Supplementing IRS Data with External Credit Report Data in Employment Tax-Predictive Models

Curt Hopkins and Ken Su (IRS, Small Business/Self-Employed Division)

Introduction

This project helped determine the value of adding business credit reporting data to existing IRS data in forecasting the likelihood of unpaid employment taxes. A credit bureau[1] provided business credit data and their three business credit scores for use in this project. We took data analysis and modeling approaches to reveal the relationship between unpaid employment tax and credit bureau credit information.

We used a specific point in time and balance due to define a noncompliant Form 941 employment tax return. Our time period target is the fourth quarter of 2012. The external data set includes eight quarters of information prior to this quarter. We define a noncompliant return as one with a balance due at first notice of at least $1,000 for data exploration and $5,000 for modeling. About 17.7 percent of records in our data set of about 288,000 businesses were noncompliant per the lower threshold and 7.3 percent per the higher threshold (unweighted).

We chose the employment tax arena for three reasons. First, prediction of employment tax liabilities is well understood by SB/SE Collection staff. Second, quarterly returns match the frequency of the credit scores in the provided dataset. Third, quarterly returns allowed for prediction across shorter time horizons, reducing the effect of external economic events.

Data Description

This project involved data from two sources: the IRS and a credit bureau. IRS data represent a sunk cost for research; it will be paid for whether these data tables are used in a specific model or not. The credit bureau data were available under an existing contract with Research, Applied Analytics, and Statistics (RAAS). Future use of credit bureau data will have an additional cost; if it benefits predictive models then a return on investment calculation will be appropriate.

IRS Data

The IRS data used in this project encompassed transcribed return data, payments, subsequent condition and transaction data, and other indicators regarding businesses the IRS already tracks in order to administer the tax system.

External Credit Bureau Data

The data from the credit bureau were new to this project and require more description. The data were provided to us by the RAAS staff based upon a stratified sample[2] devised by them for another project. The sample included 32 strata designed to represent many market subsegments including:

[1] We do not wish to name the specific credit bureau used. It was a major business in the market.

[2] The 32 strata weights were not provided, so all analyses and models are unweighted. This could have had a significant impact on the results.

- Four categories based on number of employees;
- Sole proprietorships and other business types;
- Semiweekly depositors and other deposit requirements; and
- Four definitions of compliance issues.

We used three business credit risk scores and nineteen other pieces of credit reporting data. The three credit scores include an overall business creditworthiness score (CredScore), one for financial risk prior to granting credit (FinRisk), and one to rate businesses that already owe a debt (CollPred). None of these is a product name from the credit bureau.

The first two scores are scaled from 1 to 100, with lower scores representing poorer credit.[3] The CredScore looks at risk in terms of severe payment delinquency. The credit bureau optimizes FinRisk to predict the potential to default on an obligation. CollPred ranges from 100 to 900, with the lower scores correlated with the lowest likelihood of payment. However, CollPred is not broken into classes (see below); it is purely an ordinal score.

The credit bureau summarizes the first two scoring systems into Risk Classes. Those Risk Classes with the lowest scoring businesses include a higher percentage of poor-performing businesses as defined by the target of the models. CredScore poor performers would have severe payment delinquency and FinRisk poor performers would be those defaulting on a debt. Table 1 summarizes the Risk Classes.

TABLE 1. Risk Classes and Percent of Poor-Performing Businesses

Risk Class	CredScore	Delinquents	FinRisk Score	Defaulters
High	1–10	50.8%	1–3	35.3%
High-Medium	11–25	19.1%	4–10	10.0%
Medium	26–50	10.0%	11–30	2.9%
Low-Medium	51–75	4.4%	31–65	1.1%
Low	76–100	1.7%	66–100	0.6%

SOURCE: Credit bureau white papers

We include 19 key business credit risk factors from the external data set in our analysis:

1. Count of new trades under delinquency;
2. Count of continuous trades under delinquency;
3. Days Beyond Terms (DBT) of combined trades;
4. Count of aged trades;
5. Count of aged trades under delinquency;
6. Count of trades under Days Beyond Terms (DBT);
7. Count of total trades under delinquency;
8. Count of banking liability relationship;
9. Count of leasing trades;
10. Count of leasing trades under delinquency;
11. DBT of additional trades reported within the last 4 months;
12. Count of UCC[4] filings—reported as write-offs or skips;

[3] There are also scores indicating missing and out-of-range values. We treated those scores as missing data in this project.

