Supervised Machine Learning:

Classification

Predicting Outcomes of Coups d'état with SPEED Dataset

by Vitaly Vigasin

Overview

This analysis is based on the data available from The Social, Political and Economic Event Database (SPEED) available for download from the website of the Cline Center For Advanced Social Research of University of Illinois:

"SPEED is a technology-intensive effort to extract event data from a global archive of news reports covering the Post WWII era. It is designed to provide insights into key behavioral patterns and relationships that are valid across countries and over time. Within SPEED, event data is generated by human analysts using a suite of sophisticated tools to implement carefully structured and pretested protocols. These protocols are category-specific electronic documents that are tailored to the information needs of a particular category of events (civil unrest, property rights, electoral processes, etc.). SPEED data will produce insights that complement those generated by other components of the SID project (constitutional data, archival data, survey-based data, etc.) because event data generates "bottom-up" observations from news reports. In

generating these event data SPEED leverages tens of billions of dollars that have been invested in compiling news reports from throughout the world."

Source: https://clinecenter.illinois.edu/project/human-loop-event-data-projects/SPEED

The dataset used in this analysis consists of **62,141** records corresponding to news articles and information bulletins automatically scraped from 4 sources, namely *New York Times*, *Wall Street Journal, Foreign Broadcast Information Service (CIA)*, and *Summary of World Broadcasts (BBC)*.

The records contain information about various events throughout the world from 1946 to 2005 as recognized and classified by an automated process (called PETRARCH) along with corresponding metadata in a total of **106** columns.

Among this tremendous amount of information lies data on **747** *coup d'état* events (as recorded in the dataset) that happened in the world during the aforementioned period. That very data were the focus of this analysis as described below.

More on the PETRARCH software and the The Computational Event Data System used for obtaining and coding the dataset can be obtained from the following sources:

http://eventdata.parusanalytics.com/

https://github.com/openeventdata/petrarch2

https://science.sciencemag.org/content/353/6307/1502

Goal

The goal of this research was an attempt to build a model capable of predicting the outcome of a coup d'etat based on the historical data.

As the dataset lists all *failed* coups d'etat marked with a value in a special column, that column was used as the target for the classification model.

Data Preparation and Analysis

The cleaning phase consisted mostly of identifying and removing columns that contain metadata irrelevant for the purpose of this research. The SPEED Codebook as well as an auxiliary MS Excel file were consulted in order to establish the meaning of each column, and an additional dataframe was created as partially shown below for an easy reference:

[62]:
```
df_codebook
```

[62]:

Variable	Variable Description
AD_TACT	Tactics advocated in political exp
AD_VIOL	Were violent acts advocated?
AID	Article id#
AMBIG_INI	Ambiguous initiator/target distinction
AMBIG_WGT	Weight to adjust for double counting initiators
ANTI_GOV_SENTMNTS	Event rooted in anti-government sentiments?
ARRESTS	Were arrests made?
ATK_TYPE	Type of political motivated attack
CLASS_CONFLICT	Event rooted in class-based conflict?
CODE_DAY	Day of coding
CODE_MONTH	Month of coding
CODE_YEAR	Year of coding
COUNTRY	Country in which the event occurred
COUP	Did this event involve a coup?
COUP_FAILED	Did this event involve an unrealized coup?
COWCODE	Cowcode of country where event occurred
DAM_PROP	Was property damaged in the event?
DATE_TYP	Type of date information available
DAY	Day of event - multi-day=average
DAY_SPAN	Longest span of event, in days
DSA_TYPE	Broad category of disruptive state act
E_LENGTH	Length of event, truncated
ECO_SCARCITY	Event rooted in ecological resource scarcities
EV_TYPE	Event type
EVENT	Is this coding for a destabilizing event?
EVENTID	Event identification number
EXP_TYPE	Reduced political expression type
FROM_EID	Id of event linked from
G_LVL_I	Level of government initiator
G_LVL_T	Level of government target
G_LVL_V	Level of government victim
GOV_I1	Type of government initiator, 1
GOV_T1	Type of government target
GOV_V1	Type of government victim
GP3	Country name(caps and lower case)
GP4	Lowest level entity name (caps and lower case)
GP7	Latitude
GP8	Longitude
GP_TYPE	Type of geo-political entity targeted

In total, **48** columns were eliminated, of which some were an easy target as they contained metadata such as news source, date of the historical event, geographical coordinates, event identifiers etc. while others were rejected after an prolonged deliberation.

