White Elephants kill. Can we anticipate them?

Andrei Bartra and Angelo Cozzubo

Executive Summary. For any Public Budget Office, one of the main concerns is how to deal with the opportunity costs of financing a subset of public investment projects. Having a project that is not executed within the expected time frame generates enormous opportunity costs. Our machine learning model will focus in the situation of infrastructure construction contracts between the Ministry of Economics (MEF) and the private sector in Peru. For this country, previous studies by Bancalari (2019) and Lagunes (2019) have showed that incomplete projects in Peru, 'White Elephants', have increased under-five mortality rates and that the role of monitoring by civil society organizations backed up by the public anti-corruption agency can drive considerable efficiency gains.

Our model uses administrative information for all national and regional level projects between 2013-2018[1] to predict that a project turns out into a *White Elephant*; that is, it does not complete its execution plan in the corresponding year. Using the project information merged with data from national household surveys, we obtain a dataset unit at the project-month level which includes 856 variables and 634 344 observations. This corresponds to a total of 20 232 projects in eight years.

We follow five steps for processing the data before computing the predictive models: (i) missing value imputations, (ii) outlier detection and imputation, (iii) correlation filter, (iv) removing low variance variables and (v) using polynomials and logarithms transformations. As we are trying to predict a continuous variable (a ratio), our Machine Learning problem is one related to regression techniques. For this problem, we decided to apply, on one hand, a simpler and more intuitive LASSO; while, on the other hand, we focus on computationally more intensive random forest.

The hyperparameters tuning was done by 5-fold Cross Validation. In the LASSO, we tuned the level of shrinkage; while in the Random Forest we tuned the number of estimators, the minimum sample split and the minimum samples at each final node (leaf). The chosen score metric was the R2 while the test fit of our models was assessed in a random 20% subset as well as in a out-of-time fashion using the last year of available data.

The results show that the Random Forrest (RF) performed better than Lasso Regression. RF produced an R^2 of 49.6% in the test sample, while Lasso produced an R^2 of 2%. Nonetheless, in the Out-of-time sample, both models show a considerable reduction in performance. RF reached an R^2 of 29.9%, while Lasso had a negative score showing that the model was not useful for predicting future projects.

The design of the model was developed in coordination with the MEF. They require a model that allows them to get early warnings to intervene projects with high probability of not meeting the execution goal and to optimize the budget allocation process. Up to date, this work has been done by "hand-inspection" lacking a systematic approach and without a predictive perspective. Thus, we consider the machine learning pipeline will be a considerable advantage for the Ministry responsibilities. The limitations of our work were time and data-access related. For future work, we suggest considering (i) improvement in cluster techniques, (ii) supervised models for classification of projects, (iii) more time series operators and (iv) including other data sources. Finally, we stick to Deon Checklist to assess the ethical component.

Key Words: Public Finance, Budgeting, Corruption, Developing Countries.

Github: https://github.com/acozzubo/CAPP-30254-Cozzubo-Bartra

[1] For the time-series operators of the projects of 2013, we consider data from 2012.

1. Background and overview of solution

For any Public Budget Office, one of the main concerns is how to deal with the opportunity costs of financing a subset of public investment projects. Having a project that is not executed within the expected time frame, or even that is not executed at all, generates enormous opportunity costs to the government and its citizens.

Our study will focus in the situation of infrastructure construction contracts between the Ministry of Economics (MEF) and the private sector in Peru. This country has experienced a rapid growth in the last two decades, cataloged as the 'Peruvian Miracle'. However, its infrastructure deficit is predicted to remain at $110 billion, being one of the highest in the region (CEPAL, 2011; Penaranda, 2019). In this sense, studying bottlenecks that may delay or become a prohibitive cost for new construction contracts is crucial for the Peruvian case.

Besides the transaction cost argument, there are welfare reasons to worry about this problem as demonstrated by Bancalari (2019) in her study of incomplete sewerage projects in Peru, named by her 'White Elephants'. She documents "an increase in under-five mortality in districts that experienced greater sewerage diffusion. The result is linked to hazards from the construction works and was exacerbated by delays and mid-construction abandonment. The potential health benefits of sewerage fail to manifest even after completion of projects due to lack of household connectivity". In this sense, the costs are not only monetary but can also include lives.

In Perú, the delay of projects is aggravated because of deep corruption problems (Levin, 2019) and the lack of experienced teams (OECD, 2019), as it introduces more inefficiencies and increases the execution times. Is not uncommon to hear of "ghost projects" that were budgeted but never existed in real life. Furthermore, Perú suffers from tight budget constraints due to is low tax burden which is the third lowest in Latin America[2].

In recent years, accordingly to the MEF data, the government has registered high levels of non-executed budget on public investment. For instance, in 2019, only 66% of the allocated resources for public investment was successfully executed and near $4 billion had to be returned to the public treasure.

