percentage of all registrations are bogus, the hit-rate of randomly selected firms will be high. This is also the lowest-effort approach, so any alternative will have to have a higher accuracy rate to be worthwhile.

Third, we compare our model suggested inspections with a business rule based (Group C in Table 1) or a 'red flag' approach. A firm can be flagged for inspection if it crosses a pre-set threshold on one or more parameters of interest. For example, when input credit claimed or reported turnover or their ratio exceeds a threshold. We understand that a majority of the department's current efforts identify candidates for investigation using a similar red flag approach with inputs from the relevant ward official. As we do not have data on physical inspections (successful and unsuccessful) carried out by the department using red flag reports, we create a sample red flag report based on discussions with the department's data mining team. We select firms that claim large input tax credits but do not have a commensurate amount of tax paid in cash. We select firms that have a total ITC larger than Rs $\sim$10 million (median ITC from first list, although we filter out the top 5000 taxpayers to exclude genuine firms). Finally, we select the top 50 firms in descending order of the ratio of total ITC claimed to total reported sales amount (across the firm's lifetime). We refer to these as rule-of-thumb criteria.

Fourth, we compare cancellation probability of firms in our lists to all firms not included in our lists. Since list firms were selected based on high model-predicted risk scores, we expect them to be canceled at a much higher rate than those firms not included in the lists, which have much lower risk scores.

Our inspections do not incorporate the latest available returns at the time of inspection— i.e. the inspections were carried out in 2021-22 but are based on tax return data only through FY 2019-20. Additionally, our lists consist of risky firms that were not yet identified as bogus. Therefore, these results understate the model efficacy as tax officials have already identified some very risky firms and canceled them. Model risk scores of our firms that were field inspected (Table A.5a) is lower than model risk scores of our cross-validated predictions(Table A.5b). Implementing our approach using current data should lead to further improved predictions and accuracy.

3.6 RQ2: Does ML based detection improve enforcement?

Following the empirical validation of the model's efficacy in identifying fraudulent firms through multiple benchmarks, we transition to an investigation of the model's impact on field enforcement outcomes. To this end, we conduct two distinct comparative analyses: the first juxtaposes "shadow lists" against their corresponding "main lists," while the second is a comparison across the main lists themselves, leveraging the variation in the timing of list dissemination and subsequent inspections.

In the first comparison, we compare enforcement outcomes between main lists and their analogous shadow lists. These shadow lists are constructed using the same selection criteria as the main lists but are derived from a separate, holdout dataset that was intentionally withheld from the tax authority. Thus, the shadow lists function as a randomized control group for the main lists, allowing us to evaluate the causal effect of tax department inspection on various enforcement outcomes. Since we do not have inspection outcomes for the shadow lists, we cannot identify bogus firms on the shadow lists. We instead compare common outcomes across the two lists. The primary outcome is the cancellation status of a firm (as of December 2022). Other outcomes, contingent upon observing a differential impact in cancellations, include the recovery of foregone revenue and downstream effects on trading partners.

The second comparative analysis is of cancellation times within the main lists. We analyze the temporal distribution of firm cancellations across the four sequentially shared lists.[17] If the inspection process accelerates the cancellation of fraudulent firms, firms on earlier lists—and thus subject to earlier inspection—would have earlier cancellation dates. To isolate the effect of list dissemination on cancellation timing while controlling for the inherent likelihood of cancellation (which is expected to differ between lists due to their construction in descending order of the likelihood of being bogus), we use the Cumulative Distribution Functions (CDFs) of cancellation timings. These CDFs are computed exclusively for firms that eventually undergo cancellation, ensuring that each CDF is equal to 1 at the terminal time point available in the dataset. This enables a meaningful comparison of conditional cancellation timings across lists.

Furthermore, we can compare these list-specific CDFs to the CDF of cancellation timings for the remaining firms in the primary sample that were not part of the disseminated lists but were nonetheless canceled within the study period.

[17] To clarify, firms known to be already canceled were omitted from the disseminated lists as inspecting them would be redundant and infeasible. For analysis, however, we create all lists as if they were compiled simultaneously and no exclusions are made, to make the comparison between lists valid. So all firms analyzed were active as of March 2021.

4 Results

4.1 RQ1A: Model predictions are more effective in the field compared to cross validation precision

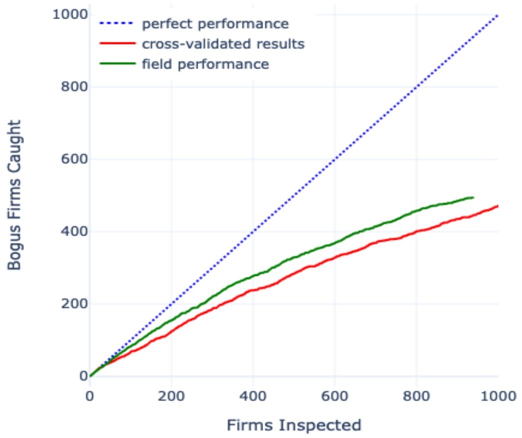


Figure 1: ***Model performance on top 938 predictions.*** *Notes: ordered by model score. We sent 1,205 firms in the 'model score' category and got back outcomes for 938 of them.*

In the field, our model holds up to (in fact performs slightly better than) the cross-validated performance. Figure 1 shows that our field inspections had a hit rate of 53% in the top 938 firms while the cross-validated predictions suggest a 47% hit rate. Note that by definition cross-validated predictions measure hit rate by the share of firms amongst the top K firms that were inspected, classified as bogus AND canceled (since our labels capture both). Field inspections on the other hand measure hit rate by the share of firms in the top-K that were found to be non-existent (all were inspected). It is important to note that we shared our lists of firms for field inspection after filtering out firms that were already canceled. So our field inspection hit rate is downward biased.

Our inspection results hide meaningful heterogeneity driven by jurisdiction authority. First, Table A.7 in the Appendix shows that only 30.3% of our predicted risky firms are state registered despite the fact that the labeled set had 89% such firms. Second, the bogus detection rate for center registered firms is much higher at 60.5% (the corresponding hit rate for state registered firms is 34.7%).

4.2 RQ1B: ML model is more accurate than the rule-of-thumb approach

Across the firms in the lists that were recommended for inspection by our ML tool (left-hand bar in Figure 2)[18], we find that 53% were identified as bogus on the basis of physical inspections.[19] In the

[18] For more details, see the criteria detailed in Appendix A.

[19] The detection rate for lists 1-3 was 67%. Recall that we shared the riskiest firms first (in the initial lists) so that by construction we would expect to see higher detection rates in the initial lists. Further, in an actual field setting later lists would also incorporate additional return information (i.e., in a field setting, a list shared later would be based on additional returns data

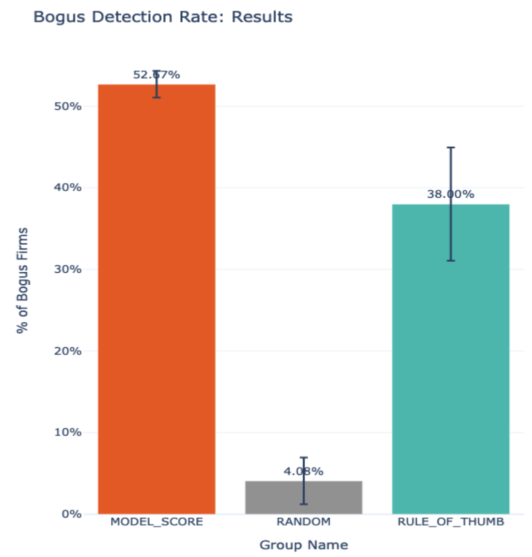


Figure 2: *Bogus detection results comparison across groups.*

rule-of-thumb criteria (right-hand bar) we found 38% of inspected firms to be bogus. Thus, the machine learning approach offers a more accurate solution than the tax department's current approach. In sum, this is a relatively low cost approach since the bulk of the effort required to generate these predictions is a one-time effort at the beginning.

The rule-of-thumb accuracy seems high by comparison to the ML model. However, it is based on a relatively small number of 50 firms, whereas the ML model accuracy is based on 938 firms. It is expected that any model's accuracy would deteriorate as it is used to make more and more predictions. Moreover, recall that the ML model was trained on labels dependent on tax department inspections, which employed these rule-of-thumb methods for inspection targeting, so it is somewhat limited in accuracy by these methods.

4.3 RQ1C: List firms are canceled at a higher rate than non-list firms

We compare the cancellation probabilities of firms in our lists to all firms not included in our lists. Since list firms were selected based on high risk scores, we expect them to be canceled at a much higher rate than those firms not included in the lists (which have lower risk scores by construction). This is indeed what we see, with list firms canceled at about 3.5 times the rate of baseline firms, see Table 3. This is an indication that the model was indeed successful at the task it was trained for - predicting which firms will be canceled.

than previous lists) which would improve the detection rate.

Group	N.Obs	Indicator for Cancellation		Combined vs. Baseline	
		Mean	**Std. Err**	**Difference**	**Std. Err**
Combined (Main)	1,377	0.415	0.013	0.298	0.013
Combined, State (Main)	450	0.376	0.023	0.259	0.023
Combined, Center (Main)	927	0.434	0.016	0.317	0.016
Baseline (All)	332,225	0.117	0.001		

Table 3: Observed Difference In Cancellation Rates Between List Firms And Non-List Firms. *Notes: The first row shows numbers for the Combined List from the Main (not Holdout) set. The next two rows disaggregate the Combined List into State and Center registered firms. The last row shows numbers for the Baseline - all firms (both Main and Holdout) not included in the lists, but meeting the basic criterion of being active as of the list compilation date - March 2021. The columns indicate the number of observations in each row, the fraction of firms canceled, the standard error, the difference between the fraction canceled and the fraction canceled in the Baseline, and the standard error of the difference. There is a much higher cancellation rate of List firms than of non-List firms. The difference is extremely significant, and exists in both State and Center. Cancellations are measured as of the final data point available in our dataset, December 2022.*

4.4 RQ2A: Cancellation rate of inspected firms is indistinguishable from that of holdout firms

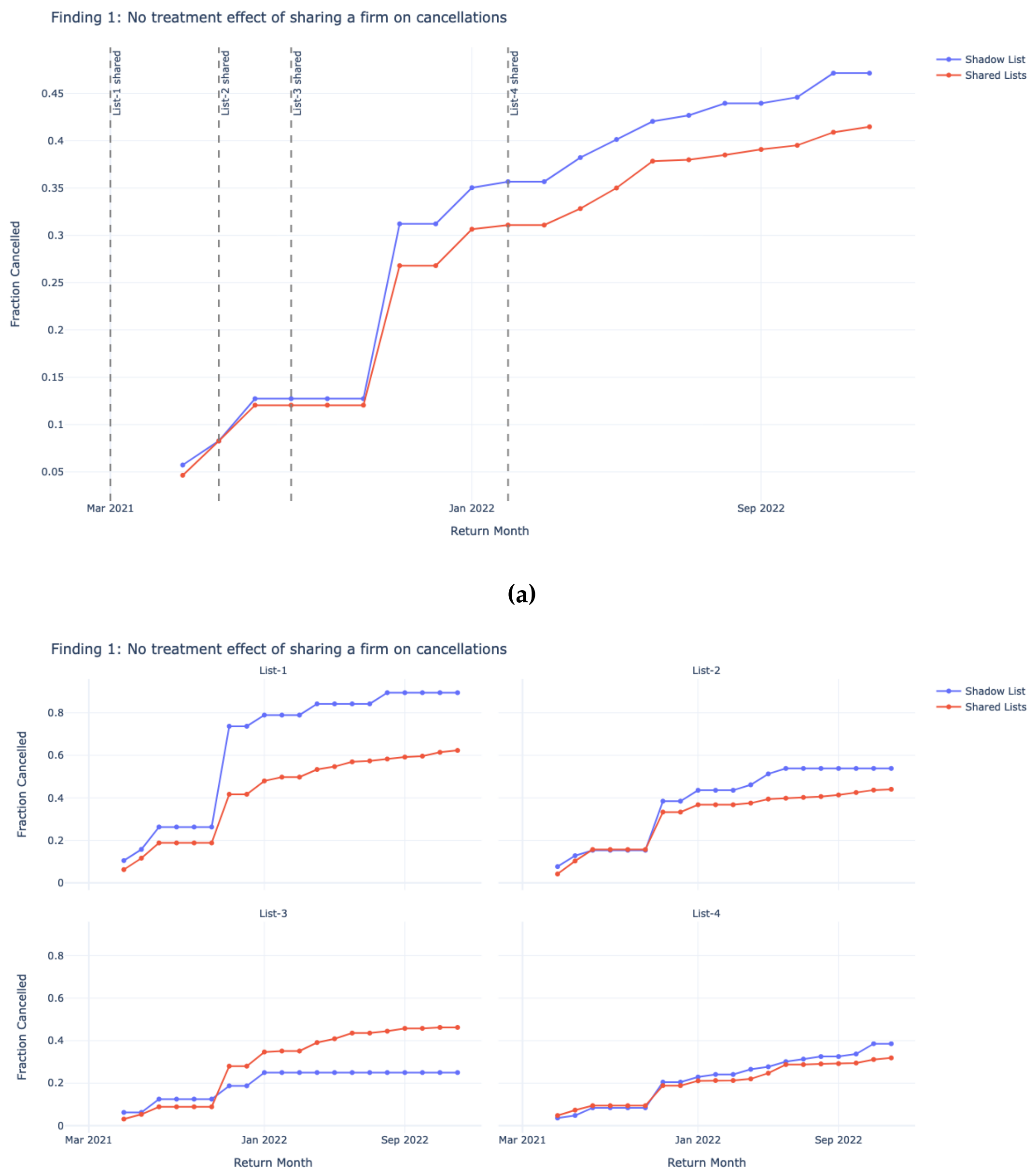
In this section, we compare firm cancellations (total, and over time) between main lists and their respective shadow lists. We conduct two sub-analyses: the first evaluates the proportion of firms canceled in each main list against its corresponding shadow list. These comparisons are made for all lists and separately for each individual list (see Table 4). We find the treatment effect of being in the main (as opposed to shadow) lists on cancellation to be -0.057 (s.e.=0.047, p=0.293). Thus firms shared with the department for inspection were 5.7 percentage points *less* likely to be canceled, although the difference is statistically insignificant. See also Figure 3.

Subsequently, we examine cancellation trends over time. To do this we constructed registration status snapshots till December 2022. We find no significant divergence, in contrast to the expectation of expedited cancellations among main list firms. Figure 3 delineates these temporal trends in cancellations. Kolmogorov-Smirnov tests indicate that any differences in the distribution of cancellation times are not statistically significant, except for List-1 (see Table A.9 in the Appendix).

The results collectively suggest no treatment effect on firm cancellations. Sharing firm lists with the tax authority for inspection does not affect the cancellation process on average. Even under weak assumptions—e.g. where an anticipated time-displacement effect would result in earlier cancellations for inspected firms identified as fraudulent, or where the diversion of the department's finite inspection resources towards listed firms would induce earlier cancellations—such outcomes are not empirically observed.

