Machine learning models to identify the risk of modern slavery in Brazilian cities

Marlu da Silva Santos
Computer Science Dept.
University of Brasília
Brasília, Brazil
marlusilva@gmail.com

Marcelo Ladeira
Computer Science Dept.
University of Brasília
Brasília, Brazil
mladeira@unb.br

Gustavo C. G. Van Erven
Computer Science Dept.
University of Brasília
Brasília, Brazil
gvanerven@gmail.com

Gladston Luiz da Silva
Department of Statistics
University of Brasília
Brasília, Brazil
gladston@unb.br

***Abstract*—The scope of modern slavery encompasses human trafficking, forced labor, debt bondage and child labor. This article proposes the use of predictive models to identify the risk of modern slavery in Brazilian cities using real socioeconomic, demographic and rescue operations data. The study uses the *embedded* technique with Lasso regularization (L1) to select variables. A comparative analyze of techniques for treatment of imbalanced data was applied and the results indicated the Random Over-Sampling (ROS) as the best one. In total, 16 models are evaluated, consisting of 8 different data sets and two classifiers: Logistic Regression (LR) and Gradient Boosting Machine (GBM). The results indicate that the GBM model has better performance and efficiency, with accuracy of 77%, AUC 80% and G-mean of 71%.**

***Index Terms*—Machine learning, Logistic Regression, Gradient Boosting, data mining, imbalanced dataset, Lasso, slavery.**

I. Introduction

The definition and reason of modern slavery, with regard to juridical and social aspects, is a research topic of interest. [1]. Studies demonstrate that are part of the scope of modern slavery: human trafficking [2], forced labor, debt bondage and child labor [3] [4]. Forced labor is conceptualized, according to ILO Convention (No. 29) [5], as being, "...all work or service which is exacted from any person under the menace of any penalty and for which the said person has not offered himself voluntarily".

The term is defined by the Article 149 of the Brazilian Criminal Law[1], what extends the meaning of forced labor and disposes of the condition analogous to that of slave, referring to someone who is subjected to forced labor or exhaustive workdays or degrading conditions of work or restriction of the locomotion of the employee due to debts contracted with the employer.

The estimation of people in conditions of slave labor has been investigated mainly by non-governmental institutions (ONG)[2,3]. The technical reports produced by these institutions are aimed at building a better understanding of the problem and also on the provision of mechanisms to aid decision-making and public policy proposal. [6] [3] [4].

In order to eradicate slave labor and preserve civil rights, various measures are motivated by government institutions[4,5] and institutions that adhere to the cause of extinction of the modern slavery. The Recommendation 203 of the ILO's 2014 protocol [7], which deals with Convention No. 29 [5], recommends that member nations should implement national policies and plans of action aimed to suppression of forced or compulsory labor in all its forms through prevention, protection and access to resources, such as compensation of victims and punishment of perpetrators.

The high rate of modern slavery is worrying. A statistical data research released by the International Labour Organization (ILO), estimated that by 2016, at least 1.3 million people were in some sort of modern slavery or degrading work in the Americas [3].

According to the data from the Brazilian Secretary of Labor, between 1995 and 2019, Brazil had 54,056 cases[6] of redemption of people in situations of work analogous to slavery, 1,744 people were rescued only in 2018. Brazil is a member of the ILO and has ratified agreements, including the Convention No. 29 [5], which refers to the commitment to abolish the use of forced labor and requires member countries to follow the Recommendation No. 203 of Protocol 2014 [7]. These are agreements that reinforce the recognition of the existing problem by the nation and the interest in its resolution.

Supervision is one of the axes that act directly in the fight against this type of crime, promoted through audits, where operations are mainly motivated by complaints, but only a small part is served due to high demand. The lack of mechanisms to assist in the choice of the regions to be investigated is a reality in Brazil.

This study proposes the use of prediction models to identify the risk of modern slavery in Brazilian cities. The risk is an indication of the chance of this type of crime occurs in a specific city and can be used to prioritize actions in certain regions or to choose which denunciations to investigate. The *CRoss Industry Standard Process for Data Mining* (CRISP-DM) [8] process reference method is used to conduct tasks involving

[1] www.planalto.gov.br/ccivil_03/decreto-lei/del2848.htm
[2] https://www.alliance87.org, https://www.minderoo.com.au/walk-free/
[3] https://ilo.org, https://www.antislavery.org
[4] Brazilian Public Ministry of Labor: https://mpt.mp.br
[5] FBI Civil Rights: https://fbi.gov/investigate/civil-rights
[6] Available at https://sit.trabalho.gov.br/radar/

978-1-7281-4550-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMLA.2019.00132

data mining. This study uses the treatment of unbalanced data, feature selection on high dimensional bases, construction and evaluation of predictive models with Logistic Regression (LR) and Gradient Boosting Machine (GBM) algorithms.