To minimize the opportunity costs, the Ministry of Economics acts in three phases: (i) selecting project proposals, (ii) approving the project and doing monitoring and (iii) intervening whenever delays or problems are detected. The first phase can be troublesome since simple and clear rules are preferred for transparency, but the selection is highly susceptible to political promises or negotiations; which make it more difficult to implement strict algorithms to select the projects.

In the second phase the decision is much simpler because it is the Ministry prerogative to decide where it wants to focus its efforts to secure timely project execution. These interventions range from technical assistance to unlock delays, to corruption prevention measures. In this regard, the research done by Lagunes (2019) shows high efficiency gains with simple treatments. He conducts a randomized control trial where he sent letters to private contractors indicating that works under their charge were being monitored by a civil society organization with the support of the country's leading anti-corruption agency. The intervention lowered the cost of public works in the treatment group thus improving its efficiency.

Following this efficiency path, our project seeks to build the best predictive model possible (given the constraints) to forecast the successful execution of infrastructure projects in Peru. We will employ the machine learning techniques and pipeline to construct an algorithm that will anticipate the appearance of 'White Elephants'. This tool will be of specific utility for the Ministry since it will let the institution focus its monitoring and intervention as preventive efforts towards projects that the model predicts will result in unsuccessful execution. Moreover, as the Ministry highly technical human capital is scarce and expensive, we aim that our tool will result in a substantive saving of public resources.

[2] https://www.bbc.com/mundo/noticias-47572413

The main audience for our analysis, the adoption and implementation of our algorithm are the public servants in the MEF. Considering we only have 4 weeks with non-exclusive dedication, our hope is not to have a perfect final model but to set a benchmark for a perfectible but technically correct machine learning model that meets the standards to be put in production. Afterwards, the Ministry may use the pipeline as a foundation for the improvement of the model or to develop other machine learning algorithms.

It is important to highlight that this project was conceived through conversations with current public servants at MEF. They generously shared with us the databases available at the short term containing the information of infrastructure projects for the last eight years.The definition of the target variable was long discussed with the MEF team, seeking that this model is truly useful in their daily work. We make sure all the external statistical and geographical information used in our model is of public access or that the Ministry have the rights to access it.

In addition to the MEF team, we believe that the audience of this work may go further. As Lagunes (2019) showed, the role of monitoring by civil society organizations backed up by the public anti-corruption agency can drive considerable efficiency gains. For this reason, we consider them as a secondary audience of interest.

As stated before, the actions the Ministry can take are clear and aim to improve the efficiency of infrastructure projects in the country. The principal would be to design preventative actions to anticipate and prevent White Elephants. Furthermore, the civil society and anti-corruption agency are also capable of acting with social monitoring and denouncing corruption and inefficiency to the authorities.

2. Data

We combine two sources of data for this project. First, we obtain administrative data from the MEF which comprises all the national and regional level public investment projects between 2013-2018. This database includes project characteristic, costs, location and the execution amounts. The budget and execution amounts are unbundled in *accounts* that allocate budget considering combinations on the following variables: Government Level, Economic Sector, Funding Source and Budget Office. For instance, a large project may have budget allocated to the central government for a component of the project another component may be allocated to the regional level. Furthermore, a project may cover different economic sectors. A port may have budget allocated for defense and budget allocated for commerce and tourism. Also, a project may have different funding sources like regular taxes, debt, and canon. Finally, there are offices that supervise the execution process, and this could also vary within a project. The possibilities for feature extraction are endless, but we kept it simple by considering only government level, economic sector (with some aggregations to reduce complexity) and funding source. Another important variable is the Executive Unit, which is the ultimate responsible of the execution of a project.

Key variables from this dataset are the total investment at the time the project proposal was accepted, the date the project was accepted, the execution performance of the Executive Unit and execution dynamics at a government level, the funding source and the economic sector. We exploit the name of the project by applying a fuzzy Jaro-Winkler matching with key words to categorize if the project was new, an upgrade of a previous work or an intangible one (non-infrastructure).

Using this information, we constructed our target variable as the ratio of the total execution (PEN) divided by the total year budget for that project (PEN)[3]. Hence, our target variable corresponds to the proportion of money executed for a given project at the end of the year. The MEF gets the value of the variables a few days after the end of the month. In order to have an implementable model, we can only consider data from January to October, having November as a blind spot. The data from November will be ready on December when no further actions can be taken.

