Predicting Intentional and Inadvertent Non-compliance

Kathleen M. Carley[1], Dawn C. Roberston, Michael K. Martin, Ju-Sung Lee, Jesse L. St. Charles, Brian R. Hirshman
Carnegie Mellon University: Center for Computational Analysis of Social and Organizational Systems

Abstract

Predictive models were developed for identifying tax-returns having errors that could be characterized as either intentional or inadvertent. Two distinct approaches were used: meta-modeling using the literature on general errors, and statistical machine learning techniques for deriving models from audited tax returns. Comparison of these models reveals that while there are commonalities, each has a strength that suggests a unique class of tax returns as possibly having errors. A combined model that links these into a single ensemble may provide the most comprehensive and reliable characterization. That reliability is partially dependent on the amount of data from which the model is built, partially on whether supervised or unsupervised learning techniques are used, and partially on the reliability of the data used to build the models. IRS audit data is suspect; examiners cannot know with any reliability what a taxpayer's motives were at the time of filing, and they have a much higher standard of evidence required to prove intentional misreporting. Thus, techniques for estimating errors that take these biases into account are needed and the ensemble modeling approach is one such technique. Even though more unsupervised techniques should be explored and a wider range of data assessed, the methods employed in this study can be instructive for the development of predictive models of taxpayer behavior.

Introduction

Tax non-compliance is socially harmful as it can reduce revenues, distort labor markets, and undermine state stability by feeding perceptions of cheating and fraud. Reducing non-compliance can be facilitated if one understands the basis for that non-compliance. Kinsey (1984) defined noncompliance with tax laws as the "failure, intentional or unintentional, of taxpayers to meet their tax obligation." Estimates of errors place the number of returns containing either an intentional or inadvertent error, or both, above 50%. Minimizing the number and size of such errors, requires attending to both types of error. This point was made in 2007, Michael Brostek in his testimony before the Committee on the Budget, U.S. Senate on tax compliance. For example, he noted (p.9) that the GAO had found that simplification had the potential to reduce the tax gap because it would reduce inadvertent errors by eliminating confusion, decrease misuse by making it harder to hide non-compliance, and increase willingness to comply due to increased understanding. In the case of simplification, the same action can reduce both intentional and inadvertent errors. However, when simplification is not possible, different strategies may be necessary to reduce the tax gap due to inadvertent and intentional errors. Educational outreach, for example, is more likely to impact inadvertent errors; whereas, enforcement, withholding, and information requirements may have a greater impact on reducing intentional errors. In order to

[1] Direct all correspondence to Prof. Kathleen M. Carley, 1923 Wean Hall – ISR-SCS, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213. Tel: 1 412-268-6016. Email: kathleen.carley@cs.cmu.edu.

provide a more nuanced approach to reducing the tax gap, that is tuned to the needs of the taxpayers understanding both intentional and inadvertent error is critical.

The majority of research on taxpayer noncompliance has been concerned with intentional errors on tax-returns (i.e., evasion). The term intentional tax error is often used synonymously with 'non-compliance' and 'tax evasion.' Intentional tax errors comprise any form of willful misrepresentation while completing a tax-return, for the purposes of minimizing the tax owed or maximizing a tax refund. Typically, these acts include under-reporting income, over-reporting deductions, and erroneously claiming credits with the intent of non-compliance. In contrast, inadvertent tax errors include mistakes, math-errors, forgetting, and unintentional mis-interpretation or misunderstanding.

Our research, conducted for the Internal Revenue Service, explores both intentional and inadvertent error. We ask, is it possible, given the information on a return, to tell whether an error is intentional or inadvertent? Thus, this work addresses the lack of knowledge concerning unintentional errors on tax-returns, and may provide potential guidance to examiners, while helping the Service better meet the taxpayer needs by identifying factors that lead to inadvertent error.

The goal is to determine when it is possible to predict intentional and inadvertent errors given only the information available on a tax-return. Once the contributing factors to the commission of errors are identified, the IRS can address these factors with the intent of reducing future errors. Also, profiles resulting from these models may be used in a similar fashion. This would enable more customized support to taxpayers. In addition, models gleaned from this study could be used in simulation models of taxpayer behavior enabling the IRS to explore the potential impact of various services and interventions.

Background on Non-Compliance Modeling

Theories of non-compliance generally break down into those that emphasize economic deterrence and those that emphasize fiscal psychology (Milliron & Toy, 1988). Researchers in the economic deterrence paradigm tend to employ expected utility theory and view the taxpayer as a rational actor seeking to maximize personal gain by minimizing taxes paid. However, the evidence is mixed and taxpayers often fail to behave in an objectively rational manner. Researchers in the fiscal psychology paradigm tend to employ prospect theory (e.g., Kahneman & Tversky, 1979) and consider factors such as the cost of compliance and social context (Smith & Kinsey, 1987). Supporting evidence includes the generally high rates of compliance and the fact that compliance increases with the expectation of a refund and as knowledge of the tax law increases. Additionally, from a purely empirical perspective, there exist key correlates of non-compliance, of general intentional non-compliance, and of inadvertent error. For example, income level, youth and unfamiliarity with the tax laws, and gender are all highly correlated with non-compliance. Despite this body of information, no clear single picture of the correlates of non-compliance exists.

This lack of a single clear picture suggests that a multi-modeling perspective is needed. We developed the first principles models using the open-source literature, which includes the results of psychology experiments and social empirical research. These models were developed in order to identify factors, outside of those derivable directly from the tax-returns that might account for errors. Further, it was felt that such models might provide greater insight into why errors occurred. Since the rational for intentional and inadvertent errors such first principles models should help distinguish the two types of error. The statistical machine learning models were

developed in order to identify factors that were directly derivable from tax-returns. Such models were expected to be potentially predictive, but more related to tax law in its current form and with less ability to predict the impact of changes. Since the statistical distribution of intentional and inadvertent errors was likely to be different the statistical models should help distinguish the two types of error.

Modeling Errors

In this study we take a dual teaming approach. We have two teams, working independently, from different sources, to develop models of error. Team A works from the open-source literature and has developed a model of intentional error and another of inadvertent error from theory using only the data and information in the published literature, much of which does not consider a taxpayer applications. These are referred to as the first principles models. Team B works from the Exam Office Automation Database (EOAD) and the Individual Return Transaction Files (IRTF) database provided by the IRS and, utilizing statistical and machine learning approaches, estimates a set of empirical models which are then combined into a unified empirical model. The first principles and the empirical models are then compared and contrasted by Team C, who using a subset of the empirical data applies the models from teams A and B to that data and creates a combined model.

Compliance was modeled first for the tax-return as a whole, and then for specific line items. Two line items have been modeled to date. The first line item examined was the earned income tax credit (EITC) as it is one of the most adjusted line items. The second, is wages, salaries and tips. Other potential line items to be modeled in the future include those found to be critical in the first principles intentional error model: capital gains, self-employed, farm income, student loans and social security income.

Data Used by Teams B & C[2]

The IRS EOAD data contains 2.66 million records containing 2,379,523 exams with corresponding line items and valid incomes, filing statuses and timeliness codes from the period 2002-2007, most of which were in 2006-2007. Of these only the data from 2006 and 2007 was used as that matched with the IRTF. It is important to note that these are operational exams, and the returns included are those that were thought to be non-compliant. As such, this is a biased sample; however, it was the only available data with any non-researcher-proposed indication of error. Having such an indication is a requirement for the specific statistical learning models employed in this exploratory study.

