Statistical Learning to Predict Firms Extortion

Pedro Armengol

June 2018

Abstract

Just as persons, firms are more or less prone to be victimized depending on their behavior, location or attributes. For this project, we merge victimization surveys with firms census, in the context of Mexico, to predict extortion through applying cutting-edge statistical learning techniques. Our best model predicts extortion with XX percent of accuracy (XX percent points better than the baseline estimations). An extrapolation of the results was done to obtain incidence at firm level for the whole universe. The spatial correlation between our predictions and other sources of data is positive and strong. Further research describes next steps to explore the role of extortion as a structural cause of firm attrition.

Introduction

A firm, like a person, can be victim of several types of crimes. Particularly, a firm can be more or less prone to be victimized depending on its characteristics, such as where the firm is located, what it produces, how much it sells, how old it is, profitability margin, and so on. All these characteristics together make each firm a unique profit generating entity that represents an opportunity for rent extraction (Burrows & Hopkins, 2005; Farrell, Phillips & Pease, 1995; Gill, 1998; Hopkins & Tilley, 2001; Mugelini, 2013a; Perrone, 2000; Sparks, 1981).

Extracting rents through exploiting a competitive advantage in violence, say through threats, is called extortion; it is a similar concept to black mailing but involves the use of violence in order to obtain the rents from firms or persons.

In order to participate in extortion industry, rent-seekers/ criminals make a trade-off assesment between profitability and risk before demanding a ransom: how dangerous is it to proceed vs. how much resources can they obtain from proceeding. According to Becker (1968), the decision-maker trade-off for extorting a firm can be expressed as

$$\Pi_{h,i} = E(Profitability|Extortion)_i - E(Punishment)_j * Pr(Caught)_{h,j}$$

Such that the expected profits for criminal or criminal group h for extortion a firm i is $\Pi_{h,i}$ and is equal to the expected profit of extorting firm i minus the expected punishment in region j multiply by the probability of criminal h to be caught in region j.

In some countries the trade-off turns out to be alarmingly positive, where the victimization rate can be as high as 6.6 percent of the firms (INEGI, 2016). It is expectable to have a particularly large under-reported number of this type of crime given the modus operandi of extortioners, which is to threaten the victims into silence. Likely, extortion could impose a psychological stress burden to the firms' management and employees as well as increase the financial and/or operations costs, such that the capacity of a firm to sustain the same levels of competence could be dramatically limited by this type of victimization.

This study uses statistical learning to model $E(Profitability|Extortion)$ and $Pr(Caught)$, such that, through the use of victimization survey data merged with firms census data, in the context of Mexico, we are able to predict for which firms i the likelihood of being extorted is greater. Where $E(Profitability|Extortion)$ is modeled using the characteristics of the victimized firm and $Pr(Caught)$ is modeled with its location, variations in the geographical location of the firms can be associated with different rules of law, cultural acceptance of extortion and presence of more or less sophisticated organized crime groups (the natural candidates to commit extortion).

An extrapolation of the results is done for the whole country, Mexico, using the weights of the best model estimated in the victimization survey sample. Lastly, short-term applications and further extensions of the research are discussed.

The remaining sections of the paper are the following ones: Data explains the type of information that this research uses as well as descriptive statistics of the predicted variable. Empirical strategy and Results discuss the prediction techniques, the structure of the results and accuracy of the predictions. Policy applications discusses the short-term uses of this research. Further research describes next steps to explore the role of extortion as a structural cause of firm attrition and, lastly, is the conclusion.

Data

This project will use the economic census and victimization surveys (both described below) to fit the likelihood that a firm will be victimized based on its characteristics and location. The victimization surveys will tell us if a firm was victimized for a sample of around 33 thousand firms that are representative of the firm's universe (around 4.5 million), such that, for a given sample, we have information about victimization prevalence and incidence. On the other side, for the universe, we have information on size, age, incomes, expenses, profits and so on, where the universe of the census includes the sample of the victimization survey.

The available data about the problem allows a "state of the art" statistical learning solution, where there is a labeled column of victimization (including victimization by extortion) and more than one hundred features of the firm recorded on the census (both sources of information are merged through a firm ID).

