Money-laundering detection using Machine Learning with public data

François Chesnay
fchesnay@stanford.edu

Sebastian Hurubaru
hurubaru@stanford.edu

CS 229 (Aut 2019) Final Project
Mentor: Leo Mehr

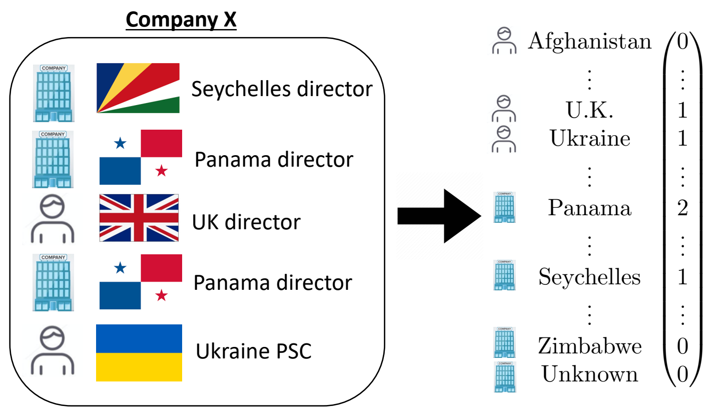
Predicting money-laundering

- According to the United Nations Office on Drugs and Crime [1], it's estimated that 2% to 5% of the global GDP—or $800 billion to $2 trillion, proceeds from drugs and cybercrime to people trafficking —is laundered each year.
- UK companies have been widely used as vehicles for money-laundering [2][3], but illegal activities leave traces which are particularly well-suited for patterns' recognition by computers.
- Our objective is to detect money-money laundering using as input UK corporate data (the nationality and status individual/corporate of its officers and beneficial owner) to perform binary classification of companies.
- An overview work undertaken in fraud detection using artificial intelligence is presented in [4], with good performance using Deep Autoencoder Networks described in [5], and models adapting to the ever changing nature of fraud in explained [6]. Our work is seminal, as we are the first to use company registry.

Data from Company House

- The names, nationalities and status (individual/corporate) for the officers and beneficial owners of UK limited companies and partnerships downloaded from the UK company registry, Company House [7] including companies involved in the Troika laundromat constructed through a search algorithm and a list of companies allegedly involved in money-laundering and other criminal activities by scraping newspapers' articles and Organized Crime and Corruption Reporting Project's website.
- Our dataset comprises 111,147 corporate entities, split between train (100,309), dev (5,280) and test (5,558).

Features extraction and labelling



Convert the raw inputs of the model (nationality and status corporate/individual of the officers and the beneficial owners) for each corporate using a bag of words approach into a vector of size 492.
Use a similar approach to create a vector with regions.

Using Snorkel [8][9], we create 4 labelling Functions to transform our unlabelled to labelled data for entities: (i) with a corporate as beneficial owner, (ii) part of the troika laundromat according to our search algorithm, (iii) with individuals beneficial owners in high risk countries, and (iv) with same officers or beneficial owners as an entity involved in criminal activities.

Labeling Training Data
snorkel
with Labeling Functions (LFs)

Models

- We perform classification using logistic regression, a fully connected neural network with 2, 5 and 10 layers (using Adam optimization with 512 hidden units, a batch size of 128 and train for 500 Epoch), a convolutional neural network (custom version of LeNet-5 where we only use convolution over one dimension and treated the list of features as a sequence), and a SVM with Gaussian Kernel.
- Given the relative simplicity of our input data, we do not expect significant better performance by adding more layers to the neural network..
- We expect the neural networks to perform marginally better than the other models given their capacity to learn more complex functions.
- We train our models on (i) all countries (ii) mix of countries and regions (iii) fully aggregated countries. We add an extra region representing the list of countries blacklisted by the EU as uncooperative.

Results

- The performance is very similar using all countries and a mix of regions and countries. We are presenting the results for the mix of regions and countries for the logistic regression, the fully connected NN and the SVM with Gaussian Kernel. The performance is lower when countries are aggregated fully into regions.
- The CNN performs significantly better on the dataset with all countries, whereas the weakest was the SVM with the sigmoid kernel.
- Except for the CNN, the results below are for the mix of countries and regions.

Logistic regression

Train	Precision	Recall	f1-score	Accuracy
0	0.96	1.00	0.98	0.96
1	0.93	0.59	0.72	
Dev	Precision	Recall	f1-score	Accuracy
0	0.96	1.00	0.98	0.98
1	0.95	0.57	0.71	
Test	Precision	Recall	f1-score	Accuracy
0	0.96	1.00	0.98	0.96
1	0.92	0.57	0.70	

Fully Connected Neural Network with 2 layers

Train	Precision	Recall	f1-score	Accuracy
0	0.98	0.98	0.98	0.97
1	0.83	0.81	0.82	
Dev	Precision	Recall	f1-score	Accuracy
0	0.98	0.98	0.98	0.98
1	0.82	0.78	0.80	
Test	Precision	Recall	f1-score	Accuracy
0	0.98	0.98	0.98	0.98
1	0.82	0.78	0.80	

Convolutional Neural Network

Train	Precision	Recall	f1-score	Accuracy
0	0.98	0.98	0.98	0.97
1	0.83	0.82	0.83	
Dev	Precision	Recall	f1-score	Accuracy
0	0.98	0.98	0.98	0.97
1	0.81	0.80	0.80	
Test	Precision	Recall	f1-score	Accuracy
0	0.98	0.98	0.98	0.97
1	0.82	0.78	0.80	

