Closed for Business

Predicting Business Failure Through The Use of Machine Learning and Local Government Administrative Datasets

Callahan Coplai, MSUI '17
Research Practicum | Master of Science in Urban Informatics
Northeastern University

Abstract

Local governments generate an enormous amount of data about properties within their jurisdiction simply through day-to-day administrative activities. For example, code enforcement violations are recorded and available in nearly every local municipality, often dating back many years or even decades. Using a variety of these administrative datasets, we aim to identify a novel approach for predicting business failures. Local administrative datasets from the City of Kalamazoo, Michigan were utilized to test the predictive power of multiple models, including logistic regression, one-class support vector machine, and a cross-validated ensemble of random forests. Different time spans of input data were also tested, using both 1-year and 2-years to evaluate the difference in predictive power. Our models sought to optimize the number of failed businesses correctly predicted rather than pure accuracy. This goal was driven by the potential applicability of such a model to business support and intervention strategies by nonprofits, business consultants, and local governments rather than the traditional application of business failure models in banking and finance. Our initial results show promise in identifying differences between active and failed businesses. In this preliminary result, we found significant relationships between business failure and specific predictors, including property tax delinquency, code enforcement violations, failed inspections, geographic location, and specific types of 911 calls. Finally, we will discuss considerations for improving data collection at the local government level, next steps for future modeling work, and potential applicability of such a model for local government staff and others.

1. Introduction

Predicting business failure through the use of data is no new idea. Banking institutions, creditors, financial firms, and business entities themselves have a significant stake in being able to identify what businesses are on an upward trajectory and which are more likely to fail. This type of predictive analysis has long relied on the use of internal financial data about the performance of a given individual company (e.g., profit and loss ratios, cash flow projections, market viability, competition, etc.). This paper proposes a new approach for identifying what businesses may be at-risk of failure through the use of administrative data gathered and maintained by nearly every local government in the United States. Data regarding code enforcement violations, inspections, permits, property taxes, assessing, ownership information, bill payments, crime, fire and EMS calls, and other data all could be used as "predictors" of businesses that may be struggling. While these predictors don't offer the level of insight into a firms financials that traditional accounting ratios do, they are worth investigating as proxies for identifying a firm's management (or lack thereof). The benefit of these inputs as opposed to accounting ratios is the near universal availability of such data (many small and private companies are not required to report detailed financial information that would be used as inputs to a traditional prediction model).

Thus, by using these administrative datasets in combination with historical data about known businesses that have "failed" between 2014-2017 in the City of Kalamazoo, MI, we will attempt to present a novel method for identifying at-risk companies. Using supervised machine learning techniques, we will determine if the administrative datasets collectively offer predictive value into identifying businesses that are at-risk of failing. This work represents a starting point into investigating non-traditional inputs to this prediction problem. The hope is that this work motivates future work into additional methods, datasets, and potential interventions to support firms that are at-risk.

1.1 Literature on Business Failure Prediction

Substantial research and existing efforts exist to identify businesses that might be at risk of failure. "Over the last 35 years, the topic of business failure prediction has developed into a major research domain within corporate finance. Many academic studies have been dedicated to finding the best corporate failure prediction model" (Balcaen and Ooghe, 2006). Most of the previous work in this area is focused on predicting business failure through the use of company financial information and financial ratios (Sun et al., 2013). This is reasonable for large financial institutions that have regular access to this data through their role as a financier of business ventures. Some of this data is also publicly available for companies that are required to report it (e.g., publicly-traded [large] companies), however, for smaller businesses and for non-financial institutions, conducting any sort of business failure prediction is very difficult. This lack of insight into business performance makes it difficult for local governments or other entities with a stake in supporting the business community to target their resources. As such, this study is unique in the sense it attempts to provide a methodology that relies on readily available administrative data maintained by nearly all local municipalities.

1.1.1 Defining "Failure"

At the base of the business failure problem is the definition of what constitutes a "failure". This definition is contingent upon the data available to identify a failure, as well as the more operational thought around what it really means for a business to have failed.

Financial distress or failure is often defined when there is a clear indicator of difficulty in a business, such as bankruptcy. In addition to this legal bankruptcy definition, when financial data about a firm is available, this allows for the identification of other, less binary definitions of distress, including negative book net assets (i.e., 'accounting bankruptcy'), inability to fulfill a contract to repay principal and interest (i.e., 'technical bankruptcy'), or inability to pay outstanding debt after liquidation (i.e., 'business failure') (Sun 2013).

Non-traditional definitions have also become more common in recent years as the datasets being used to make predictions have changed. Wang et al. (2015) are a prime example of this evolution. They developed a model for predicting "business failure" through the use of Foursquare mobile local-based checkin data for restaurants in New York City. While the dataset utilized was novel in a similar sense of the dataset used for the currently proposed study, their study was limited by their definition of “business failure”, which they chose to be any company that had less than 1 Foursquare checkin per day during their study period.

1.1.2 Data Used

As has been mentioned above, accounting and financial data are the common inputs into traditional business failure modeling efforts. Edward Altman developed the “Altman Z-Score” in his pioneering work in the 1960s, which utilizes five financial ratios; working capital to total assets, retained earnings to total assets, earnings before interest and tax to total assets, market value of equity to total liabilities, and sales to total assets (Altman, 1968). Together, these ratios measure a company’s profitability, leverage, liquidity, activity, and help predict likelihood of insolvency. Many additional financial ratios have been tested to varying degrees of success in the time since Altman’s work (Tinoco & Wilson, 2013).

Outside the realm of single-firm financial ratios, other datasets have been less common to include. TInoco & Wilson tested the utility of a hybrid approach to data inputs, using the traditional accounting information in concert with macroeconomic variables. The reasoning in testing additional data inputs was, “financial statements do not include all the information that is relevant to the prediction of financial distress, and market variables are very likely to complement this deficiency”. They found that the combination of traditional data with the macroeconomic variables improved model performance.

1.1.3 Methods

Just as the input datasets stemmed from Altman’s research, so too did the traditional methodology. Altman utilized z-scores and a mult-variant discriminant analysis approach to conduct his modeling, and William Beaver employed a univariate technique in his 1966 work (Beaver, 1966). These approaches remained almost unchanged for decades, and then there was an explosion of varying statistical modeling techniques and approaches from the 1990s to present day. Machine learning methods, artificial intelligence, hybrid and ensemble approaches, and dynamic modeling and decision implementations have become much more commonplace in recent years (Sun).

The logit model is the most common statistical single classifier method in the business prediction literature, although multiple discriminant analysis (MDA) is also common. Of the vast number of machine learning and artificial intelligence methods, decision trees, neural networks, case-based reasoning, evolution algorithms, support vector machines (SVM), and many others have been tested successfully to predict business failure (Sun). Finally, ensemble methods that combine multiple approaches and exploit each base classifier’s unique information can help reduce error variance. More detail on the logit, SVM, and random forest (combination of decision trees) methods used in this work are provided later.

1.1.4 Difference in Proposed Approach

The present study proposed in this paper differs from traditional and even non-traditional approaches in relation to the data used and, as a result, the definition of failure. The study does not differ on the methods used however, employing three fairly standard predictive modeling techniques that are well-suited to imbalanced classification problems such as business failure prediction.