Group	N.Obs	Indicator for Cancellation		Main Lists vs Holdout Lists	
		Mean	Std. Err	Difference	p-value
Combined (Main)	1,377	0.415	0.013	-0.057	0.253
Combined (Holdout)	157	0.471	0.040		
List-1 (Main)	223	0.623	0.033	-0.271	0.034
List-1 (Holdout)	19	0.895	0.072		
List-2 (Main)	261	0.441	0.031	-0.098	0.126
List-2 (Holdout)	39	0.538	0.081		
List-3 (Main)	225	0.462	0.033	0.212	0.211
List-3 (Holdout)	16	0.250	0.112		
List-4 (Main)	668	0.319	0.018	-0.067	0.309
List-4 (Holdout)	83	0.386	0.054		
Combined, State (Main)	450	0.376	0.023	-0.115	0.107
Combined, State (Holdout)	55	0.491	0.068		
Combined, Center (Main)	927	0.434	0.016	-0.027	0.667
Combined, Center (Holdout)	102	0.461	0.050		
Baseline (All)	332,225	0.117	0.001		

Table 4: Treatment Effect on Proportion of Canceled Firms. *Notes: This table compares the cancellation rates for firms between the main and shadow lists. Each pair of rows indicates one list, broken down by Main and Holdout. The columns indicate the number of observations in each row, the fraction of firms canceled, the standard error, the difference between Main and Holdout lists, and the (permutation test) p-value for equality of means from a t-test. List-1 shows a higher cancellation rate for shadow firms not subjected to departmental inspection; although significant at the 0.05 level, this could be coincidental. List-3 shows the opposite - a higher, though insignificant, cancellation rate for main list firms. Rows labeled "State" or "Center" refer to firms registered at the State or Center respectively. We show that there is no significant effect for either. The last row provides cancellation numbers of the baseline - all firms (both Main and Holdout) not included in the lists, but meeting the basic criterion of being active as of the list compilation date - March 2021. Note the cancellation rates for the baseline are much lower than those for list firms. Cancellations are measured as of the final data point available in our dataset, December 2022. A negligible number of cancellations were subsequently reversed; however, these are still included in the cancellation count, thereby implying that the metrics reflect firms that were ever canceled up to the examined time frame.*



(a)

(b)

Figure 3: ***Firm cancellations over time in Main and Shadow lists.*** *Notes: Each subplot represents data pertaining to a specific list. The top panel (a) is for the Combined List, and the bottom (b) for the separate Lists. The red line marks the proportion of cancellations over time among firms in the main list, while the blue line does the same for firms in the corresponding shadow list. Aggregated data reveals a subtly elevated likelihood of cancellation for firms on the shadow lists. List-1 notably exhibits a higher cancellation rate for shadow firms compared to main firms. For Lists 2 and 4, the cancellation proportions are relatively similar, whereas List 3 demonstrates an increase in cancellations among firms on the main list. For statistical significance of these differences, see Table A.9.*

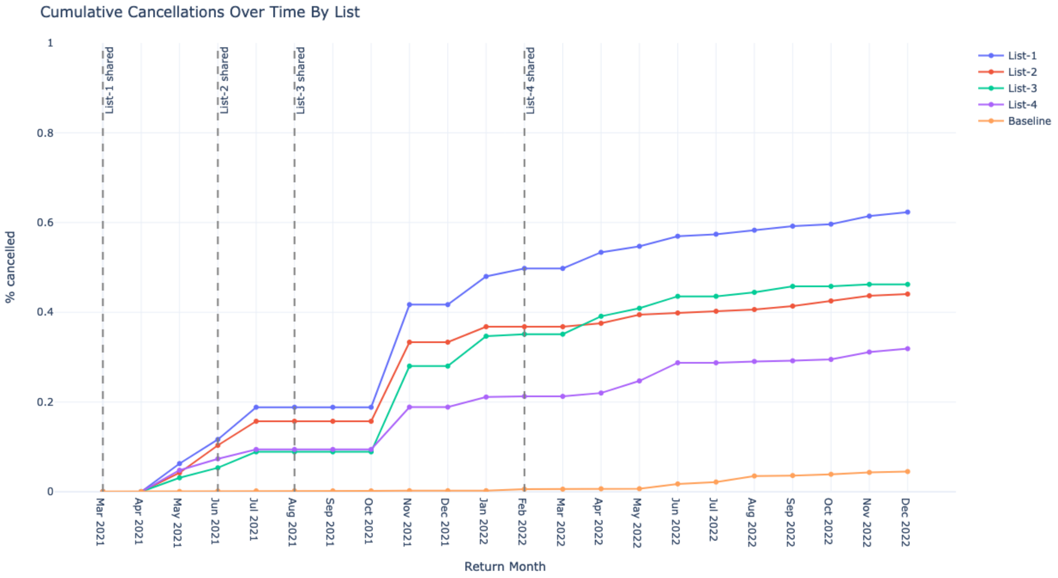


Figure 4: ***Cumulative cancellations over time By list number.*** *Notes: The horizontal axis denotes time, while the vertical axis represents the cumulative proportion of firms canceled over time in each list. For reference, the baseline for all other non-list firms, active as of the list compilation date - March 2021, is also provided. Dotted vertical lines indicate the times at which the lists were shared. As expected, earlier lists manifest a higher cancellation rate, a byproduct of their construction based on higher risk scores from the predictive model. This characteristic, however, is abstracted from subsequent analyses that are conditioned on cancellation of a firm, and so exploit variation solely in the timing of cancellations. The somewhat erratic trend in cancellations stems from the fact that cancellation data came from snapshots that were not available every month, specifically between July and November 2021.*

4.5 RQ2B: No effect of sharing a firm earlier rather than later on time of cancellation

We next direct attention to the timing of cancellations among firms listed for inspection in the main lists. As elaborated in the methodology section, our focus is on the Cumulative Distribution Functions (CDFs) of cancellation timings. We take this approach to neutralize the impact of the inherent cancellation probability (see Figure 4), thus permitting a pure examination of the *timing* of cancellations.

Despite the time lag in sharing and inspecting firms from later lists, the CDFs across all lists are remarkably similar (see Figure 5). In fact, the distributions of cancellation timings for List 3 and especially List 4 closely align with those of lists 1 and 2, even before the sharing and inspection of firms from lists 3 and 4, at which time of course the composition of the lists was not known to the department so could not have affected cancellations. This indicates that early sharing has no economically significant effect on the timing of cancellations, conditional on a firm being canceled. These findings present robust evidence against the hypothesis that inspections are the primary driver of cancellations.

We conducted k-sample Anderson-Darling tests to compare the distributions of cancellation timings among lists. For lists 1-4, the statistic was A^2=5.65, with a p-value = 0.001. When including the baseline, the statistic was A^2=294.36, with a p-value $\ll 0.001$. Though the tests yield significant results, note that the difference between the distributions is not always in the expected direction. For example List-2 firms

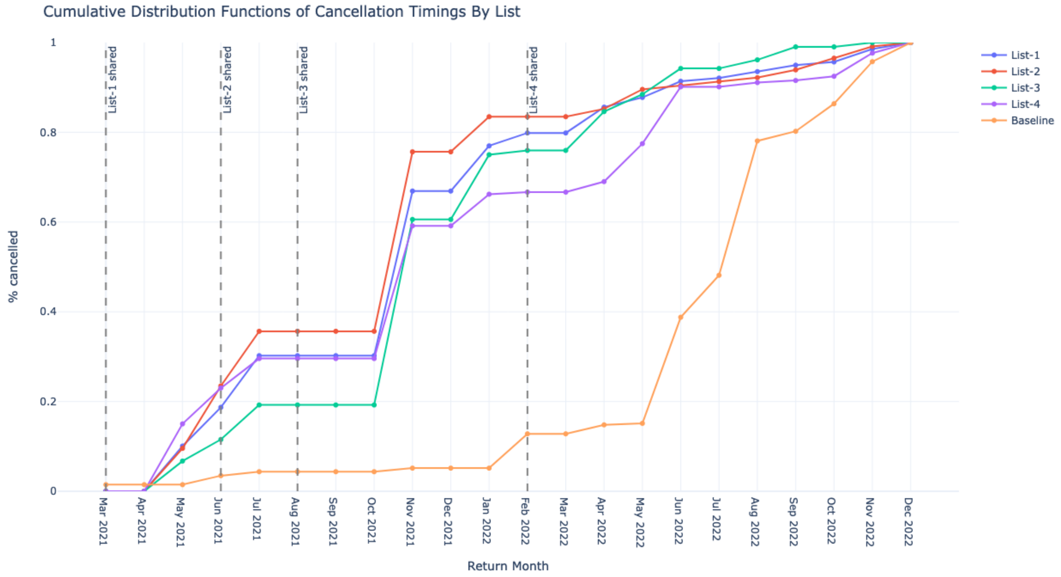


Figure 5: ***Cumulative Distribution Functions of cancellation timings By list number.*** *Notes: This figure shows the CDFs of cancellation timings, conditional upon the firms being canceled. Each curve represents one of the four lists. Dashed lines annotate the time points at which the lists were shared, commencing with List 1 and concluding with List 4. A baseline distribution is also included, depicted in orange, which represents the cumulative share of suo-moto cancellations over time for all non-list firms in our dataset. Despite the time lag in sharing and inspecting firms from later lists, the CDFs across all lists are remarkably similar. The somewhat erratic trend in cancellations, especially July-November 2021, stems from the fact that cancellation data came from snapshots that were not available every month.*

List	Inspection Result	No. of Firms (Row Total)	No. of Firms Canceled (% of Row Total)	No. of Firms Filing Final Month (% of Row Total)
List-1	Non-Existent (bogus)	174	124 (71.3%)	38 (21.8%)
	Existent	49	13 (26.5%)	32 (65.3%)
List-2	Non-Existent (bogus)	161	98 (60.9%)	58 (36.0%)
	Existent	139	8 (5.8%)	128 (92.1%)
List-3	Non-Existent (bogus)	143	94 (65.7%)	47 (32.9%)
	Existent	156	12 (7.7%)	144 (92.3%)
Combined 1-3	Non-Existent (bogus)	478	316 (66.1%)	143 (29.9%)
	Existent	344	33 (9.6%)	304 (88.4%)

Table 5: Inspection Result by List and Subsequent Firm Activity.

are canceled earlier than List-1 firms, and List-4 firms are canceled earlier than List-3 firms in the first months. Also note that most List-4 firms are canceled before the list was even shared, as would be very unexpected if inspections were the main drivers of cancellation.

While our primary enforcement outcome of cancellation did not show a significant effect, we still studied secondary outcomes like foregone revenue recovery or downstream effects on trading partners for our final research question. See Section 4.7.

4.6 RQ2C: Inspections finding bogus firms do not always lead to cancellations

Conditional on being found bogus on inspection, the majority of firms were indeed canceled. Firms inspected and found non-existent are very often (66%) later canceled, certainly more frequently than firms found existent (9.6%). However, many firms which were found to be non-existent are not canceled, and some continue to file returns (see Table 5). If even firms found bogus on inspection are not always canceled, this is a possible explanation for why our model-targeted inspections did not drive model-targeted cancellations as we expected.

Notes: Rows are subdivided according to the inspection result across individual and combined lists. We only use Lists 1-3, for which we have inspection results. Columns are as follows:

- Total Number of Firms: Count of firms for each inspection status category.
- Number of Firms Canceled: Count of canceled firms as of the most recent data. Percent of row total in parentheses.
- Number of Firms Filing Final Month: Count of firms that filed a return in the last available month of data. Percent of row total in parentheses.

4.7 RQ2D: Inspections do not increase revenue recovery from bogus firm beneficiaries

As we do not know the bogus firms in the holdout set, we limit our analysis to only firms that we know were canceled and not firms that were found bogus and canceled. After cancellations, we aimed to analyze the tax revenue recovered from firms that “purchase" from bogus firms. The department can identify these beneficiaries from the detailed sale invoices of bogus firms. After a firm has been classified as bogus, officials responsible for the jurisdictions where the beneficiary firms are located should initiate tax recovery from these beneficiaries. However, lack of automation results in limited visibility into the amount of tax recovered that can be directly attributed to bogus interactions. Nevertheless, we still analyze differences in coarse grained revenue recovery proceedings that the department has engaged in with these beneficiaries.

Specifically, we look at three outcomes. First, whether the department sent out any discrepancy notice to these beneficiary firms post our inspections. Second, did the beneficiary make any additional tax payments against the said notice. Finally, did the department temporarily prevent the beneficiary from using its input credits at any point after the inspection.[20] In Table 6, we compare the share of beneficiaries at each of these stages between our main lists (shared with the department) and holdout lists, and also against a baseline of all active firms as of March 2021.

First, we find that beneficiaries of the bogus firms are more likely to be on the official radar. For example, $\sim$6% of the beneficiaries (see results for combined lists) make additional tax payments compared to only 1.7% of the baseline (last row). Second, we do not find evidence of inspections leading to an increase in recovery proceedings. Specifically, we don't see a difference in outcomes between the main and hold-out lists. It would be interesting to explore the impact of improved sharing of inspection findings within the tax department, we leave this for future work.

[20] We were able to obtain these outcomes from GST reports made available by the department.

				Any Notice Issued (post-inspection)				Additional Tax Payment, Against Demand				Tax Credits Blocked (post-inspection)			
						Main Lists vs. Holdout Lists				Main Lists vs. Holdout Lists				Main Lists vs. Holdout Lists	
Group	N. Obs.	N. Obs. (Linked)	N. Buyers	Mean	Std. Error	Difference	p-value	Mean	Std. Error	Difference	p-value	Mean	Std. Error	Difference	p-value
Combined (Main)	571	528	11,822	7.8%	0.3%	0.4%	0.64	5.5%	0.5%	-0.8%	0.59	6.7%	0.2%	0.4%	0.60
Combined (Holdout)	74	67	1,368	7.5%	0.7%			6.3%	2.0%			6.3%	0.7%		
List-1 (Main)	139	128	2,300	8.2%	0.6%	1.4%	0.52	7.3%	0.5%	0.6%	0.61	6.4%	0.5%	1.9%	0.29
List-1 (Holdout)	17	16	221	6.8%	1.7%			6.7%	1.0%			4.5%	1.4%		
List-2 (Main)	115	112	3,309	10.2%	0.6%	1.8%	0.17	7.7%	0.6%	0.7%	0.97	7.5%	0.5%	0.9%	0.46
List-2 (Holdout)	21	20	645	8.4%	1.1%			7.0%	3.7%			6.7%	1.0%		
List-3 (Main)	104	97	1,950	6.6%	0.6%	-7.5%	0.03	7.4%	0.4%	0.9%	0.48	7.5%	0.6%	-1.0%	0.82
List-3 (Holdout)	4	4	71	14.1%	4.2%			6.5%	1.3%			8.5%	3.3%		
List-4 (Main)	213	191	4,263	6.4%	0.4%	1.0%	0.44	7.1%	0.2%	0.5%	0.52	5.8%	0.4%	-0.4%	0.74
List-4 (Holdout)	32	27	431	5.3%	1.1%			6.6%	0.7%			6.3%	1.2%		
Combined, State (Main)	169	151	4,036	6.9%	0.4%	0.5%	0.62	6.8%	0.4%	1.3%	0.30	5.2%	0.3%	1.8%	0.07
Combined, State (Holdout)	27	22	561	6.4%	1.0%			5.5%	1.0%			3.4%	0.8%		
Combined, Center (Main)	402	377	7,786	8.3%	0.3%	0.1%	0.94	7.2%	0.3%	-0.1%	0.94	7.5%	0.3%	-0.8%	0.39
Combined, Center (Holdout)	47	45	807	8.2%	1.0%			7.3%	1.0%			8.3%	1.0%		
Baseline (All)				5.9%	0.0%			1.7%	0.0%			1.0%	0.0%		

Table 6: Treatment Effect on Recovery from Bogus Firm Beneficiaries. *Notes: This table compares the action taken towards revenue recovery for firms which are listed as the beneficiaries of the transactions reported by the main and shadow lists. Each pair of rows indicates one list, broken down by Main and Holdout. The columns indicate the number of observations in each row (i.e the number of canceled firms from the list), the number of observations where at least one beneficiary was listed on their tax return, the total number of beneficiaries across all such observations, the fraction of beneficiaries where the department took issued a notice, received an additional payment of tax against such notice, or blocked the tax credits available to the beneficiary following such notice. Also included are the standard errors, the difference between Main and Holdout lists, and the (permutation test) p-values for equality of means from a t-test. The relevant number of observations here is the number of canceled firms out of all the firms included in that list because the department may only undertake any recovery from a beneficiary after the bogus firm has been canceled.*

5 Discussion

5.1 Reconciling predictive accuracy with absence of treatment effects

At the heart of this paper is a conundrum. The problem of identifying bogus firms is assumed to be a targeting problem – finding which firms are bogus is the bottleneck and most challenging step. We introduce a model which improves targeting. It is accurate at identifying bogus firms, as measured on the test set (Section 4.1). It is accurate at identifying bogus firms, as measured in field inspections (Section 4.2). It is even accurate at identifying firms much more likely to be canceled, compared to the general population (Section 4.3). The conundrum arises because despite such promising results there is no causal effect of employing it, on the primary desired outcome - cancellations (Section 4.4 and Section 4.5). We offer two explanations: overfitting to proxy-labels, and inadequate enforcement.

