**Audit lead selection and yield prediction from historical tax data**

Trevor Chan1,2, Cheng-En Tan1,2, Jae Choi3, and Ilias Tagkopoulos1,2\*

1Department of Computer Science, University of California, Davis 95616, USA

2Genome Center, University of California, Davis 95616, USA

3California Department of Tax and Fee Administration (CDTFA)

\*Corresponding author (itagkopoulos@ucdavis.edu)

**ABSTRACT**

Tax audits are a crucial process adopted in all tax departments to ensure tax compliance and fairness. Traditionally, tax audit leads have been selected based on empirical rules and randomization methods, which are not adaptive, may miss major cases and can introduce bias. Here, we present an audit lead tool based on artificial neural networks that have been trained on an integrated dataset of 93,413 unique tax records from 8,647 restaurant businesses over 10 years in the Northern California, provided by the California Department of Tax and Fee Administration (CDTFA). The tool achieved a 40.1% precision (F1-score of 0.42) on classifying positive audit leads, and the corresponding regressor provided estimated audit gains (MAE of $145,865). Finally, we evaluated the statistical significance of various empirical rules for use in lead selection, with two out of five being supported by the data. This work demonstrates how data can be leveraged for creating evidence-based models of audit selection and validating empirical hypotheses, resulting in higher audit yields and more fair audit selection processes.

**1. INTRODUCTION**

Tax audits are an important part of tax collection as the main process to correct errors, increase compliance, and ensure a fair taxation system [**1**]. The process usually starts with candidate business or taxpayer selection through different criteria, then an experienced auditor evaluates the selection criteria to decide which returns to audit. A critical part of the process is the criteria and process by which the tax agencies select the audit candidates so that the utility of the audit process is maximized in terms of fairness, time spent, and outcome. Challenges in audit selection include the lack of universally accepted criteria of an optimal audit selection process, the complexity of the tax system and latent variables that make it hard to generalize, and the incompatibility of data that are usually not ready for direct application of machine learning methods. In addition, the audit selection process needs to protect privacy, reduce bureaucracy, ensure fairness, and avoid any bias, while maximizing resource utilization and benefits for the taxpayer and state [**1**].

For these reasons, agencies such as the IRS use computational tools to assist in their audit selection process. Most of those tools use different scoring systems [**2**] to identify discrepancies or rank reports based on the expected numbers. For instance, discriminant function (DIF) systems statistically determine the accuracy of a tax return over several hundred variables [**3**], while systems that use unreported income discriminant function (UIDIF) score a return based on unreported income potential, which is decided by examining expense and income ratios [**4**]. Similarly, systems use information returns processing (IRP) modules to cross-reference third-party data e.g. taxpayer income, employment, etc. with reported auditee information [**5**]. An auditor would usually take these reports into account, together with contextual and complementary sources of information, such as related activities, investment transactions, and social security information to form a recommendation on whether an audit would be warranted. This process creates a structured approach for audit selection, with some predefined criteria and quantifiable evidence. At the same time, auditor intuition is still the determining factor, given that current computational methods are limited in their scope, use simple computational methods that underfit in terms of model complexity, are usually based on empirical rules that have not been statistically validated for their accuracy, and do not learn from historical data and audit outcomes [**6**].

