Big Data with Deep Learning for Benchmarking Profitability Performance in Project Tendering

for

Expert Systems with Applications

(Journal)

By

Muhammad Bilal [1,*]; ***Lukumon O. Oyedele*** [2,*];

(Authors)

**** Corresponding Author***

Affiliations

[1] Associate Professor of Big Data & Artificial Intelligence (AI),
Big Data Enterprise and Artificial Intelligence Lab (Big-DEAL)
Bristol Business School, University of the West of England (UWE), Bristol,
Frenchay Campus, Coldharbour Lane, Bristol, BS16 1QY, United Kingdom
Tel: +44 (0) 117 32 84253
E-mail: Muhammad.Bilal@uwe.ac.uk

[2] Director of Big Data Enterprise & Artificial Intelligence Lab (Big-DEAL)
Professor & Chair in Enterprise and Project Management
Bristol Business School, University of West of England, Bristol
Frenchay Campus, Coldharbour Lane, Bristol, BS16 1QY, United Kingdom
Tel: +44 (0) 117 32 83443
E-mail: Ayolook2001@yahoo.co.uk; L.Oyedele@uwe.ac.uk

Big Data with Deep Learning for Benchmarking Profitability Performance in Project Tendering

Abstract

A reliable benchmarking system is crucial for the contractors to evaluate the profitability performance of project tenders. Existing benchmarks are ineffective in the tender evaluation task for three reasons. Firstly, these benchmarks are mostly based on the profit margins as the only key performance indicator (KPI) while there are other KPIs fit to drive the evaluation process. Secondly, these benchmarks don't take project context into account, thereby restricts their predictive accuracy. And finally, these benchmarks are obtained from small subsets of data, making it hard to generalise. As a result, estimators cannot probe into tenders to judge the strengths and weaknesses of their bids. This advancement is critical for not only choosing more lucrative opportunities but also driving negotiations during the tendering process.

This study aims to develop a benchmarking system for tender evaluation using Big Data of 1.2 terabytes, comprising 5.7 million cells. A holistic list of seventeen (17) KPIs is identified from the email data using Text Mining approaches. Besides, eight (8) key project attributes are chosen for ensuring context-aware benchmarking using Focused Group Interviews (FGIs). At the crux of this work lies the proposition of a deep ensemble learner based on the decomposition-integration methodology. In the decomposition stage, the model predicts several attribute-specific benchmarks for each KPI using our proposed context-aware algorithm. In the integration stage, deep neural network-based learners are trained to generate final project-sensitive KPI benchmark. The learner is deployed in the `Spring` tool to support the tender evaluation of power infrastructure projects. A tender of $60km$ underground cabling project is evaluated using the proposed learner. The system spontaneously identified KPIs in the tender that require further attention to achieve greater profitability performance.

Keywords: *Big Data, Project Tendering, Text Mining, Deep Learning, Benchmarking, KPIs*

1. Introduction

1.1. The Issue of Profitability Performance

The project-based industries are becoming more competitive with firms operating at low profit margins (Fadhil Dulaimi, 2005; Tam et al., 2004). Despite immense efforts, these firms struggle to make reasonable profits. Table 1 verifies this fact for the top construction contractors in the UK. Majority of these firms went into losses. Their combined profit for the year 2017 is -5.41%, which signifies that these firms incurred £1.94 billion lost despite a total turn over of £36billion. For similar reasons, some project-based industries like construction are often ranked highest by the company insolvency index, with 14.4% firms getting bankrupt yearly (Alaka et al., 2018). Such margin erosion across the bulk of their projects calls for the adoption of digital technology to achieve precision in their project planning and control activities.

The profitability performance in power infrastructure projects is trickier as these projects tend to be more complicated and span a wider geographical area. This geographic dispersion brings lots of risks to projects that are hard to quantify at the early tendering stage (Cheng & Roy, 2011). The unanticipated risks occur during the project delivery that are often covered from the planned margins (Makovšek, 2014; Taroun, 2014). Common risks include weather surprises, schedule changes, increased demand for resources, inoperable land for plants, plant damages and fixed outage dates. Also, the clients use the competitive nature

Table 1: Top 10 Construction Players in the UK—Source:TCI (2017)

Rank By Turnover	Company Name	Turnover (£m)	Pre-tax Profit (£m)	Pre-tax Profit Margin (%)
1	Balfour Beatty	8,683.00	8	**0.09**
2	Carillion	5,214.20	146.7	**2.81**
3	Kier Group	4,211.00	-15.4	**-0.37**
4	Interserve	3,685.20	-76.4	**-2.07**
5	Morgan Sindall	2,561.60	43.9	**1.71**
6	Amey UK	2,531.00	-43.9	**-1.73**
7	Laing O'Rourke	2,513.20	-245.6	**-9.77**
8	Galliford Try	2,494.90	135	**5.41**
9	Mitie	2,126.30	-42.9	**-2.02**
10	Mace	2,041.10	10.7	**0.52**
Total Pre-tax Profit Margin (%)				**-5.41**

of the industry to push margins further down. They mostly prefer cheaper tenders. The contractors in the pursuit of winning competitions submit unrealistic bids that eventually end up into losses. Projects began with certain planned margins gets completed with entirely different (low) margins. A robust evaluation system is required for not only quantifying but also validating the core constituents of tenders like the costs, profit, risks, and opportunities.

1.2. The Need for Tender Evaluation

Tender evaluation is an important task for both clients and contractors. The clients wish to get their projects completed faster and cheaper. They need an evaluation system for choosing the right contractors. Tender evaluation from the clients' perspective has been an active area of research (Chen & Pan, 2018; Rao et al., 2018; Samuel, 2018; Watt et al., 2010; Wong et al., 2003; Watt et al., 2009a). The contractors, on the other hand, need an evaluation system to ensure the submitted bids have high quality and profitability performance. Tender evaluation from the contractors' perspective is rarely studied. At present, estimators prepare tenders (as shown in Fig.1), which either lack the form or the substance. They often fail to win the bid due to the inability of their tenders to drive the tender negotiation task. Such instances are undesirable as tendering costs these firms a huge fortune. In the case of winning tenders, contractors find it hard to complete projects due to unrealistic projections. An evaluation system is crucial to support the contractors using data-driven insights during the tender evaluation task to ensure high-quality bids are submitted before the clients.

1.3. Focus of the Paper & Research Methodology

This study aims to harness Big Data with Deep Learning for the development of a robust tender evaluation system. The idea is to facilitate contractors during the evaluation process towards tender completeness and accuracy. The system will allow estimators to understand bids, and to compare and contrast various aspects of tenders using a RAG (red/amber/green) colouring scheme. The underlying objectives of the study are as follows:

1. Identify key tender elements for driving holistic evaluation task
2. Develop deep ensemble learner for tender evaluation using Big Data
3. Evaluate deep ensemble learner for benchmarking projects' profitability performance

This study has employed data collection and analysis strategies for both qualitative and quantitative data. The qualitative data, including emails of the management, sales and delivery personnel of a UK construction contractor are analysed using Text Mining to decide key performance indicators (KPIs) formulating a robust

A	B	C	D	F
	TENDER SUMMARY			
	DIRECT COSTS:			Total
1	Materials			£ 365,956.49
2	Added for Fixed Price			
3	Labour (Incl. Supervision, O.T., T.T., N.P.T., etc.)			£ 84,945.80
4	Added for Fixed Price			
5	Expenses (Incl. Fares, O/A., Accommodation, etc.)			
6	Added for Fixed Price			
7	Bdgs., Plant, Transport, Tools, etc.			£ 62,377.42
8	Sub Contracts			£ 18,904
	TOTAL NET COST (a)			£ 532,183.71
	ONCOST:	% Cost (a)	% Gr. Amt. (b)	
9	Section O/H			
10	Cont. to Branch/Div. O/H			
11	Cont. to Head Office O/H			
12				
13				£ -
	MARK UP:			£ 532,184
14				£ 59,230
15	Gross: Oncost Mark-Up	11.13%		
	GROSS AMOUNT SCHEDULED WORK (b)			£ 591,414
	GRAND TOTAL ©		**TENDER VALUE**	**£ 591,413.88**

Figure 1: An Example Tender Summary of Power Infrastructure Project

project tender. Likewise, focus group interviews (FGIs) were held to choose key project attributes for context-aware benchmarking. An algorithm for deriving attribute-specific benchmarks across eight key project attributes is developed. The algorithm is used to generate benchmarks from Big Data of 1.2 Terabytes. The algorithm yields several benchmarks during the decomposition stage that are then collated with deep neural networks at the integration stage. The proposed model is implemented into the `Spring` system that is a web-based project management tool for supporting whole-life activities of planning and controlling mega construction projects. The system is tested with a case study of $60km$ underground cabling project. During the tender evaluation task, the system is found incredible in guiding estimators to improve the project tenders. The proposed benchmarking approach can be tailored to automating a wide range of similar project-related tasks.

1.4. Contributions

This study is unique in the sense that Big Data with Deep Learning is used for the first time to develop a benchmarking system for tender evaluation. While tender evaluation from a client perspective is largely explored, this study examines the issue from the contractors' perspective. The purpose is to develop an objective system to facilitate contractors in producing high-quality project tenders. Besides, the idea of Opportunity-On-A-Page (`OOAP`) dashboard is introduced for utilising the proposed leaner in real-life scoring of project tendering tasks. Using `RAG` (red/amber/green) encoding, the `OOAP` allows the estimators in apprehending the strong and weak aspects of their tenders at a glance. This study has enormous implications for knowledge and practice.

1.5. Organisation of Paper

The next section provides an extant literature review on KPIs, key project attributes, and the need for Big Data with Deep Learning in tender evaluation. Then, the 5-fold research methodology used in this study is described in Section 3. Section 4 and 5 describe Text Mining and FGIs for identifying critical components of a tender. The description of databases and Big Data integration strategy is explained in Section 6. The proposed deep ensemble learner is introduced in Section 7. The deployment of deep ensemble learners is explained in Section 8. The system is evaluated by a case study of a $60km$ power infrastructure project in Section 9. Section 10 highlights the implications of this research. And finally, conclusions, limitations and areas of future study are presented in Section 11.

Table 2: Uniform Rate-Based Profit Margin Bechmark for Tender Evaluation

Sr.#	RAG Score Description	RAG Score	Profitability Performance
1	Profit margin less than 10%	Red	Poor
2	Profit margin between 10% and 14%	Amber	Average
3	Profit margin 15% and above	Green	Good

2. Literature Review

2.1. Benchmarks and KPIs for Project Performance Evaluation

Project-based industries, like construction, have become mostly competitive. These firms continuously assess project performance to achieve excellence and steadily deliver higher performance. According to Kim et al. (2018), there are four strategies to evaluate project performance. Firstly, firms adopt standards promoted by the Project Management Institute (PMI) and the Global Alliance of Project Performance Standards (GAPPS). Secondly, firms evaluate project objectives against key performance indicators (KPIs) and benchmarks. Thirdly, firms benchmark project performance with standard best practices or lessons learned. And finally, comparing planned vs actual performance using the Earned Value (EV). Existing project management tools implement these strategies in one or another way to support project management tasks (Vischer, 2018). These strategies provide continuous guidelines to enhance the performance evaluation process.

This study aims to revitalise benchmarking and KPIs-based strategy for tender evaluation. Benchmarking and KPIs work in tandem. While KPIs quantify performance from a specific dimension, benchmarks provide a logical framework to distinguish the good and bad performing KPIs (Colwill & Gray, 2007; Busby et al., 2013). The selection of relevant KPIs to benchmark performance is of paramount importance to an effective evaluation system (Lu et al., 2015). KPIs can measure performance along strategic, operational and tactical objectives set out by the project teams Bassioni et al. (2004). Firms use these metrics to understand past performance in a retrospective manner. This enables comparing and contrasting projects to learn lessons and identify best practices. Benchmarking and KPIs strategy is studied by various researchers for (i) developing early warning systems (Kim et al., 2018), (ii) understanding resources efficiency (El-Mashaleh et al., 2007), and (iii) enabling greater control to achieve the smart allocation of project finances (Busby et al., 2013) in the literature. This strategy is recognised to work well in facilitating decision-making tasks towards sustainable project performance.

2.2. Limitations of Existing Tender Evaluation Approaches

The first gap in the literature is the fact that tender evaluation is mainly studied from client's perspective (Falagario et al., 2012; Kissi et al., 2017). The approaches devised so far cannot be adapted for contractors since the evaluation criteria greatly vary. As a result, current project management software used by contractors don't provide an objective mechanism to cover their tender assessment tasks. The evaluation process, based on personal judgements, usually lead to inconsistent outcomes. In general, existing approaches fall short in the following three areas:

2.2.1. The Issue of Key Performance Indicators (KPIs)

The choice of accurate KPIs is crucial for the development of a holistic tender evaluation mechanism. By KPIs, we mean items that can support the assessment of tender documents from several critical dimensions. Traditional tenders used by contractors merely capture costs and margins which are used in tandem to calculate the project value. Contractors use profit margins listed in tenders as the single KPI driving their evaluation decisions (Farooq et al., 2018; Domingues et al., 2017). They bid for tenders with substantial margins or skip the projects otherwise. Table 2 shows an example benchmark used by contractors for profit margins. This monolithic approach is not capable of driving a reliable evaluation process since profit margin alone can't explain all aspects of a project (Alaka et al., 2017). In particular, project complexity involving risks, opportunities, and innovations. These KPIs are indispensable for robust tender assessment, but barely, any existing approach uses these KPIs and thereby lack reliable grounds for tender evaluation (Watt et al., 2009a; Kissi et al., 2017; Watt et al., 2009b). Our discussions with industry professionals revealed that most contractors want to see KPIs quantifying risks, opportunities, innovations and stretched margins during the evaluation process. A project with a reasonable margin can still fail due to the huge risks involved in the project delivery. These KPIs can provide multiple assessment criteria for exploring tenders from several dimensions.