This research deals with a real-world problem that occurs in several regions of the globe. The proposal addressed in this work is of interest of institutions, that have in their scope, the search for solving the problem of contemporary slave labor. The Brazilian Public Ministry of Labor is considering the use of the risk given by our model as another guideline to plan actions in the fight against modern slavery in Brazil.

This paper is structured as follows. Section II presents a brief explanation of the theory used in the study. Section III discusses the related works. Section IV presents the research methodology used. Section V analyzes the results. Finally, the conclusions and future work are presented.

II. Background

This section addresses the main articles, researches and academic works of major relevance used for the development of this work. In the first moment the contextualization of feature selection is done, representing an important step in the analysis and preparation of the data. Subsequently, the techniques for imbalanced data sets are presented, and then the literature review is performed on the supervised algorithms that will be used in the models. At the end of this section, techniques for model evaluation are presented.

A. Feature selection

Different strategies can be applied to different data types, the main ones being *wrapper*, *filter*, *embedded*. In the *wrapper* strategy, a algorithm searches a subset of features. The quality of the subset of selected features is evaluated based on a specific learning algorithm. This process is repeated many times until the convergence criterion is achieved. This strategy offers low performance to be applied with high data dimension [9]. Another strategy is the *filter*, which consists of two steps. The first one is ranking the importance of the feature based on evaluation criteria, which can be univariate (individual variable rankings) or multivariate (batch rankings). In the second step the lowly ranked features are filtered out. [9] [10].

The *embedded* strategy embeds the feature selection into model learning. This strategy proves to be more efficient. The most commonly used techniques are regularization models [11], which try to fit a learning model by minimizing the fitting errors and forcing feature coefficients to be small (or exact zero) simultaneously [9] [10]. Two regularization models used to reduce the number of attributes with little significance are *Least absolute Shrinkage and Selection Operator* (LASSO) [11] and Ridge [12]. In general, regularization techniques use the cost function penalty. For this study the regularization will be used for variable selection and reduction of *overfitting* [13]. The general formulation can be given by equation 1.

$$\min_{f \epsilon H} \left[\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \right] \tag{1}$$

Where, $J(f)$ is the penalty given for the cost function $L(y_i, f(x_i))$. H is the space of functions $J(f)$.

The LASSO is a method whose purpose is to minimize the residual sum of the squares, being able to be used for model estimation. Produces good capacity for interpretation, with general application in models, including for the feature selection and regularization [11].

Unlike LASSO, which penalizes according to the sum of the absolute values of the coefficients, reducing the quadratic error, Ridge Regression penalizes the size of the regression coefficients by means of a Ridge estimator, decreasing the importance of the attribute [13] [12].

B. Imbalanced datasets

In machine learning approaches that use binary classification, class imbalance generates performance problems for the model. The majority classes tend to influence outcomes when compared to minority classes [14], [15], [16]. The treatment for unbalanced classes can be performed at the data level itself [17], in which it was the choice of this study.

Different resampling techniques can be used for data balancing, such as random undersampling (RUS) and random oversampling (ROS) approaches. In the RUS technique, the majority class is reduced to the size of the minority class, while the ROS makes copies of the minority class data in a random way. Other intelligent sampling techniques are derived from these main ones, namely Synthetic Minority Oversampling Technique (SMOTE) [18], Adaptive Synthetic Sampling Approach (ADASYN) [19], Cluster [20] and SMOTE-Tomek, which combines the SMOTE technique with data cleaning techniques (Tomek link) [21] to reduce the overlap generated by the oversampling methods. Good results were obtained in the study by Zhu *et al.* [17] using the resampling techniques ROS and SMOTE-Tomek with Random Forest classifier [22].

C. Classifiers

We use Logistic Regression (LR) as the classifier in this research because it was largely used on related works. An additional classifier, Gradient Boosting Machine (GBM), was used in comparison. In fact, we considered several others traditional classifiers. However, we decided to choose only these two classifiers in this study because they performed better than the others and to make benchmarking analysis between balancing techniques simpler.

LR belongs to the Generalized Linear Models (GLM) group and is strongly related to problems that needs to describe relations between independent variables (predictors) and dependent variables (response) with binary or dichotomous output, as opposed to linear regression that commonly the target variable is continuous [23]. The logistic model is represented by equation (2).

$$P(X) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \tag{2}$$

Where the terms α and β_i are represented by unknown parameters which are required to be estimated based on the values obtained of X_i. $P(X)$ is the probability of the dependent variable be 1 given the set of variables X_i.

