A Machine-Learning Analysis of Post-Soviet Power Plants

Maggie Shen, Aleksander Sas

Abstract

Abstract Decarbonization scenarios make predictions based on significant assumptions about the future, so they often fail to account for the uncertainties involved in project realization and the decision making process of financial institutions. In this study we built a series of machine-learning models (LightGBM & XGBoost) to identify key factors affecting the probability of power plant project realization in the post-Soviet space, giving special analytical focus to Central Asia and the Caucasus. We found that the most important factors influencing project realization are largely project-level – fuel type, total cost, capacity, ownership, and finance models – as well as a country's political and regulatory stability. Contrary to global energy sector trends in the last decade, renewable projects in Central Asia, especially small- and large-scale hydropower, are the least likely of any fuel type to be commissioned. The findings point to high risks of fossil fuel lock-in, however we conclude that important limitations in this model necessitate further literary analysis of these regions' energy histories and development pathways to provide more evidence for specific policy recommendations.

1. Hypothesis

According to Alova et al. (2021), which predicts Africa's generation mix in 2030 with a similar machine-learning approach, project-level factors inform the chances of project realization more than country-level indicators like corruption or regime-type. As such, we hypothesize that Central Asia and the Caucasus' power plant mix will be determined by the financing type, ownership model, and fuel type of each project.

2. Methods & Data

The following section analyzes a dataset of planned and historic power plant projects active between 2006 and 2020, using machine learning techniques to predict the chances of individual power plant commissioning and the factors most impactful to its probability modeling (SHAP value analysis). Due to limited publicly-available data on active power plant projects in Central Asia, the following section will discuss the factors relevant to calculating project realization rather than predict the total generation mix and capacity of the region, which was the initial goal of this project.

The dataset is structured wherein each row represents a specific power plant, with 8 project-level and 16-country level indicators. For the project-level factors, we used the CSIS Reconnecting Asia database, which lists 14,000+ infrastructure projects active since 2006, ranging from roads and transmission lines to large-scale hydroelectric dams in the project pipeline since the early 1960s; however, this dataset was massively incomplete, leading us to find price, ownership, and financing data from power-technology.com as well as websites for the largest international financial institutions in the region (Asian Development Bank & European Reconstruction and Development

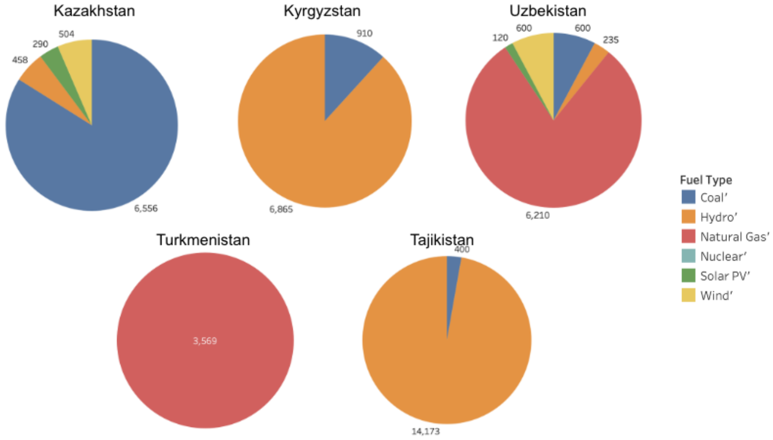
Bank, for two). At the country-level, the World Bank has excelled data on a state's economic, demographic, and political characteristics, which we combined with more specific data on governance from FreedomHouse and the Harvard Atlas Project.

In the codes attached to this paper, we started by cleaning the data using One Hot Encoder and Standard Scaler. One Hot Encoder converts seven categorical variables each into a one-hot numeric array: countries, fuel type, finance, ownership, Landlocked, Borders Russia, and Borders China. We standardized the following columns using Standard Scaler "capacity(MW)," "total_cost (USD)," "Agriculture in GDP," "Gas Rents(% of GDP)," "Oil Rents(% of GDP)," "Population (millions)," "GINI index," "Political Stability," "Corruption(0-100)," "External debt/GDP ratio," "Imports of Goods and Services as % of GDP," "Exports of Goods and Services (% of GDP)," "Democracy Score," "Economic Complexity," "Population Density/km2." The data we ended up using to train machine learning models have 93 row and 48 columns as we excluded plants with status "Announced/Under Negotiation" from our data and labeled the rest with 0 and 1—0 for status "Delayed" and 1 for status "Completed" or "Under Construction."

To learn the impact of various plant-level and country-level factors on plant status, we conducted binary classification using LightGBM and regression using XGBoost, both followed by a SHAP value summary plot to explore feature importance. For the LightGBM binary classifier, we chose Binary LogLoss and Mean Square Error(MSE) as the evaluation metric. Binary LogLoss stopped decreasing around 0.57 in multiple trials; Mean Square Error stabilized around 0.3. For the XGBoost regressor, we used LogLoss and Root-mean-square deviation(RMSE). RMSE stopped improving after 0.31, and Logloss stopped decreasing around 0.52 in multiple trials. Following the XGBoost training process, we also plotted two graphs of importance by "weight," or the number of times a feature appears in a tree, and by "gain," "or the average gain of splits which use the feature." Since the weight graph provides similar information as our XGBoost model SHAP summary plot and the gain graph provides limited insights in our case, we will mainly focus on the SHAP summary plot.

