An application of an ensemble of prediction methods for estimating the cost of buildings, land, and infrastructure for Australian Department of Defence projects.

Adrian Mitchell and John Millhouse.

QinetiQ Australia

asmitchell@qinetiq.com.au

Developing early conceptual cost forecasts of new ideas is an essential activity for organisations with long-term (twenty or more years) planning horizons. Predicting the likely cost of ideas before the organisation establishes a project, requirements are frozen, the scope is understood and agreed, and design activities have commenced is inherently risky yet unavoidable for long-term budgeting.

This paper proposes the use of an ensemble of machine learning or artificial intelligence prediction methods for generating conceptual cost estimates. The findings for estimating the cost of constructing buildings and infrastructure and purchasing land are outlined in this paper, along with comparisons to linear regression methods.

Bottom-Up Methods

Simply accumulating cost data using traditional or bottom-up costing methods, as employed by accountants or quantity surveyors, is infeasible at the early conceptual stage, because.

- There usually is not enough time for cost practitioners to develop a detailed work breakdown structures. In addition, there is typically no project office to support fine-grained costing efforts at the conceptual stage.
- The cost practitioner must define every cost element through numerous rate and volume calculations.
- Subject matter experts could base educated guesses on similar past projects of comparable scope, but subjectivity would undermine their confidence in cost forecasts.
- Detailed specifications or requirements are typically limited or non-existent.
- The time-consuming nature of traditional approaches limits the number of options that can be provided to decision makers.
- The use of lookup tables or reference material to generate rate and volume calculations often requires costly specialist skills.
- Bottom-up estimates are usually not informed by statistical forecasting or analysis techniques to assess likely prediction errors.
- In the absence of a repeatable software-driven process, costing relies heavily on the cost practitioners' expertise and adherence to cost best practices. In the authors' experience the resulting cost forecasts are mostly guesswork.

The track record of traditional costing methods is peppered with gross underestimates. Optimistic cost forecasts for large, complex projects even in mature planning, design or production stages are problematic enough; unrealistic conceptual estimates and commensurate funding shortfalls must accept a significant share of the blame.

Researchers find that cost practitioners underestimate costs in almost 9 out of 10 civil infrastructure projects, and for a randomly selected project, the likelihood of actual costs being larger than estimated costs is 86 per cent[1]. Moreover, the propensity for facilities and infrastructure project cost overruns has remained similar since the 1950s[2].

There is a substantial body of academic and industry research into the reasons for consistent project cost growth. One finding suggests the root cause is behavioural biases, and traditional cost methods produce overly biased or optimistic cost forecasts[3].

The argument proceeds that forecasting errors should be expected to occur randomly, and project 'before and after' costs should approach a normal distribution. However, project cost results typically exhibit highly skewed distributions, such as a log-normal, suggesting that behavioural biases are perhaps at play.

There is little objective evidence to suggest that detailed bottom-up estimates are more accurate than those that might be produced by other methods. Research in 2011 by the Jet Propulsion Laboratory involving 507 persons found that bottom-up estimating was often little better than guessing[4]. The study found that "…deep decompositions do not improve accuracy…", that they are "…more time consuming than helpful", and "…compound psychological effects" by biasing the cost practitioner towards optimistic outcomes. This industry finding is supported by academic literature: simple models tend to produce more accurate results than detailed bottom-up models[5].

Therefore, different cost methods are required to reduce the chances of optimism bias on the part of the cost practitioner and project sponsors.

Parametric Methods

Unlike the construction industry, the ruling paradigm for generating conceptual estimates in defence industry and government organisations is the application of statistical or parametric cost methods. The defence industry generally acknowledges this costing approach as best practice[6].

Statistical or parametric methods typically use various types of linear regression models from quantitative data, including dummy variables to represent the qualitative differences in the dataset. Some of the benefits of this approach include as follows:

- In a properly developed statistical cost model there is a high-quality link between macro technical requirements and likely costs, making it much easier to estimate conceptual designs.