The algorithm Gradient Boosting Machine (GBM) was introduced by Friedman [24], having Boosting, as the main base technique. Typically, the GBM algorithm gives a prediction model in the form of an ensemble of decision trees. The Boosting algorithm can be seeing as an iterative functional gradient descent algorithm, that optimizes a cost function over a function space based on the negative gradient direction. It transforms weak prediction models in the new trees that are added to the prediction model.

D. Evaluation of models

An evaluative approach of models contributes to the comparison of different algorithms and techniques. The confusion matrix is based on a prediction problem, for cases where there is a class to be predicted (dichotomous). The matrix will consist of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [13].

Derived metrics from the confusion matrix are commonly used on the evaluation models, as described by Wang *et al.* [16] and Powers *et al.* [25], such as precision, accuracy, and recall. Precision is obtained by the ratio of correct predictions, over true positives, formulated by the equation (3).

$$precision = \frac{TP}{TP + FP} \tag{3}$$

The sensitivity rate, also called a *recall*, is obtained by the ratio of true positives to true positives and false negatives, represented by equation (4).

$$recall = \frac{TP}{TP + FN} \tag{4}$$

The specificity is the ratio of true negatives over the total of true negative and false positive that is the true negative rate (TNR) given by (5).

$$specificity = \frac{TN}{TN + FP} \tag{5}$$

Accuracy is defined by the proportion of true (positive and negative) results, over the total number of predicted and original cases. Accuracy is mathematically formulated by equation (6).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

In a binary classification the *F1-score* or *F1-measure* considers the values of *recall* and the *precision* the score being a harmonic mean, with a result between 0 and 1, where its best value is obtained in 1. The F1-measure is formulated by the equation (7), if precision and sensitivity have the same weighting [13].

$$\textit{F1-measure} = \frac{2 \,.\, precision \,.\, recall}{precision + recall} \tag{7}$$

In unbalanced data approaches the evaluation metrics interfere in the results. Zong et al [15] and Liu et al [14] use geometric mean (G-mean), whose purpose is to maximize the accuracy in each of classes, maintaining balance.

$$\textit{G-mean} = \sqrt{recall \,.\, specificity} \tag{8}$$

The *Receiver Operating Characteristic* (ROC), popularly known in the literature as the ROC curve, is a form of representation of a binary classifier system that results in a performance graph, aiming to summarize classifier performance between error. Initially the ROC curve was designed for signal-detection theory, later it was first used by psychological and medical studies [26] [27] [13]. *Area Under the Curve* (AUC) is a metric used to measure performance under a ROC curve [28].

III. Related works

At our best known, we did not find papers on the Scopus and Web of Science databases describing studies dealing with modern slavery that directly use machine learning techniques. However, we found four works with some similarity that use statistical methods, in specific Logistic Regression.

The first, in [29], estimates and identifies vulnerability to slave labor and sexual exploitation of adults and children. For slave labor in adults, a census, conducted between 2014 and 2016, is used in 48 countries, based on the responses mapped and quantified to slavery by country and continent. For the cases of sexual exploitation of adults and children, the basis of estimates for slave labor in adults and a human trafficking database with 21 (twenty one) variables are used, where 3 models are constructed using logistic regression, by means of the *odds ratio* (probability) and not a binary classification, being able to identify the vulnerability rate.

On the second work [30], risk models prediction are designed for slave and forced marriage, resulting in a global report [4]. Using as a method the census database conducted between 2014 and 2016, the same also used by [29]. The work used variable dimensions for the following vulnerabilities: governance; social access (areas with malnutrition index); inequality; legally unfit (as immigrants); and effects of conflicts. Applying weights in certain groups of countries and using a logistic regression technique, the probability of forced slave labor and forced marriage is calculated. Two groups of models are developed, individual-level and multi-level models.

In the works of Datta and Bales [31] [32], estimation techniques were applied to predict the number of slaves in European countries, using as reference the work of [3] [29]. The Table I presents the techniques used by related works.

Table I: Related works and techniques used

References	Techniques	Evaluation
[29]	*Logistic regression*	standard error
[30]	*Logistic regression* *Multilevel model*	*log-likehood* *wald test* *AIC* *BIC* R^2 *AUC*
[31], [32]	Statistical estimates	standard error

New approaches can be proposed, based on a careful selection of features, since none of the papers presented the application of techniques to evaluate the selection of variables. In both related works they used the same census as primary source of data.

No work was found in the literature for the subject under study, which uses data mining or ML techniques and comparison between classifiers. In this way, the solution proposed by this work, incite the state of art of this problem. Proposing the use of models to predict the risk of slave labor in cities, using a set of real data of operations to rescue slave labor. Through supervised learning, different classification techniques are proposed in a comparative manner, using variable selection and use of different model evaluation metrics, thus providing greater accuracy.

