STEPS

RESULTS

INITIAL STEPS

- *Using the World Bank Development indicator data set, we filtered data sets for years 2005 to 2019 and the countries in USAID Dataset*
- *Reducing 1400 different indicators to 200 plus indicators manually*
- *Check for multi-collinearity to reduce any highly correlated variables - reduced our variable space to 65 variables*
- *Combining the world bank development indicator dataset indicator with CPI, we were able to run **XGBOOST** model to get 30 percentile of most important variables (indicators) which reflect a change in corruption levels in a country*

TEMPORAL ANALYSIS

TOPIC MODELLING

- *Cleaning the USAID Project Dataset to get the desired columns, i.e., Project Description to label each project with different topics*

FEATURE ENGINEERING

- *Reduction of topic vectors and other variables in the USAID Dataset to country level data by creating features around statistical measures of central tendency for each column (K-means, Hierarchical clustering)*
- *Reducing the world bank data indicator dataset to country level dataset by creating features around statistical measures of central tendency for each column and to preserve the history, we created features such as Central Moving Averages at Country Level for different years (Logistic Regression, Decision Tree)*

SUPERVISED MACHINE LEARNING

TARGET VARIABLE ENCODING

UNSUPERVISED MACHINE LEARNING

MACHINE LEARNING

- *Unsupervised Learning - Fusing the USAID country Level dataset and WDI at country level to cluster them using unsupervised and supervised learning to find countries that have similarities into the same cluster*
- *Supervised Learning - Running a classification model with the Target Variable to understand how important different factors in the project data play in country's change in the level of corruption.*

COUNTRIES WITH SIMILARITIES IN ONE CLUSTER

FINDING OUTLIERS

EXTENT OF SUCCESS OF ANTI CORRUPTION INVESTMENTS