2.2.2. The Issue of Context in Benchmarking

While KPIs quantify critical tender elements, benchmarks provide definitive boundaries for empirically evaluating KPIs. Benchmarks can instantly reveal the performance of KPIs in bids. Currently, there is not much work on benchmarking in the literature. Few authors have employed benchmarking for assessing other project dimensions like waste management (Lu et al., 2015, 2011; Tam et al., 2007) and bid price evaluation (Zhang et al., 2015; Wong & Ng, 2010). In practice, contractors frequently use in-house benchmarks for KPIs like profit margins, as shown in Table 2. However, these benchmarks are commonly derived using a uniform fixed rate, which cannot reflect the real-life project complexities. E.g. profitability performance varies by different projects (Rui et al., 2017). Small projects tend to have higher margins than large projects. A contractor won't bid for small projects until there is more profit in it (Peterson, 2005). The evaluation system based on uniform rate would classify a small project with 15% as a good project, which, in reality, has poor profit, keeping in view an average margin of 36% for small projects. The issue of context is vital to developing reliable benchmarks for KPIs. This inability to factor in context is the fundamental reason for existing benchmarks being unable to drive the tender evaluation process (Lu et al., 2015).

2.2.3. The Issue of Data

Another limitation of benchmarking-based evaluation systems is their ability to generalise and adapt. Benchmarks tend to obsolete frequently and require refreshing when new data are captured. The volume of data also plays a decisive role—most benchmarking solutions are derived from small subsets of data (Lu et al., 2015, 2011; Tam et al., 2007; Zhang et al., 2015; Wong & Ng, 2010). The data used by these systems are often qualitative, gathered through literature, case study, interviews or site surveys. Therefore, the derived benchmarks tend to inherent subjectivity and biases as well as lack generalisability. Most projects' data (i.e. 85%) arising from unstructured sources are barely used by these systems (Bilal et al., 2016). These data sources can enable multi-criteria KPIs based benchmarking and tender evaluation. Contract documents enlisting retention details can be used in the tender assessment to see the implication of payments withheld by clients on project delivery. Tapping into unstructured data sources has become vital for developing a robust tender evaluation system.

2.3. The Role of Big Data and Deep Learning

Big Data is the emerging ability of firms to store, integrate and use different types of large volumes of data in their enterprise solutions (Diebold, 2000; Jacobs, 2009). The term Big Data is considered to have following three features also referred to as 3Vs—including 1) `volume` (terabytes, petabytes of data and beyond), 2) `variety` (different formats like text, sensors, audio, video, graphs and more), and 3) `velocity` (continuous streams of the data). A systematic approach to harness Big Data for business strategy and advantage is the utmost priority of many firms these days (Thomas, 2015; Agneeswaran, 2014; Bonino et al., 2013). More importantly, firms are more curious about analysing extra dimensions of data to bring precision to their project planning and control tasks (Bilal et al., 2016). This study harnessed Big Data of electricity grid projects to develop a profitability benchmarking system for project tendering. This synergistic integration is incredible to revitalise the accuracy of existing tender evaluation systems.

Machine learning (ML) is the toolbox for knowledge discovery from large amounts of data. ML offers supervised and unsupervised algorithms to perform the majority of learning tasks. In supervised learning, the model is presented with $\mathbf{x}$ input features and y output feature(s). The shape of y varies based on the learning task. In classification, y constitutes a vector of scalars, denoting the class labels, whereas it can be a series of continuous values in the case of regression. In supervised learning, the training process finds parameter values (Θ) that can best fit the output vector (y) based on a loss function $L(y, \hat{y})$. The $\hat{y}$ represents the output when $\mathbf{x}$ features are fed into $f(\mathbf{x}; \Theta)$, i.e. the model. In the unsupervised learning algorithms, data is analysed without labels to find patterns like latent subspaces. This work focuses on supervised ML which profitability benchmark prediction is modelled as the regression task.

Neural networks are supervised learning algorithms that provide the basis for all modern deep learning architectures. A neural network is made up of neurons with activation function and parameters $\Theta = \mathcal{W}, \mathcal{B}$ where $\mathcal{W}$ are the weights and $\mathcal{B}$ are the biases. The activation a in a layer is the linear combination of input features ($\mathbf{x}$) with the parameters, followed by an element-wise non-linearity (δ). This is expressed as $a = \delta(\mathbf{w}^T\mathbf{x} + b)$. Most commonly used non-linearities include the sigmoid and rectified linear units. In deep learning, we stack several such layers on top of each other like $f(x; \Theta) = \delta(\mathcal{W}^l\delta(\mathcal{W}^{l-1} \cdots \delta(\mathcal{W}^0 x + b^0) + b^{l-1}) + b^l)$, where, $\mathcal{W}$ represents weights matrix and l is total layers in the network. The layers between the input and output are the `hidden` layers. When a network involve more than one hidden layer, it is called the deep neural network (DNN). Interested readers can find more about mathematical formulation of deep learning in Goodfellow et al. (2016).

In the beginning, neural networks were considered hard to train. They became more popular after 2006 when it is realised that training neural networks in a hybrid fashion (unsupervised then supervised fine-tuning) can result in excellent performance (Bengio et al., 2007; Hinton & Salakhutdinov, 2006). Nowadays, neural networks are trained in an entirely supervised way to greatly simplify the training process. The most common neural networks architectures are convolutional neural networks (CNN) used for image processing, recurrent neural networks (RNN) used for sequence data and fully connected neural networks (FCNN) used for tabular data. We employed FCNN for predicting the profitability benchmarking in this study.

Several authors have tried to utilise Big Data for developing reliable benchmarks to carry out various project-related tasks in the construction industry. Lu et al. (2015) used Big Data of construction waste to create waste generation rates (WGRs) for the Hong Kong industry. They found that WGRs derived from Big Data are more robust and can be confidently used to benchmark the waste management performance of contractors. Bortolazza et al. (2005) showed the effectiveness of Big Data towards the percentage of plans (PPC) analysis. They found that decision trees and neural networks are great tools to glean actionable insights from more data to create strong project planning and control indicators. Lastly, Ogunlana et al. (2010) explored the significance of KPIs from stakeholders (client, contractors and consultants) perspective and pointed out the importance of context in benchmarks and KPIs. They suggested the inclusion of diverse KPIs for performance evaluation than that of the iron triangle, which only focuses on time, budget, and specifications. This research further the field by proposing a comprehensive list of KPIs to facilitate contractors during the tender evaluation task using project-specific benchmarks.

3. Research Methodology

This study has 5-part research methodology to demonstrate an end to end development of an ML system for tender evaluation. The first part involved the use of Text Mining to identify KPIs for supporting a holistic assessment process. We engaged industry professionals in this stage to make this selection. Their views about KPIs were captured through the email responses. The unstructured emails data is then analysed through Text Mining to identify the top-k KPIs for the tender evaluation. Experts are engaged to finalise the list of KPIs. The next part employed Focused Group Interviews (FGIs) for identifying key project attributes that can facilitate the context-aware reasoning for benchmarks the performance of KPIs. Their discussions were recorded, transcribed and then thematic analysis was performed to identify the final list of key project attributes. The selection of KPIs and key project attributes informed the data collection and integration strategy. We collated massive amounts of construction data using Big Data integration technologies from

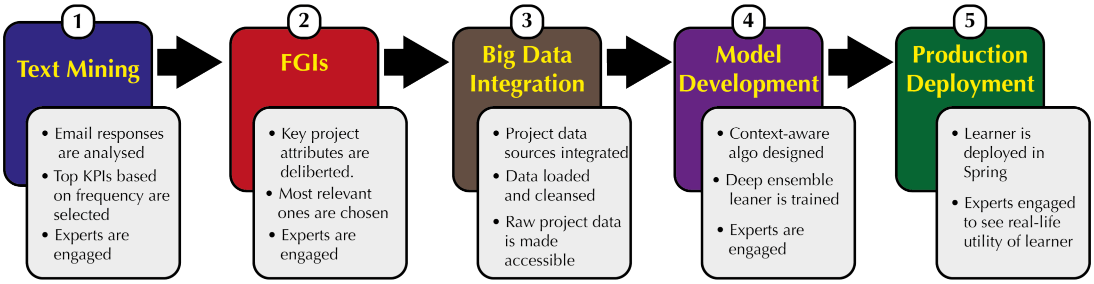

Figure 2: Proposed Research Methodology

structured and unstructured data sources. Mediator wrapper architecture is used for data integration from diverse data sources. The collected data is first filtered to select the right set of relevant power infrastructure projects. The integrated data is re-structured and stored in a relational model to support the subsequent ML tasks.

Once data is made accessible, a decomposition-integration based strategy is proposed to develop a deep ensemble learner for benchmarking KPI performance. In the decomposition stage, we proposed a floorplanning algorithm to map entire data onto the 2D plan and then extract attribute-specific KPI performance benchmarks. Recursive SQL queries were employed to derive these benchmarks where were then fed into the following integration stage. Integration stage harness the fully connected deep neural networks for training seventeen (17) models to produce project-sensitive profitability performance benchmark. Experts were engaged in the data annotation and algorithmic audit tasks. Several data augmentation tasks were performed. More importantly, embeddings were harnessed to learn high dimensional vectors, representing the inherent structure of categorical values. An extensive model training strategy is followed where different hyperparameters of deep learning algorithms were checked to identify the ones which will enable better learning of KPI benchmarks from the dataset. GPU-enabled servers from cloud were utilised for the training and evaluation of models. Oracle R Enterprise (ORE) is harnessed for production deployment where these tensor-encoded models were transported into a relational table in the Oracle database, and PLSQL package is developed to invoke these models from other applications. The outcome of the deep learning models was evaluated by industry experts across edge cases to ensure generalisability of the learner. Keras with Tensorflow backend is used for training models. These models were deployed for real-life scoring for use by end-users through the Spring system. The Spring system harnesses RAG colouring in its user interface to decode the performance of KPIs in a given project tender.

To see the real-life suitability of the proposed deep ensemble learner, we went further by testing the learner with a case study of 60km power infrastructure projects. The learner predicted a project-sensitive benchmark for all KPIs, which is deliberated by the industry experts in the Spring system. Their feedback was recorded and discussed in this study. Overall, the proposed deep ensemble learner has tremendous utility for estimators to objectively perform the tender evaluation as part of their line-of-work tasks.

4. Text Mining

Text Mining is a knowledge-intensive process to employ analytical tools for extracting meaningful insights from text documents. Unlike traditional data mining, Text Mining specialises in exploring semi-structured and unstructured text (Inzalkar & Sharma, 2015). Text Mining algorithms have a strong mathematical basis, thereby enabling quantitative analysis of the qualitative data. These approaches are widely used to solve non-trivial problems across several industries (Fleuren & Alkema, 2015; Weiss et al., 2015). Since the selection of KPIs is crucial for reliable tender evaluation, this study used Text Mining for identifying KPIs for unstructured email responses. The survey would be an excellent tool for this knowledge elicitation over emails. However, capturing all stakeholders' views using a survey would have involved asking numerous

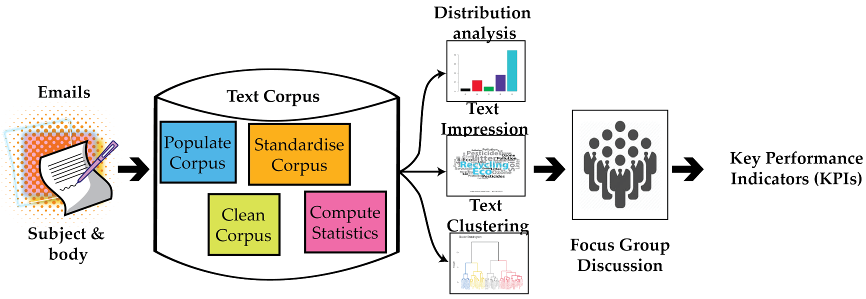

Figure 3: Proposed Text Mining Pipeline for KPIs Selection

questions. The respondents lose patience as they find many questions, especially with several unrelated to their experience. They start paying less heed to core relevant questions, which would have compromised the validity of the approach. Next, surveys are usually not useful for in-depth exploration when a study focuses on understanding the intersubjective perspectives of domain experts. Lastly, the survey is an apriori approach where researchers constrain respondents to a set of pre-selected views rather than sharing more diverse personal experiences. We, therefore, decided to capture expert knowledge from professionals through email responses to let them express their opinions on the significance of KPIs using open-free text.