[1] Bent Flyvbjerg, Mette Skamris Holm and Søren Buhl, "Underestimating Costs in Public Works Projects: Error or Lie?" Journal of the American Planning Association, vol. 68, no. 3, Summer 2002, pp. 279-295.
[2] Bent Flyvbjerg, "The Fallacy of Beneficial Ignorance: A Test of Hirschman's Hiding Hand" World Development Vol. 84, pp. 176–189, 2016.
[3] B. Flyvbjerg et al., 'Five things you should know about cost overrun', Transportation Research Part A 118 (2018) 174–190, Table 3.
[4] Jordan Gardner and Arthur Chmieleski, "Why Good Engineers Give Bad Estimates: Results of Psychological Research", 22 February 2012.
[5] B. Flyvbjerg et al., 'Five things you should know about cost overrun', Transportation Research Part A 118 (2018) 174–190, pp. 185 – 186.
[6] International Cost Estimating and Analysis Association, Cost Estimating Body of Knowledge 2013, Unit 1 – Module 2, Slide 21.

- The estimating process is more time-efficient than traditional approaches, facilitating development of multiple options with limited time and resources.
- A statistical approach supports probabilistic modelling.

The evidence that this approach has reduced project overruns seems mixed. Selected Acquisition Report milestone data for 225 planning and development projects and when normalised for changes in production quantities[7] indicates that the probability of completion on or below the original project cost forecast is only 35 per cent[8].

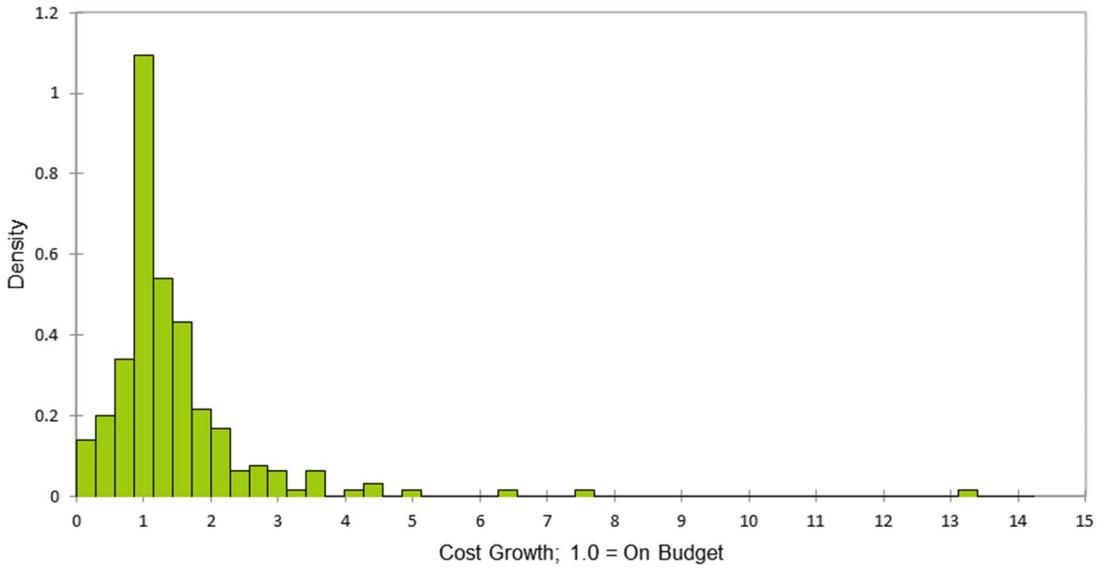

Figure 1 – United States Department of Defence Project Cost Growth Results, 1983 – 2019

While this percentage may be a disappointment, several factors should be considered:

- The cost growth outcomes for Defense projects is an improvement over sizeable civil infrastructure projects, if hardly an outstanding achievement.
- There is no certainty that conceptual estimating in the United States Defense Department is following best practice in use of statistical cost methods.
- We have no data on what confidence level is assigned to each deterministic estimate and what provisions, if any, have been made for uncertainty and risk.
- The 2019 Selected Acquisition Report indicates that estimating errors account for 6 per cent of project cost growth across 84 projects[9]. Prediction failures in relation to inflation and foreign exchange rates account for another 20 per cent, and it may be unreasonable to expect a cost practitioner to predict economic parameters years from project delivery.

