Journal of Building Engineering 26 (2019) 100850

Contents lists available at ScienceDirect

Journal of Building Engineering

journal homepage: www.elsevier.com/locate/jobe

Investigating profitability performance of construction projects using big data: A project analytics approach

Muhammad Bilal**, Lukumon O. Oyedele*, Habeeb O. Kusimo, Hakeem A. Owolabi, Lukman A. Akanbi, Anuoluwapo O. Ajayi, Olugbenga O. Akinade, Juan Manuel Davila Delgado

Big Data Enterprise and Artificial Intelligence Laboratory (Big-DEAL), University of the West of England (UWE), Frenchay Campus, Bristol, United Kingdom

ARTICLE INFO

Keywords:
Big data
Project analytics
System architecture
Machine learning
Profitability performance

ABSTRACT

The construction industry generates different types of data from the project inception stage to project delivery. This data comes in various forms and formats which surpass the data management, integration and analysis capabilities of existing project intelligence tools used within the industry. Several tasks in the project lifecycle bear implications for the efficient planning and delivery of construction projects. Setting up right profit margins and its continuous tracking as projects progress are vital management tasks that require data-driven decision support. Existing profit estimation measures use a company or industry wide benchmarks to guide these decisions. These benchmarks are oftentimes unreliable as they do not factor in project-specific variations. As a result, projects are wrongly estimated using uniform rates that eventually end up with entirely unusual margins either due to underspends or overruns. This study proposed a project analytics approach where Big Data is harnessed to understand the profitability distribution of different types of construction projects. To this end, Big Data architecture is recommended, and a prototype implementation is shown to store and analyse large amounts of projects data. Our data analysis revealed that profit margins evolve, and the profitability performance varies across several project attributes. These insights shall be incorporated as knowledge to machine learning algorithms to predict project margins accurately. The proposed approach enabled the fast exploration of data to understand the underlying pattern in the profitability performance for different types of construction projects.

1. Introduction

The construction industry is a hypercompetitive, project-based industry. Firms need to be at the best of project planning and control to make sustainable profits. A slight planning oversight can lead to severe problems during the project delivery stage [1]. Setting up the right profit margin for a construction project is one of those vital planning tasks which are carried out without considering project-specific intricacies. As a result, projects lose margins during the delivery stage due to additional costs incurred due to reworks or unanticipated new works [2]. Firms can get bankrupt due to making losses on a few construction projects. The issue of margin erosion is sometimes attributed to the poor estimation practices used currently within the industry. There are several opportunities for utilising digital technologies to revitalise the project planning and control workflows in order to remain competitive in the industry. Firms can undertake decisions like margin estimation in a more informed fashion by harnessing all types of data available for the projects.

The introduction of Building Information Modelling (BIM) has improved the availability of data in the construction sector [3]. Firms accumulate projects data, right from the opportunity selection stage to successful project delivery. This data arises in many forms and formats. Besides, the volume, variety and the speed at which this data is generated make it classified as Big Data. However, the users are getting increasingly frustrated as the data is siloed in different systems, and there is no unified view to analyse this data to answer task-specific questions accurately. Existing tools in the industry, based on the notion of project intelligence, explore these data in a backward manner, which are suitable to run regular construction project management reports [4,5]. These tools are ideal for producing listings and asking close-ended questions that involve summaries or aggregates. This overwhelming project data can be utilised in a forward manner to revitalise numerous tasks like profit margin estimation. This capability can enable the estimators to factor-in project specific insight and generate reliable estimates. This research proposes a project analytics approach as a key technology to meaningfully process all forms of construction

* Corresponding author.
** Corresponding author.
E-mail addresses: Muhammad.bilal@uwe.ac.uk (M. Bilal), L.Oyedele@uwe.ac.uk (L.O. Oyedele).

https://doi.org/10.1016/j.jobe.2019.100850
Received 15 February 2019; Received in revised form 23 June 2019; Accepted 23 June 2019
Available online 03 July 2019

data to answer task-specific questions through data-driven insights.

Big Data technologies have an enormous potential to efficiently handle large amounts of projects data using a cluster of commodity servers [6,7]. There is a huge interest in harnessing Big Data for analytics, to not only understand large volumes of data through advanced statistical and visualisation methods but also develop predictive models to inform future decisions through data-driven insights [8]. These insights are vital to inform decisions about future implications of project decisions which can empower the project team to foresee end-of-life state of projects and forecast actual project performance by then [9]. The computational requirement of interventions like project analytics thereby calls for the applications Big Data Analytics [8]. Such intervention is key to revitalise the construction industry and offers opportunities for planning and controlling construction projects through data-driven insights. This synergistic integration of Big Data technologies with project analytics promises huge opportunities for growth within the industry.

This study is part of more extensive research that aims to develop a construction simulation platform for project analytics. The platform will provide data-driven insights to different users of the system for optimally performing project tasks. This platform entails multidisciplinary efforts and knowledge from both technology professionals and project delivery experts. This study is the first step in that direction and contributes to the research by proposing, implementing and validating a Big Data architecture for construction project analytics. We developed critical components for storing and analysing project data using Big Data technologies. The prototype platform is evaluated through understanding the issue of profitability performance in construction projects. It is revealed during our exploratory analysis that profit margins tend to deviate substantially from what was initially planned in estimates. These variations need to be explored, and their relationship with specific project attributes need to be uncovered. With this objective, large volumes of construction projects data are loaded and analysed in this paper. Key findings revealed that profit margins evolve and profitability performance varies across project attributes. Such insights shall be considered while developing machine learning algorithms for forecasting margins. The proposed Big Data platform enabled us to quickly explore and understand these insights from large volumes of project data.

The remainder of this paper is organised as follows: In the next section, the research methodology of the paper is presented. Section 3 discusses the relevance of Big Data in construction industry along with strengths and weaknesses of competing Big Data platforms for project analytics. Section 4 explains the proposed Big Data architecture for profitability analytics. In section 5, the prototype deployment strategy is discussed. Section 6 describes the analysis performed over Big Data prototype to understand the profitability performance in construction projects. In section 7, conclusions are presented along with the directions for future work.

2. Research methodology and focus

Profitability performance evaluation is one of the main pillars of project analytics. To fully understand the domain of project analytics, several focused group interviews (FGIs) were carried out with the domain experts from the construction industry involving bid directors, estimators, project managers and foremen. They highlighted a large number of perspectives to project analytics based on their own level of engagement. All FGIs interactions were recorded, transcribed and analysed to extract the key theme for analysis and the development of project analytics platform. A detailed business specifications document was prepared from the findings of FGIs. in Fig. 1 shows a subset of these requirements intended for construction project planning. These specifications were then analysed with respect to their computational requirements for data storage and analysis to determine the right set of big data technologies. System analysts with varying degree of experience in enterprise IT infrastructure and software development were engaged. It is realised that the development of entire platform involves a diverse set of technology artefacts to be synergistically integrated. This led to the development of Big Data architecture to place these diverse technologies in a way that could be easily translated into a prototype to evaluate the adequacy of technology for performing analytics underpinned by various project related tasks. Once the architectural components have been agreed, a prototype implementation is performed to quickly enable construction workers to use the system to answer different types of queries and test their hypothesis.

