Predicting Completion Risk in PPP Projects Using Big Data Analytics

Hakeem A. Owolabi, Muhammad Bilal, Lukumon O. Oyedele, Hafiz A. Alaka, Saheed O. Ajayi, and Olugbenga O. Akinade

***Abstract*—Accurate prediction of potential delays in public private partnerships (PPP) projects could provide valuable information relevant for planning and mitigating completion risk in future PPP projects. However, existing techniques for evaluating completion risk remain incapable of identifying hidden patterns in risk behavior within large samples of projects, which are increasingly relevant for accurate prediction. To effectively tackle this problem in PPP projects, this study proposes a Big Data Analytics predictive modeling technique for completion risk prediction. With data from 4294 PPP project samples delivered across Europe between 1992 and 2015, a series of predictive models have been devised and evaluated using linear regression, regression trees, random forest, support vector machine, and deep neural network for completion risk prediction. Results and findings from this study reveal that random forest is an effective technique for predicting delays in PPP projects, with lower average test predicting error than other legacy regression techniques. Research issues relating to model selection, training, and validation are also presented in the study.**

***Index Terms*—Benchmark, Big Data, completion risk (CR), forecasting, predictive modeling, public private partnerships (PPP).**

I. Background

In RECENT decades, the construction industry has been caught up in the frenzy of the widespread digital revolution that is shaping global landscape [14]. More than ever, the industry is witnessing an era of vast accumulation of valuable data needed for making informed decisions [14]. The rising availability of electronic data in diverse formats [multidimensional (n-D) computer-aided design (CAD) data, three-dimensional (3-D) geometric encoded data, graphical data, video, audio, text, etc.] and sizes (terabytes, petabytes etc.) has intensified the adoption of fast technologies with strong analytical capabilities within construction industry (Caldas *et al.,* 2002). One of these frontier technologies is Big Data. Big Data are enormously large dataset that may be analyzed computationally to uncover hidden patterns, unknown correlations, trends, or preferences [88]. Typically, Big Data have three essential attributes, also known as the 3Vs, which distinguish it from traditional datasets [112]:

1) volume (terabyte, petabyte, exabyte, etc.);
2) velocity (continuous data streams and fast processing);
3) variety (disparate datasets in graphics, texts, pictures, audio, video, graphs, etc.).

These 3Vs are clearly apparent in most construction project data in recent times, providing opportunities for unraveling useful information from large data sample.

With robust analytical and data mining capabilities, Big Data conducts advanced analytics such as inferential analytics, predictive analytics, prescriptive analytics, and descriptive analytics [47], [76], [100]. While inferential analytics focuses on the interactions of explanatory variables with the target variable in the dataset (LaValle *et al.,* 2012), descriptive analytics examines what is happening now based on historical data [112]. Predictive analytics is concerned with prediction of future probabilities, trends, and patterns within a dataset [88], while prescriptive analytics adopts optimization and simulation algorithms to propose best possible outcomes and solution [16]. In this study, we examine predictive modeling of completion risk (CR) in public private partnership (PPP) projects using Big Data Analytics (BDA). Gatzert and Kosub [40] described CR in construction projects as the uncertainty that a project will be completed at a contractually agreed date. Recent literatures have examined completion risk analysis in PPP projects using various statistical tools such as Monte Carlo simulation, stochastic method, linear modeling, project evaluation review technique (PERT), critical path method, etc. (Kokkaew and Chiara, 2010, [25], [58]). Despite their immense contributions, most studies have concentrated on few project samples and limited data sources from simple relational databases (Soibelman *et al*. 2008, Kokkaew and Chiara, 2010, [49]). As such, these studies have either been adjudged deterministic or fixated on identifying generic factors influencing project delay (Kokkaew and Chiara, 2010). This is a major flaw in current completion risk analysis tools, as they remain incapable of identifying hidden patterns and trends in completion risk behavior that are relevant for accurate forecasting of completion risk across large portfolio of PPP projects. The adoption of Big Data enabled predictive modeling techniques is therefore imperative for accurate prediction of completion

Manuscript received April 5, 2018; revised August 10, 2018; accepted September 19, 2018. Date of publication November 21, 2018; date of current version April 17, 2020. Review of this manuscript was arranged by Department Editor T. Daim. *(Corresponding author: Lukumon O. Oyedele.)*

H. A. Owolabi, M. Bilal, L. O. Oyedele, and O. O. Akinade are with the Big Data Analytics Laboratory, Bristol Business School, University of West of the England, Bristol BS16 1QY, U.K. (e-mail: hakeem.owolabi@uwe.ac.uk; muhammad.bilal@uwe.ac.uk; l.oyedele@uwe.ac.uk; olugbenga.akinade@uwe.ac.uk).

H. A. Alaka is with the Faculty of Engineering, Environment and Computing, Coventry University, Coventry CV1 5FB, U.K. (e-mail: ac7485@coventry.ac.uk).

S. O. Ajayi is with the School of Built Environment and Engineering, Leeds Beckett University, Leeds LS1 3HE, U.K. (e-mail: saheed.jayi@aleedsbeckett.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TEM.2018.2876321

risk within this context. These predictive techniques will enable in-depth investigation of the dynamic interaction of underlying factors influencing project delay. In this regard, high precision analytics techniques such as deep neural network (DNN), random forest (RF), support vector machine (SVM), linear regression, and regression trees will be adopted for predictive purposes. The overarching aim of this study is therefore to develop the best BDA-based predictive model that can be used to estimate delay in PPP projects. In order to achieve the above aim, the following objectives have been identified for the study.

1) To identify the factors influencing delay in PPP projects and their dynamic interaction in large project samples.
2) To use advanced BDA techniques to predict completion risk in large portfolio of PPP projects.
3) To compare and contrast the predictive performance of these techniques toward completion risk forecasting in large project samples.

This study seeks to examine the behavior of completion risk across large PPP project portfolio. Using Big Data-driven predictive analyses, 4294 PPP projects between year 1992 and 2015 were examined across Europe for completion risk prediction. Section II of this study focused on literature review and examines the application of BDA in construction projects, smart cities, and Internet of Things (IOT). Existing techniques for completion risk evaluation in PPP projects were also discussed under the same section. While Section III presents the research methodological framework for the study; Section IV presents analysis of various predictive models for estimating completion risk in PPP projects. This is then followed by the implication for practice, while the last section concludes the study.

II. Literature Review

A. Big Data Analytics for Construction Projects and Smart Cities

The introduction of building information modeling (BIM) has helped fast-track the generation of humongous construction data across domains such as design data, enterprise resource planning systems, project schedules, financial data, and contract data among others. Many of these datasets exist in disparate formats including 3-D geometric encoded (BIM), drawing exchange format, ifcFXML (industry foundation classes XML), DWG (drawing data), DOC, XLS, PPT (Microsoft format), RVT (short for Revit), DGN (short for design), JPEG (image format), RM/MPG (video format), etc. With the emergences of sensors and embedded devices allowing facilities to generate real-time data in large volumes, variety, and under high velocity (a.k.a 3V's), the construction industry has been pushed into the Big Data era. Noticeably, despite the euphoria about BDA in the construction sector, academic literature on the topic is only gradually intensifying.

However, a quick review of construction literature revealed two emergent themes of Big Data application in the construction sector, namely Waste Analytics or Waste Management and Smart Cities vis-à-vis IOTs (Internet of Things). Lu *et al*. [66] in an investigation into construction waste performance in Hong Kong developed robust KPIs for benchmarking waste generation rate using data from waste disposal records of 5764 projects. The study found demolition works as the largest contributor to waste in Hong Kong, with new building, renovation, and maintenance contributing the least amount of waste to landfill. In another relevant literature, Bilal *et al*. [13] bemoaned existing intelligence-based waste management softwares as lacking the necessary ability to encourage stakeholders. The study also challenged the inappropriate classification of most wastes as mixed wastes under the existing waste management approaches. The study proposed a new Big Data architecture for designing-out waste from projects (by integrating Spark with BIM), and leveraged data from over 200 000 waste disposal records from 900 U.K. projects. Similarly, [66] conducted a comparative analysis of construction waste management performances in public and private projects under similar waste management governance. The study analyzed over 2 million waste disposal data from 5700 projects and concluded that construction contractors perform better on waste minimization when working on public projects than on private projects. In addition, Brown *et al*. [17] investigated the readiness of the construction sector for the adoption of BDA using sentiment analysis. Other relevant studies on Big Data in construction and engineering projects include Hampton *et al*. (2011), [14], [114].

Conversely, BDA along with the wide adoption of embedded devices in hard infrastructures have also intensified discussions on Smart Cities and IOT (Zanella *et al*., 2014, [19]). Chiang and Zhang [21] described smart cities as urban locations that use advanced communication technologies to collect and leverage electronic data via sensing devices. Through sensors, physical objects are able to stay connected through the Internet and transmit data online (IOT) in way that helps manage public assets, improve operational and resource efficiency [89]. Within the construction sector, smart cities and IOT have become a new and exciting area attracting noticeable research interests [41], [68], [81], [89]. For instance, whilst Bibri (2018) examined the state-of-the-art sensor-based Big Data application that are enabled for IOT in a sustainable environment, Osman (2018) investigated the necessary attributes of BDA algorithms suitable for developing city level smart information services. Also, in a new study done by Rathore *et al*. (2018) on exploiting IOT and BDA, sensors deployment at smart home, smart parking, surveillance, weather, vehicular networking, etc., were used to collate real-time data for developing a smart digital city service including graphically represented smart transport system. In addition, Alshawish *et al*. (2016) demonstrated practical applications of Big Data in a smart city under real-life situations, including smart energy, smart traffic systems, and smart public safety, by reviewing Big Data algorithms, city data collection, analysis, and optimization protocols. Similarly, Ming *et al*. (2018) analyzed the intentions behind smart city development in a city using Taiwan as a context and proposed a hierarchical model of smart city systems and data flow platform that leverages city sensor devices. However, while other studies have continued to examine Big Data, IOT, and smart cities within construction and engineering literature [20], [41], [89], [114], there remains

a dearth of relevant literature leveraging data from PFI/PPP projects on Big Data application despite the significant public resources involved.

B. Existing Techniques for Evaluating Completion Risk in PPP Projects

Earlier studies have examined completion risk in PPPs including Kokkaew and Chiara (2010), [6], [38], Ye and Tiong (2003), Hoffman (2008). Fight [38, 9] defines completion risk as *"the risk that projects do not yield (sufficient) revenues as a consequence of time and budget overruns."* Similarly, Kokkaew and Chiara (2010) refer to completion risk as the uncertainty of construction completion. For the purpose of this study, completion risk is considered as the uncertainty that a project will be completed at a contractually agreed deadline (Project Delay). Many literatures (i.e., [101], Hoffman, 2008, [49], [91]) have attributed completion risk to a number of factors within the construction process such as defective design of project, delayed access to project site, shortage in skilled labor, etc. Additionally, studies have suggested a number of techniques for completion risk evaluation in construction projects (Ye and Tiong, 2003, Jannadi and Almishari, 2003, Kokkaew and Chiara, 2010, [25], [58]). For instance, Ye and Tiong (2003) argued for the use of incentive schemes (bonuses) to project participants toward ensuring timely completion. The incentive scheme was assumed a function of time and other factors (such as complexity of project, source of revenue, etc.), and calculated thus:

$$B\left(t,\lambda_1,\lambda_2,R\right)=\begin{matrix}\lambda_1 R\left(T_s-t\right)\left(0\le t<T_s\right)\\ \lambda_2\ R\left(T_s-t\right)\left(T_s\le t<\infty\right)\end{matrix} \tag{1}$$

$$\frac{\lambda_1 R\left(T_s-t\right)\ \left(0\le t<T_e\right)}{\lambda_2 R\left(T_s-t\right)\ \left(T_e\le t<T_s\right)}$$

$$B\left(t,\lambda_1,\lambda_2,R\right)=\lambda_2 R\left(T_s-t\right)\left(T_s\le t<T_1\right)$$
$$\lambda_2 R\left(T_s-T_1\right)\left(T_1\le t<\infty\right). \tag{2}$$

The immense contribution of the U.S. navy in 1950s also saw the development of a tool for planning and coordinating large-scale projects, known as PERT. PERT presents a network diagram that provides a visual depiction of the critical paths in a project schedule and the sequence in which they must be completed. PERT is calculated as follows:

$$\text{Mean duration of activity } i\rightarrow \mu_i=\left\{\frac{a_i+4m_i+b_i}{6}\right\}$$

$$\text{Variance of activity } i\rightarrow \text{Var}_i=\left\{\frac{b_i-a_i}{6}\right\}$$

$$\text{Mean of critical path}\rightarrow \bar{\mu}=\sum_{j\in C}\mu_{j,}$$

where C is a set of critical activities

$$\text{Variance of critical path}\rightarrow \sum_{j\in C}\text{Var}_j.$$

Other completion risk analysis techniques have also been proposed such as linear-scheduling model, critical path method (CPM), Gantt Chart, vertical production method, line of balance, etc. However, despite their wide adoption overtime, André [6] argued the reliability of current risk analyses techniques, with their associated inaccuracies regarding completion risk, is limited by the use of out-dated analysis techniques (see Table I for exiting techniques for project scheduling and completion risk analysis). With the vast accumulation of project data in the construction industry, current risk analysis techniques and softwares, including COMFAR III Expert (UNIDO, 1994), CASPAR (Willmer, 1991), EVALUATOR (Abdel-Aziz and Russell, 2006), and INFRISK (Dailami *et al*., 1999), lack the technological capabilities to hold and analyze large volumes of disparate project data at high speed. As such, a BDA predictive modeling of completion risk remains the realistic option.

C. Big Data Predictive Analytics Techniques

BDA is predominantly employed for either inference (understanding the influence of explanatory variables over response variable) or prediction (predicting values of the response variable). Since the aim of this study is twofold i.e., understanding the interactions of explanatory variables on completion risk in PPP projects (inference), as well as devising a robust completion risk prediction model (prediction), a mix of parametric and nonparametric techniques are used for predictive modeling. These techniques are discussed in depth in the subsequent sections to fulfill the purpose of this study.

