CS209A FINAL PROJECT

COUP D'ETATS PROJECT

Authors:

Adriana Trejo-Sheu

Ana Vitoria

Reem Al-Khalqi

Introduction to Data Science

Harvard Universty

December 2021

Contents

1 Introduction

Original Problem statement: Coup d'états are the sudden overthrowing or transfer of political power of a country's leader and it is an event that has consequences for a country's well-being. In this exploratory data analysis, we use the data set provided by the Cline Center for Advanced Social Research from the University of Illinois Urbana-Champaign to analyze recorded characteristics of the coup and the fate of the leader. We are mainly interested in understanding what constitutes a successful coup.

We will begin our project by analyzing the data we have and creating a baseline model. This will help us re-frame the problem statement to better meet the goals of this project and see what potential directions of interpretations and methodologies that we can go with the given data.

2 EDA Analysis

We will be doing a series of EDA analysis to explore the content of the dataset and this will help minimize our predictor list, do model selection, and determine how to extend past our baseline model chosen. We did not have to impute any missing data and we are able to take the dataset as is as we do our EDA. Our predictors are mostly categorical data that are nominal. We do have a couple of ordinal predictors such as year, month, and day. There is an associated Juptyer notebook that walks through our steps with comments and descriptions of what we found. We will summarize our main and most important findings here.

2.1 Cowcodes vs Country Names

The data provides both the cowcodes and the country name. The cowcodes is a unique country code number based on the Correlates of War (COW) country code list. The data lists 136 countries and 134 cowcodes. When inspected, we found that there were a couple of countries where each was listed twice as two different countries but share the same cowcode. These countries were Cote d'Ivoire and Ivory coast (cowcode 437) and Kyrgyzstan and Kyrgyz Republic (cowcode 703). Essentially, these are the same country, simply imputed with different names. Therefore, we actually have 134 countries and not 136. Hence we will keep cowcodes as variable to indicate the country.

2.2 Realized vs Unrealized Coups

From the Cline Center website, a coup is labeled as "organized efforts to effect sudden and irregular (e.g. illegal or extra-legal) removal of the incumbent executive authority of a national government, or to displace the authority of the highest levels of one or more branches of the government." A coup is considered to be realized or unrealized. A realized coup would be a success in the removal of the incumbents or the removal of the incumbents ability to control the state. An unrealized coup is considered to be a conspiracy or an attempted coup, meaning an unsuccessful removal of an incumbent's power.

Note: We will be using realized coup or successful coup interchangeably to mean a success. A coup technically includes conspiracies and attempts but these will be described as either unrealized, unsuccessful or as they are mentioned. Lastly, the event_type column from the dataset notes a successful or realized coup as 'coup'.

2.3 Choosing the Response Variable

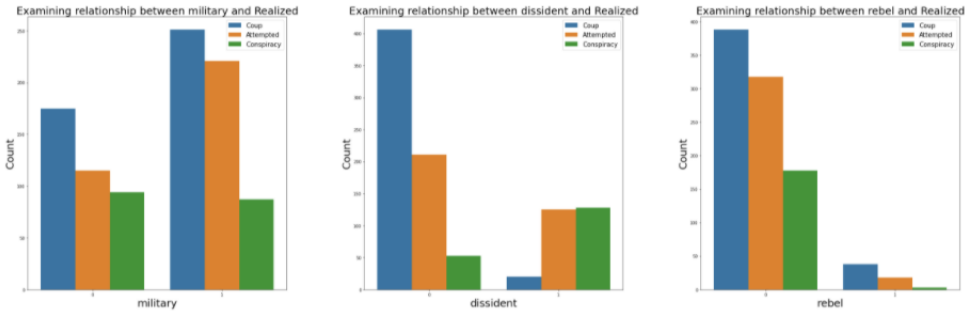
The main goal of this project is to be able to predict a successful coup. Our response variable should signify whether or not a coup was realized or unrealized. From the dataset given, event_type, realized and unrealized all describe the success or failure of a coup. The event_type variable gives three outcomes (coup, attempt, conspiracy), which we can use to determine the success. The realized column gives an array of ones and zeros with one indicating a successful coup and zero indicating a failure (attempt or conspiracy). The unrealized column is the exact opposite of the realized column and gives a one to failed coups (attempts, conspiracies).

From assessing these three different columns, we plan to use the realized column as our response predictor. The event_type can be used to supplement that column for understanding deeper what might have caused failures as well as for graphing purposes.