Of these 2,379,523 returns, all of which are in 2006-2007, 65,547 were marked as having intentional errors, 1.22 million tax-returns were marked as having unintentional or inadvertent errors, and the remaining were not marked with either type of error by the examiners. This is a second source of bias, the examiners cannot know the intent for sure, have incentives not to mark a tax-return as having an intentional error, and the taxpayers have incentives to provide support for inadvertent error. Consequently, even among this non-representative sample, there may be fewer tax-returns marked as containing an intentional error than is actually the case. These records include, 1.12 million campus (correspondence) examinations, 216,774 field exams, and the remainder are office, no-show, no-response or undeliverable mail. Although not itself a

[2] The EOAD and IRTF data were held in a secure facility on a stand-alone machine following the CASOS technical control policy guidelines. Only members of the CASOS team at CMU who were cleared by the IRS to handle sensitive data were allowed access.

source of bias, the type of exam is indirect information about the likelihood of error and is information that would not be available with a tax-return not in this operational set.

The EOAD data set contains two tables, E and C. The C table contains tax-return data without specific line item information. Example fields are exam date, adjusted gross income, and preparer. The E table contains information about the line items examined during the audit. Every line item examined is included in this table and some fields included are monetary adjustment by line item, reason for the adjustment and line item identification. The C table was cleaned and duplicate keys and records were removed. All records without valid filing statuses or adjusted gross income fields were dropped, resulting in 2.48M records left. The C and E sets were combined in such a way that the tax-return information was preserved from C along with summary information from the line item set.

Intent for the tax-returns was determined by the intent from the corresponding line items. If a tax-return had at least one line item issue that was considered intentional, the whole tax-return was marked as intentional. If a return had at least one unintentional line item, then it was considered to have inadvertent errors. This procedure resulted in some tax-returns being marked as containing both intentional and inadvertent errors. Note, an alternative would have been to consider all of the returns where the error led to an underpayment of taxes to include intentional errors. In Figure 1, the distribution of level of error by level of adjusted gross income is shown. As can be seen, most of the errors result in under-reporting of income (right hand side); however, both under and over-reporting occur at all income levels. Based on our research on the general factors leading to intentional and inadvertent errors, and discussions with examiners, we found that it should not be assumed that all cases of under-reporting are intentional, nor that all over-reporting is inadvertent. In both cases, there are a number of factors that can lead to inadvertent errors; in particular, the complexity of the return.

In Figure 1, it will be seen that there are returns with an error of 0 dollars. A return that is marked with an error of size 0 is one that after the exam; either it was determined that no adjustments need to be made or the adjustments were such that those in the positive direction cancelled those in the negative leading to 0 total adjustment.

When analyses of individual line items were done, expected burden was used to determine complexity. Information from an IRS provided burden study was used in conjunction with an estimation of the number of lines of direction one needed to read to fill out that line item. This results in an estimation of low to high complexity per line item using a five point scale. For the return as a whole, its complexity was set based on the complexity of the line items used. To minimize error this was turned in to a three point scale: as follows:

- Low complexity – Form 1040, 1040A, or 1040EZ w/o schedules
- Intermediate complexity – Form 1040A with schedules and 1040 with schedules A,B,D, Additional Child Tax Credit, Educational Credits, Child Care Credit, Credit for the Elderly or EIC
- High complexity – Form 1040 with schedules C,E or F or other schedules and all other specific Forms 1040, e.g. 1040PR, etc.

We only have the line items examined to determine which schedules were used. As such, it is likely that we are underestimating complexity.

The IRTF data came in several tables as it is a much larger database. It includes information about all tax-returns from the years 2006 and 2007. In the IRS IRTF data there are 139M records that exist in both years. The records were matched via keys for EITC eligibility and age (which were calculated from the return year.)

The IRTF data has fewer variables per tax-return and the data is less in-depth than the EOAD set. However it does contain returns not examined. We used only those records in the IRTF that could be matched to records in the EOAD. There were a few key pieces of data gleaned from this set for use with the EOAD data when modeling intent. Those included date of birth, additional preparer information and additional line items.

For the purposes of this study, for each variable, the data was binned into predetermined categories and then into the "super bins" which were used in our models. The purpose of binning is three fold: first, it reduces error by decreasing the granularity of the data, second it enables comparability with existing studies in the literature, and third it enables the results to be used directly in the simulation model and by field operatives.

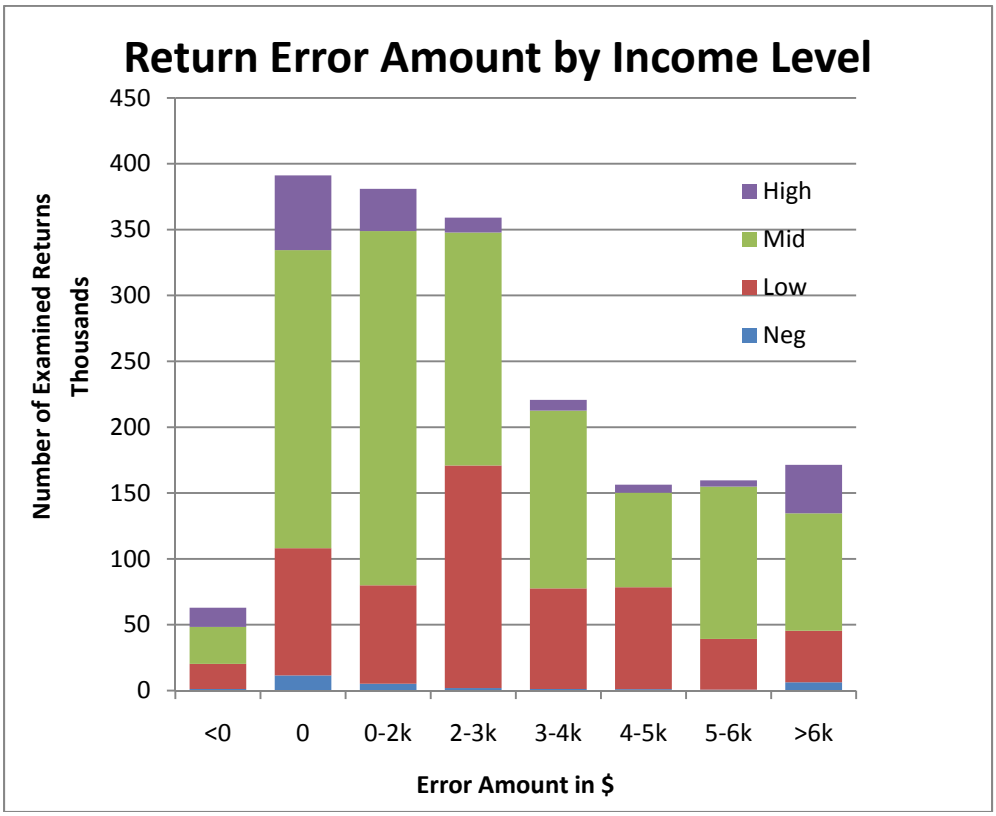

Figure 1. Distribution of Under/Over-Reporting (Loss) by Income Level

The income field used for our analysis was the adjusted gross income reported on the return. See table 1. Additional variables created at this step were itemization, preparer use, exemptions claimed and an initial capital gains variable. If the return indicated itemized rather than standard deductions, the itemized flag was set to one. Preparer use was gleaned from the preparer variables and categorized as self prepared, paid preparer use and IRS prepared. The IRS prepared tax-returns included any prepared with IRS assistance, whether by an IRS employee or the TCE/VITA programs. The number of exemptions claimed on each return was used as the exemption variable up to five. If there were more than five exemptions claimed, the variable value was set to 6. If the capital loss field was negative, then an initial capital gains flag is set to one. Later, using line item data, a more robust flag may be set.