This study utilizes historical data about businesses that are known failures through the use of a dataset on personal property tax accounts in Kalamazoo, MI. The dataset is separated into active and inactive accounts, and also provides information on delinquent tax payment over the years of 2014-2017. The inactive accounts are either those that are no longer in business or that moved out of Kalamazoo (i.e., as a result, their Kalamazoo personal property tax account is no longer active). Either case is meaningful as an account of “failure” from the City of Kalamazoo’s perspective, keeping businesses in the City and healthy is critical to the overall health and tax base of the City. This difference in definition is directly related to the data available and the perspective of a local government. They have a different consideration of what “failure” is compared to the approach of say a banking institution that wouldn’t care if a business moved locations.

This dataset was joined with a variety of additional local administrative datasets that will be used retroactively to compare historical business failure data to test the predictive power of several variables to determine if we can identify the failed businesses with meaningful accuracy. The general thinking behind such an approach is that there is insight that can be gained in evaluating a business’ data trail of interactions with the local government. For example, a business that has many code enforcement violations may be struggling and have costly fines or repairs, while a business with increasing taxable value might be experiencing success. The approach here was designed to test these possible connections between local government data and business success or failure.

2. Data Preparation

While business success and failure is a common area of concern for municipalities around the world, standard and readily available data on business failures is often not in place. In addition, local governments tend to maintain datasets in several different formats, with varying location identifiers and data input processes. The data used in this analysis from the City of Kalamazoo included the following:

- **Tax Data**: Personal property tax account list with active and inactive accounts and personal property tax delinquency list with information on which accounts were delinquent or late on their tax payments in a given year
- **Building Department Data**: Code enforcement cases, permits issued, and inspections
- **Assessing and Property Characteristic Data**: Property ownership, zoning, assessed and taxable value information
- **Public Safety Data**: 911 calls for Fire, EMS, Police, and non-emergency

Data between the years of 2013 and 2017 from each dataset was gathered and prepared for analysis. A challenge with this phase was the fact that some datasets reference something occurring at a given parcel of land, some at a specific business, some at a given address (several businesses may share an address), and some at latitude / longitude point coordinates (e.g., a crime may have happened on the road outside the business, but was unrelated to the business activities).

The core business failure dataset was the personal property tax account list, which contains data at a individual business level. These businesses are recorded with a personal property identification number (PPIN). Building department and assessing data contain information about “real property”, as opposed to “personal property” from the tax dataset. Each parcel of land within the city has a real property identification number (RPIN) associated with it. Problematically, the PPIN and RPIN numbers do not crosswalk between each other in any meaningful way. As a result, the “Address” field associated with each RPIN and PPIN were used to merge the datasets. This could be considered a fuzzy match for some data, as businesses may share an address. As such, some of the resulting data has limitations in its reliability to be describing the exact business, though there is still potential value in analyzing the condition of other businesses in immediate proximity to the business at hand.

The public safety dataset consists of a number of events (e.g., a fire, crime, emergency response) that occurred at a given point location. These points were spatially joined to the polygon (e.g., parcel of real property) that they fail within. Points not touching or within a polygon were not joined.

The output of this preparatory work to merge various administrative datasets was a near comprehensive understanding of “what occurred” (e.g., a permit was issued, a 911 call, an inspection) at a given location in a given year. Ultimately, connections between datasets were imperfect and not all data from each dataset could be reliably connected to the comprehensive dataset. However, the resulting dataset for prediction contains a meaningful and vast majority of the “activity” from 2013-2017 for each active and inactive business in our list.

2.1 Final Dataset and Predictors

The final dataset allows for binary classification modeling of active and inactive businesses by year within the City of Kalamazoo. In particular, there are 194 identified businesses that became inactive between the years of 2014 and 2016. Of these, 100 were identified as going inactive in 2014 (i.e., their last personal property tax payment was 2014), 52 went inactive in 2015, and 42 went out in 2016. There are an additional 11 businesses that have confirmed their last year of activity will be 2017. Next, there are 2,450 total active businesses, of which 984 we know have been active since 2014 (they made a payment in 2014), an additional 153 made a payment in 2015, and 168 more made their first “known” payment in 2016. This means that 1,305 total businesses have a registered personal property tax payment since 2014, and that 1,145 businesses are classified as “active” but don’t have any payments from 2014 to 2016. These businesses may simply be tax exempt, but because we don’t have a sense for the year they became active, they are excluded from the analysis. This information and the resulting year-by-year prediction datasets are summarized in Figure 1.

	Inactive	Active and First Payment Was This Year or Earlier	Percent Inactive
2014 Failure Prediction	100	984	10.2%
2015 Failure Prediction	52	1,137 (153 net new)	4.6%
2016 Failure Prediction	42	1,305 (168 net new)	3.2%
Total	194	1,305	

Figure 1: Summary of Active and Inactive (Failed) Businesses by Year

The following variables were developed from the initial raw data to be used in the predictive models:

- **Inactive**: Whether or not the business is an active or inactive record
- **Active Year and Inactive Year**: Whether or not the business became active or inactive in 2014, 2015, or 2016
- **DDA and TIF**: A set of binary variables indicating whether a given business is located in the City's Downtown Development Authority (DDA) or Tax Increment Financing (TIF) districts
- **Number of Enforcement Cases**: A set of numeric variables that indicate the number of code enforcement cases at a given location. Separate variables for each year from 2013-2016
- **Number of Permits**: A set of numeric variables that indicate the number of permits issued at a given location (variable for each year from 2013-2016)
- **Amount Billed for Permits**: A set of numeric variables that indicate the total amount of dollars billed for permits issued at a given location (variable for each year from 2013-2016)
- **Total Inspections, Inspections Failed, and Percent of Inspections Failed**: A set of numeric variables that indicate the total number of inspections, number of inspections that required another visit (i.e., failed), and the percent of total inspections that failed (variables for each year from 2013-2016)
- **SEV and TV**: The State Equalized Value (SEV) and Taxable Value (TV) of a given location (variables for each year from 2013-2016)
- **SEV and TV Growth**: Binary variables indicating whether or not the SEV or TV increased from the previous year (variables for each pair of years from 2013-2016. For example, variable name "TVGrowth13_14
- **Owner in City and Ownership Changes**: Binary variables indicating whether or not the owner of a given property is located in Kalamazoo for each year and variables indicating ownership changes from year to year (for each year from 2013-2016)
- **Zoning and Zoning Change**: Variables indicating the zoning district a given business is located within and if the zoning of a particular business changed from year to year (for each year from 2013-2016)
- **Delinquent Personal Property Tax Payments**: Binary variables indicating whether or not a business owner was delinquent on their personal property tax payments in a given year (for each year from 2013-2016)
- **911 Calls at and in Proximity to Business**: Variables that show the number of 911 at a specific location and within 500 feet of a business' property by type, including calls for violent crimes, auto-related crimes, fire-related calls, property related calls, and misc./other calls (for each year from 2013-2016)

Altogether, these variables for a given year were utilized to predict which businesses failed the following year. For example, if a given business had multiple 911 calls, failed inspections, an ownership change, and did not see their property values increase, they may be more likely to fail than a business that did not have such factors. These are the general types of relationships tested through the methodology in the present paper.

2.2 Summary Statistics

Following are a snapshot of summary statistics to compare general characteristics of active and inactive (i.e., 'failed') businesses for each year in the analysis.