Victimization Survey (firms)

The firms victimization survey (ENVE for its acronym in Spanish) is a tool developed by INEGI[1] to understand the impact of the crime over the private sector (firms). The survey covers the owners' insecurity perception, the most common crimes that the firms face, the crimes characteristics as well as the related costs related. This survey has validity for all the economic units (every physical place where an economic activity is realized). The sampling was done with two criteria: size of the firm and State. These are the two representative strata[2]. The sample is composed of 33,866 firms (from a universe of 5.6 million: taking as sampling frame the economic census of 2014). There are victimization surveys for the years 2012, 2014 and 2016. A new victimization survey is coming in 2018.

The table has 156 variables related to the following categories:

- Level of victimization
- Characteristics of the crime
- Report of the crime
- Perception of insecurity
- Institutional performance
- Cost of crimes
- Cost of informal commerce and vandalism

Economic Census

Every 5 years, INEGI performs a census about economic activity at the firm unit level. The outcome of the census is a table with a great level of detail on the geographical position, type of economic activity and size of the firm. The census uses the North American Industry Classification System (NAICS) for the type of economic activity label.

The table has 96 variables related to the following categories:

[1] Statistics Ministry of Mexico.

[2] Taking in consideration the following parameters: confidence level 95%, relative error 13% and non-response rate 15%.

- Geo-location
- Type of Economic Activity (NAICS)
- Accounts (Total production, Intermediate consumption, Aggregate value, Capital formation, etc.)
- Number of employees and worked hours by gender and type (owner, direct employee, employee without constant wage, etc.)
- Wages
- Expenditures (inputs, gasoline, light, water, etc.)
- Income (goods production, goods reselling, services, renting property, etc.)
- Inventory: finished goods, goods in process, assets by type:estate, furniture, transport, etc.)

Descriptive Analysis

From the victimization survey, variables related to extortion were selected. The variables are described **Table 1**.

```
# Table 1
## Description of Dependent Variables
### columns: name, description
### row: variable
```

In **Table 2** describes basic statistics of the **Table 1** variables, where the prevalence (mean) is reported for dichotomous variables (firm was victimized or not) and the mean and standard deviation are reported for the continuous variables (how many times the firm was victimized: incidence). The number of missing values in the sample is reported as well.

Particularly, the previous estimations are important because the prevalence will be our baseline metric to compare our prediction (if we predicted only zeros in our data-set what would be the level of accuracy?).

```
# Table 2
## Summary Statistics of Dependent Variables
### Columns: name, mean, Standard Deviation, NA's
### Row: variable (by dicotomous or continues)
```

In **Table 3** (Annex), the mean, standard deviation and number of missing values is presented for each variable of the census: the firms characteristics information.

Empirical Strategy and results

In this section we describe the modelling of the victimization likelihood of a firm. Also we describe the results from estimating seven statistical learning techniques. Finally, we explain how the results/predictions where extrapolated into the whole universe given the available census data.

Modeling

First, we delimited the data-set to include only those observations with non-missing values on the dependent variable. Later on, we divided the data into training and testing set, considering a random shuffle partition of 80 percent of the data.

A prepossessing step was done before the statistical learning estimations: the data-set of predictive features was centered (mean to zero) and scaled (divided over standard deviation) as the literature suggests to control

over unnecessary noise. Also, K-means clustering extrapolation technique was applied to impute information on the missing values (again just over the predictors) in order to archive a balanced data-set.

Subsequently, the model in **Table 4** was estimated.

```
# Table 4
## Estimated Models
### Columns: name, type (Classifier or continuous)
### Row: model
#### Classifiers
##### Regularized Logistic Regression
##### Least Squares Support Vector Machine with Polynomial Kernel
##### Gradient Boosting Machines
##### Random Forest
##### CART
#### Continuous
##### Linear regression
```

Particularly, the linear regression was estimated for the continuous dependent variables (incidence). The other models were estimated for the dichotomous dependent variables (prevalence). For each model, a 25 repetition Bootstrapped cross-validation without repetition (within the 80 percent of the training data) was done. For those 25 random shuffling, all the parameters of the model's standard grid were tested[3]. Then, the average accuracy, or alike metric depending of the model, was obtained. For every model, the parameters set with the higher average accuracy over the 25 partitions were selected.