3. Result

Figure 1. Publicly-announced and active power plant projects in Central Asia (2006-2021)

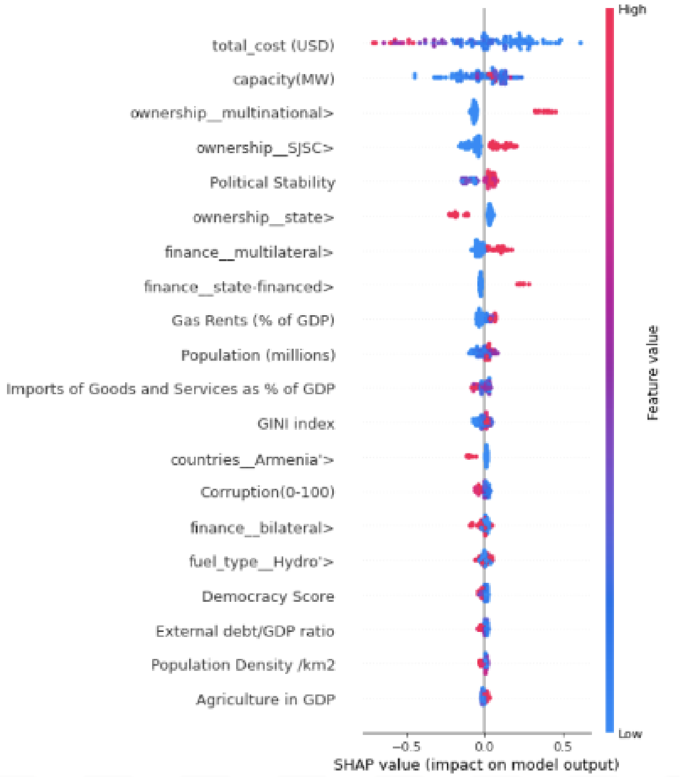


Number around the pie charts show the MW capacity of each announced/historic project fuel type. Notice that Tajikistan's announced projects dwarf other countries despite long-term insolvency in its electric sector and poor governance indicators.

Figure 1 (above) shows the generation mix of projects announced and within the CSIS Reconnecting Asia dataset, not limited by project statuses. This simple visualization indicates that countries more reliant on ODA, Kyrgyzstan and Tajikistan, with legally unbundled but still vertically-integrated power sectors, have been unable to innovate their generation mix; instead, both countries continue to rely on Soviet-era, degrading hydroelectric plants.[1] [2] Most projects included in the dataset for these two countries are classified as “revitalizations” of pre-existing hydroelectric dams built in the 60s and 70s. Tajikistan has announced by far the most new capacity of any Central Asian country due to the presence of long-term, constantly-delayed plans for the Rogun dam and other large scale hydroelectric projects in this dataset.

Kazakhstan and Uzbekistan have both recently unbundled their power sectors, with the latter in the process of implementing major cross-sectoral reforms since the installation of a new regime in 2016.[3] Figure 1 thus helps build on the hypothesis that sector liberalization is correlated with the introduction of renewable energy technologies, such as the announced 100MW Sukhondaryo-Uzbekenergo Solar PV Park in Surxondaryo, Uzbekistan and completed 100MW SES Saran Solar PV Park in Kazakhstan’s Karaganda region.[4] The latter is the largest RES facility in the whole of Central Asia.

Figure 2. Impact of features on the success or failure of project commissioning dependent on feature value



A positive SHAP value means that a feature with a high value in red will have a positive impact on a project’s realization; conversely, red feature values to the left of the axis mapped onto negative SHAP values represents a harmful impact on power plant outcomes.

Figure 2 displays the results of the model trained on a subset of completed and failed power plant projects. SHAP values are used to calculate the relative importance of a feature in determining probability. As in other machine-learning analyses of global power plant project chances, 4 out of the 5 most important features of the GBM and XGB models are at the project level: cost

[1] IEA. (2021). Kyrgyz Republic Energy Profile. Retrieved from: https://www.iea.org/reports/kyrgyzstan-energy-profile

[2] IEA. (2021). Cross-Border Electricity Trading for Tajikistan: A Roadmap. Retrieved from: https://www.iea.org/reports/cross-border-electricity-trading-for-tajikistan-a-roadmap

[3] World Bank. (2018). Doing Business 2018. Retrieved from: http://www.doingbusiness.org/reports/global-reports/doing-business-2018

[4] CSIS (2020). Reconnecting Asia Project Database. Retrieved from: https://reconasia.csis.org/reconnecting-asia-map/

(USD), MW capacity, and ownership types (multinational vs. majority state-owned joint stock companies).[5] Combining fuel types into one indicator shows that fuel jumps into the 5 most important variables, just behind plant size.