1) Regression as the Learning Problem: When learning problem is about predicting the quantitative response, the problem is referred to as regression problem. Regression analysis involves single or multiple predictors while predictive modeling. The abstract form of regression analysis is given in the following equation:

$$\boldsymbol{Y}=f\left(\boldsymbol{X}\right)+\epsilon \tag{3}$$

where $\mathbf{Y}$ is quantitative response; f is some fixed unknown function of predictors $\boldsymbol{X}$, and $\boldsymbol{\epsilon}$ is some random error term that is independent of $\boldsymbol{X}$ and has a mean of zero. In (3), $\boldsymbol{f}(\boldsymbol{X})$ provides systematic information about $\boldsymbol{Y}$ and its relationship with ρ predictors. Formally, $\boldsymbol{f}(\boldsymbol{X})$ can be expressed as shown in the following equation:

$$f(\boldsymbol{X})=\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\times\boldsymbol{x}_1+\boldsymbol{\beta}_2\times\boldsymbol{x}_2+\cdots+\boldsymbol{\beta}_{\boldsymbol{p}}\times\boldsymbol{x}_{\boldsymbol{p}} \tag{4}$$

where $\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_{\boldsymbol{p}}$ represents ρ predictors and $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_{\mathbf{p}}$ represents coefficients of ρ predictors and $\boldsymbol{\beta}_0$ is the intercept term. These coefficients quantify association between predictors and the response. In this study, coefficients are derived from a large array of PPP projects using various BDA techniques. And to assess predictive performance of model, residual sum of square (RSS) is usually employed. RSS is the square of difference of distance between the predicted value ($\hat{\mathbf{y}}$) and the actual value ($\mathbf{y}$). The following equation describes the RSS for regression analysis:

$$\boldsymbol{RSS}=\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}}(\mathbf{y}_{\boldsymbol{i}}-\hat{y}_{\boldsymbol{i}})^2. \tag{5}$$

TABLE I
EXISTING TOOLS FOR EVALUATING COMPLETION RISK IN PROJECTS

Existing Tools for Completions Risk Analysis	Origin	Features	Capabilities	Shortcomings	Literature References
Gantt Chart	Developed by Henry Gantt in 1917	Gantt displays simple activities or events that are plotted against time.	Static break down of tasks, deliverables, and milestones, analytical capabilities	Deterministic and cannot capture uncertainties in construction process.	Bossink (2004), El-Sayegh (2008); Kangari (1995), Russell and Jaselskis (1992)
Critical Part Method	Developed by Integrated Engineering Control Group (I.E.C) in 1956	It represents the longest duration in a project as a critical path , and if activities in this path are delayed will result in the overall project delay.	Uses computer algorithm, analytical capabilities	It is ineffective and cumbersome for scheduling linear continuous projects. Impact of uncertain delays is omitted	Ling and Hoi (2006); Russell and Jaselskis (1992); Dissanayaka and Kumaraswamy (1999), El-Sayegh (2008)
Program Evaluation Review technique	Developed during the 1950s by the U.S. Navy	Can handle extremely large number of activities. Also suitable for activities that are discrete in nature.	Planning and coordinating large-scale projects. Its network diagram provides visual representation of the major project activities	Useful only when major elements (events) in a have been completely identified. and cannot capture uncertainties in construction process. Sometimes relies on inspired guesses.	Le-Hoai et al. (2008), Odeh and Battaineh (2002); Yang and Wei (2010); Assaf et al. (1995)
Linear-scheduling model (LSM)	Proposed by Peer and Selinger in 1970s for analysing factors impacting construction time in repetitive building projects.	Handles few activities. It's usually executed along a linear path/space, Hard sequence logic.	Visualization features, ease of communication for specific type of projects	LSM is inefficient when scheduling complex discrete projects (i.e. bridges, buildings, etc.), weak analytical capabilities.	Van Staveren (2006), Fookes et al., (1985), Kangari (1995), Sanger and Sayles (1979)
Stochastic Critical-Path Envelope Method	Proposed by Kokkaew, N and Chiara, N (2010).	Uses simple monte Carlo simulations to randomly generate project activity durations that will later utilise CPM approach to determine project duration.	Generates a probability distribution of project duration and criticality index of project activities. Criticality index shows activity that is likely to cause delay	Lacks capacity to examine large project samples. Cannot not serve as a benchmarking tool for multiple projects.	Ng and Loosemore (2007); Shen et al. (2007); Tam and Fung (2008)
Benchmarking	Many company's In-house method of analysing completion risk	Uses completion time for similar projects to define and arrive at maximum delay time for project	Simply relies on large samples of historical data	It relies on historical data and benchmark figures that have no predictive value when considering new, large and complex projects	Chan, and Kumaraswamy (2002), Yeung et al., (2007), Bossink (2004).

BDA functions for regression of form $\boldsymbol{f}(\boldsymbol{X}) = \boldsymbol{\epsilon}(\boldsymbol{Y}|\boldsymbol{x})$ tend to minimize **RSS** among all functions from $\boldsymbol{X}$ to $\mathbf{Y}$.

This study starts predictive analysis with multivariate regression analysis as the baseline model for inferential statistics. The R function *lm()* is used for model development, with basic syntax as *lm(y ∼ x, data)*, where *y* is response, *x* are predictors, and data is dataset containing *x* and *y*. The *summary()* function retrieves the details of linear model. For attribute importance, *p*-values near the zero are used to identify predictors with superior predictive performance. The *predict()* function is used to check for test error. Predicted values are plotted to visually inspect variations in predictions. Listing 1 shows R code used to perform regression analysis in this study.

2) Regression Trees: Tree-based models can be used for regression as well as classification problems. Regression trees divide the predictor space $(\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \ldots, \boldsymbol{X_p})$ into a set of nonoverlapping J distinct regions $(\boldsymbol{R_1}, \boldsymbol{R_2}, \boldsymbol{R_3}, \ldots, \boldsymbol{R_j})$. A regression tree follows splitting rules, starting at the root and divide down the tree into smaller subsets at each split. A regression tree comprises nonleaf and leaf nodes. Nonleaf nodes are the decision paths to be followed whereas leaf nodes contain decision values. Regions in regression tree are constructed as shapes like boxes or rectangles. Regression tree algorithm tries to find the boxes (regions) that minimize the RSS, given by the following equation:

$$\boldsymbol{RSS} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{J}} \sum_{\boldsymbol{k \epsilon R_j}} \left(\mathbf{y_i} - \hat{y}_{\boldsymbol{R_j}}\right)^2 \tag{6}$$

where $\hat{y}_{\boldsymbol{R_j}}$ is the average value of response in the *j*th box. Since construction of all possible boxes for a tree is computationally infeasible, greedy algorithms such as recursive binary splitting are used to construct trees in a reasonable computation and time. During recursive binary splitting, every predictor $\boldsymbol{X_j}$ is selected and a cut *s* is defined that divides predictor space into regions, yielding greatest reduction in RSS. Finally, predictor $\boldsymbol{X_j}$ and cut point is chosen for split among predictors $(\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \ldots, \text{ and } \boldsymbol{X_p})$ that has the lowest RSS. The same process repeats for successive splits. This process of tree construction continues until stopping condition is arrived or no regions contain more than five data points. Once regions $(\boldsymbol{R_1}, \boldsymbol{R_2}, \boldsymbol{R_3}, \ldots, \text{ and } \boldsymbol{R_j})$ are defined, predictions are made for incoming data by simply using the median or mode of data in the region to which new data belong. Regression trees are simplistic, easier to interpret, and have nice graphical representation.

Complexity of regression trees bears significant impact on their predictive power. The deeper the tree, the more likely for it to overfit test data; hence, poor predictive performance. To this end, approaches like pruning regression trees comes in play, where larger tree is grown and is pruned back to obtain an optimal subtree. This reduction is achieved through cost complexity pruning (cp), also called as weakest link pruning. The cp considers subtrees, index by nonnegative parameter $\boldsymbol{\alpha}$. When $\boldsymbol{\alpha} = 0$, tree is deepest and complex. But as $\boldsymbol{\alpha}$ starts increasing, trees with more nodes pay more prices; hence, complexity starts decreasing. So as $\boldsymbol{\alpha}$ increases from 0, branches get pruned. Cost validation is often employed to obtain an optimal value of $\boldsymbol{\alpha}$ in regression analysis.

In this study, recursive partitioning and regression tree (rpart) library in R is used to fit regression tree model. The size of the tree is decided by cp, which is enforced via cross validation.

```
#Creating regression model & checking the sum of squared error for predictions
linearModel <- lm(DELAY ~ .-PROJECT, data = trainPPP)
summary(linearModel)
plot(linearModel)
linearPredictions <- predict(linearModel, newdata = testPPP)
linearPredictionsDF <- data.frame(pid = testPPP$PROJECT, pred_delay= linearPredictions,
ml_func="lm")
linearRSS <- sum((linearPredictions - testPPP$DELAY)^2)
rssTB <- data.frame(ml_func = "lm()", rss = linearRSS)
```

Listing 1. R code for creating and evaluating regression analysis using lm() function.

```
#Cross validating the decision trees
tr.control <- trainControl(method="cv", number=10)
cp.grid <- expand.grid(.cp = (0:10)*0.001)
trainTreeModel <- train(DELAY ~ .-PROJECT, data = trainPPP, method="rpart",
        trControl=tr.control, tuneGrid = cp.grid)
trainTreePredictions <- predict(trainTreeModel, newdata = testPPP)
trainTreePredictionsDF <- data.frame(pid = testPPP$PROJECT, pred_delay=
trainTreePredictions, ml_func="train")
trainTreeRSS <- sum((trainTreePredictions - testPPP$DELAY)^2)
rssTB <- rbind(rssTB, data.frame(ml_func = "train()", rss = trainTreeRSS))
```

Listing 2. R code for creating and evaluating regression analysis using rpart() function.

```
#Building the random forest of trees for predicting risk
forestModel <- randomForest(DELAY ~ .-PROJECT, data = trainPPP, mtry=4,
importance=TRUE, ntree = 500)
summary(forestModel)
plot(forestModel)
importance(forestModel)
varImpPlot(forestModel)
forestPredictions <- predict(forestModel, newdata = testPPP)
forestPredictionsDF <- data.frame(pid = testPPP$PROJECT, pred_delay = forestPredictions,
ml_func="randomForest")
forestRSS <- sum((forestPredictions - testPPP$DELAY)^2)
rssTB <- rbind(rssTB, data.frame(ml_func = "randomForest()", rss = forestRSS))
```

Listing 3. R code for creating and evaluating regression analysis using randomForest() function.

Regression tree is generated accordingly using *train()* function for different cp values. The tree model is used to check for test error using *predict()* function. Predicted values are plotted to visually inspect variations in predictions. Listing 2 shows R code used to achieve these steps in RStudio.

3) Random Forest: Regression trees are generally not robust. A small change in data can result in a large change in the model. Nonparametric approaches such as bagging, boosting, and RF are mostly used to overcome these limitations. We limit our discussions to RF only. RF improves the performance of regression trees by compromising interpretability, i.e., by growing many trees $\hat{\boldsymbol{f}}^{\boldsymbol{1}}(\boldsymbol{x}), \hat{\boldsymbol{f}}^{\boldsymbol{2}}(\boldsymbol{x}), \hat{\boldsymbol{f}}^{\boldsymbol{3}}(\boldsymbol{x}), \ldots, \hat{\boldsymbol{f}}^{\boldsymbol{B}}(\boldsymbol{x})$, and then using average of predictions to obtain low-variance regression model, given by the following:

$$\hat{\mathrm{f}}_{\boldsymbol{avg}}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{\boldsymbol{b}}(\boldsymbol{x}) \tag{7}$$

where B denotes the number of trees. RF grows tree by considering a subset m out of ρ predictors. The rule of thumb is to choose $\boldsymbol{m} \approx \sqrt{\boldsymbol{p}}$ predictors. RF with small m favors scenarios, with many correlated predictors.

In this study, we employed RF to see if they improve predictive performance by growing 500 trees. We used *randomForest()* function to grow trees on training data set. The RF model is used to check for test error using *predict()* function. Predicted values are plotted to visually inspect variations in predictions. Listing 3 shows R code used to model development and evaluation.

4) Support Vector Machine (SVM): SVM is an machine learning (ML) algorithm with robust regularization capabilities to generalize to the unseen data with a high degree of accuracy. SVM models can be used for both classification and regression analysis to solve complex and real-world problems. SVM outperforms on data with many attributes even if there are a small number of training examples.

```
svmFormula <- as.formula("DELAY ~ SECTOR + CONTRACT + NOD + FIMP +
                          POCI + PODV + PSSL + IMSS + NUSC + PMDS +
                          PDMD + NSAI + NDSC + PLAD + NDBW + NODP")

svmModel <- ore.odmSVM(svmFormula,data=trainPPP, "regression", kernel.function="gaussian")
svmPredictions <- predict(svmModel, testPPP[,c(1:16)], supplemental.cols="x")
svmPredictionsDF <- data.frame(pid = testPPP$PROJECT, pred_delay =
svmPredictions$PREDICTION, ml_func="SVM")
svmPredictionsRSS <- sum((svmPredictions$PREDICTION - testPPP$DELAY)^2)
rssTB <- rbind(rssTB, data.frame(ml_func = "odmSVM()", rss = svmPredictionsRSS))
```

Listing 4. R code for creating and evaluating regression analysis using odmSVM() function.

SVM works on a kernel function that transforms input data into a high-dimensional space and then finds the optimal solution to the problem. The kernel functions can be linear as well as Gaussian. Linear kernels translate to linear equations and suits multi-attribute training data. The Gaussian kernels convert training data into points in n-dimensional space and construct numerous linear equations using nonlinear boundaries within the kernel space.

SVM uses epsilon-intensive loss function for regression analysis. The algorithm works by finding a function where more data points lie inside the epsilon-wide insensitivity tube. The epsilon can be customized through SVM settings. SVM balances the margin of error with model robustness to achieve best generalization for the unseen data.

We used *ore.odmSVM()* to develop SVM model for regression analysis in this study. Automatic data preparation capabilities of ORE are used for one-hot encoding of categorical variables. The model is trained on training data and evaluated using test data utilizing the *ore.predict()* function. The predicted values are plotted in figures to inspect variations in predictions. Listing 4 shows R code for performing these steps.

5) Deep Neural Network: Working like a brain, DNN is leading among nonlinear regression techniques ([64], Huang *et al*., 2016). In DNN, response is modeled as a set of intermediate hidden layers that are the linear combination of predictors. DNN employs two obviously different transformations. First, nonlinear function $\boldsymbol{g}(.)$ such as sigmoidal is used for eliciting the nonlinearity of predictors, which is explained by the following equation:

$$\boldsymbol{h_k} = \boldsymbol{g}\left(\boldsymbol{\beta_{0k}} + \sum_{i=1}^{\boldsymbol{p}} \boldsymbol{x}_i \boldsymbol{\beta_{jk}}\right) \tag{8}$$

where $\boldsymbol{\beta}$ coefficients are similar to that of ordinary linear regression and $\boldsymbol{\beta_{jk}}$ is the effect of the jth predictor on k hidden layer. Second, linear transformation is applied to convert outcome back to actual values, using the following equation:

$$f(\boldsymbol{X}) = \boldsymbol{\gamma_0} + \left(\sum_{\boldsymbol{k}=1}^{\boldsymbol{H}} \boldsymbol{\gamma_k} \boldsymbol{h_k}\right). \tag{9}$$

DNN requires parameter optimization to reduce the sum of squared error. To this end, specialized numerical optimization algorithms such as back-propagation [64] are used. DNN overfits mostly the relationship between predictors and response due to large coefficients, which is combatted through prematurely stopping algorithm or by using penalization techniques like weight decay. DNN tries to minimize RSS for the given value of $\boldsymbol{\lambda}$ using the following equation:

$$\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} (\boldsymbol{y_i} - \boldsymbol{f_i}(\boldsymbol{x}))^2 + \boldsymbol{\lambda} \sum_{\boldsymbol{k}=1}^{\boldsymbol{H}} \sum_{\boldsymbol{j}=0}^{\boldsymbol{P}} \boldsymbol{\beta}_{jk}^2 + \boldsymbol{\lambda} \sum_{\boldsymbol{k}=0}^{\boldsymbol{H}} \boldsymbol{\gamma}_{\boldsymbol{k}}^2. \tag{10}$$

This makes model smoother and less susceptible to overfitting. Another challenge of employing DNN in regression analysis is adverse correlation effect, which is either circumvented manually or by using techniques for feature extraction like principal component analysis.

We employed *neuralnetwork()* library in R to develop DNN model. Using *caret()* function, hyperparameter tuning for decay size of DNN is calculated and accordingly model is developed. The DNN model is used to check for test error using *compute()* function. Predicted values are plotted to visually inspect variations in predictions. Listing 5 shows R code used for model development and evaluation.