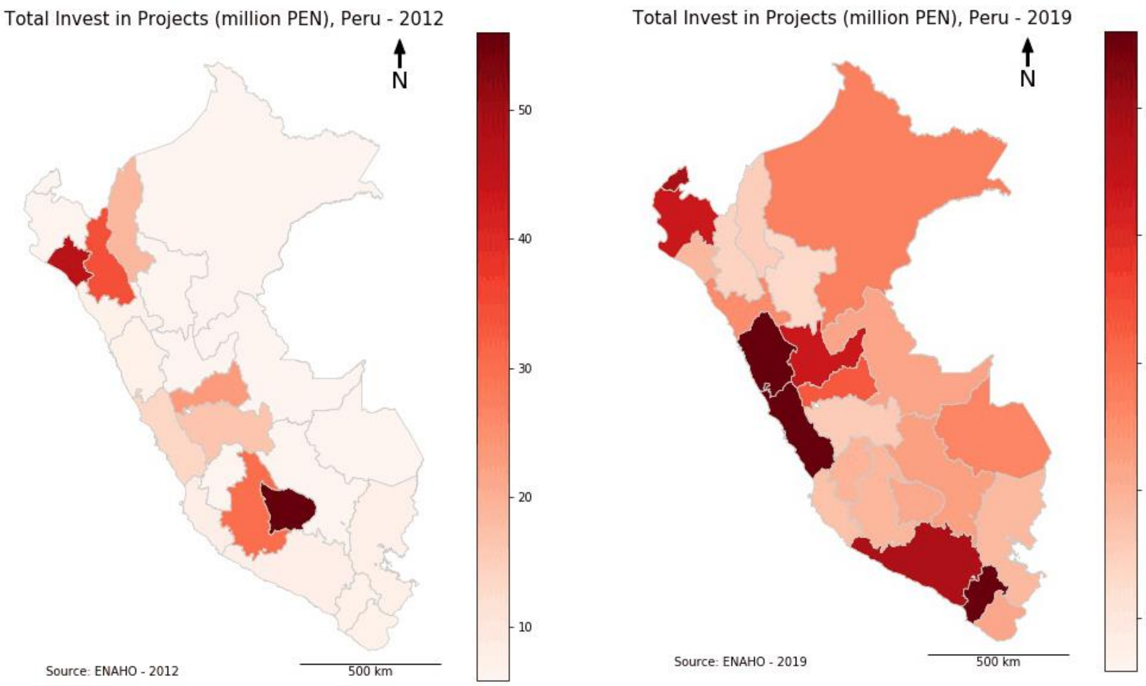
[3] As coordinated with the MEF

In addition, we merge this information using the region were the project will be deployed with data coming for the National Households Surveys 2012-2019 coming from National Institute of Statistics (INEI). We obtain more than 150 variables representative and aggregated at the regional level comprising dimensions as geography, household (hh) members, hh characteristics, hh income, hh health, hh labor and hh ethnicity. The information from these variables is lagged one year in order to ensure the MEF has access to the information when they need to run the predictions. With this final step, our dataset unit of analysis is the project-month, and it includes 856 variables and 634 344 observations. This correspond to 20 232 projects in eight years.

From Figure 1, we observe there is a strong geographic variation in the regional investments. It is important to highlight that a considerable amount of resources comes from local mining works and are invested in the region where the mineral is exploited[4]. From the map's animation (included in the power point presentation), we see that the initial years of the series show investment is focused on the poorest regions; while there is a recent increase in the overall investment and a concentration in the capital, Lima.

One interesting hypothesis commonly heard in Peru is that authorities tend to spend more when there is an election coming soon. As descriptive analysis we tested this hypothesis for the Presidential and Regional Governments Election (the latter can be found in the presentation). We used the cumulative distribution function (CDF) for the total investment in projects and for our target variable, the ratio of completion of the projects. These results are present in Figure 2 and Figure 3.

Figure 1 - Total Investment in Projects by Region


Total Invest in Projects (million PEN), Peru - 2012
N
50
40
30
20
10
Source: ENAHO - 2012
500 km
Total Invest in Projects (million PEN), Peru - 2019
N
50
40
30
20
10
Source: ENAHO - 2019
500 km

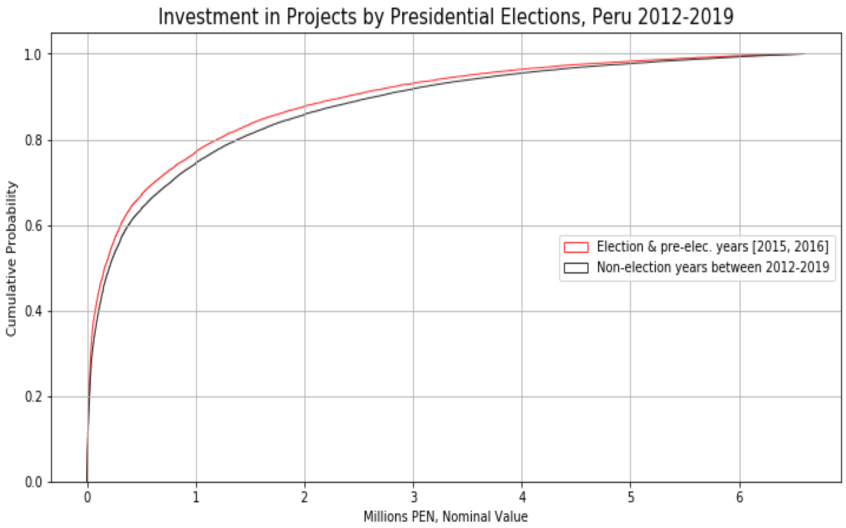
We can observe that the CDF for investment in projects in non-election years stochastically dominates[5] the election years curve. We can interpret this as in election and pre-election years smaller projects are preferred since they can be completed faster. On the other hand, the CDF for our target variable, the cumulative execution of project, shows that the election

[4] This is called "Canon" and the projects have

[5] For every percentile of the distribution the value of the non-election year execution is higher than the value of the election year execution.

