2021-05-13

A Machine Learning-Based Approach for Predictive Analysis of Cost Growth in Heavy Industrial Construction Projects

Tajziyehchi, Negar

UNIVERSITY OF CALGARY

A Machine Learning-Based Approach for Predictive Analysis of Cost Growth in Heavy Industrial Construction Projects

by

Negar Tajziyehchi

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN ELECTRICAL ENGINEERING

CALGARY, ALBERTA

MAY, 2021

Abstract

The construction industry spends billions of dollars on large-scale projects annually. These projects typically experience cost overruns, which differ across regions. For instance, Alberta's average cost growth is much higher than similar projects in the United States. This study focuses on extracting features that influence project cost growth at different phases of Alberta's construction projects, such as front-end planning, detailed engineering, procurement, construction, and commissioning. We analyzed a dataset with a number of features recorded for 139 projects, based in Alberta, between 2003 and 2019. This data is provided by the Construction Owners Association of Alberta (COAA), the Construction Industry Institute (CII) and the University of Calgary. The sample size is relatively small and high dimensional for conclusive analytics, however, the results are promising in developing useful methodologies. In this study, we first applied LASSO regression to reduce the number of features from 281 to 21 features. We then reduced the number of features to 16 based on calculating permutation feature importance using a random forest algorithm. Once we identified the features impacting project cost growth, we developed an interactive tool to illustrate permutation feature importance, partial dependence plot and the editing value of each feature alongside cost prediction. However, the extracted features are primarily from the last two phases of the projects. In order to cover the first three phases of the project, the domain expert recommends adding 29 features to the tool. The tool can help practitioners predict the cost growth of a new project based on the available data at each phase of the project, and see the impact of variations in different features on the overall project cost. The tool also provides more information about the models, and how each feature impacts the project cost growth, so practitioners can invest wisely to minimize the risk of cost overruns.

Acknowledgements

I would like to express my sincere gratitude to my supervisors Dr. Moshirpour and Dr. Jergeas, from whom I have received a great deal of support, guidance and affection. You provide me with guidance to learn how to research and be more confident. I would also like to express my appreciation and gratitude to Dr. Sadeghpour for her valuable advice and kindness. I would like to thank all my previous teachers. And I would like to thank Rebecca for her kind assistance in editing this work. I could not have completed this dissertation without your support. Furthermore, I would like to thank my grandmother for all her sweet wishes to see my success. I would like to thank my parents for their unconditional love and support. I also would like to thank my brother, Kamyar, and my sister, Gelareh, who have been the source of motivation and inspiration for me with their encouragement. And finally, I want to thank my friends who helped me during these two and a half years, making it memorable and much easier. Thank you to Elaheh, Melika, Iman, Mostafa, Abbas, Fatemeh, Reza, Behrad, Shabnam, Masoud, Yousef, Ali, Zohreh, Shohre, Zahra, Mona, Negar, Niloofar, Mihai, Tian, Danny, David, Farhad, Arash and Chargoon friends.

Table of Contents

Abstract . . . ii
Acknowledgements . . . iii
Table of Contents . . . iv
List of Tables . . . vi
List of Figures . . . vii
List of Symbols . . . viii
1 Introduction . . . 1
1.1 Problem Statement . . . 1
1.2 Motivation and Objectives . . . 3
1.3 Research Methodology . . . 4
1.4 Research Novelty and Significance . . . 5
1.5 Findings and Conclusion . . . 6
1.6 Research Contributions . . . 7
1.7 Structure of Thesis . . . 8
2 Background . . . 9
2.1 Artificial Intelligence and Machine Learning in Project Management . . . 9
2.2 Artificial Intelligence and Machine Learning in Project Cost Management . . . 12
2.3 Algorithms . . . 14
2.3.1 Filter Method . . . 15
2.3.2 Wrapper Method . . . 16
2.3.3 Embedded Method . . . 16
2.4 Artificial Intelligence and Machine Learning Tools . . . 18
2.5 Summary . . . 22
3 A Predictive Model of Cost Growth Based on Feature Selection . . . 23
3.1 Case Study . . . 24
3.2 Data Collection . . . 24
3.3 Data Cleaning . . . 26
3.4 Methodology . . . 26
3.4.1 Feature Selection Using LASSO Regression . . . 28
3.4.2 Pearson Correlation Analysis . . . 30
3.4.3 Permutation Importance . . . 30
3.4.4 Classification Results . . . 33
3.4.5 Hyperparameters Tuning . . . 35
3.5 Evaluation Metrics . . . 36
3.6 Prediction and Evaluation of the Feature Selection model . . . 39
3.7 LASSO and Hybrid Method Performance Assessment . . . 39
3.8 Summary . . . 43
4 A Predictive Model of Cost Growth Based on Domain Knowledge . . . 44
4.1 Summary . . . 47
5 An Interactive Tool for Predictive Analysis . . . 51
5.1 Methodology: . . . 52
5.1.1 Manual Feature Selection . . . 53

5.1.2 Permutation Importance 53
5.1.3 Multiple Model Comparisons 55
5.1.4 Data Point Editing 57
5.1.5 Partial Dependence Plot (PDP) 59
5.2 Evaluation 60
5.3 Summary 62
6 Conclusion and Future Work 63
6.1 Summary and Concluding Remarks 63
6.2 Contributions 64
6.3 Limitation 65
6.4 Recommendations for Future Work 66
Bibliography 67

List of Tables

3.1 Regression Model Comparison Using 29 Extracted Features by LASSO Regression 31
3.3 Important Feature Selection 33
3.2 Tuning Hyperparameters 35
3.4 Classification Model Comparison Using 16 Features Selected by Permutation Importance of Random Forest Algorithm 38
3.5 Predictive Regression Model Comparison Using 16 Features Selected by Permutation Importance of Random Forest Algorithm 40
3.6 Predictive Model Comparison: LASSO and hybrid model used in this study 41
3.7 Features and Descriptions 42

4.1 Regression Model Comparison Using Domain Expert Features 48
4.2 Regression Model Comparison Using Features Extracted by LASSO Regression from Domain Expert Features 49
4.3 Comparison of tables 4.1 and 4.2 50

5.1 Description of machine learning algorithms 58

List of Figures and Illustrations

2.1 Feature Selection Methods 15

3.1 Histogram of the Distribution of Cost Growth in Alberta 27
3.2 Bar Chart of Distribution of Cost Growth in Alberta 27
3.3 Method's flowchart 29
3.4 LASSO Path Using all 281 Features 32
3.5 Pearson Correlation 34
3.6 Permutation Feature Importance Using Random Forest Algorithm 36
3.7 Parameter Tuning for Random Forest Algorithm (80-20 train test split) 37
3.8 Parameter Tuning for Gradient Boosting Regression Algorithm (80-20 train test split) 37

4.1 LASSO Path Using 29 Features 46
4.2 Permutation Feature Importance Using Random Forest Algorithm 47

5.1 Manual Feature Selection 54
5.2 Permutation Importance Plot 55
5.3 Tool Diagram 56
5.4 Dense deep neural network with three fully connected, three dropout layers and an output layer with relu activation. 57
5.5 User Input Parameter 59
5.6 Cost Growth Prediction 60
5.7 Partial Dependence Plot 61

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
COAA	The Construction Owners Association of Alberta
CII	The Construction Industry Institute
LASSO	Least Absolute Shrinkage and Selection Operator
RMSE	Root Mean Squared Error
MSE	Mean Squared Error
RF	Random Forest
SVR	Support Vector Regression
GBR	Gradient Boosting Regression
RR	Ridge Regression
KNN	K-Nearest Neighbors Regression
ANN	Artificial Neural Network
PDRI	Project Definition Rating Index
PCA	principal component analysis
GA	Genetic Algorithm
CFS	Correlation-Based Feature Selection
PDP	Partial Dependence Plot
MDA	Mean Decrease Accuracy
MDI	Mean Decrease Impurity
ICE	Individual Conditional Expectation
SA	Sensitivity Analysis
OLS	Ordinary Least Squares

CG	Cost Growth
CA	Actual Cost
CB	Budgeted Cost

Chapter 1

Introduction

1.1 Problem Statement

The heavy industrial sector creates a substantial portion of goods-producing GDP, making them critical to economies. In the United States and Canada, the industry deals with $1,411B and $227B CAD, respectively, of the national goods-producing GDP [1]. Heavy industrial construction projects play an essential role in developing industrial sectors. The timely delivery of heavy industrial construction projects significantly affects the heavy industrial sector economy and its ability to meet demands [2]. Heavy industrial construction is a growing industry used in many sectors, including oil and gas, power plants, pipeline, refinery, and industrial buildings, among others.

One of the main issues of heavy industrial construction is related to massive cost overruns. Cost overruns occur when the overall cost far exceeds the predicted budget [3]. According to the Alberta report III, the average cost growth has decreased from 27% to 9% over the past 17 years. However, there is still potential for improvement. For example, although the mean cost growth for pipeline projects is negative (-6%), 35% of pipeline projects were over budget. Moreover, Alberta's SAGD projects experience a mean cost growth of 22%, up to 106% for some projects [4]. Thus, an investigation of cost overruns and better cost prediction is required to improve the heavy industrial project's timely delivery and profitability. This ultimately enhances investments, the number of projects, and the economy as a whole [5].

Previous research has recommended many ways to tackle this problem, from identifying factors [6] to using the data for cost growth prediction [7]. There are also studies that apply both [8]. A recent study by Haines [9] uses regression models to identify each factor's unique contribution at different project stages and predict the project cost growth of heavy industrial projects. Many

studies have also been conducted in different construction areas applying statistical analysis. Robu [1] ,[5] determined relationships between best practice used in heavy industrial projects and cost growth through pairwise inferential statistics such as the t-test and Pearson's correlation. Gazder et al. [8] identified factors influencing power plant projects' costs in Bangladesh, and applied parametric analysis using t-test, Analysis of Variance (ANOVA) and correlation coefficients. Many studies have been conducted applying machine learning models. Ambrule and Bhirud [10] applied an artificial neural network for pre-design cost prediction of reinforced concrete buildings to overcome cost estimation problems at the initial building construction stages. Another study by Juszczyk et al. [11] investigated the applicability of artificial neural networks to estimate the total construction cost of work for sports fields. Arabzadeh et al. [12] performed data-driven models, including artificial neural networks, regression, and hybrid models, in order to predict spherical storage tank projects' construction costs.

Several construction management research projects deal with high-dimensional datasets by taking advantage of the LASSO regression model to predict and tackle the curse of dimensionality. LASSO regression is capable of dealing with high-dimensional, highly correlated data [13]. For example, Tong et al.[13] used a hybrid method to predict highway cost in a high-dimensional dataset, including a low sample size of 121 highway projects with correlated predictors. After identifying significant predictors, they applied the LASSO regression in order to predict the project cost. The LASSO model also reduced the number of predictors and illustrated the relationship between each predictor and highway project cost. In another study, Zhang et al. employed LASSO regression to predict cost in the pre-construction phase. This study used LASSO regression as a way to address the prevalence of multicollinearity in empirical data.

These studies, which applied LASSO regression for prediction in construction projects, show the advantages of LASSO regression. In this research, we used LASSO regression as part of the feature selection technique, and predicted the cost growth based on other predictive models, in order to enhance the predictive performance.

We first identify factors that have the most impact on project cost growth in order to reduce overruns. More than a thousand features are recorded for one construction project. These projects' high dimensionality characteristic make it difficult to extract useful information to assist practitioners in decision-making before starting, or during a construction project. Therefore, to reduce the number of features to the specifically contributing ones, an interactive tool is developed to determine the importance of each factor, and to represent each of their unique contributions. However, not all information is available at the different stages of the project. Thus, practitioners need to know how each feature's changes can impact the output, and which selection of features results in the highest performance.

1.2 Motivation and Objectives

Estimating and predicting project cost growth is one of the most challenging, and yet essential, stages of the construction project life cycle. Due to the high-dimensional nature of construction projects, identifying which factors impact cost growth can be a challenging task. Furthermore, since construction project data is typically confidential, there is limited available data for each project. Oftentimes there is no recorded information for many features, and their contribution is missed. An efficient and effective way to identify contributing factors is to use feature selection techniques to reduce the number of features to those which effect cost growth the most. Feature selection is the process of choosing a subset from an original feature set based on a specific feature selection criterion that determines relevant features to the predictive modelling problem [14]. Feature selection is necessary for predicting high-dimensional feature sets of a construction project. Since feature selection techniques reduce the number of irrelevant features, as well as the volume of data, they can reduce learning time, improve learning performance, and enhance data visualization and data understanding [14] [15] [16].

Predictive models can be implemented once the relevant and contributing features are identified. Predictive models are often perceived as black boxes that seemed like a significant gap

between research and practice. It cannot be acceptable for stakeholders and practitioners to be predicting the cost growth without first understanding the model and how each feature can impact cost growth. As there are only a limited number of available features in each phase of the project, an effective way to improve the understanding of the model is to develop an interactive tool that predicts cost growth based on these available features. The interactive tool can illustrate prediction performance, presenting the importance of features and their contribution, to make the model more interpretable.

There are three main objectives for this thesis:

- Identifying parameters that mostly affect cost growth.
- Comparing the performance of different machine learning algorithms using parameters determined in the previous step.
- Designing a tool that provides cost growth prediction using a different selection of features. Some of these features are identified by the earlier steps in this study, and the rest are recommended by the domain expert.

Practitioners can use the results of this study to control each identified factor that impacts the project's cost growth, thereby reducing the risk of cost overrun. It is important to understand how each feature affects the cost growth in order to improve the project performance. Therefore, the tool helps to understand the relationship between each factor and the cost growth. The results also help practitioners recognize where to invest, and avoid the risk of running over budget.

1.3 Research Methodology

This research uses a dataset of 139 heavy industrial projects in Alberta. This data was gathered by the collaboration of three organizations to benchmark capital projects. The Construction Owners Association of Alberta (COAA) connects the company representatives who are providing data. The Construction Industry Institute (CII), a research unit at The University of Texas at Austin, provides

an online system for COAA member companies to input project information. A research team of civil engineers at the University of Calgary provides local support to manage data collection in the province.

In this study, the dependent variable is cost growth, which is a standard measure of cost performance in construction projects. LASSO regression is applied to find features that primarily affect the cost growth of heavy industrial projects in Alberta. Random forest calculates the permutation importance of the extracted features. Features with the highest importance are selected, which results in a decrease in the number of features and an increase in model predictive ability. The performance of extracted features is evaluated by five machine learning algorithms: random forest (RF), support vector regression (SVR), gradient boosting regression (GBR), ridge regression (RR), and k-nearest neighbors regression (KNN).

Predictive model performance can vary by considering different sets of features as an input. Therefore, in this study we provide a tool that allows users to select their own features from the 16 variables selected in the previous study (that provide the highest score) and the 29 features selected by the domain expert. One of the responsibilities of designers is to make the model comprehensible for the users. Users can understand which selection of features gives them the highest R^2 score and the lowest RMSE at different project stages. The tool also visualizes the results for all the models used in this study. Users can also see the partial dependence plot for each feature and modify data points to see the output changes.