III. Defining Key Predictors for Completion Risk Analysis Using Predictive Modeling

In order to demonstrate BDA for completion risk forecasting, data of PPP projects between 1992 and 2015 were obtained from database of the European PPP Expertise Centre, Monthly Statistics of Construction Building Materials and Components from U.K.'s Department of Business Innovation and Skills, U.K.'s Construction Industry Data, Health and Safety in Construction Sector Report of U.K., U.K.'s Office of the National Statistics, European Construction Market data (Euro Area Construction data), etc. Sixteen (16) key predictors causing time overrun in projects were used for the predictive modeling of completion risk. These factors were specifically chosen due to ability to

```
#Creating the DNN model
annFormula <- as.formula("DELAY ~ SECTOR + CONTRACT + NOD +
                                  FIMP + POCI + PODV + PSSL +
                                  IMSS + NUSC + PMDS + PDMD +
                                  NSAI + NDSC + PLAD + NDBW + NODP")
annModel <- neuralnet(annFormula, data=trainPPP, hidden=c(10,5),linear.output=T)
annPredictions <- compute(annModel, testPPP[,c(1:16)])
annPredictionsDF <- data.frame(pid = testPPP$PROJECT, pred_delay =
annPredictions$net.result, ml_func="ANN")
annPredictionsRSS <- sum((annPredictions$net.result - testPPP$DELAY)^2)
rssTB <- rbind(rssTB, data.frame(ml_func = "neuralnet()", rss = annPredictionsRSS))
```

Listing 5. R code for creating and evaluating regression analysis using neuralnetwork () function.

TABLE II
KEY PREDICTORS INFLUENCING COMPLETION RISK (DELAY) IN PPP PROJECTS

Values	Key Predictors Influencing Completion	Sources
SECTOR	Projects chosen cut across nine (9) sectors	HM Treasury (2014), NAO (2009)
CONTRACT	Projects were either procured via turnkey	PartnershipsUK.org.uk
NOD	Av. No of defects in a construction project	Buchholz (2004); Teizer et al. (2010);
FIMP	% fluctuation in construction material	Javed et al. (2013); Tam et al. (2004)
POCI	% change in inflation	Ahmed et al. (1999); El-Sayegh (2008).
PODV	% of design variations	Kangari (1995); Bossink (2004); Tatum
PSSL	% shortage in skilled labor	Tatum (1989); Bossink (2004); Tatum
IMSS	% of inferior materials supplied to site	Odeh and Battaineh (2002); Errasti et al.,
NUSC	No of unforeseen site conditions	Dikmen et al., (2007); Flyvbjerg et al., (2004)
PMDS	% of materials damaged on site	Ching (2014); Allen and Iano (2011)
PDMD	% Delay in Material delivery	Robinson and Scott (2009); Javed et al.
NSAI	No of site Accidents and injuries	Rousseau and Libuser (1997); Shen et al.
NDSC	No of days for site closure	Kaming et al., (1997); Moselhi et al., (1997)
PLAD	% of liquidated and ascertained damages	Mohamed (2002); Tam et al. (2004);Tatum
NDBW	No of days with bad weather that	Tatum (1987); Harty (2005); Tatum (1989)
NODP	Av. No of disputes among parties	El-Sayegh (2008); Russell and Jaselskis
DELAY	Delay in terms of days	Shen et al. (2007); Tam and Fung (2008)

quantify them and their potential impact on delay in construction project delivery (Kokkaew and Chiara, 2010, [32]). The factors are articulated in Table II below.

1) *Sector:* The PPP projects selected for the study cuts across nine sectors, namely housing, social care, transport, defense, education, health, waste management, public buildings, and others (comprising comprises prisons, leisure facilities, offices, housing, emergency services, courts, etc.).
2) *Contract Type:* The two principal contract types adopted in all the projects analyzed are fixed price turnkey and Design Bid Build. Fixed price turnkey ensures a contractor delivers project under a lump sum contract, while accepting completion risk (Hoffman, 2008). On the other hand, Design Bid Build, which is also known as the traditional procurement approach allows a client to contract separate parties for design and construction phases of the project [15].
3) *Average Number of defects in a construction project:* Defects in project delivery is a perennial challenge in the global construction industry. According to El-Sayegh [32], defects in construction project contribute significantly to construction delay. This could happen as a result of defects in project design or defects due to poor communication between the design managers and the contractors [117].
4) *Percentage (%) fluctuation in construction material price index:* This is often a major concern for contractors as material price fluctuation upsets prior financial forecasts and impacts project timeline, especially where contractor

has no parent company cover to bail it out in the event of financial difficulties [49].

5) *Percentage (%) change in inflation:* Similar to fluctuation in construction material price index, sudden upsurge in general inflation portends great danger to construction budget, which may result in inability to achieve critical milestones on a project [9], [80].
6) *Percentage (%) change in design variation:* Changes in project design is also a common occurrence in construction project and is mostly initiated by the client. However, studies such as [101] and [103] have argued that frequent changes in design, especially critical components of a project have direct impact on timely completion.
7) *Percentage (%) shortage in skilled labor:* The direct consequence of not having the right number of skilled manpower to deliver a project is excessive delays in achieving project completion [8].
8) *Percentage (%) of inferior materials supplied to site:* Supply chain is crucial to successful project completion and so is the quality of construction materials supplied to site [37]. Delays due to discovery of low quality materials supplied to site are not unusual and this may cause serious lag in project schedule [51].
9) *No. of unforeseen site conditions:* These can cause project delay as contractors have to confront site conditions (i.e., topography or underground conditions) not contemplated during the initial construction survey.
10) *Percentage (%) of materials damaged on site:* Kangari (1995) and [22] listed material damage on project site as one of the causes of construction time overrun. Such situations impact both project schedule and construction budget and may pose danger to the project [22].
11) *Percentage (%) delay in material delivery:* The danger of not having a reliable supply chain is unwarranted disruption in project schedule [83]. The impact of supply chain delay on a project may be viewed in terms of the percentage of construction duration that is lost to delay in material delivery.
12) *Number of site accidents and injuries:* This can be expressed in terms of man hour loss or site closure due to accidents and its impact on project schedule [58].
13) *Number of days for site closure:* This has an impact on the project timeline and does not include estimated closure due to bad weather. Site closures may occur due to industrial action by construction workers, force majeure, and closure due to potential danger to the public, etc. [36].
14) *Percentage (%) of liquidated and ascertained damages in projects:* Liquidated damages are financial penalties levied on contractor for breach of contractual obligations [45]. This has negative implications for timely delivery of a project, especially where such levy is huge enough to result in financial difficulties that prevents contractor from meeting their obligations to subcontractors [22].
15) *Number of days with bad weather that prevented site work:* Many a times, protracted and unpredictable weather conditions (high velocity wind, flood, etc.) may prevent a project from being completed on time [37].
16) *Number of disputes among parties:* This may be in form of litigation or demand for contractual settlements and is a major factor which often results in project delay (Kangari, 1995). According to Teizer *et al*. in [106], the frequency of disputed issues on a project has negative implications for timely completion.

In this study, our goal is to develop an accurate model that can be used to estimate completion risk (project delay). In order to achieve this, we assumed a linear relationship between CR and the predictors (p). The predictors (p) are thus considered as input variables $(X_{1,}X_{2,}X_3 \ldots\ldots\ldots\ldots\ldots..X_\rho)$, thereby establishing a directly proportional relationship between CR as $X = (X_{1,}X_{2,}X_3 \cdots X_\rho)$. In other to achieve this, a linear model is thus developed and formally written as follows:

$$CR = f(X) + \epsilon \tag{11}$$

where f is a fixed unknown function of $X1$, $X2 \ldots Xp$ and ϵ represents the random error term, which is independent of X and has a mean of zero. In the equation above, *f(X)* provides systematic information about the delay in PPP projects, and could be expanded to the following equation involving multiple variables to describe this relationship:

$$\begin{aligned} f(\mathrm{DELAY}) &= \beta_0 + \beta_1 \times \mathrm{NOD} + \beta_2 \times \mathrm{FIMP} + \beta_3 \times \mathrm{POCI} \\ &+ \beta_4 \times \mathrm{PODV} + \beta_5 \times \mathrm{PSSL} + \beta_6 \times \mathrm{IMSS} + \beta_7 \times \mathrm{NUSC} \\ &+ \beta_8 \times \mathrm{PMDS} + \beta_9 \times \mathrm{PDMD} + \beta_{10} \times \mathrm{NSAI} + \beta_{11} \times \mathrm{NDSC} \\ &+ \beta_{12} \times \mathrm{PLAD} + \beta_{13} \times \mathrm{NDBW} + \beta_{14} \times \mathrm{NODP} \end{aligned} \tag{12}$$

where β_i is the coefficient that will be estimated, where $i = 0, 1, 2, \ldots, p$ employing BDA from the large array of data from PPP Project samples.

IV. Research Methodology

This section explains the methodology employed in the study. After understanding the domain of CR in PPP projects, relevant data sources were identified to explore the most critical factors that lead to delay in PPP projects. The methodology steps have been described in detail under subsequent sections and shown in Fig. 1 below.

A. Databases

The predictive accuracy of the Big Data models depends on the quality and volume of PPP projects. Data of 4731 PPP projects were integrated from a large number of structured and unstructured data sources. The data were distributed in a large number of data sources. These include Oracle financials, BIM models, Primavera, Candy, Health & safety, Business objects, Customer relationship management, and a large body of unstructured documents. These sources were explored to identify relevant data, structures, and formats to enable the database design. Fig. 2 shows types and sources of data of PPP projects

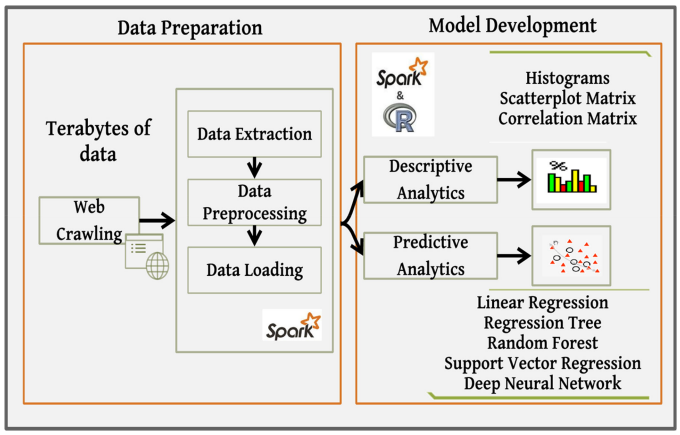


Fig. 1. Big Data Analytics workflow for predictive risk modeling.

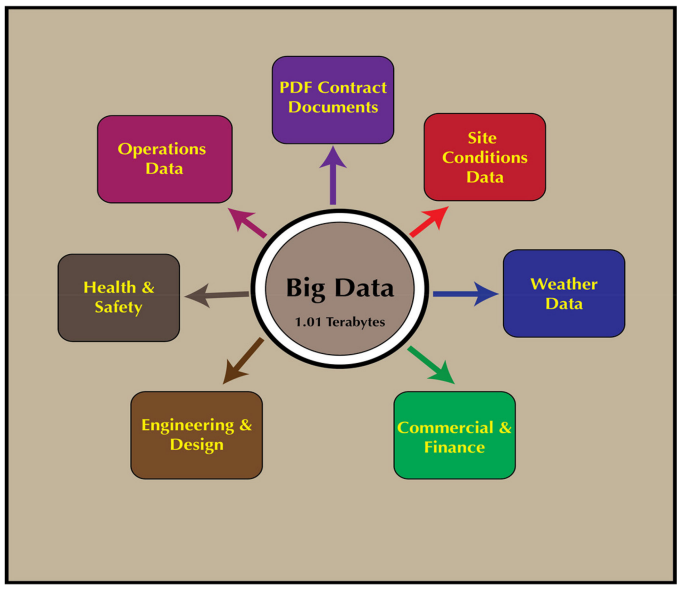


Fig. 2. Overview of Big Data of PPP Projects.

used in the study. This effort has resulted in the exploration of 1.01 terabytes of data for analysis. This data fulfills all 3V's of the Big Data that is volume, variety, and velocity.

B. Data Preprocessing and Integration

Data integration task is found the toughest in the overall risk analytics experience. A variety of syntactical and semantic heterogeneities were resolved (Halevy *et al.*, 2005, Doan & Noy, 2004). To ensure data completeness, ML programs were used to predict missing values for predictors like average defects (Bishop, 2006, Goldberg & Holland, 1988). Data were standardized with vocabularies for construction sectors and contract types. Automatic conversion is augmented to deal with inappropriate interpretations especially for date columns. The data normalization is carried out by formula given in the following equation:

$$X_i' = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \tag{13}$$

where X_i' is the scaled result of X_i, $X_{\min}$ is the smallest value of X, and $X_{\max}$ is the largest value of X. The final data analytic sample is restricted to 4294 PPP projects, which are eventually loaded onto *Apache Spark*—a resilient cluster computer engine for BDA. Table III shows distribution of projects across sectors and contract types. *SparkR* is used for data analysis and R *ggolot2* package is used for visualization.

C. Descriptive Analytics

We started with exploratory analysis to develop better understanding of the overall PPP projects data. Descriptive analytics is applied to describe main features of the dataset. Important facts are elaborated to get initial impressions of data. Numerical summaries and graphical methods are used. Histograms, boxplots, and scatterplots are drawn to see the fitness of data for predictive modeling.

D. Predictive Analytics

Descriptive analytics sets the stage for more flexible predictive analysis, where a series of predictive models were developed using various BDA techniques and evaluated for their predictive performance. The data are split across training and test sets using *sample()* function. We initially developed multivariate linear regression model to understand the interactions of predictors on response. This model is treated as the baseline model. To improve upon the predictive performance of linear model, regression trees were employed. We found different behavior of delays across different sectors and contract types, which are not fully described by the linear regression model.

Though regression trees describe nonlinearity to some extent and are highly interpretable, but they are not robust; a slight change in the data can result in a totally variant tree. To overcome these limitations in predictive modeling, we employed RF to see if they improve the predictive performance by growing 500 trees. SVM was also employed to ensure good classification of the data sample. Finally, we brought the deep learning-based predictive modeling technique called DNNs. DNN is a black box approach that knows how to process predictors to obtain more accurate matching response. For each model, hyperparameter tuning is performed and approaches like cross validation was employed to devise a robust model development strategy. These models were plotted using R *gglot2* library for evaluating their performance in terms of decreasing the test error. It is shown that RF is very robust and viable option to employ for estimating the CR in the PPP projects.

E. Attribute Importance and Ranking

Since these models employ different model development strategies, they ranked the attributes differently. To aggregate these ranking, a reliable total ranking scheme is devised. The scheme used p-value, Gini, impurity, ranked agreement factor (RAF), and percentage ranked agreement factors (PRAF) for ranking predictors for the CR prediction.

