Maddi Ross

Final Report:
Foreign Military Training Operations Cost

The Problem

In 2022, defense spending accounted for around 10% of the US budget, amounting to $750 in 2022. In the 2020 fiscal year, over $11 billion was spent on foreign military assistance. Consistent and large-scale foreign military aid has long been a part of US policy, but it is an opaque sphere of government activity. Much information about the direction of military spending is classified, as is further information on work with foreign militaries.

We will attempt to shed light on a certain sector of US-funded foreign military operations using data on unclassified training activities for foreign security forces arranged and funded by the US Department of State and US Department of Defense between 2001 and 2020. The US carries out thousands of training operations per year with foreign military and security forces as a basic and consistent part of US security strategy. We will produce a model capable of predicting the cost of a given foreign military training operation based on various objective features (year, program length, country being trained, number of individuals being trained, training type, location, etc.)

The opacity of defense industry activities means that both policymakers and invested citizens are often ignorant of the details and expected utility of military operations. This model can help both groups make more informed decisions, as in a time of growing concern about the military industrial complex, the ability to predict the cost of such operations can play an important role in deciding government policy or evaluating the necessity of proposed foreign military training operations.

Our data is limited in scope to unclassified foreign military operations, the only operations on which accurate and timely data can be gathered. These are also the only operations which undergo any sort of public review and approval process in any case. Within the scope of our data we would like to further constrain ourselves to those regular military operations that represent a fundamental part of US military and strategic doctrine, ignoring operations that are associated with particular outlier events such as war.

Dataset

The raw data came from governmental reports (as gathered by Security Force Monitor, a project of the Human Rights Institute at Columbia Law School, into a database available here). All foreign military training operations were included as reported from 2001 until 2020. The dataset is continuously updated with further information. Because of Github size limitations, we will not upload our raw, cleaned, or preprocessed data – it can be accessed at the link above.

There was no missing data in the dataset, thanks to the nature of governmental reports. The original 20 features included for the 227,832 operations in the dataset were as follows:

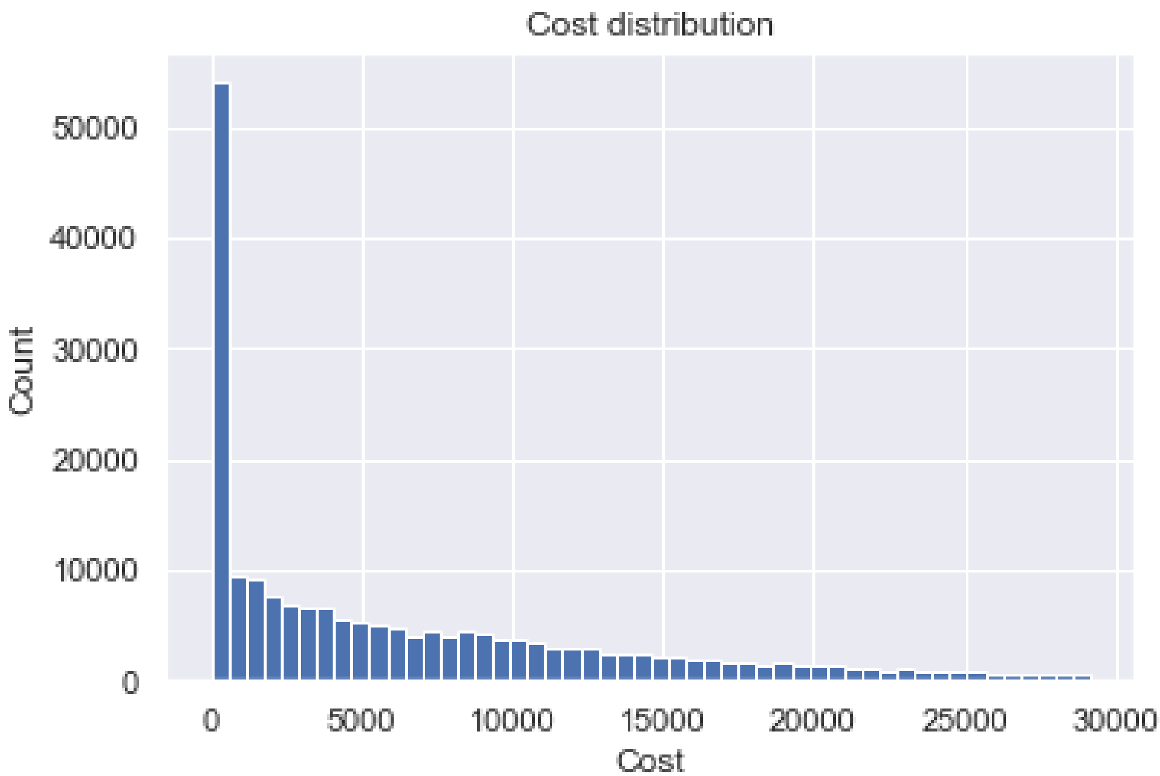
training:id:admin	a unique identifier
training:source	identifier of the document where operation info was published
qa:training_group	an organizational identifier including program region and year
training:country	the country for whom the training program was ordered
training:country_iso_3166_1	country iso identifier for training:country
training:program	the program running the operation (often includes country, region, department, and more)
training:course_title	course title given to operation
training:delivery_unit	the American unit delivering the training, if applicable
training:recipient_unit	the unit receiving American training, if applicable
qa:training_start_date	start date for training operation
training:start_date	end date for training operation
qa:training_end_date	start date for training operation (duplicate)
training:end_date	end date for training operation (duplicate)
training:location	location of training
training:quantity	number of individuals trained
training:total_cost	total cost, USD (target variable)
qa:training_source_url	link to state department document containing this data
training:status:admin	a variable tracking whether or not the data in the dataset has been checked for accuracy against the link from which it was scraped
qa:training_date_first_seen	date when information was first released
qa:training_date_scraped	date when information was scraped, added to database