[4] Uniform Commercial Code filings, required whenever a company pledges assets as collateral.

13. Count of unsatisfied UCC filings within the last 24 months;
14. Count of legal filings in past 6 months;
15. Count of bankruptcies filed within 6 months;
16. Count of tax liens filed within 6 months;
17. Count of judgments filed within 6 months;
18. Count of open and closed collection trades placed within 12 months; and
19. Total number of inquiries in the last 3 full months.

Findings

We present our findings in three types of detailed analyses:

1. In data exploration, we report bivariate tables of external data across compliant and noncompliant taxpayers.
2. Our modeling analysis includes three phases: modeling from IRS data; modeling with both IRS and credit bureau data; and predicting changes in credit bureau credit risk using IRS data.
3. The cause and effect analysis tests if there is a causal link between the credit reporting and IRS data as a time series.

1. Credit Data Exploration

In this section, we explore the relationship between the credit bureau data and employment tax compliance. The data include the nineteen credit risk factors and three business credit scores described above. We prepared the data for analysis as follows:

A. We determined the compliance level of each business in each quarter.
    i. Noncompliant businesses are those with a balance due of over $1,000 at first notice.
    ii. Compliant businesses did not have a balance due (and thus no first notice).
    iii. Those with a balance due between $1 and $1,000 are not shown in this section in order to maximize the contrast between our two groups of interest.

B. We then developed an overall profile of each group's risk factors from the credit bureau business data perspective to uncover the relationships between noncompliant taxpayers and the credit bureau's data.

Table 2 demonstrates our profile of noncompliant and compliant businesses during the four quarters of 2012. The percentage in each cell quantifies the businesses meeting the credit risk factor. For example, in the very first data cell we show that in the first quarter of 2012, 0.90 percent of businesses with a balance of at least $1,000 at first notice also have at least one new delinquent trade reported by the credit bureau. This compares to the cell below, where 0.96 percent of compliant businesses had a new delinquent trade.

These results demonstrate that these credit risk indicators did *not* significantly differentiate taxpayers with a balance due from the compliant group. We confirmed this with z-tests on the larger differences. Despite these initially negative results, we allowed consideration of the credit risk indicators in our analysis by modeling.

Continuing our data exploration, we focused on the credit bureau credit scores. Three tables below show the percent of cases with a balance due in each credit bureau Risk Class for each type of credit score. For brevity purposes, the tables show the three scores for 2012, but the results are the same in 2011 and 2013. Further, the same pattern holds for balances of at least $5,000 in each year; the rate in each cell is lower, but the near-constant rate across score ranges is the same. Ranges and descriptions are those defined by the credit bureau.