TABLE III
DATA ANALYTIC SAMPLE OF PPP PROJECTS USED FOR BIG DATA ANALYTICS

Sr.#.	*Sector*	*Contract Type*	*Number of Projects*
1	Housing	Fixed Price Turnkey (FPTK)	200
2	Housing	Design-Bid-Build (DBB)	261
3	Social Care	Fixed Price Turnkey (FPTK)	227
4	Social Care	Design-Bid-Build (DBB)	250
5	Transport	Fixed Price Turnkey (FPTK)	233
6	Transport	Design-Bid-Build (DBB)	253
7	Defence	Fixed Price Turnkey (FPTK)	243
8	Defence	Design-Bid-Build (DBB)	249
9	Education	Fixed Price Turnkey (FPTK)	219
10	Education	Design-Bid-Build (DBB)	266
11	Health	Fixed Price Turnkey (FPTK)	190
12	Health	Design-Bid-Build (DBB)	251
13	Waste Management	Fixed Price Turnkey (FPTK)	238
14	Waste Management	Design-Bid-Build (DBB)	261
15	Public Buildings	Fixed Price Turnkey (FPTK)	225
16	Public Buildings	Design-Bid-Build (DBB)	260
17	Others	Fixed Price Turnkey (FPTK)	232
18	Others	Design-Bid-Build (DBB)	236
Total Data Analytic Sample:			**4294**

V. ANALYSIS AND FINDINGS

A. Big Data Descriptive Analytics

We started with exploratory analysis to develop better understanding of the overall PPP project data. Descriptive analytics is the kind of first hand analysis applied to describe main features of the dataset. Important facts are elaborated to get initial impressions of data. Numerical summaries and graphical methods are often rampant. To showcase the analysis, correlation matrix plot is discussed here. Covariance test is performed to investigate multicollinearity among the 16 predictors in the dataset. In probability statistics and theory, covariance help describe the degree to which set of random variables deviate from their expected values [75]. According to Casella *et al.* (2013), positive covariance indicates positive linear relationship, whereas negative values mean negative linear relationship. Covariance is calculated by the following equation and color coded in Fig. 3:

$$\operatorname{cov}(X,Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x}_i)(y_i - \bar{y}_i)}{n}. \tag{14}$$

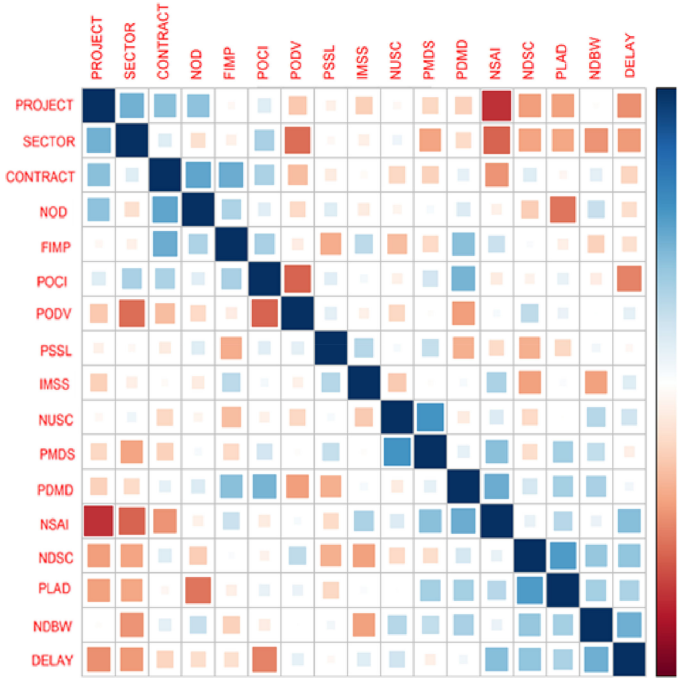


Fig. 3. Correlation plot depicts covariance between variables in PPP projects.

As shown in Fig. 3, the bright brown slots represent the positive linear relationships, whereas the blue slots depict the negative linear relationship. In addition, strong brighter colors represent the strong relationship between the variables, whereas the faded colored regions represent independent variables. It is notable from the graph that response variable (project delay) has strong relationship with most of the variables, which is a very good indicator for considering these variables in predictive modeling. However, some variables have strong covariance, like number of days with bad weather NDBW and unforeseen site condition (UNSC). This shows collinearity issue between

these variables and informs that these variables tend to add similar predictive capabilities twice. As a result, we dropped NDBW for UNSC to reduce the complexity of the model in order to achieve higher predictive performance.

B. Big Data Analytics for Estimating CR in PPP Projects

In the remainder of this paper, we discuss the development of predictive models for CR estimation. Since a single model might not be able to entirely capture the true relationship of different KPIs selected in this study with respect to delays in PPP projects, a mix of linear as well as nonlinear BDA techniques are employed during model development. These techniques have really moved our understanding of CR to the next level. In addition, a robust CR estimation model is developed for assessing delays in the future PPP projects. Subsequent sections provide more details of these models and their comparisons.

C. Multivariate Linear Regression

An important reason behind starting with linear regression is to understand the way delay in PPP projects are influenced by myriad factors. In this case, we estimated f not for the purpose of predicting CR in PPP projects. Instead, the objective is to understand the relationship between the predictor ρ and response $\boldsymbol{Y}$ or more specifically to know how $\boldsymbol{Y}$ changes as a function of ρ. So, $\hat{f}$ is not treated as a black box rather an elaborate description of its exact form. Listing 5 shows the summary of linear regression model.

As mentioned earlier, sector and contract type are categorical variables, dummy variables are created automatically for each of their elements. The intercept term ($\beta_0 = 4.028$) is implicitly added to the model. Generally, intercept term β_0 is the expected delay when all predictors equal to zero. Currently, the sector attribute contains $0 =$ hospital, $1 =$ school, $2 =$ public building, $3 =$ transportation, $4 =$ housing, $5 =$ social care, $6 =$ defense, $7 =$ waste, and $8 =$ others. The model will mislead if it is applied to dataset that contains sectors that are not representative within the training dataset. The same applies to the contract types as well. Interestingly, the model does not describe the relationship of sectors to delays, which is reported by higher p-values (0.49164, 0.35714, 0.95755, 0.33270, 0.68388, 0.47565, 0.85810, and 0.32757) of all sectors, respectively. In contrast, contract type has virtually zero p-value (0.00424), which indicates strong correlation in predicting delays. The implication of this is that delay in project varies based on contract type.

The parameter estimation is computed using ordinary least squares. The *Estimate* column shows parameter estimation for predictors and *Std. Error* displays the standard error associated with each of these coefficients. This is used for hypothesis testing, using t-distribution column *t value*, to determine if each coefficient is not statistically different from zero. And if so, then the predictor is removed from the model. Analysis shows that the associated hypothesis test p-value in Pr ($<|t|$) values are small for intercept term, contract type, number of defects (NOD), % of fluctuation in materials price (FIMP), number of unforeseen site conditions (NUSC), % materials damage (PMDS), % delay in materials delivery (PDMD), number of site injuries (NSAI), number of days bad weather (NDBW), and number of disputes among parties (NODP). Whereas the rest of the attributes are removed from the model since they have no significance in predicting delay in PPP projects. A small p-value corresponds to small probability that such a large t value would be observed under the assumption of null hypothesis. In this case, for a given $I = 0, 1, 2, \ldots, p$–1, the null and alternate hypothesis follow:

$$H_0 : \beta_i = 0 \text{ versus } H_A : \beta_i \neq 0.$$

For small p-values, as is the case with above-mentioned predictors, the null hypothesis would be rejected. Whereas for rest of predictors, null hypothesis is not rejected due to large p-values of those predictors. Dropping these columns resulted in minimal changes to the estimates as well as predictive performance of the model. The last part of the summary displays some of the vital details of regression model. Specifically, R^2, which in this context says that the model is capable to explain 69% variation in the data. And the overall p-value i.e., $<$ 2.2e–16 is small, which indicates that the null hypothesis should be rejected.

Fig. 4 shows the line plot for observed and predicted delays estimated by the linear regression, where R^2 is relatively good (69%). However, it is evident that the predictions are not uniformly accurate. To improve upon these, we employed regression trees to capture the nonlinear behavior of predictors on response.

D. Regression Trees

To explain the nonlinearity between the predictors and response variables, regression trees are fitted on the data of the PPP projects. Without hyperparameter tuning, initial regression tree only considered sector variable and ignored the rest of predictors. This is quite misleading and is tackled by appropriately configuring the regression tree for risk estimation. To this end, cross validation and cost complexity pruning parameters are optimized and the regression trees are grown for different cp values. Here, the true power of regression trees comes into play and its effectiveness to uncover nontrivial relationship of predictors could be noticed. Contrary to linear regression, regression tree utilized majority of predictors to develop very strong risk estimation model in the dataset. Similar to regression analysis, contract type is regarded as the most superior predictor in the model; hence, taken as the root of the tree. However, the second significant predictor in regression tree is considered the sector, which is totally ignored by the multivariate regression analysis. Regression tree make decisions at various levels based on the sector. So, in this case, the most complex tree is selected by the cross validation. Fig. 5 shows the line plot for observed and predicted delays for linear regression (with accuracy improved by 79%). It is evident that predictions improved significantly. To improve upon the regression trees, we are employing regression trees to capture the nonlinear behavior of predictors on response (see Fig. 6 for regression tree model).

E. Random Forest

Although the regression tree model developed for CR estimation has improved the test accuracy drastically, it is a nonrobust

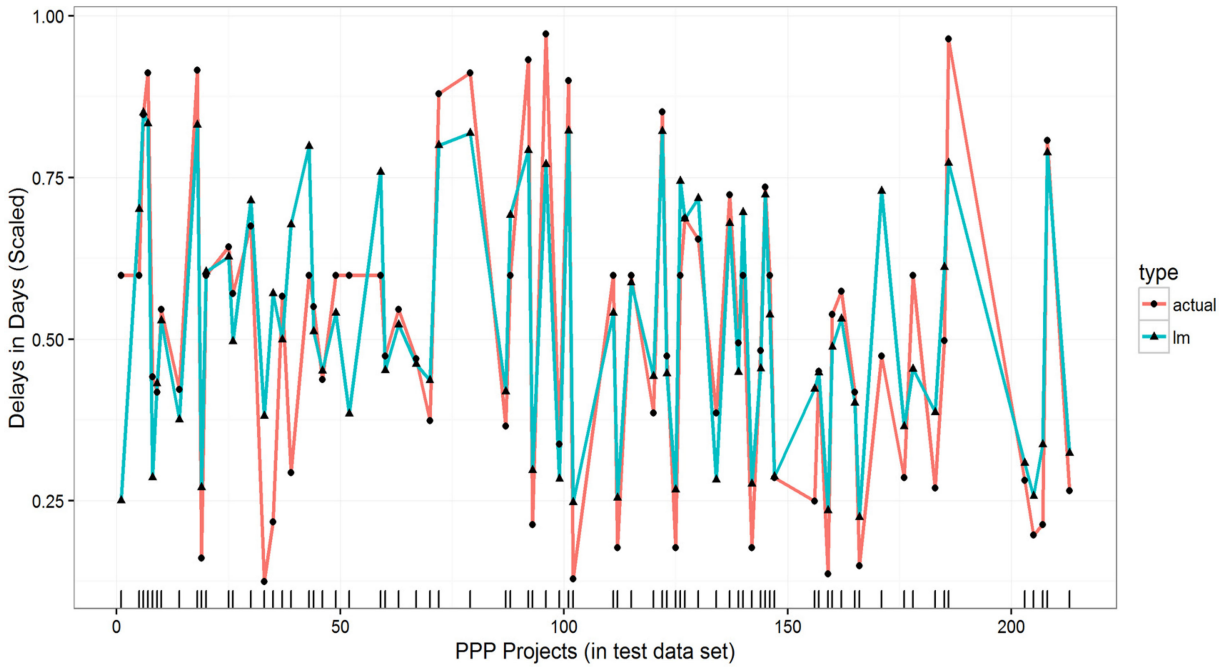


Fig. 4. Evaluating observed and predicted delays in the PPP projects.

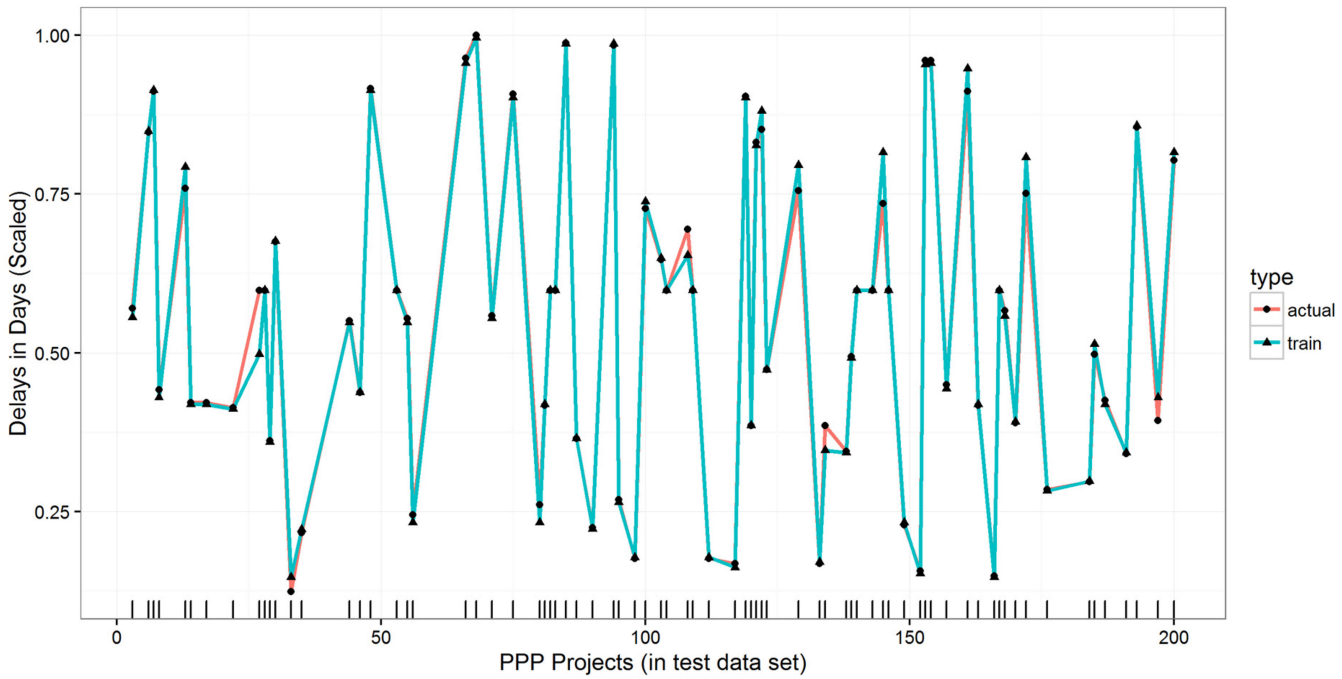


Fig. 5. Evaluating observed and predicted delays in the PPP project.

technique and a slight change in data can yield very different regression trees. Hence, we needed to improve the stability of the model by employing RF. Listing 6 shows the attribute importance summary generated by fitting a RF of 500 trees on PPP projects data, where two measures of importance are populated. The former is based upon the mean decrease of accuracy in predictions on the out of bag samples when a given variable is excluded from the model. The latter is a measure of the total decrease in node impurity that results from splits over that variable, averaged over all trees. In the case of regression trees, the node impurity is measured by the training RSS, and for classification trees by the deviance.

Fig. 7 shows the line plot for observed and predicted delays for linear regression (with accuracy improved by 81%). It is evident that predictions improved dramatically.