Table 1 - Project Cost Growth Causes (normalised for changes in production quantity)

Cause of Cost Growth	Per Cent
Engineering related changes	34
Economic forecasting issues with predicting inflation and foreign exchange	20

[7] Changes to production quantities are often beyond the power of the project to influence due to political decisions or a desire to continue a production run to meet an operational requirement.
[8] Statistical significance is P <0.0001.
[9] When data normalised for changing production quantities

Cause of Cost Growth	Per Cent
Schedule changes with cost implications	18
Support	16
Estimating errors	6
Other	6

Notwithstanding these factors, there is a downside with relying solely on linear regression models. Intuitively, it makes more sense to test various prediction methods to generate the best or most accurate prediction possible instead of just one. Linear regression is one method, and there is no evidence that linear regression by itself is more precise than other prediction methods developed over the last twenty years.

Machine Learning Methods

The authors sought to ascertain whether a modern machine learning (ML) method, or an ensemble of such methods, could match or exceed the performance of methods in current use.

ML algorithms are many and varied. The effectiveness of predictive modelling is a matter of how well a chosen ML model or ensemble of models performs when trained on datasets applicable to particular use cases. How best to choose remains an open question. Trial and error have been the ruling paradigm.

Evaluation of numerous different ML models on a given dataset enables the creation of a list of candidates sorted from most to least accurate (e.g., in terms of root mean squared error). However, the best performers may (and probably will) slide down the rankings when evaluated on a different dataset. The varying performance of prediction methods is the motivation for adopting so-called "super learner" ensembles[10].

The super learner technique begins with defining a k-fold cross validation split of the dataset, followed by an evaluation of different models (and model configurations) on the same split. The data scientist then uses out-of-fold predictions to train a 'meta-model' which is highly likely to perform better than any base model. This approach is an example of a general method called 'stacked generalisation' or 'stacking' for short. Typically a linear model is used as the meta-model.

The following table outlines some of the critical functional differences between cost prediction methods. In addition, a linear regression by itself needs high-quality quantitative data where outliers are often treated with suspicion by the cost practitioner. If outliers are excluded from the model, the ability of the model to produce a realistic prediction is undermined, even though the model may be statistically pleasing. By contrast, a super learner is less susceptible to outliers due to the ensemble of different methods used to train and test data.

Table 2 – Method Comparison

Can the Method Model?	Super Learner	Bottom-Up	Linear Regression
Complex, massive, messy datasets	Yes	No	No

[10] First proposed by Mark van der Laan, Eric Polley, and Alan Hubbard from Berkeley in their 2007 paper titled *Super Learner*, published in the journal *Statistical Applications in Genetics and Molecular Biology*, Volume 6 Issue 1.

Can the Method Model?	Super Learner	Bottom-Up	Linear Regression
…when relationships are not clear	Yes	No	Yes
…despite missing values	Yes	No	No
Model qualitative data	Yes	No	Not only
Missing categorical values	Yes	No	No
Non-linear relationships	Yes	No	No
Forecast residual and error results	Yes	No	Yes
Model feedback loops	Yes	No	No
Small data sample (rows) with many predictors (columns)	Yes, but less accurate	Yes, but not accurate	Yes, but not valid
…and are the results easy to interpret?	Not always	Yes	Yes, with simple models
Quick turnaround estimates	Yes	No	Yes

In 2021-22, QinetiQ Australia embarked on a project with the Australian Department of Defence to develop a super learner software application to forecast conceptual whole-of-life costs for a range of building, land, and infrastructure products. The application (named EstatiQ) allows subject-matter experts, without any costing expertise, to enter a range of mostly qualitative information into the application to generate estimates. The time to generate a model is a few minutes, permitting options development without the pretence of a lengthy, detailed and expense bottom-up model building process unlikely to be more accurate than any other method.

Unlike bottom-up methods, ML cost forecasts can be demonstrably "accurate" within the constraints of available evidence and analytical technologies. The cost practitioner can judge accuracy in terms of dispersion and bias of forecasting "errors" within a model by comparing different modelling techniques using standard metrics such as the root mean squared error or coefficient of determination (a.k.a. R2, a measure of how well the data explains the forecast).

EstatiQ Building Model

The building model uses 28,000 rows of Australian Department of Defence construction cost and technical data. The range of data fields was progressively refined during a year-long process of cleaning and enriching the dataset, feature engineering, model training and model testing. The results are outlined in the following figure.