1.4 Research Novelty and Significance

This study's hybrid model predicts the construction cost growth and identifies important factors that influence cost overruns. Our model reduced the number of features from 281 to 16 contributing features, which not only make the model more accurate but also reduces the training time and leads to a more concise model. Additionally, we compared the prediction performance of six different conventional machine learning algorithms. We demonstrated that the random forest algorithm

achieves the most desirable cost growth estimates based on the R^2 score and RMSE. Furthermore, the research demonstrates that the combination of identifying important factors by LASSO regression, calculating feature importance, and predicting project cost growth based on predictive models such as random forest, outperforms similar research that uses exclusively LASSO regression.

Our tool can produce information about each feature, since most similar real-world datasets with tabular features and continuous dependant variables can rarely investigate each factor's contribution in a predictive model. For instance, a user can interactively evaluate the impact of increasing or decreasing a feature on cost growth using different methods. This tool can automate, facilitate and expedite the cost growth prediction. Several models can be compared to select the least expensive alternative.

1.5 Findings and Conclusion

This research identifies several features connected to the cost growth of Alberta's heavy industrial projects. Calculating the permutation importance of each feature determines important relationships with cost growth. The two most prominent features are the total recordable incident cases and the procurement phase's actual duration. The former represents the importance of safety that might result from the flawlessness of the equipment. The latter illustrates the importance of the procurement phase to the project cost growth.

Furthermore, in this research, a tool is designed to predict the project cost growth based on the available features. However, the highest performance is achieved when using 16 features extracted by feature selection techniques. Practitioners attempting to reduce the project's cost growth can use this tool to determine where to focus and invest, thereby increasing project performance.

We can expand the current research in different ways using the available data or increasing the sample size to other regions. The following ideas can be investigated in further research.

The results of this research are limited to projects in Alberta; thus, in further research, data from other locations could be used. Given more data, we can explore potential advantages of

using deep learning and neural networks, and compare results with the traditional machine learning approaches.

Although the tool gives a good insight into the model to help users make decisions, it could be enhanced. This tool only predicts and analyses contributing factors to the cost growth of projects; this needs to be expanded to cover other project performance areas, such as schedule and safety for example

Moreover, as the tool's design is modular, new algorithms and features can be added in future. The functionality of the tool can also be improved. In this study, the partial dependence plot only supports one feature at a time; however, future studies could plot the PDP for two features to see their interaction.

1.6 Research Contributions

This thesis has three significant contributions.

1. Theoretical contribution
    - Proposed a hybrid approach that can increase the cost performance by considering only the most important features.
    - Identified a few features that can achieve the highest possible performance to predict cost growth from a high-dimensional dataset.
2. Practical Contribution
    - Presented the analytical techniques and tools that can be used in industry to provide an interactive predictive model for large-scale projects. Need to be tested on future project.

1.7 Structure of Thesis

This thesis is organized into six chapters.

Chapter 2 includes four sections. The first section is a literature review of the use of artificial intelligence and machine learning in project management. The second is primarily focused on construction cost management. The third section presents a background of three main feature selection techniques (filter, wrapper, and embedded). The fourth section comprises a literature review, which discusses the artificial intelligence and machine learning tools that are making the models more interpretable.

Chapter 3 describes the methodology applied to identify factors that impact the cost growth of the project. These factors are selected based on two feature selection approaches: LASSO feature selection and calculating permutation feature importance. Five machine learning algorithms are also compared in this chapter.

Chapter 4 applies a similar approach to Chapter 3 to validate the feature selection methods used in this study. In this chapter, the domain expert suggests features to use. This chapter also provides a brief description of different project phases.

Chapter 5 presents the analytical techniques and tools that provide an interactive predictive analysis of large-scale projects using a different selection of features. These features can be from different stages of the project life cycle. The developed tool uses five conventional machine learning approaches and deep learning algorithms to predict project cost growth. These models are developed and compared to allow practitioners to predict the cost growth of heavy industrial construction projects in Alberta using their feature of interest. The importance and the contribution of each feature also can be illustrated using different features such as permutation feature importance, partial dependence plots, and data point editing.

Chapter 6 serves as the thesis's conclusion, which includes a summary of each chapter. This chapter indicates the limitations, contributions and significance of the project. Recommendations also will be offered on how this thesis can be extended for future work.

Chapter 2

Background

In this chapter, the main focus is on the related work in the literature, specifically pertaining to artificial intelligence and machine learning tools and techniques in construction cost management and other similar fields. Furthermore, this chapter briefly explains the background knowledge about different feature selection techniques.

2.1 Artificial Intelligence and Machine Learning in Project Management

The construction project performance literature is primarily divided into two themes. The first, and most common, is the determination of key factors that affect project success or failure [5]. Due to the high dimensionality nature of such projects, research has been conducted using reduction techniques to reduce data to that which contributes to successful factor identification. Reduction approaches are categorized into feature selection [17] and feature extraction [18]. The former is used to reduce the data by selecting a subset of features. The latter reduces the number of features by combining them to create a smaller dataset. The feature extraction approach is commonly used as a pre-processing step for prediction, which is another goal of the construction performance research area.

Project performance prediction benefits from the use of machine learning in different areas, such as cost [6], schedule [19], safety [20], and productivity [21]. Furthermore, a variety of machine learning methods are used to assess project performance prediction, ranging from regression and neural networks, to some hybrid models.

For example, Wang and Gibson [7] investigate the relationship between pre-project planning and project success. In this study, two different machine learning approaches are considered to pre-

dict project success (cost and schedule): statistical regression analysis and artificial neural network (ANN). ANN performs better than statistical regression due to the following: ANN does not need a predetermination of the relationship between input and output; there is no need for the statistical distribution of data; and it can learn and update itself. An interesting outcome of this project is that the Root Mean Square Error (RMSE) in ANN is larger in value than the RMSE in the linear regression model. However, by taking out two projects with outstanding cost growth, the result is reversed. This research used the Project Definition Rating Index (PDRI) as input, and both ANN and simple linear regression models show that PDRI has a positive relationship on project success.

A study by Chua et al. [6] adopts the neural network approach to determine key management factors that influence budget performance. To initially identify factors, 27 project management factors were selected to be trained for the cost performance of the project. During the training, the insignificant factors in the model, identified after several stages, were discarded. That is, the eight highest average percentages of changes in these factors were identified as the most important factors to affect the cost performance. In another study [22], a hybrid model was developed to predict the cost performance of the commercial building projects by integrating principal component analysis (PCA) with a support vector regression (SVR) model. Due to the use of SVR, this model can handle high-dimensional datasets with low samples; however, the linear relation between the input features reduces the model's generalization ability. In this way, via PCA, 35 out of 64 new uncorrelated features were created to improve the generalization ability of the model [22].

When it comes to a higher number of features, the prediction capacity of a machine learning model becomes less accurate, requires more time, and decreases comprehensibility. Therefore, to fully exploit machine learning algorithms, data preprocessing techniques such as feature selection are essential [23].

In the field of safety, Poh et al. [20] used a combination of the Boruta feature selection technique and decision tree, to develop a model for predicting safety leading indicators, and ultimately to predict the occurrence of accidents in construction projects. The Boruta algorithm, which is a

wrapper method built around random forest, helps to reduce the number of features to those that have the most impact on the target [24]. The decision tree was trained using the reduced features and the top nodes of the tree were selected. Using selected nodes, different models were trained with five popular algorithms. Random forest emerged as a reliable model for the estimation of safety risks of construction sites, based on its highest accuracy and weighted-kappa statistics in comparison with other models. In order to reduce the dimensionality of organizational competencies feature space, Tiruneh and Fayek [17] applied a wrapper method on a fuzzy inference system using a genetic algorithm (GA) as a searching algorithm. GA reduced the number of features from 103 to 16, which resulted in a less complex model with higher computational speed, as well as better generalization ability.

Another study [25] developed a model to predict future bridge deterioration, using a combination of continuous and discrete data types, including some irrelevant and redundant data. To address this issue, a discretization technique, along with some feature selection techniques, were used. Discretization is a method of finding a set of cut points in order to transform a continuous numerical feature into discrete intervals. The result shows that the implementation of ChiMerge FD and the LASSO feature selection methods obtained the highest average accuracy improvement. Chi et al. [26] applied data mining techniques to identify key project work functions and their correlation with cost performance. Four types of ranking algorithms were applied (correlation-based feature selection (CFS) algorithms, wrappers for feature subset selection algorithms, information gain attribute evaluators, and symmetrical uncertainty attribute evaluators) in order to select those functions that are highly-correlated with project performance outcome. These selected functions were considered as inputs for the decision tree to analyze the relationship between the key work functions, cost performance, and to predict the best matching project performance. Tsehayae and Fayek [21] applied a correlation feature selection (a kind of filter feature selection) to reduce a feature space, including 176 features, and determine key influencing factors in construction labor productivity.

2.2 Artificial Intelligence and Machine Learning in Project Cost Management

Previous research has recommended many ways to tackle the problem of cost overrun, identifying factors [6] to use the data for cost and cost growth prediction [7]. There are also studies that apply both [8]. Many studies have been conducted in different construction cost areas applying statistical analysis. Robu [1], [5] determined relationships between best practices used in heavy industrial projects and cost growth through pairwise inferential statistics such as the t-test and Pearson's correlation. Gazder et al. [8] identified factors influencing power plant project costs in Bangladesh, and applied parametric analysis using t-test, analysis of variance (ANOVA) and correlation coefficients.

Many studies have been conducted applying machine learning models, from linear regression to artificial neural networks. A recent study by Haines [9] uses regression models to identify each factor's unique contribution at different project stages and predict the project cost growth of heavy industrial projects. Ambrule and Bhirud [10] applied an artificial neural network for pre-design cost prediction of reinforced concrete buildings to overcome cost estimation problems at the initial building construction stages. Another study by Juszczyk et al. [11] investigated the applicability of artificial neural networks to estimate the total construction cost of work for sports fields. Arabzadeh et al. [12] performed data-driven models, including artificial neural networks, regression, and hybrid models, in order to predict spherical storage tank projects' construction costs. Chakraborty et al. [27] proposed a hybrid model for probabilistic prediction of construction project costs throughout the initial design stage using light gradient boosting and natural gradient boosting. They applied game theory (SHapley Additive exPlanations) to make the model more interpretable for practitioners. This method describes each case's prediction by calculating each feature's role using Shapley values from coalition game theory. The Shapley value is calculated by the average marginal contribution of each feature value over all possible subsets of features.

Wang et al. [28] used the Project Definition Rating Index(PDRI) in order to predict cost and schedule success, with a sample size of 92 building projects in Taiwan. Artificial Neural Network

ensemble models and support vector machine classifiers were implemented and tested. The SVM model provides the highest outcome with a total 92% accuracy for the estimation of project cost performance. The adaptive boosting model of the ANNs, with a cumulative accurateness of 80%, offers the highest prediction outcome in project schedule performance. The findings of the study demonstrate that better preparation at the early stage of the sample projects leads to improved final results.

Elmousalami [29] compared 20 machine learning models, based on the mean absolute percentage error (MAPE) and adjusted R2, to emphasize the importance of ensemble models. The study provided a comparison of AI technologies to establish a credible conceptual cost prediction model. Different models were implemented employing tree-based models, ensemble methods ANNs, SVM, etc. The findings showed that XGBoost, with 9.091 and 0.929 percent respectively for MAPE and adjusted R2, were the most reliable and appropriate approaches. However, the results were not interpretable, which was the main drawback of the ensemble approach.

Fan and Sharma [30] present a case study of 26 construction projects, including real estate companies and construction enterprises, to predict the project cost. There are 15 features for each project, which are reduced to nine new features transformed by principal component analysis. When SVM and Least Squares Support Vector Machine are compared, standard SVM accomplished higher prediction accuracy. Zhang et al. (2017) [31] developed a LASSO regularized regression for forecasting the cost of highway construction projects for state highway agencies. They implement a parametric model of cost to predict cost in the initial stages of the project and prevent cost overruns. This study compared the LASSO and ordinary least square model's performance, concluding that the LASSO performs better than the ordinary least square. In a similar study, Tong et al. [13] achieved the same results: that LASSO outperforms OLS in several aspects, including automatic feature selection, the capability to deal with collinearity, and the numerical stability of model prediction. Their research works on a case study of 121 highway projects to find the relationship between predictors and highway engineering budgets using LASSO and ordinary

least square regression. However, they implement a hybrid method that benefits from the use of gray relation analysis and LASSO regression, which is superior to both LASSO and OLS.

Rafiei and Adeli [32] used data of 372 low- and midrise buildings (three to nine stories) to estimate the construction cost of projects. This study's variables include physical and financial real estate units, and economic variables and indexes influencing the construction costs. They proposed an unsupervised deep Boltzmann machine learning approach to derive related features from the input data. Moreover, a back-propagation neural network or support vector machine turns the trained unsupervised model into a supervised regression network. In relevant research by the same authors, a new building unit's sale price was predicted at the design phase or before construction. A nonmating GA was developed to reduce data dimensionality and increase the probability of finding the best subset of features to produce precise estimation results using the GA's global optimization capability [33]. The research goal was to help construction companies decide whether to begin a construction project based on the sale market's magnitude [34].

2.3 Algorithms

The domain knowledge is extremely useful in pruning the data and recognizing the most important features. However, in a number of cases the high dimensionality of the data makes it difficult to exclusively use the available domain knowledge to identify features contributing to the target variable [35]. Therefore, research in different fields such as machine learning, statistics, and pattern recognition, is typically focused on feature selection. Feature selection has applications in many fields, such as chemistry, astronomy, array analysis, plasma physics, and remote sensing [36], and is commonly used in areas where there are numerous features and few samples [37].