F. Support Vector Machine (SVM)

Since SVM has huge adaptability and can generalize to new data with higher accuracy, the SVM algorithm is used to train

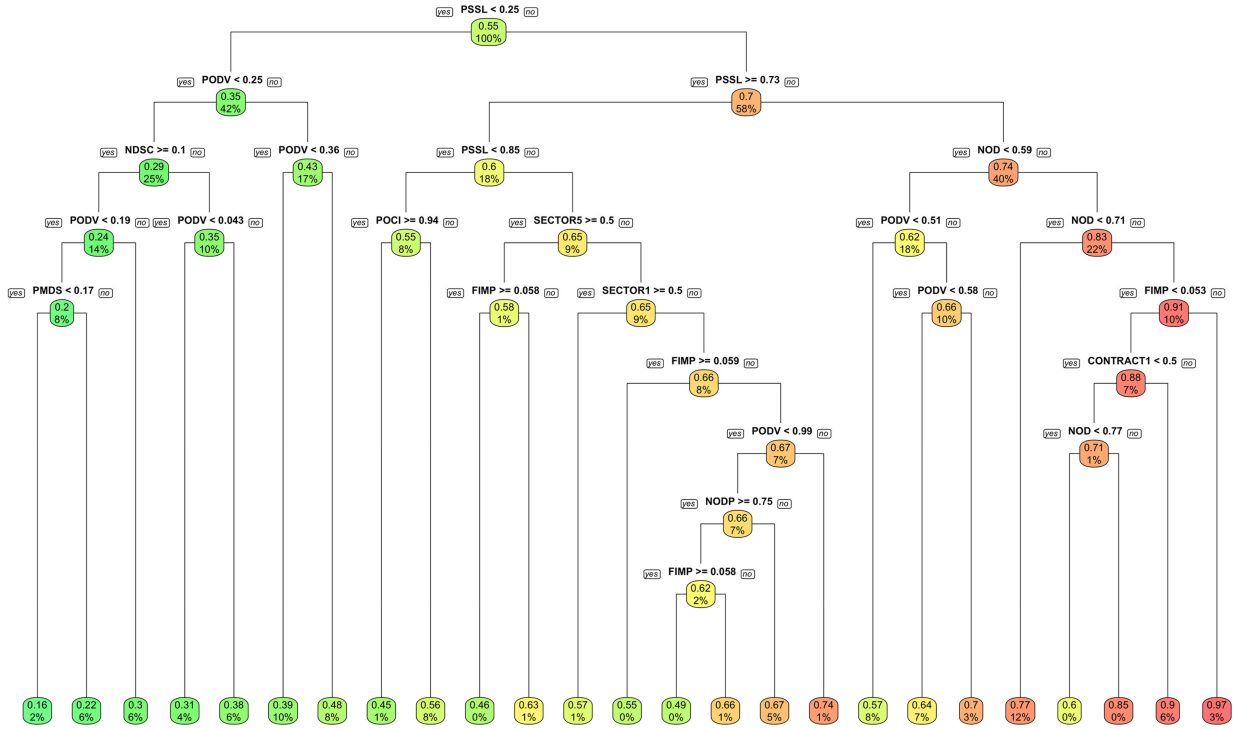


Fig. 6. Regression tree model for predicting delays in the PPP projects.

```
Call:
lm(formula = DELAY ~ ., data = trainPPP)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73954 -0.07309  0.00192  0.05645  0.71047

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3757610  0.0172054  21.840  < 2e-16 ***
SECTOR1      0.0072317  0.0105143   0.688  0.49164
SECTOR2     -0.0096546  0.0104830  -0.921  0.35714
SECTOR3      0.0005468  0.0102727   0.053  0.95755
SECTOR4      0.0100831  0.0104072   0.969  0.33270
SECTOR5     -0.0043815  0.0107594  -0.407  0.68388
SECTOR6     -0.0073937  0.0103638  -0.713  0.47565
SECTOR7      0.0018448  0.0103175   0.179  0.85810
SECTOR8      0.0103248  0.0105442   0.979  0.32757
CONTRACT1    0.0142038  0.0049633   2.862  0.00424 **
NOD          2.4175803  1.1975378   2.019  0.04361 *
FIMP        -0.1712855  0.0804999  -2.128  0.03344 *
POCI         0.0040691  0.0076285   0.533  0.59380
PODV         0.7175872  1.1699697   0.613  0.53970
PSSL        -1.1511576  0.6664203  -1.727  0.08421 .
IMSS        -0.4591695  0.9651310  -0.476  0.63428
NUSC         5.9326532  1.2714082   4.666 3.21e-06 ***
PMDS        -1.9396845  0.8848595  -2.192  0.02846 *
PDMD        -4.8422194  0.5106527  -9.482  < 2e-16 ***
NSAI        -4.1452338  1.2620944  -3.284  0.00103 **
NDSC        -1.0147283  1.0164924  -0.998  0.31824
PLAD        11.3065432  1.3579018   8.326  < 2e-16 ***
NDBW        -6.7006330  0.5704384 -11.746  < 2e-16 ***
NODP         0.0012832  0.0074467   0.172  0.86320
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1283 on 2756 degrees of freedom
Multiple R-squared:  0.6927, Adjusted R-squared:  0.6902
F-statistic: 270.1 on 23 and 2756 DF,  p-value: < 2.2e-16
```

Listing 6. Summary of the fitted multivariate regression model for risk estimation.

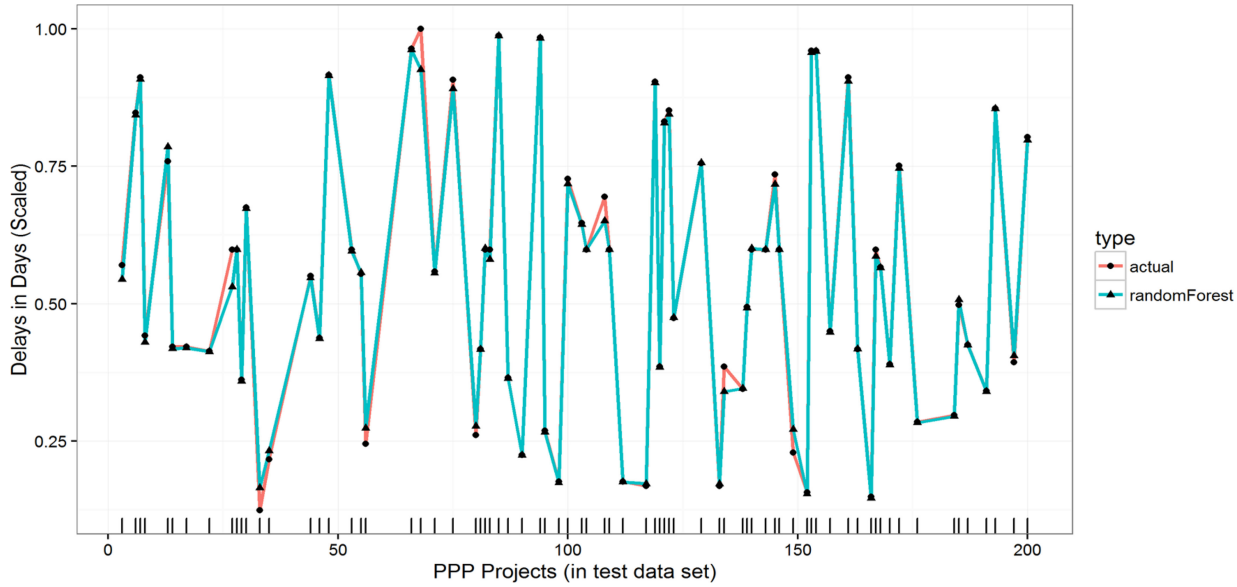


Fig. 7. Evaluating observed and predicted delays in the PPP projects.

```
          %IncMSE IncNodePurity
SECTOR   55.36064     48.883647
CONTRACT 21.08694     13.600641
NOD      12.49698      6.409105
FIMP     11.29658      5.433949
POCI     11.94632      6.791739
PODV     13.71843      6.911212
PSSL     13.60090      8.059655
IMSS      7.80614      3.806822
NUSC     13.55149      6.490616
PMDS     11.16227      3.717878
PDMD     12.70519      8.271368
NSAI     16.43001      7.282481
NDSC     12.01905      5.592730
PLAD     13.95778      7.496080
NDBW     14.17712      6.556476
NODP     12.86448      6.163925
```

Listing 7. Summary of the attribute importance by random forest in risk modeling.

a predictive model to see its prediction capabilities. We started off with SVM for regression analysis using linear kernel, which did not perform very well initially. The error loss was substantial. The Gaussian kernel was used which improved the model accuracy significantly. The algorithm started learning patterns into the data with respect to CRs. For hyperparameter settings such as epsilon, manual approach was adopted at first, and different combinations of values were tested. This approach was cumbersome due to training model for every possible combination. The SVM supported automatic parameter tuning which was then used. This system-generated hyperparameter mode of SVM was found more reliable and efficient since it used advanced optimization algorithms to identify the best values to maximize model accuracy.

SVM solved the problem by defining an n-dimensional tube around the data points to determine the vectors that yield the most extensive intervals. The coefficient vector was extracted from the SVM model to see the importance SVM was giving to each predictor for predicting the delays in PPP projects. Listing 7 above shows the attribute importance summary generated by the trained model using the Monte-Carlo Sensitivity Analysis (M-CSA). The overall accuracy of the model is 52%. Fig. 8, therefore, presents the line plot for observed and predicted delays for SVM, which outperforms the linear regression but could not uplift the predictive accuracy as the tree-based models yielded for predicting the delays in the PPP projects. Although the SVM showed inadequacy in predictive power in this study, the mathematical model underpinning the algorithm suits the classification problem more than the regression analysis.

G. *Deep Neural Network (DNN)*

Finally, to check if the deep learning technique can enhance the predictive performance of the CR estimation model, DNN

	Monte-Carlo Sensitivity
PDMD	0.092
NODP	0.091
NDSC	0.087
PMDS	0.087
PLAD	0.086
FIMP	0.086
NDBW 0.085	
NUSC	0.085
PSSL	0.085
IMSS	0.072
NSAI	0.066
POCI	0.053
NOD	0.010
CONTRACT	0.005
PODV	0.003
SECTOR	0.003

Listing 8. Summary of the attributes importance by SVM in risk modeling.

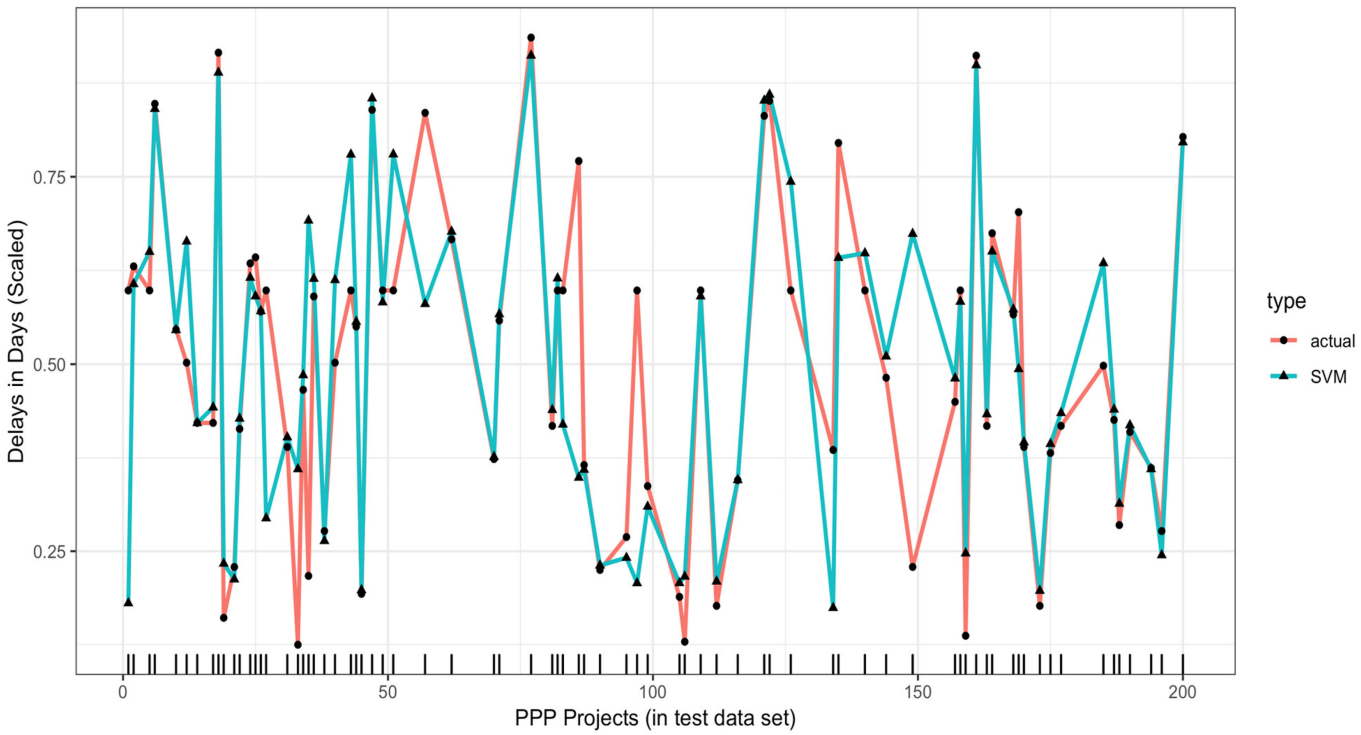


Fig. 8. Evaluating observed and predicted delays in the PPP projects.

is used. Two hidden layers of 10 and 5 nodes, respectively, are defined for the DNN model. The resultant model is shown in Fig. 9. We can see that the model is not interpretable. This is because neural network is a black box methodology to predictive modeling. It is applied in situation where the objective of the research is to make reliable predictions. So all the predictors are taken as input to the neural network. Nonlinear sigmoidal transformation is done on predictors and the weights of the hidden layers are computed. These weights are eventually converted back to the linear transformation. Fig. 10 shows the line plot for observed and predicted delays for linear regression (with accuracy improved by 13%). It is evident that the predictions look very bad. This is partly due to the fact that DNN suits classification problems more than regression problems.

H. Comparison of Big Data Analytics Techniques

1) Comparison Based on RSS: In this section, we set out to compare the five predictive models employed in the study using two major comparison indicators: residual sum square

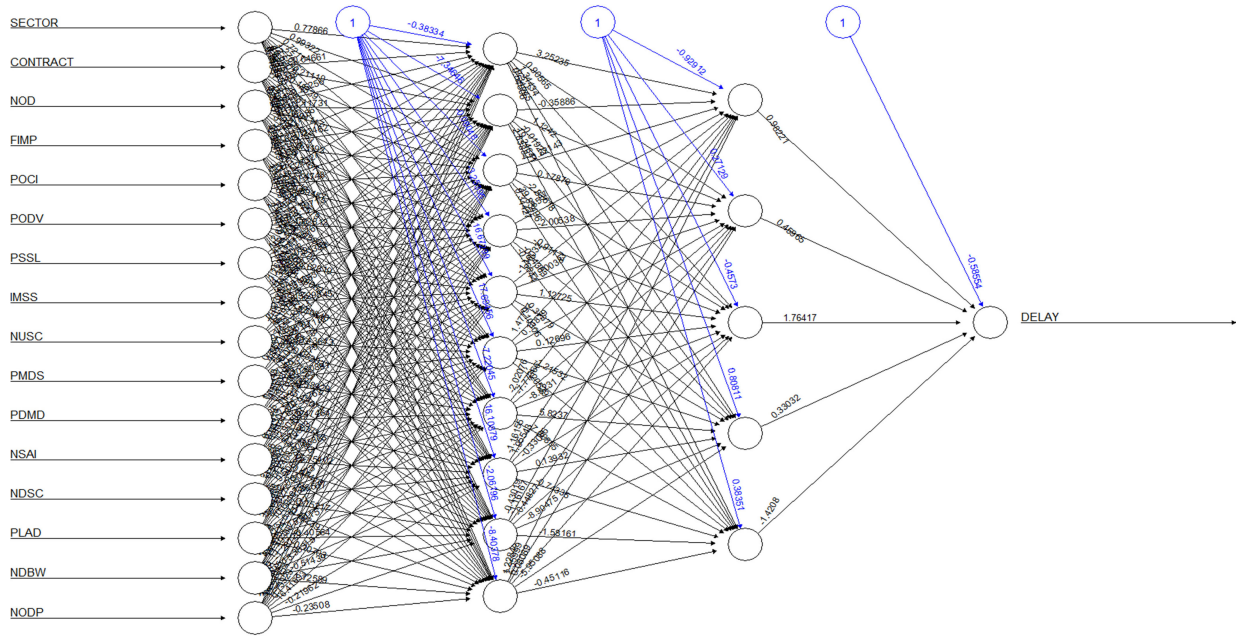


Fig. 9. Deep neural network model for forecasting delays in the PPP projects.