5.1.1 Overfitting to proxy-labels

The first cause is overfitting to proxy-labels. We use the term "overfitting" loosely, since the model technically fits well to its training labels and the 'output' of the field inspections, i.e. identifying a non-existent firm, without yielding the desired outcome, which is cancelation and eventual recovery of lost taxes (a longer process depicted in Figure A.9 in the Appendix).

The ML model is adept at predicting firms that are already on a trajectory toward cancellation, regardless of any subsequent inspection activities. Thus, they show up as accurate predictions on the test-set and inspections and are eventually canceled. But this effect is not causal (see Sections 4.4 and 4.5), they would have been canceled anyway. How could this come about? During model training we lacked an ideal target variable that would categorize firms as either 'bogus' or 'legitimate' based on field inspections, since records of past inspection results are not kept by the department. Consequently, we resorted to using cancellations with retroactive effect as a proxy variable to signify a firm's bogus status (these are known as 'suo-moto' cancellations in the tax department). This is not a baseless choice - if the purpose of the department's inspections is to find and cancel bogus firms, we would expect being found bogus upon inspection and being canceled to be closely related. Still, this measure is likely only an imperfect proxy for a firm's true bogus status. It is imperfect in a few important respects, and the model may have overfitted to each one of those imperfections:

- Firms are canceled only after being targeted by the department for inspection and found bogus. Thus our label is at best "inspected and bogus", and not just "bogus". The model cannot distinguish those two, so could overfit to the "inspected" part, yielding firms that are likely to be inspected by the department anyway. This issue arises from the so-called "one-sided labels" problem discussed in Mittal et al. (2018).

- Firms could be already determined by the department as bogus and set to be canceled, but this is not recorded yet in our data. The model could overfit to these. After all, the firms most likely to be canceled without already having been canceled, are the ones just about to be canceled.

- Firms could be suo-moto canceled even without being bogus. For example due to not filing for a long time. We find some support for this in our data. We see that many firms listed as active had in fact stopped filing returns several months prior to list sharing (see Appendix C.2).

By overfitting to the imperfections, the model could identify firms that are at risk of (suo-moto) cancellation in addition to those that are genuinely bogus. We had no data on which firms were on track to be canceled, so we could not exclude them. So it is not surprising that inspections found firms to be non-existent despite them being listed as active, and that these firms were eventually canceled. Hence the good model accuracy. But targeting those firms for additional inspections would not alter their predestined cancellation trajectory - hence the lack of causal effect on cancellations. Another way to phrase this effect is as a form of selection bias, where our model performs the selection. We formalize this hypothesis in Appendix C.1.

5.1.2 Lacking enforcement action

The second cause is lacking enforcement action by the department, in response to inspection results. It is possible that inspections were either not acted upon by the department or were inaccurate themselves. We find empirical support for limited departmental action to cancel bogus firms. Section 4.6 shows many firms found non-existent upon inspection, which are still not canceled and filing returns months later. Our data covers almost a two year period post inspection. It is possible that the effects of inspections on firm cancellations require an even longer duration to manifest. This highlights some of the pitfalls in implementing predictive models in regulatory settings.

To summarize, some of our list firms seem likely to be canceled irrespective of model-driven inspection activities and for some of our list firms that merited cancellation department action was lacking. These explanations are not mutually exclusive and we find evidence that both are likely in play.

5.2 Implications of our findings

5.2.1 Importance of evaluating ML solutions in the field

Our findings are relevant for ML practitioners who deploy ML models in the field. Applying ML techniques promises substantial advancements in predictive accuracy and cost-efficiency compared to conventional rule-of-thumb methods or even expert opinions. However, robust performance on the test set may not translate to real-world impact. Moreover, these issues can not be uncovered in test set evaluations.

First, we document potential challenges in assimilating new tools into existing operational frameworks. Despite its high accuracy, the machine learning model did not produce a measurable increase in a key tax department outcome—tax collections. Therefore, careful planning and understanding of context is needed in order to ensure take-up and proper functioning. For our purposes, it could have meant following the first list to its natural conclusion (instead of e.g. focusing on sharing newer lists).

Second, we recommend caution in the use of imperfect proxies as labels. The model might overfit to the proxy-label, which in our case was a particular type of canceled firm. The model may then replicate existing procedures and capture what would have occurred irrespective of its deployment. To mitigate this problem, one could also build indicators for cases where we know that the proxy-label is insufficient. For instance, our model appeared to also catch firms whose cancellations were already in process as they had ceased filing returns. To avoid this, we should not have inspected these firms. These firms are very likely to be canceled regardless and our inspections are not sufficient to confirm their bogus nature as they are likely to be non-existent at their stated location.[21]

5.2.2 Simple ICT solutions may have greater return

Given the inefficiencies in firm cancellations and ambiguities in recovery procedures from non-existent firms, simpler technological interventions may offer more immediate and greater returns on investment. First, an area for further research is better information sharing within the tax department and across central and state tax authorities. We have anecdotally observed that (after the detection of a bogus firm) the flagging of beneficiaries to the relevant ward officials within the department is a limiting factor. Given complex bogus trading networks that pass tax credits across jurisdictions, ward officials need timely access to information on whether a firm in their jurisdiction is connected to a verified non-existent firm. Facilitating communication between different tax wards could speed up administrative processes. Similarly, streamlined communication between state and central tax authorities may address bottlenecks in firm cancellation procedures.

Second, a number of important feedback parameters used in the model can expedite gathering relevant transaction information. These important feedback variables can be tracked as standalone dashboards or included in the current ad-hoc reporting system. For example, monitoring firms that do not have any third party reported buyers more closely could provide early indicators of potential problems. See Appendix C.4.2 for details. Finally, a more comprehensive process monitoring system can be established to track a case from the detection of a non-existent firm, identification of beneficiaries, assessment, collection demands and eventual recovery. See Appendix C.4.1 for details.

6 Conclusion

Our study aimed to assess the efficacy of machine learning models in identifying firms likely to be fraudulent and, consequently, subject to cancellation. While our model exhibited strong predictive accuracy in both simulated and field evaluations, we did not observe any impacts on tax collections, a key metric for tax authorities. This underscores two key challenges: first, the presence of a form of selection bias in the training data where the model identifies firms that are already on a trajectory to cancellation, and second, institutional frictions that impeded the quick cancellation of identified bogus firms.

A range of promising research directions emerges from our study. First, future research could revisit this study's framework but with a focus on directly influencing firm cancellations. This could involve either collaborating with the department to randomize postponing firm cancellation, or constructing a

[21] Note that the proxy-label issue does not explain the field impact challenge described above, as we do not find any impact even on firms that were still filing.

better model restricted to still-operating firms (or with true rather than proxy-labels) and implementing randomization similar to our approach. Second, an investigation into revenue recovery post-cancellation could shed light on the fiscal implications of pursuing bogus firms at all. Third, examining the effects on beneficiaries is crucial for a holistic understanding of the model's impact. Specifically, future work should explore whether inspecting and canceling bogus firms displaces fraudulent activity to other bogus firms, or rather deters such activity altogether. Fourth, measuring the effectiveness of the ICT solutions to facilitate department action and cooperation. These avenues not only extend the scope of our current study but also hold the potential for addressing some of the operational and policy-related constraints we encountered.

In summary, our study offers an initial examination of the complexities involved in transferring a machine-learning-based predictive tool to the field in a policy context. While our model excelled in accuracy, both on the test-set and in the field, operational and systemic barriers limited its effectiveness in practice, highlighting the essential need for field evaluations to complement simulated assessments. By illustrating these challenges and proposing avenues for future research, we hope to encourage a more nuanced and effective application of machine learning in public policy contexts.

References

Athey, S., S. A. Cole, S. Nath, and S. J. Zhu (2023): "Targeting, Personalization, and Engagement in an Agricultural Advisory Service," *SSRN Electronic Journal*. 3

Bachas, P., A. Brockmeyer, A. Ferreira, and Sarr (2022): "How to Target Enforcement at Scale: Evidence from Tax Audits from Senegal," Tech. rep. 3

Battaglini, M., L. Guiso, C. Lacava, D. L. Miller, and E. Patacchini (2022): "Refining Public Policies with Machine Learning: The Case of Tax Auditing," . 2

Behl, M. (2021a): "DGGI raids across three states unearth Rs 144 cr bogus billing," The Times of India, accessed: 2021-07-11. 1

——— (2021b): "Ludhiana: DGGI busts Rs 630 crore bogus billing nexus, prominent businessman arrested," The Times of India, accessed: 2021-11-20. 1

Bellon, M., E. Dabla-Norris, S. Khalid, and F. Lima (2022): "Digitalization to Improve Tax Compliance: Evidence from VAT e-Invoicing in Peru," *Journal of Public Economics*, 210, 104661. 3

Bureau, T. H. (2024): "Fake Invoices: 29,000-plus Firms Busted since May 2023," *The Hindu*. 1

Carrillo, P., D. Donaldson, D. Pomeranz, and M. Singhal (2022): "Ghosting the Tax Authority: Fake Firms and Tax Fraud," . 3

Dave, S. (2021): "Input tax credit blocked for even minor lapses," The Economic Times, accessed: 2022-01-19. 3

Dhasmana, I. (2021): "GST technical glitches behind input tax credit frauds: CAG report," Business Standard, accessed: 2021-03-26. 4

Dzansi, J., A. Jensen, D. Lagakos, and H. Telli (2022): "Technology and Tax Capacity: Evidence from Local Governments in Ghana," . 3

Fan, H., Y. Liu, N. Qian, and J. Wen (2020): "Computerizing VAT Invoices in China," . 3

GSTN (2024): "GST Goods and Services Tax," Goods and Services Tax Network, accessed: 2024-03-22. 1

IMF (2018): "In the Trenches," *IMF Finance & Development*. 1

Mahajan, A. and S. Mittal (2017): "GST Explainer: Value Added Tax 2.0," Ideas for India, accessed: 2023-03-08. 3

Mittal, S. and A. Mahajan (2017): "VAT in Emerging Economies: Does Third Party Verification Matter?" *SSRN Electronic Journal*, available at SSRN: https://ssrn.com/abstract=3029963 or http://dx.doi.org/10.2139/ssrn.3029963. 2, 4

MITTAL, S., O. REICH, AND A. MAHAJAN (2018): "Who Is Bogus?: Using One-Sided Labels to Identify Fraudulent Firms from Tax Returns," in *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, Menlo Park and San Jose CA USA: ACM, 1–11. 2, 4, 5, 6, 7, 10, 22

OKUNOGBE, O. AND F. SANTORO (2022): "The Promise and Limitations of Information Technology for Tax Mobilisation," . 3

PRABHAKARAN, G. (2022): "GST input tax credit: Why tasking the recipient with the responsibility of ensuring supplier compliance may be draconian," The Economic Times/Rise, accessed: 2022-02-14. 3

PTI (2021): "GST officers detect Rs 4,000 crore of input tax credit fraud in April-June," The New India Express, accessed: 2022-05-21. 1

SCIKIT-LEARN DEVELOPERS (2021): "Scikit-learn (Python) documentation for RandomForestClassifier," Accessed: 2021-03-08. 7

SHAH, P. (2023): "Ease of GST compliance: Still a distant dream," The Economic Times/Rise, accessed: 2023-11-11. 3

WU, Y., B. DONG, Q. ZHENG, R. WEI, Z. WANG, AND X. LI (2020): "A Novel Tax Evasion Detection Framework via Fused Transaction Network Representation," in *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, 235–244. 2

ZHENG, Q., Y. LIN, H. HE, J. RUAN, AND B. DONG (2019): "ATTENet: Detecting and Explaining Suspicious Tax Evasion Groups," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China: International Joint Conferences on Artificial Intelligence Organization, 6584–6586. 2

A List Construction Details

We weight the ranking of suspicious firms by the square root of the total input tax credits available to the firm over its lifetime. This decreases the hit-rate of the model but we expect to find larger offenders. We also filter for firms above a certain threshold in terms of the probability score. On top of this, we add a few sanity conditions to select firms for physical inspection. The firm needs to be active as per the latest available registration status record and the firm should not have been inspected already.

Besides the firms that we want to inspect for model evaluation, we also share firms that would help us compare the model effectiveness to other non-ML approaches and firms that would help us understand the bogus firm behavior. Below, we describe the construction of each list. With each subsequent list, the probability of being bogus for a given firm steadily decreases as we are moving down the list of suspicious firms. The results should be interpreted with this in mind.

A timeline of when each list was shared with the department and feedback received is in Table A.1 below. The list-wise balance test for randomization follows in Table A.3.

Sr. No.	No. of Firms in List	Date (List Shared)	Date (Results Received)
1	223	March 19, 2021	April 19, 2021
2	300	May 31, 2021	July 15, 2021
3	299	August 03, 2021	January 25, 2022
4	1,057	February 18, 2022	December 27, 2022 *

Table A.1: Timeline of lists shared and results received. *Notes: Only partial results were received for the fourth list.*

A.1 List 1, shared March 2021

List 1 was shared to determine the effectiveness of the model. We applied two filters to the firm population. First, we selected top 2,000 firms ranked by (predicted probability) $*$ sqrt (ITC). Within this group we narrowed to firms ranked in the top-22,000 only by predicted probability. The thresholds were somewhat arbitrary where we were trying to build a big enough list for the first round of inspection.

223 active firms satisfied this criteria. The feedback received from ward officials show that only 49 firms had a legitimate presence at the declared place of business with evident activity. 49 firms already had cancellation processes pending, either initiated by the taxpayer or the department. Remaining 125 firms had no physical presence.