To this end, data of email replies were gathered from industry professionals, asking them about what should comprise KPIs for project tender. Accordingly, 267 answers are recorded. The participants include employees of a leading UK construction firm. Out of all responses, 20% replies were received from the management, 30% from estimators and 50% from the delivery personnel including PMs, QS, CMs, etc. This distribution represents good diversity based on their roles and daily tasks. The following script shows a small section of an example response from those emails. There were $1,349.34$ words used on average in these emails, thereby constituting a considerable data for exploration. Manual text exploration was considered an error-prone and inefficient approach. A Text Mining approach is therefore adopted to review the textual contents from these emails. Fig. 3 shows an overview of the proposed Text Mining pipeline. We began by creating the text corpus, which is a repository geared for statistical analysis of text sources. Afterwards, various data cleansing and standardisation operations were performed. Lexical analysis is performed to break down sentences into words. Then, word-stemming is applied to find the root word for these terms.

> *"I would like to talk about the inclusion of* `retention` *to the KPIs we want to use for tender evaluation. Retention is a % of the sales amount held up by the client for a while. The % and period are agreed at the onset of the project. This amount can affect the cash flow of the project. If this amount is significant, then the reported margin on that project cannot include that amount which means the company cannot reflect that amount in their books yet. Now a construction company always have a few projects running, imagine the implication of all projects having a certain amount of retention being withheld by the different clients for years. This can adversely affect company books, i.e. several figures in red signifying figures waiting to be paid. This can be the difference in a company being in profit or being in the loss."*

An initial review revealed large disparities and grammatical errors in the emails. E.g. `Margin` was used differently by different people as `Profit Margin`, `Margin (%)`, `Margin` and `Markup`. Such heterogeneities usually stem from people paying less attention while writing emails (Sakurai & Suyama, 2005). A concept dictionary is developed to standardising vocabulary. Preferred terms used by experts are used to identify the KPIs. The concept dictionary was found phenomenal for tackling such heterogeneities in the text. Furthermore, some text cleanup operations on corpus are performed like excluding punctuations, stop words, white spaces, and numbers. This way, data is eventually made suitable for onward computation needed for

Figure 4: Wordcloud - Possible List of KPIs for Project Tenders

understanding the text. Term-document matrix (`TDM`) is constructed, whose rows represented terms and columns the emails, such that each entry in TDM represented the frequency of the given term in the given document. The `TDM` provided basis for the top-k analysis we employed to identify the KPIs for evaluation during the process.

Once data is prepared, several visualisations are produced, including wordcloud and barplot. Wordcloud is a great tool for having an initial impression of text contents (Cui et al., 2010). It reveals the most frequent terms used in the corpus. We used it to show all KPIs used in the emails. Additionally, barplot is used for filtering KPIs for *top-k* words used in the corpus. Fig. 4 displays a list of KPIs mentioned in emails using wordcloud. There were 233 KPIs in total, 87% of which comprises project costs and expenditure types. Majority of management personnel ((i.e. 93%) highlighted high-level KPIs like `materials`, `risks`, `opportunities`. The estimators mostly suggested a mix of high-level KPIs (57%) and detailed KPIs (43%). Project delivery personnel (i.e. 95%), on the other hand, mostly prescribed detailed KPIs for tender evaluation like `cables` and `conductors`.

Obviously, `Profit`, `Sales` and `Costs` are mostly recommended KPIs. The breakdown of KPIs like `Costs` including `Plant`, `Labour`, `Material`, `Subcontract` were secondly recommended. Overall, `Costs` are the frequently spoken KPIs in the corpus. Interestingly, a reasonable number of managerial staff talked about `Risks`, `Opportunities` and `Innovations`. These KPIs are rare though but reveal important information to aid tender evaluation. This way a large number of KPIs are identified from the text. Since a tender can't capture all KPs, a tradeoff is agreed concerning the complexity of documents and the ease of evaluation. It was highlighted that complex tenders are hard to analyse are consume more time. So finally top 20 KPIs were agreed based on their frequency distribution. Majority of these KPIs are high-level and capture many detailed-KPIs indirectly. Fig. 5 shows the distribution of top 20 KPIs using barplot.

It is pointed out that popular (top-k) terms might not be the most crucial KPIs for tender evaluation. We supported our selection criteria with opinions from the industry professionals to make sure a robust list of KPIs is compiled. To this end, the barplot is deliberated with industry experts to ensure the right KPIs

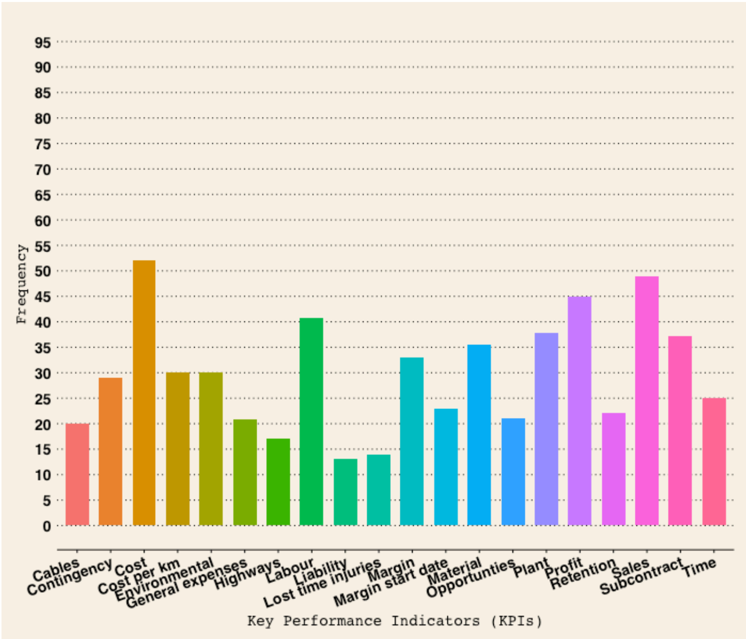

Figure 5: Barplot - Top 20 KPIs by Frequency Distribution

were chosen for tender evaluation. Their discussions have resulted in making useful adjustments to the list. KPIs like `Cables` were merged within `Materials` to avoid confounding impact. Similarly, `Time` is declared more of an attribute than a KPI, so it is dropped from KPIs and added to project attributes for enabling context in benchmarking. The `Lost Time Injuries (LTI)` refers to an injury that causes an operative absent from work for more than a day. As a KPI, `LTI` is removed from the list due to having cofounding influence on `risks`. Other KPIs like `Liability`, `Highways`, `Environmental` were merged into `risks` based on same argument. `Stretched margin`, which was not in the top-20 list, is advised for inclusion as it describes a maximum margin possible at project completion. Similarly, `margin start date` is included to see when exactly the project will start generating profit.

A focused group interview was conducted to find out the organisation of KPIs onto the tender summary. Initially, four clusters were proposed, including `Sales` (comprising NSV), `Costs` (including Labour, Plant, Material, Subcontract and General Expenses), `Profit` (including Margin) and `Miscellaneous` (comprises Retention, Cost per km, Risks, Contingency, Opportunity, Innovation, Stretch Margin and Margin Start Date). But `Miscellaneous` cluster was made redundant after the following changes. `Retention` is merged into `Sales` since it denotes a proportion of sales withheld by clients. `Cost per km` is clustered under `Costs` due to its reliance. Likewise, `risks` and `contingency` were moved to `Costs` as they represent amounts to cover `Costs` of events incurred by detrimental hazards or unforeseen circumstances. `Opportunity` and `Innovation` were added to `Profit` as they represent extra profits that can be realised; if not realised the profits doesn't suffer any more cost than it would normally. Finally, `Stretched Margin` and `Margin Start Date` were affixed with `Profit` due to their relevance. Table 3 displays the final organisation of 17 KPIs. This study utilises these KPIs for benchmarking the profitability performance of project tenders.

5. Focus Group Interviews (FGIs)

The second phase of research methodology employed several FGIs to understand the context for profitability benchmarking. The idea was to learn from the real-life experience of practitioners. FGIs were pivotal to understand key project attributes for contextual reasoning. FGIs were chosen over one-to-one

Table 3: Final List of KPIs for Tenders Evaluation

Sr.#	KPI Name	KPI Description
1	**NSV**	**Net sales value of the project**
2	Retention	% of sales withheld by the client for an agreed period
3	**Cost**	**Total cost of the project**
4	Cost per km	The amount per 1km of the route length
5	General expense	The amount allocated to others such as travel expenses, council fines, wastes etc.
6	Plant	The amount allocated to structures, machineries & specialist vehicles
7	Materials	The amount allocated to the parts needed to complete the project
8	Labour	The amount allocated to resource on the project
9	Subcontracts	The amount allocated to specialist work or work given to other contractors
10	Risk pot	The amount allocated to cover detrimental hazards to a project
11	Contingency	The amount allocated to unforeseen circumstances
12	**Profit**	**Profit to be made from the project**
13	Margin	Profit expressed as percentage
14	Opportunity pot	Margin to be realised from cutting cost
15	Innovation pot	Margin to be realised from using new techniques
16	Stretch margin	Margin + Opportunity pot + Innovation pot
17	Margin start date	The first day of realising the profit

Table 4: Details of the FGI participants

Sr No	Team	Expectations/themes	Paticipants	Experience (Years)	Firm Type	Background	Role
1	Management	KPIs, key project attributes pruning and organisation	5	18	Contractor	BSc Economics	Finance Director
2				13	Contractor	Accounting and finance	Business development director
3				22	Contractor	Accounting and finance	Cabling finance manager
4				14	Contractor	Site management	Resource manager
5				24	Contractor	Accounting and finance	Project Monitoring Officer (PMO)
6	Sales	KPIs, key project attributes pruning and organisation	5	25	Contractor	BSc Civil Eng	Quantity surveyor
7				12	Contractor	Construction quantity surveyor	Senior quantity surveyor
8				16	Contractor	Draughtsman	Project Design Manager (PDM)
9				18	Contractor	BSc Civil Eng	Senior Engineer
10				22	Contractor	BSc Civil Eng	Senior Engineer
11	Estimators	KPIs, key project attributes pruning and organisation, data labelling, crafting validation sets, and Spring evaluation	5	23	Contractor	OHL site operative	OHL estimator
12				19	Contractor	Finance and Accounting	Regional Estimating Manager
13				11	Contractor	BSc Civil Eng	Cabling estimator
14				25	Contractor	Business Analyst	Bid manager
15				22	Contractor	Business and Management	Bid Manager
16	Project Delivery	KPIs, key project attributes pruning and organisation	5	16	Contractor	OHL site operative	Technical director
17				20	Contractor	Business management	Project manager
18				14	Contractor	Construction site operative	Project manager
19				15	Contractor	Construction site operative	General foreman
20				20	Contractor	Project planning	Project planner

Table 5: Key Project Attributes for Contextual Reasoning

Sr.#	Attribute Name
1	Project size
2	Region
3	Project type
4	Business stream
5	Sector
6	Work type
7	Contractual type
8	Project duration

interviews as they allow participants to share their own experiences and respond to the views expressed by others. FGIs also facilitated group thinking with more deep-felt insights and a broader range of perspectives on the subject of contextual reasoning that can't be achieved with one-to-one interviews. The validity and applicability of the key project attributes were also authenticated before they were used to develop a reliable benchmarking algorithm. The perception and expectation of industry practitioners were also better understood. The FGIs were supervised proactively to maintain openness and ensure the contribution of all participants.

Overall, four FGIs were conducted with a total of 20 participants. Their selection was influenced by their role in the overall tendering process. Interactions were recorded and later compared with the notes taken, to ensure that all necessary information was captured. Transcripts were segmented for thematic analysis to compile a comprehensive list of key project attributes. A coding scheme was formulated. The critical project attributes found in the literature were also confirmed, with the addition of two crucial factors, i.e. project duration and contract type. In this study, a thematic analysis – that is, an exploratory qualitative data analysis approach – was employed (Guest et al., 2011). An exhaustive comparison to examine the structure and relationships among the themes was carried out. A thematic map was generated to provide an accurate representation of the transcripts. The final list of key project attributes taken from the thematic map is shown in Table 5.

6. Databases and Big Data Integration

The reliability of benchmarks depends on amounts of KPIs' data available for all project attributes. Data integration from diverse project sources was the key challenging task in this research. Data has resided in Google earth PDF route files, Oracle financials, telematics, Primavera, Candy, health & safety, think risk, business objects, project control database, customer relationship management, and other large bodies of unstructured documents. The specifications of these sources were explored to identify KPIs data from these data sources. Fig 6 describes data sources of linear projects investigated in this work. An overall data consisted of 5.7 million cells, summed to 1.2 terabytes in size are analysed. A mediator-wrapper strategy is used for interacting with data sources through a unified interface. Data is loaded into the Hadoop data warehouse (HIVE) for parallel preprocessing. Apache Spark is employed to perform computations. This data fulfils all 3V's of the Big Data, so this study qualifies this data as the Big Data.