The profitability performance varies across the construction projects. There is not much literature on this topic to understand the relationship between profit margins and key attributes of construction projects. To understand the relationship between profitability performance and project attributes, data from a large number of systems is integrated into the prototype Big Data enabled storage layer. Some of these systems include Telematics, Oracle financials, CRM, Google Earth, KML, Business Objects, General Ledger, Mobile Inventory, Project Control Database, Procurement & Payments System, Job Costing Reports, Primavera, and Digital Briefcase. A generic schema to standardise the data elements from these diverse data sources is established and this data is integrated into the platform for analysis. Various data elements required for project analytics for 2709 projects were curated. It comprised 5.7 million cells. These data elements were combined by these projects. For this research, project attributes like region, client, voltage, duration, start and end dates, workstreams, work types, and project types were needed that were combined from additional sources like Oracle financials and CRM. The final list contained 1,048 projects where all project attributes needed in this analysis were considered.

```
SELECT COUNT(*) "Sample",
   ROUND(STDDEV (profit_margin), 2) "SD",
      ROUND(AVG(profit_margin), 2) "Mean",
         MEDIAN (profit_margin) "Median",
            MIN (profit_margin)||' - '||MAX (profit_margin) "Range"
FROM projects_jcr_data_final;
```

Listing 1. SQL command to retrieve summary stats for all projects.

This data is investigated to understand how profit margins fluctuate with respect to various project attributes. Structured query language (SQL) queries were used to analyse data from the platform. Listing 1 shows an example SQL query used to retrieve the summary statistics. The overall aim was to retrospectively investigate this relationship in the past projects and use that knowledge to inform the decision-making for new opportunities. The results were discussed with industry experts working on construction projects to validate the legitimacy of our findings. A number of flaws in the data analysis were highlighted in our initial discussions of results. They guided the analytics process to include the most relevant projects in the analysis. As a result, several projects were excluded before the final analysis are carried out. The results presented in this study are based on the final set of 1048 projects which remained the most relevant after a series of filters are applied.

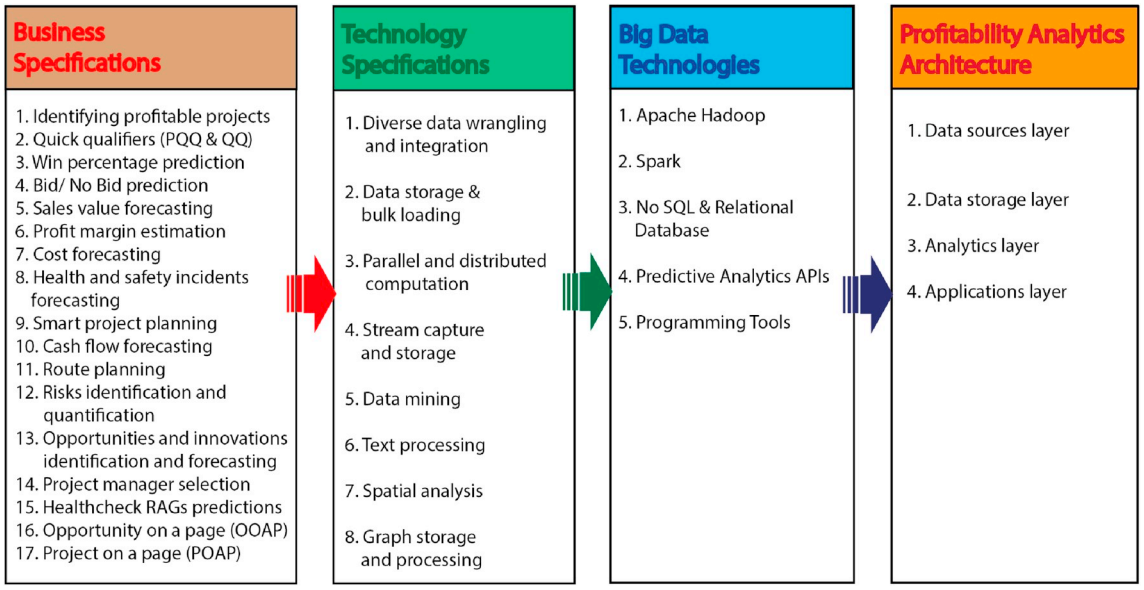

Fig. 1. Business specifications to Big Data architecture.

3. Big Data Analytics—Literature review

The term Big Data is first coined by Diebold [7] as an emerging availability of the massive amounts of potentially relevant datasets. Laney [10] identified Big Data to have three defining characteristics—also called 3Vs of Big Data—including (1) Volume (terabytes, petabytes of data and beyond) (2) Variety (heterogeneous formats like text, sensors, audio, video, graphs and more) (3) Velocity (continuous streams of the data). According to Jacobs [4]; Big Data is the kind of data whose size compels the community to look beyond state-of-the-art data management and analysis technologies. So, Big Data is relative and will always be there. Tomorrow's data might not work with today's technologies. Gartner definition of Big Data is high volume, velocity and variety of information that necessitate cost-effective, innovative forms of information processing for enhanced insights and decision-making [11]. Systematically analysing Big Data to identify underlying trends is the top strategic agenda of many modern businesses [12–14].

There is a growing curiosity in the construction industry to utilise the information in Big Data for analytical purposes [15]. Big Data Analytics is the enabling toolbox for knowledge discovery from massive datasets. Companies today are not only interested in describing past data through exploratory analytics, but they are more keen about uncovering the latent trends to forecast future events through predictive analytics [5,8]. Business insights are hidden inside the data. These insights could have the potential to reshape the entire business through the data-driven decision-making. The early identification, understanding and reaction to hidden trends in data is a competitive edge for companies [10,16].

Various Big Data platforms have been developed so far with varying capabilities. The selection of a right tools for the given application requires an in-depth knowledge of these platforms [12]. Notably, in the case of analytical applications like the profitability prediction and monitoring, the capability of the tools to adapt to emerging workload outweighs the rest of selection criteria. Big Data platforms are of two kinds, the horizontal scaling (HSPs) and the vertical scaling platforms (VSPs) [5]. The HSPs distribute processing across multiple servers and scale out to the huge workloads by adding more machines to the cluster. The VSPs carry out the computation on a single server and scaling up to the emerging workload by upgrading the processor or memory or hardware. For the sake of this study, the scope of discussions is kept limited to HSPs. Notably, the Apache Hadoop and Berkeley Big Data Analytics Stack (BDAS). Their selection is informed by the data management and computational requirements of the proposed profitability analytics architecture. These provisions include (but not limited to) supporting iterative algorithms, handling compute-intensive tasks and providing near real-time visualisations to support end users tasks. Interested readers are suggested to read more about other competing Big Data platforms in Ref. [5].

The synergistic integration of digital innovations with project analytics can revolutionalise the industry. The resulting intervention will provide firms with solutions to optimise projects through data-driven insights which will likely to enhance the overall profitability performance and efficiency of the industry.

4. Proposed Big Data Architecture

In this section, the proposed Big Data architecture for project analytics is explained (see Fig. 2). Business specifications were used to identify core computational and storage requirements, which led to inclusion of various components in the architecture. These components are organised into several layers based on their relatedness. There are five key layers in the architecture.