TABLE 2. Percent With a Credit Risk Factor by Employment Tax Compliance Category

Compliance Category	1Q2012	2Q2012	3Q2012	4Q2012	Average
Business With a New Delinquent Trade					
Balance Due > $1,000	0.90%	0.83%	0.67%	0.53%	0.73%
Compliant	0.96%	0.88%	0.64%	0.54%	0.75%
At Least One Delinquent Continuous Trade					
Balance Due > $1,000	19.20%	19.75%	20.11%	21.32%	20.10%
Compliant	19.28%	19.97%	20.36%	21.69%	20.32%
At Least One Delinquent Trade >30 Days Beyond Terms					
Balance Due > $1,000	16.15%	16.58%	17.01%	18.02%	16.94%
Compliant	16.12%	16.60%	17.09%	18.01%	16.96%
At Least One Aged Trade					
Balance Due > $1,000	45.35%	45.55%	42.71%	43.20%	44.20%
Compliant	45.03%	44.94%	42.28%	43.02%	43.82%
At Least One Delinquent Aged Trade					
Balance Due > $1,000	12.14%	12.06%	12.97%	13.61%	12.70%
Compliant	12.36%	12.10%	12.99%	13.64%	12.77%
At Least One Trade Not Beyond Term					
Balance Due > $1,000	26.86%	27.27%	28.18%	29.58%	27.97%
Compliant	26.71%	27.21%	28.10%	29.58%	27.90%
Any Delinquent Trade					
Balance Due > $1,000	26.86%	27.27%	28.18%	29.58%	27.97%
Compliant	26.71%	27.21%	28.10%	29.58%	27.90%
Banking Liability					
Balance Due > $1,000	0.32%	0.34%	0.35%	0.36%	0.34%
Compliant	0.38%	0.40%	0.40%	0.41%	0.40%
Reported Leasing Trade					
Balance Due > $1,000	2.24%	2.26%	2.31%	2.41%	2.31%
Compliant	2.30%	2.33%	2.36%	2.45%	2.36%
Delinquent Leasing Trade					
Balance Due > $1,000	0.003%	0.003%	0.003%	0.005%	0.004%
Compliant	0.000%	0.000%	0.000%	0.004%	0.001%
Combined Trades under DBT					
Balance Due > $1,000				4.68%	
Compliant				4.68%	
At Least One UCC Filing Within the Last 24 Months					
Balance Due > $1,000	20.93%	21.64%	22.55%	23.03%	22.04%
Compliant	20.68%	21.35%	22.24%	22.65%	21.73%
At Least One Legal Filing in Past 6 Months					
Balance Due > $1,000	3.15%	3.48%	3.64%	3.82%	3.52%
Compliant	3.06%	3.49%	3.71%	3.83%	3.52%
Bankruptcy Filing in Past 12 Months					
Balance Due > $1,000	0.12%	0.13%	0.15%	0.17%	0.14%
Compliant	0.15%	0.15%	0.15%	0.18%	0.16%

SOURCE: IRS Compliance Data Warehouse (CDW) and credit bureau data.

TABLE 3. Unpaid Tax Rate Within Each FinRisk Risk Class

FinRisk		Percent With an Unpaid Balance > $1,000				
Score Range	Risk Class	1Q2012	2Q2012	3Q2012	4Q2012	Average
1 – 3	High	28.4%	28.3%	28.2%	27.9%	28.2%
4 – 10	High-Medium	28.9%	28.8%	28.9%	28.9%	28.9%
11 – 30	Medium	28.9%	29.2%	29.2%	29.2%	29.1%
31 – 65	Low-Medium	29.0%	28.9%	28.8%	28.7%	28.8%
66 – 100	Low	28.8%	28.8%	28.9%	29.0%	28.9%

SOURCE: IRS CDW and credit bureau data.

TABLE 4. Unpaid Tax Rate Within Each CredScore Risk Class

CredScore		Percent With an Unpaid Balance > $1,000				
Score Range	Risk Class	1Q2012	2Q2012	3Q2012	4Q2012	Average
1 – 10	High	28.7%	28.7%	28.9%	28.7%	28.8%
11 – 25	High-Medium	29.2%	29.1%	28.6%	28.8%	28.9%
26 – 50	Medium	28.9%	28.9%	29.1%	29.2%	29.0%
51 – 75	Low-Medium	28.8%	29.0%	29.0%	28.9%	28.9%
76 – 100	Low	28.9%	28.7%	28.8%	29.0%	28.9%

SOURCE: IRS CDW and credit bureau data.

The credit bureau does not provide category definitions for the CollPred Score. We created simple categories to see if the same pattern held. In this case, the percentages show those who paid a balance due within the next six months (to parallel the definition of this score provided by the credit bureau).