Table 1. Income Bins	
Initial Bins	Super Bins
AGI < $0	Negative
AGI = 0	Low
$0 < AGI <$15k	Low
$15K < AGI < $30k	Middle
$30K < AGI < $50k	Middle
$50k < AGI < $80k	Middle
$80k < AGI < $120k	Middle
AGI > $120k	High

Another variable which required binning was the monetary adjustment of each overall tax-return: rar_ovedef_amt. See Table 2. When this field is negative, it indicates that the exam resulted in a lower tax liability than the original return indicated, i.e. the filer is owed a refund. If it is positive, then the taxpayer owes additional money to the IRS.

Bins were set so that there was an approximately uniform distribution.

Table 2. Monetary Adjustment Bins
Due/Owed Bins
Owe < $0
Owe = $0
$0 < Owe <$2k
$2k < Owe < $3k
$3k < Owe < $4k
$4k < Owe < $5k
$5k < Owe < $6k
Owe > $6k

After the initial adjustments and additions to the tax-return set, the line item set, E, was addressed. Of the line items included, 11.3M corresponded with tax-returns from C and were used. The first thing done was a determination of intent by reason code and by penalties. Very few line items were assessed penalties: 82k. Each line item had a reason code assigned by the examiner. These reason codes were split into intent groups after correspondence with the IRS. Possible values were intentional, non-intentional or inadvertent, neutral, possible intentional and "discard." It is important to note that only a subset of the reason codes were used to distinguish between intentional and inadvertent. If a line item had a penalty associated with it, it was also considered intentional. Later study revealed that this may not always be accurate. Finally, 57% of tax-returns are marked as having inadvertent errors and 4% are marked as having intentional errors.

Both the first principle and the empirical models used the same bins if they used the same variables. There are however, some differences in variables available to the two modeling teams. For example, first principle models considered information about gender which is not readily available from the tax-returns. Whereas, the empirical models information on the level of the monetary return that is not readily available without access to the tax-returns. By combining the models a more comprehensive view of the correlates of non-compliance is possible.

Additional information from the IRTF data set was fused with the EOAD data. We were only provided with a subset of the IRTF database, as such the corresponding records for some of the tax-returns in the EOAD were not available. Hence the set of tax-returns used from the EOAD was pared down to just those 1.9M records for which IRTF data was also available. The IRTF set contains information about the superset of taxpayers including the date of birth and additional line items used: EITC, student loan interest, capital gains and Social Security benefits. The taxpayers' ages and filing statuses were added to the tax-return data set. Ages were binned accordingly: under thirty, between thirty and sixty, and over 60 years of age.

In the EOAD data the rate of inadvertent and intentional errors as marked by the examiners increases with income (AGI) when looking at the percentages from the actual tax-return errors. See Table 3. The exception is the negative income category which has an even higher rate of error than the high income group. Note that the error rate is significantly lower across the board for

intentional error as compared with inadvertent error. In part, this is due to a reluctance of examiners to mark a return as containing an intentional error.

Table 3. Empirical Distribution of Inadvertent Error by Adjusted Gross Income Level					
Intent/Income	**Negative**	**Low**	**Middle**	**High**	**Total**
Inadvertent	23498	270356	630648	117910	1042412
Not Inadvertent	5133	308807	424153	46392	784485
Intentional	4671	10290	46396	14110	75467
Not Intentional	23960	568873	1008405	150192	1751430
Total	28631	579163	1054801	164302	1826897
Inadvertent%	82%	47%	60%	72%	57%
Intentional%	16%	2%	4%	9%	4%

In Figure 2, the percentage of errors of each type by income level is shown. As can be seen, the distributions are different for intentional and inadvertent errors. In general, there is a greater tendency to label tax-returns as containing intentional errors if the reports income is high or negative and inadvertent as low. This may reflect a bias on the part of the examiners due to the fact that the tax loss is higher in the negative and high income areas, or it may reflect a greater lack of financial literacy at low income levels. This difference in the distribution, and the lack of clarity on its cause, is one of the factors suggesting the need for a more comprehensive model of errors; rather than simply assuming that underpayment are intentional errors.

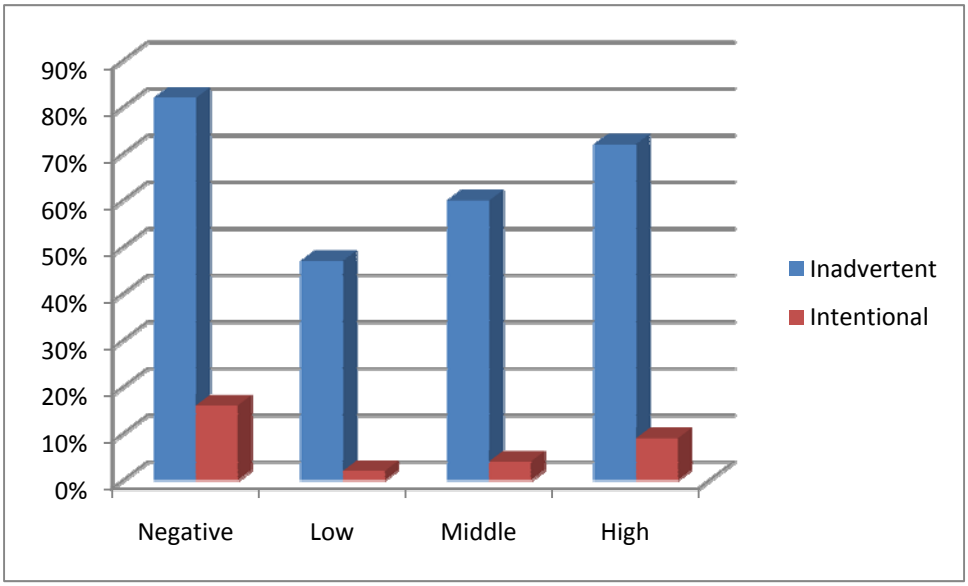

Figure 2. Percentage of Labeled Errors by Adjusted Gross Income Level

Model Details

The first principle and machine learning models employ different variables due to the way in which they are constructed. These differences are summarized in Table 4. These first principle models did not make use of the EOAD/IRTF data. The intentional error model contains variables that are available on the tax-return and so can be applied to the combined EOAD/IRTF data. The

inadvertent error model, at this point, contains less of that information and so cannot be applied to the EOAD/IRTF data as easily. As part of the next phase we will impute the relation between the EOAD/IRTF data and the first principles inadvertent error model. In this latter case, future work will seek to find a mapping between the variables in the inadvertent error first principle model and those items available on tax-returns.