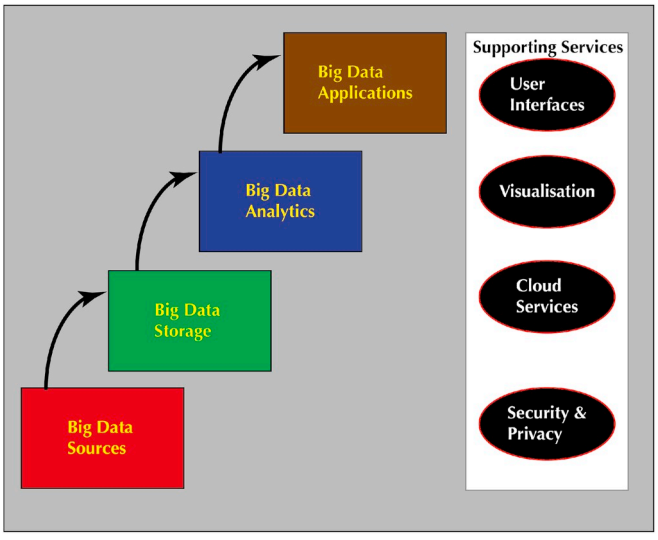

Fig. 2. An overview of proposed big data architecture for project analytics.

4.1. Big data sources layer

This layer is conceived to handle the data integration from diverse construction projects data produced by the construction industry. This layer handles two types of data. Firstly, the historical data of all construction projects, including design, opportunities, bidding, estimates, finances, project plans, job costing reports, health and safety, to name a few. This data is captured once and used several times in exploratory and predictive analytics. Traditional Big Data technologies like relational databases can handle this kind of data. Secondly, the real-time data of construction projects, which has to be captured at real-time. It usually involves streaming data received from various sources like job costing reports, telematics or site progress tracking. Specialised Big Data technologies like Kafka [17], Flume [18], Sqoop [19], and Flink [20] are designed to handle diverse streaming data. This study used Flume for capturing projects data from job costing reports to monitor the profitability performance with every transaction occurring on the project to report the live updates. Most data sources, whether streaming or historical, require data standardisation for which we employed domain-specific ontologies.

4.2. Big data storage layer

This layer is responsible for handling the data storage of construction projects. The data is loaded to the staging area for pre-processing, standardisation and cleansing. Data inside the staging area is deleted once it is processed and stored into the actual tables used within the application. The persisted data stay longer as it is utilised in the development of predictive models. This data is often of good quality as most issues related to standardisation, missing values and outliers have been fixed during the pre-processing stage. We used Hadoop Distributed File System (HDFS) for managing the data within the staging area while employed Resource Description Framework (RDF) enabled Network Data Model (NDM) for storing the persistent data. Additionally, this layer also stores the data related to predictive models. This include model weights along with pre-processing used during the model training. Most of the models-related data is stored using predictive model mark-up language (PMML) for interoperability with diverse machine learning systems.

4.3. Big Data Analytics layer

This layer is responsible for enabling technologies to develop and test predictive models to understand not only large amounts of projects data but also forecasting critical components of construction projects. These tasks entail huge expressivity to describe the analytical problem and then run it to perform specific analysis. For the sake of this paper, we analysed the profitability performance across various project attributes and used Structured Query Language (SQL) for performing the descriptive analytics to understand this relationship. Parallel processing and distributed computation are enabled at this layer through Big Data technologies like Apache Spark. Spark is reported to outperform other parallel processing tools like Map Reduce due to its built-in capabilities of in-memory computation Dean [16]. We harnessed the Spark capabilities like SparkR, PySpark, MLLib, and GraphX to support various types of user analysis. These technologies enable the data processing underpinning different analysis at speed for effective project planning and control.

4.4. Big data applications layer

This layer enables access to analytics services via domain-specific end-user applications. Various types of applications can harness the project analytics architecture including mobile, web and desktop. We developed a web application to interact with project analytics services to efficiently plan and control construction projects. Java Enterprise Edition (Java EE) features like Web Services, Java Server Faces, Java Servlets, Enterprise Java Beans and Java Messaging were utilised. JDeveloper 12c is used for developing the prototype system. JUnit is employed for testing the functionality of the architecture, and Weblogic 12c server is used to deploy the application.

4.5. Big data supporting services layer

This layer provides the services related to the adaptive user interface, interactive visualisation and security of the system. All other layers harness these services. User interface services include page templates, skins and declarative components to support the development of domain-specific applications. The UI controls can be adapted according to end-user proficiency. The architecture is designed to harness the enterprise software security features of the Java EE, which underlies the Java platform security (JPS) to ensure secure web services. Besides, data security is ensured using virtual private database (VPD), which is a state-of-the-art security configuration to safeguard data from unauthorised users. The VPD allows row-level security (RLS) where users can only see rows which they are granted the access.

5. Prototype deployment of architecture

The prototype deployment of the proposed system is shown in Fig. 3. Key technologies used in deployment include:

1. **Virtualisation platform:** VM hypervisor is configured on a standalone host with heightened specifications, i.e., Quad Core CPU, 128G memory and 20 TB storage. We used virtualisation to ensure better management of resources and agility to upgrade or downgrade the prototype. Five VMs are created. Two VMs (primary & standby) were used to manage web and database servers each. This is to ensure data protection and maximum availability.
2. **Database server:** Oracle database 12c is used with Real Application Cluster (RAC) configured to take care of the storage related to the project analytics application.
3. **Hadoop cluster:** One VM is used for Hadoop cluster to store diverse data coming from external systems. This data is mainly stored as

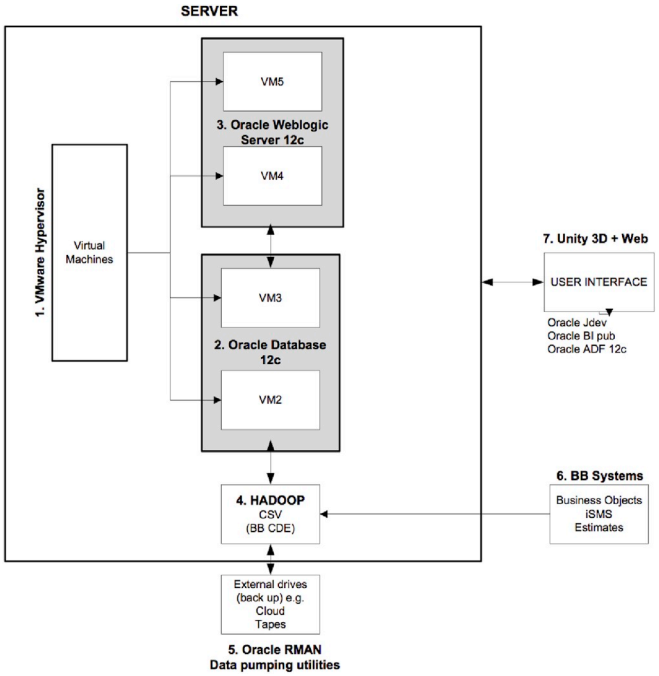

Fig. 3. Block deployment diagram of proposed architecture.

CSV files over the Hadoop distributed file system (HDFS) for parallel processing using Spark jobs.

4. **External systems:** External systems in the deployment indicate sources of construction projects data. Mostly, these systems manage data in their local formats that need to be transformed in CSV using a standard vocabulary and loaded on the HDFS.
5. **Data migration utilities:** These utilities are to keep the backups of the entire architecture. Recovery Manager (RMAN) and data pump utilities are employed. RMAN takes physical backups of the data and is the most reliable recovery tool. However, the backup sizes are large, which is not suitable for daily backups. For regular backups, data pump is used that extracts the logical backup of the entire database.
6. **User interface utilities:** We used Unity 3D and JDeveloper 12c for developing prototype application for project analytics system. Business Intelligence (BI) publisher is used for dashboards and reporting.