TABLE 5. Unpaid Tax Rate Within Each CollPred Risk Class

CollPred		Percent Paying the Balance Due Within 6 Months				
Score Range	Risk Class	1Q2012	2Q2012	3Q2012	4Q2012	Average
1–10	Very Low	6.8%	7.2%	7.3%	7.0%	7.1%
11–15	Low	7.4%	7.4%	7.5%	7.6%	7.5%
16–20	Low-Medium	7.6%	7.6%	7.5%	7.5%	7.6%
21–25	Medium	7.5%	7.5%	7.5%	7.6%	7.5%
26–50	Medium-High	7.4%	7.4%	7.4%	7.4%	7.4%
50+	High	7.7%	7.8%	7.8%	7.5%	7.7%

SOURCE: IRS CDW and credit bureau data

We found no indication that noncompliance increased with the Risk Classes defined by the credit bureau or the priority ranges set in a similar fashion. We confirmed this with chi-square tests showing that these percentages mimic a uniform distribution.

2. Analysis by Modeling

In the second set of analyses, we built models to determine the additional benefit of including credit reporting data in predicting future employment tax delinquencies.

Phase I

Initially, we built a model to predict which Forms 941 for the fourth quarter of 2012 would owe at least $5,000 at first notice using IRS information available from prior returns and other information known at the end of the prior quarter (third quarter 2012). We do include the Form 941 tax return for the third quarter of 2012, acknowledging that it is filed one month into the fourth quarter.

Phase II

Starting with the Phase I model, we then allowed consideration of credit bureau data up to the third quarter of 2012. The data included both the credit risk indicators and the three credit bureau scores for each business (detailed previously), as well as derived information such as score ranges and changes in scores across quarters.

Phase III

We then built models to predict the change in credit bureau Risk Classes from the third to fourth quarters of 2012. If the credit bureau scores predict taxpayer behavior, then we believe the relationship will hold true in the other direction and IRS data can support adequate predictive models.

General Modeling Methodology

After data preparation, we built logistic regression models using a modified stepwise method. The initial variables under consideration came from factor analysis of the available variables against the target variable. We allowed as many as 50 factors[5] in order to provide a broad selection of variables. We did not use the factors themselves, but rather selected the variable most correlated with the dependent variable from within each factor. Our intention was to have many variables available for stepwise consideration while minimizing autocorrelation among the available variables.

Methodology

We started with available information from each source, then transformed, binned, and made indicators from it, and then used a standard methodology to provide modeling variables and evaluate the results. Over one-hundred models were created at different times in this project; only results from the models determined best in their specific phase (based on diagnostic tests) are included in this report.

Variable Creation

Beginning with variables transcribed from tax returns, business entity information, and subsequent transaction and status changes, we expanded the variables by various types of recoding to support our modeling efforts. We used the following general techniques on the IRS data:

- Data Transformations

  For dollar amounts and counts of events (e.g., tax deposits, number of returns filed), we kept the raw data and added transformations by natural logs and square roots to provide three versions of each amount. We made a fourth version of tax return data by dividing dollar line items by the total wages reported on the return; this gives a less volatile amount, generally between 1 and 100 percent of the total wages.

- Data Binning

  Based on our experience with modeling payment compliance, we also converted dollar amounts to bins (ranges). This was especially useful for accounts receivable in the prior four quarters, as prior noncompliant behavior (e.g., owing $3,000 to $4,000 in the third prior quarter) is a good indicator of future noncompliance.

- Indicators

  We also set up indicators for specific conditions (e.g., prior installment agreements, prior notices of Federal tax lien, bad checks, and bankruptcy). Many of the “raw” variables from the IRS Compliance Data Warehouse (CDW) are themselves indicators (e.g., filing requirements, and specific transaction codes).

[5] No factor was included if its eigenvalue was below 1.0.

- Differences Across Quarters

  The techniques described above generally reshaped data within a specific quarter. We also compared quarters to each other and computed differences in dollar amounts, delinquencies, and counts between quarters. We set indicators for compliance events, such as a sudden drop in tax deposits or skipped filing. Finally, we set indicators and continuous variables based on the amount of variation in monthly and quarterly wages, deposits, and other payments.

