The Causes of Revolutions

LSESU Data Science Society — Lent Term 2023 Project Presentation

Executive Summary

- Research question: What are the correlates of social unrest?
- Data: World Bank, Google Trends, and various conflict datasets
- Approach: Deployed multiple models to identify factors that predict revolutions
  - Used linear/logistic regression models and decision tree to understand the most important predictors
  - Developed a neural network model for prediction
- Results: most important factors — GDP, school enrollment, FDI

Theory: three leading causes of revolution

- Demographic structural model: population changes → imbalance in dist. of resources, power, and opportunities, causing social conflict (Goldstone 1993)
- Political fractionalization: radical political/insurgent factions → civil unrest (Bates 2008; Fearon and Laitin 2003; Goldstone 1991)
- Geopolitical theory: imbalance in dynamics of power — defined as as the control of strategic resources such as oil, water etc. (Collins 1980, 1995)

- Applicability in today’s hyperinformation era?

Data from a wide range of datasets

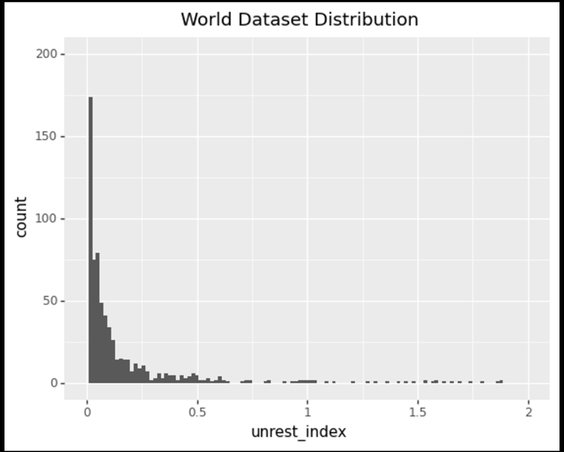
- **Dependent variable:** ‘unrest’ (broadly conceived)
- **Independent variable:** ‘causes of revolutions’
- **Datasets:**
  - The Armed Conflict Location & Event Data Project (ACLED)
  - Reported Social Unrest Index (RSUI)
  - Cross-National Time-Series Data (CNTSD)
  - World Bank Economic Data (WB)
  - Google Trend Keyword searches
    - keywords: protest, revolution, riots, strike, unrest, violence

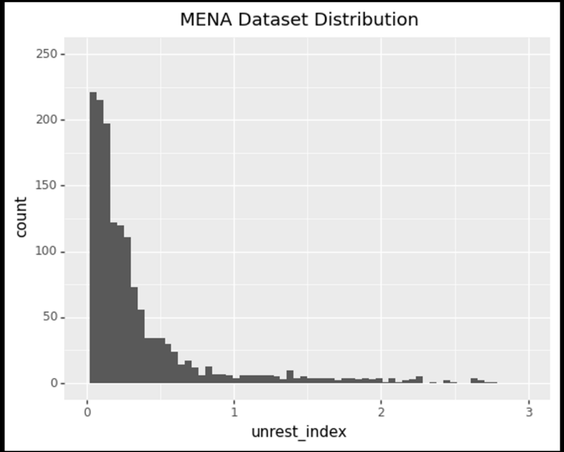
We create an ‘unrest index’ from data…

- CNTSD (sum: 370)
  - Assassinations (25)
  - Strikes (20)
  - Guerrilla Terrorism/Guerrilla Warfare (100)
  - Government Crises (20)
  - Purges (20)
  - Riots (25)
  - Revolutions (150)
  - Anti-Government Demonstrations (10)
- ACLED (sum: 290)
  - Battles (150) — match to “revolutions” from CNTSD
  - Explosions/remote violence (75) — match to “guerrilla terrorism/guerrilla warfare”, but discounting for only instances of remote violence from CNTSD
  - Violence against civilians (25) — match to “assassinations” from CNTSD
  - Protests (20) — match to “strikes/purges” from CNTSD
  - Riots (25) — match to “riots” from CNTSD
- RSUI

…and scale accordingly

- Standardize ACLED, RSUI, CNTSD data to mean 0 and standard deviation 1
- Subtract each data point in the ACLED, RSUI, CNTSD by **the minimum value of the dataset** to make the minimum values of each dataset equal 0
- Take the **average** of ACLED, RSUI, CNTSD for final **unrest_index** dependent variable, separating across World (yearly) and Middle East/North Africa (MENA) (monthly)