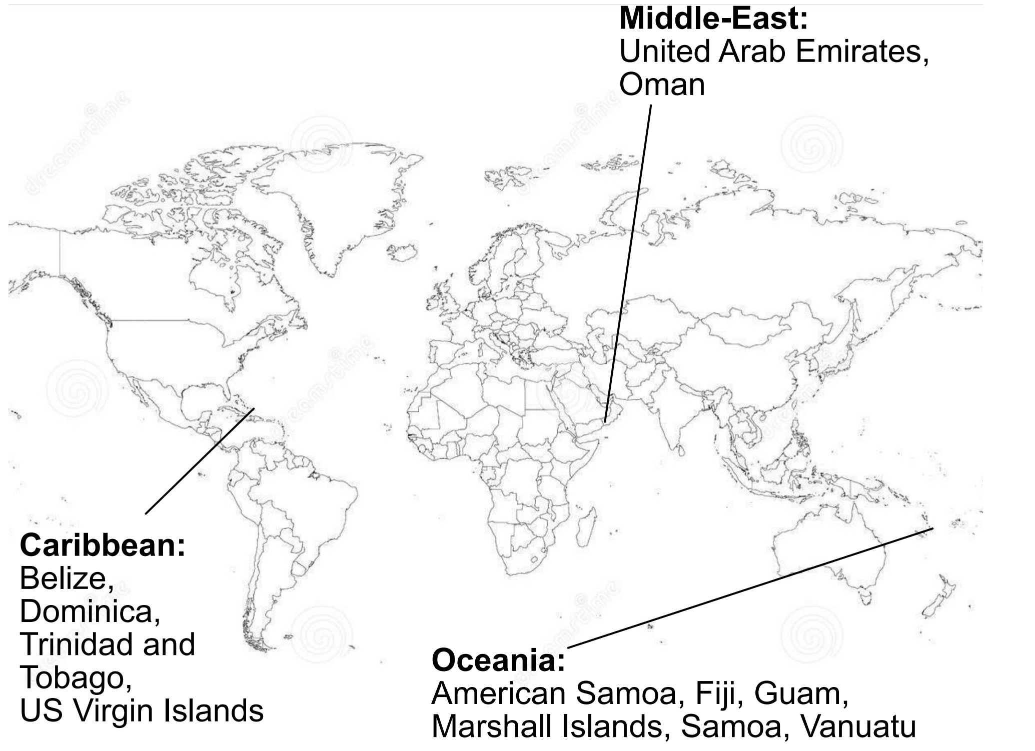
SVM with Gaussian Kernel

Train	Precision	Recall	f1-score	Accuracy
0	0.96	1.00	0.98	0.96
1	0.93	0.60	0.73	
Dev	Precision	Recall	f1-score	Accuracy
0	0.96	1.00	0.98	0.96
1	0.94	0.57	0.70	
Test	Precision	Recall	f1-score	Accuracy
0	0.96	1.00	0.98	0.96
1	0.93	0.57	0.71	

- We have an imbalanced dataset, therefore the best criteria to assess the performance of our models are the recall and f1-score.

Discussion

- Machine learning can detect companies and limited partnerships more likely to be used for illegal activities using the nationality and legal personality (individual or corporate) of its officers and beneficial, as illustrated in red and orange in the map below.
- The best performing models are the fully connected NN and the CNN with a f-1 score of 0.80, demonstrating the existence of patterns in our data.
- We will do further tests to assess the impact of adding the EU blacklist of uncooperative countries (shown on the map below) to assess the impact of their inclusion on the prediction:



Future Work

- Add new inputs to improve the explanatory and predictive power of our model: (i) events extracted from the company registry or from the press through web scraping and (ii) addresses to have a further dimensions to perform network analysis on entities.
- Generate with NLP a written explanation substantiating the rating given with an audit trail of the most important information obtained through web scraping, and provide as output a percentage of suspicion rather than a binary classification.

[1] United Nations Office on Drugs and Crime: https://www.unodc.org/unodc/en/money-laundering/globalization.html

[2] Getting the UK's House in Order. Global Witness: https://www.globalwitness.org/documents/19717/Getting_the_UKs_House_in_Order_xZZxobR.pdf

[3] At your service. Investigating how UK businesses and institutions help corrupt individuals and regimes launder their money and reputations. Transparency International, October 2019: https://www.transparency.org.uk/wp-content/plugins/download-attachments/includes/download.php?id=9299

[4] Efstathios Kirkos, Charalambos Spathis, and Yannis Manolopoulos. Data mining techniques for the detection of fraudulent financial statements. *Expert systems with applications*, 32(4):995–1003, 2007.

[5] Marco Schreyer, Timur Sattarov, Damian Borth, Andreas Dengel, and Bernd Reimer. Detection of anomalies in large scale accounting data using deep autoencoder networks. *CoRR*, abs/1709.05254, 2017.

[6] Maria Jofre and Richard Gerlach. Fighting accounting fraud through forensic data analytics. *CoRR*, abs/1805.02840, 2018.

[7] Companies House: https://beta.companieshouse.gov.uk

[8] Snorkel: Rapid Training Data Creation with Weak Supervision. Alex Ratner, Stephen Bach, Henry Ehrenberg, Jason Fries, Sen Wu, Christopher Ré. VLDB 2018.

[9] Data Programming: Creating Large Training Sets, Quickly. Alex Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, Christopher Ré. NeurIPS 2016

Audio Presentation

Link to the poster presentation: https://www.youtube.com/watch?v=mo7CMWuJ-Ok