Among these different techniques, we created a rich data set with over 550 variables considered for inclusion in the various models.

Credit Bureau Data

Using the ordinal ranges described above for the three credit bureau scores, we transformed this data in a similar way.[6] Raw scores were transformed by natural logarithms and square roots. We added binning of the first two scores based on the Risk Classes from the credit bureau and also created comparisons across quarters for changes—increases and drops in score ranges. In total, we added over 150 variables to the data set for the three scores from 11 quarters in our time window.

Variable Reduction

The data sets now contain many closely correlated variables and we set out to reduce these to a vital few for the models. For this, we used an existing program created by the Strategic Analysis and Modeling Group within Collection that creates factors and calculates the correlation of each member of the factor with the target variable.[7] With the factors made, we selected the variable with the greatest correlation to the target as our initial candidate and tried different numbers of factors to increase the potential candidates or reduce multicollinearity among the candidates.

Variable Selection

Variables selected for inclusion in the model were determined in iterations composed of two steps. The first step was stepwise logistic regression with an entry value of .01 and exit of .10. These values allow variables with strong predictive power into the model and tend to keep the variable in the model unless other (later) additions make it redundant. The second step required manual intervention. We evaluated the variables included in the initial model and, based on experience and graphical analysis, tested substitute variables. If the substitution resulted in better diagnostics, the change was incorporated into the next iteration of the model. It is quite possible (and did happen) that a square root transformation of a dollar variable was most closely correlated in the factor analysis, but with the inclusion of many other variables in the model, only an indicator was needed and not the continuous variable. We repeated these steps dozens of times to achieve our best models of future noncompliance and used this iterative process in each of the three modeling phases of this project.

Model Evaluation

In all models, we used nine standard (and one custom) diagnostics to determine if the model improved over the prior version. Here is a short summary of the diagnostics used in this project:

> **Akaike Information Criterion (AIC)**
>
> A measure of relative quality of model fit providing a means to compare models with differing numbers of independent variables.
>
> **Schwartz Criteria (SC)**
>
> SC is another model fit statistic similar to AIC, but it penalizes models more for including additional variables without a corresponding increase in predictive ability.

[6] While not strictly appropriate for ordinal data, we wanted to explore these alternatives due to the poor relationship found between the raw scores and compliance levels in the first section of our analysis.

[7] Balance due of $5,000 or more at first notice.

Somers’ D

This test of rank-order (ordinal) correlation varies from -1 to +1 with zero indicating no correlation.

Area Under Curve (AUC)

The AUC measurement can be used to compare multiple models. Based on the Receiver Operator Curve, the greater this value (on a 0.5 to 1.0 scale), the better the model’s predictive power. A value of 0.5 is considered to represent a random model, no better than flipping a coin.

Hosmer-Lemeshow Test (H-L)

The H-L Test compares the number of correct predictions in each decile (based on model score) to a theoretical distribution using a chi-square goodness of fit test.

Deviance

A measure used to help determine if a more complex model should be used. p-Values above 0.05 indicate that we should not reject the current model in favor of a more complex one.

Model Results Test—Percent in Top Decile

In this simulation of workload selection, the highest scores would have been sent forward for fieldwork. We compared the percentage of target cases found in the top decile as a proxy for moving the best selection (most true positives and fewest false positives) to the field.

Also considered during model development:

Global Likelihood Ratio (GLR)

This measure shows whether at least one of the independent variable coefficients is significantly different from zero.

Variance Inflation Factor (VIF)

VIF is a standard measure of multicollinearity among the independent variables in a model.

Model Development Findings

We present our modeling findings as a single table at the start of this section for ease of reference. Some findings and all conclusions draw across phases.