High-dimensional datasets usually include features that are not relevant to the target variables. Using all of the features might decrease the performance and lower the speed of a machine learning algorithm [38]. Moreover, the model might suffer from overfitting. These challenges of high-dimensional datasets are called the curse of dimensionality [39]. To solve the curse of dimen-

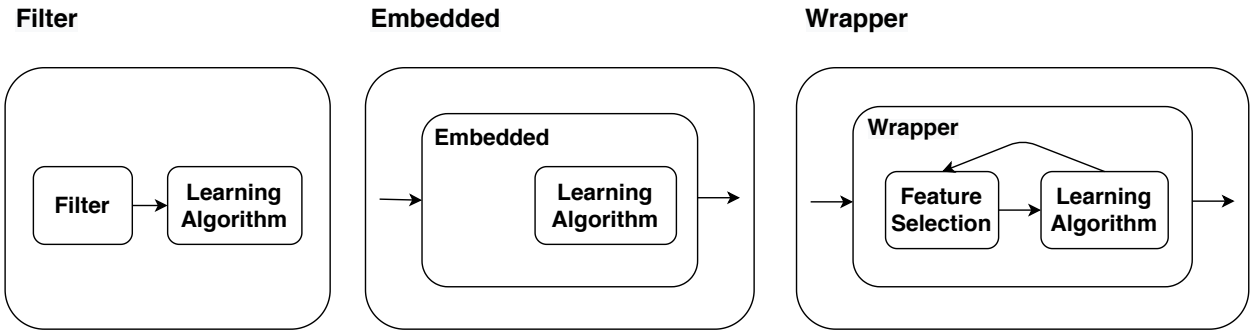

Figure 2.1: Feature selection techniques concerning learning algorithms [43]. Filter methods have no relation with the learning algorithm; on the other hand, wrapper and embedded methods are dependent on the learning algorithm. These methods are commonly used as a preprocessing technique.

sionality, feature selection techniques are useful to choose relevant and non-redundant features to identify a subset of features that improve the performance of a predictive model [20], [14], [40], [41]. Relevant features are those that have information about the target variable [42]. Redundant features are two input features that are entirely correlated [43] [44].

There are three categories of feature selection techniques (filter, wrapper, and embedded models) depending on their relationship with learning methods [14], [45], [37]. Although there are several feature selection techniques, there is no single best method [43].

2.3.1 Filter Method

Filter methods are the earliest approach of feature selection within machine learning; however the selection of features is not dependant on the machine learning algorithm [46][47]. Alternatively, features are selected based on a score assigned to each feature. This score is calculated through a statistical measurement based upon the relevancy or separation of the feature set. The measures might range from correlation measurements to probabilistic independence measurements. The allotted score describes the relative power of the feature set and target value. After the scores are assigned to features, they are ranked and either chosen or eliminated from the feature set [43].

2.3.2 Wrapper Method

In wrapper methods, the feature selection process is based upon a particular machine learning algorithm attempting to fit a given dataset, as illustrated in figure 2.1. An induction algorithm is used to estimate the merit of the feature subset [46]. Wrapper methods assess various models by applying procedures that add or remove features to gain the optimal selection and maximize model performance [48]. There are generally three possible procedures: forward selection, backward elimination and step-wise selection. One of the most popular procedures is backward elimination, which begins the exploration with all features and reduces them by applying a feature scoring model until the optimal subset selection. Forward selection begins with one feature, iteratively added into larger subsets, whereas backward elimination starts with all features and gradually eliminates features. [43].

2.3.3 Embedded Method

Embedded methods are models in which the feature selection process happens directly in the model fitting process. This method includes the feature selection algorithm as part of the machine learning algorithm [49]. Examples of embedded methods include tree-based models, LASSO[50], elastic net [51], etc.

Filter methods are computationally faster and less complex than embedded and wrapper methods; however, the most significant drawback of this method is that it is less accurate than others [17]. In high-dimensional datasets, the time complexity is almost as important as model accuracy [14]. Wrappers usually achieve the highest accuracy, but they are computationally complex and, as a result, slower than the others. Therefore, in this study we take advantage of LASSO feature selection, an embedded method that combines the quality of both [17].

LASSO is a kind of regularized linear regression analysis approach that minimizes the sum of squared residuals. LASSO does feature selection in a way that excludes useless variables from the equation by shrinking the coefficient of some features to zero when the sum of the absolute

values is less than a specific amount [52]. One of the major advantages of LASSO, which results from diminishing and removing the coefficients, is reducing variance with a small increase in the bias. This is helpful in high-dimensional datasets where the sample size is small. Furthermore, the interpretability of the model can be increased as the irrelevant features are removed [53]. LASSO is also useful when there is collinearity among dependent variables. In such cases, LASSO selects one feature and removes others [52].

Different fields of study benefit from applying LASSO regression for feature selection. For example, the Genome-Wide Association Study (GWAS) is applied to recognize the associated genes of Parkinson disease detection. In accordance with GWAS, Rafieipour et al. [37] applied two feature selection methods to choose gene subsets that are involved in the detection of Parkinson's disease. They applied Perturbation-based Feature Selection (PFS) and Hilbert-Schmidt Independence Criterion Lasso (HSIC-Lasso) feature selection methods on the GSE6613 dataset with 55 normal people and 50 people with early-stage Parkinson's. In this work, the chromosome corresponding to selected features by PFS and HSIC-Lasso is evaluated to clarify the role of the chromosome in Parkinson's disease. The feature selection methods are applied on three classifiers: SVM, Random Forest, and AdaBoost. Experimental results show that the PFS method has 86.03% accuracy on the SVM method, whereas HSIC-Lasso shows 94.9% accuracy on SVM.

There are also three categories of feature selection approaches (supervised, unsupervised, and semi-supervised models) concerning utilized training data (labeled, unlabeled, or partially labeled) [14]. In construction, researchers primarily use supervised learning; other fields with similar (high-dimensional) data take advantage of unsupervised learning. In one study, for example, [54], supervised learning is applied over unsupervised learning in a high-dimensional healthcare dataset to find features with the highest importance. More specifically, k-means clustering as unsupervised learning is used to solve the problem of imbalance data, and then a random forest classifier is applied to extract important features.

2.4 Artificial Intelligence and Machine Learning Tools

There are two primary goals defined for predictive modelling. The first involves generating accurate predictions, and the second goal may be the interpretation of the model. In connection with machine learning, interpretability means the ability to make a model understandable for humans [55]. It is also defined as helping to understand the reason behind decisions [56]. However, in order to achieve the highest accuracy, models become more complex, which results in lower interpretability [48]. There are instances for which knowing the performance is sufficient. For example, interpretability is not expected when the prediction result has no critical consequence or is well studied. On the other hand, there are instances that have no room for error, such as financial lending and healthcare, and any machine learning model might result in an incorrect prediction. Practitioners in these domains should be conscious of the process of achieving the prediction results even if the model works properly [57]. In healthcare, for example, machine learning models are often used to help with treatment decisions (which are critical for patients). However, doctors do not easily trust the predictive model, unless they know how predictions are generated, because the consequences are critical for the patient.

Machine learning algorithms usually work in two steps: training and predicting the model. In the first step, learning is done using training data, and in the second, the trained model will predict the unseen data [58]. These trained models are not always interpretable. Although there are some models such as decision trees [59] [60], random forests, and linear regressions that are informative, they cannot always help understand the model. For example, although the decision tree is interpretable, it is still difficult to elicit reasoning for the outcome when the number of nodes increases. Random forests also include several decision trees, which decrease the interpretability of the model. A random forest can provide information on the importance of each feature as well as a linear regression equation, which can determine the global importance of each feature by the weights of coefficients. However, neither can provide information on how the model makes decisions on a specific case [58] [61].

Various tools and visualization techniques have been developed to investigate the process of machine learning decision-making, and many researchers supported the need for having more interpretable models [62], [61]. Although there are many studies that have been done on predictive tools, such research is lacking in the construction industry. This section discusses some of the research that is most relevant to our work. In [63], the what-if tool is based on a perturbation method to help users with interactively probing machine learning models. This tool provides the functionality for users to perform counterfactual reasoning, and explore decision boundaries as well as simulation of model behaviour. The tool also allows users to compare two models and plot an inference score. This model was intended for non-technical users, however it was discovered that some prerequisites needed technical expertise (such as that of computer science students and data scientists, who have some machine learning experience). As a result, the tool was ultimately a better fit for more technical users. Another study by Cheng et al. [64] proposed a decision explorer with counterfactual explanations (DECE) to improve understanding a machine learning system's decision rules. DECE has a similar approach to the what-if [63] tool, focusing on counterfactuals that are separate from the available dataset. It also focuses on the visualization of subgroup counterfactuals in order to analyze a model's decision boundaries. The tool uses tabular data for binary classification problems. DECE has been used in three different scenarios and domains (finance, education, and healthcare) to demonstrate the effectiveness of the tool. The concept of counterfactual explanations was first introduced by Wachter et al. [65] in order to solve an optimization problem. The counterfactual explanation is one of the approaches that provides an actionable and human-friendly explanation as an interpretation method. Counterfactual explanations help users to achieve the desired outcome (a flip can explain this in binary classification) by minimum changes of the input [64].

Using clinical data, Krause et al. [58] developed a web-based interface to combine partial dependence and localized inspection in their pipeline. The prospector supports interactive partial dependence diagnostic, and visualizes it, in order to understand the contribution of each feature

on the dependant variable prediction. In order to visualize PDP, the prospector used a heat-map based design, in which prediction probabilities are encoded based on colours. The PDP includes a histogram of the observed values beneath the PDP to assign importance and validate insights. This tool is designed to allow the simultaneous comparison between multiple models. This comparison would be helpful to determine whether models are underfitted or overfitted. Although the tool has been evaluated on clinical data, it can be used for other domains as well. A partial dependence plot (PDP) is a kind of interpretational tool that illustrates the marginal effect of one or two features on the predicted result, to show the relationship between a dependent variable and a feature when using machine learning algorithms [66]. Moreover, the prospector supports localized inspection to help scientists understand the reason behind each data point's impact on prediction. One of the negative aspects of this prospector is its limitation in handling more than one class. Zhao et al. [57] developed a similar tool (iForest) using partial dependence; however, its focus is primarily on interpreting random forest algorithm from various aspects. The paper explains three categories of interpretation techniques (feature analysis, model reduction, and case-based reasoning) that can be used for random forest algorithm. Partial dependence and features importance rely on the category of feature analysis. Feature importance can be evaluated based on mean decrease accuracy (MDA) or mean decrease impurity (MDI). Unlike MDI (specially designed for tree-based algorithms), MDA is a model-agnostic approach (also applied in our study), which uses permutation to measure how much the performance metric decreases. More decrease in the performance metric indicates the importance of features.

A similar approach to partial dependence is an individual conditional expectation (ICE). ICE visualizes the estimated functional relationship between a feature and the target for each observation. Therefore, each observation is illustrated by a line in the plot, and each line shows prediction changes of the observation based on changes in a feature value. This approach is proposed by Goldstein et al. to extend the PDP [67]. Comparing PDP and ICE, PDP is the average of the observations line in the ICE plot. Although the ICE plot can explore heterogeneous relationships,

revealing more information about each instance and the exceptions (which cannot be understood by the average), it makes the interpretation more difficult than PDP. That said, the ICE plot can also include PDP.

Another similar method to partial dependence is sensitivity analysis (SA), used to measure the impact of changes in output when the values of the features are changed in their range. SA uses mean or median instead of calculating probabilities used in partial dependence for all data points. This method can be used for almost all supervised learning models. Many studies have been done using SA as a feature selection technique [68], [69]. A study by Tamagnini et al. [61] proposed a model that works on binary classifiers and binary features using instance-level explanation to change the label by altering the smallest number of features of an instance's vector. In another study [70], a manifold tool is provided, which offers sophisticated visualization focusing primarily on multiple models comparison. They developed and diagnosed models in three iterative processes (inspection, explanation, and refinement).

In another study, Krause et al. [71] proposed an interactive feature selection tool named INFUSE that uses feature selection techniques to rank features, and uses selected features as an input for the predictive models. Feature selection techniques exclude irrelevant features from the dataset. Choosing the best method is challenging, as there are several feature selection techniques that are not interpretable for the users. To solve this problem, INFUSE helps users to understand how features are ranked by applying different techniques. Thi study uses feature selection techniques including information gain, fisher score, odds ratio, and relative risk. The main goal of INFUSE is to portray the predictive capability of each feature. Yang et al have developed a similar method for visualizing hierarchical feature reduction [72]. The tool is based upon hierarchical clustering, and categorizes features according to their similarities.

2.5 Summary

This chapter presented the relevant literature on the use of machine learning and artificial intelligence in construction management and construction cost management. Moreover, background knowledge of the key concept, such as different feature selection techniques (including filter, wrapper, and embedded feature selection methods), was provided. A comprehensive survey on the existing methodologies to interpret and illustrate predictive models was conducted.

Chapter 3

A Predictive Model of Cost Growth Based on Feature Selection

This chapter outlines the methodology used for the data gathering, processing and analytics of the research.

The construction industry spends billions of dollars on large-scale projects annually. These projects typically experience cost overruns. To solve this issue, it is essential to identify the key factors that contribute to project cost growth. The data provided, by the Construction Owners Association of Alberta (COAA) and the Construction Industry Institute (CII), shows that Alberta's average cost growth is much higher than similar projects in the United States. It is therefore desirable to improve Alberta's project performance. There are 139 samples of projects in Alberta, and the data is high-dimensional, making it difficult to extract useful information for cost growth prediction. Dimensionality reduction techniques, such as feature selection, contribute to obtaining the most important features impacting cost growth. This study uses two steps to identify 16 out of 281 significant features. Initially, 21 features were determined by Lasso. The R^2 score and RMSE are calculated for five different models in three train and test split models. Random forest had the highest score, using more than 80 percent of data for training. The permutation importance of each feature is calculated using random forest, and 16 variables are then extracted. These features are applied as an input for five machine learning algorithms to evaluate the variables' predictive ability.

3.1 Case Study

A case study of a dataset containing over 200 feature dimensions reveals how our method significantly improves the effectiveness of feature selection techniques in high-dimensional datasets. The projects used in this study are heavy industrial construction projects based in Alberta. The dataset comprises 281 feature dimensions and 139 instances. A considerable amount of the data (more than 70 percent) comes from projects in the oil and gas sector, such as oil sands SAGD, pipeline, oil sands mining/extraction, oil and gas exploration, oil refining, and oil sands upgrading and tailing. A smaller proportion of the data is provided by non-oil and gas projects, including chemical manufacturing, electrical generating and other non-oil and gas projects. Therefore, the study results address these many kinds of heavy industrial projects.

3.2 Data Collection

The data used in this study was the result of a partnership between three organizations to benchmark capital projects. The Construction Owners Association of Alberta (COAA) connects the company representatives who provide data. The Construction Industry Institute (CII), a research unit at The University of Texas at Austin, provides an online system for COAA member companies to input project information. A research team of civil engineers at the University of Calgary provides local support to manage data collection in the province.

This thesis uses a dataset collected from heavy industrial construction projects in Alberta in three phases starting from 2005; however, the collected data is representative of projects as early as 2000. Three industry reports have been produced based on this dataset, including projects through to 2019: COAA 2009 [73], COAA 2014 [3], and COAA 2019 [4]. Unlike these industry reports, which are primarily focused on the Alberta heavy industrial construction sector [9], the goal of this thesis is to provide a relatively broad application.

The Alberta companies collaborating with the research team were instructed and guided before, and throughout, the process of collecting data and recording information from their projects. In

some cases, the companies also received the necessary training to add the project data directly to the database system maintained by CII, with the research team guiding them throughout the process. In other instances, access to internal company files was granted to the Schulich School of Engineering research team to proceed with the data collection and record the projects' data within the database [9].