IV. Methodology

In this section, the methodology followed in this study is discussed. The datasets used are: the Atlas of the Municipalities of Brazil[7] and Slave Labor Rescue Operations[8] from 2003 to 2018. The tools used are python[9] programming language with packages scikit learn, imbalanced-learn, matplotlib, plotnine and multiprocessing. An overview of the flow is shown in Fig. 1.

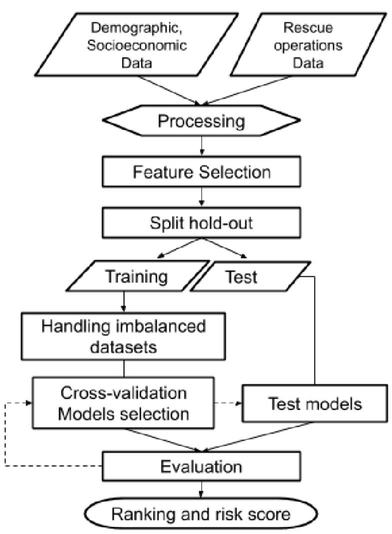


Figure 1: Overview of steps

- In the first stage, information is collected from the work groups involved in the eradication of slave labor, characterizing the understanding of the business and the problem to be addressed, as presented in section I;
- After understanding the business, the understanding of the data is carried out. In this stage, the databases are collected and analyzed, an initial pre-processing of data cleaning is accomplish. The data has a scale between 0 and 100, representing the percentages. Two data sets are used to form the final data set. The data of rescue operations, with 3318 records, representing 44238 people rescued, having 4 dimensions: the name of the city, state code, number of the rescue operation, year of the rescue and the number of workers rescued. Atlas data, which show demographic, educational, income, labor, housing and vulnerability indicators, are based on the country's last census (2010), representing 5565 records and 237 indicators (dimensions).
- With the initial cleaning performed, the next step is to perform the data preparation, which will result in a unified version of the data set classified with the target variable dependent and the independent variables. The data are classified with values 0 (rescue cases not confirmed) or 1 (rescue cases confirmed). At the end of the classification, 767 (13.7%) of cases were confirmed and 4798 (86.22%) were not unconfirmed. The selection of variables is necessary because the data set have 237 variables, of which 232 are continuous and 5 are categorical. For dimensionality reduction, an embedded feature selection technique using L1 (lasso) regularization is applied, resulting in 16 variables, which are presented in Table II. The final data set presents dimensions of 5,565 x 17, where 5,565 represents the number of cities. In the dimension of the selected features were added to the target variable (dependent), totaling 17 dimensions. The data description is shown in Table III, which has the frequency for each selected variable, maximum, minimum, mean, standard deviations and respective distributions by quartiles.

Table II: Features selected

Features selected	
Feature name	**Description**
prentrab	Percentage of income from earned income
parede	Percentage of the population living in households with walls that are not masonry or suitable
t_atraso_1_basico	percentage of the population aged 6 to 17 years attending basic education, which is 1 year behind
t_agua	percentage of the population living with running water
pren10ricos	Percentage of total income appropriated by the 10% of the population with the highest per capita household income
t_super25m	Percentage of the population aged 25 or over
trabsc	Percentage of employed persons aged 18 or over who are employed without a portfolio
t_freq4a5	Rate of 4 to 5 year old population attending school
ren2	Percentage of employees with income of up to 2 minimum salary
pren40	Percentage of total income appropriated by 40% of the population with the lowest per capita household income
rind	Average household income per capita of the extremely poor
pren20	Percentage of total income appropriated by 20% of the population with the lowest per capita household income
t_flpre	Pre-school net attendance rate
p_transf	Percentage of employees in the manufacturing industry
ren1	Percentage of employees with income of up to 1 minimum wage
t_env	Aging rate

- The modeling stage is aimed at the construction of risk prediction models of slave labor. According to the study drawn up, the choice was made to use Logistic Regression [33] [34] and GBM for classification. The data set is divided into training (70%) and test (30%) according to the hold-out approach. The test is performed on an unbalanced 30% data set. Because it is a classification problem

[7] Available at http://www.atlasbrasil.org.br/2013/pt/download/
[8] Provided by Brazilian Public Ministry of Labor (MPT)
[9] https://www.python.org

Table III: Data description.