2.4 Preliminary selection of predictors through visualization

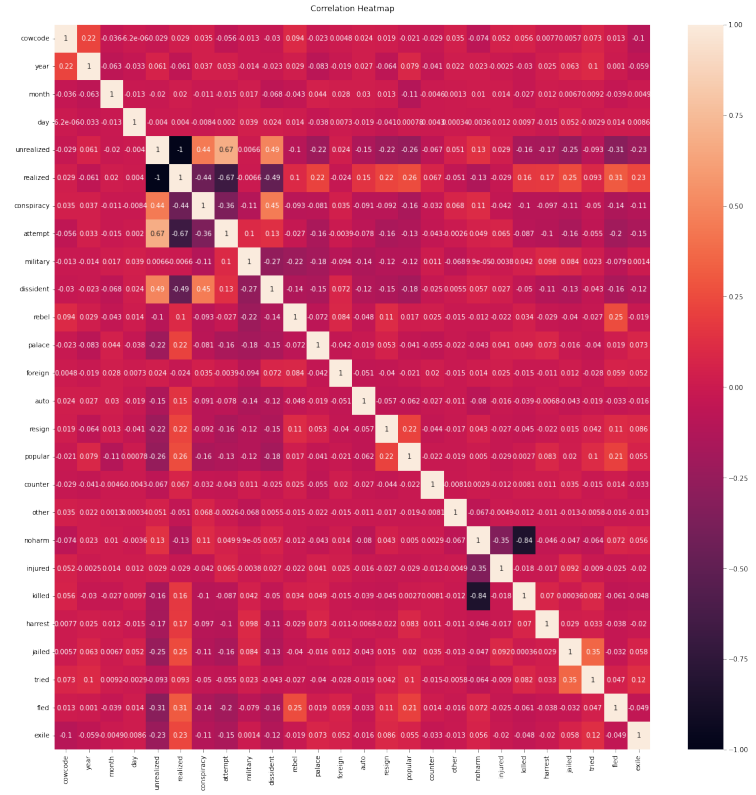
In our initial exploration, we looked at how each predictor impacted the outcome of our response variable, realized. In this case we plotted each predictor that was identified against event_type rather than realized. This decision was in order to get a deeper understanding as to how each predictor affected each of the potential coup outcomes. It is an important distinction that we will use later on in the project when assessing the impact or influences of predictors. An example of a sample of our plots are shown below:

Figure 1: Sample of Subplots showing each Predictor against Event Type



Moreover, we looked at the collinearity of all predictors in the dataset to better identify highly correlated variables.

Figure 2:Collinearity heatmap of all predictors in Coup d'états dataset



We can see from the collinearity heatmap that realized and unrealized have a collinearity of -1 which supports our decision in using one as our response variable and removing the other from the predictors list. Moreover, we can see that, unsurprisingly, noharm and killed have a collinearity of -0.84. The reason that they don't have a -1 collinearity is because there are other fates considered for the leader including home arrest, jailed, exiled and fled. There are several other predictors that have a noticeably high collinearity and being able to see the collineartiy between all these predictors will allow us to streamline the feature selection process.

Some of our initial findings from these visuals were that certain predictors (auto, resign, popular, kill, exile, injured, noharm, harrest, counter, other, jailed, tried, fled) were seen to have an effect on whether a coup was successful or not. This was seen in a group of predictors that were further analyzed through lasso and by taking the mean. We took the mean from a dataframe that only had successful coups or realized marked as 1 to see if there was an overwhelming presence of these variables in the sub-success dataframe. There were some values that were high (meaning they were above 0.2), such as noharm and killed.

Due to definitions that explicitly talk about what happens to the executive during the coup along with the preliminary plots, we identified a list of predictors to look at more in-depth. These include: auto, resign, popular, counter, other, noharm, injured, killed, harrest,jailed, tried fled, exile. In our numerical analysis, we saw that noharm was present in 94% of columns, so we decided to remove it. We also saw that that at least 1 of the listed predictors above were present in 70% of successful coups and only 11% in unsuccessful ones. From these high values relating to the successful coup versus the low ones in unsuccessful, we decided to remove them from our baseline model. We still however performed an initial model with these to see what their coefficient values would be. We also performed a lasso penalization to see if lasso removed them.

2.5 Use of Lasso for Predictor selection

We perform cross-validation to pick the best lambda across 10 values between 0.00001 and 10000. The best lambda is 0.001. After choosing this as best hyper-paramenter, we perform Lasso to see which variables it suggests us to drop.

If we feed all variables (with the exception of: country, coup_id, event_type, unrealized, attempt, conspiracy because of collinearity and because country is expressed already by cowcodes) lasso decides to drop the variables rebel and other.