This virtualisation based prototype deployment enabled the architecture to ensure maximum availability and fault tolerance. A test was carried out to check disruption during the long-running SQL queries for project analytics by turning off the primary VMs. The prototype automatically redirected queries to standby VMs. This test is repeated ten times, and average query completion time is recorded. The query results were correct due to redirection to other VMs. However, the execution time has increased by 17%. There are strategies like caching to handle such delays. Overall, we found the prototype reliable to such failures.

6. Profitability performance analysis

In this section, the proposed architecture is evaluated using exploratory data analysis, and some preliminary results are drawn. Past ten years of data of construction projects are loaded and analysed. The goal of this evaluation is to confirm the adequacy of the proposed architectural components for descriptive analytics. In this context, some results are provided that are obtained next. Interestingly, the findings of this research reveal that profit margins vary significantly across various project attributes.

6.1. Distribution of profit margin

We begin our analysis by looking at the overall distribution of profit margin. Profit margin is an essential metric that carries critical insights to understand profitability performance and quickly identify when projects are not performing well. Since this is a complex topic, we needed to understand the distribution of profit margin to understand profitability performance in construction projects. To this end, we aggregated data on project costs from Oracle financials with data from sales systems to compute profit margins. Only completed projects were considered in this analysis.

Table 1 shows the summary statistics for 1048 projects. Median and average profit margins are 19.46% and 21.48%, respectively. Their closer proximity indicates the data distribution is sound for predictive modelling. However, the data spread is long as depicted by a range of − 93.81% (projects having big losses) to 100% (projects with enormous profits) with a large standard deviation of 26.58%. This reveals an enormous shift in profitability performance across projects. This variance is often attributed to using uniform rates during margin estimation. The acceptable tolerance for variance in profit margin shall be in ± 20 range [21], which is obviously not the case in the underlying data.

Table 1
Summary statistics of all projects.

Projects	Sample	Mean	SD	Median	Range
All	1048	21.48%	26.58%	19.46%	−93.81%–100%

We explored this trend further by dividing projects into several groups. Fig. 4 shows inverted histogram to visualise this distribution. For typical data distribution, histogram forms a symmetrical bell-curve. Fig. 4 shows skewed distribution towards the profit side, which gives the impression that the company accrued profits on many (86%) projects. Only a few (14%) projects are on the loss side. This interpretation is misleading as the company actual profit starts after 8% after the company overheads are subtracted from the profit. With this insight, the distribution shifts and 24% projects fall on the loss side, whereas 76% on profit. So, 793 projects made profits and 250 projects ran into losses. The results of the binomial statistical test of ($Pr < 0.02$) means that there is a statistically significant fluctuation between projects that run into losses to the ones that made profits. If we expand the eligible profit boundaries by adding 20% to both sides of average margin (21.48%) as expected tolerance, 56% projects fall on the profit side whereas 44% on the losses side, which reveals a huge disparity and poor profitability performance practices. Our discussions with the estimators revealed that poor estimation from constructors or clients side often lead to such situation in projects. It sometimes put projects in severe cash flow problems during the project delivery stage. More reliable approaches are requested for profit margin estimation in the construction projects.

6.2. Profit margin trend

It is a general perception that the profit margin remains in a constant range across construction projects. The most prevalent range within the sector falls between 10% and 15% of the total sales value. Some industries like Transport have revealed that gross margins bear no relationship with the time, and it neither increases nor decreases with time. Oil and gas industry have also reported similar findings. There is not much literature in the construction industry to explain this trend. We did analyses to evaluate profit margins trend in construction projects.

Fig. 5 explains this trend for the past ten years of completed construction projects. It is evident in the chart that the profit margin has gradually increased. The rate of change, though, has sharply decreased. A closer look at the chart reveals two major trends. The first one is a sharp and abrupt increase in margins between 2008 and 2012. Margins increased from 10% to 19.89% during this time. The second trend is for projects between 2013 and 2017, which small but steady increase. Overall, the average profit margin evolved from 9.82% to 21.48%, with an average rate of change of 1.17% yearly. This rate increased abruptly in 2010 and 2012 by 3.45% and 3.8%, respectively. As a whole, the rate of change was high until 2014. After that, profit margins converged around 21%. The average increase in margins for the past three years is 0.06%. Data shows a slightly upward trend between profit margin and time.

We grouped projects into two clusters, namely pre-2014 and post-2014, to check the significance of difference statistically in the growth. These clusters were studied using one-way ANOVA *t*-test. The test results revealed a p-value of $P_r < 0.02$ that indicate a significant statistical difference. The reason behind the sharp upward trend in pre-2014 is deliberated with industry experts. They listed several reasons. The most apparent one was the new entry into this business. The firm was eager to get more projects without care for huge profits. Margins evolved as the firms established in this business which is revealed by the data. However, they still struggle to decide right profit margins, which impact their bidding by losing many opportunities. This evolving trend shall be considered in forecasting for profit margins of new projects.

6.3. The impact of region on profit margins

Every region has its own peculiarities so a region may affect profit margins of construction projects. Some researchers have found its

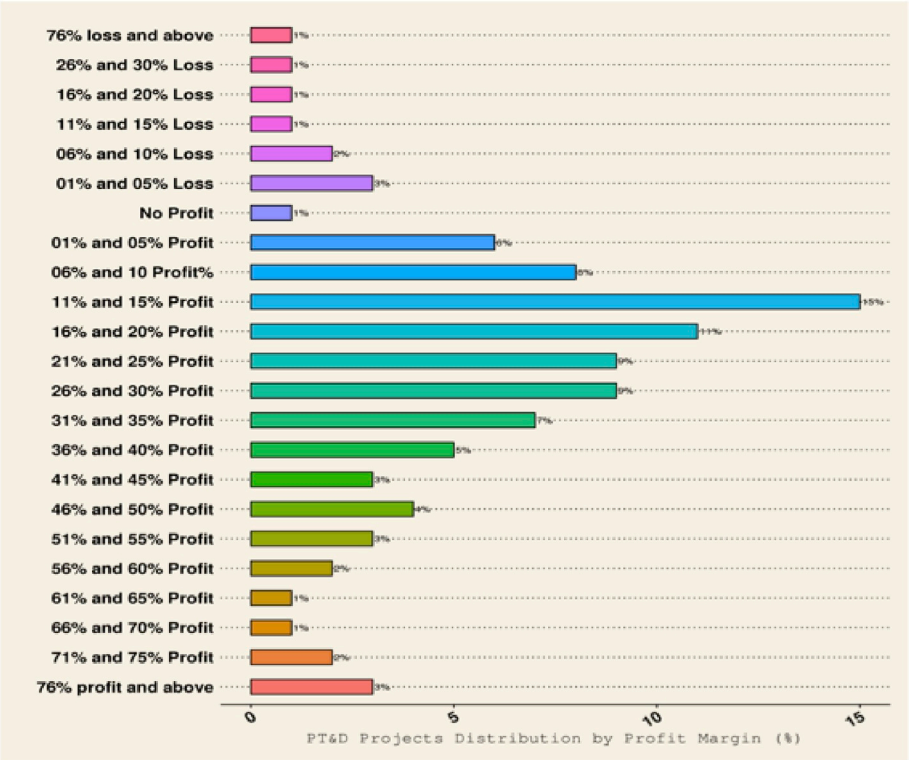

Fig. 4. Profitability performance distribution in construction projects.