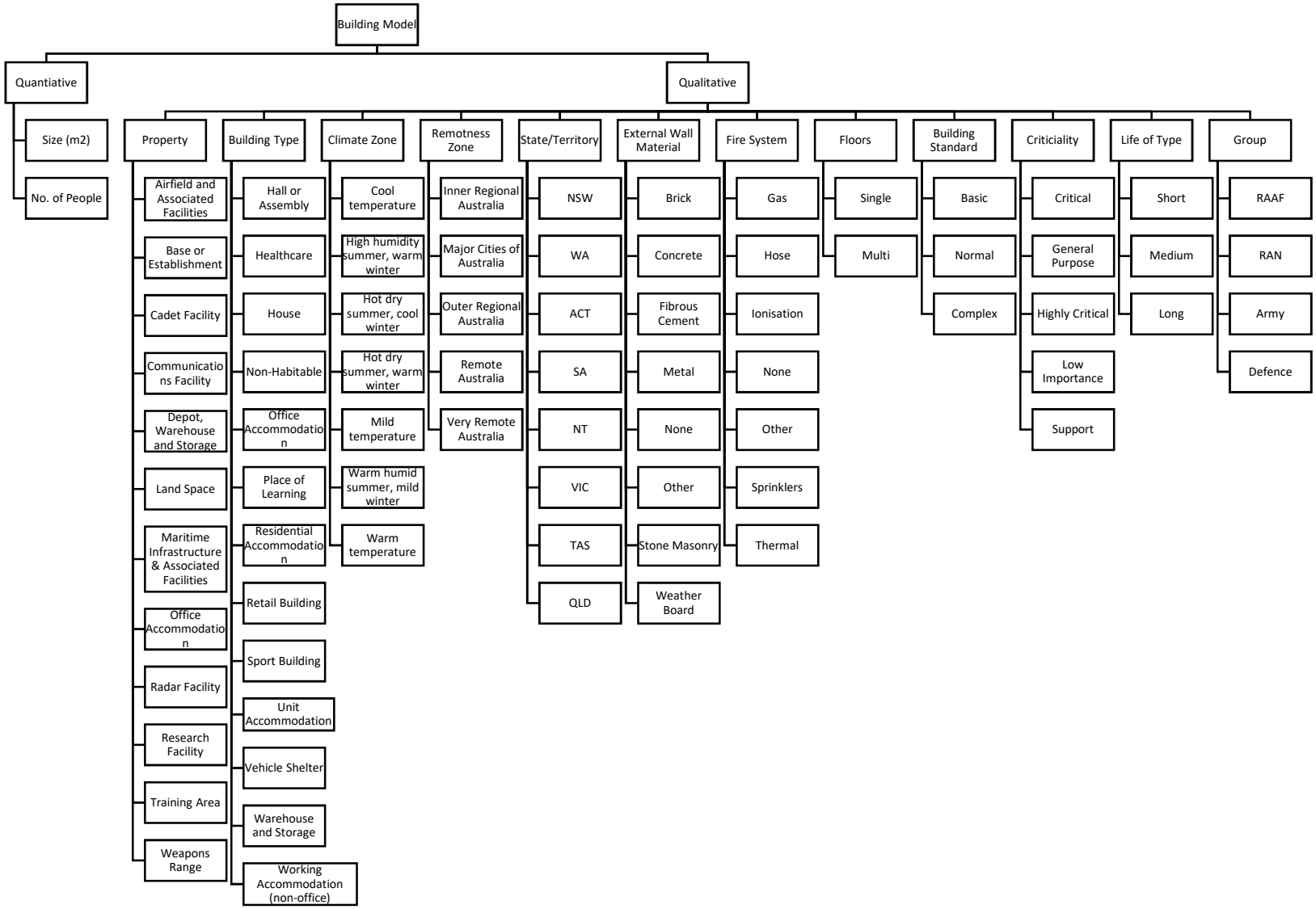

Figure 2 – Construction Cost Building Dataset

Due to data quality challenges and the "curse of dimensionality", the initially numerical Life-of-Type and Floors fields had to be transformed into qualitative (a.k.a. categorical) fields. The curse of dimensionality occurs when there is a paucity of detailed and consistent data; as dimensionality increases, the number of data points required for good performance of any ML algorithm increases exponentially. For example, the data for Life-of-Type has values ranging from 5 to 99 years, but there are not enough examples in the dataset of most of the in-between possibilities to support reliable predictions. The overall accuracy of the model improves measurably on substitution of "Short", "Medium" and "Long" categories for numeric Life-of-Type values and "Single and "Multi-Storey" for numeric building Floors. Predictions using these categorical data fields also tend to be more reliable and consistent.