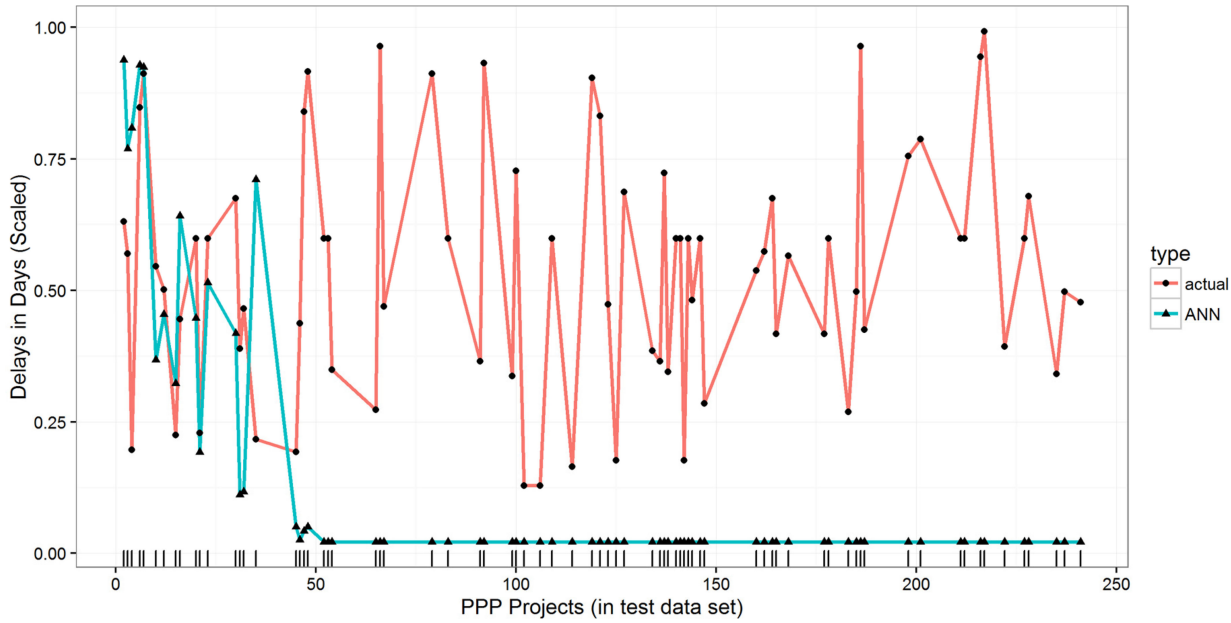


Fig. 10. Evaluating observed and predicted delays in the PPP projects.

(RSS) and PRAF. While the RSS compares the predictive performance of the model (flexibility and interpretability were examined separately), PRAF compares each predictor's importance in forecasting project delay. Based on results from data analysis, RF show the least residual error, with an error margin of 1.03 and is considered good in flexibility. This is immediately followed by decision tree with RSS score of 2.17. Linear regression, SVM, and DNNs, however, showed profound weakness in predictive performance with large error margins of 23.20, 25.64, and 469.56 between the data and the estimation models, respectively. Table IV below shows detailed comparison of the predictive modeling techniques. Further details of the results are discussed in greater detail in the next section.

2) Percentage Rank Factor: Going further, in order to have an overall agreement in the ranking of all predictors, the RAF and PRAF [11], [115] were applied. RAF and PRAF are mathematically computed using (15) and (16), respectively

$$\mathrm{RAF} = \frac{\Sigma\,(\mathrm{LR})\,(\mathrm{RT})\,(\mathrm{RF})\,(\mathrm{SVM})\,(\mathrm{DNN})}{N} \tag{15}$$

$$\mathrm{PRAF} = \frac{\mathrm{RAF}_{\max} - \mathrm{RAF}_i}{\mathrm{RAF}_{\max}} \times 100\% \tag{16}$$

TABLE IV
COMPARISON OF BIG DATA ANALYTICS TECHNIQUES BASED ON RSS

	Big Data Analytics Techniques	*RSS*	*Flexibility*	*Interpretability*
1	Random Forest	1.03	Good	low
2	Decision Tree	2.17	average	high
3	Linear Regression	23.20	Low	high
4	Support Vector Machine	25.64	High	Low
5	Deep Neural Network	469.56	High	low

where RAFmax = maximum RAF, RAF_i is the RAF for criteria i, N = number of variable predictors ranked, and $\Sigma(\mathrm{LR})(\mathrm{RT})(\mathrm{RF})(\mathrm{SVM})(\mathrm{DNN})$ = sum of the order of rankings of linear regression, regression trees, RF, SVM, and DNN. An absolute rank difference of 2, for example, implies more agreement as to the importance of the predictor than when the absolute rank difference is 3. The rank agreement factor may be >1, with a higher factor indicating more disagreement [11]. For the 16 predictors affecting project delay, the maximum $\mathrm{RAF}_{\max} = 2.00$. An RAF of zero implies perfect agreement. The result RAF for the models is shown in the 14th column of Table V. In addition, a cursory look at results of the PRAF in Table V shows the five most important predictors contributing to project delay to be as follows:

1) "percentage shortage in skilled labor;"
2) "percentage delay in material delivery;"
3) "number of site Accidents and injuries;"
4) "percentage of design variations;"
5) "percentage of liquidated and ascertained damages in projects."

These predictors are further enumerated in the discussion section.

Additionally, the study ranked the significance of predictors under each of the five models using, *P*-value (linear regression), Gini (regression tree), impurity (RF), Monte Carlo sensitivity analysis (SVM), and weight (DNN). For the linear regression model, the study conducted a one-sample *t*-test to derive *p*-values for each predictor at 95% confidence level. If the mean difference is significantly different from the hypothesized value ($<.05$), it means that the value is statistically important in affecting project delay at the 95% confidence level (see column 3 and 4 of Table V for *P*-value of each predictor and their ranking). Going further, with regression tree, the study also evaluated the importance of some variable X_k when predicting *Y* by adding up the decreases in weighted impurity for all nodes *t*, where X_k is used (averaged over all trees in the forest, but actually, we can use it on a single tree)

$$I(X_k) = \frac{I}{M} \sum_m \sum_t \frac{N_t}{N} \Delta i(t) \tag{17}$$

where the second sum is only on nodes t based on variable X_k. If $i(\cdot)$ is Gini index, then $I(\cdot)$ is called Mean Decrease Gini function. In addition, in order to identify which of the predictor variables are most important for predicting project delay in PPP projects, we used RF to derive the mean decrease impurity importance of each predictor from assemblages of randomized trees. The ranking of each predictors derived from this process are shown in column 7 and 8 of Table V. Regarding SVM, sample data were smoothly segregated based on sectors and contract types. In case of DNN, hidden layers are involved with complex interactions; hence, getting a single value for attributes is not realistic. As such, zero is set as the weight and rank of these attributes in DNN to carry out the overall ranking process.

VI. DISCUSSION

This section discusses results from the study and started by comparing the predictive performance of the five models (RF, linear regression, decision tree, SVM, and DNNs) in forecasting delay in PPP projects, their flexibility and interpretability, respectively. Based on evidences shown in Table IV, a cursory look at the RSS of the five analytical models suggest that RF has the best predictive performance in terms of reducing error in the model to 1.23. This is followed by decision tree with RSS score of 2.17. Linear regression, SVM, and DNNs however show profound weakness in predictive performance with large error margins of 23.20, 25.64, and 469.56 between the data and the estimation models, respectively. According to Theobald [104], RSS is a measure of the variability or error in the dataset which is not captured in the model. A small RSS therefore suggests a tight fit of the estimation model to the data used for analysis [109]. This suggests the capability of RF in this study to explain a greater amount of the dataset. However, considering that RSS alone may not be entirely suitable to judge the correctness of the models [2], flexibility, and interpretability of the five models were also considered in the study. Although SVM and DNNs showed high flexibility as evidenced in Table IV, this is only attributed to their ability to accept and review new data streaming in and thus help provide a progressively realistic assessment of a model [46]. However, whilst RF is considered good enough in terms of flexibility [26], [33], [85], decision tree and linear regression are rated average and low, respectively, in model flexibility. Additionally, this study examined users' ability to interpret the model, which is also an important factor in deciding which model may be suitable for forecasting CR. As represented in Table IV, the results show that while decision tree and linear regression are high on interpretability, which confirms their wider uptake in risk analysis, RF, and DNN models are rated very low in interpretability. However, in the overall, and based on

TABLE V
PRAF OF THE FOUR BIG DATA PREDICTIVE MODELS AND THEIR LEVEL OF SIGNIFICANCE (P-VALUE)

Sr.#.	Predictors	Ranking of Factors by Models										Sum Ranks (Σ)	RAF	PRAF	Overall Ranking Order
		Linear Regression		Regression Tree		Random Forest		Support Vector Machine		Neural Network					
		P-value	Rank	Gini	Rank	Impurity	Rank	M-CSA	Rank	Weight	Rank				
1	Percentage shortage in skilled labour	0.08421	4	153.1195	2	8.05966	4	0.085	9	0	0	13	0.81	69.11	1
2	Percentage Delay in Material delivery	< 2e-16	1	55.2312	8	8.27137	3	0.092	1	0	0	19	1.19	54.75	2
3	Number of site Accidents and injuries	0.00103	2	114.3011	5	7.28248	6	0.066	11	0	0	21	1.31	50.10	3
4	Percentage of design variations	0.5397	5	173.7692	1	6.91121	7	0.003	15	0	0	24	1.50	42.97	4
5	Percentage of liquidated and ascertained damages in projects	< 2e-16	1	41.30585	10	7.49608	5	0.086	5	0	0	26	1.63	38.21	5
6	Number of unforeseen site conditions	0.5938	5	132.5914	4	6.79174	8	0.053	12	0	0	28	1.75	33.46	6
7	Percentage fluctuation in construction material price index	0.00424	2	0.562405	16	13.6006	2	0.005	14	0	0	29	1.81	31.08	7
8	Percent change in inflation	0.04361	3	141.6945	3	6.40911	11	0.010	13	0	0	29	1.81	31.07	8
9	Average number of disputes among parties	3.21E-06	1	62.27339	7	6.49062	10	0.085	8	0	0	30	1.88	28.71	9
10	Number of defects in a construction project	4.48423	5	10.30915	15	48.8836	1	0.003	16	0	0	30	1.88	28.70	10
11	Number of days with bad weather that prevented site work	0.03344	3	93.32449	6	5.43395	14	0.086	6	0	0	30	1.88	26.33	11
12	Percentage of materials damaged on site	< 2e-16	1	21.89769	14	6.55648	9	0.085	7	0	0	32	2.00	23.95	12
13	Number of days for site closure	0.02846	3	52.33817	9	6.16393	16	0.087	4	0	0	34	2.13	19.20	13
14	Projects were either procured via turnkey or Design Bid Build	0.8632	5	40.06702	11	3.71788	12	0.091	2	0	0	34	2.13	19.18	14
15	Projects chosen cut across 9 sectors of the economy	0.63428	5	36.84742	12	3.80682	15	0.072	10	0	0	37	2.31	12.07	15
16	Percentage of inferior materials supplied to site	0.31824	5	22.02123	13	5.59273	13	0.087	3	0	0	42	2.63	0	16

its seeming higher predictive performance (least test error) and flexibility, this study therefore suggests RF for predicting CR in large portfolio of PPP projects. According to Liaw and Wiener [59], RF provides a powerful approach to data exploration; analysis and predictive modeling of uncertainty (see also [96]). With a high error detection rate and easy identification of anomalies and outliers in data [78], RF will enable automatic identification of significant predictors influencing PPP project delay [5]. RF is therefore considered a desirable technique capable of helping to make more accurate decisions toward minimizing time wastage in delivering projects.

The second phase of data analysis in this study examines the key predictors contributing toward delay in PPP projects out of the 16 predictors investigated (14 numerical and 2 categorical predictors). As evidenced in Table V, results of PRAF calculation performed on the data relating to the 16 predictors indicate that overall there are five most important predictors contributing toward project delay. These are as follows.

1) *Percentage shortage in skilled labor:* After extensive data analysis, the study identified percentage shortage in skilled labor as the first most significant factor contributing to delay in construction projects with a PRAF score of 69.11. This confirms that shortage in skilled workers creates bottlenecks with various implications on project cost, quality, productivity, and timely completion, as suggested in Teizer *et al*. in [106] (see also [56]). Usually, the construction industry employs subcontractors, direct labor, and third party services including project management, and sustainable solutions. However, the recent global recession coupled with the increased demand for quality infrastructures (Mackenzie *et al*., 2001) has contributed to the massive shortage of skilled work force in the global construction industry [2]. According to Larsen *et al*., [56], the huge number of skilled workers that left the construction industry at the wake of the financial crisis had a major impact in the industry's completion rate, with more companies identifying insufficient skilled workers as one of the major causes of schedule overrun in projects [54]. This situation is also worsened by the insufficient number of new recruits joining the industry through apprenticeship, resulting in growing skill gap in areas such as carpenters, millwrights, and electrical technicians among others [1].
2) *Percentage Delay in Material delivery:* Percentage delay in material delivery was identified as the second most important predictor of project delay in this study showing a PRAF score of 54.75. Existing studies such as [1], [25], [111] have also highlighted the above perspective and suggested that timely completion of projects is often contingent upon trouble-free supply to project site. As argued by [2], the supply chain is an important stakeholder in construction project delivery and ensures the right construction material and quantities are delivered in a timely fashion at the right location. Al-Hazim *et al*. [2] identified some causes of delays in material delivery as high demand for construction material, long procedure of purchasing order, poor communication between the contractor and the supplier among others (see also [25], [49]). Besides being a major cause of CR, delay in material delivery to site also results in significant cost overrun to the contractor in terms of wasted productive time for workers waiting for materials, penalties in liquidated and ascertained damages in the event of project's failure to meet completion deadline, etc. [56].
3) *Number of site accidents and injuries:* Number of site accidents and injuries was ranked as the third important predictor of project delay with a PRAF score of 50.10. This confirms studies such as [69], [87], [111], which have emphasized construction site accidents as one of the important factors contributing to project delay. Ching [25] suggested that unsafe behavior is the most significant contributor to construction site accidents with a resulting impact on timely completion of projects. According to [56], in most instances of site accidents, the project manager is often obliged to either temporarily suspend site activities or in a number of fatal cases, call indefinite site closure to allow proper investigation and assessment of such accidents. This results in man-hour loss and causes disruption to schedule of projects' activities [111].
4) *Percentage of design variations:* Another important predictor of project delay is the percentage of design variations carried out on the project with a PRAF score of 42.97. Design variations are a general phenomenon in construction projects [10]. Variations have to do with the amendments to original project design and ultimately the project scope (Kangari, 1995). Variations are a contentious issue in construction project and often cause disputes among project stakeholders [1], [102]. In most instances, variations in project are initiated by client [111]. This happens because, often times, many clients do not fully make up their mind about what they want in terms of project's designs and other aspects, until the construction commences [111]. As such, they tend to make their decisions as the project's construction process progresses, while proposing different variations to original project scope and design. Variations have serious implications for timely completion of projects and the more or bigger the variations implemented on a project, the higher the potential for CR [102]. A number of studies have suggested better engagement between the client and contractor at the preconstruction stage may reduce the number of potential variations to a project's scope [1], [78], [101].
5) *Percentage of liquidated and ascertained damages in projects:* The study identified percentage of liquidated and ascertained damages (LAD) as the fifth most important predictor of project delay with a PRAF value of 38.21. According to Hampton *et al.* (2010), liquidate and ascertained damages arises from failure of the construction contractor to successfully put the project into operations at the agreed deadline. LAD is often contractual, and the penalty for it is expressed as a financial liability to the contractor [46]. As argued by [82], except where a project contractor is a big construction firm with strong financial

capabilities, a huge financial penalty in liquidated damages may cause financial distress to the contractor, which may also affect its ability to deliver the project as scheduled. As suggested by [12] and [49], many SME contractors in the construction industry had gone bankrupt due to incurring heavy financial liabilities via liquated damages, while eventually failing to deliver such projects at their deadlines. Studies such as [1] and [95] argued that quick resolutions of contractual issues without recourse to lengthy court actions will mitigate the impact of LAD.