A.2 List 2, shared May 2021

In List 2, for the model evaluation, we increased thresholds from 2,000 to 3,000 for firms ranked by (predicted probability) $*$ sqrt (ITC), and from 22,000 to 40,000 for only predicted probability. 471 active firms satisfied this criteria. We shared the top 250 ordered by sqrt-ITC with the department.

We also added a small sample of firms to evaluate a simple non-ML approach for a baseline comparison of the model effectiveness. We selected firms that use large amounts of ITC but do not have a commensurate amount of tax paid in cash. This is an important adhoc rule used by the department. We first filtered out firms that are not in the top-5,000 taxpayers in the state by tax paid in cash over the firm's lifetime. We then selected firms that had claimed total lifetime ITC greater than the median of the total ITC claimed (approximately ₹9.8 million) by the 223 firms in List 1. Out of these firms, we selected top-50 firms in descending order of ratio of total ITC claimed to total reported sales amount across the entire firm lifetime. We label this set of criteria as Turnover-ITC.

A.3 List 3, shared August 2021

To continue the model evaluation, we included the remaining 221 firms that were not shared of the 471 firms in List 2. Of these, 105 firms (47.5% of total) were found to be non-existent or non-functional and another 16 firms that fit the criteria had been canceled or suspended before the list was shared.

We also wanted to evaluate firms that sell to the bogus firms. Our prior belief was that firms that sell to non-existent firms are more likely to be bogus themselves. We focused on trading partners of firms that were found non-existent in the first 2 lists. We then filtered firms with positive ITC claimed as per the consolidated returns because firms with zero ITC are not directly passing fake credits. We then split the group into two categories. We selected the top 65 firms by sqrt-ITC rank and categorized them as high risk sellers to bogus firms. We selected the bottom 25 firms by sqrt-ITC rank and categorized them as low risk sellers to bogus firms.

Finally, we included 4 firms which were registered on the same PAN number as one of the earlier detected non-existent firms. All 4 of these firms were found non-existent.

A.4 List 4, shared February 2022

To continue the model evaluation, we included 296 firms by increasing the thresholds from 3,000 to 4,000 for firms ranked by (predicted probability) $*$ sqrt (ITC) AND from 40,000 to 51,000 for only predicted probability. We also added 381 firms by not using the ITC weighted rank. Specifically, we included firms in the top 45,000 ranked by model probability with lifetime ICT over ₹1 million.

The current model predictions optimize for accuracy by averaging over the entire history of tax returns available till date. However, department experts hypothesized that a typical bogus firm files most of its ghost transactions over the first few return periods and subsequently either goes dormant or files for de-registration. To evaluate our model efficacy for such a hypothesis, we simulated the prediction using only a limited number of returns. Specifically, we construct 3 groups, of 300 firms each, using the first 3 returns only, first 6 returns only, first 9 returns only.

Additionally, we shared 50 randomly selected firms registered on the mobile number or email address of a known non-existent firm each (100 firms total). Finally, we included randomly select 100 firms, to construct a baseline.

A summary of the goal of each group, method of selection, expected results and size is listed in Table A.2 below. Unfortunately, we did not receive complete feedback for any of the above groups. As a result, we can only make limited inference about the effectiveness of these channels.

Goal	Method of selection	Expected accuracy	Size
Establish a baseline for comparison	Firms with same registration parameters (mobile number, email) as known bogus firms, randomly selected	Low	100 firms
	Any firm not yet inspected, randomly selected	Very low	100 firms
Test effectiveness on early detection	High-risk firms utilizing only early returns	Moderate	330 firms
Test model effectiveness	High-risk firms utilizing all available returns	High	527 firms

Table A.2: Details of subgroups selected for List-4.

Table A.3: List-wise balance tests for randomization between Shadow lists and Main lists.

Firm Attribute	**List-1**		**Shadow-1**		**Difference (Main - Shadow)**	
	Mean	St. Dev	Mean	St. Dev	Coef.	p-value
Model Score	0.10	0.05	0.08	0.04	0.02	0.10
Registration Month	21.83	11.47	21.26	11.31	0.57	0.84
If registered with State authorities	22.87%	42.09%	36.84%	49.56%	-13.97%	0.27
Turnover, till FY20	232.2M	1,227.8M	119.4M	86.8M	112.8M	0.40
If Turnover (till FY20) is not available	4.48%	20.74%	5.26%	22.94%	-0.78%	1.00
Total Tax Paid (Cash), till FY20	0.6M	4.9M	0.2M	0.8M	0.4M	0.93
Total ITC Claimed, till FY20	23.8M	76.3M	15.4M	12.4M	8.3M	0.55
Tax-to-Turnover Ratio, till FY20	0.41%	4.11%	0.13%	0.49%	0.00	0.91
log(Turnover, till FY20)	18.29	1.08	18.35	0.75	-0.05	0.84
log(Total Tax Paid (Cash), till FY20)	6.60	4.37	6.52	4.77	0.08	0.93
log(Total ITC Claimed, till FY20)	16.24	0.98	16.30	0.73	-0.06	0.79
log(Tax-to-Turnover Ratio, till FY20)	0.00	0.03	0.00	0.00	0.00	0.91
No. of Monthly Sales Statements Filed, till FY20	5.75	5.31	6.53	4.79	-0.78	0.54
No. of Consolidated Returns Filed, till FY20	6.04	5.27	6.79	4.22	-0.75	0.54
If registered as Composition firm	0.00%	0.00%	0.00%	0.00%	0.00%	
If migrated from VAT to GST	11.21%	31.62%	15.79%	37.46%	-4.58%	0.67
If registered as proprietorship, partnership or Pvt Ltd firm	95.96%	19.72%	94.74%	22.94%	1.23%	1.00
N. Observations	List-1 (Main)	223	List-1 (Shadow)	19		

(a) List-1 Firms.

Firm Attribute	**List-2**		**Shadow-2**		**Difference (Main - Shadow)**	
	Mean	St. Dev	Mean	St. Dev	Coef.	p-value
Model Score	0.04	0.01	0.04	0.01	0.00	0.21
Registration Month	17.02	8.25	18.31	9.00	-1.29	0.35
If registered with State authorities	25.67%	43.77%	33.33%	47.76%	-7.66%	0.34
Turnover, till FY20	438.5M	1,253.5M	689.3M	2,611.0M	250.8M	0.30
If Turnover (till FY20) is not available	0.00%	0.00%	2.56%	16.01%	-2.56%	0.03
Total Tax Paid (Cash), till FY20	1.2M	4.2M	3.7M	13.9M	2.6M	0.03
Total ITC Claimed, till FY20	60.3M	97.4M	91.9M	299.5M	31.7M	0.16
Tax-to-Turnover Ratio, till FY20	0.57%	4.75%	0.32%	0.87%	0.00	0.69
log(Turnover, till FY20)	19.21	1.01	19.06	1.23	0.15	0.41
log(Total Tax Paid (Cash), till FY20)	9.81	4.11	10.34	4.07	-0.53	0.45
log(Total ITC Claimed, till FY20)	17.42	0.94	17.26	1.13	0.16	0.34
log(Tax-to-Turnover Ratio, till FY20)	0.00	0.04	0.00	0.01	0.00	0.71
No. of Monthly Sales Statements Filed, till FY20	13.70	7.71	13.15	8.28	0.54	0.70
No. of Consolidated Returns Filed, till FY20	15.48	7.70	14.23	8.22	1.25	0.36
If registered as Composition firm	0.00%	0.00%	0.00%	0.00%	0.00%	
If migrated from VAT to GST	6.90%	25.39%	10.26%	30.74%	-3.36%	0.50
If registered as proprietorship, partnership or Pvt Ltd firm	88.89%	31.49%	89.74%	30.74%	-0.85%	1.00
N. Observations	List-2 (Main)	261	List-2 (Shadow)	39		

(b) List-2 Firms.

Firm Attribute	**List-3**		**Shadow-3**		**Difference (Main - Shadow)**	
	Mean	St. Dev	Mean	St. Dev	Coef.	p-value
Model Score	0.05	0.03	0.05	0.03	0.00	0.71
Registration Month	22.18	8.62	18.06	9.44	4.12	0.07
If registered with State authorities	27.56%	44.78%	25.00%	44.72%	2.56%	1.00
Turnover, till FY20	73.6M	68.2M	65.0M	52.0M	8.6M	0.64
If Turnover (till FY20) is not available	2.22%	14.77%	0.00%	0.00%	2.22%	1.00
Total Tax Paid (Cash), till FY20	0.3M	1.1M	1.0M	2.7M	0.8M	0.04
Total ITC Claimed, till FY20	9.1M	6.4M	9.4M	4.5M	0.3M	0.87
Tax-to-Turnover Ratio, till FY20	0.76%	6.73%	59.78%	237.96%	-0.59	0.04
log(Turnover, till FY20)	17.67	1.04	17.45	1.44	0.22	0.45
log(Total Tax Paid (Cash), till FY20)	7.83	4.59	7.72	5.02	0.12	0.93
log(Total ITC Claimed, till FY20)	15.71	0.89	15.81	0.93	-0.10	0.68
log(Tax-to-Turnover Ratio, till FY20)	0.01	0.05	0.15	0.59	-0.14	0.03
No. of Monthly Sales Statements Filed, till FY20	8.53	6.51	11.94	7.18	-3.41	0.05
No. of Consolidated Returns Filed, till FY20	9.88	6.85	13.50	7.33	-3.62	0.04
If registered as Composition firm	0.89%	9.41%	0.00%	0.00%	0.89%	1.00
If migrated from VAT to GST	5.33%	22.52%	12.50%	34.16%	-7.17%	0.24
If registered as proprietorship, partnership or Pvt Ltd firm	96.89%	17.40%	87.50%	34.16%	9.39%	0.03
N. Observations	List-3 (Main)	225	List-3 (Shadow)	16		

(c) List-3 Firms.

Firm Attribute	**List-3**		**Shadow-3**		**Difference (Main - Shadow)**	
	Mean	St. Dev	Mean	St. Dev	Coef.	p-value
Model Score	0.03	0.01	0.03	0.02	0.00	0.05
Registration Month	18.50	9.77	17.45	9.81	1.05	0.33
If registered with State authorities	40.42%	49.11%	37.35%	48.67%	3.07%	0.61
Turnover, till FY20	175.8M	721.4M	116.4M	267.2M	59.5M	0.42
If Turnover (till FY20) is not available	1.20%	10.89%	2.41%	15.43%	-1.21%	0.63
Total Tax Paid (Cash), till FY20	0.7M	3.3M	0.6M	2.4M	0.1M	0.88
Total ITC Claimed, till FY20	22.2M	64.7M	15.0M	34.1M	7.2M	0.28
Tax-to-Turnover Ratio, till FY20	0.59%	1.73%	0.63%	1.78%	0.00	0.83
log(Turnover, till FY20)	17.60	1.56	17.54	1.36	0.06	0.74
log(Total Tax Paid (Cash), till FY20)	9.19	4.15	8.64	4.75	0.56	0.27
log(Total ITC Claimed, till FY20)	15.58	1.43	15.40	1.39	0.18	0.28
log(Tax-to-Turnover Ratio, till FY20)	0.01	0.02	0.01	0.02	0.00	0.83
No. of Monthly Sales Statements Filed, till FY20	11.03	7.84	10.87	7.37	0.16	0.87
No. of Consolidated Returns Filed, till FY20	13.08	7.88	13.72	7.57	-0.64	0.48
If registered as Composition firm	0.60%	7.72%	2.41%	15.43%	-1.81%	0.10
If migrated from VAT to GST	10.78%	31.03%	10.84%	31.28%	-0.06%	0.94
If registered as proprietorship, partnership or Pvt Ltd firm	100.00%	0.00%	100.00%	0.00%	0.00%	
N. Observations	List-4 (Main)	668	List-4 (Shadow)	83		

(d) List-4 Firms.

B Additional Results

B.1 Model performance

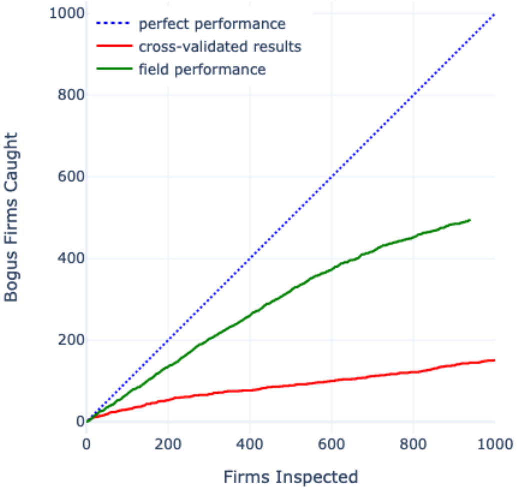


Figure A.1: ***Model performance on top 938 predictions.*** *Notes: Ordered by Predicted Probability weighted by Sq. Root ITC Claimed*

Table A.4: Characteristics for top-N firms ordered by model score.

Firms Inspected	Bogus Firms Caught	Average Predicted Probability	Bogus Firms Caught (per inspection)	Cumulative Input Credits Claimed
10	9	0.2886	0.90	116.3M
20	18	0.2485	0.90	315.9M
50	43	0.1878	0.86	534.9M
100	83	0.1492	0.83	938.0M
200	155	0.1155	0.78	1,901.0M
500	329	0.0764	0.66	6,266.9M
938	494	0.0536	0.53	11,071.1M

(a) top-N Field-Inspected Firms

Firms Inspected	Bogus Firms Caught	Average Predicted Probability	Bogus Firms Caught (per inspection)	Cumulative Input Credits Claimed
10	10	0.7178	1.00	0.0M
20	20	0.6470	1.00	0.0M
50	38	0.5168	0.76	405.4M
100	67	0.4298	0.67	734.6M
200	124	0.3658	0.62	834.9M
500	284	0.3070	0.57	1,156.9M
938	445	0.2766	0.47	1,232.1M

(b) top-N Firms Selected by Cross-Validated Scores

Table A.5: Characteristics for top-N firms ordered by model score & weighted by sq. root of ITC claimed.

Firms Inspected	Bogus Firms Caught	Average Predicted Probability	Bogus Firms Caught (per inspection)	Cumulative Input Credits Claimed
10	6	0.2211	0.60	205.5M
20	15	0.1823	0.75	946.0M
50	35	0.1356	0.70	2,098.8M
100	68	0.1077	0.68	3,814.3M
200	135	0.0886	0.68	5,714.2M
500	321	0.0665	0.64	9,706.6M
938	494	0.0536	0.53	11,071.1M

(a) top-N Field-Inspected Firms

Firms Inspected	Bogus Firms Caught	Average Predicted Probability	Bogus Firms Caught (per inspection)	Cumulative Input Credits Claimed
10	7	0.3934	0.70	467.8M
20	12	0.3780	0.60	770.0M
50	19	0.3131	0.38	999.4M
100	31	0.2525	0.31	1,752.7M
200	53	0.2026	0.27	2,885.6M
500	89	0.1451	0.18	3,858.6M
938	145	0.1144	0.15	4,910.0M

(b) top-N Firms Selected by Cross-Validated Scores

B.2 Model performance segmented by approval authority

For all subsequent results, we label firms registered with the State Tax Authority as 'state-registered' and those registered with the Central Tax Authority as 'center-registered'. Table A.6 shows the overall split of the training data, and firms labeled as bogus during training, between the two tax authorities. A breakdown of the inspection results by tax authority follows in Table A.7.