Data Cleaning, Wrangling, Exploration

We dropped our most extreme cost outliers from the dataset.

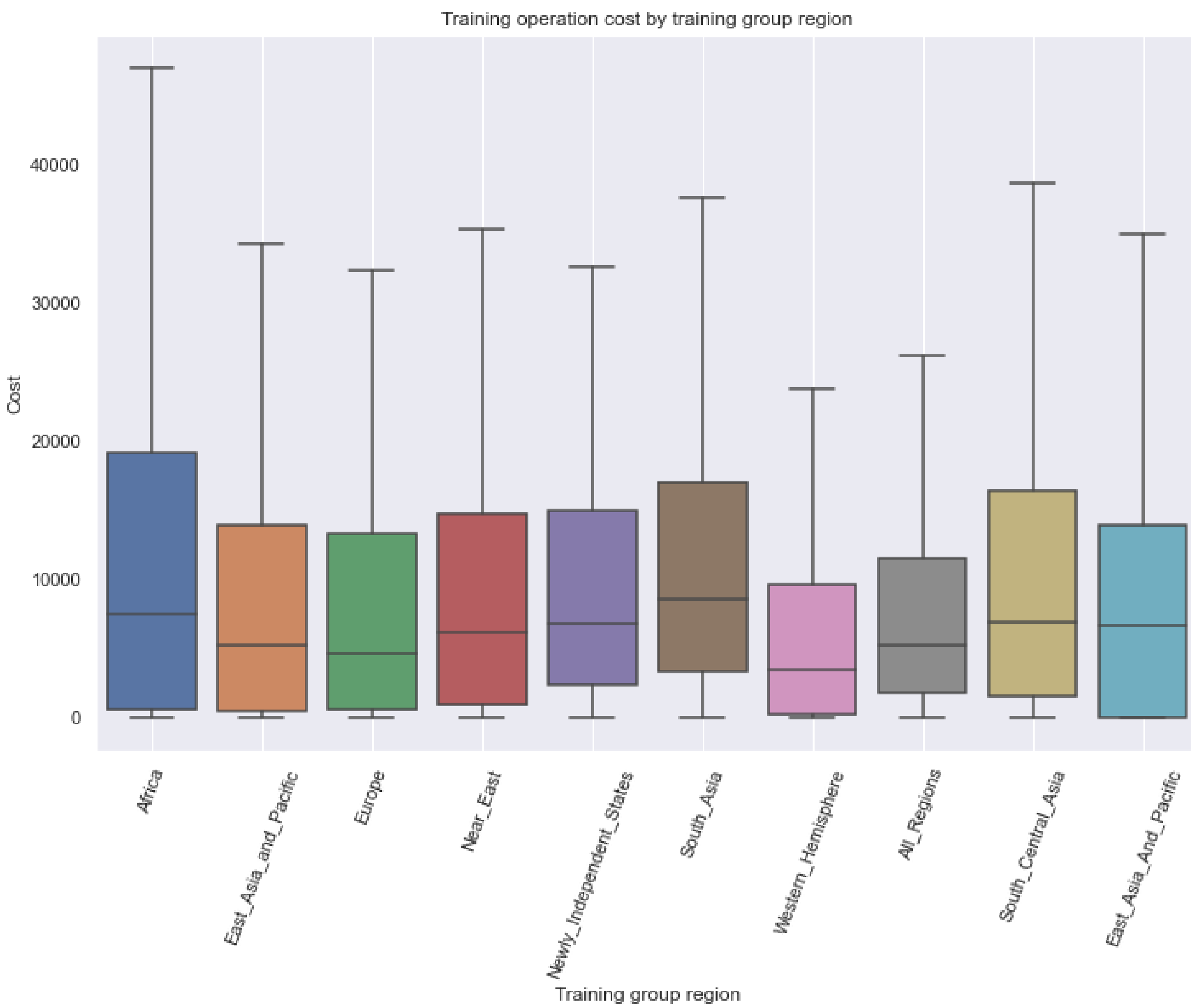
There are many force majeure situations in politics that could lead to extremely high costs for varying political and military reasons that would be much more difficult to model or predict. In this data, for example, we see uniquely numbers of individuals and training costs associated with Iraq in the early years of the Iraq war. The inclusion of this data would only distort the modeling of our primary area of interest: the commonplace foreign military training operations that the US undertakes regularly and consistently as a part of its normal foreign policy, separate from particular outlier events such as wars (one would expect, for example, that Ukrainian data for the year 2022 would also represent outliers). This allows our model to focus on predicting the costs of the US's **consistent** foreign policy and defense strategy commitments based on data over the last 20 years.