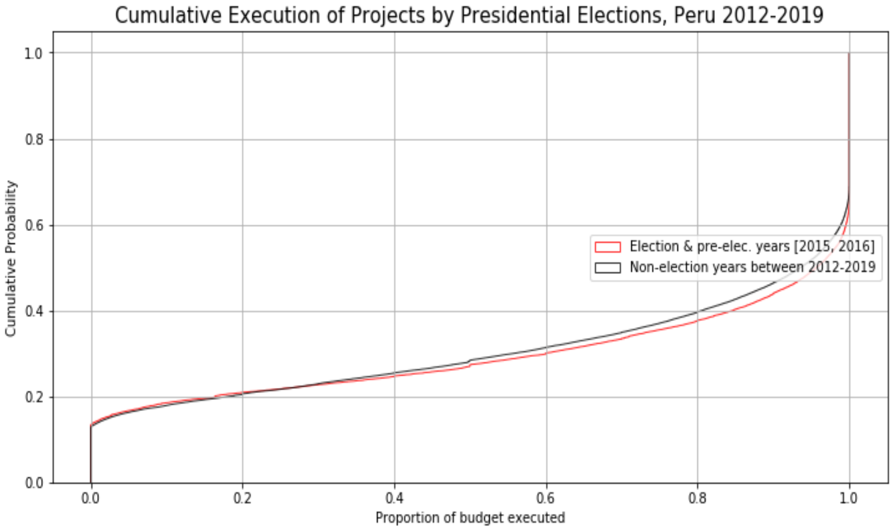
years curve stochastically dominates the non-election years curve. That is, the completion of projects is accelerated in the election and pre-election years, as we can believe is an attempt to catch up with campaign promises.

Figure 2 - Cumulative Distribution of Investment by Presidential Election years



Note: Trimmed at the upper 5%. Source: MEF Administrative Data 2012-2019

Figure 3 - Cumulative Distribution of Target Variable by Presidential Election years



Source: MEF Administrative Data 2012-2019

3. Machine Learning and Details of Solution

As we discussed during the quarter, the main workload of a machine learning pipeline relies on the preprocessing, data cleaning and data wrangling. Having our database merged we applied five steps in our preprocessing phase:

i. Missing value imputation: we start by imputing the missing values in our variables with zero. We added dummy variables for each variable imputed to indicate which rows had been imputed for the model to be able to consider this step (i.e. as a relevant dummy variable in the LASSO or as a relevant subdivision of the tree for the Random Forest).

ii. Outlier imputation: for the continuous variables, we trimmed the upper 1%. That is, we imputed values above the 99th percentile with the 99th percentile. This was important in our case since we have considerably large projects in the right tail of the distribution.

iii. Correlation filter: We developed a correlation filter to reduce the complexity of our model in the preprocessing discarding predictor variables that are similar with a greedy approach algorithm. We obtain the correlation coefficient for a pair of predictors and, if it happens to be above certain threshold (0.2), we discard the predictor with the lowest correlation with the target variable. If the correlation coefficient of this pair is below the threshold, we keep them both. We iterate this process for each pair of features[6].

iv. Removing of constant variables: Some of our dummy variables has no variation and will not add up to the predictive capacity of the model. Hence, we discard these variables.

v. Taking logarithms and polynomials: For the variables expressed in millions of PEN, we decided to take natural logarithms to compress the extremely high variance of the variable. This is a common step done in regression analysis since it helps "normalize" the distribution and interpret the results in terms of percentage change which makes more sense for this kind of variables. For tree-based algorithms, this does not have any effect as it is a monotonic transformation. Also, for continuous variables we applied squared polynomials to try to capture non-linear relationships.

As we are trying to predict a continuous variable (a ratio), our Machine Learning problem is one related to regression techniques. For this problem, we decided to apply, on one hand, a simpler and more intuitive LASSO. This estimator performs selection and shrinkage of the coefficients, while the final form of the model is "observable" and interpretable due to its linearity. On the other hand, we fitter a more computationally intensive model, a random forest. This model would be better at exploiting non-linear relationships in the data, but we lose the ease of interpretation attribute.

For both models, we considered the attributes filtered by the preprocessing steps and kept only 100 variables. To the subset of continuous variables of them, we applied polynomial expansions that considered square transformations. This model was trained using the 80% of the data and the evaluation was done using the 20%. We also considered 2018 as the out-of-time sample to test the implementation of the model with a more acid criterion.