	prentrab	parede	t_atraso_1_basico	t_agua	pren10ricos	t_super25m	trabsc	t_freq4a5	ren2	pren40	rind	pren20	t_flpre	p_transf	ren1	t_env
count	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0	5565.0
mean	68.48	5.36	18.29	85.59	38.20	5.49	25.22	78.45	82.37	12.11	32.03	3.72	54.77	9.61	39.30	8.39
std	10.79	9.41	4.01	14.7	5.91	3.25	9.85	15.46	10.35	3.25	9.60	1.54	15.92	8.92	21.58	2.42
min	27.43	0.00	4.33	0.15	22.26	0.28	3.03	13.03	36.49	0.00	0.00	0.00	3.91	0.00	4.53	1.46
25%	61.06	0.41	15.72	79.65	34.17	3.24	17.64	70.37	74.90	9.920	27.44	2.45	44.26	3.33	19.60	6.78
50%	70.58	1.64	18.80	90.28	37.62	4.81	24.75	82.14	83.86	12.17	32.51	3.72	55.56	6.53	35.81	8.38
75%	76.60	5.82	21.03	96.26	41.63	6.95	32.04	90.13	91.57	14.45	37.09	4.90	66.07	13.31	58.55	9.96
max	95.24	82.74	32.31	100.00	75.34	33.68	62.23	100.00	99.140	22.50	70.00	9.26	100.00	65.11	89.33	20.42

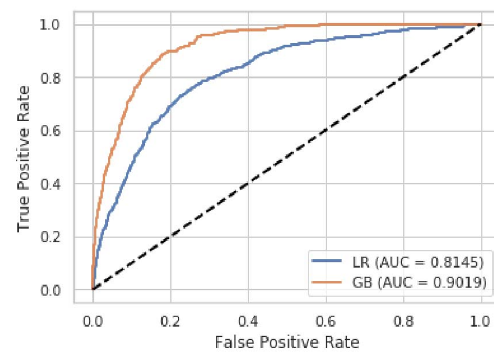
in unbalanced data, the study raised in Section II suggests the balancing of data in training. Due to the need of works with similar context to this study, it was decided to perform the benchmarking to choose the best technique of data balancing for the reality of the problem. The training and test is performed for an unbalanced set (base) and for the RUS, ClusterCentroids, ROS, SMOTE, SMOTEEN, SMOTETomek and ADASYN balancing techniques. In total, 16 models (8x2) are evaluated, based on 8 different data sets and 2 classifiers. Results are presented in Tables IV and V for training and testing respectively. Sampling methods are evaluated through the arithmetic means of the techniques (LR and GBM), for $k = (1, ..., n)$, where k is the classification technique and n is the maximum number of techniques by classification, in this study $n = 2$. The mean values are obtained for each metric (accuracy, auc, precision, recall, f1 and G-mean), formulated by $(\sum_{k=1}^{n} metric_k)/n$, where $metric_k$ is the metric result for each classification technique. The sampling method is chosen using the criterion of better means. With the choice of data set to be used, the training is performed using cross-validation with 10-folds. The model is tested using the data test (30%) which is unbalanced and unknown by the trained model.

- The evaluations of the models will be measured by the accuracy and F1-measure, precision, recall, G-mean. The performance will be verified by the use of the ROC curve.
- In the final step, the implementation of the selected model is applied, the results will be displayed for interested parties. An additional step is applied to the model, performing discretization [35] for a modern slavery scale [29].

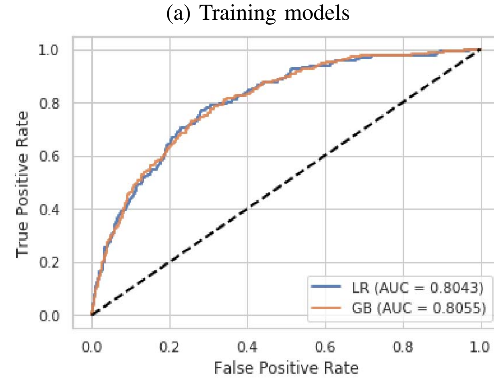
V. Results

The values of the classification algorithms LR and GBM for each type of training and testing data balancing technique, used in benchmarking, are presented respectively in Tables IV and V. The results of the averages obtained for each type of training and testing data balance technique are presented in Tables VI and VII. In training, the results suggest that the SMOTEEN, SMOTETomek, SMOTE and ROS balancing techniques obtained better results in the AUC, F1 and G-mean metrics.

The selection of the best technique for balancing the data for the context of this work was validated with the use of the test database. The data from Table V are used to compose the averages presented in Table VII. The results show that an unbalanced base has low recall rates, F1 and consequently



(a) Training models



(b) Test models

Figure 2: The ROC curve for the models in the ROS datasets

low value for G-mean, suggesting that the use of accuracy alone is not a form of evaluation for unbalanced data set. On the contrary, when testing data sets with balancing techniques *recall* values, F1 and G-mean are better. The best technique evaluated for accuracy, AUC, F1 and G-mean was ROS.