Then the original dataset was filtered so that it only contains rows containing coup d'etat events along with remaining **58** columns.

The resulting dataset contained **747** coup d'etat events, out of which **370** were marked as unsuccessful attempts by the coding software.

The records named a total of **128** world states were the coups d'etats occurred, however, a look at the list of the countries revealed potential errors as among such countries the US, the United Kingdom, Israel, Japan, and even Latvia and Finland were listed, clearly contradicting the known historical reality.

As was confirmed in a very courteous and informative email reply from a representative of the Cline Center for Advanced Social Research, those were indeed errors that resulted from the automated nature of data collection and classification process.

As a result, the following obvious records were further eliminated from the dataset: 'United States', 'United Kingdom', 'Ukraine', 'Latvia', 'Finland', 'Japan', 'Israel', 'Italy'.

However, some of the erroneous records may still be present somewhat biasing the dataset. As just one example, even though Russia (or rather former Soviet Union) indeed experienced a coup d'etat event in 1991, there is no known such event that happened in 1946 in that country but nevertheless reported by the dataset. The validity of the recorded events in other countries is not at all a given either. Nevertheless, as the main goal of this exercise is the every possibility of finding a combination of independent variables that may have the power of prediction in terms of this research, the resulting dataset now truncated to the size of **722** rows was accepted as final with

Finally, the missing values throughout the dataset were replaced with zeros.

Feature Selection

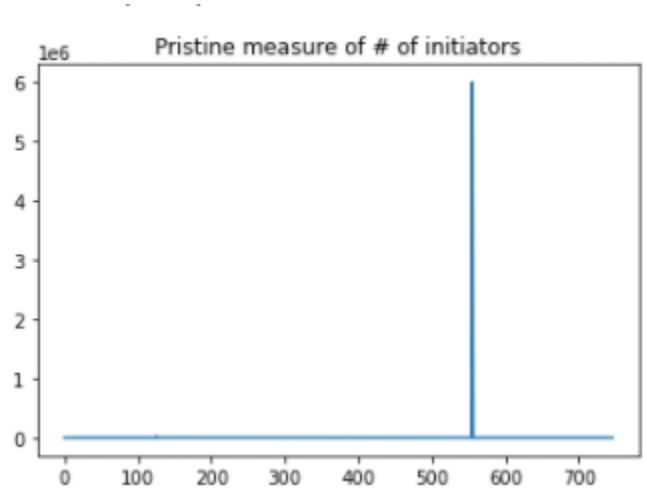
The feature selection stage began from identifying columns that contained no data (in the subset of the original dataset that has now been reduced to only coups d'etat-related records). A total of **17** columns were identified as such and removed from the dataset.

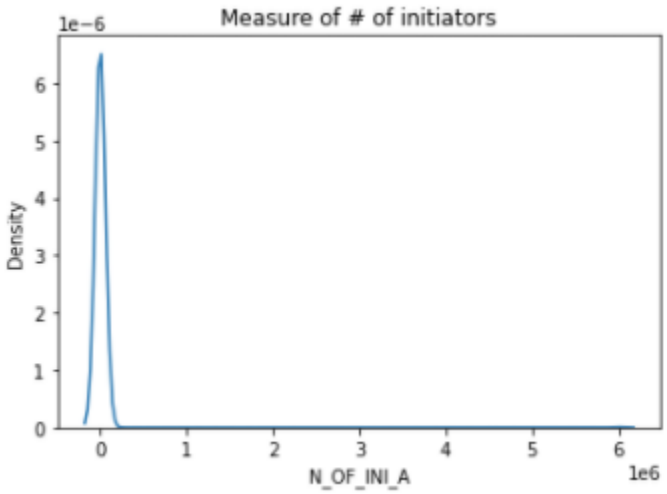
Further, columns denoting geographical locations (country and region) were also removed, as we should assume that even if some world regions may be prone to violent changes of power more than the others, the geographical location itself has nothing to do with the *chances of success* of such an event.

Further, the *'Event rooted in anti-government sentiments?'* parameter was also ignored as it turned out to be of a single value ('yes') for all of the records.

Next the *'Longest span of event, in days'* column was removed from consideration as redundant to another parameter called 'Length of event'.