Data cleaning methodology

- All data combined into final merged dataset
- **Two merged datasets:** MENA (monthly), World (yearly)
- N/A values: interpolation or removal
- Google Trends data used for testing whether online search activity predicted revolution
  - Range: 0–100 for search activity, indexed to maximum search activity across search space
    - Dataset many “<1” values, which were simplified to 0.5
  - Data is monthly
    - → for the yearly worldwide dataset, take a yearly average

Final dataset sample

Unrest index

Google Trends data

date	Country	country_cod	protest	revolution	riots	strike	unrest	violence	BN.CAB.XOK	BX.KLT.DINV	EG.CFT.ACCS	EG.ELC.ACCS	EN.ATM.CO2	EN.POP.DNS	EN.POP.SLUI	EN.URB.LCTY	EN.URB.MCT	unrest_index
2016-06-01/	Algeria	DZA	0	5	0	26	1	3	-16.289765	-0.3240121	99.5	99.1866608	0.33628389	16.6026256	24.98121	2592330	6.55569862	0.07034985
2016-07-01/	Algeria	DZA	0.5	4	0	20	0	5	-16.296788	-0.211703	99.5083333	99.2002932	0.33517861	16.6304825	24.65576	2595105.67	6.55179696	0.11645936
2016-08-01/	Algeria	DZA	0	4	0	20	0	4	-16.30381	-0.099394	99.5166667	99.2139257	0.33407334	16.6583394	24.33031	2597881.33	6.5478953	0.06032861
2016-09-01/	Algeria	DZA	0	3	0	15	0	3	-16.310832	0.01291503	99.525	99.2275581	0.33296806	16.6861963	24.00486	2600657	6.54399364	0.27585623
2016-10-01/	Algeria	DZA	0.5	4	0	13	0	5	-16.317854	0.12522407	99.5333333	99.2411906	0.33186279	16.7140532	23.67941	2603432.67	6.54009198	0.12500542
2016-11-01/	Algeria	DZA	0	6	1	14	0	11	-16.324877	0.23753311	99.5416667	99.254823	0.33075751	16.74191	23.35396	2606208.33	6.53619033	0.07083475
2016-12-01/	Algeria	DZA	0	5	1	17	0	7	-16.331899	0.34984215	99.55	99.2684555	0.32965224	16.7697669	23.02851	2608984	6.53228867	0.21912433
2017-01-01/	Algeria	DZA	0	6	0	14	0	8	-16.338921	0.46215119	99.5583333	99.282088	0.32854696	16.7976238	22.70306	2611759.67	6.52838701	0.24900066
2017-02-01/	Algeria	DZA	0.5	4	0	12	0	13	-16.345944	0.57446023	99.5666667	99.2957204	0.32744169	16.8254807	22.37761	2614535.33	6.52448535	0.088294
2017-03-01/	Algeria	DZA	0.5	5	0	17	0	7	-16.352966	0.68676927	99.575	99.3093529	0.32633641	16.8533376	22.05216	2617311	6.52058369	0.0243829
2017-04-01/	Algeria	DZA	0	5	0	12	1	8	-16.359988	0.79907831	99.5833333	99.3229853	0.32523114	16.8811945	21.72671	2620086.67	6.51668203	0.07257049
2017-05-01/	Algeria	DZA	0	4	0	16	0	5	-16.36701	0.91138735	99.5916667	99.3366178	0.32412586	16.9090514	21.40126	2622862.33	6.51278038	0.0991274
2017-06-01/	Algeria	DZA	0	5	0	21	0.5	4	-16.374033	1.02369639	99.6	99.3502502	0.32302059	16.9369083	21.07581	2625638	6.50887872	0.12980563
2017-07-01/	Algeria	DZA	0.5	4	1	16	0	4	-16.090654	0.99866001	99.6	99.3740203	0.32326351	16.9648017	21.07581	2628449.25	6.50520095	0.08285458
2017-08-01/	Algeria	DZA	0.5	4	0	16	1	3	-15.807275	0.97362363	99.6	99.3977903	0.32350644	16.992695	21.07581	2631260.5	6.50152318	0.19034912
2017-09-01/	Algeria	DZA	0	4	0	12	0	4	-15.523897	0.94858726	99.6	99.4215603	0.32374936	17.0205884	21.07581	2634071.75	6.49784541	0.36875065
2017-10-01/	Algeria	DZA	0	6	0	11	0	6	-15.240518	0.92355088	99.6	99.4453303	0.32399229	17.0484818	21.07581	2636883	6.49416764	0.20530514
2017-11-01/	Algeria	DZA	0	7	0	11	0	9	-14.95714	0.89851451	99.6	99.4691003	0.32423521	17.0763751	21.07581	2639694.25	6.49048987	0.08156994