Preprocessing of projects' data was another big leap in this research. Projects were filtered to ensure homogeneity in the analysis. Projects that tend to mislead results were excluded. Projects were filtered by their completion, which is determined from null values for pending costs or work in progress (WIP) or retention or capital employed or unpaid sales. This pruning reduced projects to $2,709$. Besides, projects only involving `Cabling`, `Overhead lines (OHL)` or `Substation` were selected. Among these, projects involving `fault & services`, `maintenance`, `supply only`, `bundled projects`, `overhead costs`, `internal projects` or `non-projects` (staff or resource training) works were also eliminated. International projects

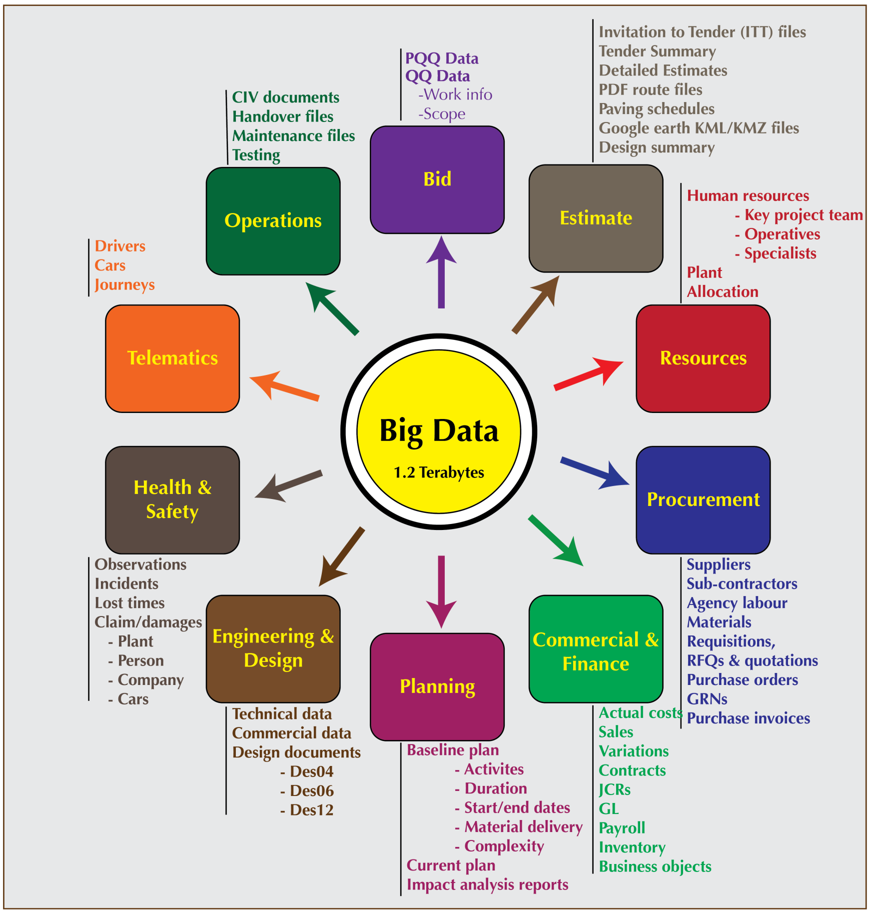

Figure 6: Big Data Sources of Linear Construction Projects

were also excluded. Apart from project pruning, lots of data disparities like non-standard values, missing values, and outliers were rectified. Projects, where data of KPIs were missing, were populated from other sources. Data ranges were validated, and erroneous entries were fixated. These manipulations excluded another 71, eventually left with 2, 443 projects in total. Project durations were populated from project plans. Project sizes were computed from financials. Margins at completion by fiscal periods were derived from job costing reports (JCRs). Data values indicating regions, contracts, voltages, workstreams, work types, and project types were standardised. A generic extensive `SQL` library is developed for `HIVE` to carry out similar data processing tasks. A sample of the raw data extracted from the integration of a large number of projects data sources is shown in Fig. 7. The choice of data elements included in the analysis is largely informed by the Text Mining and FGIs carried out during the research. Most monetary values were transformed into percentages and then the data is utilised by the ML models to derive KPI performance benchmarks.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Project Id	Project size	Project duration (mnths)	Sector	Region	Project type	Work type	Client	Business stream	Contract type	NSV	Retention	Costs
2	ANAG5820	Large	24	Road	Scotland	New build	Rural	Client 1	Cabling	NEC Option A	£16,661,120.47	£1,332,889.64	£15,433,307.49
3	ANAG5806	Medium	11	Transport	Scotland	New build	Urban	Client 2	Cabling	NEC Option B	£4,336,054.06	£346,884.32	£4,770,370.71
4	ANAG5750	Medium	9	Telecomm	Scotland	New build	Urban	Client 5	Cabling	NEC Option C	£1,580,583.46	£126,446.68	£1,655,699.33
5	ANAG5734	Large	20	Road	Scotland	New build	Urban	Client 7	Cabling	NEC Option C	£9,445,523.76	£755,641.90	£8,966,206.74
6	ANAG5640	Medium	10	Transport	North West	New build	Rural	Client 11	Cabling	NEC Option A	£3,577,600.78	£286,208.06	£2,521,634.88
7	ANAG5540	Medium	9	Transport	Scotland	New build	Rural	Client 7	Cabling	FIDIC contract	£2,473,575.15	£197,886.01	£2,454,544.43
8	ANAG5403	Medium	9	Telecomm	Scotland	New build	Rural	Client 3	Cabling	NEC Option B	£2,937,291.55	£234,983.32	£3,486,764.42
9	ANAG5375	Medium	14	Telecomm	Scotland	New build	Urban	Client 3	Cabling	NEC Option B	£2,434,240.93	£194,739.27	£2,629,887.33
10	ANAG5220	Medium	13	Road	Scotland	New build	Urban	Client 1	Cabling	NEC Option B	£3,381,942.98	£270,555.44	£2,485,595.17
11	ANCG5999	Medium	16	Transport	Midlands	New build	Rural	Client 15	Cabling	NEC Option A	£3,083,441.96	£246,675.36	£3,091,685.04
12	ANCG5998	Medium	14	Transport	Midlands	New build	Urban	Client 2	Cabling	NEC Option A	£3,236,077.64	£258,886.21	£3,462,916.69
13	ANCG5997	Large	21	Telecomm	Midlands	New build	Rural	Client 8	Cabling	NEC Option E	£10,103,841.03	£808,307.28	£8,590,029.32
14	ANCG5984	Medium	13	Telecomm	Midlands	New build	Urban	Client 13	Cabling	FIDIC contract	£2,652,490.02	£212,199.20	£2,714,529.98
15	ANCG5915	Medium	9	Road	Midlands	New build	Rural	Client 14	Cabling	NEC Option B	£2,990,262.82	£239,221.03	£3,667,931.64
16	ANCG5911	Medium	10	Transport	Midlands	New build	Urban	Client 10	Cabling	NEC Option B	£3,496,721.62	£279,737.73	£4,075,727.65
17	ANCG5904	Medium	8	Transport	Midlands	New build	Rural	Client 9	Cabling	NEC Option B	£2,517,885.83	£201,430.87	£2,394,494.17
18	ANCG5903	Medium	11	Telecomm	South	New build	Rural	Client 9	Cabling	Framework contrac	£3,750,172.54	£300,013.80	£2,718,940.94
19	ANCG5888	Medium	9	Telecomm	Midlands	New build	Rural	Client 1	Cabling	NEC Option A	£2,482,490.56	£198,599.24	£2,197,979.29
20	ANCG5887	Medium	12	Road	Midlands	New build	Urban	Client 2	Cabling	Framework contrac	£2,450,443.16	£196,035.45	£1,906,748.36
21	ANCG5881	Medium	18	Transport	Midlands	New build	Rural	Client 5	Cabling	NEC Option C	£6,122,042.56	£489,763.40	£4,981,624.59

	A	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	Project Id	Cost per km	General expense	Plant	Materials	Labour	Subcontracts	Risk pot	Contingency	Profit	Margin	Opportunity pot	Innovation pot	Total length in river (m)	Stretch margin	Margin start date
2	ANAG5820	£2,572.22	£349,362.36	£26,060.31	£14,950,063.60	£107,821.22	£0.00	£249,916.81	£249,916.81	£1,227,812.98	7%	£30,695.32	£4,297.35	0	£1,262,805.65	12/05/2015
3	ANAG5806	£973.55	£168,406.18	£84,417.22	£4,331,554.57	£177,520.43	£1,657.05	£130,081.62	£21,680.27	-£434,316.65	-10%	£10,857.92	£1,520.11	0	-£421,938.63	11/08/2012
4	ANAG5750	£649.29	£108,211.54	£135,221.42	£1,080,708.89	£310,591.58	£20,965.90	£47,417.50	£0.00	-£75,115.87	-5%	£1,877.90	£0.00	0	-£73,237.97	23/04/2014
5	ANAG5734	£2,241.55	£239,121.46	£183,131.58	£7,937,960.46	£356,319.73	£249,673.51	£141,682.86	£94,455.24	£479,317.02	5%	£11,982.93	£1,677.61	0	£492,977.56	29/01/2017
6	ANAG5640	£840.54	£71,329.36	£224,038.26	£1,792,523.44	£363,657.62	£69,187.54	£107,328.02	£0.00	£1,055,965.90	30%	£26,399.15	£3,695.88	0	£1,086,060.93	20/02/2016
7	ANAG5540	£2,454.54	£140,223.43	£208,495.58	£1,182,845.22	£365,136.85	£537,865.41	£74,207.25	£0.00	£19,030.72	1%	£475.77	£0.00	200	£19,506.49	15/05/2016
8	ANAG5403	£1,660.36	£187,702.60	£246,386.71	£1,751,874.20	£422,942.16	£868,542.75	£44,059.37	£0.00	-£549,472.87	-19%	£13,736.82	£1,923.16	290	-£533,812.89	17/07/2013
9	ANAG5375	£1,143.43	£155,206.51	£317,080.70	£1,321,443.37	£417,717.23	£418,439.52	£73,027.23	£0.00	-£195,646.40	-8%	£4,891.16	£684.76	0	-£190,070.48	19/03/2014
10	ANAG5220	£872.14	£77,332.45	£331,051.43	£605,934.24	£427,945.40	£1,043,204.61	£101,458.29	£0.00	£896,347.81	27%	£22,408.70	£3,137.22	0	£921,893.72	13/04/2014
11	ANCG5999	£1,437.99	£172,859.32	£401,774.27	£413,728.55	£574,914.98	£1,528,407.92	£92,503.26	£0.00	-£8,243.08	0%	£206.08	£0.00	320	-£8,037.00	15/09/2016
12	ANCG5998	£577.15	£186,202.66	£358,816.94	£1,456,893.09	£519,847.55	£928,100.86	£97,082.33	£0.00	-£226,839.05	-7%	£5,670.98	£0.00	0	-£221,168.07	18/07/2015
13	ANCG5997	£2,489.86	£271,889.60	£390,933.01	£332,487.08	£548,620.31	£1,445,906.32	£151,557.62	£191,972.98	£1,513,811.71	15%	£37,845.29	£5,298.34	800	£1,556,955.34	26/10/2014
14	ANCG5984	£2,088.10	£187,024.80	£504,428.32	£981,376.10	£567,354.39	£474,332.37	£79,574.70	£0.00	-£62,039.96	-2%	£1,551.00	£0.00	0	-£60,488.96	05/11/2015
15	ANCG5915	£1,833.97	£216,810.40	£405,216.76	£869,726.46	£532,426.36	£1,643,751.66	£44,853.94	£0.00	-£677,668.82	-23%	£16,941.72	£2,371.84	0	-£658,355.26	09/06/2016
16	ANCG5911	£1,509.53	£159,307.76	£462,678.03	£1,958,940.41	£629,322.17	£864,802.85	£104,901.65	£0.00	-£579,006.03	-17%	£14,475.15	£2,026.52	100	-£562,504.36	07/07/2013
17	ANCG5904	£798.16	£86,254.82	£331,303.10	£526,962.51	£480,999.68	£954,617.37	£75,536.57	£0.00	£123,391.66	5%	£3,084.79	£431.87	0	£126,908.32	10/08/2016
18	ANCG5903	£647.37	£158,892.82	£377,162.13	£397,101.33	£487,163.34	£1,298,621.32	£112,505.18	£26,251.21	£1,031,231.60	27%	£25,780.79	£3,609.31	0	£1,060,621.70	11/09/2017
19	ANCG5888	£686.87	£149,163.05	£326,161.82	£291,965.23	£463,718.86	£466,970.33	£74,474.72	£0.00	£284,511.27	11%	£7,112.78	£995.79	0	£292,619.84	11/12/2015
20	ANCG5887	£719.53	£138,214.89	£361,193.70	£452,604.38	£506,330.28	£448,400.17	£73,513.29	£0.00	£543,694.80	22%	£13,592.37	£1,902.93	350	£559,190.10	15/11/2016
21	ANCG5881	£1,018.74	£159,846.75	£450,362.53	£593,434.25	£497,989.49	£1,279,991.57	£183,661.28	£104,074.72	£1,140,417.97	19%	£28,510.45	£3,991.46	0	£1,172,919.88	19/10/2013

Figure 7: Sample Data of Power Infrastructure Projects Integrated from Diverse Big Data Sources