Several data mining and machine learning methods have been proposed to address these limitations, and a multitude of case studies have been conducted regarding audit selection. Greene et al. (1992) showed that genetic algorithms were able to effectively predict audit yields and at the time they were better performing than algorithms based on statistical pattern recognition and symbolic concept acquisition [**7**]. Fischthal (1988) registered a patent in the United States for a neural network-based fraud detection system, which uses a clustering technique to generate a collection of classes from historical data [**8**]. Yu et al. (2003) introduced a data mining application to detect fraud in Chinese commercial enterprises [**9**]. In that work, the established workflow considered domain expertise, data mining algorithm, and design architecture, with a case study that resulted in an 85-90% accuracy rate for 500 businesses and 100 variables over one year. Micci-Barreca and Ramachandran (2004) conducted an audit selection case study in Texas using an SPSS-based data mining suite [**10**]. Compared against two other Texas-based audit selection strategies in terms of sales tax adjustments, the data mining tool returned higher tax adjustments in selected audits than the other two strategies. Gupta and Nagadevara (2007) performed a case study of the tradeoff between precision and sensitivity of 8 developed classification models on audit selection in the value-added tax (VAT) system of India [**11**]. Attempting to maximize precision, they found that all models developed through data mining techniques were better than random selection. Wu et al. (2012) introduced a framework that utilizes association rules to minimize losses from VAT evasion in Taiwan [**12**]. The results show that the proposed data mining technique enhances the detection of tax evasion, and therefore can be employed to effectively reduce or minimize losses from VAT evasion. Hsu et al. (2014) conducted a pilot study of the efficiency of classification models on real-world data from the Minnesota Department of Revenue [**13**]. The study experimented with a combination of decision tree, Naïve Bayes, multilayer perceptron, and support vector machine classification techniques on 10,943 samples over the years 2004-2006 and achieved an increase in 63% efficiency. Results were validated by the Minnesota Department of Revenue. Silva et al. (2016) conducted a case study of financial fraud using Bayesian networks on taxpayer data from Sao Paulo, Brazil [**14**]. The study achieved a performance rate of 41% on a dataset of 25,322 returns over 35 numeric variables. Rahimikia et al. (2017) investigated the effects of a hybrid model - support vector machine, multilayer perceptron, logistic regression with harmony search optimization - to detect corporate tax evasion in the food and textile financial sectors of Iran [**15**]. The models were tested across two different datasets with 2000-3000 business samples each over 21 financial variables and found that multilayer perceptron combined with harmony search optimization outperformed all other methods. Hajek and Henriques (2017) employed feature selection and classification (Bayesian belief networks, ensemble methods) for detection of financial fraud and found that interpretable Naïve Bayes outperformed all methods in terms of misclassification costs for a dataset of 622 businesses [**16**]. de Roux et al. (2018) proposed a cluster-based unsupervised learning approach for tax fraud detection on Urban Delineation tax in Colombia [**17**]. They evaluated their model on 1,367 tax returns with 3 variables and performed both quality review and an expert review of their model results. They found that their model accurately predicts non-underreporting returns. Most of the prior work, while pioneering, they are limited in that they lack a comprehensive, cohesive dataset with audit results validation. In addition, empirical hypotheses are not usually assessed, and all samples (instances) are treated equally, which limits the emergence of stronger patterns that take place when a higher resolution binning takes place. Another ubiquitous limitation is the lack of feature selection and explainability of the corresponding models, which makes their adoption in a practical setting both difficult and dangerous, given potential artifacts and unwanted biases that may take place due to overfitting and limited data representation.

Here, we address these challenges by constructing a cohesive compendium and then applying machine learning techniques to develop an audit selection predictor for sales tax. In collaboration with the California Department of Tax and Fee Administration (CDTFA), we integrated 93,413 unique tax records from 8,647 businesses over 10 years into a dataset with 20 return-related variables and 5 registration-related variables. After categorizing data based on business type, location, and other features, we evaluated the statistical significance of several hypotheses. We then used feature selection together with artificial neural networks to create an audit selection predictor that achieves substantially better performance than the baseline (**Figure 1**).

**2. METHODS**

2.1 Data Processing

The initial dataset had 108,162 unique tax records from 10,067 businesses over 57 years. Data were organized into a matrix where each row corresponded to an anonymized business and each column corresponded to one of the 42 features (see **Supplementary Material, Table S1**). The data consisted of three data types – categorical, numeric, and datetime. Categorical data were converted to numerical data via one-hot encoding. This was done to capture information on which businesses submitted late returns. Certain features were not used in the analysis, listed in **Table S2** and more information on their removal is given in **Section 1.2** of Supplementary Materials. Regarding encoding, categorical data was encoded with a one-hot encoding scheme. Because the categorical features DasZipID and DasCityID were both anonymized, we were unable to extract additional information, such as latitude or longitude or census data. We note that city ID and zip code ID features were also not used during analysis, as they increase the dimensionality of the dataset and contributed to overfitting (**Figure S19**). Post one-hot encoding the dataset contained 37 features. While this entire dataset was used for classification, only positive audit records were used for regression. For the audit lead prediction, post-filtration (removal of features with 0 standard deviation) the dataset contained 30 features, while for audit yield prediction post-filtration yielded 26 features.