Variable	Active Businesses				Failed Businesses			
	2014	2015	2016	Avg.	2014	2015	2016	Avg.
% in the DDA	18.2%	19.0%	18.7%	18.63%	30.0%	21.1%	16.7%	22.6%
% Delinquent on Pers. Prop. Tax	19.6%	19.8%	16.9%	18.77%	42.0%	24.5%	23.3%	29.9%
Avg. # of Code Enf. Cases	0.16	0.12	0.14	0.14	0.24	0.10	0.05	0.13
Avg. # of Permits	1.22	1.0	0.97	1.06	0.56	1.50	0.42	0.83
Avg. % of Inspections Failed	10%	7.8%	6.3%	8.03%	8.6%	8.9%	1.5%	6.3%
Avg. SEV	$358,172	$346,100	$351,664	$351,978	$70,486	$333,356	$150,560	$184,801
Avg. TV	$353,634	$333,356	$339,675	$342,221	$68,805	$328,015	$148,893	$181,904
% of Businesses with Increase in SEV from Previous Year	54.8%	34.0%	31.8%	40.20%	33.6%	26.4%	27.9%	29.3%
% of Businesses with Increase in TV from Previous Year	54.8%	41.6%	40.8%	45.73%	34.6%	33.9%	37.2%	35.2%
% of Owners in Kalamazoo	62.2%	62.0%	64.5%	62.90%	70.3%	62.3%	65.1%	65.9%
% of Businesses with Ownership Changes from Previous Year	8.0%	8.9%	7.4%	8.10%	5.0%	9.4%	7.0%	7.13%
Avg. # of Auto-Related 911 Calls	0.18	1.12	1.08	0.79	0.40	2.11	0.79	1.10
Avg. # of Auto-Related 911 Calls within 500 feet	58.85	43.43	42.54	48.27	56.74	44.38	38.79	46.64
Avg. # of Fire-Related 911 Calls	0.13	0.90	0.87	0.63	0.02	1.67	0.47	0.72
Avg. # of Fire-Related 911 Calls within 500 feet	6.27	4.98	4.58	5.28	5.08	7.15	4.09	5.44
Avg. # of Property-Related 911 Calls	0.46	3.26	3.54	2.42	0.13	7.96	1.79	3.29
Avg. # of Property-Related 911 Calls within 500 feet	36.09	37.69	40.38	38.05	33.16	43.96	43.95	40.36
Avg. # of Violent Crime 911 Calls	0.21	1.62	1.67	1.17	0.07	4.23	0.85	1.72
Avg. # of Violent Crime 911 Calls within 500 feet	20.43	18.96	19.75	19.71	17.20	23.06	18.38	19.55

Figure 2: Summary statistics of predictors of active and failed businesses by year

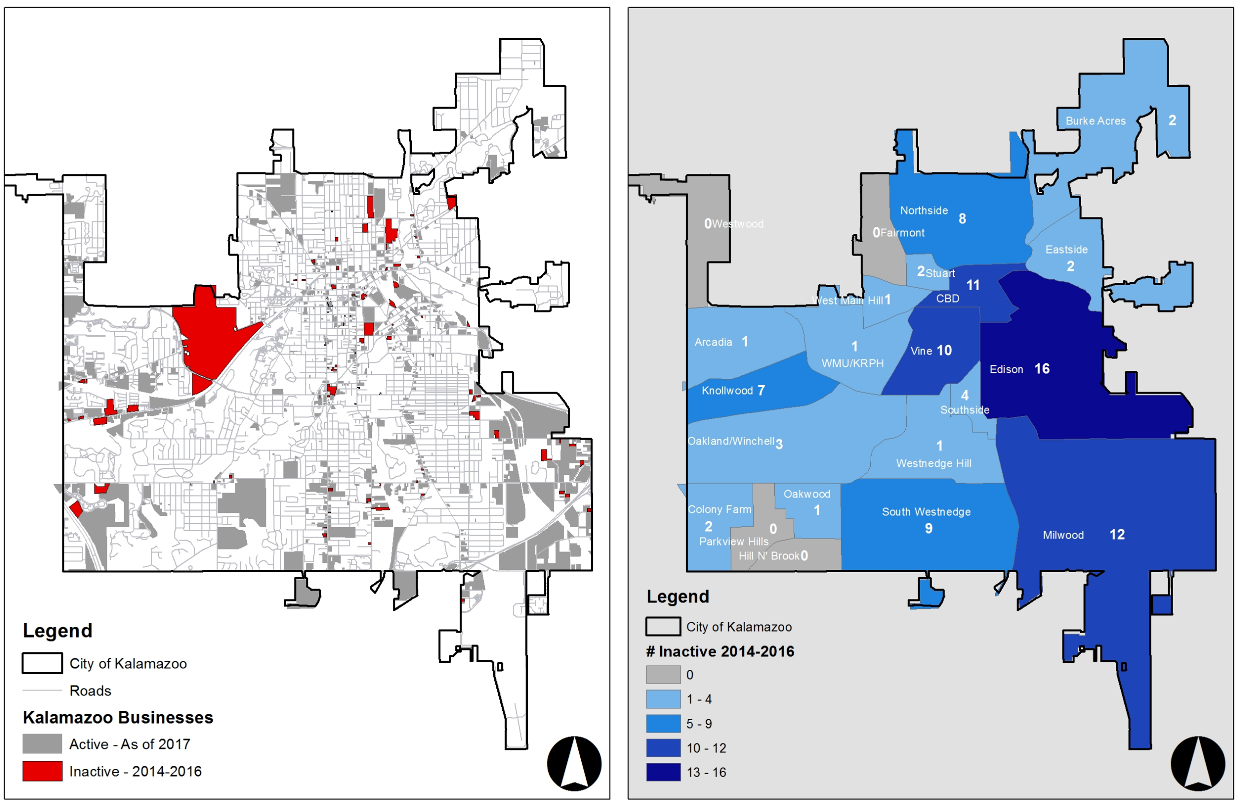
In summary, we found that active businesses and failed businesses differ in respect to a variety of our predictors. In particular, a greater proportion of failed businesses are located in the downtown (DDA), were delinquent on their personal property taxes, had lower SEV and TV, didn't see increases in SEV and TV as often, and had more 911 calls of every type analyzed at their location. Failed businesses also had less permits issued on average compared to active businesses and are slightly more likely to have their owner in Kalamazoo. These relationships show the initial promise in developing a predictive model that can reliably identify the distinction between an active business and one that is no longer active.

2.3 Geospatial Distribution

Only two variables in the dataset utilized in this study had a geographic component, the DDA and TIF variables, which are binary variables that indicate whether or not a given business was located within the City's Downtown Development Authority (DDA) or Tax Increment Financing (TIF) districts. These districts share roughly the same boundary and are representative of "downtown" businesses.

Figures 3 and 4 below show the geospatial distribution of active and inactive businesses in Kalamazoo from the datasets utilized for this research. You can see the clustering of failures by neighborhood, with businesses around the central business district (CBD), Vine, and Edison neighborhood experiencing the greatest number of failures. The areas of the city with the largest number of failures are unsurprisingly the areas of the city with large business districts, core commercial corridors, or large industrial complexes (e.g., Milwood). While not implemented in this study, future work evaluating the impact of location on business failure may be able to identify relationships between success or failure at different spaces in the urban environment.

Figures 3, 4: Map of individual business locations (left) and distribution showing number of failures by neighborhood



3. Methodology

In addition to the summary statistics that give a sense of the relationship between the independent variables and business failure, a number of specific statistical models were developed to test the predictive power of this data. In each case, three iterations of each model were run, one using 2013 data to predict failures in 2014, one using 2014 data to predict failures in 2015, and one using 2015 data to predict failures in 2016. In addition, the models were tested to see the impact of using two years of input data rather than one (e.g., using 2012 and 2013 data to predict failures in 2014). The three models used were logistic regression (or a ‘logit’ model), a one-class support vector machine (SVM) model, and cross-validated random forest models. These models were all chosen for their strengths in classification modeling, and in particular in relation to binary classification. In addition, the unique structure of the dataset, being substantially imbalanced (i.e., significantly more active businesses than inactive each year) necessitated the use of these machine learning techniques that are best suited to handle this kind of problem. An overview of each modeling technique used is provided here.