The Fig. 2 presents the Receiver Operating Characteristic (ROC) curves of positive classes of training and test models using ROS balancing, in our study correspond to binary values 1. As can be observed, in the models (AUC) is lower for the LR model compared to the GBM model with the highest AUC. With a higher value of AUC indicates better model predictability, with better rates of true positives.

The results of the metrics obtained for the two models in the validation under test are presented in the Table VIII. It is possible to observe that GBM presents better results. The LR model has values with recall and G-mean metrics greater than GBM. The GBM model has better performance with superior results in accuracy, AUC and F1.

Table IV: Benchmarking of Balancing Techniques in the Training Dataset

	accuracy		auc		precision		recall		F1		G-mean	
Method	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*
base	0.8652	0.8652	0.8105	0.8030	0.5953	0.5980	0.2037	0.2074	0.3022	0.3057	0.4461	0.4499
RUS	0.7279	0.7092	0.7972	0.7821	0.7240	0.7033	0.7387	0.7316	0.7302	0.7154	0.7280	0.7089
ClusterCentroids	0.7521	0.7468	0.8235	0.8331	0.7554	0.7533	0.7494	0.7387	0.7509	0.7429	0.7522	0.7468
ROS	0.7473	0.8303	0.8144	0.9019	0.7469	0.8069	0.7488	0.8690	0.7475	0.8365	0.7473	0.8294
SMOTE	0.7618	0.8126	0.8304	0.8900	0.7547	0.7929	0.7763	0.8471	0.7649	0.8187	0.7617	0.8119
SMOTEEN	0.8436	0.8938	0.9129	0.9567	0.8636	0.8943	0.8912	0.9417	0.8770	0.9173	0.8251	0.8753
SMOTETomek	0.7660	0.8226	0.8330	0.8943	0.7589	0.7996	0.7804	0.8615	0.7693	0.8293	0.7659	0.8217
ADASYN	0.7369	0.8032	0.8018	0.8739	0.7308	0.7763	0.7514	0.8536	0.7409	0.8130	0.7368	0.8015

Table V: Benchmarking of Balancing Techniques in the Test Dataset

	accuracy		auc		precision		recall		F1		G-mean	
Method	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*	*LR*	*GBM*
base	0.8790	0.8766	0.8030	0.8029	0.5468	0.5135	0.1682	0.1826	0.2573	0.2695	0.4061	0.4221
RUS	0.7425	0.7155	0.8044	0.7928	0.2881	0.2645	0.7259	0.7211	0.4125	0.3870	0.7353	0.7179
ClusterCentroids	0.6958	0.6293	0.7884	0.7532	0.2508	0.2157	0.7259	0.7500	0.3728	0.3351	0.7085	0.6775
ROS	0.7371	0.7742	0.8043	0.8054	0.2824	0.3066	0.7211	0.6442	0.4059	0.4155	0.7302	0.7146
SMOTE	0.7353	0.7550	0.8004	0.7994	0.2784	0.2866	0.7067	0.6490	0.3994	0.3976	0.7228	0.7070
SMOTEEN	0.6137	0.6634	8022	0.8035	0.2230	0.2449	0.8461	0.8173	0.3530	0.3769	0.7009	0.7241
SMOTETomek	0.7383	0.7622	0.7998	0.7918	0.2793	0.2895	0.6971	0.6250	0.3989	0.3957	0.7202	0.6990
ADASYN	0.7149	0.7389	0.7981	0.7930	0.2697	0.2816	0.7548	0.7067	0.3974	0.4027	0.7317	0.7248

Table VI: Training - The mean values for each sampling method using the classification models (LR and GBM)

Method	$\overline{acc}$	$\overline{auc}$	$\overline{precision}$	$\overline{recall}$	$\overline{F1}$	$\overline{G\text{-}mean}$
base	0.8652	0.8067	0.5967	0.2056	0.3039	0.4480
RUS	0.7186	0.7897	0.7137	0.7352	0.7228	0.7184
ClusterCentroids	0.7494	0.8283	0.7543	0.7441	0.7469	0.7495
ROS	0.7888	0.8581	0.7769	0.8089	0.7920	0.7883
SMOTE	0.7872	0.8602	0.7738	0.8117	0.7918	0.7868
SMOTEEN	0.8687	0.9348	0.8790	0.9165	0.8972	0.8502
SMOTETomek	0.7943	0.8636	0.7793	0.8210	0.7993	0.7938
ADASYN	0.7701	0.8378	0.7535	0.8025	0.7770	0.7691