From the code in the jupyter notebook, we notice that the predictors (auto, resign, popular, killed, exile, noharm, harrest, counter, other, tried, fled) have coefficients that are dominating over than remaining variables coefficients (cowcode, year, month, day, military, dissident, palace, foreign) because they could have a greater influence on the prediction of the response. This could potentially support the reasons for dropping them as they might give away the response variable, as it was described earlier in section 2.4.

Hence, for our baseline model we consider to keep only: cowcode, year, month, day, military, dissident, palace, foreign.

2.6 Baseline Model

As a Baseline model we chose Logistic Regression.

Logistic Regression is quite apt for this dataset and for our chosen response variable, realized. Given that our response variable is of a binary type (coup has been realized = 1, the coup has not been realized = 0), using a linear regression is not the ideal method because we have to classify the presence of the response. Linear regression is not bounded by 1 or 0 values and could potentially return values in-between that we cannot use to classify.

With logistic regression, we are modelling the log-odds of the predictors:

$$ln(\frac{P(Y=1)}{1-P(Y=1)}) = \beta_0 + \beta_1 X \tag{1}$$

...where a unit change in X is associated with a e^{β_1} change in the odds of success (Y=1).

Table 1: List of Predictors and their Descriptions

Predictor	**Description**
cowcode	A unique country code number based on the Correlates of War (COW) country code list. It is used to identify the country where a coup event occurred. Please note, these codes are slightly different from
month	Month of the coup event.
day	Day of the coup event.
year	Year of the coup event
military	A dummy variable where one indicates a military coup/attempt/conspiracy and zero otherwise.
dissident	A dummy variable where one indicates a dissident coup/attempt/conspiracy and zero otherwise.
palace	A dummy variable where one indicates a palace coup/attempt/conspiracy and zero otherwise.
foreign	A dummy variable where one indicates a foreign-backed coup/attempt/conspiracy and zero otherwise.

Given this, our baseline model can be described by the following formula:

$ln\frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + \beta_1 cowcode + \beta_2 year + \beta_3 month + \beta_4 day + \beta_5 military + \beta_6 dissident + \beta_7 palace + \beta_8 foreign$

The Logistic Regression performed with the 8 predictors mentioned above yields a test score of 68.55% and a training score of 70%.

This yields the following beta coefficients:

Coefficients	Value	Name of Predictor
β_0	0.01257692	intercept
β_1	-0.0002	cowcode
β_2	0.0004	year
β_3	-0.0072	month
β_4	0.0054	day
β_5	-0.5706	military
β_6	-3.0051	dissident
β_7	1.1339	palace
β_8	-0.0188	foreign

By looking at these coefficients, we can say that:

- if the variable **year** increases, the odd of a success of coup increases as well. Meaning, that the more recently performed in history attempts have the most likelihood of success compared to the ones performed earlier.
- if the variable **day** is closer to 31, the odd of a success of coup increases. Meaning, that the later a coup attempt is performed in a given month, the more likely it is that it succeeds.
- if the variable **month** is closer to 12, the odd of a success of coup decreases. Meaning, that the later a coup attempt is performed in a given year, the less likely it is that it succeeds.
- if the variable **military** is 1, the odd of a success of coup decreases. Meaning, that if the attempt of the coup was organized by the military, i.e. a military attempt, the less likely it is that it succeeds.
- if the variable **dissident** is 1, the odd of a success of coup decreases significantly more than the other predictors. Meaning, that if the attempt of the coup was organized by dissidents, it less likely it is that it succeeds.

- if the variable **palace** is 1, the odd of a success of coup increases. Meaning, that if the attempt of the coup was a palace attempt, it less likely it is that it succeeds.
- if the variable **foreign** is 1, the odd of a success of coup decreases. Meaning, that if the attempt is backed by foreign forces, it is less likely to succeed.

3 New Project Statement

From our EDA, we are able to update the original project statement and discuss what directions and methods we plan to use. While we are still interested in understanding what constitutes a successful coup, we want to look at more than just what makes a coup successful generically. What if there is a difference in what determines a coup post WWII? Would a scenario that made a successful coup in Chile work in Russia? Would it work in South Africa? We have the ability to make predictions of what success means as a whole but also tailor it to more specific scenarios. To accomplish these questions, we plan to improve upon our baseline logistic model by doing further feature analysis with interaction terms, improving our selected model with Random Forest, Bagging, Adaboost models through hyper-tuning and cross-validation, and using additional models, such as ensemble and mixture of experts with different heterogeneous models. The ensemble methods use a combination of simpler learners to provide us with a more precise model that will improve our outcome predictions. Moreover, the mixture of experts is an ensemble technique which uses multiple simple learners to specialize in different parts of the data plus a manger model that will decide which specialist to use for each input data.