The hyperparameters tuning was done by 5-fold Cross Validation. In the LASSO, we tuned the level of shrinkage; while in the Random Forest we tuned the number of estimators and the minimum sample split. The score metric was the R^2 as it gives us a bounded value of what proportion of the variance in the target variable, we are capable to explain with our model.

4. Evaluation and Results

To evaluate the performance of the model we will use R^2 metrics as we are solving a regression problem and the MEF is also familiarized with the metric. The score can be easily interpreted as the share of the variance from the target variable that our model is able to explain. In order to assess the performance using a benchmark metric, we run a simple linear regression (OLS) with the final variables of our pre-processing. We obtained the following results: 6.56% in the Train sample, 7.4% in the Test sample and 0.3% in Out-of-time sample.

For the LASSO, the best performing model in the train sample was with a shrinkage parameter value of 0.06. This model has an R^2 of 2.2% in train sample, 2.2% in test sample and a negative score in the OOT sample. Due to its lower predictive capacity in comparison with the benchmark OLS, this model was discarded.

For the Random Forest, the best performing model in the train sample was with 100 tree estimations and a minimum sample to split of 200 observations. This model has an R^2 of 52.4% in train sample (x8 the benchmark), 49.6% in test

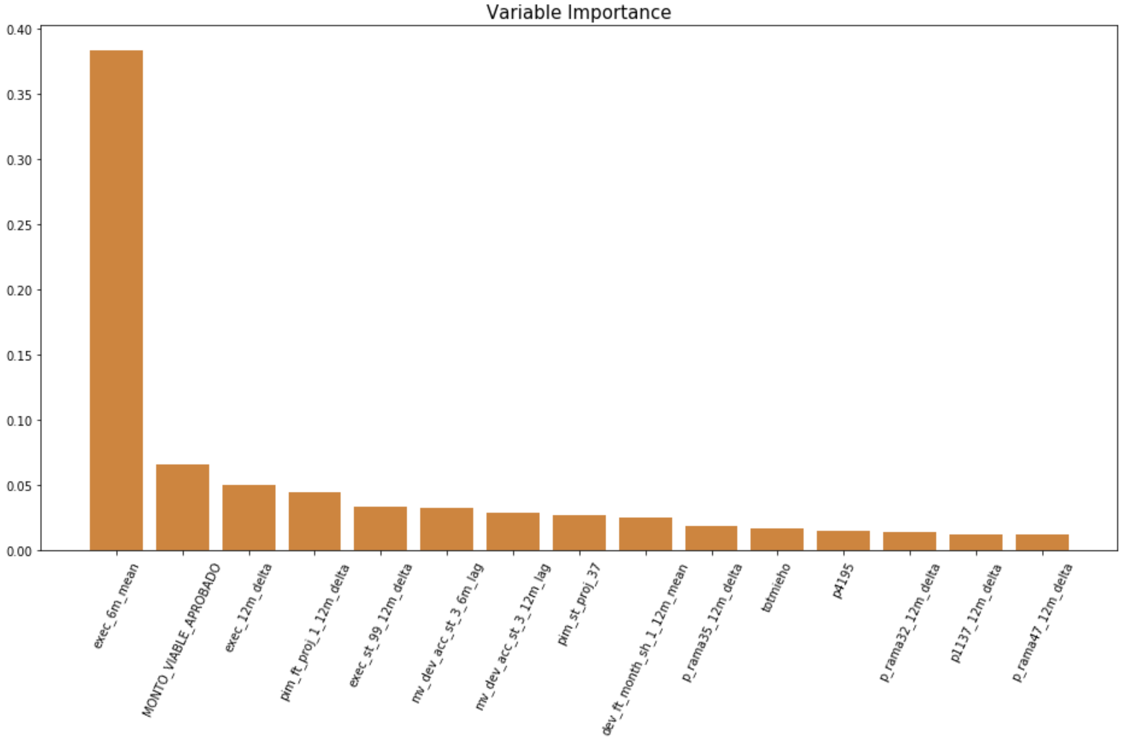
[6] As suggested by Sellmair (2018) and Obi (2019), training prediction models with too many correlated features may reduce the model accuracy. Feature selection and dimensionality reduction are two paths to overcome this problem. Moreover, these techniques help to prevent overfitting, helps with the parsimony of the model and helps with the computational efficiency.

sample (x7 the benchmark) and 29.9% in the OOT sample (x100 the benchmark). In the annex, we show a plot of the cumulative distribution function that compares the estimated versus the true execution ratio. The model is not able to replicate the modes at 0 and 1 values. Beyond that, the distribution follows the behavior of the true value.

Unfortunately, we were not able to solve the dramatic decline in performance in the OOT, which is the most important metric as it resembles the performance of the model in a deployment situation.