Some parameters demonstrated extremely narrow frequency distribution, as shown below, and were eliminated as a result:





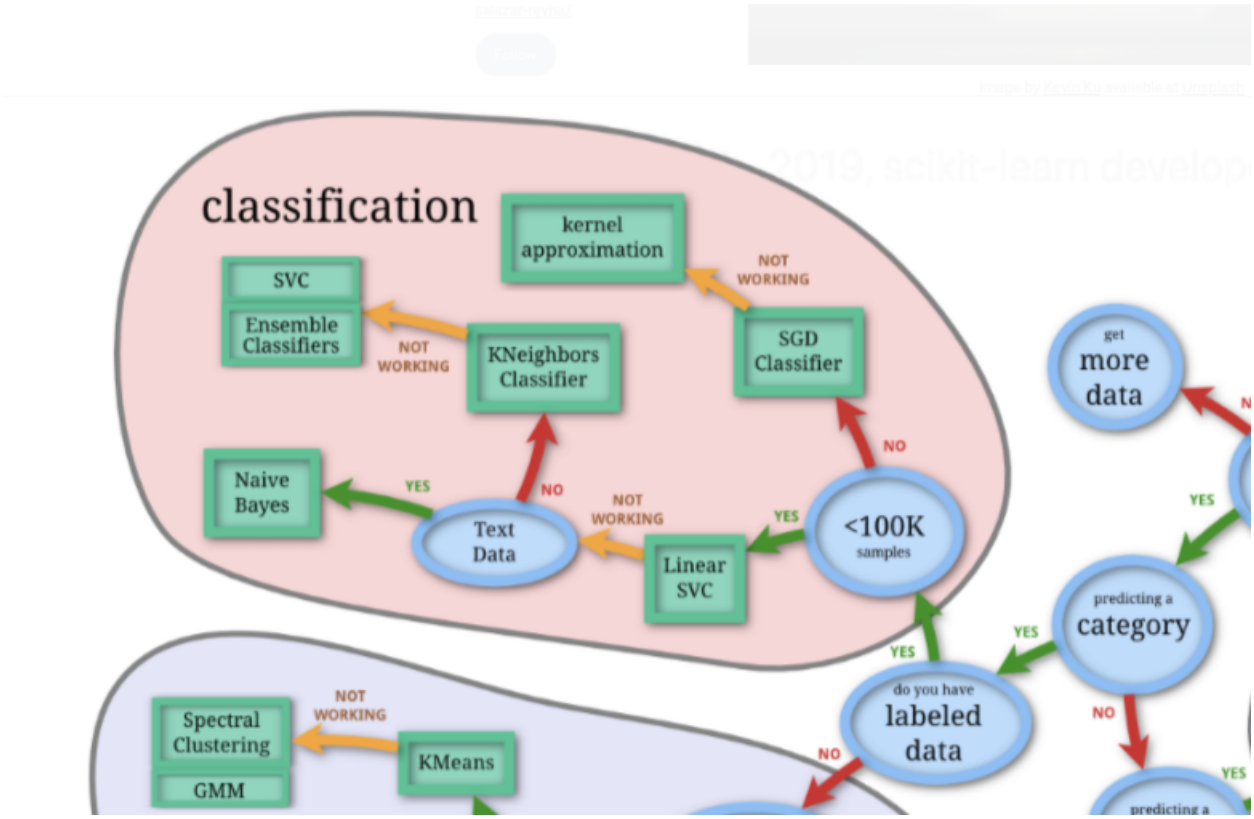
Finally, the *'Type of geo-political entity targeted'* parameter was also ignored as it was the same (entire nation) for all of the events in question.

At this stage, **30** features list of which is shown on the next page remained available for selection:

Variable	Variable Description
AD_TACT	Tactics advocated in political exp
ARRESTS	Were arrests made?
CLASS_CONFLICT	Event rooted in class-based conflict?
E_LENGTH	Length of event, truncated
ECO_SCARCITY	Event rooted in ecological resource scarcities
G_LVL_I	Level of government initiator
G_LVL_T	Level of government target
G_LVL_V	Level of government victim
GOV_I1	Type of government initiator, 1
GOV_T1	Type of government target
GOV_V1	Type of government victim
HUMAN_T1	Type of human target
HUMAN_V1	Type of human victim
INI_TYPE	Type of initiator
LINK_TYPE	Type of event link
LINKED	Is this a linked coding?
NGOV_I1	Type of non-government initiators
PERS_SECURITY	Event rooted in desire for personal security?
POL_DESIRES	Event rooted in desire for political rights?
POSTHOC	Post hoc reaction?
PUB_ORDER	Event rooted in desire to maintain pub order?
RETAIN_POWER	Event rooted in desire to retain political power?
RETRIBUTION	Event rooted in desire for retribution?
SC_ANIMOSITY	Event rooted in socio-cultural animosities?
TAR_GPOL	Was the target a geo-political entity
TAR_TYPE	Type of target
VIC_TYPE	Type of entity victimized
VICTIM_EFFECT	Impact of event on victim
WEAP_GRD	5 category weapon variable
WEAPON	Type of weapon used

Model Selection and Evaluation

The well-known diagram from the *'Choosing the right estimator'* page on the official SciKit-Learn website was the starting point of choosing a model, with LinearSVC and KNeighborsClassifier being the main contestants.