Whether the data was collected by the research team or by the companies themselves, it was important that the recorded information from different projects and companies was comparable. Therefore, industry partners helped create a strict set of definitions to ensure existing project documentation could meet the new requirements. These data collection requirements and definitions were provided to companies to prevent inconsistencies. For example, there can be different interpretations of whether labour hours qualify as direct or indirect work; this definition is included in COAA (2019) and explains the complete requirements of each type of employees' hours [9].

The data was validated after it was entered into the database using the same process developed and utilized by the CII since the 1990s. Schulich School of Engineering representatives performed the validation, using this predetermined process, to ensure all required information was collected. They also evaluated the accuracy and rationality of the data, in addition to confirming their compatibility with the pre-established definitions. This process was followed by a final validation step performed by a third party at CII to guarantee that the data is reliable and meets all requirements.

The sample size is small, despite collecting data over a period of 16 years from 2003 to 2019, due to the difficulties in obtaining data on construction projects. Construction projects usually continue for a long time, but they are infrequent, and there are aspects that are not easy to measure. Confidentiality and privacy are other concerns that limit the availability of data [5]. Consequently, much of the research in this area suffers from insufficient data.

3.3 Data Cleaning

This study uses a sample size of 139 heavy industrial projects. All projects are based in Alberta, and the project's data was obtained during three phases. The data is a mix of continuous and discrete variables, including redundant and irrelevant variables. The first step of data pre-processing, which is data cleaning, identifies categorical and numerical data, and transforms the categorical data to numerical. For example, in terms of project type, zero is dedicated to oil and gas projects and one is devoted to non-oil and gas projects. A similar approach is used for other categorical variables. According to a previous study [9], some variables are changed. For example, although there are different project types in the original dataset, we only consider them as "oil and gas" and "non oil and gas" projects. This dataset comrpises 1206 features. The percentages of missing values for each feature are calculated, and those with more than 70 percent of missing values are removed, reducing the number of features to 281. These features have very different magnitudes from one another, and contain significant outliers. Each feature is therefore scaled and transformed to a number between zero and one to prepare the feature selection dataset. The two outliers of the cost growth are detected and removed. The cost growth distribution is shown in figures 3.1 and 3.2. The distribution of cost growth is skewed to the right, and the values fall between approximately -0.5 and 0.98. According to figure 3.2, although about 50 percent of projects have negative cost growth, more than 40 percent of projects experience high (10-20 percent) and very high (20 percent or more) cost overruns.

3.4 Methodology

In this study, the dependent variable is cost growth, which is a standard measure of cost performance in construction projects. This variable is not part of the original dataset and it is calculated by the following equation, where CG is Cost Growth, CA is Actual Cost and CB is Budgeted Cost.

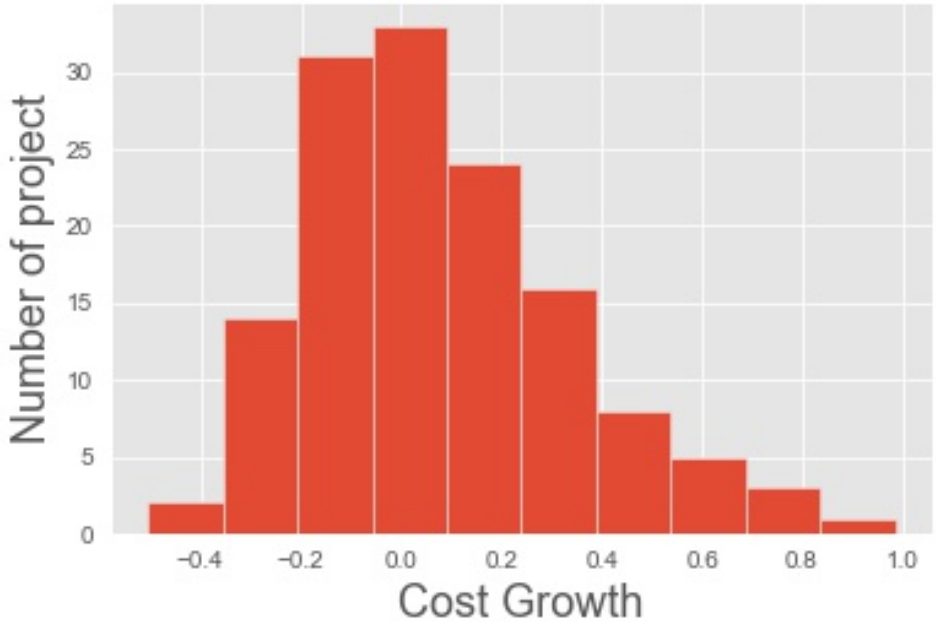

Figure 3.1: Distribution of Cost Growth in Alberta. The peak of the data happens at about zero to 0.1 cost growth. The cost growth data are right-skewed in this histogram, and the data spread is from approximately -0.5 to 0.98.

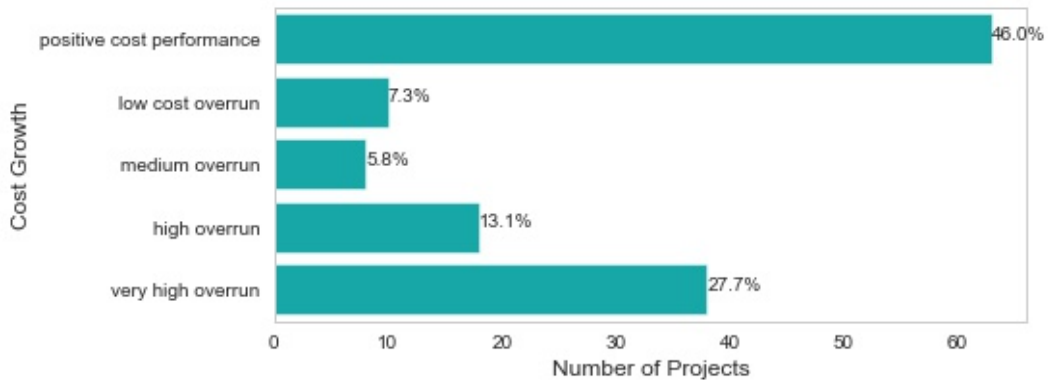

Figure 3.2: Bar Chart of Distribution of Cost Growth in Alberta. The bar chart compares the cost growth of different construction projects. The cost growth is divided into five groups: Positive cost performance (projects with negative cost growth), low cost overrun (0–5%), medium overrun (5–10%), high overrun (10–20%), and very high overrun (20% and more). Positive cost performance group is the most common, followed by very high overrun, high overrun, and others. The cost growth range from about -0.5 to 0.98.

$$CG = \frac{CA - CB}{CB} \tag{3.1}$$

Based on a previous study [9], several highly correlated variables in this dataset need to be eliminated. Otherwise, fluctuations in one variable influence other associated independent variables, leading to an increase in prediction variance and a decrease in prediction accuracy. The LASSO Regression will effectively solve this dilemma [13].

In this study, LASSO regression is applied to find features that primarily affect the cost growth of heavy industrial projects in Alberta. Random forest calculates the permutation importance of extracted features. Features with the highest importance are selected, which results in a decrease in the number of features and an increase in model predictive ability. The performance of extracted features is evaluated by five machine learning algorithms: random forest (RF), support vector regression (SVR), gradient boosting regression (GBR), ridge regression (RR), and k-nearest neighbors regression (KNN).

3.4.1 Feature Selection Using LASSO Regression

First, we use LASSO regression to extract features that have the highest effect on cost growth. LASSO works well if there are a large number of features in the dataset. Using LASSO regression, features that are not relevant to the cost growth will result in zero coefficients by applying the LASSO penalty (L1 penalty) that penalizes the sum of their absolute values. A tuning parameter λ transforms all coefficients and controls the amount of shrinkage. If λ is equal to zero, it works like ordinary least squares (OLS) that overfitted the data as there is collinearity in the input features [52].

In this study, a value of 0.0062 is chosen for λ (this has the highest R^2 score among other λ values) through cross-validation that zeroes the coefficients of unimportant features. Features with more eminent coefficients are considered more important. Figure 3.4 describes the LASSO path, showing the coefficient of each parameter according to changes in λ parameter. Each colored

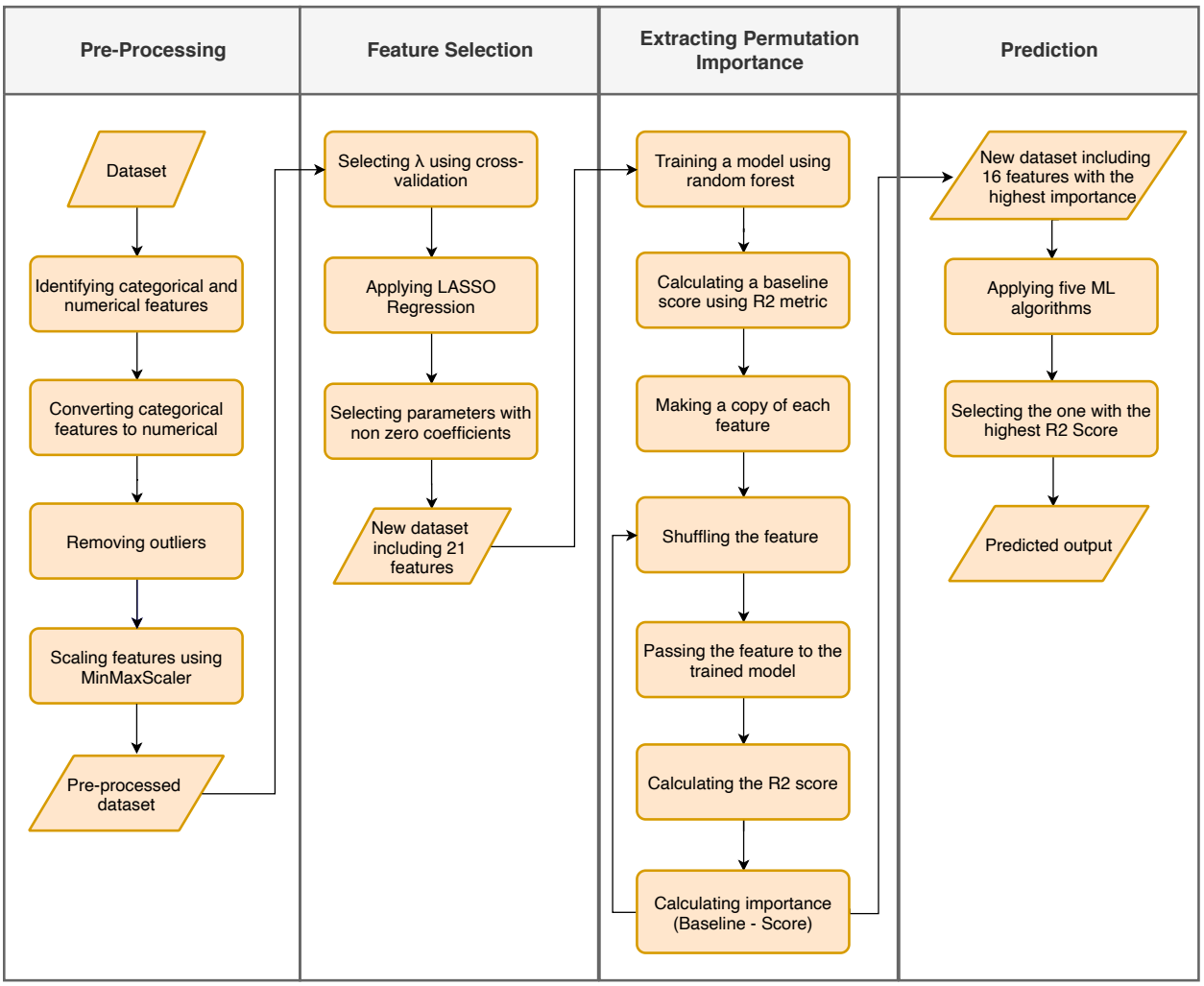

Figure 3.3: Method's Flowchart, describes the method in four phases including (1) pre-processing, (2) feature selection, (3) extracting permutation importance, and (4) prediction. In the first step, two outliers of the dataset are removed, and 137 construction projects are used for the second step. In the second step, redundant and irrelevant features are excluded, and the number of features is reduced to 21. In the third step, 16 important features are selected and used for predicting project cost growth. In the final step, five conventional predictive models are compared, and the random forest algorithm achieved the highest performance

line in the plot represents a regression model parameter. The y-axis shows the coefficient of each parameter. When the λ is zero, we have all of the features in the model; however, as the λ grows, fewer parameters have non-zero coefficients. The dotted line in the plot represents 21 non-zero features when the λ is 0.0062. These features are used as an input for five machine learning algorithms. Table 3.1 reports the R^2 score and Root Mean Square Error (RMSE) of all these models using LASSO regression features. The scores represent a mean after ten iterations (the random state has changed each time). Since there are so few samples, and the results of these two metrics vary significantly based on different proportions of data used for training models, models are trained with 70, 80, and 90 percent of data separately, and tested on the remaining unseen data. Although RMSE is reduced, by increasing the number of training data from 80 to 90 percent, there is no increase in the percentage of R^2 score. RF and SVR have better results than other algorithms. RF achieved the highest R^2 score when we trained the model with 90 percent of samples. Therefore, RF is selected to calculate the permutation importance of features.

3.4.2 Pearson Correlation Analysis

After applying LASSO regression, Pearson's correlation test evaluates the linear correlation between the two variables. The coefficient values of -1 to 1 indicating power, and the sign indicating the direction of the relationship. According to a previous study by [9], there is multicollinearity among independent variables, and some features are highly correlated. Applying LASSO regression, correlated features are removed from the equation. According to figure 3.5, the correlation between independent variables is less than 0.53.