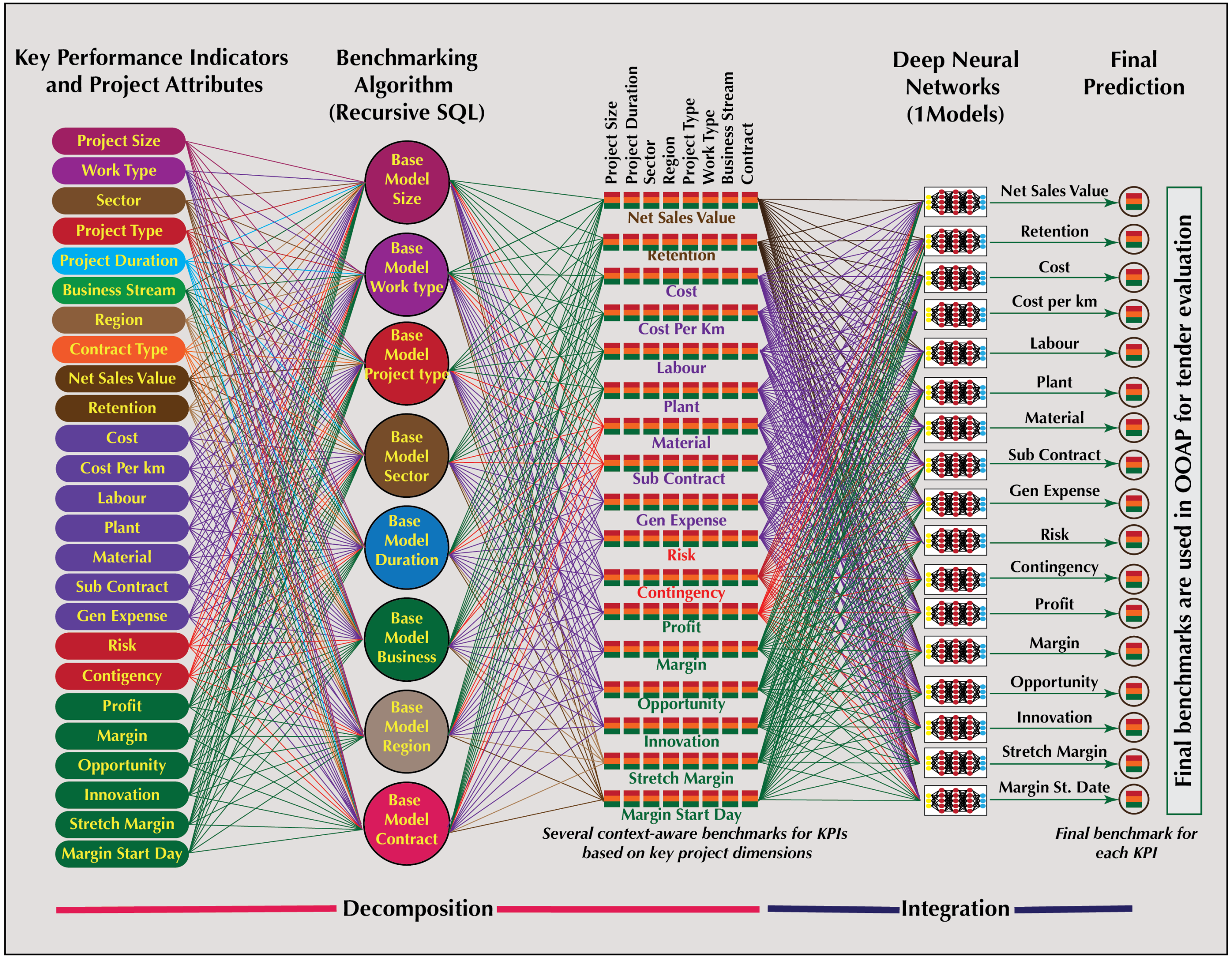

Figure 8: Proposed Deep Ensemble Learner

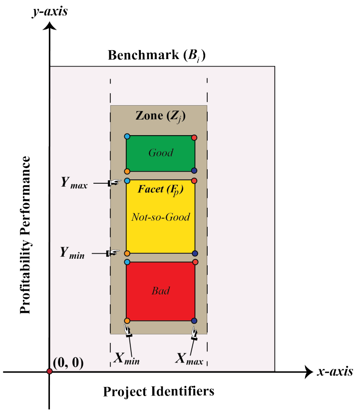

Figure 9: Proposed Data Representation using Floorplanning

7. Decomposition-Integration based Deep Ensemble Learner

The next step in the proposed methodology involved the development of the deep ensemble learner. Ensembling is popular ML strategy for developing more reliable learners from several base learners (Friedman et al., 2001). This idea is reported to have significantly improved the predictive accuracy for many ML tasks (Wang et al., 2012, 2011, 2009). Ensemble learners outperform traditional leaners where predictions from base learners diverge substantially, which is the case with estimating benchmarks for KPIs. Ensembling works best with tree-based or neural network-based algorithms (Pino-Mejías et al., 2008). This study employs two-staged decomposition and integration approach to devising the proposed learner. The learner first yields several attribute-specific benchmarks and then shrinks it into a final project-sensitive benchmark to support the tender assessment process.

Fig. 8 illustrates the architecture of our proposed learner. The learner takes in the percentage-transformed and normalised data of power infrastructure projects (see Fig. 7) during the decomposition stage and yields attribute-specific benchmarks for all the KPIs. A benchmark defines the good, average and bad performance criteria for KPIs. It comprises numerical thresholds (i.e. `thresh1` and `thresh2`) against which the performance of KPIs can be measured objectively. The decomposition stage harnesses recursive `SQL` queries to the finest, which are formulated based on our proposed context-aware benchmarking algorithm, described in the following subsection. The decomposition algorithm generates several benchmarks for each KPIs based on key project attributes. The output from this stage is then stuck through the fully-connected deep neural network, comprised of several layers, during the integration stage. The integration stage collates these attribute-specific benchmarks and generates one project-sensitive benchmark for each KPI. These benchmarks are then utilised by the Spring system for supporting the tender evaluation process. The user interface (UI) of Spring uses red/amber/green (RAG) colouring approach to visualise the performance of KPIs against a benchmark during tender evaluation. This intelligent colour coding of KPIs informed by deep ensemble learner helps the estimators to identify strengths and weaknesses of tenders quickly. The following subsections explain these stages in more detail.

7.1. Proposed Decomposition Algorithm

The proposed learner takes the project context into account, which is defined based on eight key project attributets, agreed during FGI with industry professionals. Context has enormous significance for a robust evaluation system. E.g. tenders of mega projects tend to have low profit margins. It would be useless to trigger a red alarm for a mega project using a uniform averaged-benchmark for margin, which is likely to have higher thresholds. Our analysis of most projects revealed that context inclusion is vital for deriving reliable benchmarks for all selected KPIs. We proposed the decomposition stage in the learner to enable context-aware reasoning in our proposed methodology. The algorithm models entire data as a `2D` surface, and

Algorithm 1: Context-aware benchmarking algorithm

Data: $\mathcal{P}$ data of past projects
Result: $\mathcal{B}$ benchmarks

procedure ***computeContextualBenchmarks** ($\mathcal{P}$ **projects' data**)*
  $\mathcal{K} \leftarrow [k_1, k_2, k_3, \cdots, k_n]$ such that $k \in K$;
  $\mathcal{C} \leftarrow [c_1, c_2, c_3, \cdots, c_m]$ such that $c \in C$;
  $\mathcal{B} \leftarrow [\mathcal{B}_{kc1}, \mathcal{B}_{kc2}, \mathcal{B}_{kc3}, \cdots, \mathcal{B}_{kcj}]$ where j is $\mathcal{K} \times \mathcal{C}$;
  $\mathcal{L} \leftarrow [good, average, bad]$;
  for k ***in*** $\mathcal{K}$ **do**
    for c ***in*** $\mathcal{C}$ **do**
      sort & index $\mathcal{P}$ by c;
      cluster $\mathcal{P}$ for distinct values of c;
      for l ***in*** $\mathcal{L}$ **do**
        compute x and y boundaries for k, c, and l;
        impute boundaries if x and y are still empty ($\emptyset$);
        generate facet f for benchmark $\mathcal{B}_i$;
      end
      $\mathcal{F} \leftarrow \mathcal{F} \cup f$;
      $\mathcal{Z} \leftarrow resolveClash(\mathcal{F}, c)$;
    end
    $\mathcal{B}_i \leftarrow resolveClash(\mathcal{Z}, k)$;
    $\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{B}_i$;
  end
  return $\mathcal{B}$
end procedure

attempts to find boundaries (i.e. x and y coordinates) for benchmarks to separate KPI performances. Three fuzzy labels, including `good`, `average`, and `bad` were chosen to denote performance. The algorithm traverses the data like decision trees with slight alterations. Traditional decision tree algorithm underperform during extrapolation due to missing values. They are likely to yield undesirable blank benchmarks for certain edge cases including the missing values. On the contrary, the proposed algorithm extends search space and returns expected value (`EV`) for all cases, which is likely to be the best predicted benchmark.

7.1.1. Benchmarking as the Floorplanning Problem

The algorithm uses the floorplanning technique to learn attribute-specific benchmarks. The floorplanning is a field in convex optimisation where spatial units are used to model a computational problem. The optimisation algorithm aims to find the optimal size and placement of spatial units within an outer parcel of a fixed perimeter. The proposed algorithm uses similar logic. It models the entire benchmarking space $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3, \cdots, \mathcal{B}_n\}$ comprises n benchmarks for k KPIs, where n is the number of project attributes, $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \cdots, \mathcal{C}_n\}$, for contextual reasoning. A benchmark $\mathcal{B}_i$ is made up of zones, $\mathcal{Z} = \{\mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3, \cdots, \mathcal{Z}_j\}$, where j is the number of unique values in the attribute. E.g., if $\mathcal{C}_i$ is *Business stream* then $\mathcal{Z}_1$ is *Cabling*, $\mathcal{Z}_2$ is *Transmission*, and $\mathcal{Z}_3$ is *Substation*. The zones ($\mathcal{Z}$) horizontally divides 2D plan of a KPI into j slices. These zones are further split into facets $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \cdots, \mathcal{F}_p\}$ where p denotes three performance labels. The proposed algorithm computes x and y boundaries for facets ($\mathcal{F}_p$). The benchmark ($\mathcal{B}_i$) surrounds j zones ($\mathcal{Z}_j$) which in turn includes p facets ($\mathcal{F}_p$). The algorithm is based on few assumptions. A facet $\mathcal{F}_p$ cannot span several zones in $\mathcal{Z}$. Likewise, a zone $\mathcal{Z}_j$ cannot span several benchmarks. The $\mathcal{B}_i$ is the external parcel of 2D plan where x and y axes are derived. The lower-left corner of $\mathcal{B}_i$ is fixated at $(0, 0)$. Unlike the actual floorplanning, the proposed algorithm doesn't optimise the placement of rectangles; rather, it uses spatial querying to compute benchmark boundaries. Fig 9 elaborates the proposed representation used in the algorithm.

7.1.2. The Proposed Algorithm

Listing 1 outlines our proposed algorithm for decomposition. Let $\mathcal{K}$ denotes KPIs and $\mathcal{C}$ the key project attributes. The algorithm returns $\mathcal{C}$ benchmarks for $\mathcal{K}$ KPIs. All computation is enclosed in three loops. The first loop iterates over $\mathcal{K}$ KPIs and returns $\mathcal{C}$ benchmarks for each KPI. The second loop iterates over $\mathcal{C}$ values of a given project attributes and maps data as `2D` plan. To this end, it `sorts`, `index` and `clusters` data based on previous project performance. Then, project identifiers are mapped onto `x-axis` and scaled KPI range to `y-axis`. Next, data are grouped by $\mathcal{C}$ to support benchmark computation. The innermost loop extracts performance boundaries, as thresholds for `good`, `average` and `bad` facets, accordingly. These boundaries might be null or overlap with one another. This occurs if data is imbalance or projects data has quality issues. The algorithm employs data imputation for blank benchmarks by extending search space from facet ($\mathcal{F}$) data to zone ($\mathcal{Z}$) data or in the worst case to benchmark $\mathcal{B}_i$ data. This ensures appropriate threshold values are always returned. However, the accuracy drops whenever the algorithm widens the search space. The algorithm also applies a simulated-annealing based optimisation to yield non-overlapping crisp boundaries for benchmarks. Since overlaps usually occur in facets and zones, the algorithm uses a `resolveClash` function to stop overlaps. The final benchmarks are guaranteed to be concrete by these augmentations.

We used recursive `SQL` queries to implement all steps in our proposed algorithm. The algorithm draws eight benchmarks for each KPI by harnessing Big Data of power infrastructure projects. Fig. 10 visualises some benchmarks for the `profit margin` KPI. These visualisations clearly reveal that profitability performance varies significantly by project attributes. The use of averaged margin (i.e. 21.48% as the case in this study) for deriving benchmark thresholds to plan and control project performance is misleading. This context-aware reasoning is the novel aspect of the decomposition stage in our proposed learner. However, the inclusion of context results in several benchmarks for each KPI. Several benchmarks would confuse estimators during the tender evaluation process. A project might be considered `good` by one benchmark (`say region`) but `bad` by another (`say contract`). A clever scheme needs to be incorporated in the leaner for combining several attribute-specific benchmarks to generate one project-sensitive and context-aware benchmark. To this end, we are proposing the integration stage, where fully-connected deep neural networks-based models are exploited for consolidating KPI benchmarks. The following subsection explains the proposed integration stage in detail.

7.2. Proposed Integration Approach

The integration stage involved the training of seventeen (17) deep learning models. These models takes as input the key project attribute along with the output from the decomposition stage which is an attribute-specific benchmark containing `performance label` (categorical), `thresh1` (numerical) and `thresh2` (numerical). These models outputs a unified benchmark for respective KPIs, in the form of `performance label` (category), `thresh1` (numerical) and `thresh2` (numerical) values. During evaluation, a KPI will be classified as poorly designed if its performance is below the `thresh1`, it will be considered average performing if the performance lies between `thresh1` and `thresh2`, and good if it's performance exceeds `thresh2`. To prepare the data for training these models, we engaged industry experts to help us annotate all projects given eight attribute-specific benchmarks. They deliberated attribute-specific benchmarks and then agreed on the most appropriate thresholds for all KPIs to be used to evaluate one project. In this way, the entire dataset of power infrastructure projects is labelled with context-aware thresholds.