Source: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

However, 3 additional ensemble algorithms were added to the competition mainly based on the fact that they were the one covered in the 'Supervised Learning: Classification' course as part of the IBM Machine Learning Professional Certificate studies.

The complete list of the algorithms tried and their parameters is as follows:

```
estimators=[
    ('LinearSVC', LinearSVC(max_iter=10000)),
    ('KNeighborsClassifier', KNeighborsClassifier(n_neighbors=5)),
    ('RandomForest', RandomForestClassifier(n_estimators=n_estimators, random_state=42) ),
    ('AdaBoost', AdaBoostClassifier(n_estimators=n_estimators, random_state=0) ),
    ('Bagging', BaggingClassifier(base_estimator=SVC(), n_estimators=n_estimators,
random_state=0))
]
```

Additionally, StackingClassifier was run combining the effects of all of the aforementioned classifiers with LogisticalRegression as the final estimator.

Remarkably, all of the models demonstrated similar scores with varying parameters such number of trees with KNeighborsClassifier showing consistently worse performance with varying number of neighbors:

estimator	score
LinearSVC	0.645329
KNeighborsClassifier	0.574394
RandomForest	0.633218
AdaBoost	0.655709
Bagging	0.634948
Stacking	0.660900

At the same time, the stacking estimator consistently produced better results.

As the result, KNeighborsClassifier was retired and the score was somewhat improved:

estimator	score
LinearSVC	0.645329
RandomForest	0.612457
AdaBoost	0.657439
Bagging	0.636678
Stacking	0.664360

By 'score' the output of the standard *score()* method of each classifier is meant here.

To compare, the accuracy scores produced by the *accuracy_score()* function in the sklearn.metrics package were also tried, with the following results:

estimator	score
LinearSVC	0.685596
RandomForest	0.673130
AdaBoost	0.674515
Bagging	0.653740
Stacking	0.695291

As a final step of this stage, various combinations of features selected from the list of total 30 independent variables selected during the Feature Selection stage were tried with all of the models described above. None of the (random) combinations of 5 to 10 features produced a better result so eventually the whole set of available features was selected as the final input to the model.

Final Model

The following final configuration was selected as a result of the trials described in the previous paragraph:

Transformer: **OneHotEncode**r

Estimator: **StackingClassifier** (LinearSVC + RandomForest + AdaBoost + Bagging with LogisticRegression as final)

Highest accuracy score achieved on full ground truth : **0.695291**

Key Findings

The main outcome of this research is that it seems likely to predict an outcome of such a relatively rare, violent and unusual event as a coup d'etat based on the available historical data.

The power of prediction, however, is not very high and leaves rooms for hopes for an improvement.

Going Forward

There are two obstacles for any new attempt to achieve a better result:

- Quality of the existing data
- Availability of *relevant* data

The first problem is well understood, although the extent of it is not known. As has already been pointed out, the underlying data has been proven to be not free of errors, and the potential impact of some of the most obvious ones was mitigated by removing the corresponding rows. However, there is no guarantee that the remaining data is correct. Confirming that would require a tremendous amount of work, if at all possible.

The second problem has to do with the very fact that we do not know what factors in reality have the greatest impact on the possibility to predict such an undoubtedly complex event as a coup d'etat. When it comes to predictions in the field of political science, where we deal with such material as human behaviour and tremendous complexity of human societies, predicting just about *anything* is an absolutely daunting task.

This analysis has been based on available data collected from open sources such as information releases and news articles. We can be quite sure that open sources will never be sufficient for any serious political analysis (which leaves plenty of room for job security of the members of the clandestine intelligence community for the foreseeable future).

But even if the open data will remain our only option in terms of this research, it remains to be seen whether the parameters that can be chosen as the independent variables from the SPEED dataset in its current form are indeed the ones that we may call good predictors.

Nevertheless, this exercise served as a good proof of concept with moderately convincing results. There may be more exciting time repeating it when and if better data become available.