VII. Implication for Practice

Events in the industry over time had prompted arguments about how best to estimate project delay to enable benchmarking for future project delivery and help improve procurement policies [37], [60], [64]. Industry stakeholders, especially public sector clients had clamored for realistic forecasting and benchmarking of project delays [78], [84], [94], [102]. This comes amidst recent statistics suggesting delay as a recurring decimal within the construction industry [10], [54], [83]. By proposing a Big Data predictive modeling approach, this study provides a reliable technique for CR forecasting by comparing the predictive performance of five advanced analytical techniques (DNNs, SVM, RF, linear regression, and decision tree). The study focused on 16 drivers of project delay and proposed RF as the best possible analytic technique for predicting CR. This is based on evidences from the study, which shows that RF model has the least residual error with good flexibility, and such a good fit for predicting and benchmarking CR. This is against the low performances of other four predictive models. It therefore has significant implication for construction industry stakeholders in terms of choosing the right model that helps accurately predict the possibility of delay in PPP projects. Based on the evidences from the study, five key predictors with significant impact on delay were also considered:

1) "percentage shortage in skilled labor;"
2) "percentage delay in material delivery;"
3) "number of site accidents and injuries;"
4) "percentage of design variations;"
5) "percentage of liquidated and ascertained damages in projects."

These results show that construction industry stakeholders will benefit more from including the evaluation of these predictors in their strategic framework for risk evaluation and monitoring. This is considered crucial toward addressing the growing concern about CR in the industry, especially when considering mega PPP projects. According to recent statistics from KPMG global Infrastructure Report (2015), only 25% of projects delivered globally in the last 3 years came within 10% of completion deadline. This excessive time overrun on projects have far-reaching negative implications especially in the case of PPPs, where taxpayers' money is often exposed.

Additionally, this study emerges at an opportune time for policy makers and industry stakeholders to reflect on the performance of historical PPP projects in terms of delay and ultimately redesign procurement policies to meet existing realities. The Big Data predictive modeling technique will thus be useful at the procurement stage of PPP projects, to estimate the potential delay in projects using critical input variables. Looking at a 2005 report by one of the Not for Profit organizations in the U.K. (The Tax Payers Alliance), statistics show the total net cost overrun for 305 public sector projects was over £23 billion above initial estimates, with a significant chunk of the cost attributed to project delays. By estimating potential delay in future projects, policy makers and contractors will be able to adopt effective project management strategies that can deliver cost savings on future public procurements. Similarly, considering that 80% to 90% of construction costs in PPPs are financed through banks' limited recourse funds, CR forecasts can enable financiers to make informed decisions concerning loan life and refinancing for PPP investments. With a Big Data enabled prediction of CR, new industry standards in terms of average delay in various types of PPP projects across different sectors can also be established as best practice for the construction industry. Additionally, the study offers new opportunities to project-based firms, public sector clients, contractors, financiers, and other relevant stakeholders for developing increased capabilities relevant for managing CR during construction phase of their projects.

VIII. Conclusion

Accurate prediction of potential delays in PPP projects is considered vital for providing valuable insights that are relevant for planning and mitigating CR in future PPP projects. This study examined BDA-driven predictive modeling of CR (project delay) in PPP projects. In order to forecast potential delay in PPP projects, predictive performance of five advanced BDA techniques, namely DNNs, RF, SVM, regression trees, and multivariate linear regression were compared. Using huge datasets from 4294 PPP project samples across Europe between 1992 and 2015, 16 predictors influencing delay in PPPs (i.e., percentage (%) shortage in skilled labor, number of site accidents and injuries, etc.) were employed to identify underlying pattern in project delay and its' relationship with the identified influential predictors. The data were analyzed using two categorical variables, namely contract type and sector to introduce dimensions for analyzing the rest of the predictors and to uncover nonobvious correlations. With minimum, maximum, and average values for each predictor produced from various construction industry data and government statistical reports, trends showing the behavior of delay were generated across the entire dataset.

After extensive analysis of the projects' data, results show that out of the five BDA techniques, RF has the best predictive performance for forecasting delay across large samples of projects. RF showed minimum RSS error with high predictive performance accuracy compared to the three remaining analytics techniques. Evidences from the study also show that five predictors significantly with delay across the five models:

1) "percentage shortage in skilled labor;"
2) "percentage delay in material delivery;"
3) "number of site accidents and injuries;"

4) "percentage of design variations;"
5) "percentage of liquidated and ascertained damages in projects."

These predictors were therefore considered as key contributors to project delay in construction PPP projects. The predictors showed higher correlation coefficients with delay across five sectors (hospitals, schools, public buildings, others, defense) and the two contract types [(Fixed Price Turnkey and Design Bid Build (DBB)]. In considering contract type as an important predictor of delay, results showed massive delay in PPP projects, where the DBB approach has been used, as against the fixed price turnkey method. The statistical significance of the results was compelling to the extent that large samples of projects were discovered to have been delayed beyond 150% of construction duration. Other predictors such as number of days with bad weather preventing project work, also revealed reasonable level of correlation with delay across the dataset. This study contributes to knowledge by proposing a BDA predictive model for predicting delay in PPP projects. By unraveling the hidden correlations and patterns contributing toward delay within the construction process, the negative impact of CR on project timeline, contractual obligations, and contractors' margins can be mitigated. This study also provides valuable opportunities policy makers and other industry stakeholders to consider evidence-based industry benchmarks for delay in future PPP projects. Such move is therefore expected to offer additional benefits of efficiency in PPP procurements. This study has examined CR (project delay) within the context of construction PPP projects delivered across few countries in Europe. As such, findings from the study should be interpreted within that context. Possible areas for future research are BDA investigation of critical predictors of cost overrun in historical PPP projects, a Big Data-driven research into counter-party risk and PPP contracting toward identifying top construction contractor practices influencing liquidated and ascertained damage payments.

REFERENCES

[1] A. Adam, P. E. B. Josephson, and G. Lindahl, "Aggregation of factors causing cost overruns and time delays in large public construction projects: Trends and implications," *Eng. Construct. Architectural Manage.*, vol. 24, no. 3, pp. 393–406, 2017.

[2] N. Al-Hazim, Z. A. Salem, and H. Ahmad, "Delay and cost overrun in infrastructure projects in jordan." *Procedia Eng.*, vol. 182, pp. 18–24, 2017.

[3] S. M. Ahmed, R. Ahmad, D. Saram, and D. Darshi, "Risk management trends in the Hong Kong construction industry: a comparison of contractors and owners perceptions," *Eng. Construct. Architectural Manage.*, vol. 6, no. 3, pp. 225–234, 1999.

[4] H. N. Ahuja and V. Nandakumar, "Simulation model to forecast project completion time," *J. Construct. Eng. Manage.*, vol. 111, no. 4, pp. 325–342, 1985.

[5] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Comput. Statist. Data Anal.*, vol. 52, no. 4, pp. 2249–2260, 2008.

[6] K. André, "Dealing with completion risk," *Risk Manage.*, 2013. [Online]. Available: www.ampsdelft.nl/onderzoek_en_publicatie/ControllersMagazine_ENG.pdf. Accessed on: Feb. 23, 2016.

[7] S. A. Assaf and S. Al-Hejji, "Causes of delay in large construction projects," *Int. J. Project Manage.*, vol. 24, no. 4, pp. 349–357, 2006.

[8] A. A. Aibinu and G. O. Jagboro, "The effects of construction delays on project delivery in Nigerian construction industry," *Int. J. Project Manage.*, vol. 20, no. 8, pp. 593–599, 2002.

[9] S. A. Assaf, M. Al-Khalil, and M. Al-Hazmi, "Causes of delay in large building construction projects," *J. Manage. Eng.*, vol. 11, no. 2, pp. 45–50, 1995.

[10] E. Allen and J. Iano, *Fundamentals of Building Construction: Materials and Methods*. Hoboken, NJ, USA: Wiley, 2011.

[11] A. U. Elinwa and M. Joshua, "Time-overrun factors in nigerian construct. industry," *J. Construct. Eng. Manage.*, vol. 127, no. 5, pp. 419–425, 2001, doi:10.1061/(ASCE)0733-9364(2001)127:5(419).

[12] M. Backstrom, "An examination of the independent certification processes of a construction contract," *Building Construct. Law J.*, vol. 29, no. 5, pp. 406–416, 2013.

[13] M. Bilal *et al.*, "Big data in the construction industry: A review of present status, opportunities, and future trends," *Adv. Eng. Informat.*, vol. 30, no. 3, pp. 500–521, 2016.

[14] M. Bilal *et al.*, "Analysis of critical features and evaluation of BIM software: Towards a plug-in for construction waste minimization using big data," *Int. J. Sustain. Building Technol. Urban Develop.*, vol. 6, no. 4, pp. 211–228, 2015.

[15] L. Bing, A. Akintoye, P. J. Edwards, and C. Hardcastle, "The allocation of risk in PPP/PFI construction projects in the UK," *Int. J. Project Manage.*, vol. 23, no. 1, pp. 25–35, 2005.

[16] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Inf., Commun. Soc.*, vol. 15, no. 5, pp. 662–679, 2012.

[17] B. Brown, M. Chui, and J. Manyika, "Are you ready for the era of 'big data'," *McKinsey Quart.*, vol. 4, no. 1, pp. 24–35, 2011.

[18] G. Carter and S. D. Smith, "Safety hazard identification on construction projects," *J. Construct. Eng. Manage.*, vol. 132, no. 2, pp. 197–205, 2006.

[19] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi, "Long-range communications in unlicensed bands: The rising stars in the IoT and smart city scenarios," *IEEE Wireless Commun.*, vol. 23, no. 5, pp. 60–67, Oct. 2016.

[20] S. Chakrabarty and D. W. Engels, "A secure IoT architecture for smart cities," in *Proc. 13th IEEE Annu. Consumer Commun. Netw. Conf.*, Jan. 2016, pp. 812–813.

[21] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[22] B. A. Bossink, "Managing drivers of innovation in construction networks," *J. Construct. Eng. Manage.*, vol. 130, no. 3, pp. 337–345, 2004.

[23] D. Baloi and A. D. Price, "Modelling global risk factors affecting construction cost performance," *Int. J. Project Manage.*, vol. 21, no. 4, pp. 261–269, 2003.

[24] J. L. Burati, Jr., J. J. Farrington, and W. B. Ledbetter, "Causes of quality deviations in design and construction," *J. Construct. Eng. Manage.*, vol. 118, no. 1, pp. 34–49, 1992.

[25] F. D. Ching, *Building Construction Illustrated*. Hoboken, NJ. USA: Wiley, 2014.

[26] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Found. Trends Comput. Graphics Vision*, vol. 7, nos. 2/3, pp. 81–227, 2012.

[27] K. Davis, W. B. Ledbetter, and J. L. Burati, Jr, "Measuring design and construction quality costs," *J. Construct. Eng. Manage.*, vol. 115, no. 3, pp. 385–400, 1989.

[28] S. M. Dissanayaka, and M. M. Kumaraswamy, "Evaluation of factors affecting time and cost performance in Hong Kong building projects," *Eng. Construct. Architectural Manage.*, vol. 6, no. 3, pp. 287–298, 1999.

[29] I. Dikmen, M. T. Birgonul, and S. Han, "Using fuzzy risk assessment to rate cost overrun risk in international construction projects," *Int. J. Project Manage.*, vol. 25, no. 5, pp. 494–505, 2007.

[30] R. G. Eccles, "The quasifirm in the construction industry," *J. Econ. Behav. Org.*, vol. 2, no. 4, pp. 335–357, 1981.

[31] A. Errasti, R. Beach, A. Oyarbide, and J. Santos, "A process for developing partnerships with subcontractors in the construction industry: An empirical study," *Int. J. Project Manage.*, vol. 25, no. 3, pp. 250–256, 2007.

[32] S. M. El-Sayegh, "Risk assessment and allocation in the UAE construction industry," *Int. J. Project Manage.*, vol. 26, no. 4, pp. 431–438, 2008.

[33] J. S. Evans, M. A. Murphy, Z. A. Holden, and S. A. Cushman, "Modeling species distribution and change using random forest," in *Predictive Species and Habitat Modeling in Landscape Ecology*. New York, NY, USA: Springer, 2011, pp. 139–159.

[34] J. B. Fan, Y. Chikashige, C. L. Smith, O. Niwa, M. Yanagida, and C. R. Cantor, "Construction of a Not I restriction map of the fission yeast Schizosaccharomyces pombe genome," *Nucleic Acids Res.*, vol. 17, no. 7, pp. 2801–2818, 1989.

[35] P. G. Fookes, W. J. French, and S. M. M. Rice, "The influence of ground and groundwater geochemistry on construction in the Middle East," *Quart. J. Eng. Geol. Hydrogeol.*, vol. 18, no. 2, pp. 101–127, 1985.

[36] B. Flyvbjerg, M. K. Skamris Holm, and S. L. Buhl, "What causes cost overrun in transport infrastructure projects?" *Transport Rev.*, vol. 24, no. 1, pp. 3–18, 2004.

[37] I. W. Fung, V. W. Tam, T. Y. Lo, and L. L. Lu, "Developing a risk assessment model for construction safety," *Int. J. Project Manage.*, vol. 28, no. 6, pp. 593–600, 2010.

[38] A. Fight, *Introduction to Project Finance*. Oxford, U.K.: Butterworth–Heinemann, 1999.

[39] J. H. Friedman and W. Stuetzle, "Projection pursuit regression," *J. Amer. Statist. Assoc.*, vol. 76, no. 376, pp. 817–823, 1981.

[40] N. Gatzert and T. Kosub, "Risks and risk management of renewable energy projects: The case of onshore and offshore wind parks," *Renew. Sustain. Energy Rev.*, vol. 60, pp. 982–998, 2016.