Table 4. Variables Used by the Different Models

Variable	1st Principle Intentional	1st Principle Inadvertent	Machine Learning
EITC	no	no	yes
Age	yes	yes	yes
Burden/Complexity	no	yes	yes
Late	no	yes	yes
Filing Status	no	no	yes
Itemization	no	no	yes
Exemptions	no	no	yes
Preparer	no	no	yes
Error Amount	no	no	yes
Income	yes	yes	yes
Gender	yes	yes	no
Belief in obey law	yes	no	no
Education	yes	yes	no

Team A: The first principle models, as they are derived from the general literature and not the EOAD/IRTF data provide a principled way of characterizing errors that can be applied to any return. The model of intentional errors from first principles predicts the probability that individuals will commit some error as determined by their socio-demographic traits, namely gender, age, education, and income, as well as their attitudes toward obedience to the law (Lee & Carley, 2009). This model incorporates scientific findings from several published papers on tax evasion and represents their weighted average taking into account their similarities to the recent U.S. population. In Figure 3 the intentional error model, for the likelihood that the return contains a intentional error, as derived from the open source literature is shown. As can be seen tendency to believe that laws should be obeyed, age, and indirectly education are primary drivers.

The inadvertent error model from first principles takes into account issues of literacy, the relative complexity of the tax law, stress due to time of filing, and basic socio-demographic correlates of error to predict tax-payer mistakes. The basic inadvertent error model is shown in Figure 4. In this case general socio-demographic traits have a diagnostic role only to the extent they correlate with financial literacy and the expectation to receive a refund. In general, the dominant factor in producing an inadvertent error is task complexity; in other words, the burden in filling out the relevant line items.

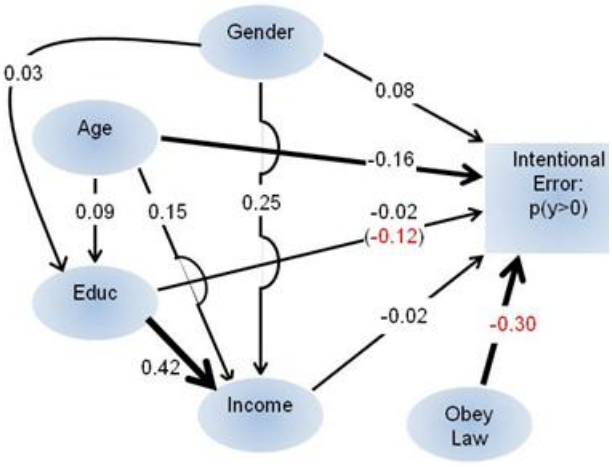

Figure 3. First Principles Intentional Error Model

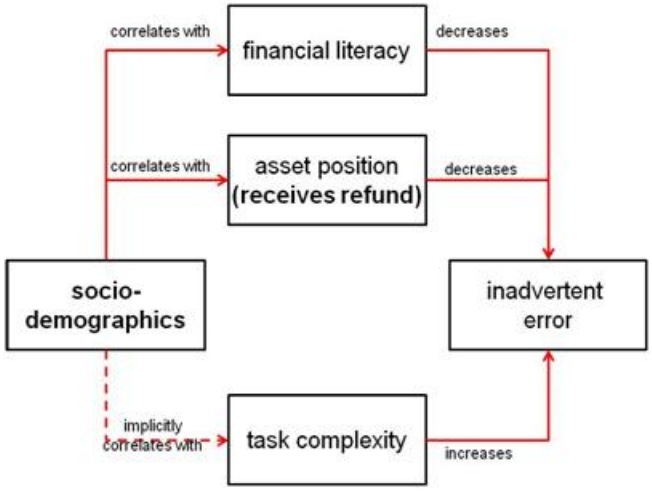

Figure 4. First Principles Inadvertent Error Model

Taken together, the two first principles models suggest about a 45-50% error rate of which about 30% are inadvertent and 20-30% are intentional. And, although we have not yet been able to estimate it, these models suggest that there are likely to be returns with both intentional and inadvertent error; particularly, when the complexity of the return is high.

Team B: The empirical model of errors is a composite model employing three machine learning and statistical techniques: the Proc Logistic regression model developed in SAS, a Bayesian Network Prediction model, and a j48 decision tree classifier with multi-boosting. The models for error were formulated with ten explanatory variables and a binary response variable. For one set, the response variable is intentional error and the other set has inadvertent error. The ten explanatory variables are: income, error amount as determined by the exam, complexity (burden), late code, preparer used, exemptions, filing status, age, EIC and itemization.

Proc Log is a linear regression procedure used to model dichotomous outcomes of interest, such as the error variables. A linear function is produced to model the relationship between the explanatory and dependent variables. The error variables were coded as "0" for no error and "1" for an error in order to be used with Proc Log. Proc Log can produce several "goodness of fit" indicators, but Proc Log was used primarily to produce classification tables for the IRS data. Once the classification tables were produced from the labeled set, they were used to predict outcomes in both the labeled and unlabeled sets for the intentional and inadvertent models.

The other software used for prediction was the Belief Net (BN) Power Constructor. This Bayesian network predictive software uses a conditional independence based algorithm to construct a directed acyclic graph. Given the binned variables, this software can produce a graph that will calculate error probabilities for each tax-return. Like the Proc Log classifiers, the resulting models are applied to the labeled and unlabeled sets for comparison. The predictive software (SAS and BNP) uses a tolerance of .5 to determine whether the model predicts that a particular tax-return has an error. Changing this tolerance lowers or increases the threshold for prediction. We used a tolerance of .5 for inadvertent errors and .1 for intentional errors. This difference is a direct result of the fact that there are so few known cases of intentional error.

The models are learned using data gleaned from the EOAD and IRTF data sets provided by the IRS. The EOAD data is split into two sets - "labeled" and "unlabelled." The labeled set is

further divided into two overlapping sets – "intentional" and "inadvertent." This was done at the full tax-return level and by line item. The unlabelled set had neither intentional nor inadvertent errors. The data was again split by four income groups: negative, low, medium and high. Each of these income groups has a substantially different profile in terms of taxpaying behavior and so errors. These splits were applied overall and by line item. Several line items or issues associated with each tax-return were derived from the line item set. These include tips, self-employment income, farm income, alimony, as well as another indicator for capital gains. In lieu of learning separate models for exam types, such as field or campus, we simply controlled for complexity.

Comparing, Contrasting and Testing the Models – Team C: The first principle intentional error models and the empirical models for intentional and inadvertent errors are applied to the labeled sets to determine how well the models work. This is done for the overall tax-return and by selected line items. After the models are assessed using the labeled data, they are then applied to the unlabeled sets to determine how many of these forms can be characterized. Finally, to create a composite model, the predictions of the various independent models are combined. Both intersection and union are explored.

Model results are strongest when controlling for income as cause, type and level of error are different. There is substantial overlap among models suggesting a class of cases where there is strong ability to discriminate between intentional and inadvertent errors. However, each of the models has a different strength with respect to the cases with less clear signals. Hence, a composite model, formed by combining the diverse models provides a more comprehensive assessment.

Results

Both first principle and machine learning models were built separately for inadvertent errors and for intentional errors. These models suggest that it is possible to discriminate apparently intentional from inadvertent errors for most returns. Of the 1,042,412 tax-returns marked as inadvertent by the examiners, 81% are predicted to be inadvertent using machine learning models. Of the 784,485 tax-returns marked as intentional by the examiners, approximately 50% are predicted to be intentional using the machine learning models. Of the records marked as both intentional and inadvertent by the examiners, approximately 84% are predicted to be both inadvertent and intentional using the respective machine learning models. Using the first principles models, a higher percentage of the tax-returns are marked as containing intentional errors.