3.4.3 Permutation Importance

We take advantage of permutation importance to increase the performance of the final model. To find feature importance, each is permuted to make noise from the same distribution. Then the R^2 score of the model is calculated. A decrease in prediction score before and after the permutation is considered the permutation importance of a feature. In this study, random forest algorithm is

Table 3.1: Predictive Regression Model Comparison Using 29 Extracted Features by LASSO Regression. Compares the R^2 score and RMSE of five conventional models in three different distribution of training and test data in order to select a model to calculate feature importance

Train-Test-Split	**90% train-10%test**			**80% train-20%test**			**70% train-30%test**		
Model R2 Score (LASSO Features)	Train	Test	RMSE	Train	Test	RMSE	Train	Test	RMSE
Support Vector Regression	0.59	0.35	0.1964	0.60	0.42	0.2024	0.61	0.35	0.2155
Random Forest	0.88	0.42	0.1940	0.88	0.42	0.2070	0.87	0.30	0.2234
Gradient Boosting	0.87	0.37	0.2027	0.79	0.37	0.2161	0.80	0.30	0.2229
Ridge Regression	0.61	0.34	0.1962	0.62	0.41	0.2047	0.63	0.38	0.2107
K-Nearest Neighbors	0.99	0.27	0.1631	0.99	0.28	0.1752	0.99	0.25	0.1819

used to calculate the importance of features. The importance of features can be calculated through impurity-based feature importance [74] or permutation importance. The impurity-based might result in considering high values for features that are not predictive of the dependent variable. Therefore, we calculate permutation importance using random forest algorithm. According to figure 3.6, removing each of the first 16 variables results in a decrease in performance. The greater the decrease, the more important the feature. According to table **??**, the first 16 variables that create the highest R2 score and the lowest RMSE are selected to predict project cost growth. The description and ranges of the 16 variables are shown in table 3.7. There are also negative values for permutation feature importance. In these instances, the shuffled data predictions appear to be more accurate than the actual data. This occurs when the feature is not important, however the predictions on noisy data can accidentally be more accurate. This is common when the sample

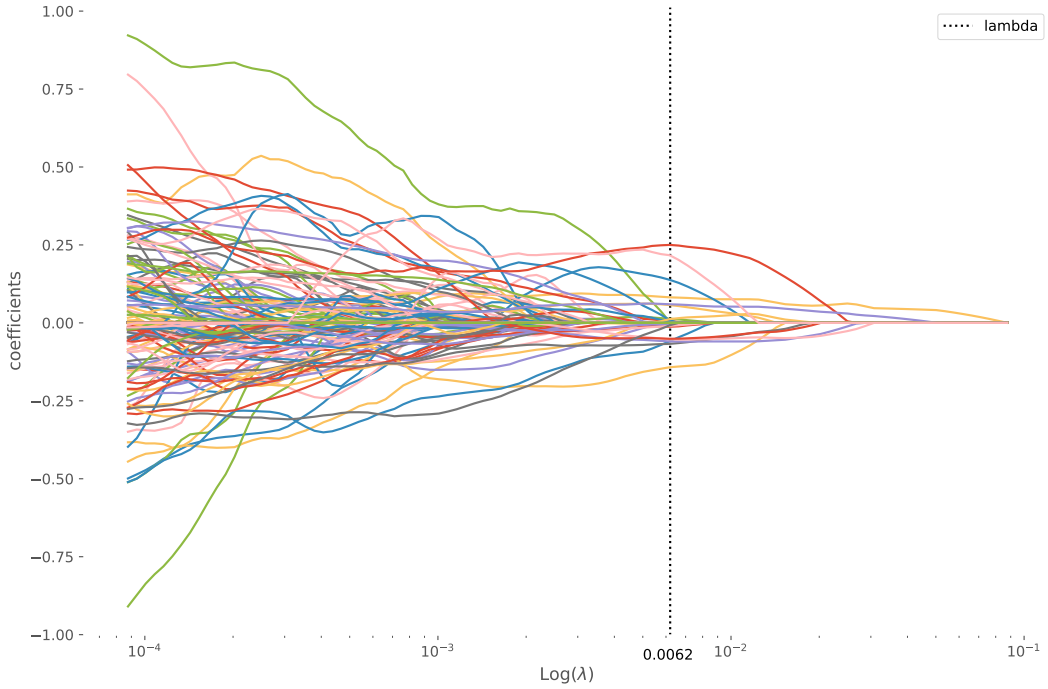

Figure 3.4: LASSO Path Using all 281 Features. The figure shows the LASSO path for the cost growth of construction projects. There are 281 features entered the model, and by increasing the λ, features coefficients decreased to zero. The dotted line cut 21 features that are likely to create the highest model performance.

size is small.

Number of Features	Train R^2 Score	Test R^2 Score	RMSE
1	0.38	0.12	0.2555
1-2	0.69	0.17	0.2462
1-3	0.72	0.20	0.2424
1-4	0.80	0.40	0.2100
1-5	0.81	0.41	0.2076
1-6	0.82	0.42	0.2048
1-7	0.86	0.41	0.2075
1-8	0.86	0.42	0.2061

1-9	0.86	0.42	0.2051
1-10	0.87	0.42	0.2057
1-11	0.87	0.42	0.2064
1-12	0.87	0.42	0.2054
1-13	0.87	0.42	0.2057
1-14	0.87	0.428	0.2051
1-15	0.88	0.421	0.2063
1-16	**0.88**	**0.429**	**0.2050**
1-17	0.88	0.420	0.2064
1-18	0.87	0.427	0.2055
1-19	0.87	0.427	0.2056
1-20	0.88	0.428	0.2051
1-21	0.87	0.409	0.2089

Table 3.3: Important Feature Selection: The table illustrates that the selection of variables with the positive importance has the highest R^2 score and the lowest RMSE

3.4.4 Classification Results

In this section, the cost growth is divided into five groups to make the problem classification. These groups are divided based on the percentage of the cost growth. Positive cost performance (i.e., projects that have negative cost growth values), low cost overrun (0–5%), medium overrun (5–10%), high overrun (10–20%), and very high overrun (20% and more). Table 3.4 represents the results of the 16 features selected by regression models. The performance of KNN increased the most and Ridge Classification (RC) provides the highest accuracy among others. After RC, Gradient Boosting Classification (GBC) has the highest performance. Although classification models do not provide an exact number of cost growth, they can provide a more accurate range (i.e. to

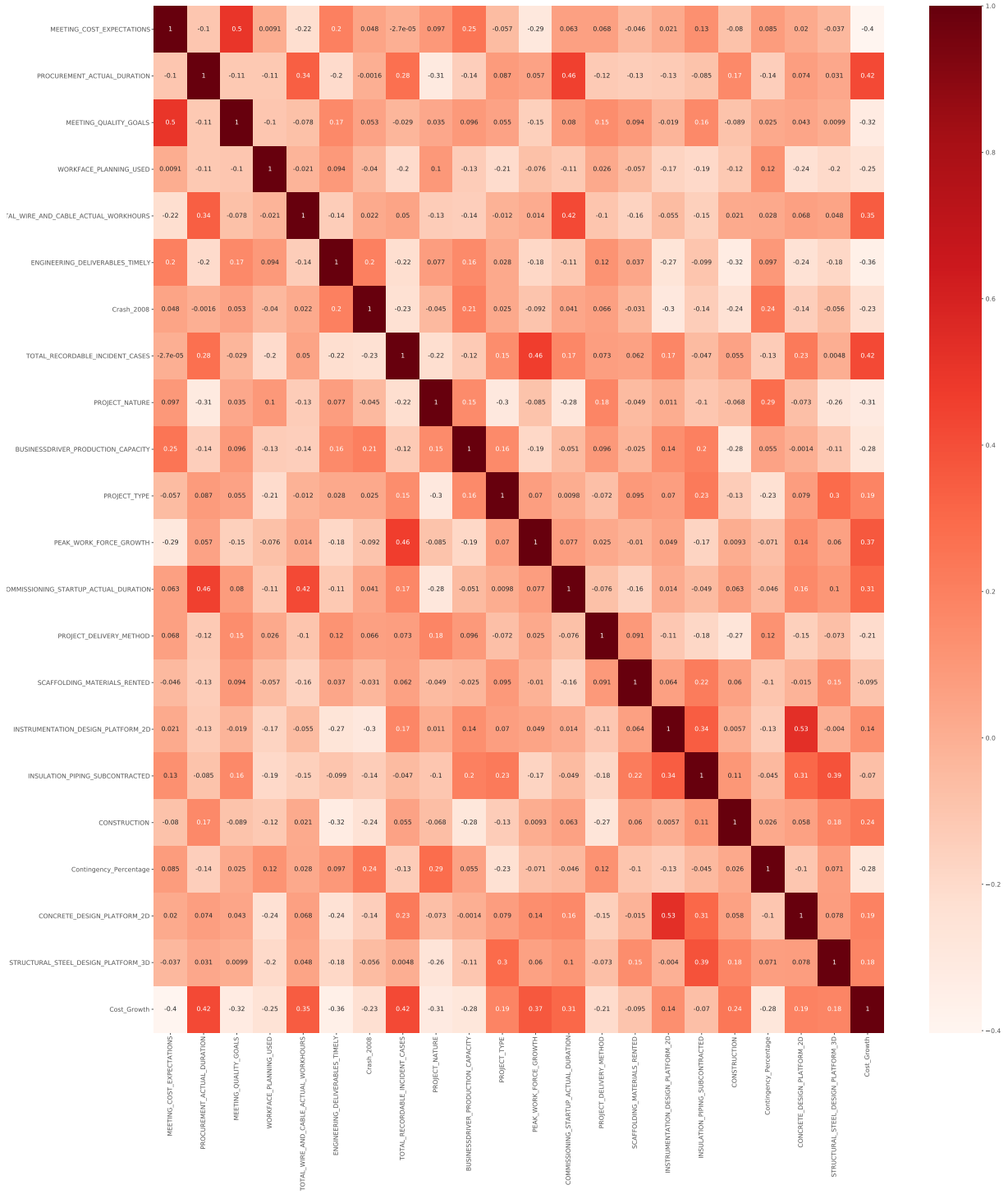

Figure 3.5: Pearson Correlation

Table 3.2: Tuning Hyperparameters
Bold values are selected as tuned parameters
n_estimator for both RF and GBR represent the number of trees to be used in the forest
n_neighbors is the number of neighbors to be used by k-neighbors regressor

Trial	**RF** (n_estimator)	R^2 **score**	**GBR** (n_estimator)	R^2 **score**	**KNN** (n_neighbors)	R^2 **score**
1	40	0.52	10	0.42	9	0.32
2	50	0.55	20	0.49	**10**	**0.37**
3	**57**	**0.56**	**31**	**0.51**	11	0.35
4	65	0.56	35	0.50	12	0.34
5	75	0.54	40	0.48	13	0.31

predict whether a project experience medium or high cost overrun) which can be practical at the first stages of a project.

3.4.5 Hyperparameters Tuning

Tuning parameters are of paramount importance in attaining the model's best results. Tuning parameters are also crucial as they can help to prevent the overfitting that occurs when a model learns the training data too well (even learning noises, for example); although the model can work well on training data, it has little ability for generalization. In this study, to avoid overfitting of the tree-based models (RF and GBR) [75], we adjusted the max_depth of eight and three, respectively. For SVR models, linear kernel performs better than other kernel types (polynomial, sigmoid, etc.) in this problem. According to figure 3.7, after a certain number of n_estimatores, increasing the number of estimators does not have a big impact on R^2 score. Therefore, a value of 57 is chosen in order to optimize the parameter. Based on figure 3.8, a higher performance for the GBR algorithm is achieved by setting the n_estimator between approximately 25 and 35 (these hyperparameters are chosen when the random state is equal to zero). Table 3.2 outlines other hyperparameters that are more precisely tuned and optimized.

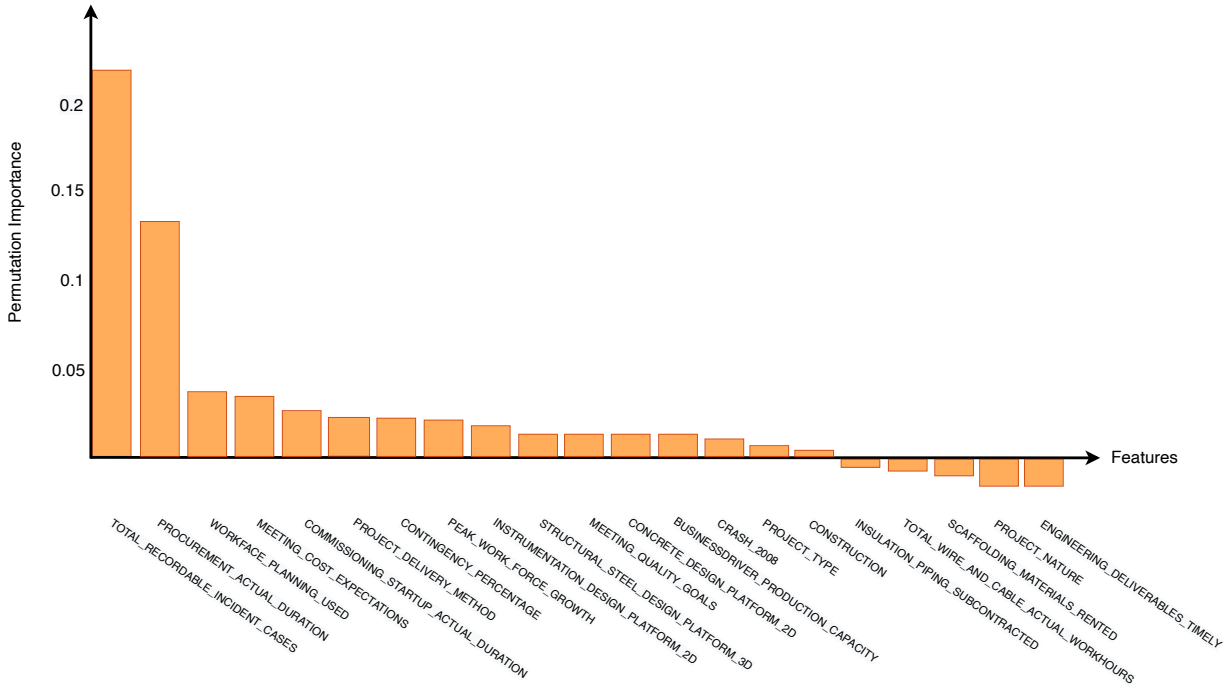

Figure 3.6: Permutation Feature Importance Using Random Forest Algorithm. Illustrates the importance of each of the 21 features for predicting cost growth with a random forest model. The most important features are total recordable instance cases and procurement actual duration associated with the highest error after permutation.