Certain refinement criteria were chosen for the categorical features in our preliminary analysis. We extracted only single-location businesses that had recorded filed returns. Only businesses classified as individual (IND), corporation (CORP), partnership (PART), and limited liability corporation (LLC) were extracted, which make up 75% of the dataset (Figures S4-S7). After these refinement criteria were applied, the dataset was reduced to a total of 93,413 unique tax records from 8,647 businesses over 10 with 20 return-related variables and 5 registration-related variables. Of these businesses, only entries with nonnegative audit information were considered (510 businesses, with 6,301 tax records). We applied the analysis on this dataset, and the results are presented in **Supplementary Materials, Section 2.3.2, Figures S20 and S21**. To increase the performance and specificity of our predictive methods, we further filtered out businesses that are not within the 95% confidence interval for the various variables (see **Supplementary Material, Figures S8 and S9**). The final dataset that is used in our classification analysis presented here has 2,614 tax records from 210 unique businesses, all of them audited over the past 10 years, with 30 features total (**Table S11**). The final dataset used in our regression analysis contains 998 tax records from 76 unique businesses, audited over the past 10 years, with 26 features total (**Table S11**). Finally, quantile, min-max, z-score normalization methods were applied and evaluated during the analysis, with min-max normalization performing the best and selected for the analysis presented here (**Table S12, Figure S22**).

2.2. Data Visualization

Principal component analysis (PCA), as well as t-Distributed Stochastic Neighbor Embedding (t-SNE) were both employed to visualize data. As several features were sparse, these features (**Table S3**) were removed for better visualization. Additionally, data visualization with all tax record data was poor (**Figure S13**), so data was first merged by business – numerical features were summed across the returns per audit period and normalized by the number of returns per audit period. Data were then hierarchically clustered for both PCA and t-SNE (**Figure 2** and **Figure S16**), with the elbow and gap statistic methods [**18**] used to identify the optimal number of clusters (**Figure S15**).

2.3 Binning and Hypothesis Testing

The validity of empirical rules from auditors and other hypotheses was tested as follows. First, each business was placed on one of the 250 bins that were created based on the 10 unique North American Industry Classification System (NAICS) codes and the 25 unique City IDs in the dataset (**Supplementary Materials, Figure S1**). Then, various hypotheses related to gross sales, taxable sales to gross sales ratio, business classification, and late penalties were tested for different thresholds in the distribution (**Tables 1** and **2**). Statistical validation was performed by calculating FDR-adjusted p-values [**19**], and the non-parametric Wilcoxon rank-sum test.

2.4 Audit Lead Predictor

To identify positive audit leads, we trained a feed-forward Artificial Neural Network (ANN) on the 2,613 tax records of our dataset. Random search was used to select its hyperparameters, and a stratified ten-fold cross-validation was employed for training. As multiple returns in the dataset are mapped to a given business, returns were grouped before being split into training and testing sets to avoid overlap of returns in both the training and testing sets/folds. In terms of labels, audits with a yield above the specified dollar threshold were classified as positive, while those with less than the threshold were classified as negative, as per CDTFA instructions on the cost of conducting an audit. The final architecture had 30 input nodes, 2 hidden layers with 19 and 8 nodes respectively, and 1 output node. The sigmoid function was used as an activation function for the nodes in each layer. The data was split 90%-10% between training/testing and validation sets. There was a high-class imbalance, with only 23% of labels being positive, hence over-sampling of minority and under-sampling of the majority class was performed [**20**]. Each fold was trained over 1000 epochs with a batch size 32, the Adam optimizer and binary cross-entropy as the loss function [**21**]. A dropout of 0.3 and L2 regularization was used to reduce overfitting. For early stopping, a delta of 0.0001 and patience, the number of epochs with no improvement after which training is stopped, of 10 was used [**22**].

2.5 Audit Yield Regressor

To identify the expected return of an audit, we created an audit yield regressor for positive audits. Due to the fragmentation of the dataset, using the same predictor approach as a classification for this task led to low regression performance (**Figure S23**). For that reason, we grouped audit cases based on the validated hypothesis rules and categorical variables (**Supplementary Materials, Section 2.4.2, Figures S24-S26**). Extreme gradient boosted trees were trained on each case subset, and hyperparameters were selected by random search. We used ten-fold cross-validation for training, with the same 90%-10% split between training/testing and validation sets. We also had returns grouped by the business before being split into training and testing sets to avoid overlap of returns in both the training and testing sets/folds. As the audit yield is the same for multiple returns, each regressor was evaluated based on the prediction mean for a given business. The best performing model used a learning rate of 1, using 50 trees. Each tree had a max depth of 2 with 20% of the samples and 30% of features in each training set used per tree. An L1 regularization value of 4 and a linear loss function was used. Recursive Feature Elimination (RFE) was used for feature selection [**23**].