Of the tax-returns marked as inadvertent, 2% are predicted to be intentional by the empirical models. There are two possibilities: 1) the flags which are set by the examiner are wrong, or 2) the flags are correct and the intentional error models are "over" predicting. If the flags are wrong, then this 2% means that these models identify an additional 2% of the cases as containing intentional errors. If the flags are correct, this 2% error means that we would expect these intentional error models to incorrectly suggest that returns might contain intentional errors 2% of the time for returns already selected as thought to contain an error. This would be the cap on the inaccuracy of these models.

We expect that refined models that look at line items, and explore the correlations among those, may increase further the predictive value of the results. We also expect that combining the final models from Teams A and B will result in a better general model that can be used by the

Service in a variety of ways, including compliance related education for both IRS enforcement staff and the taxpayer.

We now turn to a more detailed analysis of the modeling results for inadvertent and then intentional errors. In this more detailed analysis we consider both the labeled and the unlabeled exams.

Inadvertent Errors

We developed from the open literature a general or first principles model of inadvertent errors. This is shown in Figure 4. As can be seen, two factors that drive inadvertent errors, are complexity of the problem and financial literacy. The greater the complexity, the less literate the individual, the more likely that an inadvertent error will be made.

Predicting Inadvertent Errors (.5 Tolerance)

Labeled Set – Only Those Classed as Inadvertent

Accuracy results from applying the learned models to the known or labeled set of tax-returns are shown in Table 5. Note that the predictive models return a percent likelihood of error for each tax-return. The tolerance for these outcomes is set at the default of .5. At .5, the sum of the percentage of correct positives and correct negatives are usually maximized, for both types of error. The tolerance is not a confidence interval. It is simply a cut-off point for whether the exam is predicted to have an error or not. Moving away from .5 increases the number of likelihood of false positives or false negatives. In an operational context, a different tolerance might be used for intentional errors if, e.g., the policy was to examine all possible cases of intentional error even if there is a high chance that the error if there was one was not intentional. Similarly, for inadvertent errors, a policy that education never hurts, might use a tolerance that produces a high level of false positives.

In table 5,8,11 and 14 the percentage errors for labeled tax-returns is shown. To generate the values shown the following factors were considered. Note that there are two ways for a model to match the conclusions of the examiner. A model can label the tax-return as having the same type of error (inadvertent or intentional) as marked by the examiner. We refer to these as Confirmed Errors. Or, a model can label the tax-return as not having an error of that type and the examiner also marks the tax-return as not having an error of that type. These are Confirmed Non-Errors and will not be reported. Similarly, there are two ways in which the models can mis-match the examiners. A model can label the tax-return as having an error of that type and the examiner did not mark it as such. We refer to these as Potential Errors as they are tax-returns that the models would also characterize as having an error of that type. Or, a model can label the tax-return as not having an error of that type but the examiner did mark it as having an error of that type. We refer to these as Mistakes.[3] It should be noted that the difference between the %of returns marked as having that type of error by the examiner (see table 3) and the percentage of the returns that are Confirmed Errors, are the mistakes. The basic idea behind this demarcation is that although examiners may under-report errors, if they do mark an exam as containing a particular type of error they are unlikely to be wrong. The percentages under mistakes can be thought of as the

[3] From an experimental perspective, these categories of match and mismatch are the same as the traditional False+ and False- distinction that is used when ground truth is known. Since, there is reason to suspect that the examiner markings contain errors, we use the term match and mismatch instead of correct and false. In summary, Potential Errors are, from a ground truth perspective, false positives and Mistakes from a ground truth perspective are false negatives.

minimum level of inaccuracy expected when these models are used. Another feature of many of these tables is that we present results for both an intersected and a union approach on confirmed errors. In tables with these combinations, a ∩ is used for intersection on confirmed and a ∪ to denote union on confirmed. This refers to the way in which the models were combined for the confirmed errors, as well as confirmed non-errors. In the case of the union, the potential errors are those cases where none of the models suggested it was not in error.

Since more exams are marked as inadvertent than as intentional by the examiners, 57% and 4% respectively, if a model were to exactly match the examiners the maximum number of labeled tax-returns they would label as intentional would be 4%. A model that exactly matches the examiners would for inadvertent errors have a higher percentage of the returns characterized as Confirmed Errors and for intentional errors have a higher percentage characterized as Confirmed Non-Errors. The sum of Confirmed Errors and Potential Errors is the percentage of labeled exams a model suggests has that type of error. The maximum possible predicted error that can be confirmed is also shown in these tables.

Looking at table 5, we can see that for inadvertent error the minimum level of inaccuracy is highest when the return is from someone in the middle income area. In contrast, for negative and high income cases the models tend to mark as inadvertent the same cases marked by the examiners. Moreover, the models suggest that over 90% of these returns contain inadvertent errors. Whereas, the models suggest that 40-50% of the low income returns and 65-75% of the middle income returns contain inadvertent errors.

Table 5. Inadvertent Error Predictions by Models Independently and Collectively Given Labeled Tax-Returns

Income	Negative		Low		Middle		High	
	Confirmed Error	Potential Error	Confirmed Error	Potential Error	Confirmed Error	Potential Error	Confirmed Error	Potential Error
BNP	80.96%	16.51%	30.03%	16.51%	49.93%	23.51%	71.76%	28.24%
PL	80.98%	16.51%	29.80%	17.16%	49.42%	24.39%	70.29%	26.64%
BNP ∩ PL	80.51%	16.96%	28.24%	18.52%	47.02%	26.46%	70.29%	28.24%
BNP ∪PL	81.43%	16.06%	31.59%	15.14%	52.33%	21.44%	71.76%	26.64%
Confirmed Maximum	82.00%		47.00%		60.00%		72.00%	

The accuracy is highest for negative and high incomes. Also, there is a great deal of overlap between the two models. However, the percentage of false positives is quite high. Increasing the tolerance or threshold for a positive result will minimize the false positives, but at a cost to overall accuracy. If examiners have a tendency to mark exams as inadvertent, even if they are not, then these models can be interpreted as suggesting that for low and middle income cases, 16% and 10% of the cases respectively, may have been erroneously marked as inadvertent.

Unlabelled Data

Table 6 contains the models' predictions for inadvertent error in the unlabeled set of tax-returns. Note that the percentages predicted are higher than the actual percentages based off the

labeled set. This may be, because something about the tax-return or the tax-payer alerted the examiner that this case was inadvertent so they just didn't mark it. Or, this may be because there were other signals during the exam for the labeled cases that suggested they were intentional.

Table 6. Percentage of Unlabeled Exams Predicted to be Inadvertent by Models				
Model/Income	**Negative**	**Low**	**Middle**	**High**
BNP	83.4%	69.0%	81.4%	100.0%
PL	80.3%	66.9%	84.0%	94.9%
BNP ∩ PL	76.3%	62.8%	76.6%	94.9%
BNP ∪ PL	87.4%	73.1%	88.8%	100.0%

Profiles of Tax-returns with Errors Where the Errors are Likely to be Inadvertent

Because so many examined returns have inadvertent errors, picking definitive profiles is challenging. Many returns have both intentional and inadvertent errors. Nevertheless, trends definitely emerge. Illustrative profiles by income level are shown in Table 7. For all income groups, higher burden is associated with inadvertent error. We note, that the first principle model for inadvertent error also suggests the complexity (and so burden) is a major contributor to inadvertent error. In this table NA means not applicable.