The following diagram outlines the results of the building model. Linear regression is noticeably less accurate than the super learner.

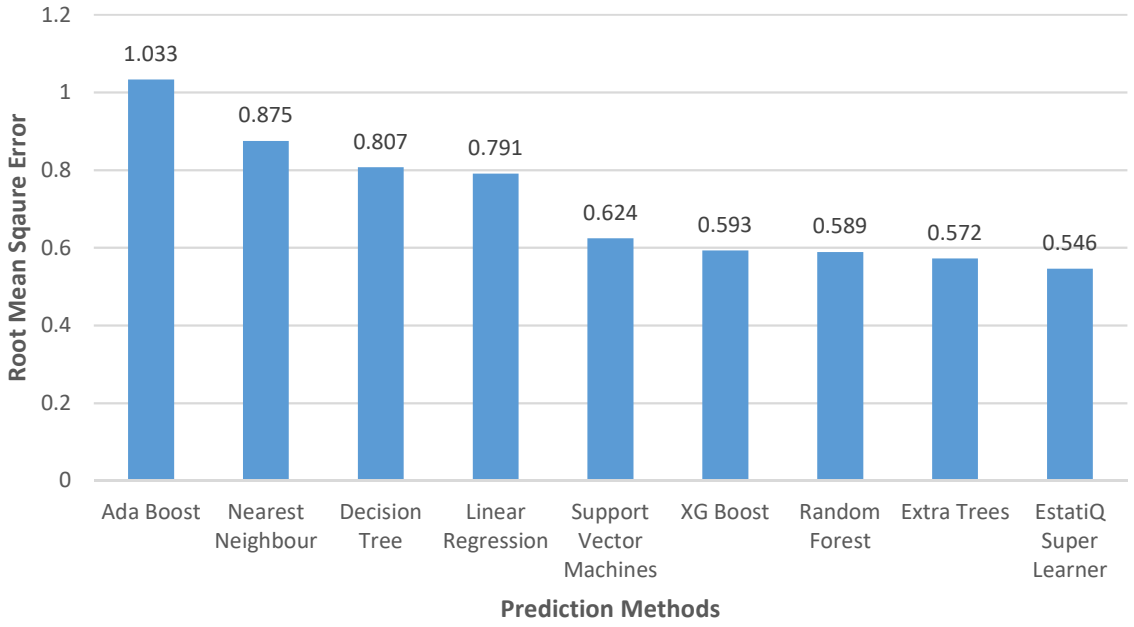

Figure 3 – Building Model Results

The coefficient of determination for the building super learner model is 95 per cent. Despite this result, the super learner struggles to make predictions beyond the scope of the training data due to:

- The heavy reliance of the super learner on ensemble methods such random forest regression,
- the comparatively weak performance of the linear regression method, and
- A lack of reliable data for buildings larger than 10,000 square metres. More data will need to be collected to improve model prediction results for large buildings.

EstatiQ Land Model

The land model makes use of 106,000 rows of $ per m^2 data. In addition to climate and remoteness zones and state and territory variables outlined in Figure 2, a socio-economic index by postcode

from the Australian Bureau of Statistics is included in the dataset as a proxy for local spending power. Other variables are outlined in the following table.

Table 3 – Land Categories

Land Usage	Land Size
Commercial	Vast (>1,000,000 m2)
Health, Care and Community	Large (>100,000 m2)
Industrial	Medium (>1,000 m2)
Mixed Use - Urban	Small (>100 m2)
Other	Tiny (<100 m2)
Parks/Reserves	
Primary Production	
Residential	

Predicting land value with a high degree of accuracy is a difficult task[11]. Initial testing and training yielded questionable results. The size of land measured in m^2 showed a surprisingly weak correlation with purchase prices, as outlined in the following figure.