Regarding the feature importance, the most relevant variable is the average execution from the last 6 months. This makes sense as it captures the inertia of the execution levels of the project. Then, the second variable in importance is the approved amount of the whole project. Larger projects get more attention from authorities and are designated to the best performing executive units; therefore, they tend to have higher levels of expected execution. The next variables are time series operators of different execution categories. At the tail of the top 15 variables, we observe several socio-demographic variables as the number of household members and the economic sector where the household chiefs are employed.

Figure 4 - Feature importance in Regression Forest



5. Policy Recommendations

First, we must bear in mind that this model is not a structural neither a causal one. Hence, a feature with a high importance at predicting the target cannot be considered as the cause of the target.

The main policy recommendations that we can derive from this exercise goes along the efficiency on the budget allocation. These models let us have a quick a systematic way to flag those projects that more probably will have delayed executions. That is, projects that are more prone to be White Elephants.

In this sense, our model serves at giving early warnings to the MEF for intervening projects with low expected execution. As we know that public resources are scarce and highly valuable, the model can serve as an important improvement for targeting the monitoring work of the MEF.

Furthermore, the model has the convenience that it estimates the expected execution which allows the MEF to apply this estimation in their budget allocation models.

It is important to contextualize that currently the MEF does this work by “hand inspection”. That is, analysts check the status of each project each of them having different criteria and heuristic indicators for when a project might turn into a White Elephant. In this sense, up to date, the monitoring of projects has not been a predictive work, but one mostly based on the data availability at the moment and variation based on subjective criteria.

We know that this is not the final model that the ministry might employ. As policy recommendation we also suggest improving this algorithm with more data coming from several non-public datasets available to the MEF. Also extending the series of administrative data for previous year to 2012 would be beneficial to the prediction accuracy.

6. Ethics

Regarding the ethical component of our project, we employ the Data Science Checklist from Deon to assess this component[7].

1. Data Collection
    a. Informed consent: If there are human subjects, have they given informed consent, where subjects affirmatively opt-in and have a clear understanding of the data uses to which they consent?

No humans subject. Information from surveys is aggregated at the regional level and have consent from INEI.

   b. Collection bias: Have we considered sources of bias that could be introduced during data collection and survey design and taken steps to mitigate those?

No bias in administrative data since it is not a sample but the whole universe. Data from INEI is a representative sample.

   c. Limit PII exposure: Have we considered ways to minimize exposure of personally identifiable information (PII) for example through anonymization or not collecting information that isn't relevant for analysis?

All data is anonymized.

2. Data Storage
    a. Data security: Do we have a plan to protect and secure data (e.g., encryption at rest and in transit, access controls on internal users and third parties, access logs, and up-to-date software)?

Data is public. No need to restrict.

   b. Right to be forgotten: Do we have a mechanism through which an individual can request their personal information be removed?

No individual data. This petition should be done to INEI.

   c. Data retention plan: Is there a schedule or plan to delete the data after it is no longer needed?

No need to erase the data as it is of public access.

3. Analysis

[7] https://deon.drivendata.org/#data-science-ethics-checklist

a. Missing perspectives: Have we sought to address blind spots in the analysis through engagement with relevant stakeholders (e.g., checking assumptions and implications with affected communities and subject matter experts)?

Beyond the scope and timing of the project. However, we obtain enriching feedback from analyst at MEF

b. Dataset bias: Have we examined the data for possible sources of bias and taken steps to mitigate or address these biases (e.g., stereotype perpetuation, confirmation bias, imbalanced classes, or omitted confounding variables)?

No bias in the datasets as MEF data is administrative, while INEI is representative at the regional and national level. No individual information. Confounding variables will not be a problem since we are not supporting a causal claim.

c. Honest representation: Are our visualizations, summary statistics, and reports designed to honestly represent the underlying data?

As far as our knowledge permit, yes. They were done honestly, and the code is open for replication.

d. Privacy in analysis: Have we ensured that data with PII are not used or displayed unless necessary for the analysis?

Yes. Data is anonymized.

e. Auditability: Is the process of generating the analysis well documented and reproducible if we discover issues in the future?

Yes. Open data and code at Github repository.

4. Modeling

a. Proxy discrimination: Have we ensured that the model does not rely on variables or proxies for variables that are unfairly discriminatory?

No individual data.

b. Fairness across groups: Have we tested model results for fairness with respect to different affected groups (e.g., tested for disparate error rates)?

No human groups involved. Prediction is at the project level.

c. Metric selection: Have we considered the effects of optimizing for our defined metrics and considered additional metrics?

R2 and MSE gave us very similar results. R2 is preferred since it is bounded.

d. Explainability: Can we explain in understandable terms a decision the model made in cases where a justification is needed?

Yes. Elaboration of this in the feature importance and policy recommendations.

e. Communicate bias: Have we communicated the shortcomings, limitations, and biases of the model to relevant stakeholders in ways that can be generally understood?

Limitations are made explicit below. Results will be communicated with MEF authorities.