**3. RESULTS**

**Dimensionality reduction reveals higher audit yields clusters and underlying rules.**

The dataset includes a variety of different types of information pertaining to location, business type, audit, and return history (**Figure 2A**). Dimensionality reduction and subsequent visualization reveal clusters that have statistically significant over-representation of positive (or negative) audits results. Further dissecting those clusters uncovers the rules that are common to their members. For example, cluster 1 was the cluster with the highest audit success rate (27.8% vs. 19.5% which is the average). **Table S5** provides a list of those clusters and underlying rules, while **Figures S16–S18** depict such clusters with different normalization schemas. Considering all returns and features, and by ordering the numerical features for each return, patterns became evident. For example, pattern 24 in **Tables S6** and **S7** contains 1,290 returns, with 141 of them been audited. In these 141 returns, 116 (82.2%) of them contained a positive audit yield and 97 (68.8%) of them contained audit yield greater than the threshold, a significantly higher success rate for selecting a non-zero audit yield (82.2% vs. 33% on average, p-value = 3.8∙10-35) or a positive audit, i.e. one with a yield higher than the threshold (68.8% vs. 23.6% on average, p-value = 1.78∙10-31). In addition, the audit ratio (141/1290 = 11%) is also significantly higher than the overall audit ratio (6,301/93,413 = 6.75%), which means that the CDTFA has empirically also focused on the returns with this specific pattern.

**Empirical based hypotheses validated by data**

Next, we investigated the rules by which tax departments employ an in-house method to detect anomalies on tax reports and select businesses to audit. **Table 1** depicts five such rules that are empirically chosen and are used in practice. We evaluated these hypotheses on varying thresholds depending on the mean of the respective bin (see **Methods**). **Figure 3** presents a visual representation with box plots in the cases where the threshold is equal to the mean of the distribution for the respective bin. Our results show that two out of the five hypotheses pan out for a given value of the threshold. The higher propensity of positive audits was recorded when the reported gross sales are lower than the mean and when there have been late return penalties. Interestingly, the ratio of taxable vs. gross sales is not a good indicator of audit success, and only those businesses with no variance in their reported gross sales had a statistically significant over-representation of positive audits. We also tested the inverse hypotheses, which were all rejected as expected (**Table 2**).

**65% of the audit leads produced by the classifier are positive.**

The audit lead classifier was able to achieve average training and testing accuracies of 65.3% and 63.6% across all ten folds, while on the validation set it achieved an accuracy of 61.2%, and F1 score of 0.42. **Figure 4** shows the architecture, ROC curve, precision-recall curve, and confusion matrix for the audit lead classifier. Additionally, running RFE on the audit lead predictor revealed some unexpected features being selected, such as the tax amount due, the type of the business (mobile food or individually owned), the amount of sales tax included in gross sales, and whether the business owner filed their taxes quarterly or annually. **Figure 4D** shows how the removal of each feature impacts the performance of the classifier, as a function of the F1 score.

**30% of all variation in audit yield can be measured by an audit yield regressor.**

Next, we investigated whether we could predict the actual amount of a positive audit, given its characteristics. After experimenting with several binning scenarios and regression techniques, we concluded that binning based on the validated hypotheses together with a boosting tree method produces optimal results (see **Methods**). The selected audit yield regressor achieved an average training and testing R-squared of 0.31 and 0.28 across all ten folds and achieved an R-squared of 0.30 on the validation set. **Figure 5** shows the architecture and pipeline, predicted v. actual scatterplot, and a residual scatterplot for the selected NAICS code-based audit lead regressor. Similarly to the classifier, running RFE on the audit lead predictor revealed that the important features are the type of the business, catering option, pre-paying and filing characteristics, the amount due, and its use tax purchases (**Figure 5D**).

**DISCUSSION**

This work is building upon decades of empirical tax audit selection and attempts to use computational methods to automate pattern analysis and outlier detection. There are several lessons that we learned on this journey. First, there is a need for standardization, integration, and preprocessing of tax records so that advanced machine learning techniques can be applied. Although this need is less dire than other domains, such as life sciences or business development, due to the digitization and nature of the tax records, we have encountered a number of issues, ranging from unmappable audit yield to audit period timestamps and lack of a large enough pool of accounts that were audited to data anonymization and lack of access to dark pools of data that contain important features for building a positive audit predictor.