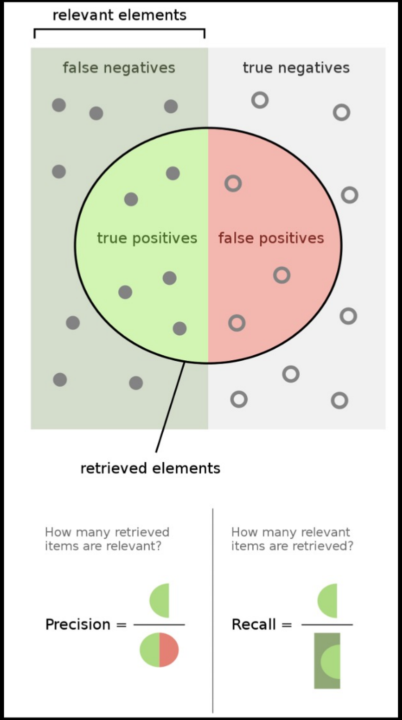
Economic data with World Bank code for column names

We use two main approaches for analysis: linear and logistic regression

- **Linear regression:** continuous $unrest\_index$ variable
  - Strength: easy to apply
  - Weakness: ind. variables are not independent of each other → no possibility of causal inference
- **Logistic regression:** binary variable (0 or 1 — yes or no conflict) prediction
  - Strength: could be a simpler approach than using linear $unrest\_index$?
  - Weakness: simplification into binary variable required
- Run with **regularization** technique
  - avoids the “overfitting” problem
- **Overall:** too many problems to be useful

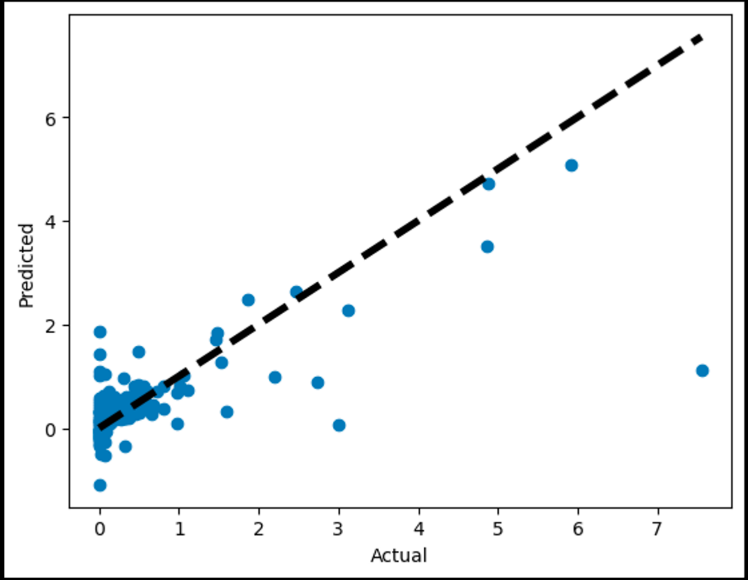
Aside: Logistic regression overview

- Logistic regression solves the binary classification problem
- Success or failure of logistic regression?
  - → use F1-score
- $F_1 = \frac{precision \times recall}{precision + recall}$



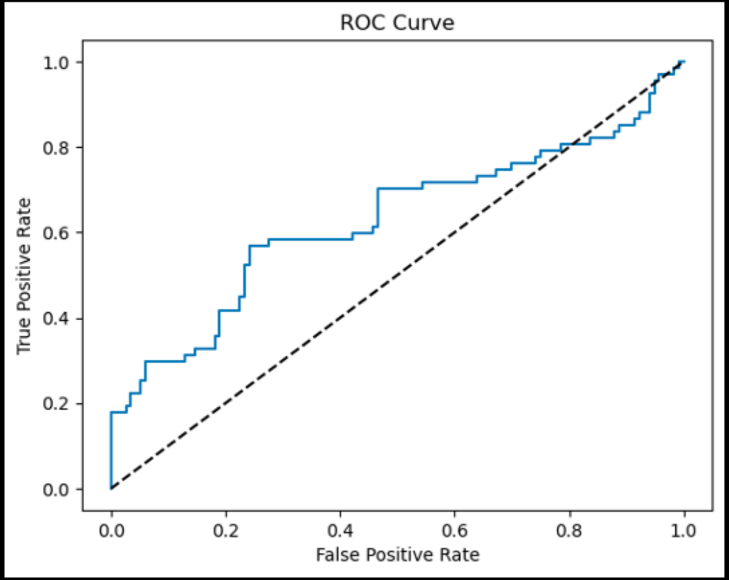
Results: MENA (linear regression model)