Table 7. Profiles Consistent with Inadvertent Errors								
	Age	**Use Paid Preparer**	**Itemized**	**Income**	**Late**	**Burden**	**EIC**	**FS**
Low	Mixed	Less Likely	More Likely	Mixed	Mixed	High	More Likely	Mixed
Middle	Older	Less Likely	More Likely	Higher	More extensions	High	More Likely	Joint – More likely
High	Slightly Older	Slightly more likely	Mixed	NA	Mixed	High	NA	Mixed
Negative	Slightly Older	Slightly more likely	Mixed	NA	Mixed	High	Mixed	Joint – More likely

Burden is consistently higher for erroneous tax-returns. Although it is not always higher for every single tax-return, when looking at the percentages of erroneous tax-returns versus ones without inadvertent error, a clear pattern is shown. For example, 90% of non-erroneous tax-returns in the negative group are in the lowest burden group. 80% of those in the error group were in the highest burden group. Also, the percentage of those married, filing jointly, increases in each erroneous group. This may be a result of more opportunity for error as more lines of tax-returns must be completed compared with those filing singly or a head of household. Also, younger taxpayers (in the < 30 bin) have lower percentages of erroneous tax-returns. Again, this may be due to younger people having less complicated tax situations in general.

Predicting Intentional Errors (.1 Tolerance)

Labeled Set – Only Those Classed as Intentional

The accuracy results from applying the learned models to the known or labeled set of tax-returns are shown in Table 8. While the accuracy percentage is quite high (80-90%), there are many false negatives. Essentially, the models under-predict intentional errors at the .5 level, resulting in a high number of correct negatives. When the tolerance is set to .1, then a wider net is cast and more tax-returns will be classed as intentional. This lowers the number of cases where a model does claims there is no intentional error and the examiner marks the exam as containing an intentional error (false negatives). And, it increases the number of cases where a model claims that the error is intentional and the examiner does not (false positives). While there is a great deal of overlap in the Bayes Net and Proc Log models, the first principles model (FP) has different yet still similar results. Overall by combining the models – a stronger result is produced.

We set the tolerance lower for intentional than for inadvertent errors for two reasons. First, there were simply far fewer tax-returns marked as intentional. Second, by setting it lower, the overall mismatch with the examiners is lower. However, even though the overall mismatch is lower the number of returns where a model suggests there is an intentional error and the examiner does not will be higher. Thus, we erred on the side of forecasting potential errors.

In table 8, we see that the first principle model, and the union of models with the first principle models tends to predict more intentional errors, and to have lower minimum levels of inaccuracy. As with the inadvertent errors, the models are better for negative and high income than for low income.

Table 8. Intentional Error Predictions by Models Independently and Collectively Given Labeled Tax-Returns

Income	Negative		Low		Middle		High	
	Confirmed Error	Potential Error	Confirmed Error	Potential Error	Confirmed Error	Potential Error	Confirmed Error	Potential Error
BNP	14.17%	56.02%	1.03%	4.36%	2.72%	13.14%	5.71%	24.92%
PL	14.28%	55.43%	1.03%	4.66%	2.68%	13.06%	6.09%	27.04%
FP	13.15%	59.41%	1.43%	42.20%	2.61%	37.80%	4.79%	33.43%
FP∩PL	11.86%	71.35%	0.94%	42.59%	2.02%	41.14%	3.73%	46.45%
FP ∪ PL	15.58%	43.49%	1.53%	4.27%	3.27%	9.72%	7.15%	14.01%
FP∩BNP	11.78%	71.98%	0.93%	42.52%	2.03%	41.40%	3.65%	44.52%
FP ∪ BNP	15.54%	43.45%	1.53%	4.04%	3.30%	9.54%	6.85%	13.82%
BNP∩PL	13.48%	62.01%	0.93%	5.33%	2.57%	14.28%	5.42%	29.93%
BNP ∪ PL	14.97%	49.45%	1.14%	3.69%	2.83%	11.93%	6.38%	22.03%
∩ all	11.29%	73.61%	0.85%	42.68%	1.95%	41.74%	3.51%	48.03%
Union All	15.77%	39.13%	1.55%	3.46%	3.34%	8.92%	7.29%	12.51%
Confirmed Maximum	16.00%		2.00%		4.00%		9.00%	

Unlabeled Data

Since the predictive models for intent determine so few errors, lowering the tolerance to .1 results in percentages of erroneous tax-returns more in keeping with the actual exam error percentages. These results are shown in Table 9.

Profiles of Tax-returns with Errors Where the Errors are Likely to be Intentional

By income level, the profile of tax-returns with intentional and non-intentional errors are somewhat different. For all four income groups, markers for intentional error include self preparation, age above thirty years, high complexity and no EITC. For all income groups except low income, itemized deductions were also well represented. One consistent difference is the representation of head of household filers. They are consistently more represented in the "no error" group. Less of them and more married taxpayers appear in the group making intentional errors. It should be noted that people may claim the Head of Household status who are not eligible to do so. This was not controlled for. If, we could determine that this claim was wrong then that might move some of these cases to the intentional error category; if, in fact, this error was not inadvertent. However, the complexity of determining eligibility for head of household status in and of itself, is likely to increase both intentional and inadvertent errors. Taking into account errors on other factors, such as head of household status, is a point for future research. These profiles for intentional errors are shown in Table 10.

Table 9. Percentage of Unlabeled Exams Predicted to be Intentional by Models

Model/Income	Negative	Low	Middle	High
BNP	39.3%	2.0%	6.3%	7.0%
PL	30.5%	2.3%	5.7%	9.0%
FP	35.7%	30.2%	23.4%	25.1%
FP∩P PL	14.7%	0.9%	2.5%	2.2%
FP ∪ PL	51.5%	31.5%	26.7%	32.0%
FP ∩ BNP	13.9%	1.0%	2.4%	2.1%
FP ∪ BNP	61.1%	31.2%	27.3%	30.1%
BNP ∩ PL	23.4%	1.2%	4.9%	5.5%
BNP ∪ PL	46.4%	3.1%	7.2%	10.6%
Intersect all	11.0%	0.7%	2.0%	1.6%
Union All	64.5%	32.1%	27.7%	33.1%

Table 10. Profiles Consistent with Intentional Errors

	Itemized	Late	Exemptions	Error Amount	Filing Status
Low	No	Extension and No File	<2	Very High and Low	Single and Married-J
Middle	Yes	Extension	Mixed	Very High and Low	Single and Married-J
High	Yes	Mixed	Mixed	Very High and Low	Married-J
Negative	Yes	Extension	Mixed	High	Married-J

Line Items - EITC

The first line item modeled was the Earned Income Tax Credit. This was because it is one of the most examined line items, being concentrated in low and middle income groups. It is also one of the most complex of the line items and as such, according to the theoretical first principles models, the likelihood of both intentional and inadvertent errors is likely to be higher than for other line items. Over 940K EITC line items were examined in the set. The average adjustment was -$2, 285 and the total was -$2.15B. Almost all returns were labeled as containing inadvertent errors (99%+) while there were very few returns marked as containing intentional errors (<1%) for all income groups except the high income group. Due to the nature of the EITC line items, there are no tax-returns that employ this line item that are in the High Income bracket. The models behaved accordingly. We note that, the distribution of errors for the EITC line item is not symmetric about zero; i.e., in most cases the errors result in tax-loss (under-reporting). The distribution is slightly more symmetric for taxpayers with low income for those taking the EITC than for other income levels. As with the entire tax-returns we do not make the assumption that errors resulting in under-reporting are intentional.