3.1 Logit Model

The logit model is commonly used across many disciplines when the core task at hand is to predict a binary outcome variable based on a set of predictors. It has been used regularly as a technique for business failure prediction and modeling (Altman & Saunders 1998, Vuran 2009, Sun 2013, Wang et al. 2015). The outcome of a logit model provides a probability estimate for how likely a record (in our case a business) is to be classified as either a ‘1’ or a ‘0’. At this point, cutoff values are chosen to transform these probabilities into binary classifications. Although, the probabilities themselves are useful to get a sense of what businesses the model has classified most strongly as likely to fail or most likely to stay active. For the sake of prioritizing outreach efforts and intervention strategies, the probabilities themselves may be more useful than the 1 or 0 classifications, but for the sake of this ex ante modeling, the binary classifications are helpful to understand the efficacy of the method.

Our first iteration will utilize a probability estimate of 0.5 as the cutoff to classify a business as a likely Inactive, though we’ll use a cutoff optimization technique to test other cutoffs as well to see if we can improve performance. We utilized two of these “optimized” cutoff points for each iteration of the logit model, one that optimizes both the number of actives and inactives correctly predicted, and one that optimizes the sensitivity (i.e., recall) only. We’ll utilize the specificity, sensitivity, precision, accuracy, and F1-Score to evaluate performance of our logit models for each year (Alice 2015).

3.2 One-Class Support Vector Machine

In machine learning, support vector machines (SVMs) are what are known as supervised learning models. Supervised models can be used when there is some data that is already classified, and then we use that data to train to model to make a prediction on unclassified data. Unlike a logit model, the resulting classification is not probabilistic, and rather provides a simple binary classification output (Cortes and Vapnik 1995). The benefit of an SVM approach is it is not likely to result in over-fitting even for small samples. It has been used previously to predict business failure and bankruptcy and showed promising results, outperforming logit and neural network approaches (Sun, 2013).

One-Class SVMs are a particular version of an SVM designed for anomaly detection. This is useful in situations where “you have a lot of ‘normal’ data and not many cases of anomalies you are trying to detect” (Microsoft 2017). In a One-Class SVM, the model is trained using only one of the classes of data (often the ‘normal’ class) and then when the abnormal class is fed in, the model attempts to detect these outliers because they should be dissimilar to the normal class.

3.3 Random Forest with Cross-Validation

The final modeling approach employed was also the most complex. We employed and compared five variants of a 10 x 10 repeated cross-validated Random Forest model using a standard approach, under-sampling, over-sampling, Random Over-Sampling Examples (ROSE), and Synthetic Minority Over-Sampling Technique (SMOTE). A Random Forest operates as an ensemble learning method for classification that uses multiple decision trees to train the data and outputs the mode of the class of the trees in aggregate (hence the name 'forest').

These variations of the standard Random Forest are designed to help deal with problems of imbalanced data. With under-sampling, a random subset of samples from the 'normal' class are chosen to match the number of abnormal cases. Here, we risk losing valuable information from the active businesses in our dataset. With over-sampling, we randomly duplicate known failed businesses so that the number of failures matches the number of active businesses. The risk with this method is that of overfitting the model as a result of such duplications.

To address the problems with both regular over- and under- sampling, we also tested some hybrid techniques. ROSE and SMOTE are two of the most common hybrid techniques. ROSE generates artificial balanced samples according to a smoothed bootstrap approach that aids estimation and accuracy evaluation of imbalanced classification problems (Lunardon et al. 2014). SMOTE is a hybrid technique that combines synthetic over-sampling of the abnormal class and under-sampling the normal class to achieve better classification performance (Chawla et al. 2002).

3.4 Evaluation

To evaluate the success of the various modeling techniques employed, a suite of evaluation metrics were used. The performance of each of these techniques was evaluated through the use of Sensitivity, Specificity, Precision, Accuracy, and the F1-Score (Glander 2017). These metrics can be explained by the following:

- **Sensitivity** is the proportion of true positives that are correctly identified as such (e.g., a failed business is correctly predicted as failed)
- **Specificity** is the proportion of true negatives that are correctly identified as such (e.g., an active business is correctly predicted as active)
- **Precision** is the number of positive predictions divided by the total number of positives predicted (e.g., of all the predicted inactives, how many were actually inactive?). Low precision can indicate a large number of false positives, which would be active businesses predicted as inactive.
- **Accuracy** is the total share of correct predictions (e.g., how many active and inactive businesses were correctly predicted as such?)
- **F1-Score** measures the balance between precision and sensitivity

(Altman & Bland1994, Brownlee 2014)

While all five of the above metrics were calculated for all of the models, the F1-Score (and thus precision and sensitivity) was prioritized when decisions of model optimization had to be made. This is because sensitivity allows us to be sure that the largest share of inactive businesses were correctly predicted as inactive as possible. Optimizing precision ensures that we are limiting false positives and prioritizing true positives (e.g., failed businesses accurately predicted as failed).

Finally, to understand what variables are most important to the “decisions” being made by each of the modeling techniques, we gathered the variable importance metric for each model and each year. The variable importance allows us to evaluate if there is in fact consistency in the predictions across modeling techniques and year-to-year.

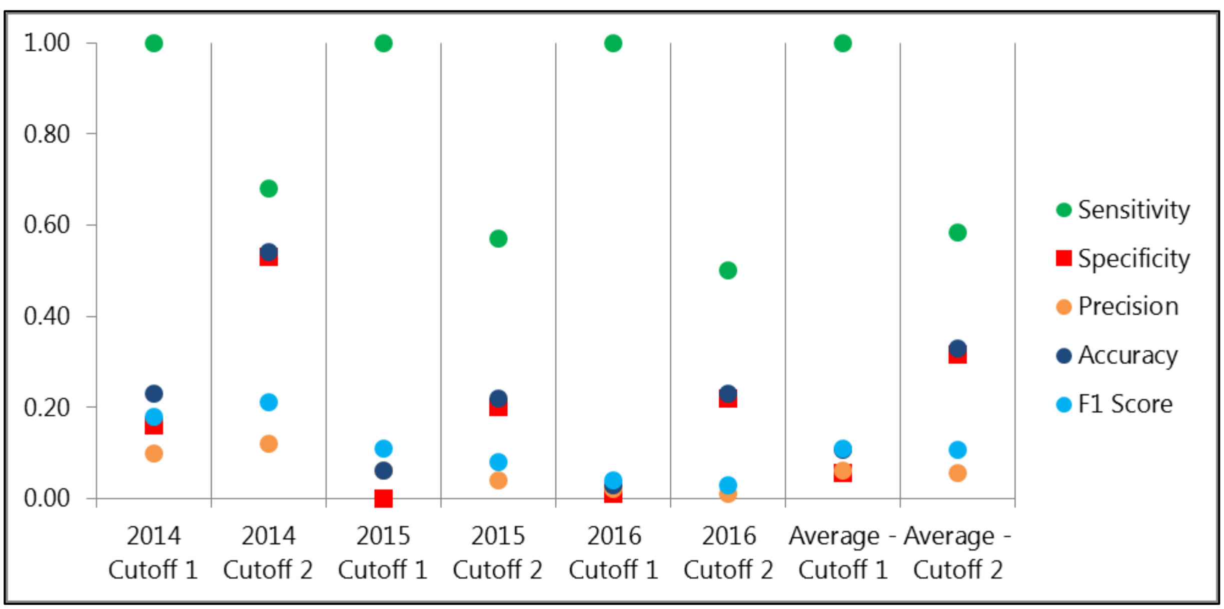
4. Results

Each modeling technique was employed to predict business failures in 2014, 2015, and 2016, once using one year of predictor data and once using two years. An average of our five evaluation metrics was calculated to determine the performance of each model over the course of the three years (e.g., accuracy of 0.7 in 2014, 0.9 in 2015, and 0.8 in 2016 is a 0.8 average performance). As mentioned, in evaluating the models, the optimization of F1-Score is top priority. A brief overview of performance of each modeling technique is provided below, followed by an overall comparison of each of the three techniques.