Since key project attributes were categorical like the business stream, holding Cabling, Substation and Overhead lines (OHL) categories, it was needed to convert these attributes into some numerical form. Neural networks under the hood perform lots of mathematical manipulations during the training process, which would not be possible with categorical literals in the data. Merely integer enconding these attributes would be an option. However, it would be illogical as encoded categories won't have revealed any important information or insights to the model. `Cabling` encoded as 150 and `OHL` as 50 in `business stream` attribute doesn't means that `cabling` is 3 times larger or significant than `OHL`. Another alternative was to employ `one-hot encoding` that works well for attributes with fewer values, but the attributes

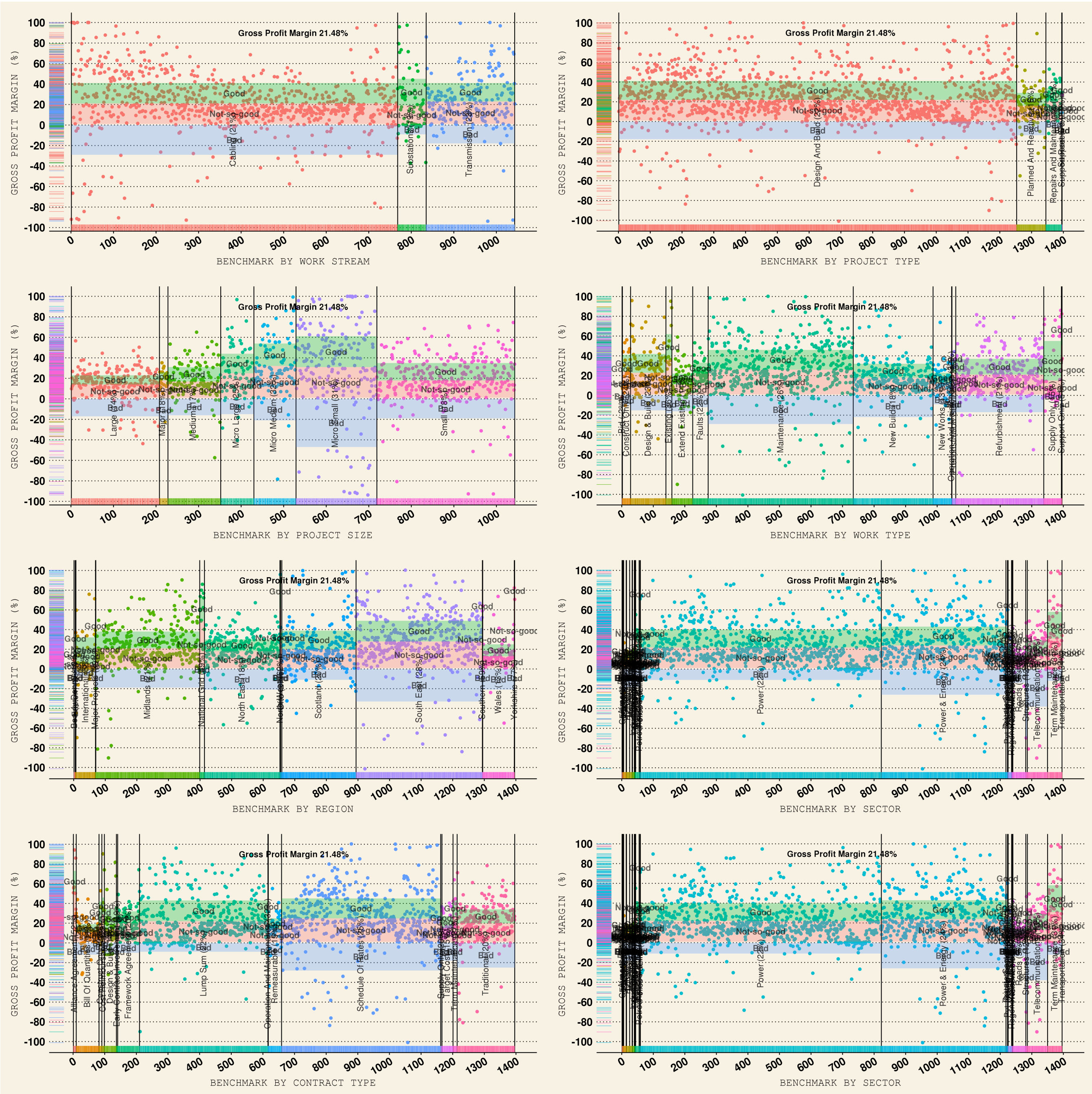

Figure 10: Attribute Specific Benchmarks for Profit Margin KPI

like Client can take up to 1000 values. This approach is lazy and would have resulted in sparse matrices; hence, unnecessary memory-intensive computations. The most appealing option that was finally taken up in this research was that of embeddings. A primary reason behind this modelling decisions was to capture intrinsic properties of categorical attributes to aid to models' performance. In embedding, categorical values of an attribute are mapped onto a dense vector consisted of real numbers (its embedding) to learn the semantics of each category. The embeddings of size 8 for business stream would look like $[1.624, -0.612, -0.528, -1.073, 0.865, -2.302, 1.745, -0.761]$. We randomly initialised embeddings for key project attributes, and then learned their dense representations via the training process, before training our deep neural networks.

Data normalisation is also exercised for the thresh1 and thresh2 input features to ensure zero mean and unit standard deviation. The data is then split into training (70%), validation (20%) and test (10%). We engaged industry professionals for crafting good validation and test sets that can enable the algorithm to train models which have reasonably better generalisation capabilities. One guiding principle was to train models on projects of earlier dates and test on the most recent projects. This was to mimic the real-life complexities and check models' production deployment capacity, which was one of the main objectives of this study. The prediction problem is modelled as the regression problem where these models will predict thresh1 and thresh2 for the good, average and bad KPI performance. Mean squared error (MSE) and R^2 are employed as error and accuracy matrices to ensure that these models are advancing in the right directions during the training process.

We followed a systematic approach to train these seventeen (17) deep neural networks. Grid search is applied to check for the most optimal values of hyperparameters during the training phase. Lots of models were trained using different combinations of hyperparameters across initialisation type, scaling, activation functions, epochs, number of layers and number of nodes. Learning rate finder based on differential learning rate annealing is employed for finding the right step size during the training process. Stochastic gradient descent with restarts (SGDR) is implemented for training the model. Kaiming and Glorot initialisations are found to work well over random or uniform approaches for initialising the model parameters. These tasks entailed enormous processing which is carried out on cloud-based servers with massive NVIDIA GPU compute capabilities. The learners' error, accuracy, response time, along with hyperparameter details, were recorded. The configuration of deep neural networks with the best accuracy is selected to separately train, intensively investigate and eventually deploy these models in the Spring system. Table 6 shows seventeen (17) deep learning models trained during the integration stage along with architectural details. Lots of adjustments were performed to develop models with the highest predictive accuracy and generalisation capability. Keras library is used for programming deep learning models. The library provided high-level methods to develop different architectures of deep learning models. Keras can be configured with several numerical optimisation engines like Theano, Tensorflow, and Microsoft Cognitive Toolkit (CNTK). This study utilised Tensorflow as the backend computation engine. R interface for Keras is configured using RStudio over a cloud server.

Oracle R Enterprise (ORE) with custom PLSQL library is used for production deployment. The Spring system invokes deep ensemble leaner through PLSQL interface in an integrated fashion. The learner takes key project attributes for a given opportunity and then returns a project-sensitive benchmark of all KPIs to support the evaluation process. Spring exploits Java EE Expression Language (EL) constructs to enforce these predicted benchmarks using RAG colour coding in the user interface to support user task and speed up tool adoption and use.

8. Production Deployment

The proposed deep ensemble learner is deployed in the Spring system that is designed to facilitate staff in performing the whole-life tasks of construction projects using Machine Learning (ML). The entire functionality of Spring can be divided into two main construction stages, i.e. sales and delivery. Spring facilitate users during these stages through two dashboards. The first is Opportunity-On-A-Page (OOAP)

Table 6: Details of deep neural network architectures used in integration layer

Sr.#	**KPIs**	**ML Approach**	**Scaling type**	**Solver**	**Learning Rate**	**Weight initialisation**	**Hidden layers**	**Number of nodes**	**Activation function**	**Epochs**	**Dropout**	**Training Error**	**Test Error**	**Training Accuracy**	***Test Accuracy***	**Solution Status**
1	NSV	DNN	normalization	Adam	0.5	kaiming uniform	5	49, 49, 49, 49, 49, 1	relu, relu, relu, relu, relu, linear	250	0.5	12.56	39.26	87.44	*60.74*	Maximum iterations reached
2	Retention	DNN	standardization	Adam	0.3	kaiming uniform	5	49, 49, 49, 49, 49, 1	relu, relu, relu, relu, relu, linear	250	0.3	0.59	6.53	99.41	*93.47*	Optimal (objMinProgress)
3	Cost	DNN	rescaling	Adam	0.3	kaiming uniform	3	49, 49, 49, 1	relu, relu, relu, linear	250	0.3	1.18	5.40	98.82	*94.60*	Optimal
4	Cost per km	DNN	standardization	Adam	0.5	kaiming uniform	3	49, 49, 49, 1	relu, relu, relu, linear	250	0.5	2.03	4.67	97.97	*95.33*	Optimal (objMinProgress)
5	Gen Expense	DNN	standardization	Adam	0.5	kaiming uniform	5	49, 49, 49, 49, 49, 1	relu, relu, relu, relu, relu, linear	250	4.23	4.23	15.65	95.77	*84.35*	Optimal
6	Plant	DNN	standardization	Adam	0.5	kaiming uniform	3	49, 49, 49, 1	relu, relu, relu, linear	250	0.3	6.67	12.55	93.33	*87.45*	Optimal
7	Material	DNN	standardization	Adam	0.2	kaiming uniform	3	49, 49, 49, 1	relu, relu, relu, linear	250	0.2	3.57	9.78	96.43	*90.22*	Optimal
8	Labour	DNN	standardization	Adam	0.2	kaiming uniform	3	49, 49, 49, 1	relu, relu, relu, linear	250	0.3	4.88	10.09	95.12	*89.91*	Optimal
9	Sub-contract	DNN	standardization	Adam	0.3	kaiming uniform	3	49, 49, 49, 1	relu, relu, relu, linear	250	0.3	10.22	14.68	89.78	*85.32*	Optimal
10	Risk	DNN	normalization	Adam	0.2	kaiming uniform	5	49, 49, 49, 49, 49, 1	relu, relu, relu, relu, relu, linear	250	0.2	10.99	18.69	89.01	*81.31*	Maximum iterations reached
11	Contingency	DNN	normalization	Adam	0.2	kaiming uniform	5	49, 49, 49, 49, 49, 1	relu, relu, relu, relu, relu, linear	250	0.4	2.97	16.03	97.03	*83.97*	Maximum iterations reached
12	Profit	DNN	rescaling	Adam	0.4	kaiming uniform	3	49, 49, 49, 1	relu, relu, relu, linear	250	0.3	4.61	6.59	95.39	*93.41*	Optimal (objMinProgress)
13	Margin	DNN	rescaling	Adam	0.4	kaiming uniform	3	49, 49, 49, 1	relu, relu, relu, linear	250	0.4	22.63	77.37	99.01	*93.41*	Optimal (objMinProgress)
14	Opportunity	DNN	standardization	Adam	0.2	glorot_uniform	4	49, 49, 49, 49, 1	relu, relu, relu, relu, linear	250	0.2	2.58	15.27	97.42	*84.73*	Maximum iterations reached
15	Innovation	DNN	normalization	Adam	0.2	glorot_uniform	4	49, 49, 49, 49, 1	relu, relu, relu, relu, linear	250	0.2	6.30	19.91	93.70	*80.09*	Maximum iterations reached
16	Stretch margin	DNN	normalization	Adam	0.5	kaiming uniform	3	49, 49, 49, 1	relu, relu, relu, linear	250	0.3	1.91	12.90	98.09	*87.10*	Optimal (objMinProgress)
17	Margin start date	DNN	standardization	Adam	0.5	kaiming uniform	3	49, 49, 49, 1	relu, relu, relu, linear	250	0.4	0.64	6.96	99.36	*93.04*	Optimal (objMinProgress)

