6 October 2024
Sujith Roy S

Rapid Feasibility Assessment: Modeling the Likelihood of Infrastructure Project Delivery

Objective
To assess the feasibility of predicting the likelihood of infrastructure project delivery using data from the Pipeline dataset and easily accessible external sources.

Overview
The Pipeline dataset contains information on 5,940 infrastructure projects, including project status, funding status, procurement methods, timelines, and geographic details (latitude, longitude).

Approach
The brute force approach involved creating a weighting system to assign likelihoods based on project sector, region, funding source, and duration spent in specific phases. However, the dataset lacks crucial information, such as how long a project remains in specific statuses (e.g., "On hold," "In Development") or reasons behind the statuses. These factors would have made the modeling more straightforward.

While it is possible to manually extract status durations, there were insufficient data points to model the 'likelihood' of completion directly. Thus, we focused on an alternative approach—predicting total delay as a proxy for project delivery likelihood.

To support this, several features were engineered from the Pipeline data:

- complete: Indicates whether the project is complete.
- planned_duration: Planned project duration.
- actual_duration: Actual duration of the project.
- budget_min and budget_max: Estimated budget range.
- funding_status_indicator: Whether the funding source is confirmed.
- project_status_indicator: Indicates whether the project is "In planning."
- total_delay: Calculated as the difference between the actual completion date and the estimated completion date for completed projects.

Predicting total delay allows us to estimate the likelihood of project delivery, with higher delays implying lower chances of on-time completion. The predicted delays were then scaled to a likelihood percentage (0-100%) for each project region and sector, normalizing the values to assess the probability of on-time delivery.

Methodology I (LM1)
We developed a linear regression model (LM1) to predict total_delay using features from the Pipeline data:

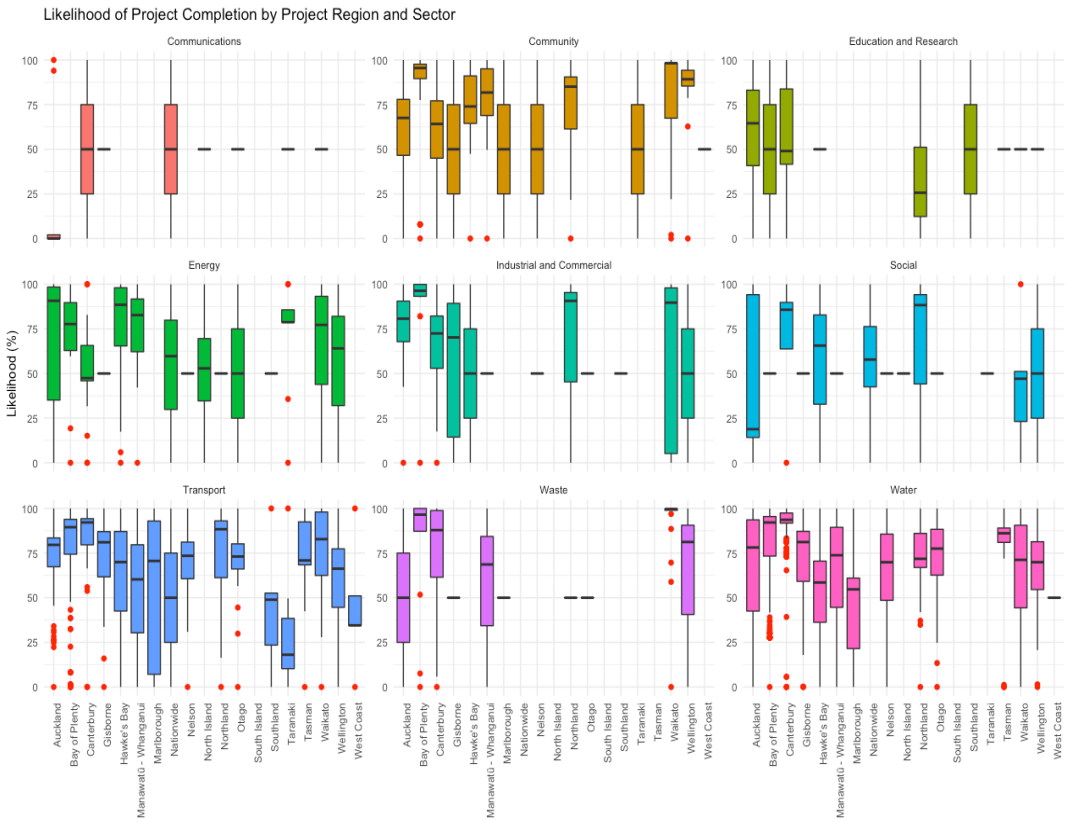
- avg_group_delay (average regional delay)

- budget_min and budget_max (estimated project budget)
- planned_duration (planned project duration)
- funding_status_indicator (whether funding is confirmed)
- project_status_indicator (project's planning status)

The predicted delay values were scaled within each region and sector, converting them to a likelihood percentage for project delivery.