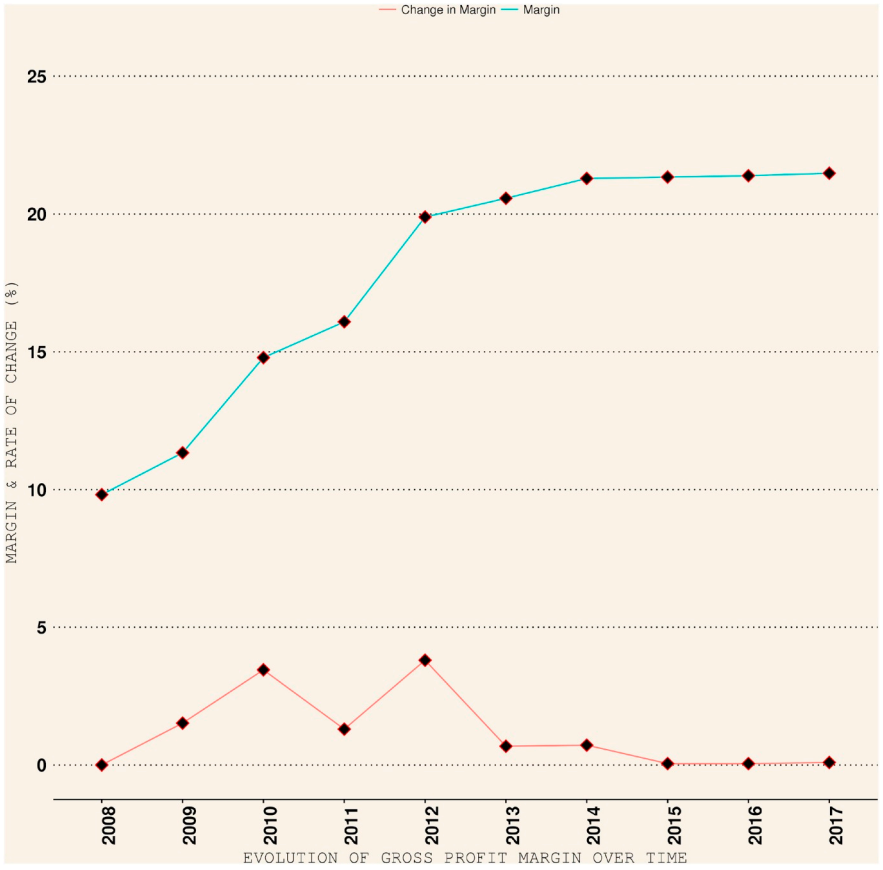

Fig. 5. Profit margins trend.

Table 2
Summary statistics based on region.

Regions	Number of Projects		Project Value		Profit Made		Summary Stats of Gross Profit Margin (%)			
	By Value	By (%)	By Value (in 100 k)	By (%)	By Value (in 100 k)	By (%)	Average	SD	Min	Max
Region 1	2	0.19	£0.55	0	£0.29	0.01	61.22	48.89	26.65	95.79
Region 2	313	30.01	£1217.10	10.31	£229.32	11.46	30.09	32.65	−92.23	100
Region 3	277	26.56	£4195.26	35.53	£738.06	36.9	20.09	23.41	−92.74	90.24
Region 4	37	3.55	£586.54	4.97	£97.49	4.87	18.98	16.48	−1.82	77.41
Region 5	187	17.93	£1904.80	16.13	£314.04	15.7	17.62	18.48	−57.4	70.91
Region 6	195	18.7	£3827.93	32.42	£615.83	30.79	15.14	24.12	−93.81	100
International	32	3.07	£75.86	0.64	£5.33	0.27	10.91	26.95	−44.67	97.39

association with different project activities. Flyvbjerg et al [22]. identified varying cost overruns across different regions. Olaniran et al [23]. observed that project location affects technical and installation tasks in projects. Alexander Hall & Delille [21] identified weather, surface condition, soil condition, remoteness, and availability of materials are key to determine the profitability performance in projects. Currently, there is limited literature on how profit margins get affected by different region-specific characteristics. This section provides some basic insights across seven regions in the dataset. Table 2 shows the summary statistics across these regions. A cursory look reveals that data distribution is uneven across regions. Region 2 has the largest share of projects at about 30.01%, followed by Region 3 with 26.56%, Region 6 with 18.7%, Region 5 with 17.93%, Region 4 with 3.55%, International with 3.07%, and Region 5 having 0.19%.

Most projects are from Region 2, Region 3, Region 5 and Region 6 whereas, the mainstream profit arises from Region 3 (36.09%) and Region 6 (30.79%). The average profit margin varies significantly across different regions. Region 1 has the highest profit margin of 61.22%; however, the number of projects undertaken in Region 1 are few, and hence the region can't be considered the most profitable region. Pointedly, projects in Region 1 are mainly Minor – small projects. Region 2 and Region 3 have a large number of projects and have the highest profit margins of 30.09% and 20.09%, respectively. Wales/South West and Region 5 s these regions with profit margins at 18.98% and 17.62%, respectively. Region 6 and International projects are least profitable at 15.14% and 10.91%, respectively. Some of the projects in the Region 2, Region 3 and Region 6 have margins less than − 90%, which shows serious profitability issues in those regions.

Except for the Region 1, the majority of regions have incurred losses on projects. Some regions like Region 4 shows both extremes losses (−1.82%) and huge profit (100%). There are several reasons for such extreme profits. However, this analysis shows huge variations in terms of profit margins and losses by project regions. The statistical One-way ANOVA test is performed to see how profit margins based on regions. The result ($P_r = 4.514e - 11$) reveals significant statistical variance in profit margins by regions. This shows that regions influence profit margins distinguishably. This interaction must be captured in profit estimation systems.

6.4. The impact of project type on profit margins

Project types impose specific design and operational specifications. These include (i) greenfield and (ii) brownfield projects. Greenfield projects are newly conceived projects, which are often carried out in the undeveloped sites. These projects tend to be flexible regarding design changes and can easily be adapted to the current legislation needs or operational requirements. These projects also need access routes to sites and require planning permissions before their execution. On the downside, it is always hard to find locally skilled labour or historical data explaining the ground realities. In contrary, brownfield projects are aimed at modifying existing facilities. Modifications include extensions or renovations. These projects tend to have historical data, prior permissions, existing infrastructure, and in some cases, prior local experience. However, these projects have limited design flexibility, predefined operational efficiency, and in many cases, produce waste due to demolition. Some industries have explored this relationship [24,25]. It is revealed that brownfield projects often incur cost overruns whereas companies make good profits on greenfield projects. We explored this peculiarity in construction projects data.

Table 3 shows the summary statistics based on project type. The dataset included a large number of greenfield projects which constituted 90% whereas remaining 10% were brownfield projects. Accordingly, greenfield projects have high total value, i.e.(78.9%), and average profit margins of (82.83%). Brownfield projects constituted 21.1% of project value and made 17.17% of the total profit. The average profit margin for greenfield projects is higher (22.4%) with projects fall with extreme profits within the range of − 93.81% and 100%. Similarly, brownfield projects have smaller average profit margins (13.11%) with projects fall in extreme range of − 54% and 89%. This fact confirms the findings of other industries. The industry experts highlighted that brownfield projects are susceptible to huge cost overruns. The spread of margins for project type is wide due to bigger values of standard deviation, i.e., 27.12% and 19.15%. These wider spreads also expose unreliable approaches used in the industry for predicting profit margins.