Another observation is that the current process is distributed, subjective, and less driven by data analysis. Although there are tools that analyze data for identifying outliers and potential leads, these tools are usually limited in both scope and capabilities, mostly relying on decision trees and simple regression models. Usually, empirical rules are subjective and have never been subjected to statistical tests to assess their validity. Our analysis shows that, as expected, a small subset of features is sufficient to achieve accurate audit lead prediction, although what these features are and the how the methodology needs to be refined to broaden its scope and accuracy, is work in progress. Given the revolution in machine learning that we have witnessed the past decade and the plethora of data that tax departments have in their disposal, it creates a perfect storm for a disrupting technology such as deep learning to operate on the variety of contextual data that are currently available to achieve precision taxation and compliance at scale.

Future work includes the amelioration of both data and computational resources. In terms of data, we will expand the scope of this work to other sectors beyond restaurants, and other districts beyond Northern California. That would lead to generalizable methods that may consist of a core module that can be sector and location agnostic, and specialized learners that fine-tune predictions based on those characteristics. In terms of computational methods, we will work on adopting time series techniques, such as recurrent neural networks and Long Short-Term Memory (LSTM) architectures, as well as the integration of multiple other sources of information, coming from employment, property, and income tax records, as well as receipts or similar non-tax information. As filtration removed many tax records above the 95% confidence interval, it would be interesting performing a similar analysis on these data points. Considering features from multiple data streams would allow our models to make more informed predictions. Additionally, techniques that increase the explainability and trustworthiness of the results would be helpful, as well as a more in-depth look at the candidate business patterns to understand why they have been selected should be done.

Finally, a key topic that was not addressed here, but it is of high importance is algorithmic bias. From dataset population bias to aggregation bias, it is important to take steps to assess and improve fairness [**24**] to ensure our predictive models achieve non-discriminatory results. Steps such as better preprocessing and adding fairness constraints, as well as introducing fairness penalties to our predictive models would be of great interest [**25-29**]. The work here and extensions we propose above will pave the way and contribute towards a better, more precision and fair tax system that will adapt to new developments and events while adopting an evidence-based approach to public service.

**ACKNOWLEDGEMENTS**

We would like to thank Aimee Olhiser, Nick Maduros, Trista Gonzalez, Joshua Canham, Jason Mallet, Chris Drews, and other collaborators at the California Department of Fee and Tax Administration for their comments and advise on this project.

**REFERENCES**

1. Alm, J. “Tax Compliance and Administration,” in Handbook on Taxation; eds. Hildreth, W.

B., Richardson, J. A., pp. 741-768. Marcel Dekker, Inc. 1993.

2. The Tax Audit Process: Procedure, Rules & Guidelines for IRS to Audit.

www.taxdebthelp.com/tax-problems/tax-audit/audit-process. Accessed 16 Apr. 2020.

3. Wedick, J. L. Looking for a Needle in a Haystack – How the IRS Selects returns for

Audit. The Tax Advisor, New York, 14, (11), 673-675. 1983.

4. Cyr, D., Eckhardt, T., Sandoval, L. A., and Halldorson M. Predictors of Unreported Income:

Test of Unrepeorted Income (UI) DIF Scores. Internal Revenue Service Research Conference. June 11-12, 2002.

5. Information Returns Processing | Internal Revenue Service. Internal Revenue Service (IRS),

www.irs.gov/privacy-disclosure/information-returns-processing. Accessed 16 Apr. 2020.

6. González, P. C. and Velásquez, J.D. Characterization and detection of taxpayers with false

invoices using data mining techniques. Expert Systems with Applications 40, 5. Apr

2013, 1427–1436.

7. Greene et al. Learning audit selection rules from data: A genetic algorithms approach.

Studies in Management Science and Systems, North-Holland Press, Amsterdam (1991).

8. Fischthal, S., 1998, Neural network/conceptual clustering fraud detection architecture. US

Patent 5822741. Oct. 13.

9. Fan, Y., Zheng, Q., and Xiao-Ling, J. "Data mining application issues in fraudulent tax

declaration detection," *Proceedings of the 2003 International Conference on Machine*

*Learning and Cybernetics (IEEE Cat. No.03EX693)*, Xi'an, 2003, pp. 2202-2206 Vol.4.

10. Micci-Barrec, D., Ramachandran, S. (2006) Analytics Elite. Predictive Tax Compliance

Management. http://www.spss.com Accessed 16 Apr. 2020.