- $R^2_{train} = 0.61$
- $R^2_{test} = 0.48$
- Statistically significant variables:
  - School enrollment (+'ve)
  - CO2 emissions (–'ve)
  - Exchange rate (+'ve)



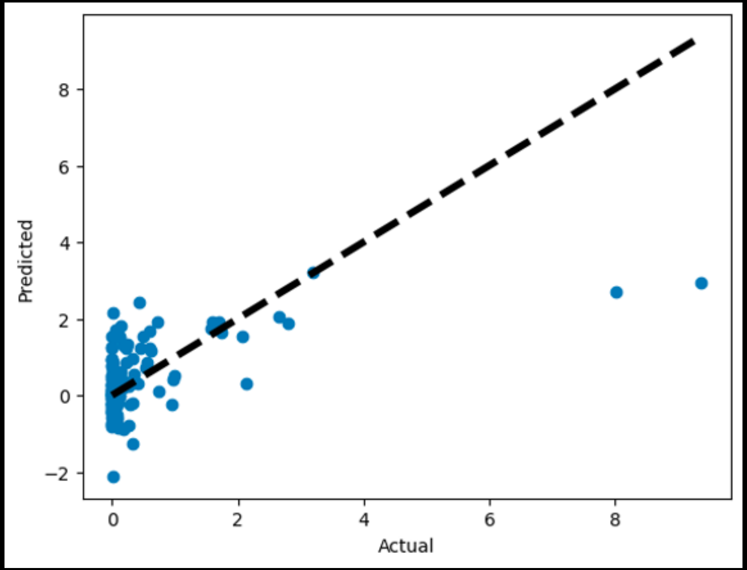
Results: MENA (logistic regression model)

- $F_1 = 0.42$ (out of 1)
- Statistically significant variables:
  - GDP (+'ve)
  - Exchange rates (+'ve)
  - "riots" Google search term (+'ve)



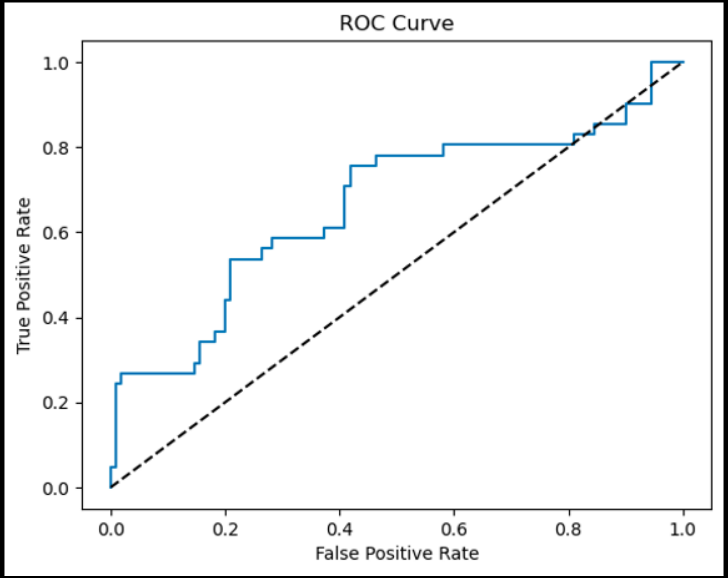
Results: World (linear regression model)

- $R^2_{train} = 0.46$
- $R^2_{test} = 0.16$
- Statistically significant variables:
  - School enrollment (–'ve)
  - Income share held by lowest 10% (–'ve)