Finally, there is a chance that the model is used for a purpose different from what is designed. For example, the MEF may target interventions aimed to reduce delay times when the model predicts execution percentage.

7. Limitations, caveats, suggestions for future work

Time was our main limitation since this project was done in four weeks with partial dedication as we were taking a complete set of courses for the quarter. One limitation of our project is that we consider only regional and national level projects. For future work it be recommended to include province and district level projects. This impose a whole new set of adjustment, since national surveys are usually not representative at this level of disaggregation.

More, for projects that are done in more than one location (in our case more than one region) there should be a better adjustment of the variables. We considered these projects as equally distributed between the geographical units and assigned a homogenous weight to obtain averages. However, the MEF knows the relative importance of each geographic unit for these broader projects and may have a better method for obtaining these predictors.

For future work, we suggest considering:

- Cluster techniques for budget offices and economic sectors
- Supervised model for the classification of projects in replacement of our Jaro-Winkler approach.
- More time-series operators: time-beta, moving averages differences, among others
- Include other relevant data sources as other surveys (i.e. ENDES) or non-public information available to MEF.

References

Bancalari, P. (2019) Can White Elephants Kill? Unintended Consequences of Infrastructure Development in Peru. Job Market Paper. London School of Economics and Institute for Fiscal Studies

CEPAL (2011) The economic infrastructure gap in Latin America and the Caribbean. Bulletin FAL No. 293 (1). ECLAC. United Nations.

Lagunes, P. (2018) Guardians of accountability: a field experiment on corruption and inefficiency in local public works. Manuscript. School of International & Public Affairs, Columbia University

Levin (2019) Corruption in Peru: An overview of systemic corruptionand an interview with former prosecutor José Ugaz. Financial Crime Digest, June 2019.

Obi, B. (2019) Feature Selection and Dimensionality Reduction Using Covariance Matrix Plot. Towards Artificial intelligence. Medium.

OECD (2019) Public Procurement in Peru. Reinforcing Capacity and Co-ordination. OECD Public Governance Reviews.

Penaranda, C. (2019) Peru works to make up its infrastructure deficit by looking to private sector. The New Economy.

Sellmair, R. (2019) How to handle correlated Features? Rmarkdown script using data from multiple data sources. Kaggle.

Annex – Random Forest Goodness of Fit

Figure 5 - True vs Expected Execution Ratios Comparison - Test Sample, Random Forest Model

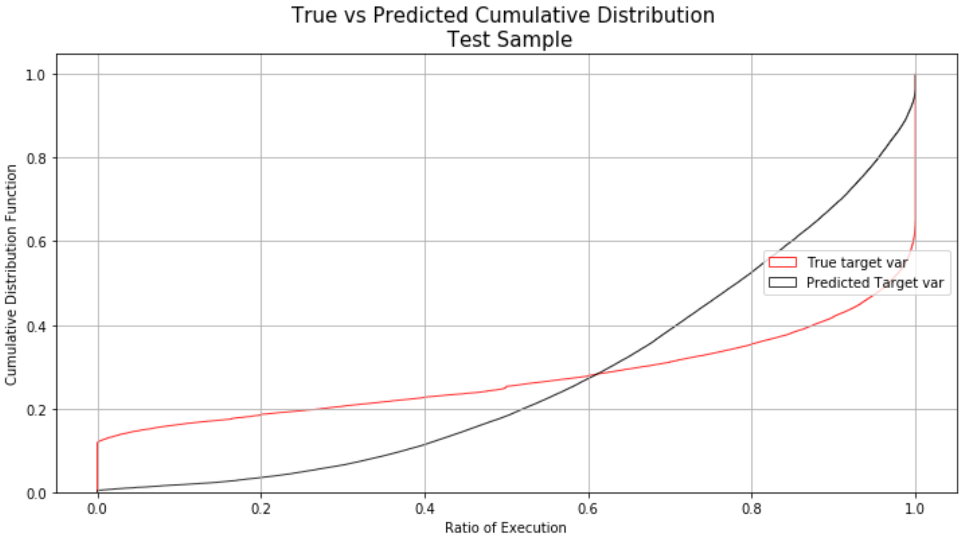


Figure 6 - True vs Expected Execution Ratios Comparison - OOT Sample, Random Forest Model