	State	Center	% State
Training Data	307,362	141,304	68.51%
Labeled Set	4,293	544	88.75%

Table A.6: State-Center split for all firms in training data. *Notes: Our labeled set is dominated by state jurisdiction firms (89%). Yet our predictions are dominated by the center (60% vs 35%). This is despite 69% of our training data being from state jurisdiction firms.*

	Center-registered firms			State-registered firms			All firms			
Group	**count**	**n. of bogus**	**share bogus**	**count**	**n. of bogus**	**share bogus**	**count**	**share state**	**n. of bogus**	**share bogus**
Model Score	653	395	60.50%	285	99	34.70%	938	30.3%	494	52.70%
Random	20	1	5.00%	29	1	3.40%	49	59.2%	2	4.10%
Rule of Thumb	23	10	43.50%	27	9	33.30%	50	54%	19	38.00%

Table A.7: State-Center split for inspection results. *Notes: There are large differences in share of bogus firms between the two authorities. We have left the investigation of the effect of approval authority on enforcement outcomes for a later exercise.*

Identifier	No. of firms	No. of firms inspected	No. of firms found bogus
PAN	4	4	4
Email	7,477	18	2
Mobile	3,327	12	2

Table A.8: Inspection results for firms matched on key variables. *Notes: We randomly inspected firms with common identifiers (PAN, mobile or email) as our field-verified bogus firms. The hit-rate was 100% for firms with common PAN but at a small sample (4 firms). We are surprised at the considerably lower hit-rate of about 12% for the other two identifiers, even though they were randomly selected. We believe this may be due to the common use of shared mobiles/emails for registering firms. Therefore, simple random inspection is not effective. We propose ranking firms based on a combination of these features for inspection and tracking. We updated our model with these combination features and found them to be the most important features out of 556 features.*

B.3 Cancellation timing

KS Test Results	N. Obs. (main)	N. Obs. (shadow)	Statistic	p-value	Statistic Location	Statistic Sign
List-1, Shadow-1	223	19	0.320	0.043	Dec 2021	-1
List-2, Shadow-2	261	39	0.140	0.473	Jul 2022	-1
List-3, Shadow-3	225	16	0.212	0.447	Dec 2022	1
List-4, Shadow-4	668	83	0.074	0.781	Dec 2022	-1
Combined	1377	157	0.062	0.616	Dec 2022	-1

Table A.9: Kolmogorov-Smirnov tests assessing equality of cancellation timing distributions between Main and Shadow lists. *Notes: Statistical significance at the 5% level is only observed for List-1, where, counterintuitively, more shadow firms are canceled than main firms. The test statistic is the maximal difference in absolute value between the CDFs, the Statistic Location is the month when that maximal difference was observed, and the Statistic Sign is the sign of that maximal difference. All uncanceled firms were attributed a cancellation time exceeding all recorded data points.*

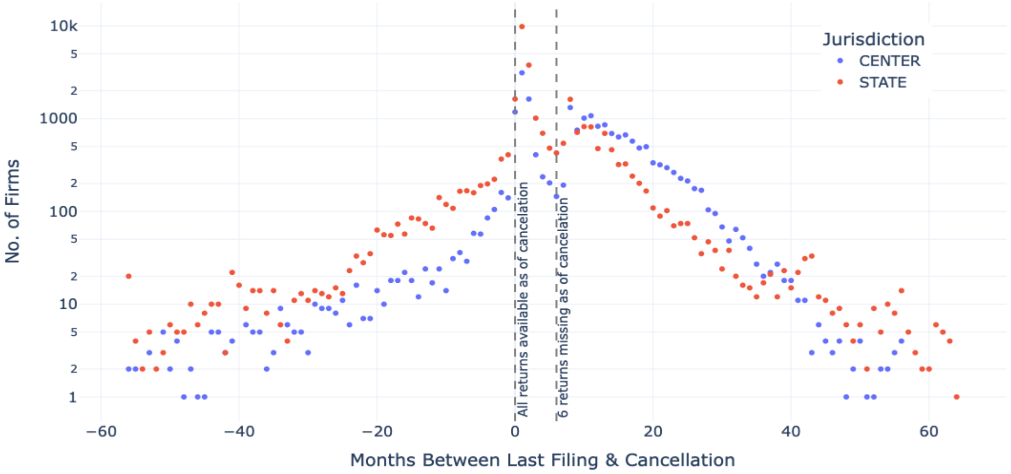


Figure A.2: ***Distribution of cancellation delay.*** *Notes: For all suo-moto canceled firms, we measure the delay between last filing and date of effective cancellation. We then plot the number of firms by delay month. Right of the first dotted line indicates the time that has lapsed between a firms' last return and when it was canceled. The firms to the left of the first dotted line are canceled with retroactive effect. By doing this, the tax officials invalidate any transactions carried out by the offending firm and provide a basis for further prosecution. We see that state jurisdiction does an order of magnitude more cancellations with retroactive effect. Center does cancellations For our modeling exercise, we use an indicator for whether the entire transaction history of a firm was deemed invalid (through retroactive cancellation with effect from date of registration) as a proxy variable for the firm being bogus.*

B.4 Inspection results and revenue outcomes for trading partners

	All Bogus Firms			Bogus firm is State-Registered			Bogus firm is Center-Registered		
	Approval Authority for Buyer			Approval Authority for Buyer			Approval Authority for Buyer		
	All (1)	State (2)	Center (3)	All (4)	State (5)	Center (6)	All (7)	State (8)	Center (9)
count	23,509	14,827	8,682	17,895	11,789	6,106	6,512	3,541	2,971
mean	0.00722	0.00590	0.00948	0.00572	0.00576	0.00562	0.01132	0.00615	0.01747
std	0.02284	0.02214	0.02382	0.02156	0.02253	0.01957	0.02495	0.01981	0.02877
25th %ile	0.00004	0.00003	0.00008	0.00003	0.00002	0.00005	0.00013	0.00008	0.00030
50th %ile	0.00026	0.00017	0.00059	0.00019	0.00014	0.00034	0.00071	0.00035	0.00289
90th %ile	0.01806	0.01207	0.02818	0.01179	0.01134	0.01252	0.03822	0.01515	0.05580
99.9th %ile	0.12629	0.12740	0.12502	0.12920	0.13340	0.10954	0.11878	0.10690	0.12663
max	0.32160	0.32160	0.22150	0.32160	0.32160	0.22150	0.27880	0.27880	0.21800

(a) for Firms that Purchased from a Bogus Firm (Buyer)

	All Bogus Firms			Bogus firm is State-Registered			Bogus firm is Center-Registered		
	Approval Authority for Supplier			Approval Authority for Supplier			Approval Authority for Supplier		
	All (1)	State (2)	Center (3)	All (4)	State (5)	Center (6)	All (7)	State (8)	Center (9)
count	5,242	2,691	2,551	2,541	1,428	1,113	3,258	1,517	1,741
mean	0.01540	0.00945	0.02168	0.00880	0.00633	0.01197	0.02119	0.01294	0.02838
std	0.03303	0.02948	0.03533	0.02460	0.02275	0.02648	0.03773	0.03536	0.03827
25th %ile	0.00010	0.00006	0.00020	0.00006	0.00005	0.00011	0.00017	0.00009	0.00044
50th %ile	0.00061	0.00031	0.00184	0.00034	0.00023	0.00056	0.00136	0.00047	0.01052
90th %ile	0.05585	0.01971	0.07195	0.02544	0.01125	0.04432	0.07332	0.04187	0.08171
99.9th %ile	0.21906	0.21360	0.22862	0.18459	0.17117	0.18515	0.25274	0.24757	0.24550
max	0.32430	0.32430	0.26100	0.32430	0.32430	0.18720	0.32430	0.32430	0.26100

(b) for Firms that Sold to a Bogus Firm (Supplier)

Table A.10: Model score distribution for partners of bogus firms.

Approval Authority of Bogus Firms	No. of Bogus Firms	Buyers			Suppliers		
		State registered	Center registered	% State registered	State registered	Center registered	% State registered
State-registered	91	11,998	6,410	65.2%	1,436	1,163	55.3%
Center-registered	387	3,585	3,240	52.5%	1,551	2,019	43.4%

Table A.11: Center-State registration split of bogus firm trading partners. *Notes: (Buyers) While a significant share of inspected bogus firms are registered with the center authorities, their beneficiaries (i.e., firms that buy from bogus firms) are more evenly split across the state and center authorities. We find that 3,585 of the 6,825 total buyers from center-registered bogus firms are state-registered. These state-registered buyers have received a total of ₹511.2 Cr. as input tax credits from the center-registered bogus firms. (Suppliers) A firm may be supplying to multiple bogus firms, both center- and state-registered ones. As such, there may be overlap between firms counted in the two rows in.*

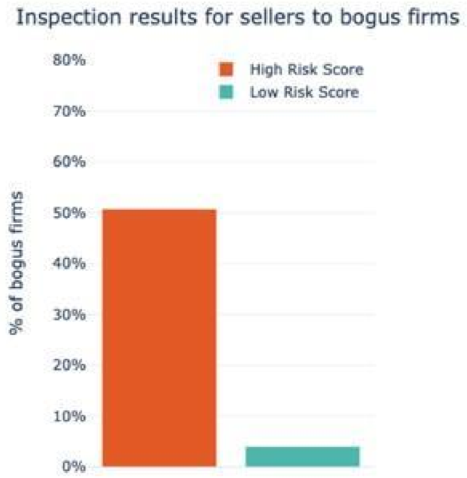


Figure A.3: ***Inspection results for sellers to bogus firms.*** *Notes: Suppliers to Bogus firms are risky and our model is able to find them - 52% (34/65) of risky sellers to bogus firms that were inspected also turned out to be bogus. In contrast only 1 out of the 25 low risk sellers were bogus in our inspections. (See Appendix A.3 for details on how these 65 firms were chosen.)*

		State-Registered Firms Only	All Firms (Baseline)	Buyers	Sellers
Notices	(2021 - 2023)	**Scrutiny (2021-2023)**	**6.69%**	**18.79%**	**32.70%**
		Show Cause Notice (2021-2023)	**1.29%**	**4.66%**	**7.76%**
		Additional Tax Paid (ITC Mismatch)	0.53%	1.19%	1.71%
		Additional Tax Paid (Liability Mismatch)	0.24%	0.41%	0.49%
		Additional Tax Paid (Reconciliation)	0.24%	0.66%	1.46%
		Additional Tax Paid (against Notice)	0.56%	1.72%	2.64%
		Additional Tax Paid (against Scrutiny)	0.46%	1.08%	1.58%
		Additional Tax Paid (Voluntary)	**4.21%**	**9.98%**	**14.34%**
	(2017 - 2021)	Scrutiny (2017-2021)	0.37%	1.19%	2.01%
		Show Cause Notice (2017-2021)	0.71%	2.52%	4.80%
ITC Blocked		**Net ITC Blocked >0 (2023)**	**1.36%**	**4.22%**	**6.08%**
		Tax Credits Blocked (2021-2023)	2.22%	7.46%	11.01%
		Tax Credits Blocked (2017-2021)	1.53%	6.12%	8.72%

Table A.12: Recovery outcomes for trading partners of firms labeled as bogus in the training data. *Notes: The tables report the percentages for three different sets of firms (a) all firms (as of March 2021), (b) buyers from firms labeled as bogus in our training data, and (c) sellers to firms labeled as bogus in the training data. We restrict our sample to firms registered with the state authorities, as the data on these outcomes is only available for this sub-sample. First, we note that relative to the population rates (for e.g. scrutiny or show cause notice or voluntary tax payment) the rates for the trading partners of bogus firms are higher, typically much higher, than those for all firms. Second, the rates for sellers to bogus firms are higher than those for buyers from bogus firms. This is further evidence that sellers to bogus firms appear to be more suspicious than buyers from bogus firms.*

B.5 Geographic distribution of inspection results and trading partners

Inspected bogus firms are concentrated in two districts We examined the geographic distribution of identified bogus firms and found that these are highly concentrated in two key districts (Districts 5 and 12, see Figure A.4). This is an interesting finding given that this distribution does not coincide with the corresponding distribution for firms in the training data (Figure A.7).

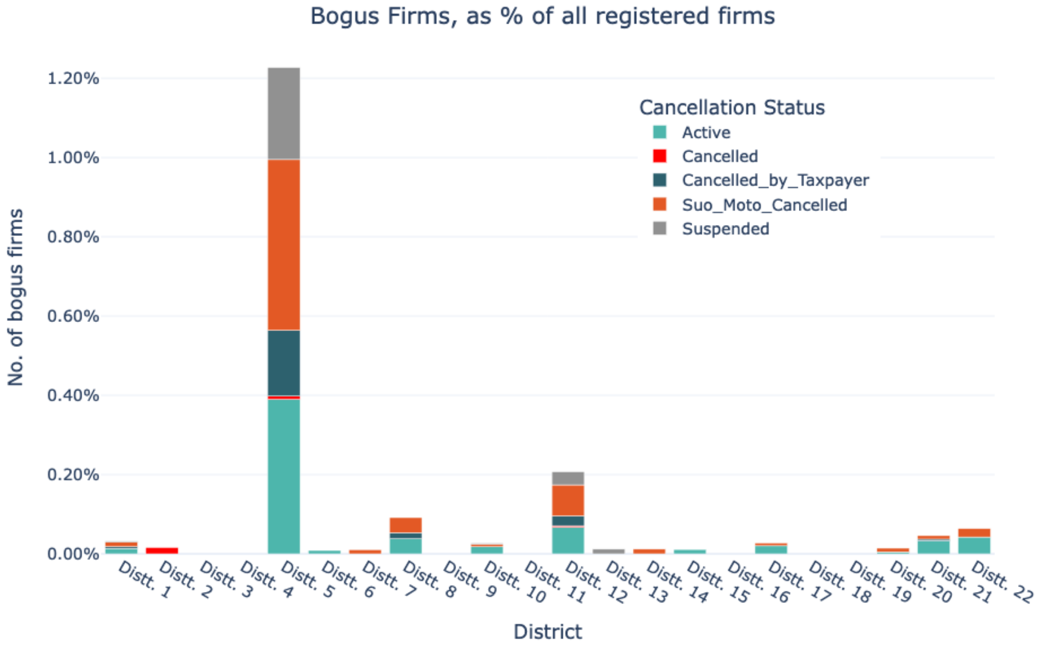
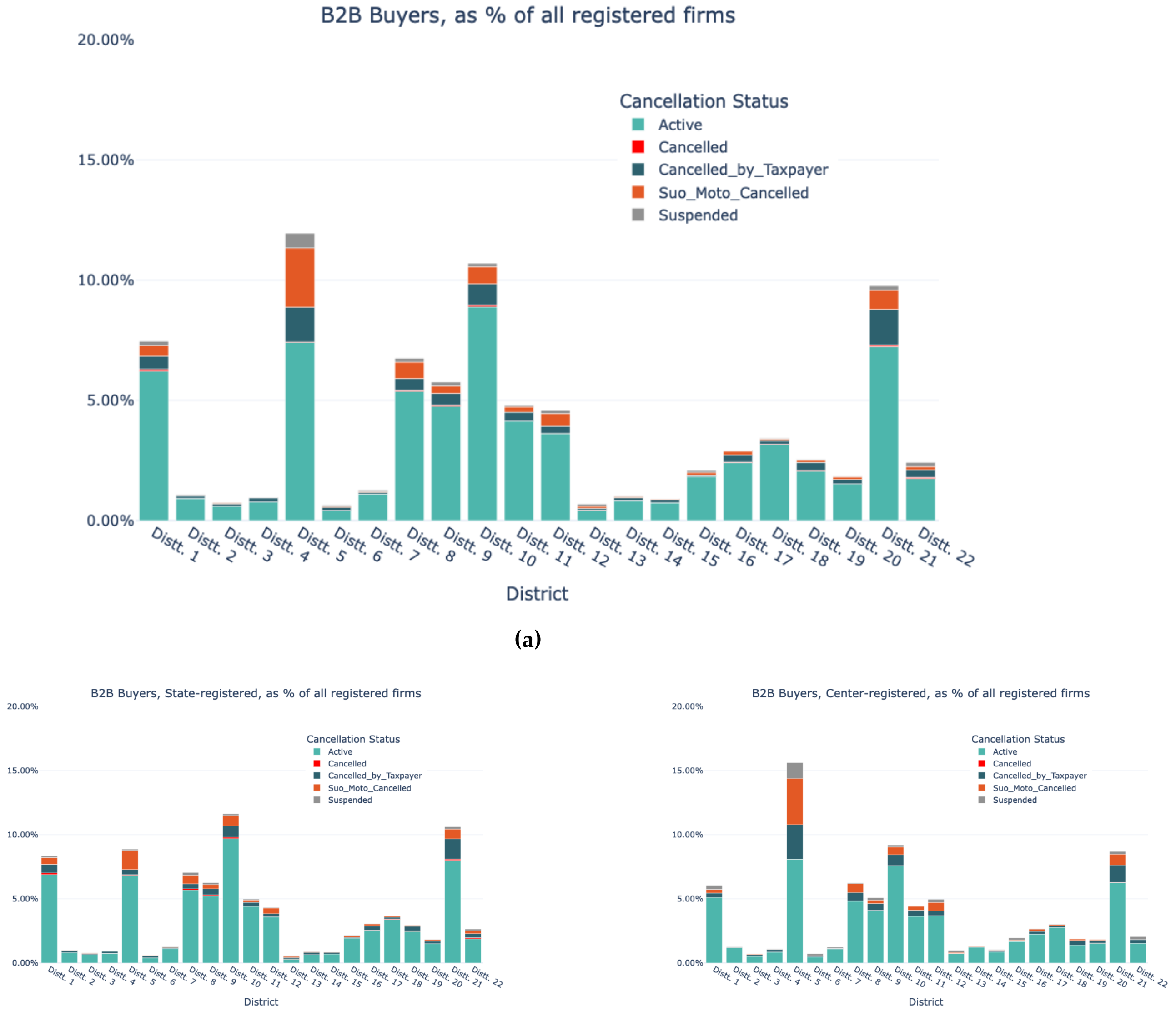