6.5. The impact of workstream on profit margins

Firms employ several workstreams to ensure financial sustainability and stability. This strategy also enables more opportunities for growth and reduces the risks of bankruptcy if a particular workstream fails. The dataset includes three workstreams, including (i) overhead power transmission lines (OHL), (ii) cabling, and (ii) substation. The OHL involves the construction of electrical power transmission and distribution projects to carry the bulk of electrical energy over the vast distance. The substation consists of the development of infrastructures to store the electricity transmitted over transmission lines. Cabling

Table 3
Summary statistics based on project type.

Project type	Number of Projects		Project Value		Profit Made		Summary Stats of Gross Profit Margin (%)			
	By Value	By (%)	By Value (in 100 k)	By (%)	By Value (in 100 k)	By (%)	Average	SD	Min	Max
Greenfield	940	90.12	£9316.81	78.9	£1656.82	82.83	22.4	27.12	−93.81	100
Brownfield	103	9.88	£2491.25	21.1	£343.54	17.17	13.11	19.15	−54.71	88.81

Table 4
Summary statistics based on workstream.

Work stream	Number of Projects		Project Value		Profit Made		Summary Stats of Gross Profit Margin (%)			
	By Value	By (%)	By Value (in 100 k)	By (%)	By Value (in 100 k)	By (%)	Average	SD	Min	Max
Transmission	209	20.04	£4364.90	36.97	£897.83	44.88	23.37	24.85	−93.81	100
Cabling	767	73.54	£7038.92	59.61	£1046.26	52.3	21.16	26.98	−92.74	100
Substations	67	6.42	£404.23	3.42	£56.27	2.81	19.21	27.29	−37.12	97.39

Table 5
Summary statistics based on sector.

Sector Types	Number of Projects		Project Value		Profit Made		Summary Stats of Gross Profit Margin (%)			
	By Value	By (%)	By Value (in 100 k)	By (%)	By Value (in 100 k)	By (%)	Average	SD	Min	Max
Others	60	5.75	£640.92	5.43	£103.77	5.19	25.33	34.78	−93.81	100
Power & Energy	883	84.66	£10,460.55	88.59	£1817.96	90.88	22.75	25.39	−92.23	100
Telecom	47	4.51	£416.83	3.53	£66.35	3.32	8.45	36.29	−92.74	59.2
Roads & Highways	43	4.12	£235.15	1.99	£5.43	0.27	8.12	9.99	−28.82	19.9
Transport	10	0.96	£54.6	0.46	£6.84	0.34	5.25	35.12	−91.98	32.2

projects involve the development of infrastructures to distribute electricity from substations to consumers. The industry experts believed that profit margins vary by workstreams. This relationship is thereby investigated and explained here.

Table 4 shows summary statistics across different workstreams. Cabling projects are abundance, covering 73.54% of data. OHL followed cabling that comprised 20.04% of data. Substation projects were fewer, i.e., 6.42%. The firm has delivered 59.61% cabling, 36.97% OHL and 3.42% substation projects. Similarly, 52.3% profit arose from cabling, followed by OHL with 44.88% and finally substation projects with 2.81% profit. Overall, OHL projects are fewer than cabling projects, but the revenue accrued from these projects is substantial. OHL projects gained more profits by having the average profit margin of 23.37%, and range of − 93.81% and 100%. Cabling projects, despite their vast number, comes next in terms of making profits (i.e. 21.16%), and range of − 92.74% and 100%. Finally, substation projects have the least profit margins of 19.21%, and range of − 37.12% and 97.39%. These insights are useful for business development teams while selecting profitable opportunities and creating estimates.

6.6. The impact of sector on profit margins

The relationship between profit margins and sector is also very complicated since sector-related factors influence project activities invariably. There is no literature to describe this non-trivial relationship in construction projects. The dataset included five sectors, namely (i) Power and energy, (ii) Roads and highways, (iii) Telecommunications, (iv) Transport, and (v) Others.

Table 5 shows summary statistics of profit margins by sectors. Sectors have a varying proportion of projects in the dataset. Majority of projects (84.66%) are power and energy projects, followed by others, Telecom and Roads & Highways sectors with 5.75%, 4.51%, and 4.12% of share, respectively. The transport sector has a smaller number of projects, comprising 0.96% share. In terms of total project value, 88.59% of projects are of Power & Energy sector, second by Others with 5.43%.

The transport sector has the smallest project value constituting 0.46% of all projects. In terms of profitability, 90.88% profit arose from Power & Energy sector, followed by Others sectors with 5.19%. Some sectors such as Telecom, Road & Highways, and Transport projects have the least amounts of profit on projects. The average profit margin in these sectors falls within the range of 5.25% and 25.33%. Others sectors highest profit margins up to 25.33%, with projects having varied profits ranging from − 93.81% to 100%. However, the number of projects in Others sector (i.e., 5.75%) is too small relative to the sectors mentioned earlier.

The data has more projects of Power & Energy sector, where the firm has made an average profit margin of 22.75%, just above the overall average profit margin of 21.48%. Profit falls drastically in Telecom, Roads & Highways, and Transport projects until 10% and below. In summary, the difference in profits made by the firm varies significantly across various sectors that reveals a sector-aware sensitivity on construction projects.

6.7. The impact of work type on profit margins

Construction projects can further be classified based on the types of work. Since different work types involve various activities, they are likely to influence profit margins differently. Construct only, Design & Build, Extend, Existing Maintenance, Pre-construction services, Refurbishment, and Supply only, are seven types of works, recorded in the dataset. There were some non-standard entries listed in the dataset, which are adequately handled through the data cleansing process. There is no literature informing how work types affect profit margins in projects. This section explains our findings to fill this knowledge gap.

Table 6 shows summary statistics by work types. In terms of sample size, 37.39% of projects are New Build, followed by Maintenance and

Table 6
Summary statistics based on work type.

Work Types	Number of Projects		Project Value		Profit Made		Summary Statsof Gross Profit Margin (%)			
	By Value	By (%)	By Value (in 100 k)	By (%)	By Value (in 100 k)	By (%)	Average	SD	Min	Max
Maintenance	307	29.43	£653.52	5.53	£142.81	7.14	29.01	34.33	−93.81	100
Preconstrution	8	0.77	£27.61	0.23	£4.16	0.21	26.99	14.7	5.69	43.96
Supply Only	38	3.64	£106.0	0.9	£8.34	0.42	23.52	33.24	−15.53	97.39
Refurbishment	300	28.76	£5426.16	45.95	£1018.94	50.94	18.46	22.72	−92.74	99.2
New Build	390	37.39	£5594.77	47.38	£826.1	41.3	17.56	19.79	−91.98	95.79

Table 7
Summary statistics based on contract types.