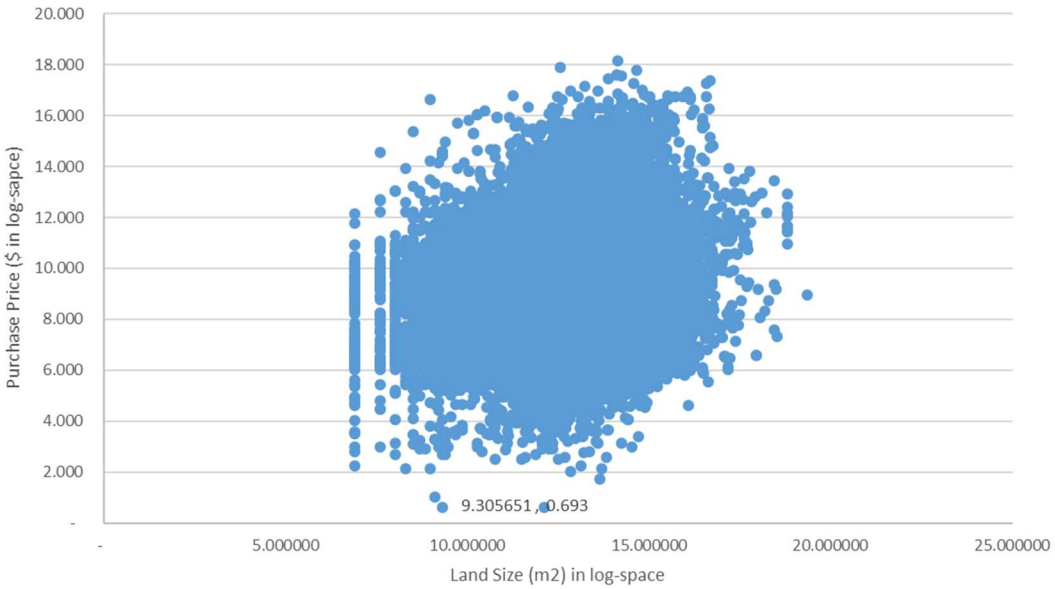

Figure 4 - Scatter Plot for Land Size and Price Data

Accordingly, the Land Size variable shown in Table 3 was created to substitute qualitative indicators for numeric values. This strategy reduced the training dataset to rows of solely categorical features, necessitating removal of linear regression from the land super learner ensemble.

The land model was reconfigured to predict $ per m2 rather than an absolute dollar price. EstatiQ takes the $ per m2 prediction and multiplies it by the user-entered land size to provide a land price.

[11] See the following report of a more extensive effort of predicting land prices, where R^2 scores of 63 per cent were achieved using log-linear regression: https://www.awe.gov.au/sites/default/files/documents/MeasuringAustralianBroadacreFarmlandValue20191213%20_v.1.0.0.pdf

The following figure shows the results for the random forest regression method using only categorical or qualitative data.

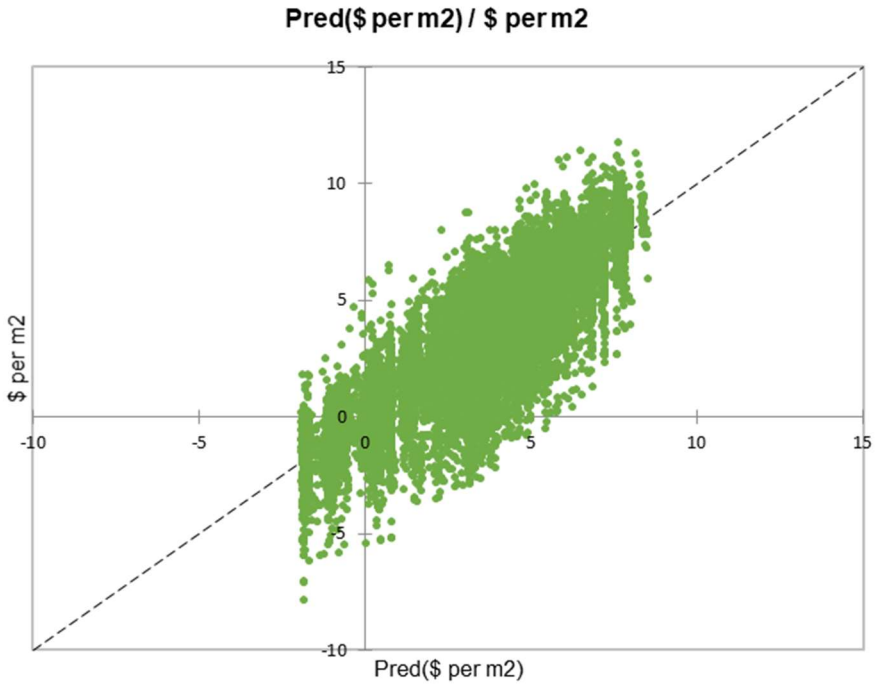

Figure 5 - Land Model Results using Categorical Data

The following figure shows the results across all prediction methods using only categorical data.