TABLE 6. Combined Model Performance and Diagnostics

Model	AIC	SC	Somers’ D	AUC	H-L	Deviance	Top Decile Percent
Phase I: Predict Balance Due on Form 941 with IRS Data							
IRS Data Only	101,618	102,353	0.72	0.86	269.0	0.36	56.5%
Phase II: Add Credit Bureau Data to Phase I Model							
Mixed Data	101,608	102,385	0.72	0.86	265.0	0.37	56.5%
Phase IIIa: Predict Worsening Credit Bureau Risk Class with IRS Data							
FinRisk IRS Data Only	120,665	120,675	0.00	0.50	N/A	0.53	10.4%
CredScore IRS Data Only	182,746	182,787	0.01	0.51	2.9	0.80	9.9%
Phase IIIb: Predict Worsening Credit Bureau Risk Class with IRS Data and Prior Risk Class							
FinRisk Mixed Data	113,706	113,840	0.34	0.67	2.9	0.50	22.8%
CredScore Mixed Data	172,586	172,752	0.32	0.66	983.0	0.77	17.3%

Phase I. Model Based Upon IRS Data Exclusively

The model from the IRS data shows solid diagnostics. Somers' D is over 0.70 and AUC above 0.85. The Deviance is close to 0.5 (optimal) and its p-value does not indicate the need for a more complex model. Finally, over 55 percent of target cases are found in the top decile.

TABLE 7. Combined Model Performance and Diagnostics

Model	AIC	SC	Somers' D	AUC	H-L	Deviance	Top Decile Percent
Phase I: Predict Balance Due on Form 941 with IRS Data							
IRS Data Only	101,618	102,353	0.72	0.86	269.0	0.36	56.5%

Phase II. Comparing Models Without and With Credit Bureau Data

In Phase II, we take the model from Phase I and supplement those models with the credit bureau information. After initial attempts at restricted variable choices,[8] using the same factor technique described for Phase I failed to improve the existing IRS data model, we allowed all credit bureau data (raw, transformed, categorized, and indicators for changes across quarters) into consideration for the model. Among these 22 variables (and dozens of derived indicators), the stepwise selection chose four to be included in the model: one based on the most risky categories of CredScore and three indicators for changes across quarters in the FinRisk Scores. No variables from CollPred made it into the final model. The variables chosen were:

CredScore_Bad032012	This indicator is set to 1 if the business was in either of the two riskiest classes based on the first quarter 2012 CredScore.
FinCat062012_Worse	This indicator is set to 1 when the second quarter 2012 FinRisk Class is at least one level more risky than it was for the same business in the first quarter of 2012.
FinCat092012_Worse2	This indicator is set to 1 when the third quarter 2012 FinRisk Class is at least *two* levels more risky than it was for the same business in the second quarter of 2012.
FinCat092012_Better	This indicator is set to 1 when the third quarter 2012 FinRisk Class is at least one level *less* risky than it was for the same business in the second quarter of 2012. Note: This indicator has a negative coefficient in the model, indicating that this condition links to a business that is less likely to owe in their fourth quarter return.

The first three variables selected have coefficients that indicate a greater likelihood of a balance due if the business was considered a poor risk nine months earlier (first condition), or had a worsening financial rating three or six months prior (second and third conditions). The last variable shows a negative coefficient, consistent with the lower risk of a balance due at the same time the credit bureau coded an improving financial rating for the business (in the prior three months).

Knowing which indicators were included in the model, we then evaluated the additional predictive power and stability of the model with this new information.

TABLE 8. Combined Model Performance and Diagnostics

Model	AIC	SC	Somers' D	AUC	H-L	Deviance	Top Decile Percent
Phase I: Predict Balance Due on Form 941 with IRS Data							
IRS Data Only	101,618	102,353	0.72	0.86	269.0	0.36	56.5%
Phase II: Add Credit Bureau Data to Phase I Model							
Mixed Data	101,608	102,385	0.72	0.86	265.0	0.37	56.5%

[8] That is, choices restricted to a single variable within each of 30 factors made from these credit bureau variables.