EITC Models – Labeled Set – Only Those Classed as Inadvertent

The BNP and Proc Log models for error on the EITC line item, unlike the corresponding models for error somewhere on the overall tax-return, does not use taking the EITC credit as a control. The EITC error results are shown in Table 11. The BNP line item model for EITC results in a much higher percentage of false positives than the full tax-return BNP model. The Proc Log model outperforms the full tax-return model significantly and, remarkably, does not overlap very much with the BNP model. This line item may be a good candidate for ensemble learning because of this lack of overlap. By combining the models in an ensemble the strengths of both individual models can be exploited. It is likely that the Proc Log model is over-estimating the likelihood of inadvertent errors. As such, in this case, it would not be reasonable to use the union of the two models as the composite model of inadvertent errors. Another important point is that the minimum level of inaccuracy is much lower for the Proc Log model than the BNP.

Table 11. Inadvertent Error Predictions by Models Independently and Collectively Given Labeled Tax-Returns for Inadvertent Error on the EITC Line Item

Income	Negative			Low			Middle		
	Confirmed Error	Potential Error	Mistakes	Confirmed Error	Potential Error	Mistakes	Confirmed Error	Potential Error	Mistakes
BNP	41.98%	0.00%	58.02%	33.53%	0.75%	65.72%	36.80%	0.12%	63.08%
PL	100.00%	0.00%	0.00%	86.57%	4.88%	8.54%	88.07%	4.23%	7.71%
BNP ∩ PL	41.98%	0.00%	58.02%	29.24%	4.93%	65.84%	32.65%	4.25%	63.10%
BNP ∪PL	100.00%	0.00%	0.00%	90.87%	0.71%	8.42%	92.22%	0.10%	7.69%

EITC Models – Unlabeled Inadvertent EITC Line Items Compared with Intent on Overall Tax-Return

As previously noted, the unlabeled set had no error designation, so for the sake of comparison, the results of the EITC line item models were compared with the intent ascribed to the overall tax-return. This compares, for a specific return the type of error on a line item with the type of error on the tax-return as a whole. For the unlabeled set, the Bayes Net model outperformed the Proc Logistic model. Proc Log tended to mark the vast majority of the line items as inadvertent, which resulted in large percentages of false positives. For the line items, the model can be applied to the unlabeled data directly, and/or in comparison with the predicted intentionality of the tax-return as a whole. In Table 12, the latter is shown. In this case, we assume that the predicted type of error for the tax-return as a whole is correct. Then if a model labels the EITC line item as inadvertent and the parent model labeled the overall tax-return to be inadvertent, we would say that is a confirmed error. If a model labels the EITC line item as inadvertent when the overall tax-return was not labeled as inadvertent then that model is suggesting there is a potential error on that line item. If a model does not label the EITC line item as inadvertent but the overall return was labeled as inadvertent then that model is either mistaken, or the source of error is on a different line item. From a conservative point of view, then, the minimum inaccuracy would be that all of these last cases are actually model mistakes and the percentage shown can be thought of as the minimum possible mistakes.

Table 12. Match of the Models Independently and Collectively for Inadvertent Errors on the EITC Line Item for Unlabeled Tax-returns Assuming that the Prediction for the Overall Tax-return holds.

Income	Negative			Low			Middle		
	Matches Overall Exam	Potential Error	Mistake	Matches Overall Exam	Potential Error	Mistake	Matches Overall Exam	Potential Error	Mistake
BNP	50.98%	4.25%	44.77%	77.55%	0.42%	22.02%	77.19%	3.56%	19.25%
PL	50.98%	48.53%	0.49%	24.74%	73.30%	1.95%	23.76%	74.74%	1.50%
BNP ∩ PL	6.70%	48.53%	44.77%	4.66%	73.31%	22.03%	5.99%	74.74%	19.27%

Inadvertent EITC Profile

The EITC profiles are somewhat different from those for the tax-returns when viewed in their entirety. See Table 13. All but two negative income tax-returns were classified as having inadvertent errors, so there was no basis for a comparison. Also, high income tax-returns were excluded as so few claimed EITC. Again, the people who tended to make errors were a bit older and the complexity somewhat higher, but it wasn't as pronounced as in the whole tax-returns. There were fewer distinguishing characteristics between those who made errors and those who didn't. This is at least partially due to the high percentage (over 99%) of those making errors.

Table 13. Profiles Consistent with Inadvertent Errors on EITC Line Item						
	Age	**Preparer**	**Complexity**	**Exemptions**	**Error Amt**	**FS**
Low	Mixed	Slightly more self	Higher	More <2	Higher	More Singles
Mid	Slightly Older	Mixed	Mixed	More <2	Higher	More Singles And HOH

EITC Models – Labeled Set – Only Those Classed as Intentional

Very few of the examiners marked the EITC line item as containing an intentional error. Table 14 clearly demonstrates the effects of "rare events" on our models. The rare event in this case is the designation of an intentional error on the EITC line item. The Bayes Net Model marked every single tax-return as not having an error and was not included. The other two models marked some, though very few, tax-returns as having intentional errors on the EITC line item. As with the tax-returns as a whole, for the EITC line item, the models are quite likely to label a return as not having an intentional error when the examiner also marks it as such. Overall, these models are more able to identify inadvertent errors on the EITC line item than on the return as a whole and are unable to identify intentional errors marked by examiners. This shortcoming is likely to be overcome by simply building the models on more data.

Table 14. Intentional Error Predictions by Models Independently and Collectively Given Labeled Tax-Returns for Intentional Error on the EITC Line Item									
Income	**Negative**			**Low**			**Middle**		
	Confirmed Error	**Potential Error**	**Mistakes**	**Confirmed Error**	**Potential Error**	**Mistakes**	**Confirmed Error**	**Potential Error**	**Mistakes**
PL	0.41%	0.00%	0.00%	0.01%	0.00%	0.58%	0.00%	0.00%	0.77%
FP	0.00%	0.00%	0.41%	0.00%	0.00%	0.58%	0.00%	0.00%	0.77%
FP ∩ PL	0.00%	0.00%	0.41%	0.00%	0.00%	0.58%	0.00%	0.00%	0.77%
FP ∪ PL	0.41%	0.00%	0.00%	0.01%	0.00%	0.58%	0.00%	0.00%	0.77%

EITC Models – Unlabeled Intentional EITC Line Items Compared with a label of intentionality on the Tax-return as a Whole

Compared with the predictions for the return as a whole, the EITC line item model produces similar results. See Table 15. Again, the Bayes Net Model was excluded as it marked no errors. This data is suggesting that even when the overall exam is likely to include an intentional error, that intentional error is likely not to be on the EITC line item. One must be cautious in over interpreting the EITC result, however, as it is based on so little data.