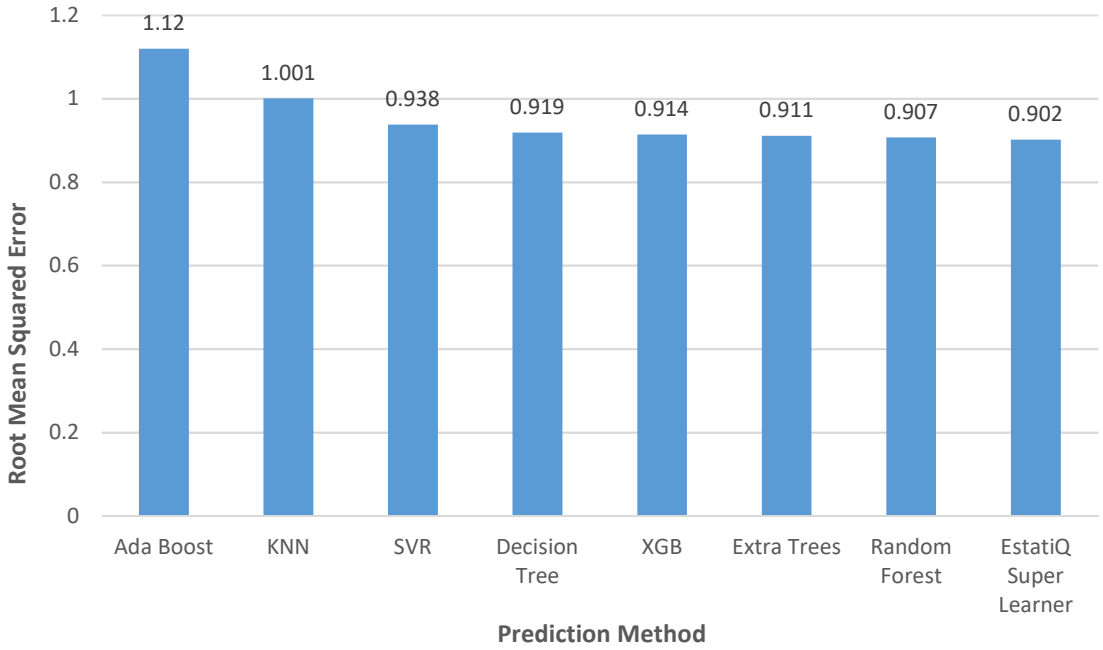

Figure 6 - Land Model Results

The coefficient of determination for the land super learner model is 85 per cent.

EstatiQ Infrastructure Model

The infrastructure model is based on 100,000 rows of data, summarised down to 9,000 construction cost rows due to missing data and other data quality issues. The infrastructure data presents two unique challenges.

1. Uncommon infrastructure categories

Infrastructure is a broad category that compasses 46 asset types, many of which are seemingly unrelated, such as: Airfield, Communications, Detached Outdoor Structure, Electrical, Exercise, Fencing, Fuel/Lubricant/Chemical, Gas, Hydraulic, Marine, Military Training, Road / Pavement / Path, Security System, Waste Storage.

Variables are created that are common across the dataset, as outlined in the following table.

Table 4 - Infrastructure Common Variables

Profile	Primary Construction Material	Construction Complexity
Elevated	Ceramic	Basic
Subterranean	Composite Rock[12]	Normal
Surface	Earth[13]	Complex
	Metal	
	Plastic	
	Silicon	
	Wood	

Table 4 results in a feature engineering process that creates a matrix of common variables cross-referencing each of the 46 infrastructure asset categories. For example, assets with a surface profile include:

- A runway that is concrete = Complex
- A road that is bitumen = Normal
- A grass playing field = Basic

This process also helps predict the cost of assets when data is limited. For instance, EstatiQ contains data for only a few dry docks. To generate the cost of a new dry dock, the EstatiQ super learner essentially combines all subterranean assets of a specific material and complexity into a single prediction, from swimming pools, bunkers, conduits, etc.