Akaike Information Criterion (AIC)

The addition of the credit bureau data improved the AIC by 10 points—dropping from 101,618 to 101,608.

Schwartz Criteria (SC)

As a modification of the AIC that penalizes models for including additional variables without a corresponding increase in predictive ability, the SC rose (i.e., worsened) by 32 points with the addition of the credit bureau data. The extra four variables moved the rating from 102,353 to 102,385.

Somers' D

This test of rank-order (ordinal) correlation varies from -1 to +1 with zero indicating no correlation. Both the IRS and IRS with credit bureau data models have a Somers' D of 0.72 indicating very good ordinal correlation.

Area Under Curve (AUC)

The addition of the credit bureau variables to the final IRS model did not change the AUC rating. It is 0.86 under both models.

Hosmer-Lemeshow Test (H-L)

Neither model had an H-L Test that would be considered acceptable. The H-L test is satisfied when the models do not reject the assumption that their decile distribution is identical to a theoretical one when tested using chi-square. In this test, both models are too good at placing the desired cases in the top decile and keeping them out of the bottom one. The H-L theoretical top decile is assumed to have 10,486 cases and the lowest 261. The models each place over 11,110 cases in the top decile and fewer than 180 in the bottom decile—with similar results in the middle eight deciles. In other words, the models performed so well at separating the desired cases by score that the chi-square test shows independence.

Deviance

This is a measure to help determine if a more complex model should be used. P-Values above 0.05 indicate that we should not reject the current model in favor of a more complex one. The IRS model has a Deviance 0.36 and the Mixed Data model of 0.37. Both p-values approach 1.0. The models are sufficiently complex.

Model Results Test—Percent in Top Decile

In this simulation of workload selection, the highest scores would have been sent forward for fieldwork. The IRS model placed 56.5 percent of the target cases (11,118 of 19,680) in the top (highest-scoring) decile. The Mixed Data model improved this slightly, placing 11 additional cases in the top decile (still 56.5 percent).

Global Likelihood Ratio (GLR)

This measure shows whether at least one of the independent variable coefficients is significantly different from zero. In comparing models, a higher score can be interpreted as having greater significance. The IRS model's GLR is 39,409 and the Mixed Data model is 39,426. The p-value of each is <0.0001, indicating that at least some of the variables in each model are significant predictors. In fact, the Wald statistics for each predictor in the IRS Data model have a p-value below 0.029, and in the Mixed Data model below 0.035, with each predictor rejecting the hypothesis that it adds no value to the model.

Variance Inflation Factor (VIF)

The addition of credit bureau data did not change the highest VIF among variables in the model. In the Mixed Data model, the same two variables[9] had the highest VIFs, although the order of them swapped. The ratings stayed on either side of 6.0, indicating a small concern with multicollinearity. This rating is not shown in the table.

[9] The two variables are indicators that: 1) the business owed at least $4,000 more in the first prior quarter than in the third; and 2) it owed at least $6,000 in the second prior quarter.

Phase III. Predicting Credit Bureau Scores or Categories Using IRS Data

If the credit bureau data were closely linked to tax compliance, it should be possible to predict certain features of the credit bureau scores using IRS tax data. Under this assumption, we tried to model the condition that the credit bureau score was in a worse Risk Class in the fourth quarter of 2012 than it was in the third. We chose movement between Risk Classes since this information was selected into the models in Phase II and should, therefore, be the most similar to IRS tax information. Only two of the three credit scores have Risk Classes.

TABLE 9. Combined Model Performance and Diagnostics

Model	AIC	SC	Somers' D	AUC	H-L	Deviance	Top Decile Percent
Phase IIIa: Predict Worsening Credit Bureau Risk Class with IRS Data							
FinRisk IRS Data Only	120,665	120,675	0.00	0.50	N/A	0.53	10.4%
CredScore IRS Data Only	182,746	182,787	0.01	0.51	2.9	0.80	9.9%
Phase IIIb: Predict Worsening Credit Bureau Risk Class with IRS Data and Prior Risk Class							
FinRisk Mixed Data	113,706	113,840	0.34	0.67	2.9	0.50	22.8%
CredScore Mixed Data	172,586	172,752	0.32	0.66	983.0	0.77	17.3%

Using just IRS data, the initial models (labeled IRS Data Only) were disappointing. Allowing one piece of credit bureau data into the mix made a large difference.