Table VII: Test - The mean values for each sampling method using the classification models (LR and GBM)

Method	$\overline{acc}$	$\overline{auc}$	$\overline{precision}$	$\overline{recall}$	$\overline{F1}$	$\overline{G\text{-}mean}$
base	0.8778	0.8029	0.5301	0.1754	0.2634	0.4141
RUS	0.7290	0.7986	0.2763	0.7235	0.3998	0.7266
ClusterCentroids	0.6625	0.7708	0.2332	0.7379	0.3539	0.6930
ROS	**0.7556**	**0.8049**	**0.2945**	0.6826	**0.4107**	**0.7224**
SMOTE	0.7452	0.7999	0.2825	0.6778	0.3985	0.7149
SMOTEEN	0.6386	0.8028	0.2340	**0.8317**	0.3650	0.7125
SMOTETomek	0.7502	0.7958	0.2844	0.6610	0.3973	0.7096
ADASYN	0.7269	0.7956	0.2756	0.7307	0.4001	**0.7282**

Table VIII: Evaluation of the data set models under test

Model	*accuracy*	*AUC*	*precision*	*recall*	*F1*	*G-mean*
LR	0.7371	0.8043	0.2824	**0.7211**	0.4059	**0.7302**
GBM	**0.7742**	**0.8054**	0.3066	0.6442	**0.4155**	0.7146

The model with best result was selected, GBM with ROS technique for balancing in training. The degree of risk for each city was determined, as shown in Fig. 3. For this, it was used the predictive power of the model to obtain the probability of each class, in this case, the class of interest with value 1. With the probability of occurrence of the class, the next step was to discretize the values in a scale of slavery, in 4 levels of risk (0 - low, 1- medium, 2- high and 3 - very high). In Fig. 3, the x axis corresponds to the States. In vertical alignment, each point represents a city for a respective State (name anonymized from State 1 to State 27). On the other hand, the y axis has the risk scores for each city. The discretization to levels of slavery reveals the clustering of cities at different levels.

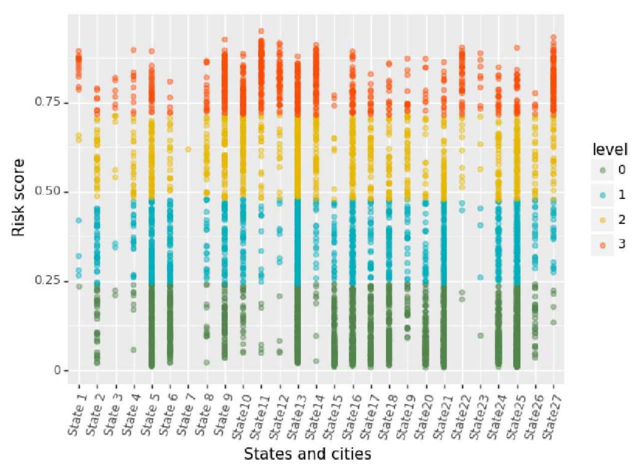


Figure 3: Level and risk score of slavery in Brazilian cities

VI. Conclusion and Future Works

This article has addressed the use of machine learning to predict the risk of modern slavery in Brazilian cities. Using L1 regularization for feature selection the data dimensionality reduction occurred. Because of it is a problem of unbalanced data and in the absence of work in similar contexts, the benchmarking with applications of the main techniques of data balancing was carried out. The contributions of this work can be summarized as following: (i) the score and level of risk for each city was determined; (ii) the main predictor variables were discovered; (iii) ML models have been validated for the proposed problem, new researches may use it as a reference; (iv) benchmarking was performed to select the most efficient method of data balancing.

The risk scores can be used for planning preventive strategic actions, which currently do not have predictive approaches. Risk scores can also be used as an input to rank the denunciations that will be attended.

It was possible to verify with the selected models that the predictions have an acceptable performance for the study proposed, which is the identification of the risk of contemporary slavery for each city. The best performing model was the GBM with 77 % accuracy, 80% AUC and G-mean with 71%. False-positive rates are more evident in the test model, which is expected because, has a low precision rate and the test base is unbalanced, resulting in F1 with 41%.

Limiting factors of the work are the models do not discriminate urban and rural conditions, placing them in a single class and the only use of socioeconomic and demographic data. As future works, new approaches using others classification algorithms and population census datasets can be exploited to construct new models. The use of GridSearch can be applied to parametric models, aiming to improve results. As next step, apply multiclass models to urban and rural regions.

REFERENCES

[1] S. Scarpa, "The Nebulous Definition of Slavery : Legal versus Sociological Definitions of Slavery," *The Palgrave International Handbook of Human Trafficking*, pp. 1–14, 2019.