3.5 Evaluation Metrics

Since cost growth is a numerical variable, coefficient of determination is used to evaluate the predictive capability of the models. Coefficient of determination (commonly known as R^2 score) is a common measure of performance that applies on regression problems [76] [77]. The best possible R^2 score would be equal to +1; however, achieving this score is not realistic with the limited number of samples used in this study. The coefficient R^2 is calculated by equation 3.2, where U is the residual sum of squares and V is the total sum of squares. This metric reflects the variability of the dependent variable that has been explained by the model's independent variables [78].

$$R^2 = 1 - \frac{U}{V} \tag{3.2}$$

Another metric used to compare models is the root mean squared error (RMSE). This met-

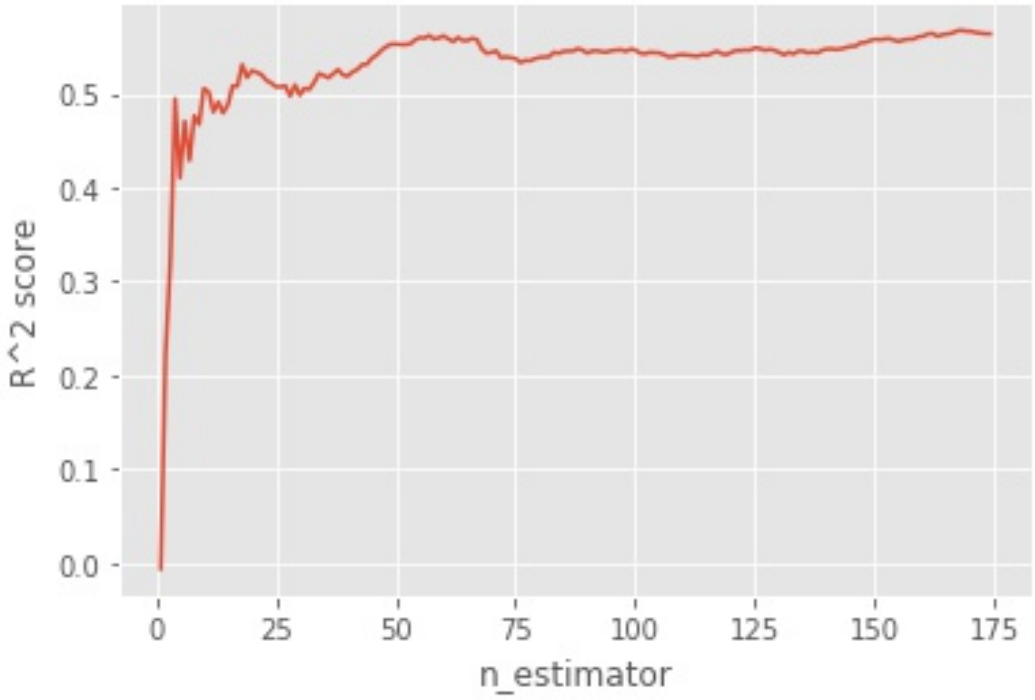

Figure 3.7: Parameter Tuning for Random Forest Algorithm (80-20 train test split)

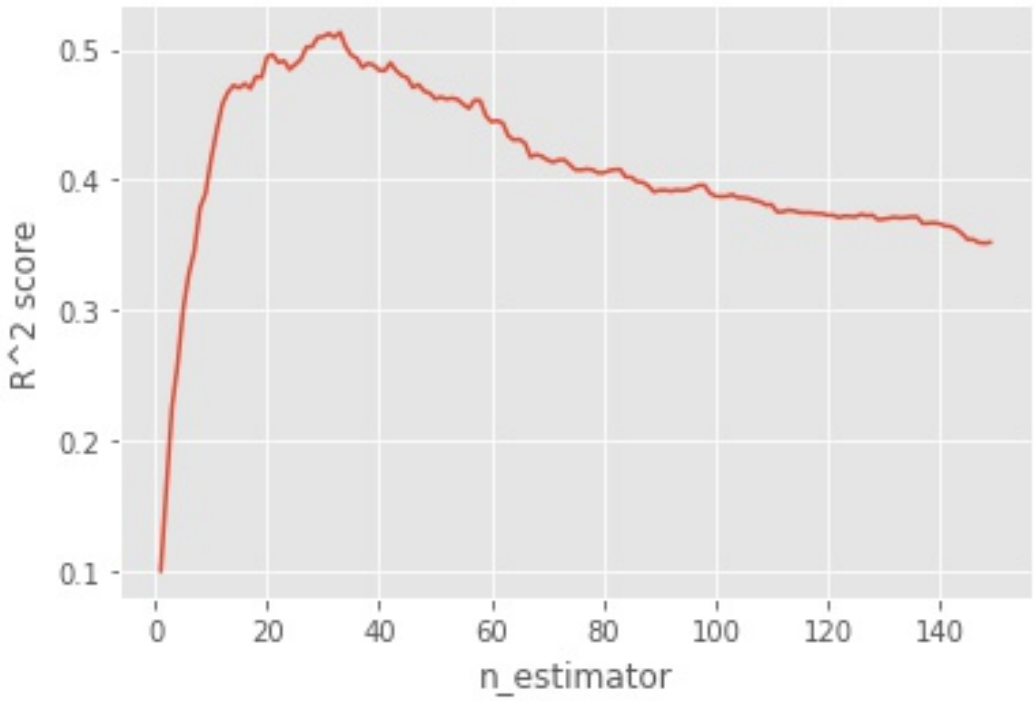

Figure 3.8: Parameter Tuning for Gradient Boosting Regression Algorithm (80-20 train test split)

Table 3.4: Predictive Classification Model Comparison Using 16 Features Selected by Permutation Importance of Random Forest Algorithm. Compares the accuracy of five conventional models in three different distribution of training and testing data in order to select a model to calculate feature importance

Train-Test-Split	**90% train-10%test**		**80% train-20%test**		**70% train-30%test**	
Model Accuracy (RF Features)	Train	Test	Train	Test	Train	Test
Support Vector Classification	0.66	0.64	0.67	0.60	0.68	0.61
Random Forest	0.98	0.58	0.99	0.56	0.99	0.54
Gradient Boosting	0.82	0.62	0.83	0.58	0.85	0.58
Ridge Classification	0.69	0.65	0.70	0.64	0.70	0.61
K-Nearest Neighbors	0.62	0.57	0.61	0.55	0.60	0.55

ric represents the square root of mean squared error (MSE) [76], which measures the difference between predicted values and actual values. In terms of RMSE, the lower the better, since zero is optimal. RMSE is the same unit as the output which makes it more interpretable than MSE. Equation 3.3 describes how RMSE is calculated. In this equation, y_i is the measured value, $\hat{y}_i$ is the predicted value and N is the number of observations. Both metrics are the standard measures of performance in regression problems.

$$RMSE = \sqrt{\frac{1}{N}\sum_{1}^{N}(y_i - \hat{y}_i)^2} \tag{3.3}$$

3.6 Prediction and Evaluation of the Feature Selection model

Our proposed method of feature selection is evaluated by applying the five previously identified machine learning algorithms (SVR, RF, GBR, RR, and KNN) to evaluate whether the extracted features are predictive. We decided to use SVR, as it is one of the most popular regressors in construction problems. RF and GBM were chosen because they use boosting that provides a more precise prediction of the dependant variable [79].

The result of the previous stage was 16 variables. We split the data into three stages of training and testing, and five models were selected and trained with these features.

Table 3.5 reports the R^2 score and RMSE of each model using different sample sizes. Models were trained with 16 parameters extracted by random forest after permutation. Comparing tables 3.1 and 3.5, using features that are extracted after permutation on random forest algorithm, the score for almost all models (80-20 and 70-30 train-test-split) except KNN has increased. As with the previous stage, random forest has the highest score. Using all 281 features to create a model, the performance was reduced from seven to 41 percent based on R^2 scores. After training models with only 16 extracted features, the R^2 score of all models had increased. For example, the performance of SVR increased about 20 to 29 percent in different splits (i.e from 18 to 40 in 80-20 train-test-split); RR increased about 7 to 18 percent; GB increased about 13 to 15 percent; KNN increased about 7 to 10 percent; and RR that has non-predictive ability provides a R^2 of 41 percent. RF and KNN have the least change, whether using all features or just extracted features; however, RF has far better performance than KNN. Comparing the results, RF works better than the other algorithms when there is a high dimensional dataset.

3.7 LASSO and Hybrid Method Performance Assessment

Elmousalami's review [80] of 70 related studies noted that hybrid models are the most commonly applied techniques, and the common parametric cost prediction modelling trend. Many studies employed hybrid models and used LASSO regression as a predictive model [13] [31]. In this

Table 3.5: Predictive Regression Model Comparison Using 16 Features Selected by Permutation Importance of Random Forest Algorithm. Compares the R^2 score and RMSE of five conventional models in three different distribution of training and testing data in order to select a model to calculate feature importance

Train-Test-Split	**90% train-10%test**			**80% train-20%test**			**70% train-30%test**		
Model R2 Score (RF Features)	Train	Test	RMSE	Train	Test	RMSE	Train	Test	RMSE
Support Vector Regression	0.58	0.39	0.1896	0.55	0.40	0.2080	0.55	0.39	0.1896
Random Forest	0.88	0.43	0.1913	0.88	0.43	0.2052	0.88	0.32	0.2191
Gradient Boosting	0.80	0.39	0.1964	0.82	0.43	0.2048	0.83	0.33	0.2186
Ridge Regression	0.57	0.38	0.1916	0.58	0.41	0.2052	0.57	0.38	0.1916
K-Nearest Neighbors	0.99	0.23	0.1659	0.99	0.26	0.1783	0.99	0.25	0.1820

study, however, LASSO is applied as part of a feature selection technique, followed by calculating permutation importance, and the predictive models are trained using different conventional models.

As shown in Table 3.6, the testing R^2 score of the hybrid model is 5 to 17 percent more accurate than the LASSO in different distributions. For example, the testing R^2 score of the LASSO and hybrid models are 30 and 43 percent, respectively, and the training R^2 scores are 51 and 88 percent in 80-20 training and testing distribution. Moreover, the RMSE has decreased in all distributions using the hybrid model. As an example, the RMSE score of the LASSO and hybrid models are 0.2256 and 0.2052, respectively in 80-20 training and testing distribution, indicating that the error has decreased. The same results are achieved in all distributions of the training and testing division. The selection of LASSO variables, followed by calculating permutation importance and predicting the cost growth using conventional methods preliminarily, can improve prediction performance.

Table 3.6: Predictive Model Comparison: LASSO and hybrid model used in this study

Train-Test-Split	**90% train-10%test**			**80% train-20%test**			**70% train-30%test**		
Model	Train R2 Score	Test R2 Score	RMSE	Train R2 Score	Test R2 Score	RMSE	Train R2 Score	Test R2 Score	RMSE
LASSO Regression	0.50	0.26	0.2201	0.51	0.30	0.2256	0.53	0.27	0.2254
Hybrid Model (Random Forest)	0.88	0.43	0.1913	0.88	0.43	0.2052	0.88	0.32	0.2191
Comparison	+0.38	**+0.17**	-0.0288	+0.37	**+0.13**	-0.0204	+0.35	**+0.05**	-0.0135

Features	**Description**	**Ranges**
Total Recordable Incident Cases	Total number of recordable incident cases	0 to 157
Procurement Actual duration	The actual duration of the procurement phase	25 to 1948 days
Workface Planning Used	Was Workface Planning used in this project?	Yes or No
Meeting Cost Expectations	Indicates the overall success of a project in terms of meeting cost expectations	1 to 7
Commissioning Startup Actual Duration	Commissioning and startup actual duration	1 to 865 days
Contingency Percentage	Percentage added as contingency to the budget	0 to 0.244
Peak Workforce Growth	Peak workforce number of employees from planned to actual	-0.469 to 1.2

Project Delivery Method	A project delivery method is a system employed by an organization or owner to design, construct, operate and maintain services for organization and finance.	Parallel primes, traditional design bid build, design build or epc, cm at risk, Other
Instrumentation Design Platform	Which design platform was used for this category in this project?	2D or 3D
Structural Steel Design Platform	Which design platform was used for this category in this project?	2D or 3D
Meeting Quality Goals	Indicate the overall success of the project in terms of meeting quality goals	1 to 7
Concrete Design Platform	Which design platform was used for this category in this project?	2D or 3D
Crash 2008	Before or After the 2008 Financial Crash	Before or after
Project Type	Project type means whether or not a project is of the type of oil and gas	Oil and Gas, and non oil and gas
Construction	Whether a project contract is lump sum or not	Lump Sum or other

Table 3.7: Features and Descriptions

3.8 Summary

Prior to implementing a model to predict project cost performance, it is vital to select a relevant and informative subset of features that are considered most useful to predict the dependant variable. This chapter outlined two approaches for the selection of contributing features to cost growth. The first is LASSO, which is a kind of embedded method. Using LASSO, we reduced the number of features from 281 to 21 features. Next, we calculated the permutation importance of features in order to reduce them to 16. These methodologies are evaluated based on different conventional machine learning algorithms.

Chapter 4

A Predictive Model of Cost Growth Based on Domain Knowledge

In this chapter, the same approach as the previous chapter is applied to another case study, limiting features recommended by the domain expert, to highlight the impact of feature selection on different volumes of data. Based on the domain expert's recommendation, 29 features from the first three phases of the construction project are selected. There is a tremendous degree of uncertainty in project cost estimation in the initial stages of construction projects, and therefore considerable interest in developing an efficient method to overcome such uncertainty[12].

Robu [5] provides the following short description of each project phase. Heavy industrial projects are comprised of five separate phases – front-end planning, detailed engineering, procurement, construction, and commissioning – each of which has different objectives and natures. The study of cost overruns based on phases can reveal more information on which phases cause cost overruns in the entire project, and subsequently where revisions can be made.

The initial phase of a construction project life cycle is front-end planning. This phase is similar to the commissioning phase (the last phase of a project) in terms of the proportion of the project funds; however, it encounters less average cost growth. The cost of the front-end planning phase is notably correlated with decreased construction cost growth. More investment in the front-end planning phase results in reduced construction cost.

The second phase of the construction project life cycle is detailed engineering, which has the second highest cost growth. About 13 percent of the project cost budget is dedicated to this phase. Many factors are involved in these cost overruns, such as unexpected site conditions or constructability, and inadequate scope definition.The third phase, which has the lowest average

cost growth, is procurement. There is negative cost growth in this phase, however, it is quite variable.

The construction phase, the fourth in the construction life cycle, represents the most significant proportion of the project budget. Therefore, overcoming the construction phase's cost growth may have the greatest impact on the overall project cost. The last phase is commissioning. This phase has a similar budget to the front-end planning, however the cost growth and variability are much higher.

The methodology requires clean data, according to the first step of figure 3.3, which can be achieved through a pre-processing step as outlined in the previous chapter. Pre-processing is a crucial stage for feature selection and prediction. We first identified categorical and numerical features, in order to convert the categorical features into numerical features. For example, a number is dedicated to each category. Outliers were then identified and removed. Features were scaled and transformed, to between zero and one using MinMaxScaler, to prepare the dataset for feature selection.

Using this pre-processed data, LASSO regression is then applied as a feature selection technique to reduce the features to those contributing to cost growth. Figure 4.1 outlines the path of a coefficient in the LASSO model. Each colored line in the plot indicates a regression model parameter. This shows the coefficients of all independent features approaching zero as the tuning parameter λ rises from the 10^{-4} to 10^{-1}, and illustrates that more regularization results in reducing the number of nonzero feature coefficients. The coefficient of many features dramatically decreased to zero after λ rises from 10^{-3} and more. The vertical dotted line explains the λ determined by cross-validation, where the tuning parameter best fits the data, and represents 11 non-zero features when the λ is 0.0038. The LASSO feature selection identified 11 of the 29 features recommended by the domain expert.