The first attempt at a FinRisk model was intercept-only (no independent variables selected). Only the intercept came in through stepwise regression, even when the level for entry eased to 0.10 and the stay criterion eased to 0.20. This held true even with all 556 IRS variables allowed into consideration. The AUC is indistinguishable from random chance and the other diagnostics are not much better. The initial CredScore model had a similar, though less extreme, set of diagnostics. Results with Somers' D at 0.01 and AUC at 0.51 are likewise not acceptable.

In the second part of this Phase, we allowed the Risk Class from the third quarter of 2012 into each model. We tried this as both ordinal and continuous data with very similar results. The continuous version is presented here. Providing the model with the prior Risk Class and predicting a more risky category the next quarter was more successful, providing a weak model. The overall model criteria (AIC and SC) came down, indicating better fit. The prediction diagnostics (Somers' D and AUC) moved to levels considered to indicate a poor model. It is important to note that almost all the improvement came from adding a piece of information not part of the IRS' tax data – the prior credit bureau Risk Class.

3. Granger Causality Testing

We built new unpaid employment tax and related credit bureau credit score time series data to perform cause and effect testing. The Granger Causality Test is a statistical hypothesis test to determine whether one time series is useful in forecasting another. The results of the Granger Causality Test are shown in the tables below.

TABLE 10. Granger—Causality Wald Test (Direction: IRS Data Predict Credit Bureau Data)

Using These Data	To Predict These Data	Chi-Square	Pr > ChiSq
IRS Compliance	CredScore	0.96	0.33
IRS Compliance	FinRisk	0.01	0.91
IRS Compliance	CollPred	0.24	0.63

TABLE 11. Granger-Causality Wald Test (Direction: Credit Bureau Data Predict IRS Data)

Using These Data	To Predict These Data	Chi-Square	Pr > ChiSq
CredScore	IRS Compliance	0.60	0.44
FinRisk	IRS Compliance	2.17	0.14
CollPred	IRS Compliance	0.72	0.40

The Wald tests show no evidence to reject the hypothesis that IRS compliance data and credit bureau scores are independent of each other at the 0.05 significance level. None of these three credit score time series can forecast our noncompliant taxpayer group or the other way around. Our results demonstrate that both data series are highly orthogonal. Credit bureau business data appear to have little predictive power in forecasting future employment tax noncompliance.

Conclusions

Our initial data exploration of credit bureau credit risk indicators and credit score risk categories showed little evidence that these data related to employment tax compliance. The inclusion of credit bureau data added a very small amount to the predictive power of the IRS-only model, resulting in just 0.1 percent more target cases in the top decile (Phase II). The six diagnostic tests are split, with one (AIC) showing an improvement, four showing no discernable change, and the last (SC) indicating that the additional predictive power does not justify the inclusion of four additional variables. Tests in the reverse direction show that IRS tax administration data and credit bureau scores show little, if any, relationship (Phase III).

Our final analysis showed no ability for credit bureau data to predict employment tax compliance, nor could IRS compliance data predict credit bureau credit scores. While not truly orthogonal, it appears that there is little relation between credit bureau business credit risk indicators or credit scores and the likelihood of owing employment taxes in the future.

Recommendations for Further Research

There is little evidence that including credit bureau data would lead to improved models for predicting employment tax compliance. Therefore, incorporating credit bureau data appears worthwhile only if there is no cost (either direct or opportunity) in bringing that data into the IRS for use in employment tax prediction. Because there is a measurable benefit in one area, it may be worthwhile to explore modeling outside the employment tax arena or testing individual (nonbusiness) credit scores at some point in the future.