11. Gupta, M., Nagadevara, V., Audit Selection Strategy for Improving Tax Compliance:

Application of Data Mining Techniques. Foundations of Risk-Based Audits. Proceedings

of the eleventh International Conference on e-Governance, Hyderabad, India, December. pp. 28–30. 2007.

12. Wu, R.-S., Ou, C.-S., Lin, H.-y., Chang, S.-I., Yen, D.C., Using data mining technique to

enhance tax evasion detection performance. Expert. Syst. Appl. 39, 8769–8777. 2012.

13. Hsu, K.-W., Pathak, N., Srivastava, J., Tschida, G., Bjorklund, E., Data Mining Based Tax

Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of

Revenue. Real World Data Mining Applications. Springer. pp. 221–245. 2015.

14. da Silva, L. S., Rigitano, H. de C., Carvalho, R. N., and Souza J. C. F. Bayesian networks

on income tax audit selection - a case study of brazilian tax administration. In Bayesian Modeling Application Workshop (BMAW), 2016. 37

15. Rahimikia, E., Mohammadi, S., Rahmani, T., and Ghazanfari, M. ‘‘Detecting corporate tax

evasion using a hybrid intelligent system: A case study of Iran,’’ Int. J. Account. Inf. Syst., vol. 25, pp. 1–17, May 2017.

16. Hajek P, Henriques R (2017) Mining corporate annual reports for intelligent detection of

financial statement fraud: a comparative study of machine learning methods. Knowl Based Syst 128:139–152. doi:10.1016/j.knosys.2017.05.001

17. de Roux, D. et. al. “Tax Fraud Detection for Under-Reporting Declarations Using an

Unsupervised Machine Learning Approach” Proceedings of the 24th ACM SIGKDD

International Conference on Knowledge Discovery & Data Mining. (2018): 215–222

doi:10.1145/3219819.3219878

18. Tibshirani, R., Walther, G., Hastie, T. Estimating the number of clusters in a dataset via the

gap statistic. J. Roy. Stat. Soc. B 63(2), 411–423 (2001)

19. Chen, S. et al. “A general introduction to adjustment for multiple

comparisons.” *Journal of thoracic disease* vol. 9,6 (2017): 1725-1729. doi:10.21037/jtd.2017.05.34

20. Chawla, N. et al. “SMOTE: Synthetic Minority Over-sampling Technique.” J. Artif.

Intell. Res. 16 (2002): 321-357.

21. Buja, A., Stuetzle, W. and Shen, Y. Loss functions for binary class probability estimation and

classification: structure and applications. manuscript, available at

www-stat.wharton.upenn.edu/~buja/

22. Lutz Prechelt. Early stopping-but when? Neural Networks: Tricks of the trade, pp. 553–553,

1998.

23. Breiman L, “Random forests,” Mach. Learn, vol. 45, no. 1, pp. 5–32, Oct. 2001.

24. Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un) fairness: Lessons for

Machine Learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, 49–58.

25. Brian d’Alessandro, Cathy O’Neil, and Tom LaGatta. 2017. Conscientious classification: A

data scientist’s guide to discrimination-aware classification. Big data 5, 2 (2017), 120–134.

26. Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification

without discrimination. Knowledge and Information Systems 33, 1 (01 Oct 2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

27. Yongkai Wu, Lu Zhang, and Xintao Wu. 2018. Fairness-aware Classification: Criterion,

Convexity, and Bounds. (2018). arXiv:cs.LG/1809.04737

28. Lingxiao Huang and Nisheeth Vishnoi. 2019. Stable and Fair Classification. In International

Conference on Machine Learning. 2879–2890.

29. Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie

Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. (2017). arXiv:cs.LG/1706.02409.

**TABLES**

**Table 1. Tax audit empirical hypotheses validation**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **q-value (Threshold = 0.1)** | **Selected Threshold for each bin (combination of CityID and NAICS)** | | | | | | |
| **ORIGINAL HYPOTHESES** | **0.25\*Mean** | **0.5\*Mean** | **0.75\*Mean** | **Mean** | **1.25\*Mean** | **1.5\*Mean** | **1.75\*Mean** |
| Reported gross sales lower than some threshold result in higher yield audits (vs. when reported gross sales is higher than the threshold) | 0.11 | < 10-3 | < 10-3 | 0.10 | 0.15 | 0.16 | 0.20 |
| Individual taxpayers are more likely to have positive audits compared to corporations | 0.17 | | | | | | |
| The ratio of is lower than some threshold result in higher yield audits | ~ 1 | ~ 1 | ~ 1 | ~ 1 | 0.50 | 0.38 | 0.67 |
| Variance of reported gross sales (in a specific period) lower than some threshold result in higher yield audits | 0.70 | ~ 1 | ~ 1 | 0.79 | ~ 1 | ~ 1 | ~ 1 |
| Late return penalties higher than some threshold result in higher yield audits | < 10-6 | < 10-6 | < 10-5 | < 10-5 | < 10-6 | < 10-4 | < 10-3 |