We dropped everything above the 90^{th} quantile in terms of cost. The distribution was heavily right-skewed and zero-inflated:



We rid ourselves of id information, data validation information, and urls (training:source, training:id:admin, qa:training:date:scraped, qa:training_date_first_seen, training:status:admin, qa:training_source_url). We also rid ourselves of features that were duplicates of other features (duplicated start and end date information, as well as training:country).

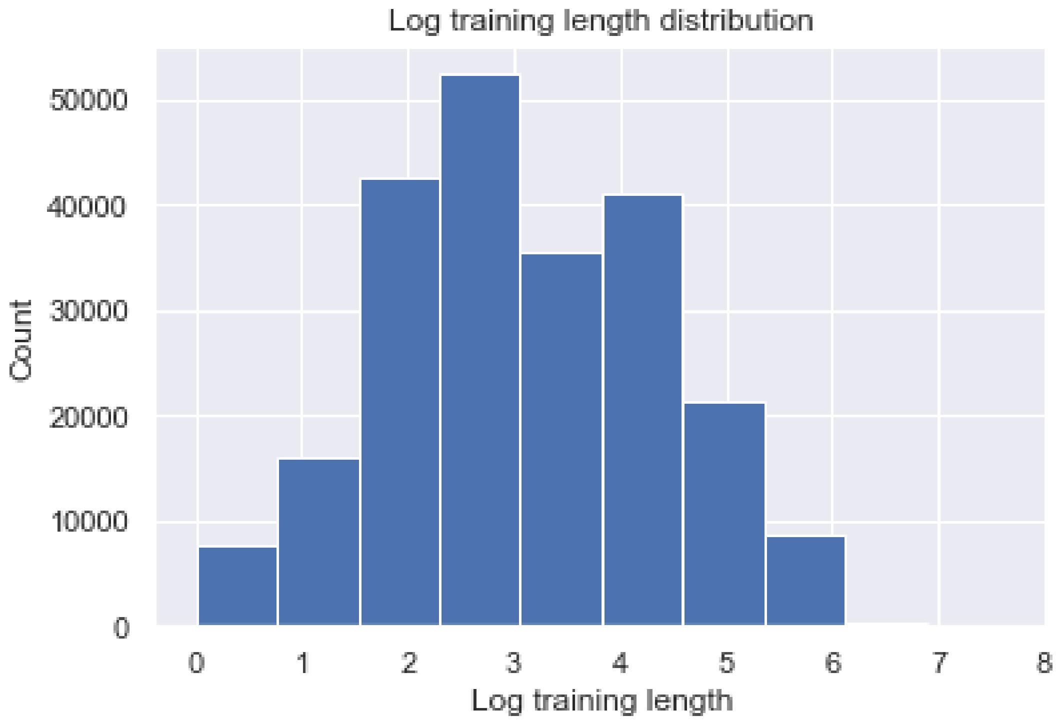
We cleaned the information on training group in order to track the region with which the operation was associated. The other data from training group with regard to date is already accounted for in the training start and end dates. Different regions have different cost distributions, all following the right skew trend. Regional data distributions are displayed below (without outliers):