The results show that higher project costs have an inverse correlation to chances of successful implementation. Yet due to differences in cost between fuel types and possibly other factors, a plant's capacity instead has a direct correlation with chances of commissioning: the largest plants have a higher chance of successful implementation. The disparity between cost and capacity in Figure 2 may also be caused by the occurrence of cost overruns in long-term project development cycles, especially in hydroelectric power stations. A subsequent version of this model ought to include time-specific features relating to project commencement and completion.

Different forms of asset ownership have varying effects on project implementation. Plants owned by multinational companies or other states (Build, Own, Operate model) have the highest degree of success, while Joint Stock Companies (JSC) with high-proportion of state-owned shares have a smaller chance of implementation than privately or foreign owned assets. Yet JSCs have a greater chance of implementation than plants directly owned by a state ministry or vertically integrated state-owned operator. Thus, market structure not only has an impact on a state's ability to diversify its fuel sources (Figure 1), it also factors into the level of enthusiasm for investment in a state's power sector.

Financing types have varying degrees of importance to project commissioning. A wholly state-financed asset, such as the Derweze Simple Cycle Power Plant in Turkmenistan, has the highest chance of being commissioned. 88% of self-financed projects in Central Asia are single-cycle natural gas power plants in fossil-fuel exporting countries, particularly Turkmenistan, showing that Central Asian states are highly able to leverage their export revenues for local electrification projects. This is also shown in the fact that projects undertaken in countries with higher levels of fossil fuel rents as a proportion of their GDP have a greater probability of successful implementation.

Other forms of funding have a significant impact on plant realization. Projects receiving multilateral funding are more likely to be commissioned than bilateral projects, despite the former having higher transaction costs than the latter.[6] Multilateral funding comes with a few advantages: for one, lenders are able to distribute risk and are thus more able to absorb complications during the project cycle; in addition, multilateral organizations offer loans with reform conditionality, meaning that recipients of aid often need to make commitments to improve transparency, governance, and procedures, all of which are shown to have a positive effect on project implementation.[7]

Country factors are less significant but still relevant to project implementation. These include political stability, GINI coefficient, corruption score, democracy score, population, debt/GDP ratio, and, apparently, whether a country is Armenia (this is an imperfect model).[8] Political stability is the most significant country-level factor—while it is fairly intuitive to recognize that politically volatile situations discourage investment, it is surprising that instability does not affect project outcomes nearly to the same extent as fuel types or project costs. In addition, as political stability has a stronger effect in project realization than democracy score, it becomes clear that strong autocracies are better for development outcomes than weak democracies or hybrid regimes. At the country level, it can be summarized that states with lower inequality, lower corruption, lower debt, and a larger population (meaning a bigger potential energy market) increase the likelihood of a project being commissioned.

4. Conclusions: Model limitations & future areas of improvement

Because of challenges encountered during data collection and model creation, this study requires a significant literature review to understand key components of Central Asia's energy system and

[5] G. Alova, P.A. Trotter, A. Money. "A machine-learning approach to predicting Africa's electricity mix based on planned power plants and their chances of success." Nat Energy 6 (2021): 158–166

[6] P. Biscaye, T. Reynolds C. Anderson. "Relative Effectiveness of Bilateral and Multilateral Aid on Development Outcomes." Review of Development Economics 21 (2016): 1425-1447

[7] P. Biscaye, T. Reynolds C. Anderson. "Relative Effectiveness of Bilateral and Multilateral Aid on Development Outcomes." Review of Development Economics 21 (2016): 1425-1447

[8] The more Armenian your country is, the less power. . .

best practice and policies for sectoral reform. In attempting to design original research through data analysis, we found five major limitations to the model's current design: first, there is a lack of comprehensive data available for power plants in Central Asia, making it necessary to include Caucasian and Easterm European states to have enough inputs to adequately train the model. Second, Central Asian countries are not transparent with their energy data, leading to two sub-issues: available data doesn't include all current projects, and governments are slow to publicly release cancellation announcements. Third, the country-level data has changed over time. Due to countries' growing experience engaging in bilateral negotiations with China through the Belt and Road Initiative, multiple states—especially Uzbekistan—have greatly improved their governance, ease of doing business, and other indicators.[9] Our model does not account for the change factor of these indicators. An area of improvement would be structuring projects in the dataset by date of project commencement/commission and modeling the rate of change of probabilities over 5 and 10 year periods to create a more accurate prediction of future scenarios. Fourth, since we are trying to predict the importance of 48 columns (features) using merely 93 rows of data, the data fitting accuracy of our models is limited. Finally, feature selection requires multiple iterations of training to test whether features have a negative impact on the model's ability to calculate probabilities; thus, more work needs to be done to fine tune and separate features (i.e., is GDP a better indicator than GDP per capita? Why is being Armenian important to the model's outcome?).

[9] D. Burghart & T. Sabonis-Helf. Central Asia in the Era of Sovereignty: The Return of Tamerlane?. (New York: Lexington Books, 2018), 195-196.