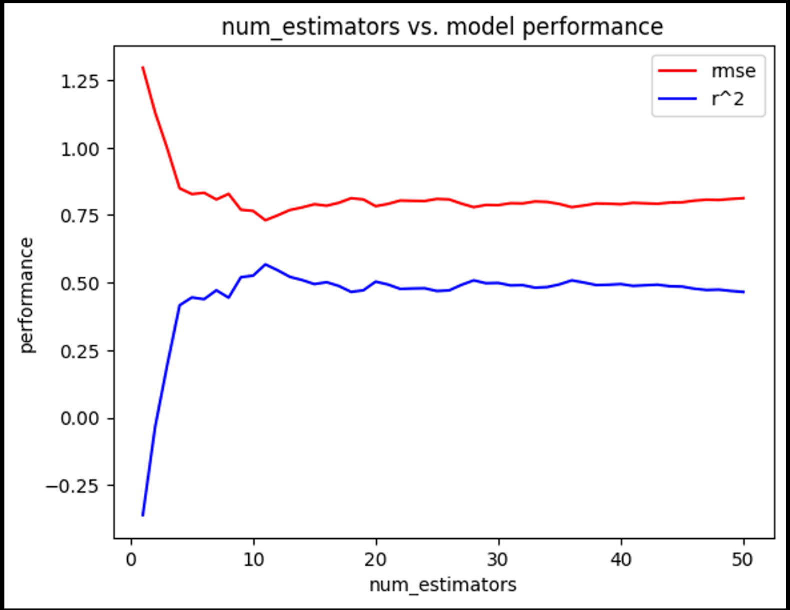


Results: World (logistic regression model)

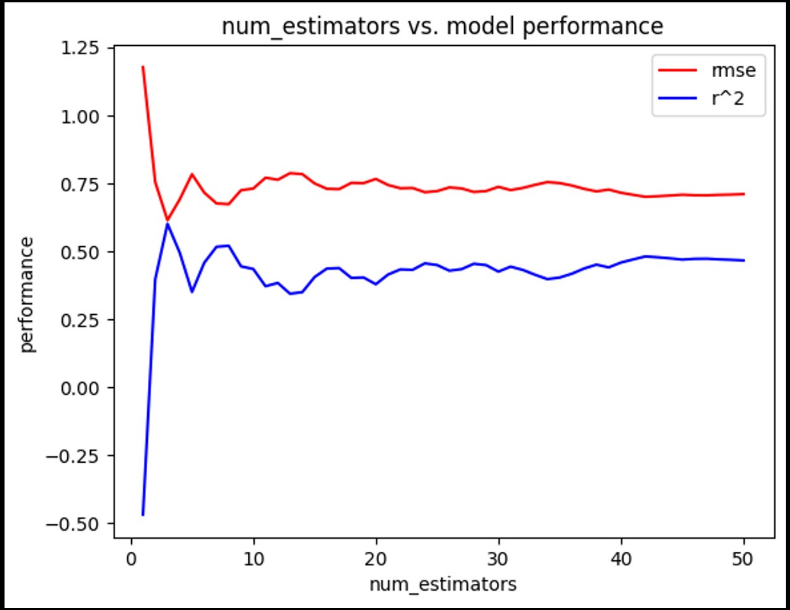
- $F_1 = 0.14$ (out of 1)
- Statistically significant variables:
  - GDP (+'ve)
  - Current account balance (+'ve)
  - CO2 emissions (–'ve)



We also use random forest models and optimise # of estimators (decision trees)