[41] A. Gaur, B. Scotney, G. Parr, and S. McClean, "Smart city architecture and its applications based on IoT," *Procedia Comput. Sci.*, vol. 52, pp. 1089–1094, 2015.

[42] D. D. Gransberg and K. Molenaar, "Analysis of owner's design and construction quality management approaches in design/build projects," *J. Manage. Eng.*, vol. 20, no. 4, pp. 162–169, 2004.

[43] G. Hampton, A. N. Baldwin, and G. Holt, "Project delays and cost: stakeholder perceptions of traditional v. PPP procurement," *J. Financial Manage. Property Construct.*, vol. 17, no. 1, pp. 73–91, 2012.

[44] S. E. Hampton *et al.*, "Big data and the future of ecology," *Frontiers Ecol. Environ.*, vol. 11, no. 3, pp. 156–162, 2013.

[45] C. Harty, "Innovation in construction: A sociology of technology approach," *Building Res. Inf.*, vol. 33, no. 6, pp. 512–522, 2005.

[46] J. J. Hopfield, "Artificial neural networks," *IEEE Circuits Devices Mag.*, vol. 4, no. 5, pp. 3–10, Sept. 1988.

[47] H. Hu, Y. Wen, T. S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

[48] C. Hendrickson, and T. Au, *Project Management for Construction: Fundamental Concepts for Owners, Engineers, Architects, and Builders*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.

[49] A. A. Javed, P. T. Lam, and A. P. Chan, "A model framework of output specifications for hospital PPP/PFI projects," *Facilities*, vol. 31, nos. 13/14, pp. 610–633, 2013.

[50] A. A. Javed, P. T. Lam, and P. X. Zou, "Output-based specifications for PPP projects: Lessons for facilities management from Australia," *J. Facilities Manage.*, vol. 11, no. 1, pp. 5–30, 2013.

[51] P. F. Kaming, P. O. Olomolaiye, G. D. Holt, and F. C. Harris, "Factors influencing construction time and cost overruns on high-rise projects in Indonesia," *Construction Manage. Econ.*, vol. 15, no. 1, pp. 83–94, 1997.

[52] A. Kazaz, and S. Ulubeyli, "Drivers of productivity among construction workers: A study in a developing country," *Building Environ.*, vol. 42, no. 5, pp. 2132–2140, 2007.

[53] S. Y. Kim, N. Van Tuan, and S. O. Ogunlana, "Quantifying schedule risk in construction projects using Bayesian belief networks," *Int. J. Project Manage.*, vol. 27, no. 1, pp. 39–50, 2009.

[54] "Climbing the curve," KPMG Global Construction Industry Report, 2015. [Online]. Available: https://www.kpmg.com/Global/. . . / global-construction . . . /global-construction-survey-201. Accessed on: Mar. 12, 2015.

[55] N. K. Kittusamy, and B. Buchholz, "Whole-body vibration and postural stress among operators of construction equipment: A literature review," *J. Saf. Res.*, vol. 35, no. 3, pp. 255–261, 2004.

[56] J. K. Larsen, G. Q. Shen, S. M. Lindhard, and T. D. Brunoe, "Factors affecting schedule delay, cost overrun, and quality level in public construction projects," *J. Manage. Eng.*, vol. 32, no. 1, 2015, Art. no. 04015032.

[57] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, "Big data, analytics and the path from insights to value," *MIT Sloan Manage. Rev.*, vol. 52, no. 2, pp. 21–31, 2011.

[58] L. Le-Hoai, Y. Dai Lee, and J. Y. Lee, "Delay and cost overruns in Vietnam large construction projects: A comparison with other selected countries," *KSCE J. Civil Eng.*, vol. 12, no. 6, pp. 367–377, 2008.

[59] A. Liaw, and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[60] J. K. Lee, "Cost overrun and cause in Korean social overhead capital projects: Roads, rails, airports, and ports," *J. Urban Planning Develop.*, vol. 134, no. 2, pp. 59–62, 2008.

[61] F. Y. Y. Ling and L. Hoi, "Risks faced by Singapore firms when undertaking construction projects in India," *Int. J. Project Manage.*, vol. 24, no. 3, pp. 261–270, 2006.

[62] E. C. Lim and J. Alum, "Construction productivity: issues encountered by contractors in Singapore," *Int. J. Project Manage.*, vol. 13, no. 1, pp. 51–58, 1995.

[63] X. Li *et al.*, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.

[64] P. E. Love, D. J. Edwards, and Z. Irani, "Moving beyond optimism bias and strategic misrepresentation: An explanation for social infrastructure project cost overruns," *IEEE Trans. Eng. Manage.*, vol. 59, no. 4, pp. 560–571, Nov. 2012.

[65] W. Lu, X. Chen, D. C. Ho, and H. Wang, "Analysis of the construction waste management performance in Hong Kong: the public and private sectors compared using big data," *J. Cleaner Production*, vol. 112, pp. 521–531, 2016.

[66] W. Lu, X. Chen, Y. Peng, and L. Shen, "Benchmarking construction waste management performance using big data," *Resources, Conserv. Recycling*, vol. 105, pp. 49–58, 2015.

[67] S. Mackenzie, A. R. Kilpatrick, and A. Akintoye, "UK construction skills shortage response strategies and an analysis of industry perceptions," *Construct. Manage. Econ.*, vol. 18, no. 7, pp. 853–862, 2000.

[68] V. A. Memos, K. E. Psannis, Y. Ishibashi, B. G. Kim, and B. B. Gupta, "An efficient algorithm for media-based surveillance system (EAMSuS) in IoT smart city framework," *Future Gener. Comput. Syst.*, vol. 83, pp. 619–628, 2018.

[69] S. Mohamed, "Safety climate in construction site environments," *J. Construct. Eng. Manage.*, vol. 128, no. 5, pp. 375–384, 2002.

[70] T. M. Mezher and W. Tawil, "Causes of delays in the construction industry in Lebanon," *Eng., Construct. Architectural Manage.*, vol. 5, no. 3, pp. 252–260, 1998.

[71] O. Moselhi, D. Gong, and K. El-Rayes, "Estimating weather impact on the duration of construction activities," *Can. J. Civil Eng.*, vol. 24, no. 3, pp. 359–366, 1997.

[72] M. A. Mustafa and J. F. Al-Bahar, "Project risk assessment using the analytic hierarchy process," *IEEE Trans. Eng. Manage.*, vol. 38, no. 1, pp. 46–52, Feb. 1991.

[73] S. Mohamed, "Safety climate in construction site environments," *J. Construct. Eng. Manage.*, vol. 128, no. 5, pp. 375–384, 2002.

[74] A. Ng and M. Loosemore, "Risk allocation in the private provision of public infrastructure," *Int. J. Project Manage.*, vol. 25, no. 1, pp. 66–76, 2007.

[75] W. K. Newey and K. D. West, "Automatic lag selection in covariance matrix estimation," *Rev. Econ. Stud.*, vol. 61, no. 4, pp. 631–653, 1994.

[76] F. J. Ohlhorst, *Big Data Analytics: Turning Big Data Into Big Money*. Hoboken, NJ, USA: Wiley, 2012.

[77] A. M. Odeh, and H. T. Battaineh, "Causes of construction delay: Traditional contracts," *Int. J. Project Manage.*, vol. 20, no. 1, pp. 67–73, 2002.

[78] R. Pal, P. Wang, and X. Liang, "The critical factors in managing relationships in international engineering, procurement, and construction (IEPC) projects of Chinese organizations," *Int. J. Project Manage.*, vol. 35, no. 7, pp. 1225–1237, 2017.

[79] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2005.

[80] J. Palomo, D. Rios Insua, and F. Ruggeri, "Modeling external risks in project management," *Risk Anal.*, vol. 27, no. 4, pp. 961–978, 2007.

[81] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, "Urban planning and building smart cities based on the internet of things using big data analytics," *Comput. Netw.*, vol. 101, pp. 63–80, 2016.

[82] K. S. Rebeiz, "Public–private partnership risk factors in emerging countries: BOOT illustrative case study," *J. Manage. Eng.*, vol. 28, no. 4, pp. 421–428, 2011.

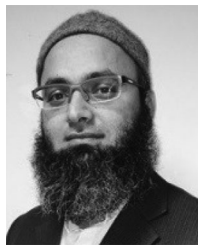
[83] H. S. Robinson and J. Scott, "Service delivery and performance monitoring in PFI/PPP projects," *Construct. Manage. Econ.*, vol. 27, no. 2, pp. 181–197, 2009.

[84] D. M. Rousseau and C. Libuser, "Contingent workers in high risk environments," *California Manage. Rev.*, vol. 39, no. 2, pp. 103–123, 1997.

[85] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 67, pp. 93–104, 2012.
[86] J. S. Russell and E. J. Jaselskis, "Predicting construction contractor failure prior to contract award," *J. Construct. Eng. Manage.*, vol. 118, no. 4, pp. 791–811, 1992.
[87] E. Sawacha, S. Naoum, and D. Fong, "Factors affecting safety performance on construction sites," *Int. J. Project Manage.*, vol. 17, no. 5, pp. 309–315, 1999.
[88] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Proc. Int. Conf. Collab. Technol. Syst.*, May 2013 pp. 42–47.
[89] V. Scuotto, A. Ferraris, and S. Bresciani, "Internet of things: Applications and challenges in smart cities: A case study of IBM smart city projects," *Bus. Process Manage. J.*, vol. 22, no. 2, pp. 357–367, 2016.
[90] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.
[91] J. S. Shane, K. R. Molenaar, S. Anderson, and C. Schexnayder, "Construction project cost escalation factors," *J. Manage. Eng.*, vol. 25, no. 4, pp. 221–229, 2009.
[92] F. J. Sanger and F. H. Sayles, "Thermal and rheological computations for artificially frozen ground construction," *Eng. Geol.*, vol. 13, no. 1, pp. 311–337, 1979.
[93] C. Semple, F. T. Hartman, and G. Jergeas, "Construction claims and disputes: Causes and cost/time overruns," *J. Construct. Eng. Manage.*, vol. 120, no. 4, pp. 785–795, 1994.
[94] L. Y. Shen, J. Li Hao, V. W. Y. Tam, and H. Yao, "A checklist for assessing sustainability performance of construction projects," *J. Civil Eng. Manage.*, vol. 13, no. 4, pp. 273–281, 2007.
[95] M. Sun and X. Meng, "Taxonomy for change causes and effects in construction projects," *Int. J. Project Manage.*, vol. 27, no. 6, pp. 560–572, 2009.
[96] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: A classification and regression tool for compound classification and QSAR modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947–1958, 2003.
[97] C. B. Tatum, "Process of innovation in construction firm," *J. Construct. Eng. Manage.*, vol. 113, no. 4, pp. 648–663, 1987.
[98] C. B. Tatum, "Organizing to increase innovation in construction firms," *J. Construct. Eng. Manage.*, vol. 115, no. 4, pp. 602–617, 1989.
[99] F. P. Tolman, "Product modeling standards for the building and construction industry: Past, present and future," *Automat. Construct.*, vol. 8, no. 3, pp. 227–235, 1999.
[100] D. Talia, "Toward cloud-based Big-data analytics," *IEEE Comput. Sci.*, pp. 98–101, 2013. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1215667
[101] C. M. Tam, S. X. Zeng, and Z. M. Deng, "Identifying elements of poor construction safety management in China," *Saf. Sci.*, vol. 42, no. 7, pp. 569–586, 2004.
[102] V. W. Y. Tam, and I. W. H. Fung, "A study of knowledge, awareness, practice and recommendations among Hong Kong construction workers on using personal respiratory protective equipment at risk," *Open Construct. Building Technol. J.*, vol. 2, pp. 69–81, 2008.
[103] J. Teizer, B. S. Allread, C. E. Fullerton, and J. Hinze, "Autonomous pro-active real-time construction worker and equipment operator proximity safety alert system," *Automat. Construct.*, vol. 19, no. 5, pp. 630–640, 2010.
[104] C. M. Theobald, "Generalizations of mean square error applied to ridge regression," *J. Roy. Statist. Soc., Ser. B (Methodological)*, vol. 36, pp. 103–106, 1974.
[105] E. A. L. Teo, F. Y. Y. Ling, and A. F. W. Chong, "Framework for project managers to manage construction safety," *Int. J. Project Manage.*, vol. 23, no. 4, pp. 329–341, 2005.
[106] R. L. Tiong, "BOT projects: Risks and securities," *Construct. Manage. Econ.*, vol. 8, no. 3, pp. 315–328, 1990.
[107] T. M. Toole, "Construction site safety roles," *J. Construct. Eng. Manage.*, vol. 128, no. 3, pp. 203–210, 2002.
[108] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., Ser. B (Methodological)*, vol. 58, pp. 267–288, 1996.
[109] W. R. True, "Weather, construction inflation could squeeze North American pipelines," *Oil Gas J.*, vol. 96, no. 35, 1998.
[110] M. T. Van Staveren, *Uncertainty and Ground Conditions: A Risk Management Approach*. Boca Raton, FL, USA: CRC Press, 2006.
[111] L. T. Van, N. M. Sang, and N. T. Viet, "A conceptual model of delay factors affecting government construction projects," *ARPN J. Sci. Technol.*, vol. 5, no. 2, pp. 92–100, 2015.
[112] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.
[113] J. Wu, S. Guo, J. Li, and D. Zeng, "Big data meet green challenges: Big data toward green applications," *IEEE Syst. J.*, vol. 10, no. 3, pp. 888–900, Sep. 2016.
[114] W. M. Chan and M. Kumaraswamy, "Compressing construction durations: Lessons learned from Hong Kong building projects," *Int. J. Project Manage.*, vol. 20, no. 1, pp. 23–35, 2002, doi: 10.1016/S0263-7863(00)00032-6.
[115] J. B. Yang and P. R. Wei, "Causes of delay in the planning and design phases for construction projects," *J. Architectural Eng.*, vol. 16, no. 2, pp. 80–83, 2010.
[116] P. X. Zou, G. Zhang, and J. Wang, "Understanding the key risks in construction projects in China," *Int. J. Project Manage.*, vol. 25, no. 6, pp. 601–614, 2007.
[117] O. Zwikael and M. Ahn, "The effectiveness of risk management: An analysis of project risk planning across industries and countries," *Risk Anal.*, vol. 31, no. 1, pp. 25–37, 2011.

Hakeem A. Owolabi received the B.Sc., M.Sc., and Ph.D. degrees in project analytics.

He is currently an Associate Professor of project analytics and digital enterprise with the Big Data and Analytics Laboratory, Bristol Business School, University of the West of England, Bristol, U.K.



Muhammad Bilal is currently an Associate Professor of big data application development with the Big Data Analytics Laboratory (BDAL), Bristol Business School, University of the West of England, Bristol, U.K.

Lukumon O. Oyedele is currently an Assistant Vice Chancellor of digital innovation and enterprise, and the Chair Professor of enterprise and project management with the University of the West of England, Bristol, U.K.

Hafiz A. Alaka is currently a Senior Lecturer of civil engineering management with the School of Energy, Construction, and Environment, Coventry University, Coventry, U.K.

Saheed O. Ajayi is currently a Senior Lecturer of construction management and architecture with the School of Built Environment and Engineering, Leeds Beckett University, Leeds, U.K.

Olugbenga O. Akinade received the B.Sc., M.Sc., and Ph.D. degrees in big data applications.

He is currently an Associate Professor of big data application development with the Big Data and Analytics Laboratory, Bristol Business School, University of the West of England, Bristol, U.K.