Table 15. Match of the Models Independently and Collectively for Intentional Errors on the EITC Line Item for Unlabeled Tax-returns Assuming that the Prediction for the Overall Tax-return holds.									
Income	**Negative**			**Low**			**Middle**		
	Confirmed Error	**Potential Error**	**Mistake**	**Confirmed Error**	**Potential Error**	**Mistake**	**Confirmed Error Overall Exam**	**Potential Error**	**Mistake**
PL	0.16%	0.33%	3.92%	0.00%	0.05%	0.13%	0.00%	0.01%	0.11%
FP	0.00%	0.00%	4.08%	0.00%	0.00%	0.13%	0.00%	0.00%	0.11%
FP ∩ PL	0.00%	0.33%	4.08%	0.00%	0.05%	0.13%	0.00%	0.01%	0.11%
FP ∪ PL	0.16%	0.00%	3.92%	0.00%	0.00%	0.13%	0.00%	0.00%	0.11%

Intentional Profiles

Very few intentional tax-returns were identified by the models, which is not surprising as intentional errors are truly rare events. Like whole tax-returns, for the EITC line item, the preparer tends to be self and the complexity high. Also, once again, head of household is not as represented in the error group. See Table 16.

Table 16. Profiles Consistent with Intentional Errors on EITC Line Item									
	Age	**Preparer**	**Itemized**	**Income**	**Late**	**Complexity**	**Exemptions**	**Error Amt**	**FS**
Low	<60	Self	No	Lower	Less on time	High	More <2	High	Very Few HOH
Middle	<60	Self	Yes	Mixed	Mixed	High	More <2	Mixed	Very Few HOH
Negative	>30, <60	Mixed	Mixed	Mixed	On Time	High	Mixed	Mixed	Married-J

Discussion

It is important to recognize that these models are not true models of error so much as models of error as determined by the IRS examiners. This is due to the data used. The EOAD data contain only operational exams. Consequently the tax-returns are not representative of the population. They were selected for examination because of some perceived noncompliance. In deciding which tax-returns to further examine, a set of selection criteria are used resulting in a set of tax-returns that are suspected to contain errors. Thus the first source of bias is selection on the dependent variable – error. Future work should take the proposed models and test against a random sample of all tax-returns.

The second limitation is the criteria for defining error. The criteria we used for asserting that the tax-return contained an intentional error was that the examiner marked it as such. If an examiner did not mark a tax-return as intentional then we would not have marked it as such. In general, examiners cannot know for sure whether an error is intentional or inadvertent. Making that judgment requires knowledge of the taxpayer's true motives at the time of preparing the return or possibly an admission of intent. However, such information is generally not available. In addition, examiners have very significant incentives not to characterize an error as intentional since that generally carries with it a higher standard of proof. While, taxpayers have every incentive to claim that they forgot, lost, or didn't know something; for one taxpayer, that may be true and inadvertent, but for another similarly situated taxpayer, it may be a simple attempt to cover up intentional noncompliance. To mitigate this bias we used a jittering approach where we tested the models by adding relabeling a few of the tax-returns as intentional or not and rebuilding the models. This did not appreciably change the results.

Discussions with examiners also led to the conclusion that expectations about the source of error and/or level of error impacted the type of exam; e.g., field or campus. This source of bias is related to the differing proportions of intentional marking given the different types of exams. To mitigate this source of bias, all tax-returns were considered collectively with controls for types of exams considered.

Conclusion

This research suggests that it is possible to identify factors associated with intentional and inadvertent non-compliance on tax-returns. From a theoretical perspective, the core difference in causes of errors from in the first principle models is that a belief in obeying laws will decreases intentional errors and is irrelevant for inadvertent errors; whereas, complexity or burden is a strong predictor of inadvertent errors and is not a direct predictor of intentional errors. The machine learning models suggest that for inadvertent errors age, use of paid preparers (no for negative and low income, yes for middle and high income), taking the EIC, and the overall burden/complexity of the exam are diagnostic suggest that the error is inadvertent; whereas filing late, taking multiple exemptions, and larger errors are diagnostic of the error being intentional.

The most challenging part of this effort has been dealing with the rare events. In general, many statistical leaning models work better when there are vast quantities of data and when the data contains a uniform set of results. While a 50/50 split on the results (inadvertent/intentional) is not required; a more even split than 99.9/.01 is helpful. Despite the rarity of the event (the intentional error) trends are definitely emerging for both inadvertent and intentional errors; however, more work is needed on the models to increase the accuracy and robustness of the results. This challenge is difficult for the tax-returns as a whole; but, it is even worse for the individual line items. One possible way of mitigating this would be to get more data. Another, would be to see if imputing labels for line items that are not labeled, when the exam as a whole is labeled would alter the results.

Our investigations suggest that the key to improved accuracy is by employing ensemble techniques that blend results from multiple diverse models. As noted previously, the various models have different strengths and weaknesses and as such tend to pick up on different aspects of the factors that lead to errors. By blending the models a more robust comprehensive pictures emerges. We note that even blending the Bayes Net, the Proc Log, and the first principles models improve the predictive model for the intentional errors. We expect the same will be true for the

inadvertent. The gains, however, will be larger for intentional than for inadvertent errors as a higher percentage of the tax-returns marked as containing inadvertent errors as opposed to intentional errors by the examiners were classified as inadvertent by the machine learning models. In addition, the gains will be larger at the individual line item level. If sufficient gain is made at the line item level it might be possible to then re-estimate the type of error for the exam as a whole using a composite of line item characteristics and overall exam characteristics. The lower accuracy of the machine learning models for intentional errors means greater room for improvement as additional machine learning techniques are employed. Although not reported here, we are currently investigating models that may reach as high as 80% accuracy.

The examination of individual line items is another way of improving the overall accuracy of the results vis-a-vie diagnosing at least those returns as containing errors that an examiner would have. With individual line items, there is still the problem that intentional errors are a rare event; however, restrictions on which taxpayers can utilize which line items does alter the proportions and makes the distribution slightly less rare. Further, by building models of errors for key line items, an overall improved ensemble model is made possible. Future work should expand on this by focusing on an exploration of additional line items and building a composite model using line item and overall predictions. Self employment promises to be a fruitful line item to consider.

Other ensemble techniques should also be used. For example, the intersection/union results for intentional models show that by adding in the first principle model accuracy can be improved. The next step here is to employ the specific coefficients from the first principle models for intentional and inadvertent models in the statistical models.

Finally, having a wider range of data would also help improve the model as it would provide more cases and examples of returns without errors. This would support the use of unsupervised learning techniques and enable us to make better use of the first principles models. Using such techniques are critical if we are to move further beyond the constraints impose by training models on the basis of exam results. The core issue will be determining the extent to which these techniques can provide useful models of error, intentional and inadvertent, that are independent of known biases.

References

Brostek, M. (2007). Tax Compliance: Multiple Approaches are Needed to Reduce the Tax Gap, Testimony before the Committee on the Budget, U.S. Senate. GAO-07-391T.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 263–291.

Kinsey, K. (1984). Survey Data on Tax Compliance: A Compendium and Review, Working paper #8716, (Chicago: American Bar Foundation, 1984).

Lee, Ju-Sung and Kathleen M. Carley, (2009). "Predicting Intentional Tax Error (Non-Compliance) Using Open Source Literature and Data", Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report CMU-ISR-09-125.

Milliron, V. & Toy, D, (1988). Tax compliance: An investigation of key features, The Journal of the American Taxation Association. 9, 84–104.

Smith, K., &Kinsey, K. (1987). Understanding taxpayer behaviour: A conceptual framework with implications for research. Law and Society Review, 21, 639–663.