Figure A.4: ***Bogus firms as % of all registered firms.*** *Notes: Inspected and identified bogus firms are highly concentrated in Districts 5 and 12.*

Buyers from bogus firms located all over the state Next, we examine trading partners of firms identified as bogus during our inspections. As noted above, there is a high concentration of inspected bogus firms in a couple of districts. However, the location of their trading partners reveals interesting patterns. There is substantial dispersion in the location of the buyers from our identified bogus firms. We find the buyers to be evenly spread across the state (see Figure A.5). This suggests that direct beneficiaries (i.e., firms that are claiming input tax credit against purchases from bogus firms) interact with bogus firms across districts quite regularly and this re-enforces the need for inter-ward communication to flag trading partners once a bogus firm is identified.



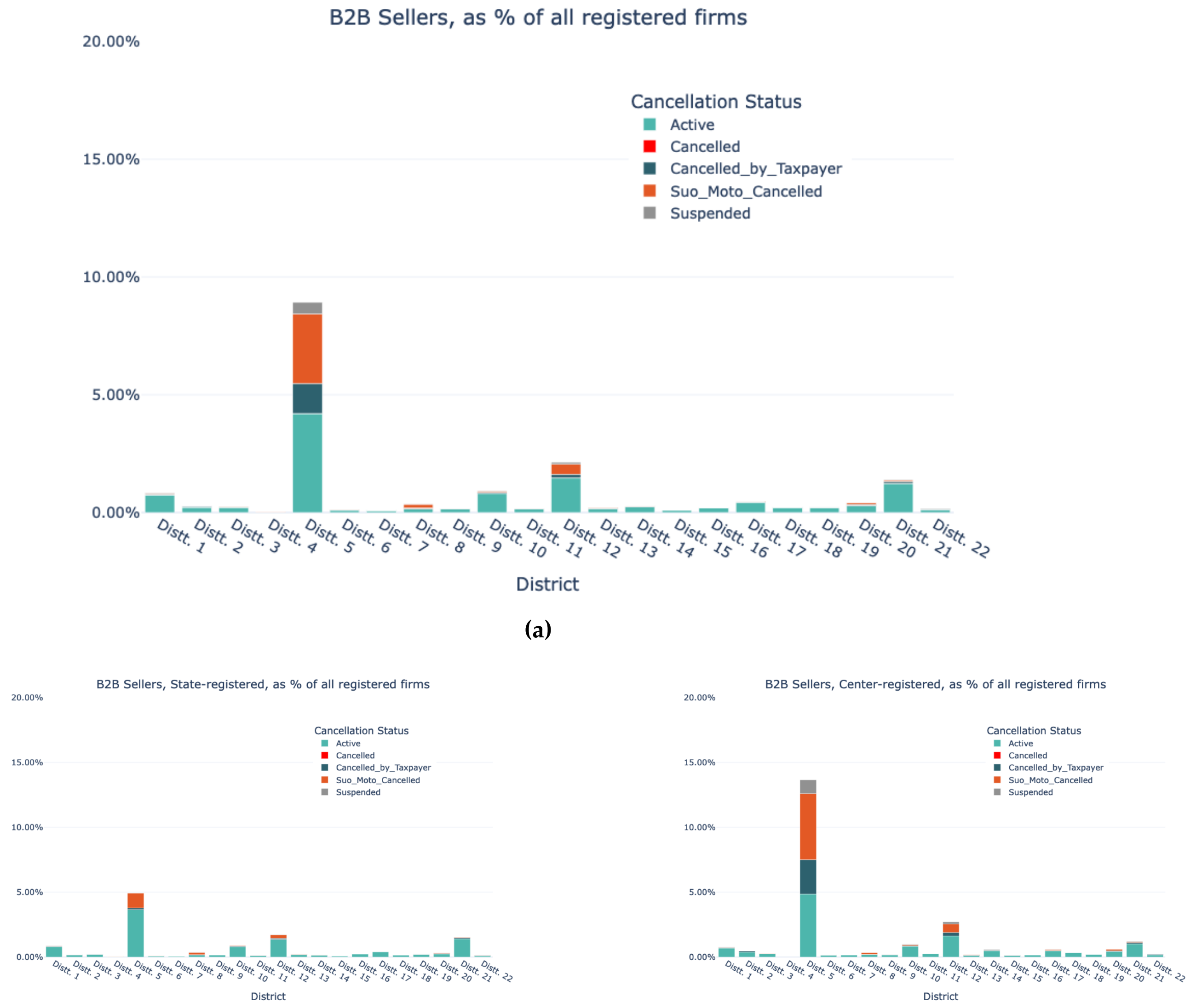
(a)

(b)

(c)

Figure A.5: ***B2B buyers from bogus firms as*** $\%$ ***of all registered firms.*** *The top panel (a) shows the distribution across districts, irrespective of approval authority. The bottom-left (b) and bottom-right (c) panels show the distribution for state-registered and center-registered buyers respectively.*

Suppliers to bogus firms concentrated in 2 districts The high incidence in Districts 5 and 12 is similar to the spatial distribution of the bogus firms themselves. This is interesting because it suggests that these sellers to bogus firms are perhaps likely to be bogus themselves.



(a)

(b)

(c)

Figure A.6: ***B2B suppliers to bogus firms as % of all registered firms.*** *The top panel (a) shows the distribution across districts, irrespective of approval authority. The bottom-left (b) and bottom-right (c) panels show the distribution for state-registered and center-registered buyers respectively. The distribution of sellers is more concentrated in two districts for bogus firms that are center-registered.*

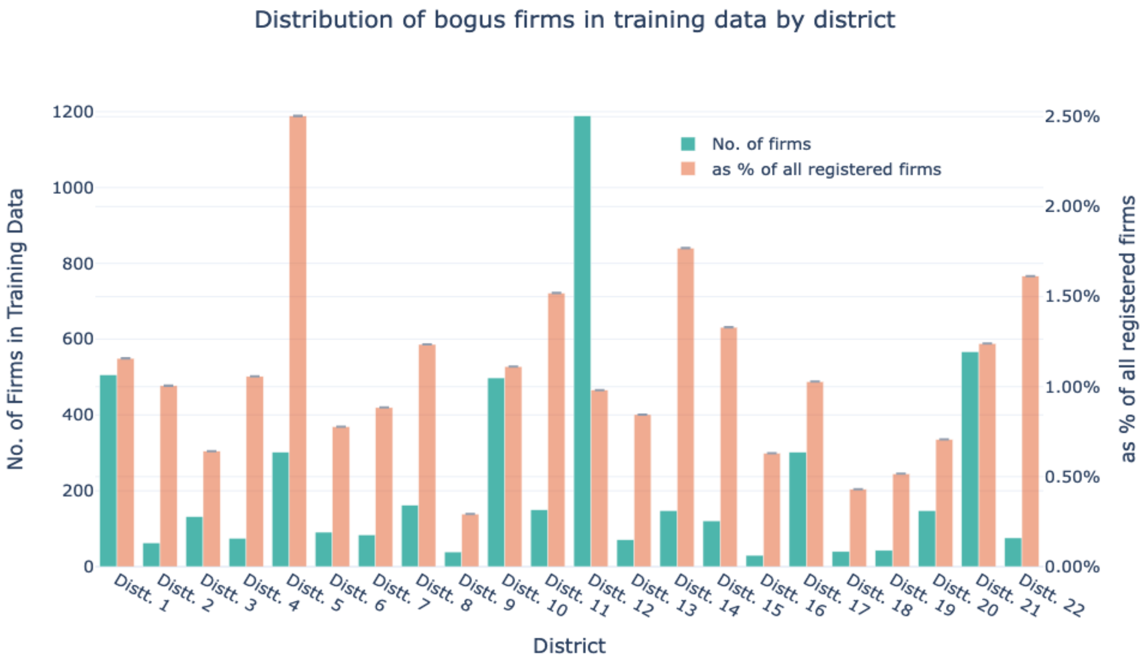


Figure A.7: ***Distribution of bogus firms by district in training data.*** *Notes: The training data do not follow the same geographic distribution as bogus firms in our lists. In particular, while the fraction of such firms is high in District 12 (as is the case with our identified bogus firms), the same is not true for District 5. Recall that our definition of bogus firms in the training data is suo moto canceled firms w.e.f. date of registration.*

C Additional Discussions

C.1 Potential outcomes framework describing selection bias

To formalize our finding that most of our observed effect is due to selection bias, we adopt the potential outcomes framework and introduce the following notation:

- Y represents the observed cancellation outcome: cancellation status, or sometimes cancellation time. Y(1) and Y(0) are the underlying potential outcomes conditional on being inspected or not. For example, $Y_i(1)$ is the cancellation status of firm *i* if it were inspected, and $Y_i(0)$ the cancellation status if it were not.
- D is an (observed) indicator, where D=1 means that a firm is risky according to the model, and D=0 means it is not.
- W is an (observed) indicator of random holdout. W=1 indicates the main set, and W=0 the holdout set. W is randomly assigned to all firms with D=1. Although W is only assigned and observed for risky firms (i.e. for those with D=1) we can conceptually also allow it to be defined for non-risky firms as well (D=0). This latter will be helpful in some of the calculations below. To summarize: (D=1, W=1) characterizes a firm that was risky and whose identity was shared with the department (and hence it was inspected, i.e. was in Lists 1-4). (D=1, W=0) means a firm was risky (D=1) but was randomly placed in the holdout set, i.e. was included in the Shadow Lists 1-4. A firm is inspected if and only if both D=1 and W=1. (D=0,W=w) characterizes a non-risky firm with list status w. Since W is independent of all potential outcomes as well as D, this additional indexing by w will not matter when we compute expectations (see below).
- ATT refers to the Average Treatment effect on the Treated E(Y(1)-Y(0)|D=1).

We decompose the total observed difference (on the main set) as follows:

$ObservedDifference = E[Y|D=1, W=1] - E[Y|D=0, W=1]$

$= E[Y(1)|D=1, W=1] - E[Y(0)|D=0, W=1]$

Here, the equality is owing to the fact that for firms in the main set (W=1), their observed outcome is the potential outcome by D, since D determines the inspections. Next, firms with D=0 are not inspected so their observed cancellation outcome is Y(0).

Since W is randomly assigned, it is independent of the potential outcomes, so the equation can further be simplified to

$... = E[Y(1)|D=1] - E[Y(0)|D=0]$

$= (E[Y(1)|D=1] - E[Y(0)|D=1]) + (E[Y(0)|D=1] - E[Y(0)|D=0])$

$= ATT + SelectionBias$

The Observed Difference (defined above) can thus be decomposed into the sum of the ATT and the difference in cancellation between risky and non-risky firms, even absent our inspections, described in Section 4.3. This last part is what we refer to as “selection bias".

The ATT can be estimated by comparing cancellation rate between main lists and shadow lists. Since W is randomly assigned, it is independent of potential outcomes. Thus,

$ATT = E[Y(1) - Y(0)|D = 1] = E[Y(1)|D = 1, W = 1] - E[Y(0)|D = 1, W = 0]$

But since W determined inspection for risky firms, i.e. those with D=1, we have:

$E[Y(1)|D = 1, W = 1] = E[Y|D = 1, W = 1]$

$E[Y(0)|D = 1, W = 0] = E[Y|D = 1, W = 0]$

So the ATT becomes:

$ATT = E[Y|D = 1, W = 1] - E[Y|D = 1, W = 0]$

Now $E[Y|D = 1, W = 1]$ is just the expected outcome over the main lists, and $E[Y|D = 1, W = 0]$ is the expected outcome over the shadow lists. We estimate both using the average of observed values.

Section 4.3 showed a large Observed Difference in cancellations between list firms and baseline firms. Section 4.4 used the comparison between main lists and shadow lists to show that the ATT for both cancellation status and cancellation time is very small, statistically indistinguishable from 0. Section 4.5 exploited variation in the timing of list sharing to again show that the ATT for cancellation time is small, and not always in the expected direction.

Despite observing large and significant differences between listed and baseline firms in terms of cancellations, our decomposition suggests that this is almost entirely attributable to selection bias. For example, for cancellation status, using the point estimates and standard errors from Section 4.4 and Section 4.3 we get:

$SelectionBias = ObservedDifference - ATT$

$\frac{SelectionBias}{ObservedDifference} = 1 - (\frac{ATT}{ObservedDifference}) = 1.17 \pm 0.28$

So the 95% Confidence Interval for $\frac{SelectionBias}{ObservedDifference}$ is (0.89, 1.45).