The permutation importance of 11 extracted features are calculated according to the third step of figure 3.3, and presented in the figure 4.2. The permutation importance of each feature is

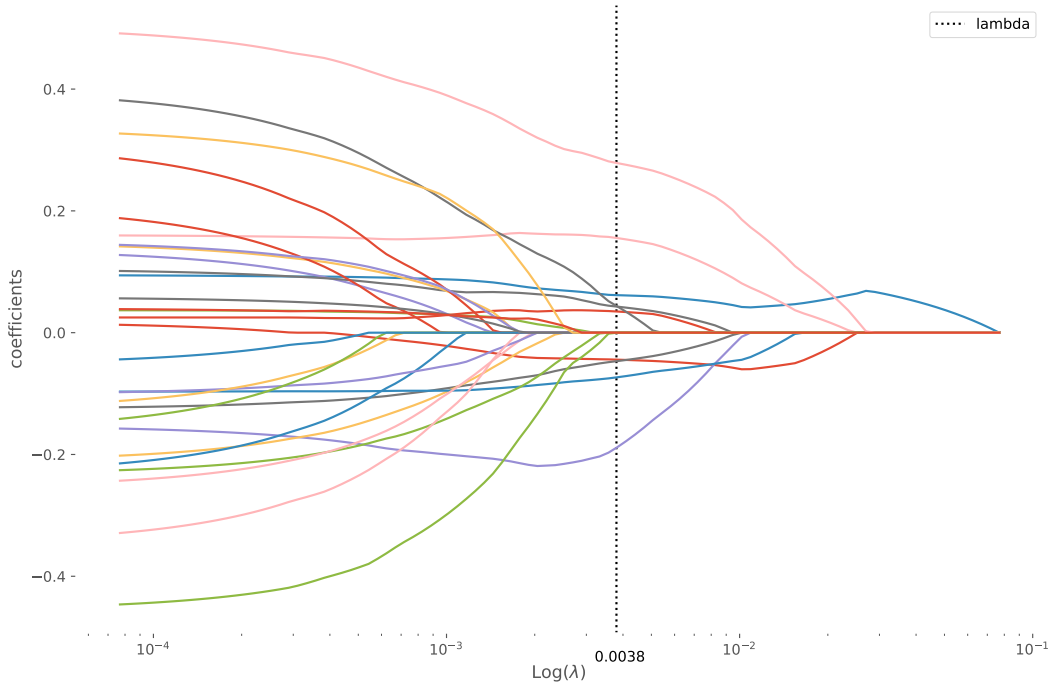

Figure 4.1: LASSO Path Using 29 Features. The figure shows the LASSO path for the cost growth of construction projects. There are 29 features entered the model, and by increasing the λ, features coefficients decreased to zero. The dotted line cut 11 features that are likely to create the highest model performance.

calculated using a random forest algorithm. As seen in the figure 4.2, the permutation importance of all features is positive, indicating that each impacts the cost growth prediction. Therefore, in this step, none of the features are removed. Although the permutation importance does not reduce the number of features in this case study, it provides valuable information about each feature. The most important feature is detail engineering actual duration, the least important is project type.

In tables 4.1 and 4.2, five conventional models were trained and tested using all 29 features recommended by the domain expert, and 11 features selected by feature selection techniques. These tables represent the performance of five conventional machine learning algorithms in terms of R^2 score and RMSE. Table 4.3 compares the performance of the models in tables 4.1 and 4.2. Table 4.3 demonstrates that by reducing to 11 important features, the R^2 score of all models has

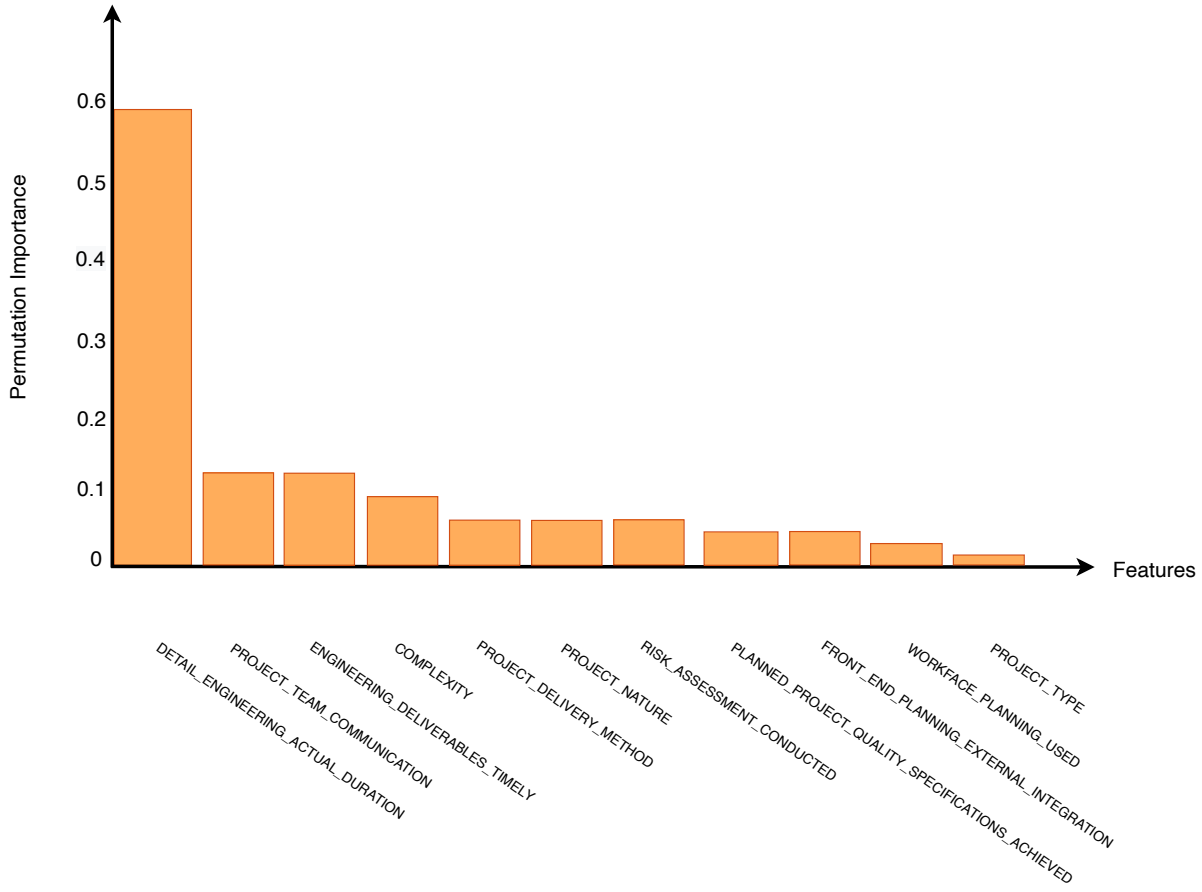

Figure 4.2: Permutation Feature Importance Using Random Forest Algorithm. Illustrates the importance of each of the 11 features for predicting cost growth with a random forest model. The most important feature is detail engineering actual duration associated with the highest error after permutation.

increased, and the RMSE has decreased. Moreover, highly correlated features are removed from the equation, and the correlation between independent variables is less than 0.5 for the 11 selected features.

4.1 Summary

A dataset of 137 heavy industrial projects was investigated for the validation of the feature selection techniques used in this study. For each project, 29 features (recommended by the domain expert) related to the first three phases of the construction project are considered for analysis. This is accomplished by applying LASSO regression and calculating feature importance, followed by the

Table 4.1: Predictive Regression Model Comparison Using Domain Expert Features. Compares the R^2 score and RMSE of five conventional models in three different distributions of training and testing data in order to select a model to calculate feature importance

Train-Test-Split	**90% train-10%test**			**80% train-20%test**			**70% train-30%test**		
Model	Train	Test	RMSE	Train	Test	RMSE	Train	Test	RMSE
SVR	0.43	–	0.2542	0.45	0.07	0.2593	0.47	0.01	0.2636
RF	0.52	0.12	0.2382	0.54	0.17	0.2472	0.57	0.10	0.2526
GBR	0.71	0.12	0.2376	0.73	0.14	0.2497	0.76	0.09	0.2531
RR	0.47	–	0.1658	0.50	0.14	0.2254	0.53	0.01	0.2230
KNN	0.99	0.20	0.2269	0.99	0.23	0.2391	0.99	0.19	0.2404

prediction of cost growth with different subsets of features.

The results show that, by reducing the number of features from 29 to 11, the performance of all models has increased. Moreover, the permutation importance plot highlights the importance of detail engineering's actual duration. According to previous studies, the more time dedicated to detail engineering, the less likely cost overrun will result.

Table 4.2: Predictive Regression Model Comparison Using Features Extracted by LASSO Regression from Domain Expert Features. Compares the R^2 score and RMSE of five conventional models in three different distributions of training and testing data.

Train-Test-Split	**90% train-10%test**			**80% train-20%test**			**70% train-30%test**		
Model	Train	Test	RMSE	Train	Test	RMSE	Train	Test	RMSE
SVR	0.36	0.16	0.2258	0.37	0.20	0.2411	0.37	0.20	0.2376
RF	0.50	0.14	0.2360	0.52	0.20	0.2430	0.54	0.13	0.2483
GBR	0.64	0.16	0.2303	0.65	0.16	0.2472	0.68	0.10	0.2515
RR	0.38	0.19	0.1464	0.40	0.23	0.2140	0.43	0.08	0.2144
KNN	0.97	0.20	0.2223	0.98	0.27	0.2307	0.98	0.22	0.2355

Table 4.3: Comparison of tables 4.1 and 4.2. The table shows that the R^2 score of all models has increased, and the RMSE has decreased in all three training and testing data distributions. These changes, however, are not insignificant.

Train-Test-Split	**90% train-10%test**		**80% train-20%test**		**70% train-30%test**	
Model	R^2 Score	RMSE	R^2 Score	RMSE	R^2 Score	RMSE
SVR	+0.16	-0.0284	+0.13	-0.0182	+0.19	-0.026
RF	+0.02	-0.0022	+0.03	-0.0042	+0.03	-0.0043
GBR	+0.04	-0.0073	+0.02	-0.0025	+0.01	-0.0016
RR	+0.19	-0.0194	+0.09	-0.0114	+0.07	-0.0106
KNN	0	-0.0046	+0.05	-0.0084	+0.03	-0.0049

Chapter 5

An Interactive Tool for Predictive Analysis

The difference between the actual cost and the budgeted cost is high in construction projects. Due to their complex nature, it can be challenging to understand the true impact of each data field on the overall execution and completion of the project. There is therefore a need for an interactive tool to allow for predictive analysis based on different scenarios and variations of significant data fields. This chapter presents the analytical techniques and tools that provide an interactive predictive analysis of large-scale projects with different features, and outlines how the tool's design can help practitioners generate insight and assist them in making strategic decisions for their company.

In the previous study [78], 16 of 281 features were selected using a combination of feature selection techniques: LASSO regression, and calculating the permutation importance using a random forest model. Using these features, five machine learning algorithms (Support Vector Regression, Random Forest, Gradient Boosting Regression, Ridge Regression, and K-Nearest Neighbors) were trained and compared. The previous study selected features primarily related to the last two phases of the projects. Therefore, the domain expert suggests 29 features from the first three phases of the project, among which three features (project type, project delivery method, and workforce planning used) exist in both selections. Moreover, cost growth (a standard measure of construction performance) is selected as a dependant variable. Based on these 42 features, the user can select desired features to train different models.

One of the main challenges of developing machine learning systems is understanding the performance across different input selections, specifically in construction management, where a different number of features are available at different stages of a project. Practitioners need to understand how each feature's changes can impact the output, and select features that result in the best performance. In a construction project area, the prediction results are as significant as how each

factor can impact the model prediction. Therefore, the visualization is of paramount importance to better understand why and how decisions are made. Interactive predictive models can provide practitioners with better insight into the predictive model. In this chapter, the interactive tool addresses two different functions: providing a manual feature selection, and trying to make the model more interpretable.

As demonstrated in previous chapters, we have 42 features to implement the tool for the first step. Sixteen out of 281 features have been selected (as they are expected to provide the highest model performance) using Lasso regression, followed by selection based on feature importance. The domain expert added 26 features from the first three phases. Users can choose any number of the total 42 available features from the total 42 to train models. Different models are trained and evaluated.

For the second step, each model acts as a black box, which means the user can't oversee how the algorithm functions. This study uses three approaches to make the model more interpretable. Calculating permutation importance is one method to explain the predictive model, which helps find the most factors contributing to the target variable. Another critical method for interpreting the regression model is determining the relationship between predictors and prediction using a partial dependence plot. Changing the input values and observing the corresponding output changes can also help users understand how a specific data point is predicted.

5.1 Methodology:

The predictive model performance can be varied by considering different sets of features. In this study, we provide a tool that allows the users to select their own features from the 16 variables (which are likely to provide the highest score) selected in the previous study, as well as the 29 features selected by the domain expert. Users can then test which selection of features gives them the highest R^2 score and the lowest RMSE at different project stages. The tool also provides a visual of results for all the models used in this study. Users can see the partial dependence plot

for each feature and modify data points to see the output. Figure 5.3 depicts the tool's diagram, as described in the following subsections.

5.1.1 Manual Feature Selection

Analysts need to know which selection of available features have the highest performance, or how changes in the selection of features can increase or decrease the model's performance. In the previous study [78], we found 16 important features that are most likely to result in the highest performance; these features are extracted by applying LASSO regression (a regularized linear regression technique), and then calculating feature importance. LASSO selects important features, removing those which are irrelevant, by reducing the features' coefficients to zero. These features are primarily related to the project's last phases; with this tool, analysts can select their desired features from all phases, making the model more domain-relevant and beyond the model performance.

In figure 5.1, the left box includes all 42 features, and the right includes selected input features of the model. Users can scroll through features, dragging them from the left box to the right and vice versa. The features marked with stars were selected in the previous study [78]. The highest performance is expected to be achieved by the random forest algorithm by having all the features with a star in the right box and training the models. However, analysts can select features based on the previous study, phases, or available information about the project to predict cost growth. After selecting features, users can press the "Train Models" button to train all models. In the next step, shown in the figure 5.2, the tool guides users to a page that reports the highest model performance and includes the permutation importance plot.