Contractual Types	Number of Projects		Project Value		Profit Made		Summary Stats of Gross Profit Margin (%)			
	By Value	By (%)	By Value (in 100 k)	By (%)	By Value (in 100 k)	By (%)	Average	SD	Min	Max
Alliance Agreement	6	0.58	£46.08	0.39	£23.04	1.15	43.67	16.67	13.85	56.59
Schedule of Rates	355	34.04	£1258.59	10.66	£217.95	10.9	26.33	32.71	−93.81	100
Lump Sum	307	29.43	£2876.94	24.36	£488.84	24.44	22.17	23.51	−91.98	97.39
Early Contract Involvement	4	0.38	£12.34	0.1	£2.12	0.11	21.32	12.29	5.69	35.09
Framework Agreement	48	4.6	£388.98	3.29	£83.47	4.17	19.79	14.42	−29.74	49.42
Traditional	144	13.81	£3151.15	26.69	£689.8	34.48	18.83	24.04	−92.74	78.4
Target Cost	24	2.3	£357.07	3.02	£52.25	2.61	15.32	9.57	4	39.59
Remeasurable	37	3.55	£774.25	6.56	£102.37	5.12	14.35	15.56	−43.67	55.35
Cost Reimbursable	77	7.38	£2207.80	18.7	£221.28	11.06	12.21	20.79	−32.45	99.2
Design & Build	26	2.49	£584.08	4.95	£108.23	5.41	10.58	23.12	−36.52	73.52
Construct Only	10	0.96	£130.49	1.11	£8.54	0.43	10.57	28.11	−36.14	52.91
Supply Only	3	0.29	£13.93	0.12	£1.36	0.07	7.48	16.18	−7.32	24.76
Term Maintenance	2	0.19	£6.35	0.05	£1.11	0.06	−1.41	38.24	−28.45	25.63

Refurbishment projects with 29.43% and 28.76% share, respectively. A few projects are of Preconstruction and Supply only types that form 0.77% and 3.64% of data. New Build projects the highest of total project value (i.e., 47.38%). They are seconded by Refurbishment projects with a value of 45.95%. Maintenance project, despite their enormous size, i.e., 29.43%, their total project value is smaller, i.e., 5.53%.

The cumulative value of Preconstruction services and Supply only projects is 1.13%. Refurbishment projects topped other work types in terms of profits made (50.94%) which are followed by New Build projects where the firm makes 41.3% of profit. The average profit margin falls in the range of 17.56% and 29.01%. Maintenance projects are found to have the highest margins (29.01%). Preconstruction services and Supply only projects have large profit margins of 26.99% and 23.52%, respectively. Refurbishments and New Build projects made relatively smaller profits of 18.46% and 17.56%, respectively.

The industry experts perceived New Build to make more profits than refurbishment ones (Chen et al., 2015). However, this trend is found wrong in the data. A primary reason for this is that the majority of New builds are high-valued projects. This fact reveals work types influence margins differently and shall be considered in the predictive modelling tasks.

6.8. The impact of contract type on profit margins

This section explores the relationship between profit margins and contractual project type. The contract type is a legal binding between the client and the contractor explicitly describing essential terms concerning the construction project. The industry professionals have preferences for certain types of contracts as they believe those types ensure planned margins. However, no literature work to corroborate or falsify this assumption. Contract types found in the dataset include: (1) Alliance Agreement, (2) Schedule of Rates, (3) Lump Sum, (4) Early Contract Involvement, (5) Framework Agreement, (6) Traditional, (7) Target Cost, (8) Remeasurable, (9) Cost Reimbursable, (10) Design & Build, (11) Construct Only, (12) Supply Only and (13) Term Maintenance.

Table 7 shows summary statistics of profitability performance of construction projects based on the contractual types. Contractual types vary based on their number of projects in the data. There are a large number of Schedule of Rates type projects, i.e., 34.04%, followed by Lump Sum and Traditional types, having 29.43% and 13.81% projects, respectively. A small number of projects are Cost Reimbursable (7.38%), Framework Agreement (4.6%), Reimbursable (3.55%) and the rest. Traditional and Lump Sum contracts have the highest project value, i.e., 26.69% and 24.44%, respectively. They are followed by Schedule of Rates, with projects of worth 10% of the entire projects.

The firm has made a considerable profit, i.e., 34.48% from the Traditional contracts, followed by 24.44% from the Lump Sum contracts. Reimbursable and Schedule of Rates contracts also have a significant profit share of 11.06% and 10.9%, respectively. The average margin of projects falls within the range of − 1.41% and 43.67%. Alliance Agreement contracts are found to have the highest profit margin of 43.67% — however, this contract has a small number of projects in the data. Schedule of Rates projects are also plentiful in data (34.04%) and the company has made huge profit margins of 26.33% in those projects. Lump Sum and Early Contract Involvement projects have excellent profitability performance of 22.17% and 21.32% respectively. Design & Build, Construction only, Supply only, and Term maintenance projects are amongst the least profitable contracts with profitability performance 10.58%, 10.57%, 7.48%, and − 1.41%, respectively.

This fact reveals a substantial diversity of making profits based on different contract types. Overall, it is clear from the results profitability performance vary significantly by varying contractual types.

6.9. The impact of duration on profit margins

Project duration is another project attribute that carries a great insight to see variations of profit margins in construction projects. Making huge profits in short-term projects than long-term projects is often more desirable. However, there exists no literature to learn how this relationship holds in construction projects. This section provides some insights to fill this knowledge gap. To carry out this analysis, some project features such as the project start and project end dates are used to form various clusters of projects by duration. A new feature is, therefore, created with following discrete categories, including (i) Up to 1 year (ii) Up to 2 years (iii) Up to 3 years (iv) Up to 4 years(v) Up to 5 years (vi) Up to 6 years (vii) and 7 years & above.

Table 8 shows summary statistics for the profitability performance based on project duration. Projects in each project duration category are not equally distributed. Projects Up to 3 years are largest in the data, i.e., 33.27%, followed by the second largest are projects of Up to 4 years (21.09%). Projects of Up to 5 years come next with 19.18%. Projects of Up to 2 years are the least, i.e., (2.97%). 9.4% of all projects are the ones that span 7 years & above.

The company has delivered projects of total project value in Up to 3 years (35.44%) and Up to 4 years (35.25%) categories. The projects Up to 5 years have a total worth of 13.25%. More profits are made in projects of Up to 3 years (33.32%) and Up to 4 years (46.25%). A good amount of profit (10.22%) is made from the projects of 7 years & above. The rate of change of gross profit margins across the project duration is significant. The average profit margins fall in the range of 6.02% and 27.62%. There is a clear variability trend in the data. Projects of Up to 3 years (27.62%), Up to 4 years (25.3%), and Up to 2 years (24.47%) have the highest profitability performance. Whereas, projects of longer duration than 4

Table 8
Summary statistics based on project duration.

Project Duration	Number of Projects		Project Value		Profit Made		Summary Stats of Gross Profit Margin (%)			
	By Value	By (%)	By Value (in 100 k)	By (%)	By Value (in 100 k)	By (%)	Average	SD	Min	Max
Upto 3 year	347	33.27	4185.25	35.44	666.61	33.32	27.62	25.15	10.5	100
Upto 4 year	220	21.09	4162.08	35.25	925.14	46.25	25.3	16.43	−71.38	53.76
Upto 2 year	31	2.97	151.14	1.28	0.93	0.05	24.47	36.18	−10.37	68.9
Upto 1 year	78	7.48	400.81	3.39	26.72	1.34	19.97	31.82	−50.31	69.59
Upto 5 year	200	19.18	1564.74	13.25	135.83	6.79	16.75	19.31	0.54	68.56
Upto 6 year	69	6.62	423.69	3.59	40.73	2.04	14.44	21.51	−4.29	44.5
7 year & above	98	9.4	920.33	7.79	204.4	10.22	6.02	43.93	−93.81	35.47

years reduce profit margins drastically. This is evident from Up to 5 years (16.75%), Up to 6 years (14.44%), and 7 years & above (6.02%).