**Table 1**: Wilcoxon-ranksum test applied to the hypotheses. A result less than 0.1 signifies a statistically significant result. Grayed-out cells mark non-significant results. “~ 1” results denote probabilities arbitrarily close to 1.

**Table 2. Tax audit empirical inverse hypotheses validation**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **q-value (Threshold = 0.1)** | **Selected Threshold for each bin (combination of CityID and NAICS)** | | | | | | |
| **INVERSE HYPOTHESES** | **0.25\*Mean** | **0.5\*Mean** | **0.75\*Mean** | **Mean** | **1.25\*Mean** | **1.5\*Mean** | **1.75\*Mean** |
| Reported gross sales higher than some threshold result in higher yield audits (vs. when reported gross sales is lower than the threshold) | ~ 1 | ~ 1 | ~ 1 | ~ 1 | ~ 1 | ~ 1 | ~ 1 |
| Individual taxpayers are less likely to have positive audits compared to corporations | 0.84 | | | | | | |
| The ratio of is higher than  some threshold results in higher yield audits | 0.29 | 0.43 | 0.12 | ~ 1 | ~ 1 | ~ 1 | ~ 1 |
| Variance of reported gross sales (in a specific period) higher than some threshold result in higher yield audits | ~ 1 | ~ 1 | ~ 1 | ~ 1 | ~ 1 | ~ 1 | ~ 1 |
| Late return penalties lower than some threshold result in higher yield audits | ~ 1 | ~ 1 | ~ 1 | ~ 1 | ~ 1 | ~ 1 | ~ 1 |

**Table** **2**: Wilcoxon-ranksum test applied to the inverse hypotheses. A result less than 0.1 signifies a statistically significant result. Grayed-out cells mark non-significant results. “~ 1” results denote probabilities arbitrarily close to 1.

**FIGURE CAPTIONS**

**Figure 1.** Overview of tax prediction pipeline – Raw data received from the CDTFA is first preprocessed via normalization, merging of business profiles and feature removal to create a compendium of tax audit data based on registration and return features. This complete dataset then undergoes one of two steps: One, the data is fed through a predictive modeling pipeline in which a feed-forward neural network first classifies an audit as positive or negative, and then second a series of extreme gradient boosted trees are used to predict audit yield based on these positive audits, and two, data is binned based on city and NAICS codes such that hypotheses generated by the CDTFA team can be tested and validated statistically.

**Figure 2.** (A) tables showing organization of the dataset, with (B) PCA and t-SNE visualization of data with minmax normalization post-filtration. K-means clustering is shown and was conducted on the PCA-data. Outliers are removed, and only datapoints with audit yield greater than $0 are shown for visualization purposes.

**Figure 3.** Statistical validation using Wilcoxon-ranksum test of all five hypotheses. Distribution of data for each threshold bin are plotted with statistical significance denoted above each pair of groups. \* denotes a p-value less than 0.1, \*\* a p-value less than 0.05 and \*\*\* a p-value less than 0.001. n.s. denotes a non-significant result. For all graphs above, the threshold was set to the mean.

**Figure 4.** (A) receiver operating characteristic (ROC) curve, (B) precision-recall curve, and (C) audit lead classifier architecture with (D) recursive feature elimination results. The red star signifies the feature added after which model performance no longer improves. (E) displays the confusion matrix for the trained classification model. The average F1 score for the positive class was 0.45 across all 10 folds, with AUROC and AUPRC being 0.86 and 0.47, respectively.

**Figure 5.** (A) regressor architecture and pipeline, (B) predicted vs. actual and (C) residual plots. Error bars represent the prediction on multiple returns, where each point represents the mean of predictions for a single business. (D) shows recursive feature elimination results for the trained regression model. The red star signifies the feature added after which model performance no longer improves.