[2] United Nations Office on Drugs and Crime, "United Nations convention against transnational organized crime and protocols thereto," *Trends in Organized Crime*, vol. 5, no. 4, pp. 11–21, 2007.

[3] International Labor Organization, "Global Estimates of Modern Slavery: Forced Labour and Forced Marriage," Tech. Rep., 2017.

[4] Walk Free Foundation, "The Global Slavery Index," Tech. Rep., 2018.

[5] Members of the International Labour Organization, "Forced Labour Convention - C029," International Labour Office, Ed. Geneva: The General Conference of the International Labour Organization, 1930.

[6] M. N. Datta, F. David, K. Bales, and N. Grono, "The Global Slavery Index," Walk Free Foundation, Tech. Rep., 2013.

[7] International Labour Organization, "R203 - Forced Labour (Supplementary Measures) Recommendation," Geneva, 2014.

[8] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0 Step-by-step," Tech. Rep., 2000.

[9] J. Li, K. Cheng, W. Suhang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature Selection: A Data Perspective," *ACM Computing Surveys*, vol. 50, no. 6, 2017.

[10] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[11] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1995.

[12] R. W. Kennard and A. E. Hoerl, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[13] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning The Elements of Statistical Learning," 2017.

[14] X.-y. Liu, J. Wu, and Z.-h. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *Ieee Transactions on Systems, Man and Cybernetics*, vol. 39, no. 2, pp. 1–14, 2009.

[15] W. Zong, G.-B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalanced learning," *Neurocomputing*, vol. 101, pp. 229–242, 2013.

[16] S. Wang, L. L. Minku, and X. Yao, "A Systematic Study of Online Class Imbalance Learning With Concept Drift," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–20, 2018.

[17] B. Zhu, B. Baesens, A. Backiel, and S. K. Vanden Broucke, "Benchmarking sampling techniques for imbalance learning in churn prediction," *Journal of the Operational Research Society*, vol. 69, no. 1, pp. 49–65, 2018.

[18] B. K. H. L. K. W. Chawla, N.V., "SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[19] H. He, Y. Bai, E. Garcia A., and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," *IEEE International Joint Conference on Neural Networks, IJCNN '08*, no. 3, pp. 1322–1328, 2008.

[20] S. J. Yen and Y. S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications*, vol. 36, no. 3 PART 1, pp. 5718–5727, 2009.

[21] I. Tomek, "Tomek Link: Two Modifications of CNN," *IEEE Trans. Systems, Man and Cybernetics*, pp. 769–772, 1976. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4309452

[22] L. Breiman, "Random forests," *Ensemble Machine Learning: Methods and Applications*, vol. 45, pp. 5–32, 2001.

[23] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. Wiley-Interscience Publication, 2000.

[24] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of statistics*, pp. 1189–1232, 2001.

[25] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," *International Cognitive Science Conference and the 2007 Human Communication Science SummerFest.*, no. December, 2007.

[26] J. A. Swets, "Measuring the Accuracy of Diagnostic Systems," *Science*, vol. 240, no. 4857, pp. 1285–1293, 1988.

[27] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," *Data Mining and Knowledge Discovery Handbook 2nd ed.*, 2010.

[28] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition Society*, vol. 30, no. 7, pp. 1145–1159, 1997.

[29] International Labour Office and W. F. Foundation, "Methodology of the global estimates of modern slavery: Forced labour and forced marriage," International Labour Office, Geneva, Tech. Rep., 2017. [Online]. Available: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/publication/wcms_575479.pdf

[30] P. Diego-Rosell and J. Joudo Larsen, "Modelling the Risk of Modern Slavery," *SSRN Electronic Journal*, 2018.

[31] M. N. Datta and K. Bales, "Slavery in Europe: Part 1, Estimating the Dark Figure," *Human Rights Quarterly*, vol. 35, no. 4, pp. 817–829, 2013.

[32] ——, "Slavery in Europe: Part 2, Testing a Predictive Model," *Human Rights Quarterly*, vol. 36, no. 2, pp. 277–295, 2014.

[33] C. Liu, "Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression," *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, pp. 227–238, 2005.

[34] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su, "A WEAKLY INFORMATIVE DEFAULT PRIOR DISTRIBUTION FOR LOGISTIC AND OTHER REGRESSION MODELS," *IMS Journals and Publications - Annals of Applied Statistics*, vol. 2, no. 4, p. 1360–1383, 2008.

[35] S. Kotsiantis and D. Kanellopoulos, "Discretization Techniques : A recent survey," *GESTS International Transactions on Computer Science and Engineering*, vol. 32, no. 1, pp. 47–58, 2006.