A similar pattern of low profits is revealed for the short-term projects of Up to 1 year (19.97%). In short, the profit margins on completion are found higher for Up to 2 years, Up to 3 years, and Up to 4 years, whereas smaller profits are exhibited for short term projects of Up to 1 year and the long term projects of 5 years and beyond. Long-term projects are susceptible to more problems due to long durations that result in cost overruns (Chen, Liu, & Wei, 2013). The analysis clearly revealed a trend in the project duration, and the way profit margins are anticipated on construction projects.

7. Conclusions

In this paper, we demonstrated the Big Data approach for project analytics to understand the profitability performance of construction projects. Project analytics enables us to look at diverse construction data in a forward-looking manner to comprehend the domain better using data-driven insights. We proposed Big Data architecture for which a prototype implementation is provided. Various components of the architecture are derived from our focused group interviews (FGIs) with domain experts to capture the requirements for project analytics. The prototype architecture is evaluated for exploratory analytics to understand the domain of profit margins in construction projects. To this end, a large number of construction projects are loaded into the Big Data storage and evaluated using advanced Structured Query Language (SQL) commands and statistical methods.

Our analysis revealed that the way firms gain profits on construction projects varies significantly by project attributes. Existing uniform rate based profit margin based profitability estimation and monitoring approaches are unreliable. It is a norm in the industry that most construction projects begin with planned margins and eventually end up having entirely different margins. This trend is largely attributed to inadequate estimation approaches like following a company or industry-wide benchmarks. These benchmarks are stale and mostly unable to capture project specific nuances. However, there is no evidence that project specific details impact the way a firm attains profit margins. We analysed various project attributes and their impact on profit margins. We found that profit margins evolve, and the profitability performance varies across several project attributes. These insights call for the consideration of project specific knowledge during training the machine learning algorithms to predict profit margins accurately. The proposed Big Data architecture enabled us to quickly explore vast amounts of projects data to see swiftly underlying trends, and understand the relationship between profit margins and project attributes.

This study is the part of a larger research project that intends to develop a construction simulation tool for project analytics. The proposed Big Data architecture will serve the majority of data storage and computational needs required to implement a furnished construction simulation tool. This intervention will enable various stakeholders involved in the delivery of construction projects to analyse data from multiple dimensions to make decisions based on evidence and machine learning insights. We plan to develop a large number of machine learning models based on similar insights to support estimators and projects in carrying out project-specific tasks in a much informed fashion.

Acknowledgment

The authors would like to express their sincere gratitude to Innovate UK through grant application number 54832–413479; file number 102473 and Engineering and Physical Science Research Council (EPSRC) through grant reference number EP/S031480/1 for providing financial support to carryout this study.

References

[1] Z. Rui, F. Peng, K. Ling, H. Chang, G. Chen, X. Zhou, Investigation into the performance of oil and gas projects, J. Nat. Gas Sci. Eng. 38 (2017) 12–20.
[2] S.-Y. Kim, T.-A. Huynh, et al., Improving project management performance of large contractors using benchmarking approach, Int. J. Proj. Manag. 26 (2008) 758–769.
[3] J.G. Motwani, V.E. Sower, A. Kumar, J. Antony, T.S. Dhakar, Integrating quality function deployment and benchmarking to achieve greater profitability, Benchmarking Int. J. 13 (2006) 290–310.
[4] A. Jacobs, The pathologies of big data, Commun. ACM 52 (2009) 36–44.
[5] D. Singh, C.K. Reddy, A survey on platforms for big data analytics, J. Big Data 2 (2015) 8.
[6] J. Bughin, M. Chui, J. Manyika, Clouds, big data, and smart assets: ten tech-enabled business trends to watch, McKinsey Q. 56 (2010) 75–86.
[7] F.X. Diebold, Dynamic factor models for macroeconomic measurement and forecasting, in: M. Dewatripont, L.P. Hansen, S. Turnovsky (Eds.), Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society, 2000, pp. 115–122.
[8] E. Siegel, Predictive Analytics: the Power to Predict Who Will Click, Buy, Lie, or Die, John Wiley & Sons, 2013.
[9] F. Provost, T. Fawcett, Data Science for Business: what You Need to Know about Data Mining and Data-Analytic Thinking, O'Reilly Media, Inc., 2013.
[10] D. Laney, 3d data management: controlling data volume, velocity and variety, META Group Res. Note 6 (2001) 70.
[11] J.S. Ward, A. Barker, Undefined by Data: a Survey of Big Data Definitions, (2013) arXiv preprint arXiv:1309.5821.
[12] V.S. Agneeswaran, Big Data Analytics beyond Hadoop: Real-Time Applications with Storm, Spark, and More Hadoop Alternatives, FT Press, 2014.
[13] D. Bonino, F. Corno, L. De Russis, Real-time Big Data Processing for Domain Experts, an Application to Smart Buildings. Big Data Computing/Rajendra Akerkar vol. 1, Taylor & Francis Press, London, UK, 2013, pp. 415–447.
[14] R. Thomas, Big Data Revolution: what Farmers, Doctors, and Insurance Agents Can Teach Us about Patterns in Big Data, John Wiley & Sons, 2015.
[15] M. Bilal, L.O. Oyedele, O.O. Akinade, S.O. Ajayi, H.A. Alaka, H.A. Owolabi, J. Qadir, M. Pasha, S.A. Bello, Big data architecture for construction waste analytics (cwa): a conceptual framework, J. Build. Eng. 6 (2016) 144–156.
[16] J. Dean, Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, John Wiley and Sons, 2014.
[17] G. Wang, J. Koshy, S. Subramanian, K. Paramasivam, M. Zadeh, N. Narkhede, J. Rao, J. Kreps, J. Stein, Building a replicated logging system with Apache kafka. Proceedings of the VLDB Endowment vol. 8, (2015), pp. 1654–1655.
[18] S. Hoffman, Apache Flume: Distributed Log Collection for Hadoop, Packt Publishing Ltd, 2013.
[19] K. Ting, J.J. Cecho, Apache Sqoop Cookbook: Unlocking Hadoop for Your Relational Database, O'Reilly Media, Inc., 2013.
[20] O.-C. Marcu, A. Costan, G. Antoniu, M.S. Pérez-Hernández, Spark versus flink:

understanding performance in big data analytics frameworks, Cluster Computing (CLUSTER), 2016 IEEE International Conference on, IEEE, 2016, pp. 433–442.
[21] N. Alexander Hall, S. Delille, Cost estimation challenges & uncertainties confronting oil and gas companies. Cost Engineering-Morgantown vol. 54, (2012), p. 14.
[22] B. Flyvbjerg, M.K. Skamris Holm, S.L. Buhl, What causes cost overrun in transport infrastructure projects? Transport Rev. 24 (2004) 3–18.
[23] O.J. Olaniran, P.E. Love, D. Edwards, O.A. Olatunji, J. Matthews, Cost overruns in hydrocarbon megaprojects: a critical review and implications for research, Proj. Manag. J. 46 (2015) 126–138.
[24] M. Skitmore, S. Stradling, A. Tuohy, H. Mkwezalamba, The Accuracy of Construction Price Forecasts, University of Salford, 1990.
[25] Z. Shehu, I.R. Endut, A. Akintoye, G.D. Holt, Cost overrun in the malaysian construction industry projects: a deeper insight, Int. J. Proj. Manag. 32 (2014) 1471–1480.