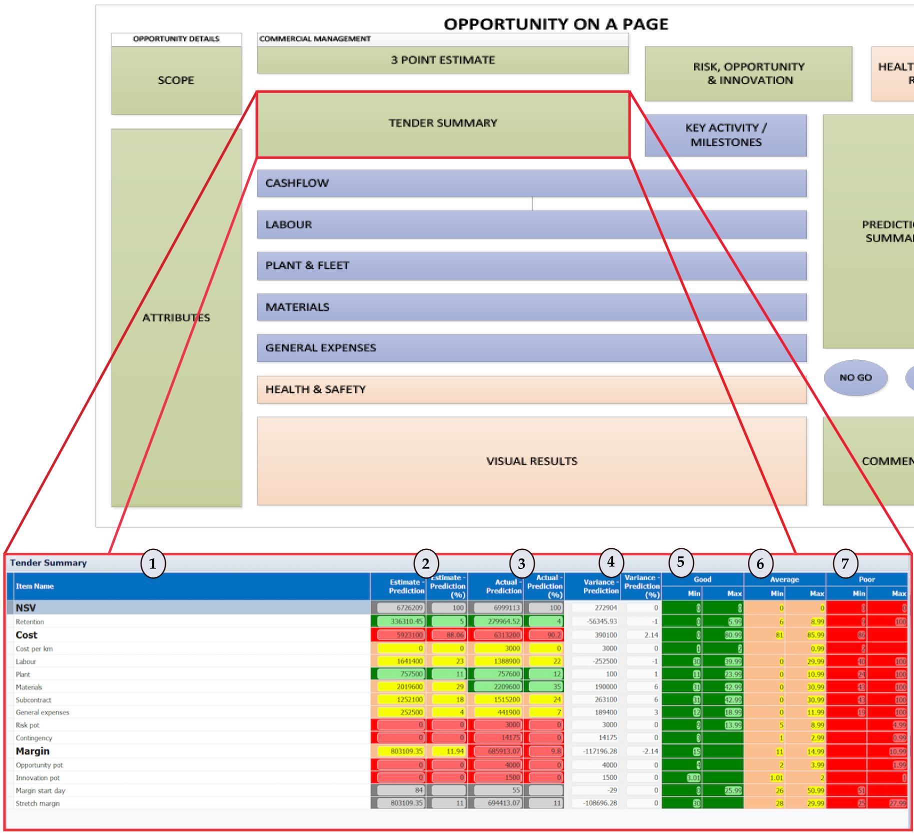

Figure 11: (a) The OOAP Layout (b) Deep Ensemble Learner in Action (Spring)

that covers all areas to support the sales activity. The second dashboard is `Project-On-A-Page (POAP)` that facilitates users during key delivery tasks. `OOAP` is the sales-phase equivalent of `POAP` used in the delivery phase. This study aims to gear `Spring` towards tender evaluation. `Spring` employs many ML models which are supervised by human experts to prepare reliable tender documents. The discussion of all ML models is beyond the scope of this study. Once tender documents are ready and available for evaluation, `Spring` uses the deep ensemble learner trough `OOAP` dashboard to generate project-sensitive benchmarks for KPI valuation. Fig. 11 displays `OOAP` layout and deep ensemble learner in action in the `Tender Summary` accordion of the tool. The deep ensemble learner has generated a benchmark to evaluate the given opportunity. Main attributes of a `tender summary` include the following:

1. **Tender items:** The first column displays the name of the KPI used to describe a project tender.
2. **Estimated predictions:** Next two columns display predictions (by value and proportion) using the proposed deep ensemble learner, trained on estimates generated by estimators. Ideally, these values shall be closer to human-generated estimates as the learner learned relationship from their experience.
3. **Actual predictions:** Next two columns display predictions (by value & by proportion) using our proposed deep ensemble learner trained over the actual finances of completed projects. These values shall be closer to facts when the project would be delivered.
4. **Variance prediction:** Next two columns report the difference between forecasts from ML models trained on actual finances and the estimates.
5. **Poor performance:** These two columns provide benchmarks for poorly designed project tenders.
6. **Average performance:** These two columns provide benchmarks for average project tenders.
7. **Good performance:** Last two columns provide benchmarks for good project tenders.

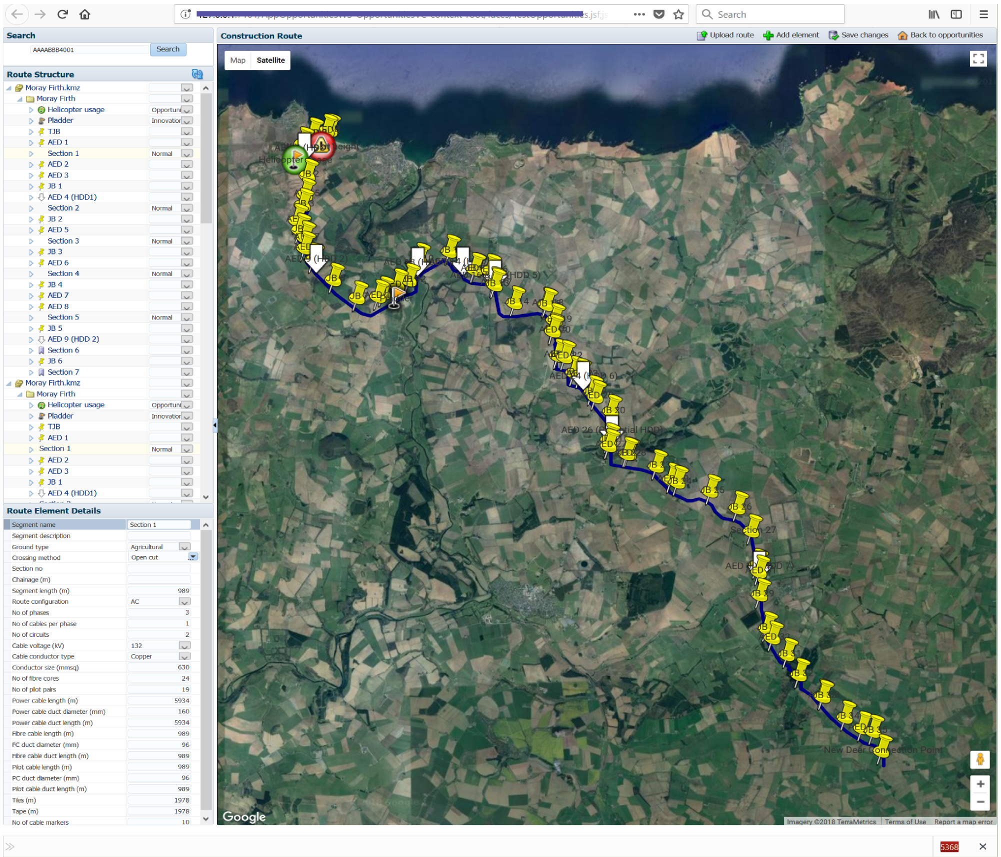

Figure 12: Case Study of Power Infrastructure Project in Spring Tool used for Model Evaluation

The proposed deep ensemble learner populates columns labelled 5, 6 and 7 in Fig. 11. `Spring` tool uses these benchmarks to inform `RAG` colour encoding shown in the `OOAP` interface. It can be seen that Spring has highlighted KPIs using red, green and amber colours. The `Tender Summary` accordion shall be green for most KPIs for a tender to be good. Otherwise, estimators can start sorting out the red fields and improve estimates for KPIs where the proportions are off. In this way, the estimators glance through the tender and quickly identify week (red) or strong (green) KPIs for the opportunity at hand.

9. System Evaluation through Case Study

While we followed a robust strategy to validate the performance of our deep ensemble learner by crafting robust validation and test sets. This section explains how we further advanced our research toward real-life deployment. We assessed a real tender of 60km cabling project. Fig. 12 shows the `Spring` design editor where the entire construction route for cabling project is displayed. `Spring` utilises advanced geospatial analysis and mining to compute many critical route statistics with a high degree of accuracy. Table 7 presents the tender summary alongside the project-specific benchmark generated by the proposed deep ensemble learner for tender evaluation. The same benchmark is also shown in Fig.11 (b)) using colour coded KPIs. This benchmark is critically analysed by senior estimators to understand the reliability and suitability of the proposed learner. Their discussions for each KPI benchmark are captured in the following paragraphs.

We found that it is challenging to define an accurate benchmark for the Net Sales Value (NSV) as project performance can be good, bad or average regardless of project sizes. Table 6 corroborated this fact that the

Table 7: Profitability performance benchmark predicted by deep ensemble learner for case study project

Sr. #	**Key Performance Indicator (KPI)**	**Poor Performance**		**Average Performance**		**Good Performance**	
		Min Threshold	**Max Threshold**	**Min Threshold**	**Max Threshold**	**Min Threshold**	**Max Threshold**
1	NSV	0	0	0	0	0	0
2	Retention	9	100	6	8.99	0	5.99
3	Cost	86	-	81	85.99	0	80.99
4	Cost per km	2	-		0.99	1	2
5	Labour	40	100	0	29.99	30	39.99
6	Plant	24	100	0	10.99	11	23.99
7	Materials	43	100	0	30.99	31	42.99
8	Subcontract	43	100	0	30.99	31	42.99
9	General expenses	19	100	0	11.99	12	18.99
10	Risk pot	-	4.99	5	8.99	9	13.99
11	Contingency	-	0.99	1	2.99	3	-
12	Margin	-	10.99	11	14.99	15	-
13	Opportunity pot	-	1.99	2	3.99	4	-
14	Innovation pot	-	1	1.01	2	3.01	-
15	Margin start day	51	-	26	50.99	0	25.99
16	Stretch margin	25	27.99	28	29.99	30	-

NSV model has the least predictive accuracy, despite extensive data augmentation and model tuning efforts. This led us to override Spring predictions by displaying `null` values. For `retention`, deep ensemble learner foretold good projects shall have this value set between 0% and 5% of the NSV. For average performance, retention shall fall in a range of 5.1% to 8% of NSV, whereas retention rate above 8% will result in poor project performance. While contractors wish retention as low as possible, these predictions seemed quite high at first. However, after a detailed data exploration, it is revealed that the client involved in this project has always imposed high retention rates in the past projects conducted by this contractor.

The benchmark revealed interesting insights into different cost categories. For the given tender to ensure good performance, cost categories such as labour, plant, material, subcontractor and general expenses have to be within the median percentages of the total cost of the project. These categories are not necessarily been at the extreme ends. The contractors shall price cost categories just right such that these tendering items are neither set too low nor too high. In either case, the project is likely to end up having a poor profitability performance. Such insights are crucial to facilitate contractors during tender negotiations and to shift the power of negotiation in the contractors' favour. They will better understand their position and allowances to push cost boundaries in either direction. While negotiating with an aggressive client, contractors can begin with a cost at the top of the good range and start cutting it down slowly until they reach the bottom of that range. Due to the boundaries being broad, the contractor can cut a lot off before they end up in the amber range, which again is quite a wide range. Likewise, the contractor can adjust their costs to get the best tender possible. A project with original KPIs at amber could be tuned such that some KPIs are in the green and some within amber before submitting it to the client.

An essential insight reported by the benchmark includes the limits for risks and contingencies associated with this project. The `risk` pot is another crucial KPI on the tender summary. It should be carefully designed. In case contractor overload risks, they lose points during tender negotiations. And if they underestimate risks, they can encounter severe problems during project planning and delivery. The benchmark revealed that risk pot should be up to the 5% of the total project cost for best profitability performance. The contingency is usually used in tandem with risk pot. For example, if the risk pot is ever lower than 5% on a project, the contingency pot should be used to make up the risk pot to 5% of the total cost of the project. It guarantees that appropriate risk pot has been included in the tender. If ever the contingency and the risk pot are lower than 5% on a project; then approval should be asked to ensure that risks are less

Weight	Feature
0.0750 ± 0.1159	Project size
0.0437 ± 0.0500	Project duration
0.0375 ± 0.0729	Work stream
0.0375 ± 0.0468	Region
0.0125 ± 0.0500	Contract
0.0063 ± 0.0250	Project type
0.0000 ± 0.1046	Work type
-0.0312 ± 0.0884	Sector

Figure 13: Significance of Attributes for Context-aware Benchmarking

than the amount requested. Spring enables such reasoning for all the KPIs to ensure proper governance and accountability.

Defining a reliable benchmark for profit margins is another tricky task, as several factors influence this judgment. By margins, we mean the actual margin that can be obtained after all projects costs, including firms overheads, are paid out. A project completed at 10% margin usually ends up having 1% or 2% margin. Forecasting models have lots to improve in this area as models rarely factor in overhead costs while preparing the cost estimates. The inclusion of context is found phenomenal in getting the right benchmarks for the profit margins. The given benchmark reveals that the firm will perform poorly if the margin on this project is set anything less than 11%, and will achieve average performance if the margin is set in the range of 11% and 15%. The tender shall aim for a margin of 15% and beyond to accomplish good profitability performance. Industry experts also vetted this fact. The learner predicts higher boundaries for small-sized projects and similar boundaries for large-sized projects of the same kind. The benchmark also provided opportunities to increase margins through the opportunity and innovation pots. The planned margin combined with additional margin obtained from the opportunity and innovation pots will become hidden margin that can go up to 30% for this project if the right resources are allocated to the project, and it is executed optimally.

9.1. Deep Learning Insights

This study was designed to train deep ensemble learner for obtaining reliable predictions for KPI benchmarks. This section slightly touches upon the need for deep learning insights towards understanding what the model has learnt from the data. Such capability is at the heart of debugging, informing feature engineering, future data collection, informing human decision-making and building trust. We employed the `permutation importance` algorithm to ask the learner about how key project attributes impact the benchmark formulation. Several other algorithms exist for performing this analysis. Our selection is mainly informed by the speed, popularity, and consistency of the underlying algorithm. A major advantage of `permutation importance` is that it can be applied to a learner without any need to make adjustments to it. The way it works is simple. `Permutation importance` algorithm randomly shuffles one attribute at a time, leaving the target and the rest unchanged, and then assess variations in the accuracy of predictions on the shuffled data.