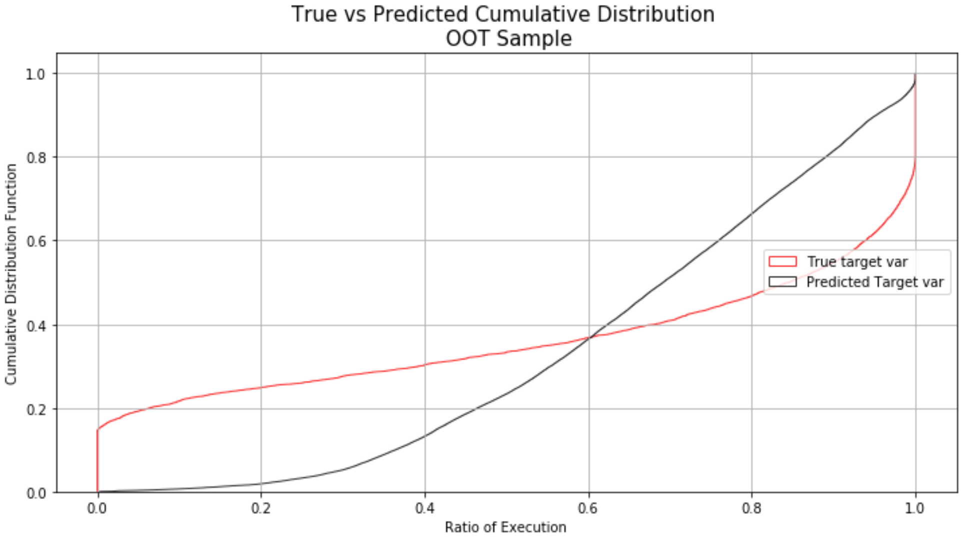


Figure 7 - True vs Predicted Target Variable Scatter - Test Sample, Random Forest Model

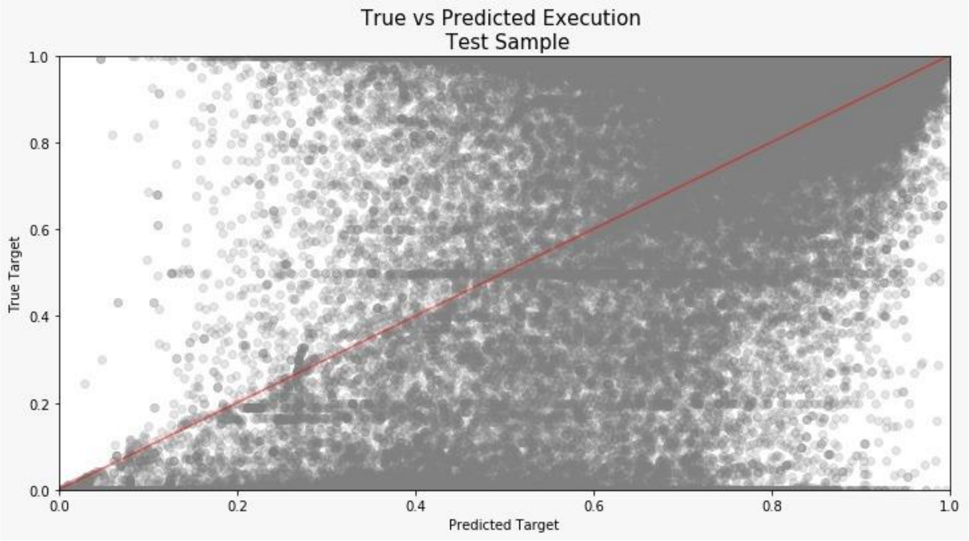
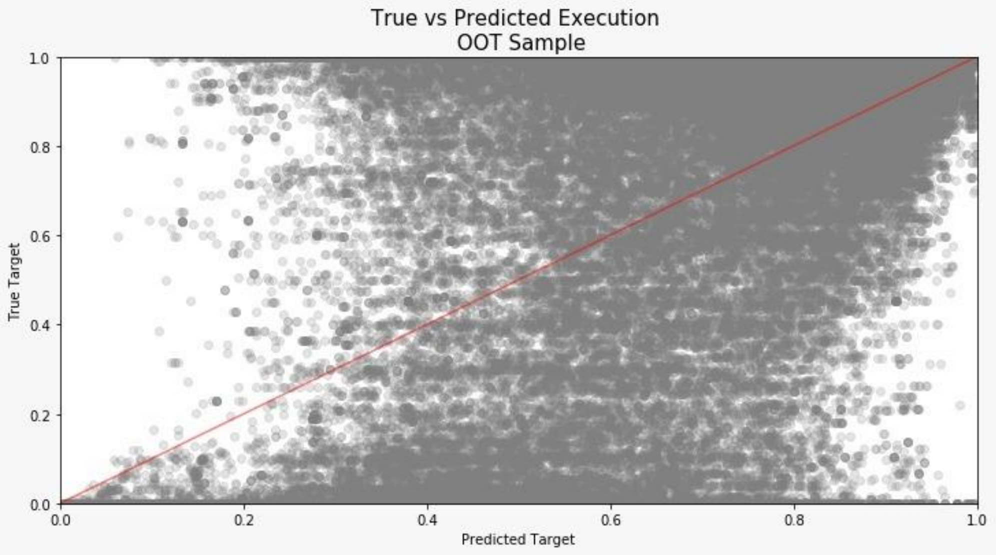


Figure 8 - True vs Predicted Target Variable Scatter - OOT Sample, Random Forest Model


True vs Predicted Execution
OOT Sample
True Target
Predicted Target