4.1 Logit Model Performance

As mentioned previously, two optimized cutoff points were utilized in our logit model predictions for each year of the model. The performance of these variants are provided in Figure 5 below.

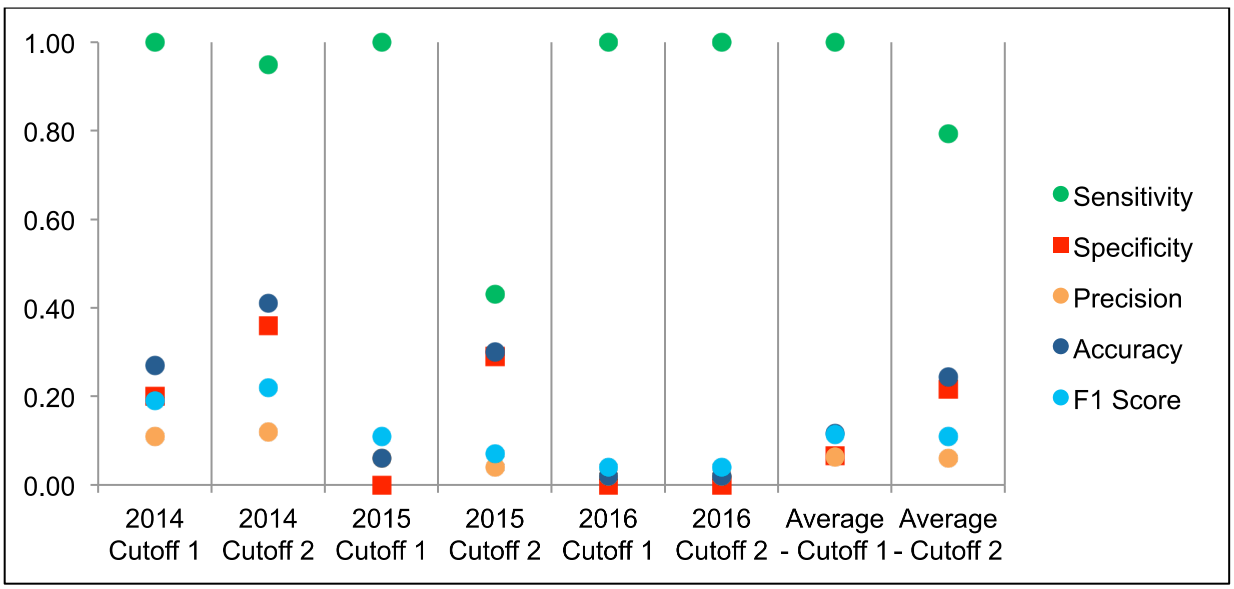
Figure 5: Performance of logit model variations by year, one year of predictor data



As you can see, the first threshold cutoff point optimized the sensitivity of the model, and was in a sense flawed by attempting to correctly classify every single inactive business. The second cutoff that optimized both correct prediction of actives and inactives overall provides some promising results. This model, on average over the course of the three years of predictions had a sensitivity of 0.58, specificity of 0.32, precision of 0.06, overall accuracy of 0.33, and F1-Score of 0.11. Finally, the various logit models found statistically significant relationships between failed businesses and delinquency on personal property taxes, increased code enforcement cases, failed inspections, property-related 911 calls, lack of SEV growth, and violent 911 calls.

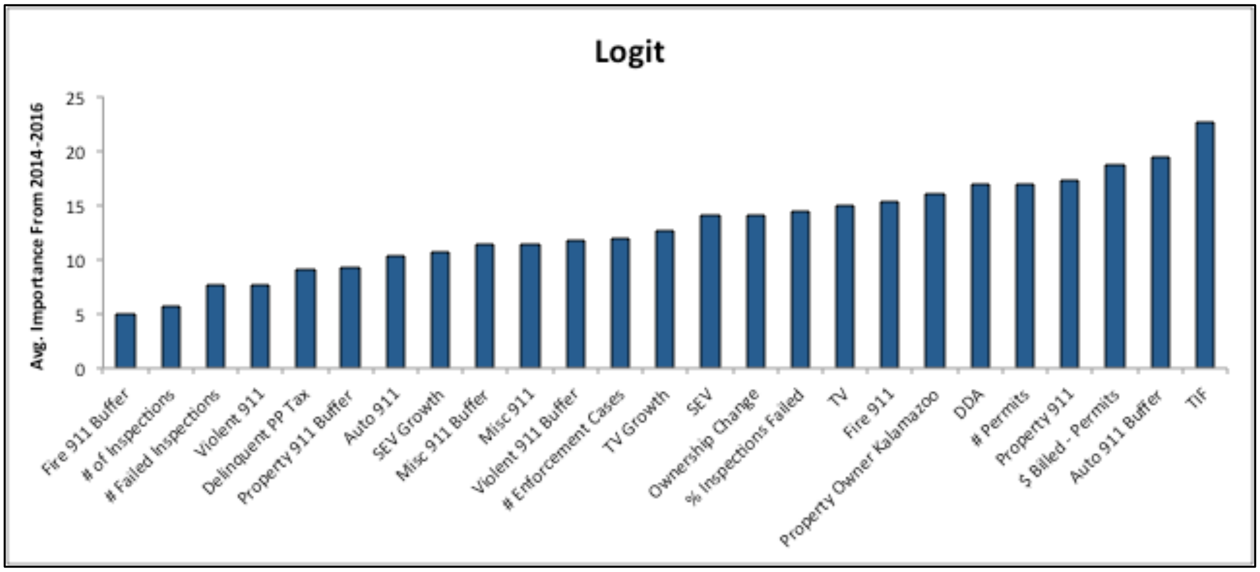
Using two years of input data instead of one, we see the second cutoff method having more success than the first cutoff method. The resulting metrics for this approach with additional input data include sensitivity of 0.79, specificity of 0.22, precision of 0.06, accuracy of 0.24, and an F1-Score of 0.11. Figure 6 illustrates these results.

Figure 6: Performance of logit model variations by year, two years of predictor data



The variable importance from the predictions for each year were averaged to get a sense of overall importance across the 3 years of prediction. What we found was that fire 911 calls in the immediate vicinity of the business (500' buffer), number of inspections and inspections failed, violent 911 calls, and delinquency on personal property taxes were the most important variables. The average ranking of each input variable can be seen in Figure 7.