C.2 Filing gap and cancellation

Our model predicts firms already on a trajectory to being canceled. How is it able to do this? Prediction of future cancellation could be a difficult task. Some traces in the data could predate and predict cancellation, and the model picks up on those traces, since its label was defined by suo-moto cancellation. One such attribute of firms is their last filing date. Firms which have not filed a return for a long time, are very likely to be canceled at some point.

Many of our list firms, though listed as active (or rather as not-canceled) at the time of list creation in March 2021, have in fact stopped filing returns months before the lists were shared. So it is not surprising that inspections found firms to be non-existent despite them being listed as active.

But what is the probability of being canceled for firms which have not filed for a long time? We

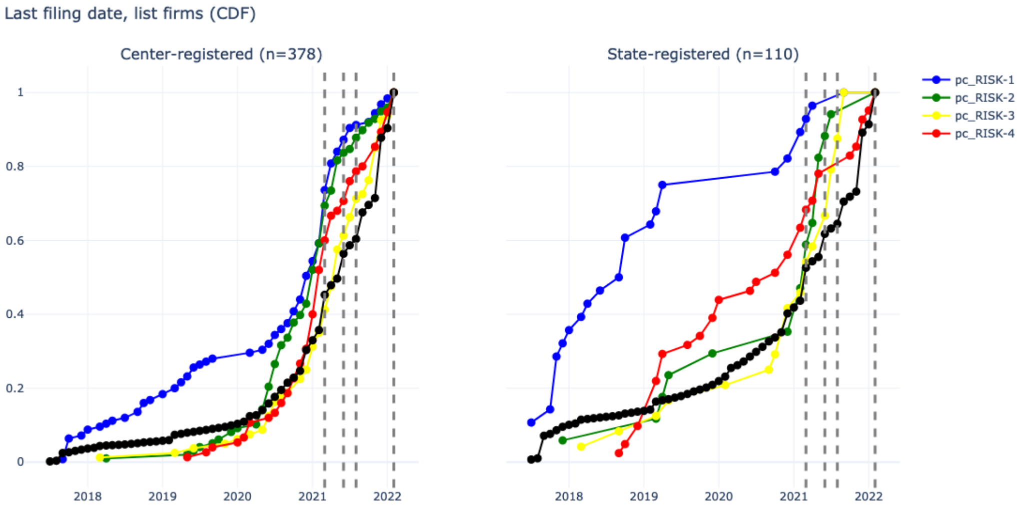


Figure A.8: ***Distribution of last filing date for list firms***. *Notes: In this chart we plot the Cumulative Distribution Function of the last filing date for list firms. The colors denote the different lists. The black line is the baseline, i.e. the last filing date for all firms active as of March 2021, which don't have the final return available. We split into center- and state-registered firms. The dashed lines denote the dates of list sharing.*

quantify this as follows. In each month where we have data, and for each firm in our dataset, we create an observation. In the observation, we list the number of months passed since this firm last filed a return before this month, and whether it was canceled by that month. For example, suppose a firm files returns every month, and is not canceled. In each month it would get an observation of (0 months since last return, not-canceled). Now suppose the firm stops filing returns, and is canceled exactly 3 months later. It would have additional observations:

(1, not-canceled), (2, not-canceled), (3, canceled), (4, canceled), (5, canceled), . . .

In fact, we don't have cancellation data for each month, so we only create observations using snapshots where we have cancellation data. For each snapshot date D with cancellation data, we only include firms which have filed at some point before D, discard all their returns from after date D, and create a single observation for each firm describing the time passed since its last return filed, and its cancellation status as of date D.

We then gather all of these observations, and calculate the probability that a firm is canceled conditional on not filing a return for M months, which we denote as the "filing gap". The results are in Table A.13.

Since we filtered out firms which were canceled as of the time of list creation, the precise probabilities we are interested in are of the form

$P(\textit{firm will be canceled in next X months} \mid \textit{firm currently not canceled, firm has not filed for Y months})$

These conditional probabilities can be calculated from the table above. For example

$P(\textit{firm will be canceled in next 3 months} \mid \textit{firm currently not canceled, firm has not filed for 9 months})$

Combined				
filing gap (Months)	Total Observations	Canceled	% Canceled	Error (% Canceled)
>21.0	530,845	463,721	87.4%	0.0%
(18.0, 21.0]	58,946	49,298	83.6%	0.2%
(15.0, 18.0]	53,284	42,817	80.4%	0.2%
(12.0, 15.0]	59,100	44,059	74.5%	0.2%
(9.0, 12.0]	62,630	43,998	70.3%	0.2%
(6.0, 9.0]	62,216	38,477	61.8%	0.2%
(3.0, 6.0]	65,716	31,218	47.5%	0.2%
(0.0, 3.0]	902,505	20,799	2.3%	0.0%
0	1,933,212	6,369	0.3%	0.0%

Table A.13: Probability of a firm being canceled conditional on the time passed since it last filed a return (the "filing gap"). *Notes: The filing gap is binned into 3-month segments. For each segment we describe the number of observations, number of observations with "canceled" status, the percent canceled out of total observations, and the standard error of percent canceled. We can see that the probability of cancellation climbs quickly between 2.3% at 3 months, to 47.5% at 6 months not filing, and then continues to climb more slowly.*

Since a canceled firm will not file future returns (neglecting very rare cases of firms overturning their cancellation), we know its trajectory in the table.

$$P(\textit{canceled in next 3 months }|\textit{currently not-canceled, hasn't filed for 9 months})$$

$$= \frac{\textit{P(canceled in next 3 months, currently not-canceled|hasn't filed for 9 months)}}{\textit{P(currently not-canceled|hasn't filed for 9 months)}}$$

$$= \frac{(P(\textit{canceled| hasn't filed for 10-12 months}) - P(\textit{canceled| hasn't filed for 9 months}))}{(1 - P(\textit{currently canceled | hasn't filed for 9 months}))}$$

$$= \frac{(70.3\% - 61.8\%)}{(1 - 61.8\%)}$$

$$= 0.22$$

Calculating this for the table above, we get numbers around 0.2 for (6,9] months onwards.

Looking at Figure A.8, since many of our firms have not filed for a long time even before list creation, these conditional probabilities might not suffice to explain the large share of firms from our lists which were canceled. More direct evidence of this not explaining the entire effect is that even restricting ourselves to list firms which have filed recently, many of the firms predicted by our model and inspected were found bogus. See Table A.14.

We failed to consider this filing gap attribute of firms when constructing our model and lists. This is a problematic oversight for two reasons. First because we might have produced better predictions by adding this feature to the model. But mainly because canceling firms which have not filed for a long time and are not expected to file again, even if they were in fact bogus, will not result in revenue recovery, since they are no longer filing returns and not claiming any input credits. We should therefore have excluded from our lists firms which have not filed for a long time before list sharing. We did filter out firms which were already canceled as of each list creation, which would account for many of these non-filing firms, but evidently not all.

	Old			Recent		
	Avg. Score	No. bogus	No. total	Avg. Score	No. bogus	No. total
Risk-1	0.120	89	99	0.091	85	124
Risk-2	0.045	53	55	0.036	89	195
Risk-3	0.060	24	27	0.046	81	178
Risk-4	0.067	4	6	0.029	38	142
Risk-4-ITC	0.029	5	7	0.028	26	105

Table A.14: Inspection results by list and filing gap. *Notes: The rows are the different lists shared. Each list is broken down to two groups by filing gap: "Recent" filing firms (right panel) are firms which have last filed in March 2021 or later. "Old" filing firms (left panel) are firms who have last filed before March 2021. The columns in each panel are the average model score for each group, the number of bogus firms (according to inspection results) and the total number of firms in the group.*

C.3 Department can use model results flexibly

As we go down the ranking, the probability of being bogus for a given firm steadily decreases. We find that 83 of the top 100 firms are bogus, 77.5% of the top 200, 69.3% of the top 400 and 57.1% of the top 800 (see Table A.4a). This is expected behavior and demonstrates that our model is correctly able to sort risky firms. Moreover, the department has the flexibility to decide when the model results stop being useful from a cost-benefit perspective. For example, the department can decide that a hit-rate of 70% is cost effective and go down the list till the hit-rate stays above that.

C.4 Automation recommendations based on our research

Using and updating the ML tool requires in-house capacity that may not be readily available with a tax department. Taking this into consideration, we provide some automation suggestions that do not require any ML related expertise. We recommend that the department maintain an exhaustive list of firms that are found to be non-existent upon physical inspection and subsequently verified as a supplier of bogus bills. Using sales statements to filter out firms that never passed any ITC will further strengthen this list. Once a robust list has been created, the following downstream reports can track fraudulent trading activity networks and catch more likely bogus firms without using ML.

C.4.1 Tracking trading partners of bogus firms

To benefit from identifying a bogus bill supplier, we need timely scrutiny of its trading network. First, we have shown that beneficiaries are spread out across districts and only 19% of firms that have supplies from a non-existent firm have received scrutiny notices (see Table A.12). Therefore, we find that beneficiaries are not flagged to the relevant ward official when a bogus firm is detected. Second, we have also found that suppliers of bogus firms are more likely to be bogus (see Appendix B.4) and focusing on them may be a low hanging fruit. Information technology can fill both these gaps.

Currently, the department separately tracks the due process initiated against firms by the type of action taken, viz. for scrutiny notice, show cause notice, ITC blocked etc. However, a key information

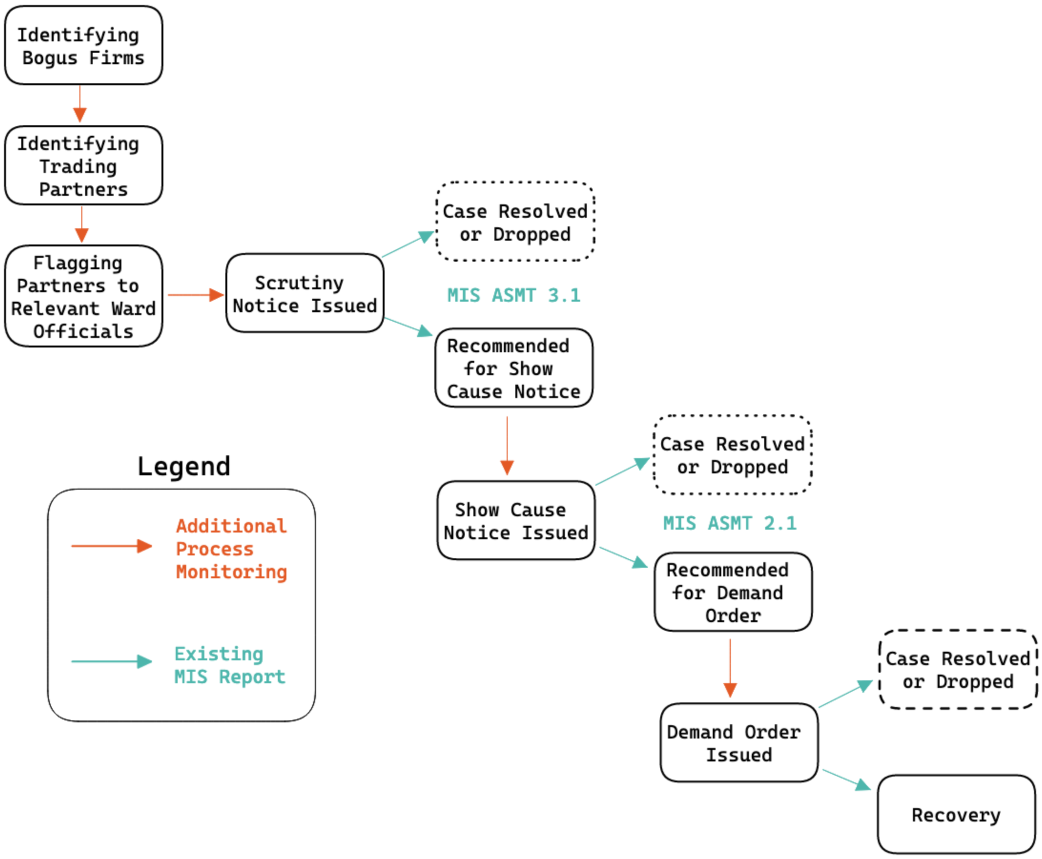
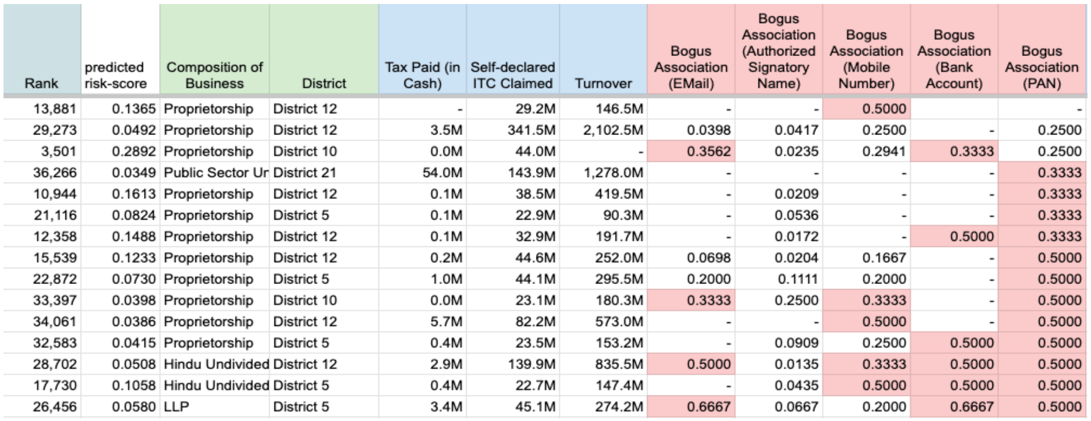


Figure A.9: *Process monitoring setup, proposed.*

gap is comprehensively tracking the list of firms that *should* be scrutinized and their current status. A process monitoring setup that tracks each case from the detection of a non-existent firm, to identification of beneficiaries, their assessment, raising demand and eventual recovery will ensure that beneficiaries do not fall through the cracks.

C.4.2 Tracking potentially bogus firms without ML

Dashboard to track firms sharing identifiers with bogus firms: To reduce supply of fake credits, a ward official needs timely access to information on whether a firm in their jurisdiction is connected to a non-existent firm. When a single entity (say an accountant) is operating many bogus firms, the firms might have common elements in their registration form. For example, they might be registered on the same PAN number, the email address and mobile number associated with the principal place of business could be common, etc. We naively expected dashboards based on shared registration details to be informative. However, we found that simply focusing on firms that share a common registration detail (mobile number, email address etc.) with a non-existent firm does not result in a high rate of detecting bogus firms. We shared a list of firms that had common identifiers with the department for inspection and found the hit-rate to be around 12% (see Table A.8). This could be because many firms share these details for operational reasons.