A small amount of date information (less than 200 rows) was corrupted when extracted from the pdf into the database, and failed to convert properly into a datetime format. We removed this data from our dataset.

A variable was created to track the operation's year (start_year).

We then created a new variable to track the length of an operation (taking the natural log of the length in days + 1 to regularize the distribution, as it was heavily right-skewed):

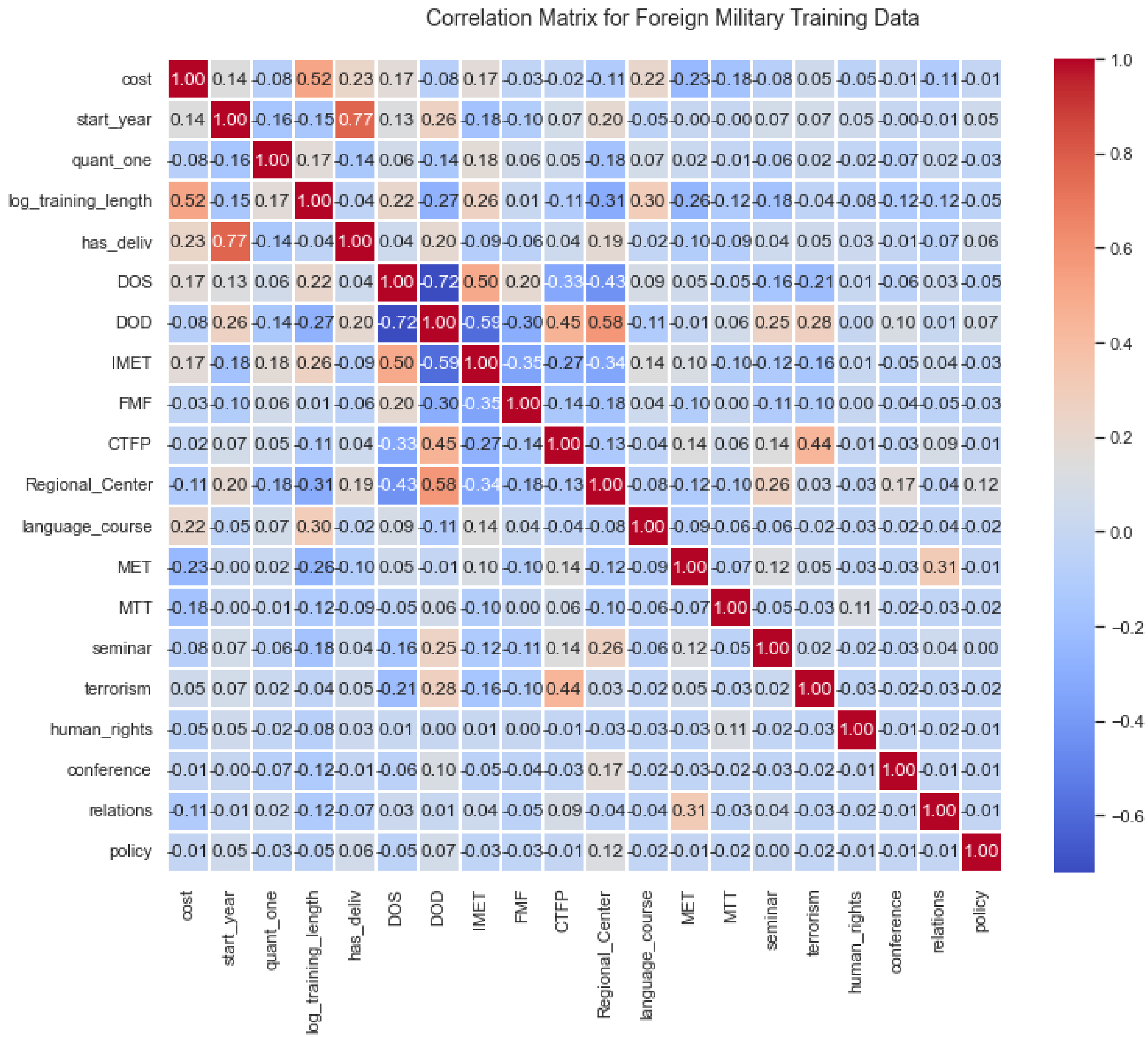


The majority of all operations trained only one individual. We therefore replaced the variable training:quantity with a simple Boolean tracking whether or not training was limited to one individual (quant_one).

We similarly created a Boolean to track whether or not there was a particular American delivery unit associated with a foreign military training operation (has_deliv).

We created new dummy columns to track the most important aspects of our most high cardinality features: location and group name. This included columns to track whether an operation was operated by and funded by Department of Defense or Department of State umbrella (or both, more rarely), what programs it was a part of (IMET, FMF, CTFP, Regional Center programs), and also some information on what kind of program it was (MET, MTT, a seminar or conference, focused on terrorism, human rights, policy, or relations).