Figure 7: Average variable importance, logit model



4.2 One-Class SVM Model Performance

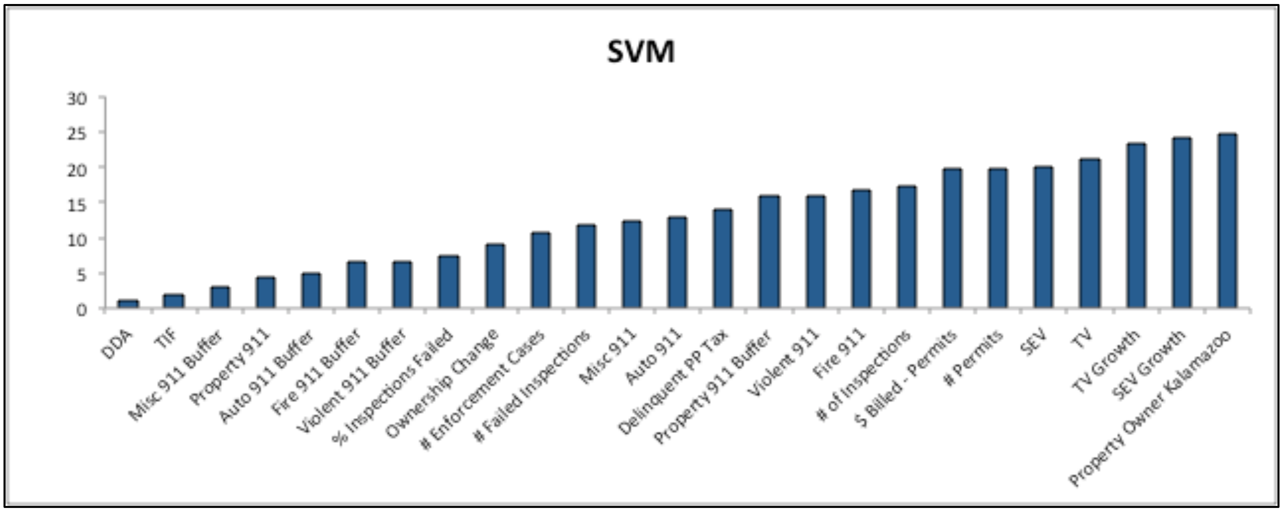
The One-Class SVM Model relied on training the model on only active businesses and then feeding the entire data set into the model for anomaly detection of inactives. The results were again moderately promising (see Figure 8).

On average, when using one year of input data, the model maintained a sensitivity of 0.45, specificity of 0.50, precision of 0.05, overall accuracy of 0.50, and F1-Score of 0.09.

Figure 8: One-class SVM performance by year, one year of input data

When using two years of input data, the model performed exactly the same in regards to the key metrics. This finding was interesting and indicates that perhaps the most recent year of data was providing the vast majority of the predictive power. The most important variables for the SVM model were whether the business was in the DDA or TIF district, and misc., auto-related, fire, and violent 911 calls 500' around the business (Figure 9).

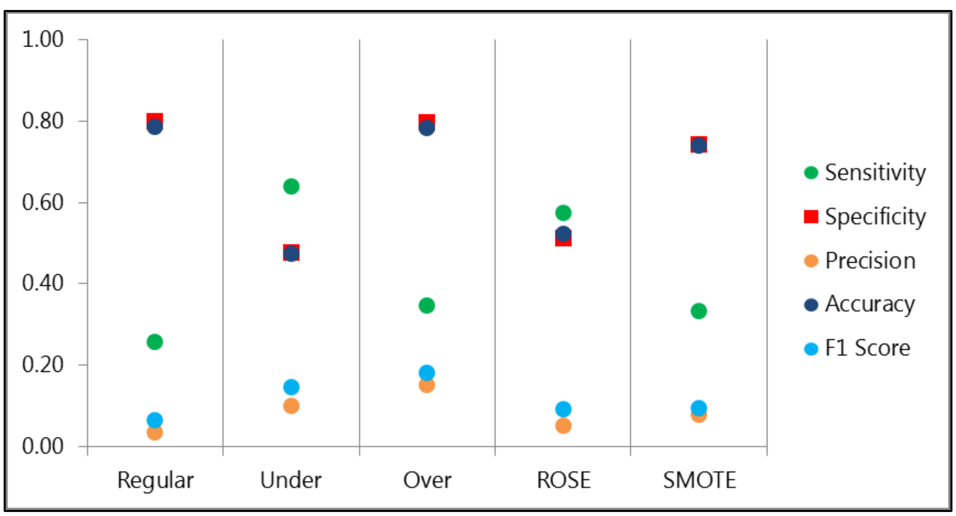
Figure 9: Average variable importance, SVM model



4.3 Random Forest Model Performance

The Random Forest technique relied on a 10 by 10 cross-validation and five different variants for prediction within each year. We utilized a standard approach, under-sampling, over-sampling, ROSE, and SMOTE techniques and compared each using sensitivity, specificity, precision, accuracy, and F1-Score. The average results of each model are included in Figures 10 and 11. What we found was that overall the under-sampled and over-sampled random forest models show the most promise for this particular classification task. The under-sampled model has the highest sensitivity on average, while the over-sampled model has the highest precision and F1-Score of any of the models tested.

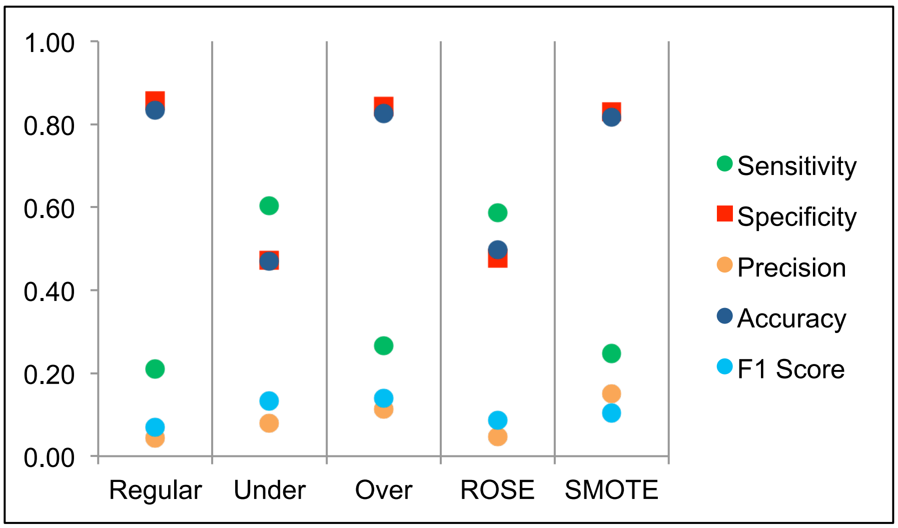
Figure 10: Graph of average performance of the random forest model variants, one year of input data



Measure	Average Performance -- Random Forest Models				
	Regular	Under	Over	ROSE	SMOTE
Sensitivity	0.26	0.64	0.35	0.57	0.33
Specificity	0.80	0.48	0.80	0.51	0.74
Precision	0.04	0.10	0.15	0.05	0.08
Accuracy	0.79	0.47	0.78	0.52	0.74
F1-Score	0.07	0.15	0.18	0.09	0.10

Figure 11: Table of average performance of the random forest model variants, one year of input data