Rank	predicted risk-score	Composition of Business	District	Tax Paid (in Cash)	Self-declared ITC Claimed	Turnover	Bogus Association (EMail)	Bogus Association (Authorized Signatory Name)	Bogus Association (Mobile Number)	Bogus Association (Bank Account)	Bogus Association (PAN)
13,881	0.1365	Proprietorship	District 12	-	29.2M	146.5M	-	-	0.5000		-
29,273	0.0492	Proprietorship	District 12	3.5M	341.5M	2,102.5M	0.0398	0.0417	0.2500	-	0.2500
3,501	0.2892	Proprietorship	District 10	0.0M	44.0M	-	0.3562	0.0235	0.2941	0.3333	0.2500
36,266	0.0349	Public Sector Ur	District 21	54.0M	143.9M	1,278.0M	-	-	-	-	0.3333
10,944	0.1613	Proprietorship	District 12	0.1M	38.5M	419.5M	-	0.0209	-	-	0.3333
21,116	0.0824	Proprietorship	District 5	0.1M	22.9M	90.3M	-	0.0536	-	-	0.3333
12,358	0.1488	Proprietorship	District 12	0.1M	32.9M	191.7M	-	0.0172	-	0.5000	0.3333
15,539	0.1233	Proprietorship	District 12	0.2M	44.6M	252.0M	0.0698	0.0204	0.1667	-	0.5000
22,872	0.0730	Proprietorship	District 5	1.0M	44.1M	295.5M	0.2000	0.1111	0.2000	-	0.5000
33,397	0.0398	Proprietorship	District 10	0.0M	23.1M	180.3M	0.3333	0.2500	0.3333	-	0.5000
34,061	0.0386	Proprietorship	District 12	5.7M	82.2M	573.0M	-	-	0.5000	-	0.5000
32,583	0.0415	Proprietorship	District 5	0.4M	23.5M	153.2M	-	0.0909	0.2500	0.5000	0.5000
28,702	0.0508	Hindu Undivided	District 12	2.9M	139.9M	835.5M	0.5000	0.0135	0.3333	0.5000	0.5000
17,730	0.1058	Hindu Undivided	District 5	0.4M	22.7M	147.4M	-	0.0435	0.5000	0.5000	0.5000
26,456	0.0580	LLP	District 5	3.4M	45.1M	274.2M	0.6667	0.0667	0.2000	0.6667	0.5000

Figure A.10: ***Mock report using identifier association parameters.***

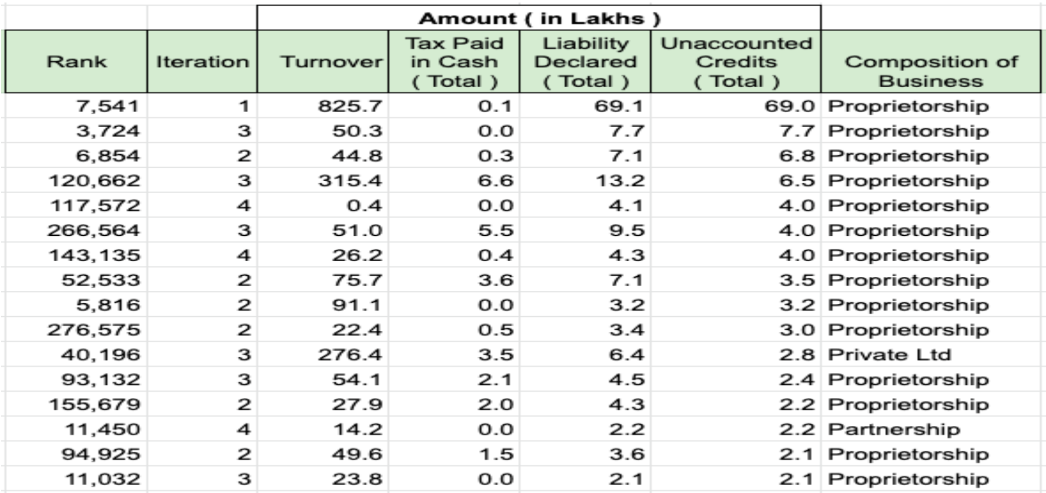
We incorporate these findings to further improve the dashboard features. We suggest measuring the share of firms with that common registration detail which have already been found to be bogus. If a large percentage of firms using a single PAN or mobile number have been found to be non-existent, there is a high probability that the remaining firms with that PAN or mobile number will also have irregularities that are worth examining. So, for every firm with a repeat registration identifier (mobile number, email address, authorized signatory, PAN or bank account number), we calculate a **bogus association feature** defined below.

$$BogusAssociation(Identifier) = \frac{\#ofnon-existentfirmsregisteredusingthegiven(identifier)}{\#oftotalfirmsregisteredusingthegiven(identifier)}$$

The higher this ratio, the riskier a firm is likely to be. For example, if a firm has a mobile number that was used by 4 other firms, and 3 of those firms turned out to be non-existent, the Bogus Association (Mobile) score for that firm would be 3/(4+1) = 0.60. This is a dynamic score, and it will change as the department discovers more non-existent firms in the network and updates the list. By looking at all of these scores at once, as shown below in Figure A.10, we can identify the riskiest firms first. In this example, for rows where the last 5 columns have been highlighted, more than 3 Bogus Association scores are above 33%. We have incorporated these features in the updated ML model which is ready for department's use.

Dashboard to track firms with unaccounted local credits: Based on discussions with department officials and data on cancellation trends and assessment notices, we find that suppliers to bogus firms are likely to be passing fraudulent credits and to have meaningful discrepancies between eligible and claimed tax credits. We recommend that the department creates a dashboard to track unaccounted local credits (see Figure A.11 for a sample). We generated the report by tracking suppliers of bogus firms and calculating their unaccounted local credits. We only focus on firms that have zero supplies from local firms but are claiming local credits.

To build this dashboard we do the following. In step 1, we filter transaction records where the counterparty is a known bogus firm. In step 2, we save this list of suppliers and track their suppliers by repeating step 1. Suppose we now discover that no in-state supplier has reported sales to these firms. We can conclude that all the local credits claimed by these firms are without an actual supply. In step 3,

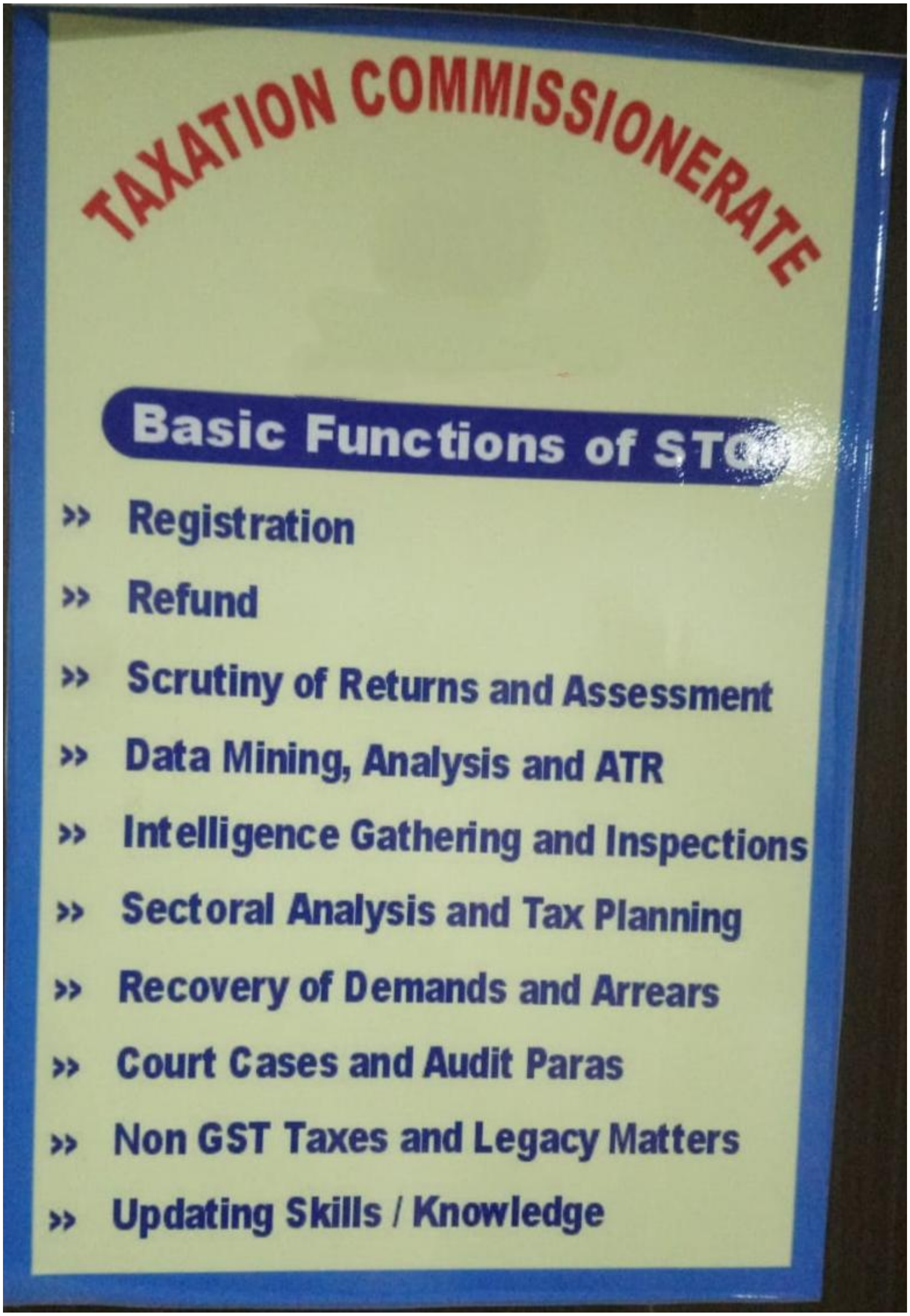


		Amount (in Lakhs)				
Rank	Iteration	Turnover	Tax Paid in Cash (Total)	Liability Declared (Total)	Unaccounted Credits (Total)	Composition of Business
7,541	1	825.7	0.1	69.1	69.0	Proprietorship
3,724	3	50.3	0.0	7.7	7.7	Proprietorship
6,854	2	44.8	0.3	7.1	6.8	Proprietorship
120,662	3	315.4	6.6	13.2	6.5	Proprietorship
117,572	4	0.4	0.0	4.1	4.0	Proprietorship
266,564	3	51.0	5.5	9.5	4.0	Proprietorship
143,135	4	26.2	0.4	4.3	4.0	Proprietorship
52,533	2	75.7	3.6	7.1	3.5	Proprietorship
5,816	2	91.1	0.0	3.2	3.2	Proprietorship
276,575	2	22.4	0.5	3.4	3.0	Proprietorship
40,196	3	276.4	3.5	6.4	2.8	Private Ltd
93,132	3	54.1	2.1	4.5	2.4	Proprietorship
155,679	2	27.9	2.0	4.3	2.2	Proprietorship
11,450	4	14.2	0.0	2.2	2.2	Partnership
94,925	2	49.6	1.5	3.6	2.1	Proprietorship
11,032	3	23.8	0.0	2.1	2.1	Proprietorship

Figure A.11: ***Mock report tracking unaccounted credits for suppliers to known bogus firms.***

for all these firms, we use the consolidated returns to compare the tax liabilities declared and tax paid in cash for these periods. The difference between these values reflects the unaccounted credits i.e., credits that don't have a verified supplier. We only compare the values under the CGST and SGST headers, as both the seller and counterparty belong to in-state firms where the sales statement is available.

Besides Tax Paid in Cash, Liability Declared and Unaccounted Credits, for each firm with zero local supplies, we also show their Turnover, the Iteration Number i.e., the number of transactions separating the firm from the known bogus firm we started with, the model risk-rank of the firm and the Composition of Business.



TAXATION COMMISSIONERATE

Basic Functions of STO

» Registration

» Refund

» Scrutiny of Returns and Assessment

» Data Mining, Analysis and ATR

» Intelligence Gathering and Inspections

» Sectoral Analysis and Tax Planning

» Recovery of Demands and Arrears

» Court Cases and Audit Paras

» Non GST Taxes and Legacy Matters

» Updating Skills / Knowledge

Figure A.12: ***A poster detailing the duties of a State Tax Officer.*** *Notes: This was photographed during one of the visits to the Taxation Commissionerate Headquarters*

D List of top features from each module

D.1 Features based on association with known bogus firms

- Share of non-existent firms, out of all firms with repeat mobile number
- Share of non-existent firms, out of all firms with repeat email address
- Share of non-existent firms, out of all firms with repeat authorized signatory
- Total CGST/SGST liable on transactions with known bogus trading partners
- Total transaction value with trading partners that are known bogus
- % of total purchase value with trading partners that are known bogus
- % of distinct trading partners in a filing period that are known bogus
- % of goods transport bills that are rejected where purchaser was bogus
- % of goods transport bills that are canceled where purchaser was bogus
- Num. of distinct trading partners in a filing period that are known bogus

D.2 Mismatch between sales statement and consolidated return

- Absolute and % difference between total liability declared (For sales)
- Absolute and % difference between total credit claimed (For purchases)
- Absolute and % difference between local credit claimed (Imputed Purchases)
- Absolute and % difference between total assessed value of inward supplies (from goods transport bills) and total taxable value of purchases (Imputed Purchases)

D.3 Network features

- PageRank value, a measure of the importance of firm in its trading network
- In-degree of the firm in its trading network
- Out-degree of the firm in its trading network
- Measure of centrality of the K-core decomposition of the graph of a firm's network
- The number of weakly connected components in the graph
- The number of 3-firm circular transactions that the firm is a party to

D.4 Registration

- Freq. of occurrence of street address listed by firm
- Freq. of occurrence of email domain listed under firm contact details
- Num. of Days between Registration-From date and Approval date
- Freq. of occurrence of most popular HSN listed by firm
- Registration-From Month
- Num. of Days between Registration-From date and Declaration date
- Earliest year of registration for a non-GST (eg. State VAT) registration
- Num. of goods declared by firm at time of registration
- Num. of entries listed by firm under Nature of Business

- If firm is composition dealer

D.5 Consolidated return

- Difference between the first days of a firm's first and last filing periods
- Ratio of difference between the first days of a firm's first and last filing periods, and total consolidated returns filed
- Difference in days between consolidated return filing date and first day of filing period
- Percentile rank of firm by total tax liability declared for the return period
- Local (intra-state) tax liability declared by firm
- Percentage value add by firm, calculated as $\frac{TotalLiabilityDeclared\breve{}TotalCreditClaimed}{TotalLiabilityDeclared}$
- Total input tax credit available claimed by firm
- Local (intra-state) input tax credit available claimed by firm
- Percentage of local tax liability paid by firm
- Total tax liability set off using input tax credit

D.6 Sales statement

- Total invoice value for taxable transactions
- Percentile rank of firm by turnover in current return period
- Sale amount to largest partner by amount of transaction
- Total central tax liability on B2B transactions
- Num. of transactions with items taxable @ 18%
- Turnover at time of GST migration (2017Q1)
- Num. of unique B2B trading partners
- Total tax liability for goods with HSN declared
- Percentile rank by total invoice value for local B2B transactions
- Total tax liability on transactions with largest partner by num. of transactions

D.7 Goods transport bills

- Total distance traveled for inward supplies
- Percentage of inward supplies against which goods transport bills issued
- Maximum distance traveled for inward supplies
- Percentile rank by median distance traveled for inward supplies
- Total tax liability for local outward supplies
- Percentage of outward supplies against which goods transport bills issued
- Total tax liability for local inward supplies
- Total assessed value of goods for outward supplies
- Business duration for inward supplies
- Num. of e-Way bills for outward supplies originating in-state