The models were selected under the criteria of having a plethora of classifiers, each one representing a different estimation technique: Regularized Logistic Regression, which runs linear regressions with penalization for correlations over predictors; Least Squares Support Vector Machine with Polynomial Kernel, which obtains support points that are able to divide hyper-plans of decision thresholds; Gradient Boosting Machines, which compile several classifiers, weighted on in-train accuracy; and Random forest and CART, which estimate decision thresholds based on variable cut-off cascades. These methods will be particularly useful in visualizing what variables are more important to detect extortion (or in other words, some of the tree nodes are good for dividing the firms that were extorted from the firms that were not).

Finally, the selected model of each type is used over the testing set, as a second out-of-sample validation.

Results

Classifiers

The model XX was chosen as the best classifier for its superior accuracy of XX percent. The accuracy of our best classifier is better than the baseline accuracy for all the dependent variables in an average of XX percentage points.

The Precision-Recall curve shows that choosing a threshold of XX brings a precision of XX and a recall of XX.

Using that model, the most important features of a firm that explain extortion are y1, y2, y3 and y4, which are consistent with the literature (**paper 1**,**paper 2**,**paper 3** and **paper 4**).

```
# Table 5
## Efficiency Metrics per Model
### Column: Statistics
#### Sensitivity
```

[3]The **R Caret Package** was use to implement the models (the standard grid is the default set of parameters).

```
#### Specificity
#### Pos Pred Value
#### Neg Pred Value
#### Detection Rate
#### Detection Prevalence
#### Balanced Accuracy
#### Positive Class
### Row: Models
##### Regularized Logistic Regression
##### Least Squares Support Vector Machine with Polynomial Kernel
##### Gradient Boosting Machines
##### Random Forest
##### CART
```

```
# Table 6
## Confusion Matrix
### Columns: Reference (0,1)
### Rows: Prediction (0,1)
```

```
# Figure 1
## Precision-Recall Curve
```

```
# Figure 2
## Variable Importance Graph (for the best classifier)
```

Continues variables

The model XX was chosen as the best continuous model for having the smallest RMSE[4]. Further, the most important features of a firm that explain extortion are y1, y2, y3 and y4 (see **Table 7**).

```
# Table 7
## Variable Importance: Continous Variables
### Columns: Continuous Dependent Variables
### Rows: Set of top five predictors (By Magnitud and Significance)
```

Extrapolation of results

The weights of the model were applied to the whole universe (the 5.6 million firms that existed in Mexico during 2014). In **Figure 3** can be observed the predicted geographical concentration of the extortion incidence during 2016.

```
# Figure 3
## Map of the Geographical Distribution of Likelihoods
```

As a robustness check, data of the SESNSP[5] for extortion prevalence was used as a second source. Notably,

[4]RSME Best Model.

[5]Where SESNSP is the Spanish acronym of the Executive Secretariat that is in-charge of aggregating crime statistics from all the crime prosecutors around the country.This data-set is composed with information of all the preliminary inquiries recorded by the States and the Federal prosecutors in Mexico. It is recorded since the 1997 and covers 22 types of crimes (homicide, rape,

this source of information was not used initially because it presents an under-reporting rate that is as high as 98 percent for extortion prevalence [6].

The spatial correlation between our predicted extortion prevalence for 2014 and the aggregated extortion prevalence of 2014 is of XX. That indicates a strong correlation between the two estimations. The result points to a close geographical variation between the two sources, evidence in our argument's direction.

Policy Applications

According to our results, it's possible to predict extortion prevalence. Also, Estevez (2016), was able to predict extortion incidence, giving a proof of the hypothesis that "repeat victimization consistently reveals that crimes concentrate within a small subset of firms."

Having information about what firms are extorted, and knowing that they are likely to be extorted several times, gives the law enforcement agencies a unique opportunity to guide their prevention efforts through the strategic allocation of resources according to victimization risk of each unit.

Further Reseach

Based on the discussed relationship between extortion prevalence and the facing of higher financial/operation costs, we should expect to predict higher attrition rates in the chunks of firms that have higher likelihood of been extorted. In further research, we are going to show that the inverse probability of a vulnerable firm to be selected in the next victimization survey of 2018 (based on the survey stratification strategy) does not correspond to the actual probability of selection: the difference of the estimated probability and the actual probability will be attributed to the attrition, consequence of being extorted.

```
# Table 8
## Expected results: Firms Attrition as a consequence of Extrotions
### Columns: Predicted and Real Survey Turnout probability, Difference, T-Test Difference
### Rows: Extortion Likelihood Strata
```

This further step will contribute to the economic theory of structural causes that explain firms attrition, particularly, in the context of countries with weak rule of law.