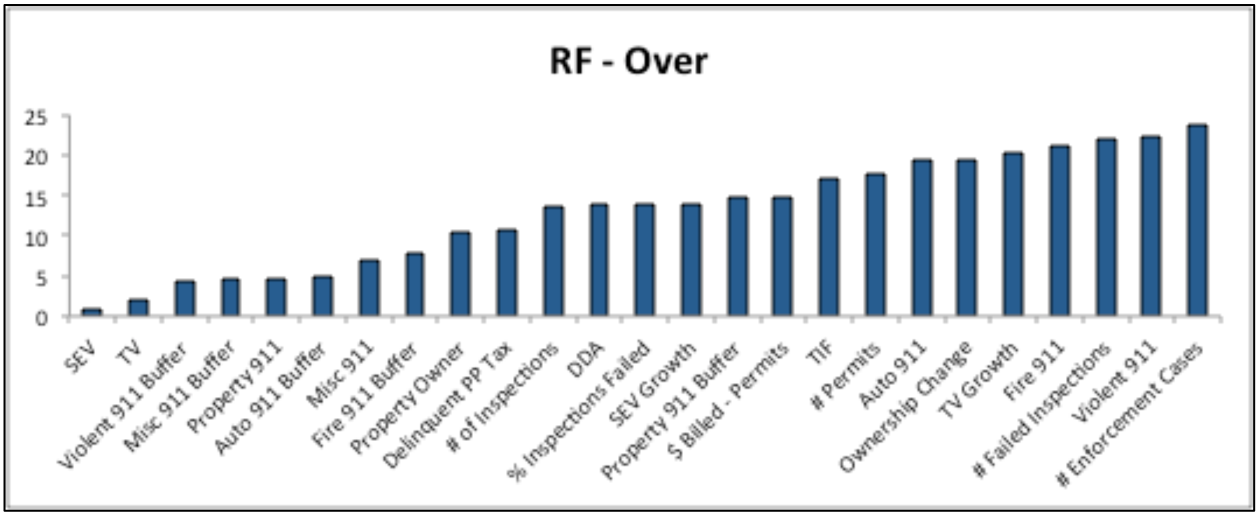
Using two years of input data, the models actually perform slightly worse than using only one year of input data. However, the under- and over-sampled models still perform the best, adding confidence to those two techniques being most appropriate for this task. Figures 12 and 13 show the results.

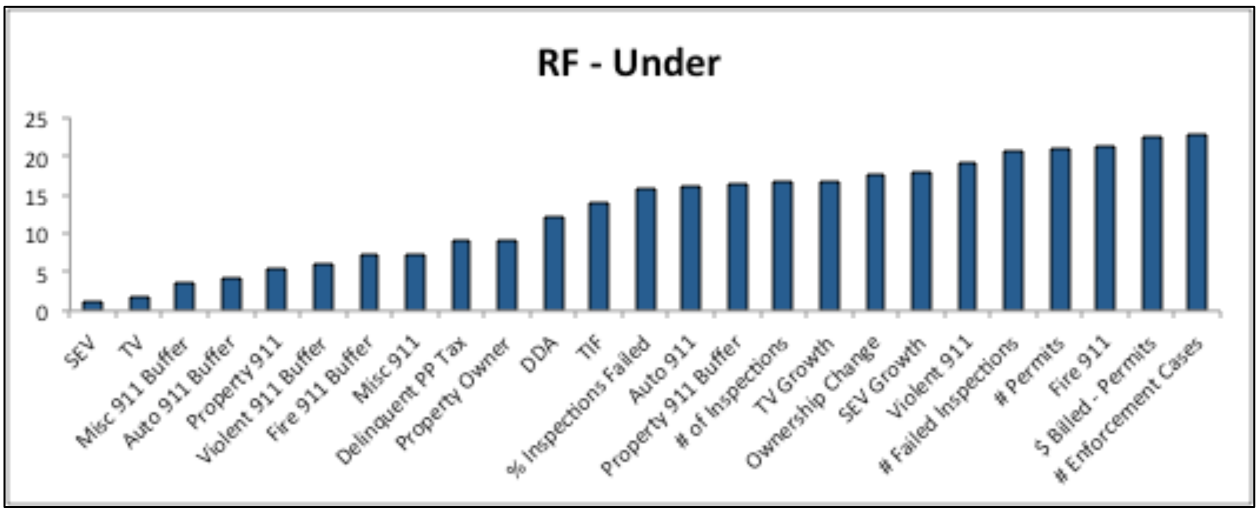


Measure	Average Performance -- Random Forest Models				
	Regular	Under	Over	ROSE	SMOTE
Sensitivity	0.21	0.60	0.27	0.59	0.25
Specificity	0.86	0.47	0.84	0.48	0.83
Precision	0.04	0.08	0.11	0.05	0.15
Accuracy	0.83	0.47	0.83	0.50	0.82
F1-Score	0.07	0.13	0.14	0.09	0.10

Figures 12, 13: Graph (top) and table (bottom) of average performance of the random forest model variants, two years of input data

These two best performing techniques have similar results when it comes to variable importance (Figures 14 and 15). We see that the state equalized value (SEV) and taxable value (TV) of the real property the business is located on being the top two predictors, and a variety of 911 call related variables following that.





Figures 14, 15: Average variable importance across three years of models, random forest over-sampled technique (top) and under-sampled technique (bottom)

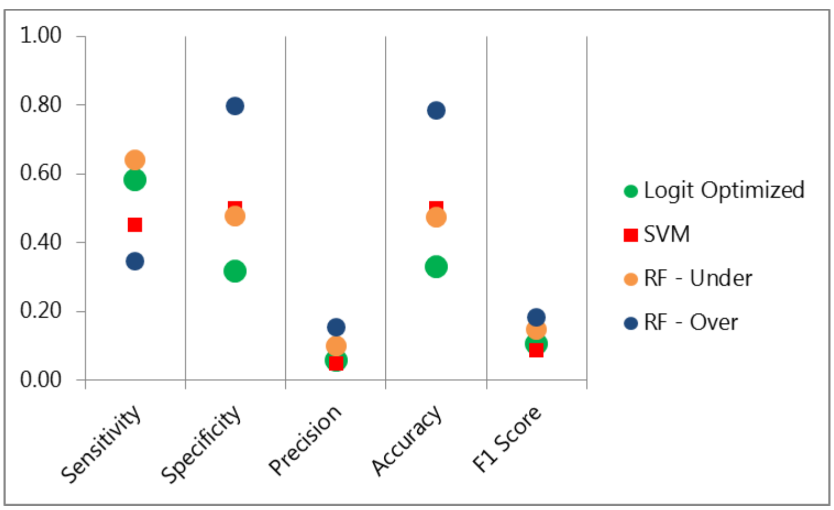
4.4 Summary and Comparison of Models

First, a comparison of the two different time spans of predictor data was completed. In this case, it appears that using two years of data did not improve upon the predictive power of the modeling techniques compared to simply using one year of data. This is a bit surprising but not necessarily a major finding, as research has shown that using one to four years of predictor data can be reasonable and the differences in performance here were not substantial (Camara-Turull et al., 2015). Figure 16 shows the difference in performance from using one year of data to two years of data (negative number means the model performed worse with two years of data).