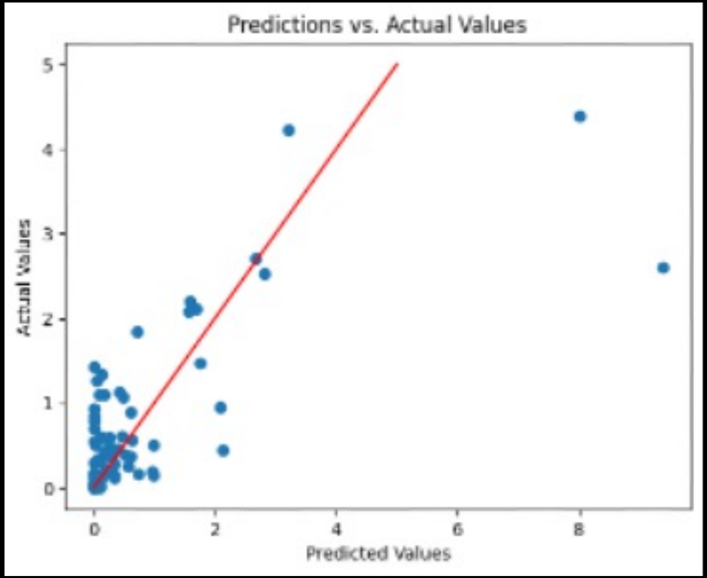
World Data



MENA Data

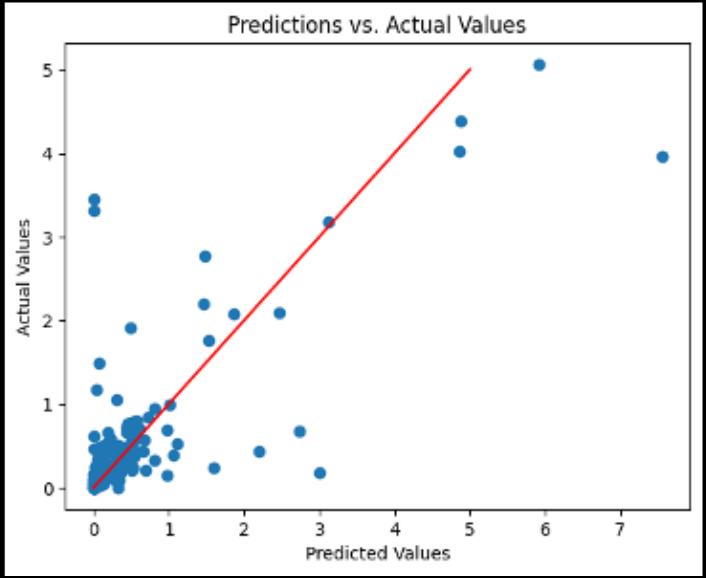
Results: World (single-run random forest)

- $R^2 = 0.56$ (out of 1)
- "Important" variables (sig. > 0.05)
  - Tax revenue (% of GDP)
  - Population of largest city
  - Prevalence of moderate or severe food insecurity



Results: MENA (single-run random forest)

- $R^2 = 0.60$ (out of 1)
- "Important" variables (sig. > 0.05)
  - GDP
  - Current health expenditure per capita
  - GDP per capita
  - "Revolution" Google search term
  - Population living in slums (% of urban population)



Results: World (long-run random forest)

- For each random state:
  - (1) find optimal number of estimators given by $num\_estimators$
  - (2) run model, taking mean root mean-squared error $rmse$, R^2, and important features
  - (3) repeat over 100 trials
  - (4) take mean $rmse$, R^2, important features
- $\boldsymbol{R^2 = 0.58, rmse = 0.67}$
- ”Important” variables (sig. > 0.05)
  - GDP

Results: MENA (long-run random forest)

- For each random state: **(same as for world dataset)**
  - (1) find optimal number of estimators given by $num\_estimators$
  - (2) run model, taking mean root mean-squared error $rmse$, R^2, and important features
  - (3) repeat over 100 trials
  - (4) take mean $rmse$, R^2, important features
- $\boldsymbol{R^2 = 0.64, rmse = 0.52}$
- ”Important” variables (sig. > 0.05)
  - GDP
  - GDP per capita
  - “revolution” Google search term
  - Population in urban agglomerations of more than 1 million

For **prediction**, we attempt to use neural network models

- We built a neural network in keras (Python) for predicting social unrest using our datasets
  - Used 3 hidden layers with a “dropout” after each hidden layer
  - Trained on training set of 80% of the dataframes (test set: 20%)
- Of course, **no easy way of finding important determinants**!
- **Overall results**
  - World $R^2_{train} = 0.77$, $R^2_{test} = 0.45$
  - MENA $R^2_{train} = 0.74$, $R^2_{test} = 0.23$

A brief comparison of model performance

Model	R_{test}^2
Linear (MENA)	0.48
Linear (World)	0.16
Decision Tree (MENA)	**0.64**
Decision Tree (World)	**0.58**
Neural Network (MENA)	0.22
Neural Network (World)	0.45

Logistic model for binary classification uses F_1 statistic: $F_1 = 0.14$ (world), $F_1 = 0.42$ (MENA)
*higher F_1 = more predictive power

Actionable insights

- **Expected:** GDP, price level (inflation) associated with an increase in social unrest; urban density associated with social unrest
- **Unexpected:** School enrollment — association unclear as both positive and negative effects from regression models
  - one would expect more education = more social unrest?
- **Interesting:** increased CO_2 emissions associated with a decrease in social unrest

Discussion

- The project was **much** more challenging than expected
- Many missing variables: we have approx. 60 independent variables, compared to approx. 340 independent variables used in IMF study (Redl & Hlatshwayo 2021)
- Usage of big data? → Google data did occasionally seem quite significant; some search terms have predictive power
- Use cases of this research? → More **verification** than prediction