PUI2015 Extra Credit

Understanding the Factors that correlate to Warranted NYC Business Tax debt

Lani M’cleod

Center for Urban Science + Progress, New York University

lrm238nyu.edu

Abstract:

This paper examines the relationship between several characteristics of business tax receivables and the warranting of such debt. Logistic regression was used to see if these characteristics could be used to predict the warranting of debt. Ultimately, this information could be used to prioritize case management and quality assurance of the warranting process.

Introduction:

The Department of Finance issues approximately 50,000 initial notices for tax due each year. About 6,800 are the result of account discrepancies rather than actual debt owed. While the number is low, reaching out and docketing the debt is labor intensive and inefficient. The benefits of identifying these receivables upfront would be time saved and a more efficient strategy for collection of remaining receivables.

Methodology

Receivables from 2013, 2014 and 2014 fiscal periods were identified as the dataset. Financial transactions related to business tax payments and return processing. Data available at the receivable level included noticing actions on the account, payments, abatements, and penalties. All these characteristics describe what the agency or taxpayer has done to pay off or verify the receivable. Other information, like the more detailed information about the customer, such as industry, filing history, filing method were at the customer or transaction level and not available in this data set.

I chose a set of categorical features: Tax type, filing period, amount due, assessment type, penalties, abatements, and balance from the dataset as the independent variables and set the existence of a warrant as the dependent variable. I used the sklearn logistic regression module to evaluate the variables. I followed an example on github from justmarkham to complete the model.

Analysis

The logistic regression model resulted in a very high score: 97.6%. 2.7% of the time the data did resulted in warrants and our data came close to matching that result. This seems to be a great result however it may be that the fact that most debts are eventually paid and not warranted is skewing the prediction.

Looking at the coefficients, we see some interesting indicators – Bank filers were less likely to be warranted, perhaps because they are more likely to have accountants verifying their filings. Receiving an abatement didn't seem to have any predictive effect. Surprisingly, none of the amount categories had a positive effect.

Bank	[0.998463989811]
GCT	[0.52095504491]
UBTI	[-0.594689082661]
UBTP	[-3.63306735308]
2012	[-0.530474551745]
2013	[-0.600821711347]
2014	[-1.57704113781]
MathError	[0.000542667197284]
1	[-0.39057423921]
2	[-1.61025374209]
3	[-0.233855670653]
4	[0.0341123303187]
5	[-0.14256135522]
7	[-0.365204724243]
500KOver	[-0.700732189748]
100to500K	[-0.169127476977]
50to100K	[-0.443928197558]
5to50k	[-0.599344196362]
5kUnder	[-0.79520534038]
LFPen	[-0.207784253128]
UPen	[0.95746641042]
LPPen	[4.2065268322]
LPAbat	[0.0]
LFAbat	[0.0]
UPAbat	[0.0]
Payments	[-1.84363810489]

Conclusion: I didn't find any conclusive characteristics to set a qa process on. My next steps would be to re-evaluate the dataset and findings. Perhaps looking exclusively at receivables that were vacated as well as customer or transaction level characteristics would yield features with stronger predictive effects.

References:

Reviewed models used for risky loan applicants and best uses of random forest method as well as logistic Regression with scikit-learn:

http://blog.yhathq.com/posts/machine-learning-for-predicting-bad-loans.html

https://github.com/justmarkham/gadsdc1/blob/master/logistic_assignment/kevin_logistic_sklearn.ipynb