Conclusion

Firms are more or less prone to be victimized depending on its characteristics. Where is a firm located, what it produce, how much it sells, how old is it, profitability margin and the other characteristics together made of every firm a unique profit generating entity that represent an opportunity for rent extraction. Where extracting rents through exploiting a competitive advantage in violence, say through treats, is called extortion.

This project will used the economic census and victimization surveys (both described below) to fit the likelihood that a firm will be victimized based on its characteristics and location. The victimization surveys will tell us if a firm was victimized for a sample of around 33 thousand units that are representative of the universe of firms (around 4.5 millions).

We modeled the likelihood of victimization of a firm. Also, we described the results of estimating seven statistical learning techniques where the models were selected under the criteria of having a plethora of

common thief, etc.) with some basic categories for each type (local or federal jurisdiction for example). The data is published in a monthly basis with one month of delayed (the information of January is published until February 20Th, for example). The granularity of this information is municipality and monthly level.

[6] Using the ENVE estimates as a contrast figure.

classifiers, each one representing a different estimation technique. Particularly, the linear regression was estimated for the continuous dependent variables, and classifiers were used for the dichotomous dependent variables.

After observing the results of the estimated models, model XX was chosen for its superior accuracy of XX percent. The accuracy of our best classifier is better than the baseline accuracy for all the dependent variables in an average of XX percentage points.

The weights of the model were applied to the whole universe. The spatial correlation between our predicted incidence for 2016 and the aggregated incidence of the SESNSP data (a second source of information) is XX.

Having information about what firms are extorted, and knowing that they are likely to be extorted several times, gives the law enforcement agencies a unique opportunity to guide their prevention efforts through the strategic allocation of resources according to victimization risk of each unit.

Lastly, based on the discussed relationship between extortion prevalence and the facing of higher financial/operation costs, we should expect to predict higher attrition rates in the chunks of firms that have higher likelihood of being extorted.

Bibliography

Burrows, J. & Hopkins, M. (2005). Business and crime. In N. Tilley (Ed.), Handbook of crime prevention and community safety (p. 486-515). Cullompton, Devon: Willan.

Estevez, P. (2016), Repeat extortion victimisation of Mexican businesses, University Collage London MPhil Tesis.

Farrell, G., Phillips, C. & Pease, K. (1995). Like taking candy-why does repeat victimization occur. Brit. J. Criminology, 35, 384.

Gary S. Becker, "Crime and Punishment: An Economic Approach," Journal of Political Economy 76, no. 2 (Mar. - Apr., 1968): 169-217.

Gill, M. (1998). The Victimization of Business: Indicators of Risk and the Direction of Future Research. International Review of Victimology, 6(1), 17-28. Retrieved from http://irv.sagepub.com/cgi/content/abstract/6/1/17 doi: 10.1177/026975809800600102

Hopkins, M. & Tilley, N. (2001). Once a Victim, Always A Victim? A Study of How Victimisation Patterns may Change over Time. International Review of Victimology, 8(1), 19-35. Retrieved from http://irv.sagepub.com/cgi/content/abstract/8/1/19 doi: 10.1177/026975800100800102

INEGI. (2016). Encuesta Nacional de Victimizacion de Empresas (ENVE - 2016). México DF.

Mugelini, G. (2013a). Measuring and analyzing crime against the private sector: International experiences and the Mexican practice. Mexico DF: INEGI.

Perrone, S. (2000). Crimes against small business in Australia: A preliminary analysis. Trends & Issues in Crime and Criminal Justice, 184 .

Sparks, R. F. (1981). Multiple victimization: Evidence, theory, and future research. Journal of Criminal Law and Criminology, 762-778. Retrieved from http://www.jstor.org/stable/1143014 doi: 10.2307/1143014

Annex

```
# Table 3
## Summary Statistics of independent Variables
```

```
### Columns: name, mean, Standard Deviation, NA's
### Row: variable (by dicotomous or continues)
```