Measure	Average Performance -- Random Forest Models			
	Logit Optimized	SVM	RF - Under	RF - Over
Sensitivity	0.21	0.00	-0.04	-0.08
Specificity	-0.10	0.00	0.00	0.05
Precision	0.00	0.00	-0.02	-0.04
Accuracy	-0.09	0.00	0.00	0.04
F1-Score	0.00	0.00	-0.01	-0.04

Figure 16: Comparison of performance using one year of data and two years

Figures 17 and 18 show the average performance of the most promising models tested when using one year of input data; the logit model using a threshold to optimize both active and inactive classification, the SVM one-class model, and the random forest under- and over-sampled models. Each of these four models has various strengths depending on the desired outcome of the prediction. For example, the optimized logit model offers a relatively high sensitivity but the lowest overall accuracy of any of these models while the over-sampled random forest model offers the highest specificity, precision, accuracy, and F1-Score. The over-sampled model's only weakness is its low sensitivity, which was a high priority performance measure for this task, as we wanted to primarily focus on classifying a high share of the inactives correctly.



Measure	Average Performance – Best Models			
	Logit Optimized	SVM	RF - Under	RF - Over
Sensitivity	0.58	0.45	0.64	0.35
Specificity	0.32	0.50	0.48	0.80
Precision	0.06	0.05	0.10	0.15
Accuracy	0.33	0.50	0.47	0.78
F1-Score	0.11	0.09	0.15	0.18

Figures 17 and 18: Graph (top) and table (bottom) of average performance of the best performing models from each modeling technique

5. Discussion & Future Work

The desired performance of our models is not necessarily to have a high accuracy, as this could be an indication of simply classifying all active companies correctly and misclassifying the inactives. For the sake of this approach, we wanted to optimize classifying inactives as inactive, which is where the real value potentially lies when it comes to conducting pre-emptive outreach to potentially help a struggling business. Thus, we are interested in models that classify the most inactives correctly, even if it provides a large amount of actual actives incorrectly classified as inactive. From an intervention standpoint, identifying potential failures and being able to outline a list of businesses that have the characteristics of failures will be extremely useful.

Using this model in the future to predict businesses that 'may fail' would allow for an immediate prioritization of resources to the companies identified as potential inactives by the model. For this use case, four specific variations show promise, the optimized logit model, the one-class SVM, and the over- and under-sampled Random Forest models. All would allow for meaningful predictions of potential "failures" into the future. Such a list of potential failures could be used by local governments, banking institutions, nonprofits, and business consultancies or support organizations to either prioritize businesses for support services and training, or to know which businesses to be wary of when it comes to investment. For investment prediction purposes, this model is not as reliable as models that have been in use for the past decades that rely on internal financial information, however, the novel use of the datasets presented in this paper show some promise as additional factors to consider in predicting business failure. Perhaps existing, traditional models can be strengthened by integrating components of the approach shown here.

Future work on this effort to predict business failures using local government administrative data could potentially be improved in two ways; new or modified statistical modeling approaches and the inclusion of additional predictors and improved data. The field of machine learning and predictive modeling offers a vast array of tools and techniques that are constantly evolving. This paper evaluated the use of three such techniques and attempted to optimize them for the prediction case at hand, however we recognize there are many additional methods to test. One technique that has been used with success in recent years is the use of ensemble modeling methods that combine the predictions of multiple models into a single prediction. This could be beneficial in this case, as the variables were given different levels of importance depending on what model we were using.

Next, we recognize that the datasets utilized are not all-inclusive when it comes to potential publicly-available information about companies to include. We utilized datasets on a number of features including code enforcement, permitting, assessing and ownership, tax payments, and 911 calls. However, additional datasets from online review/check-in websites such as Yelp, YellowPages, or FourSquare could also be included. There are plenty of additional sources of information about companies both open and closed that ought to be investigated for their predictive power.

Finally, a number of data preparation challenges were encountered during the completion of this project due to the messy and incompatible nature of datasets within even a single local government. These data consistency and compatibility issues present both a challenge but also a unique opportunity for enhancement that could result in improved ease of use for future work. Local governments sit on a mountain of data that can provide great insight, only if it is structured in a format suitable for efficient analysis. Anytime there is data with a spatial component, there should be processes and structures in place to allow for crosswalking of data to understand relationships between datasets.

References

Alice, Michy (2015). How to Perform a Logistic Regression in R. DataScience+. https://datascienceplus.com/perform-logistic-regression-in-r/

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance 23(4). 589-609.

Altman, D. G. & Bland, J. M. (1994). Diagnostic tests 1: Sensitivity and specificity. The British Medical Journal 309 (248).

Altman, E. I., & Saunders, A. (1998). Credit risk measurement: Developments over the last 20 years. Journal of Banking and Finance, 21(11–12), 1721–1742.

Balcaen, Sofie and Hubert Ooghe. 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. The British Accounting Review 38 (2006) 63-93.

Bates, Timothy. Analysis of young, small firms that have closed: delineating successful from unsuccessful closures. Journal of Business Venturing 20(3) (2005) 343-358.

Beaver, W. H. (1966). Financial ratios as predictors of failure. Journal of Accounting Research 4(1966). 71-111.

Brownlee, J. (2014). Classification accuracy is not enough: more performance measures you can use. Machine Learning Mastery. www.machinelearningmastery.com

Camara-Turull, X., Izquierdo, F., & Forradellas., S. (2015). How do different time spans affect the prediction accuracy of business failure? Fuzzy Economic Review 1 (2015) 71-89.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 16, 321–357.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. Machine Learning. 20(3), 273–297. doi:10.1007/BF00994018.

Glander, Shirin (2017). Dealing with unbalanced data in machine learning. Shirin's PlaygRound. https://shiring.github.io/machine_learning/2017/04/02/unbalanced

Halabi, Claudia and Robert Lussier. A Model for Predicting Small Firm Performance: Increasing the Probability of Entrepreneurial Success.

Halabi, Claudia and Robert Lussier. A Three-Country Comparison of the Business Success versus Failure Prediction Model. Journal of Small Business Management 48(3) (2010) 360-377.

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. R Journal. 6(1) 79.

Microsoft Azure (2017). One-Class Support Vector Machine. Microsoft. https://msdn.microsoft.com/en-us/library/azure/dn913103.aspx

Narayanan, Paul. Business Failure Classification Models: An International Survey. http://www.iiiglobal.org/sites/default/files/44businessfailuremodels_0.pdf

Parsa, H. G. et al. Why Restaurants Fail. Cornell Hospitality Quarterly 46(3) (2005) 304-322.

Stokes, David and Robert Blackburn. Learning the hard way: the lessons of owner-managers who have closed their businesses. Journal of Small Business and Enterprise Development 9(1) (2002) 17-27.

Sun, J., Huang, Q. H., & He, K. Y. (2013). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. Journal of Knowledge-Based Systems 57 (2014) 41-56.

Tinoco, M. H. and Nick Wilson (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market, and macroeconomic variables. International Review of Financial Analysis 30 (2013) 394-419.

Vuran, Bengu (2009). Prediction of business failure: a comparison of discriminant and logistic regression analyses. Istanbul University Journal of the School of Business Administration, 38(1) 47-65. http://dergipark.ulakbim.gov.tr/iuisletme/article/viewFile/1023013870/1023013084

Wang, Lu and Chong Wu. Business failure prediction based on two-stage selective ensemble with manifold learning algorithm and kernel-based fuzzy self-organizing map. Knowledge-Based Systems 121 (2017) 99-110.

Wang, Lei et al. On the brink: Predicting business failure with mobile location-based checkins. Decision Support Systems 76 (2015) 3-13.

Wetter, Erik and Karl Wennberg. Improving Business Failure Prediction for New Firms: Benchmarking Financial Models with Human and Social Capital. Journal of Private Equity 12(2) (2009) 30-37.

Williams, Densil. Resources and Business Failure in SMEs: Does Size Matter? Journal of Business Management; Fort Collins 20.2 (2014) 89-102.