The correlation matrix for these features is shown below:



Our target variable's strongest correlation was with the log of training length (0.52), and there were several other promising correlations (has_deliv, IMET, language_course, MET, MTT). Our features are interrelated, especially between programs and the particular US institution that operates them.

We then created dummy variables to track the country for which a training operation was ordered.

We began modeling with data on 196,487 foreign military operations, each with 262 features, including our target feature of cost.

Modeling

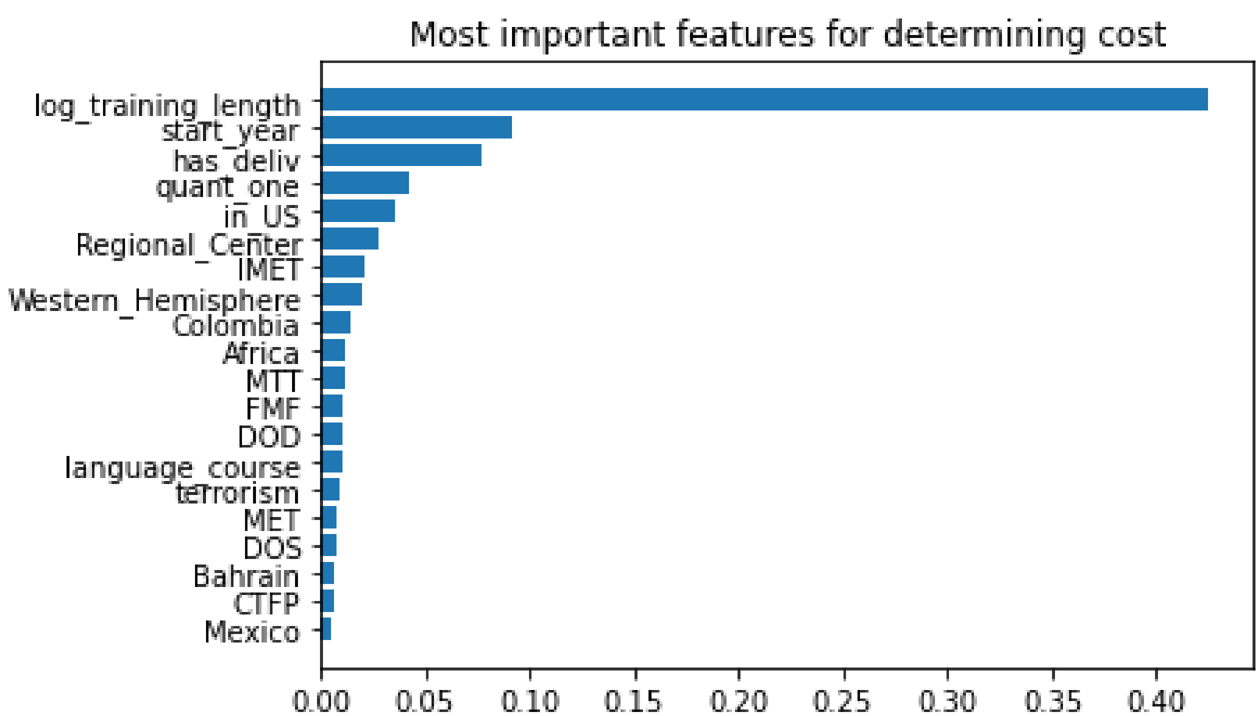
We experimented with several models, using grid search to optimize our parameters and hyperparameters. As our target variable was not normally distributed, a linear model was a poor choice for our data. We avoided the choices of certain models, such as KNeighbors Regression, because of the size of the dataset and the time required for computation. The results of three different models (RandomForest, HistGradientBoosting and XGBRegressor) on our dataset are presented below:

	Cross val r2 score	Cross val r2 std	Train score r2	Test score r2	RMSE
Random Forest	0.765	0.004	0.927	0.771	3408.439
Hist Gradient Boosting	0.653	0.003	0.662	0.659	4157.194
XGBoost	0.743	0.003	0.823	0.751	3553.726

Random Forest Regression performed best on our dataset in all aspects. It had the highest r2 scores for both training and test data, and the lowest root mean square error. Our Random Forest model was able to account for almost 80% of the variation in the cost distribution, with a mean absolute error of 1752.83.

Using our model, we can predict the cost of a foreign military training operation within an average error margin of $1752.83.

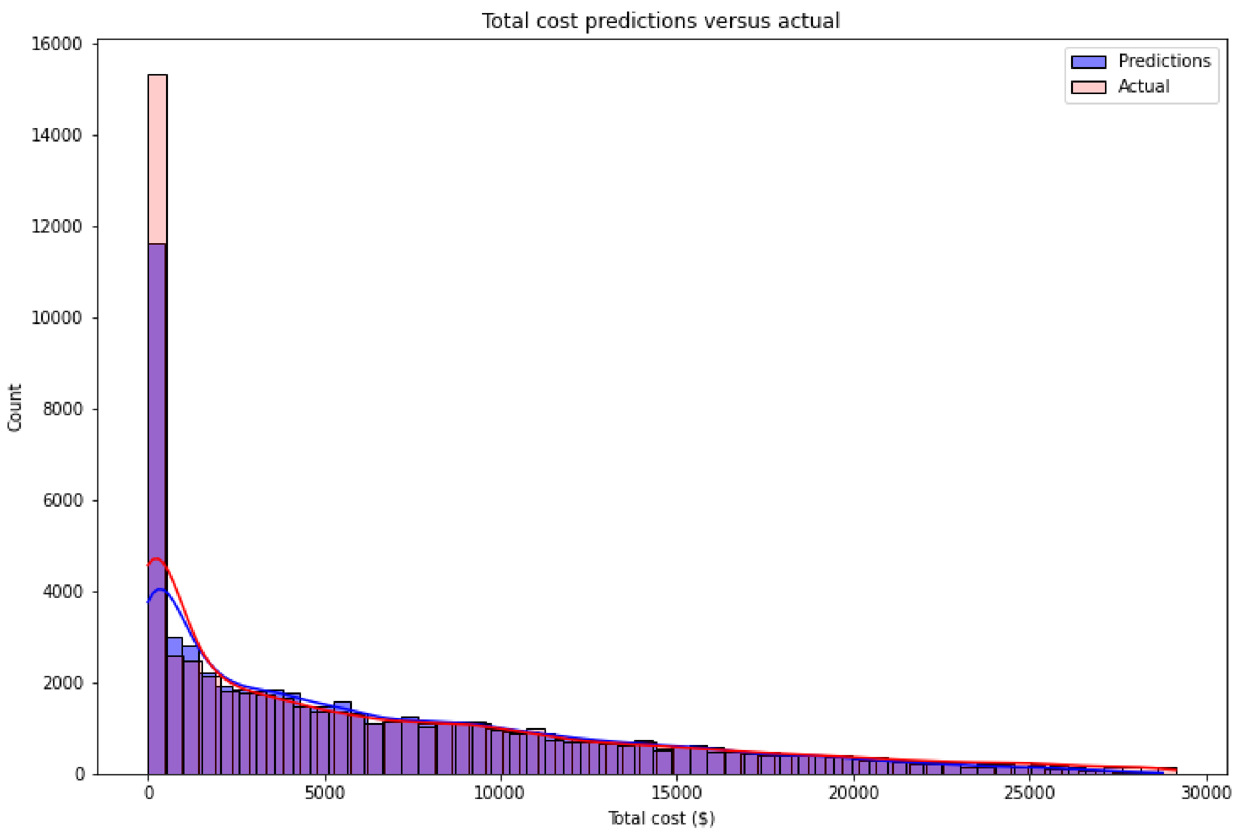
Our model also clarifies the most important features for determining cost:



Training length is responsible for nearly half of the variation in cost. This comes as no surprise, and when plotted, we can see that one-day long programs (with a training length 0) are quite often even "free," with a training cost of 0, in the sense that they require no additional budgetary commitments.

The year also plays a significant role in determining cost in our model, as does the presence of a delivery unit, or whether more than one student was involved. After the first several features, feature importance decreases greatly, and we are left with many unimportant features. However, some of these features that are currently unimportant in our model (such as specific countries), could be important in later iterations of the model done with updated operations data for later years, and thus they should not be deleted. This could occur if a specific country became a particular focus of military training operations, as an example, the way Colombia currently is (reflecting its status as the 9th most important feature for determining cost according to this model).

The distribution of our cost predictions is plotted below alongside actual costs:



Our predicted and actual distributions are practically identical for total costs above $2000. Almost all of the error in our model comes from low costs, which the model mistakenly increases. High cost programs are the ones of most concern for our stakeholders, so these areas of relative model weakness and strength are not a large issue for its predictive utility.

Conclusions:

This purpose of the model does not lie in feature optimization, that is decided by other factors, primarily political and military, that take precedence over budget in matters of US security. The model is more useful for the lawmaker or citizen interested in understanding exactly what is being proposed when foreign military training operations are discussed.

Many of these operations are in fact free, in the sense that they fit within the budgets of larger Department of State or Department of Defense programs. This should be somewhat relieving for those worried about the scale of these programs and inflating military costs. The deep right skew of the data shows that most training programs are fairly inexpensive, and our model's accuracy with the more expensive models is very high. We have excellent predictive capabilities in the area that would thus be of most concern and interest for a stakeholder.

Our model can explain 77% of the variation in the training costs for test data with a RMSE of 3408. Using cross validation on training data it performs consistently at this level, with a mean r2 score of 0.765, with a standard deviation of 0.004. It performs excellently on the training data (r2 score of 0.927), but is not deeply overfitted – by decreasing max tree depth (currently unlimited), for example, we decrease training r2 score without increasing testing r2 score. In terms of real dollars, we can predict the cost of a foreign military training operation within an average error margin of $1752.83 using this model.

The data is fairly predictable based on a number of simple factors related to the location, program type, country that is readily available for unclassified data when programs are being discussed – though at this stage it may remain inaccessible for the average citizen, a policymaker or bureaucrat involved in budgetary processes would be empowered to easily understand and visualize real costs of proposed operations and thus better evaluate their utility and necessity.

The effectiveness of such a program, of course, still remains a different concern that cannot be evaluated within the scope of our data and would be another important concern of any involved stakeholder going forward.