Fig 13 displays the importance of each key project attribute. Attributes towards the top are considered more important, whereas ones towards the bottom are less significant. The first number in the figure depicts the decrease in learners' accuracy when the attribute is randomly shuffled. As with most ML algorithms, there is always some randomness in the performance change by shuffling an attribute. We repeatedly shuffle attributes several times to average out randomness. The number after ± reports the variance in performance from one reshuffle to the next. `Permutation importance` seldom return negative values that occur when predictions on shuffled data have higher accuracy than the real data. This implies that the attribute has no importance at all in the learner. In this study, the most important attribute for profitability benchmarking is revealed to be the `project size`. That seems sensible and is also witnessed by the professionals. `Work` type is considered as the least significant to the learner. Sector attribute is found entirely irrelevant for enabling context-aware benchmarking.

10. Implication for Practice—Contractors' Control

The clients in the construction industry always control the tender negotiations process and use their role to play constructors against their commercial gains during bid competitions. The contractors, to win the tender, frequently set essential tender KPIs blindly with no real guidance as to the implication of those choices on project delivery, beyond a common confidently uttered phrase "we should be alright" taken because of years of experience. However, the implications of such practices always end up being more than what was lightly anticipated. The primary purpose of this study is to develop an objective system for pushing the negotiations' power back into the contractors' hand through data-driven insights. The proposed deep ensemble learner empowers estimators with detailed knowledge about KPIs boundaries to use them as cheat sheets against their four-headed opponents. The estimators are not just informed about a single KPI (usually margin) instead of each of wisely chosen KPIs comprising the tender summary. They can smartly move those KPIs in a way that appeals to clients without compromising tender competitiveness. It also allows the estimators to clearly pinpoint precisely where a problem might arise and potentially plan for hidden possibilities within the estimate that can avoid margin erosion.

The system gives the contractor their playing boundaries for KPIs, and if the client persists, the contractor can confidently withdraw from the process knowing that it would have brought on too much loss, and doesn't worth the revenue it brings. The industry experts revealed several such incidents. A case study is mentioned here to bolster the adequacy of the proposed system. "At the tender stage, we went through 9 rounds of tender submission with a client before being awarded the contract. The client knowing they have the power to control negotiations due to lots of contractors participated in the competition. One or more of which were cut off at every round based on closeness to the target margin supplied by the client. The target margin got smaller at each round which the client claims is based on the lowest from the previous round. So, in respect of this, going into every round, we cut down our margin closer to the target margin supplied by the client. Eventually, at contract award, we were 3% lower than where we started and had well underestimated a lot of other KPIs. No surprises that the project incurred a lot of added cost from risks alongside omitted or underestimated items at tender, delays and defects. A lesson learned performed at the project close highlighted many things that went wrong could be linked to the inability of the contractor to evaluate KPIs during the tender negotiation process." A system like the one proposed here is a great rescue for enabling estimators to adjust KPIs based on data-driven insights.

11. Conclusions, Limitations and Future Work

In this study, we reported the development of an objective system for supporting the estimators during the tender evaluation process. The focus of the research was toward contractor facilitation as most systems, developed so far, facilitate clients in tender evaluation or supplier selection tasks. Besides, most tender evaluation tools have limited accuracy due to shortcomings in the underlying data collection, analysis and model development techniques. This study exercised a five-fold methodology for developing an end to end ML system. Text mining, focused group discussions, and Big Data of power infrastructure projects are exploited at various stages of the study to achieve the stated research objectives.

At the core of this research lies the development of deep ensemble learner based on the decomposition-integration strategy. The learner generates attribute-specific benchmarks for KPIs during the decomposition stage using a custom benchmarking algorithm. These benchmarks are collated in the integration stage using deep neural networks to yield final project-sensitive KPI benchmark. The proposed learner is deployed in the Spring system to facilitate estimators in the tender evaluation process. To move this research beyond lab experiments, a real-life tender of 60km power infrastructure project is assessed using the Spring system.

The intelligent user interface of the tool facilitated estimators in quickly understanding the strengths and weakness of the tender. The proposed learner is also scrutinised for its learning using permutation importance to see the significance of key project attributes for context-aware benchmarking. The overall goal of ML in this study is to develop technology for shifting power to control the tendering process back to contractors rather than their clients. The contractors need to know KPI limits and make informed choices during the tender evaluation and negotiation process. It is noticed that engaging end-users at the early stage in ML brings great benefits. The most important one is that the ML models will be much production-ready due to timely feedback from the domain experts.

While Spring can facilitate estimators to highlight tender issues, this functionality needs to be enriched by guiding the estimators with detailed instruction involved in the mitigation to resolve identified issues. n addition, the proposed deep ensemble learner suffers poor generalisability whenever a tender involving maintenance works for power infrastructure projects is evaluated through the Spring system. It is because the data used for training learner in this study is largely of new projects. The issue of generalisability shall be resolved for the broader intake of the system in the contractors' community. We intend to collect more data in future to extend our proposed learner for maintenance works projects using transfer learning approach.

References

Agneeswaran, V. S. (2014). *Big Data Analytics Beyond Hadoop: Real-Time Applications with Storm, Spark, and More Hadoop Alternatives*. FT Press.

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Bilal, M., Ajayi, S. O., & Akinade, O. O. (2018). A framework for big data analytics approach to failure prediction of construction firms. *Applied Computing and Informatics*, .

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Oyedele, A. A., Akinade, O. O., Bilal, M., & Ajayi, S. O. (2017). Critical factors for insolvency prediction: towards a theoretical model for the construction industry. *International Journal of Construction Management*, *17*, 25–49.

Bassioni, H. A., Price, A. D., & Hassan, T. M. (2004). Performance measurement in construction. *Journal of management in engineering*, *20*, 42–50.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *Advances in neural information processing systems* (pp. 153–160).

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H. A., Alaka, H. A., & Pasha, M. (2016). Big data in the construction industry: A review of present status, opportunities, and future trends. *Advanced engineering informatics*, *30*, 500–521.

Bonino, D., Corno, F., & De Russis, L. (2013). Real-time big data processing for domain experts, an application to smart buildings. *Big Data Computing/Rajendra Akerkar; Taylor & Francis Press: London, UK*, *1*, 415–447.

Bortolazza, R. C., Costa, D. B., Formoso, C. T. et al. (2005). A quantitative analysis of the implementation of the last planner system in brazil. In *13th International Group for Lean Construction Conference: Proceedings* (p. 413). International Group on Lean Construction.

Busby, M., Chapple, E., Matthews, R., & Chapple, I. (2013). Practitioner evaluation of a novel online integrated oral health and risk assessment tool: a practice pilot. *British dental journal*, *215*, 115.

Chen, L., & Pan, W. (2018). An intelligent decision support system for improving information integrity in procuring infrastructures in hong kong. In *Proceedings of the 21st International Symposium on Advancement of Construction Management and Real Estate* (pp. 213–221). Springer.

Cheng, M.-Y., & Roy, A. F. (2011). Evolutionary fuzzy decision model for cash flow prediction using time-dependent support vector machines. *International Journal of Project Management*, *29*, 56–65.

Colwill, C., & Gray, A. (2007). Creating an effective security risk model for outsourcing decisions. *BT technology journal*, *25*, 79–87.

Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M. X., & Qu, H. (2010). Context preserving dynamic word cloud visualization. In *Visualization Symposium (PacificVis), 2010 IEEE Pacific* (pp. 121–128). IEEE.

Diebold, F. X. (2000). Dynamic factor models for macroeconomic measurement and forecasting. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society,"(edited by M. Dewatripont, LP Hansen and S. Turnovsky)* (pp. 115–122).

Domingues, S., Almeidaand, N., & Sousa, V. (2017). Gross margin risk assessment in construction projects: The influence of indirect cost deviations. *Welcome to Delegats IRC 2017*, (p. 731).

El-Mashaleh, M. S., Edward Minchin Jr, R., & O'Brien, W. J. (2007). Management of construction firm performance using benchmarking. *Journal of Management in Engineering*, *23*, 10–17.

Fadhil Dulaimi, M. (2005). The challenge of customer orientation in the construction industry. *Construction Innovation*, *5*, 3–12.

Falagario, M., Sciancalepore, F., Costantino, N., & Pietroforte, R. (2012). Using a dea-cross efficiency approach in public procurement tenders. *European Journal of Operational Research*, *218*, 523–529.

Farooq, M. U., Thaheem, M. J., & Arshad, H. (2018). Improving the risk quantification under behavioural tendencies: A tale of construction projects. *International Journal of Project Management*, *36*, 414–428.

Fleuren, W. W., & Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods*, *74*, 97–106.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* volume 1. Springer series in statistics New York.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Guest, G., MacQueen, K. M., & Namey, E. E. (2011). *Applied thematic analysis*. sage.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*, 504–507.

Inzalkar, S., & Sharma, J. (2015). A survey on text mining-techniques and application. *International Journal of Research In Science & Engineering*, *24*, 1–14.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, *52*, 36–44.

Kim, C.-W., Yoo, W. S., Lim, H., Yu, I., Cho, H., & Kang, K.-I. (2018). Early-warning performance monitoring system (epms) using the business information of a project. *International Journal of Project Management*, *36*, 730–743.

Kissi, E., Adjei-Kumi, T., Badu, E., & Boateng, E. B. (2017). Factors affecting tender price in the ghanaian construction industry. *Journal of Financial Management of Property and Construction*, *22*, 252–268.

Lu, W., Chen, X., Peng, Y., & Shen, L. (2015). Benchmarking construction waste management performance using big data. *Resources, Conservation and Recycling*, *105*, 49–58.

Lu, W., Yuan, H., Li, J., Hao, J. J., Mi, X., & Ding, Z. (2011). An empirical investigation of construction and demolition waste generation rates in shenzhen city, south china. *Waste management*, *31*, 680–687.

Makovšek, D. (2014). Systematic construction risk, cost estimation mechanism and unit price movements. *Transport Policy*, *35*, 135–145.

Ogunlana, S. O. et al. (2010). Beyond the 'iron triangle': Stakeholder perception of key performance indicators (kpis) for large-scale public sector development projects. *International journal of project management*, *28*, 228–236.

Peterson, S. J. (2005). *Construction accounting and financial management*. Pearson Prentice Hall New Jersey.

Pino-Mejías, R., Jiménez-Gamero, M.-D., Cubiles-de-la Vega, M.-D., & Pascual-Acosta, A. (2008). Reduced bootstrap aggregating of learning algorithms. *Pattern recognition letters*, *29*, 265–271.

Rao, M. K., Kumar, V., & Kumar, P. R. (2018). Optimal contractor selection in construction industry: The fuzzy way. *Journal of The Institution of Engineers (India): Series A*, *99*, 67–78.

Rui, Z., Peng, F., Ling, K., Chang, H., Chen, G., & Zhou, X. (2017). Investigation into the performance of oil and gas projects. *Journal of Natural Gas Science and Engineering*, *38*, 12 – 20.

Sakurai, S., & Suyama, A. (2005). An e-mail analysis method based on text mining techniques. *Applied Soft Computing*, *6*, 62 – 71.

Samuel, D. (2018). Risk based tender evaluation using multi-criteria decision analysis. *Proceedings of the Institution of Civil Engineers-Management, Procurement and Law*, (pp. 1–44).

Tam, C., Zeng, S., & Deng, Z. (2004). Identifying elements of poor construction safety management in china. *Safety Science*, *42*, 569–586.

Tam, V. W., Tam, C. M., Zeng, S., & Ng, W. C. (2007). Towards adoption of prefabrication in construction. *Building and environment*, *42*, 3642–3654.

Taroun, A. (2014). Towards a better modelling and assessment of construction risk: Insights from a literature review. *International Journal of Project Management*, *32*, 101–115.

TCI (2017). The construction index. URL: https://www.theconstructionindex.co.uk/market-data/top-100-construction-companies/2017.

Thomas, R. (2015). *Big Data Revolution: What Farmers, Doctors, and Insurance Agents Can Teach Us about Patterns in Big Data*. John Wiley & Sons.

Vischer, J. C. (2018). Epilogue: From building evaluation to building performance evaluation and beyond. In *Building Performance Evaluation* (pp. 309–315). Springer.

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications*, *38*, 223–230.

Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, *26*, 61–68.

Wang, Y.-R., Chen, Y.-J., & Huang, C. J. (2009). Applying neural network ensemble concepts for modelling project success. In *International Symposium on Automation and Robotics In Construction (ISARC 2009), Austin, TX: University of Texas at Austin* (pp. 199–204).

Watt, D., Kayis, B., & Willey, K. (2009a). Identifying key factors in the evaluation of tenders for projects and services. *International Journal of Project Management*, *27*, 250 – 260.

Watt, D., Kayis, B., & Willey, K. (2009b). Identifying key factors in the evaluation of tenders for projects and services. *International Journal of Project Management*, *27*, 250–260.

Watt, D., Kayis, B., & Willey, K. (2010). The relative importance of tender evaluation and contractor selection criteria. *International Journal of Project Management*, *28*, 51–60.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2015). *Fundamentals of predictive text mining*. Springer.

Wong, C., Nicholas, J., & Holt, G. D. (2003). Using multivariate techniques for developing contractor classification models. *Engineering, Construction and Architectural Management*, *10*, 99–116.

Wong, J. M., & Ng, S. T. (2010). Forecasting construction tender price index in hong kong using vector error correction model.

Construction management and Economics, *28*, 1255–1268.
Zhang, Y., Luo, H., & He, Y. (2015). A system for tender price evaluation of construction project based on big data. *Procedia Engineering*, *123*, 606 – 614.