The infrastructure data also reuses variables from the building model, including State/Territory, Climate and Remoteness Zones, Criticality Factor, Group and Asset Life of Type to provide additional explanatory power.

2. The infrastructure data set lacked sufficient sizing data to support the implementation of a plausible data imputation strategy.

A categorical sizing variable is created from residuals generated during the testing and training of the model. Residuals in the initial testing of the model are assumed to be missing sizing variables. The

[12] Concrete, asphalt or bitumen
[13] Grass or dirt.

sizing labels (Large to Small) are created and then checked against the few sizing data observations that do exist, to create handy lookup tables for the end-user. Various rounds of testing and training models were required to provide a satisfactory result.

This approach to sizing an infrastructure asset is less than ideal and would probably make a quantity surveyor shudder. However, given the current track record of infrastructure project cost overruns, there seems little point in providing any more precision in the model inputs during a proposal's conceptual phase.

As with the land model, linear regression was excluded from the infrastructure model due to model only using existing categorical data. Resulting model performance is as follows.

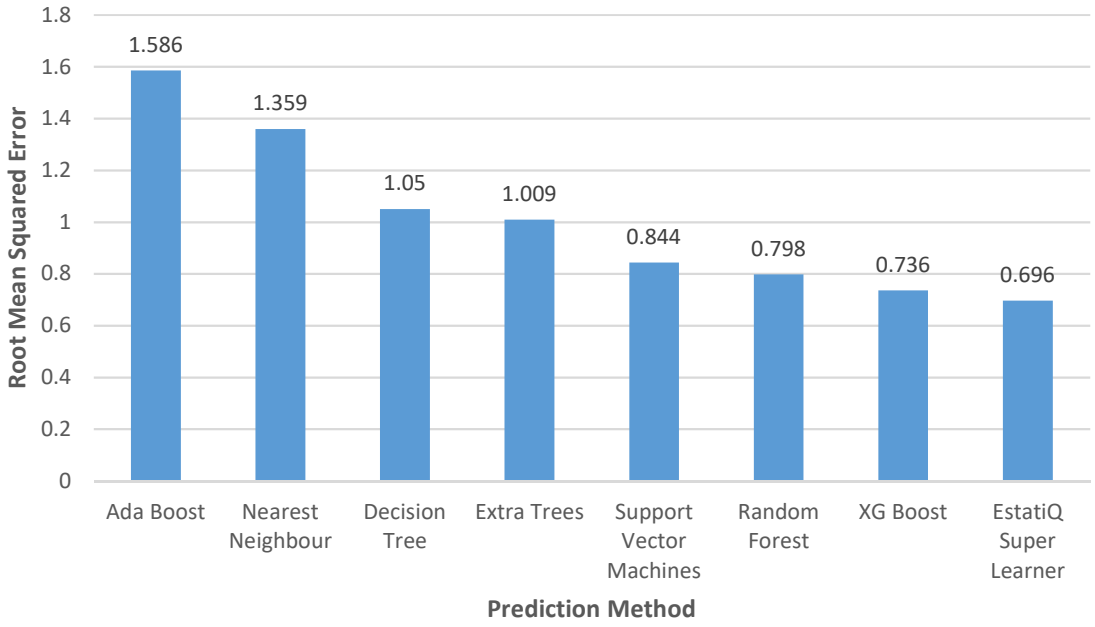

Figure 7 - Infrastructure Model Results

The coefficient of determination for the infrastructure super learner model is 88 per cent

Conclusion

This study has shown how reliance on traditional bottom-up, parametric or linear regression costing methods may limit the ability of cost practitioners to generate conceptual estimates as readily and reliably as the best approaches now available. Given past and current track record of cost overruns for large and complex projects, cost practitioners and other subject matter experts without specialist costing expertise are entitled to explore alternative paradigms. ML-based prediction methods can make effective use of a broader range of data, including creative categorical information; skilful feature engineering followed by straightforward training and testing of models can deliver highly productive and consistent cost forecasting solutions.