Results

The fit for LM1 was not strong enough to use in production or for decision-making. The likelihood distribution is all over the place failing to account for extreme cases.



Methodology II (LM2)

We extend the model to include macroeconomic factors like CPI and GDP. These macros are available on Treasury's website and more features like 'CPI_Delta Percent', 'GDP_Delta_Percent' were engineered and used in addition to features for LM1. This second model was primarily tested on Transport Sector projects, where economic indicators have a known impact on project delays.

Reasons for focusing on the Transport Sector:

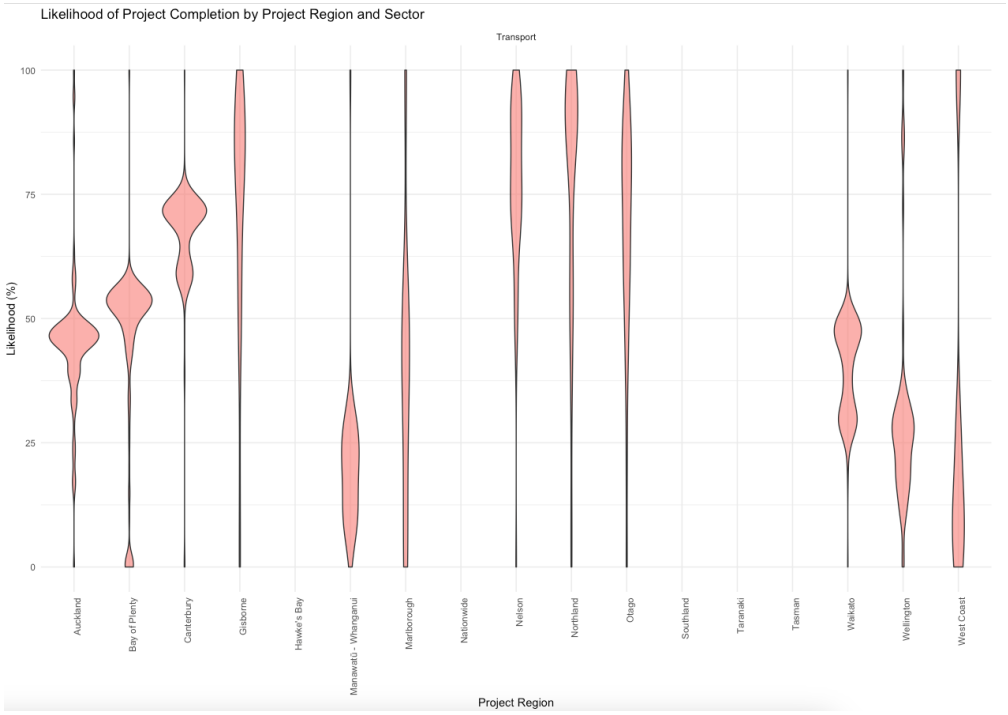
1. Transport projects are largely funded by revenue from the National Land Transport Fund (NLTF).

2. As the lead for NLTF modeling at Transport, I am aware that NLTF revenues are driven by macroeconomic factors such as CPI and GDP.
3. It is assumed that a lower-than-expected revenue due to changing macro-economic conditions may cause delays in the Transport sector projects.

Another important assumption to note is the macro data is available till 2028-06 and the same was used for periods beyond that period.

Results

The LM2 model showed an improved fit, with an R^2 value of 80%, meaning that the input variables could explain 80% of the variance in total delay. The graph below shows the distribution of likelihood project delivery across regions for the planned projects in Transport sector.



Recommendation

We strongly recommend using the LM2 model due to its higher accuracy. A similar approach can be used to for other sectors. Further optimization is possible by incorporating sector and region-specific macroeconomic factors, such as using transport-specific CPI instead of general CPI, to better capture the nuances of project delivery. Growth rates can be assumed for CPI and GDP beyond 2028 would yield more realistic outcomes.