5.1.2 Permutation Importance

Figure 5.2 illustrates the permutation importance plot of the models. If users want to know which features make the most impact on cost growth, altering features one by one seems to be an inefficient procedure. Therefore, this tool calculates the permutation importance of each feature, so

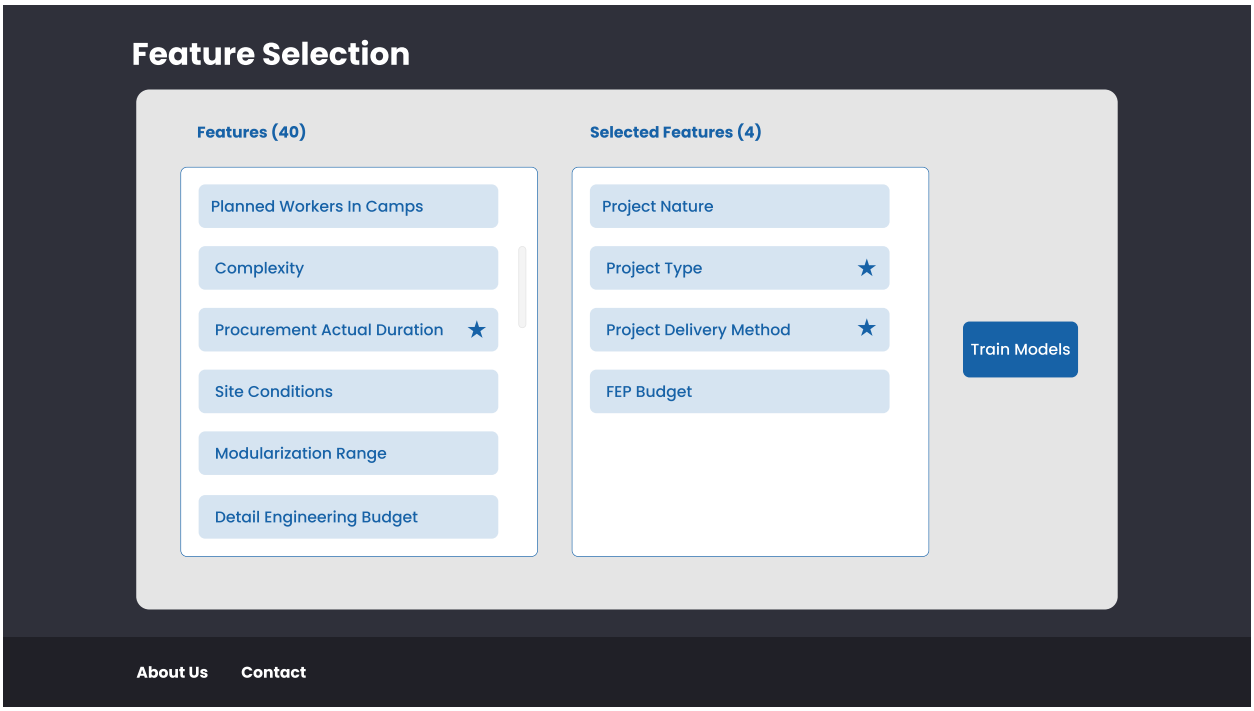

Figure 5.1: Manual Feature Selection. The tool supports users to select their features of interest to understand the predictive ability of each selection. The left box represents all features that the user can select, and the right box includes the user's features of interest. As users press the Train Model button, six predictive models will be trained based on the selected features.
Features including a star are recommended by the previous study[78] that likely has the highest model performance.

that users can add or remove features from the models based on the importance of features. The permutation feature importance is explained as a decrease in model performance, while shuffling a specific feature's values [74]. This approach also benefits from being model agnostic and can be computed repeatedly with different feature permutations, while also helping analysts identify the most important features from all models. Once the most important features from all models are known, they can return to the feature selection page and create new models. According to figure 5.2, by applying a random forest model, we can see that the first two features contribute the most to the model performance. Feature importance is one of the popular techniques to help users find powerful features for prediction [81]. However, the information provided by the feature importance is not adequate to describe how each feature value can impact the model's prediction [57]. To solve this problem, we used a partial dependence plot (described in the following sections) to il-

lustrate prediction responses caused by changes in features. Users can identified important features to manipulate data points and see how changes in those features can impact the cost growth.

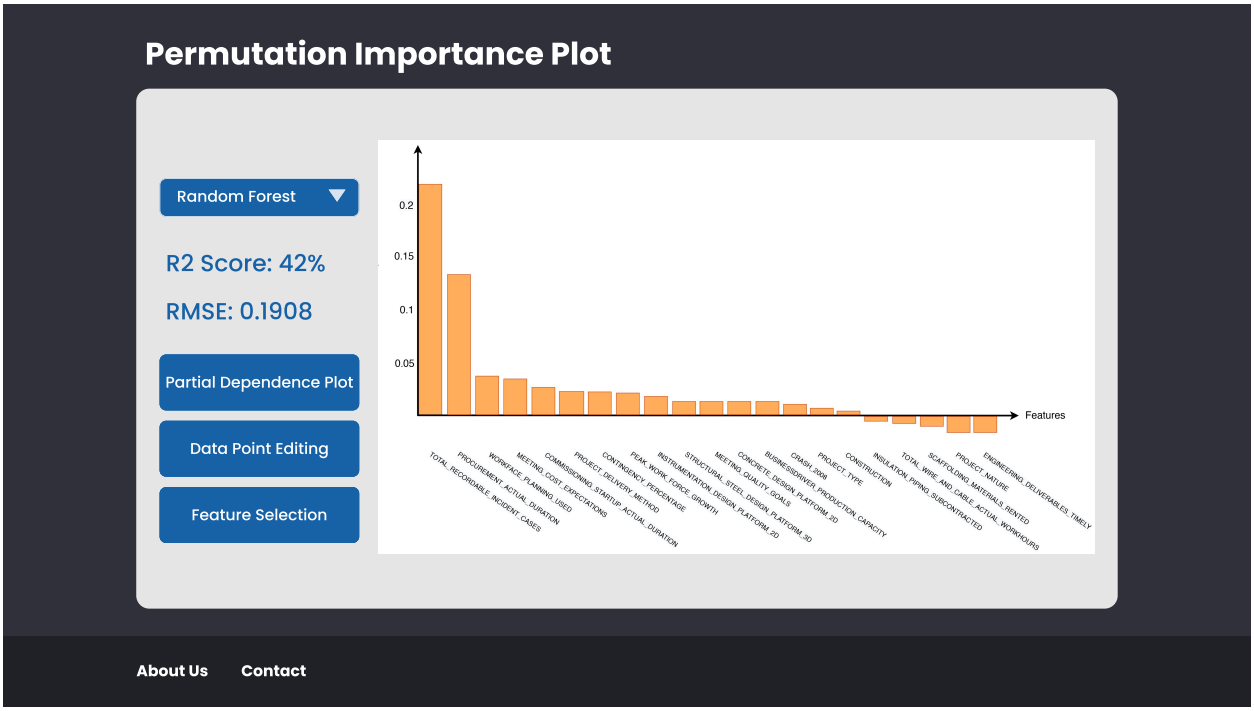

Figure 5.2: Permutation Importance Plot. After training models, the tool directs the user to the model's permutation importance plot with the highest R2 score. However, users can select different models and see the R2 score and RMSE of the models as well as the permutation importance of features based on the selected model.

5.1.3 Multiple Model Comparisons

Different models will be trained and evaluated based on the features that users select. For the selected features, the tool trains six models, five of which are similar to machine learning algorithms that are used in the previous study [78] (Support Vector Regression, Random Forest, Gradient Boosting Regression, Ridge Regression, and K-Nearest Neighbors). A short description of each of these algorithms is provided in table 5.1. The five algorithms are selected due to different reasons. Support vector machine is selected as it is the most common algorithm used in the construction performance prediction. Random forest and gradient boosting regressors are selected to see the results of tree-based algorithms. As we used LASSO, a regularized linear regression technique,

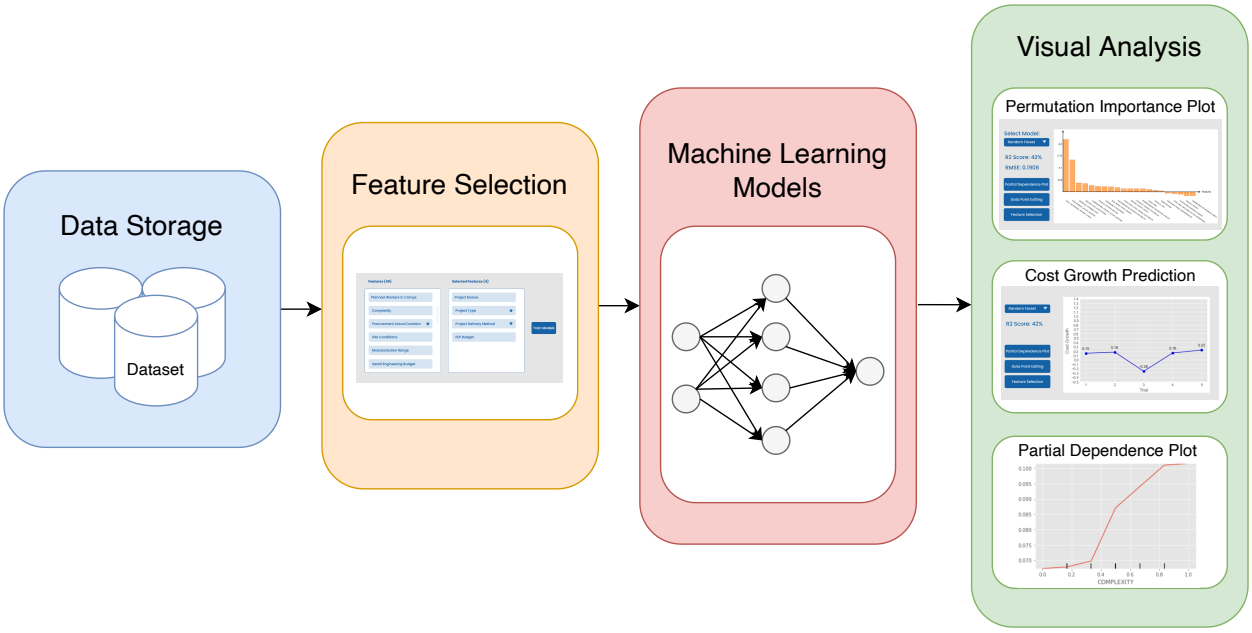

Figure 5.3: Tool Diagram

ridge regression is also selected to check another regularized model. We also used a default structure for the deep learning model. However, the first layer of the deep learning model varies based on the number of selected features. Deep learning aims to define feature hierarchies, with features from higher levels formed by the composition of lower-level features [54]. This study used a simple multi-layer perceptron (MLP) with three dense layers, followed by a dropout layer after each dense layer, as shown in figure 5.4. There are 256, 128, and 64 neurons in the first, second and third hidden layers, respectively. The combination of these fully-connected networks and dropouts are repeated three times. Dropout layers (a kind of regularization technique) are applied to prevent overfitting of the deep learning model [82]. A number needs to be assigned to each dropout layer; in our model, we assigned 0.5 to all dropout layers. Therefore, the algorithm does not rely on a specific feature, which means 50% of neurons on that layer are ignored. A learning rate of 0.0005 is selected to control the change in the model in response to the error whenever the model weights are updated. Finally, the last fully connected layer includes a single output neuron, and the activation function used in this model is ReLU.

After training the models, the one with the highest performance will be reported; however, it is

possible to check other models' results by changing the model. This model selection is available for all plots.

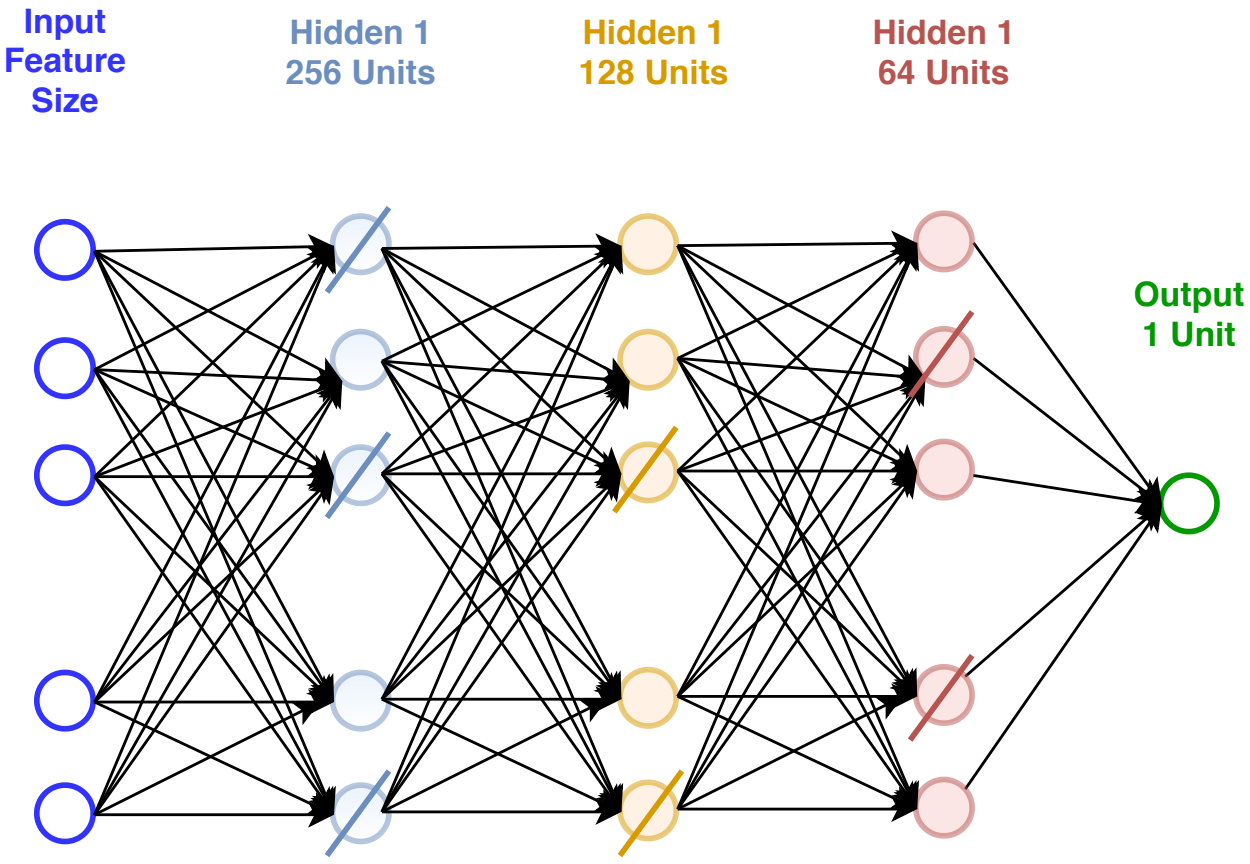

Figure 5.4: Dense deep neural network with three fully connected, three dropout layers and an output layer with relu activation.

5.1.4 Data Point Editing

Similar to [63] and [58] our tool allows users to modify the input, and see how the modification changes the output. As shown in figure 5.5, users can change the value of input parameters and see the predicted value of cost growth in the white box at the left side of the page. After modifying the values, users can then press the "Next Trial" button to open a similar page. It is possible to tweak parameters five times, and then by pressing the "Predict" button, users will be guided to the page shown in figure 5.6. The tool then illustrates the output and its prediction. It is also possible to set five data points and compare the results on a plot. Users can edit, add or remove a data

Table 5.1: Description of machine learning algorithms

Machine Learning Algorithms	Description
Support Vector Regression (SVR)	SVR is one of the most famous regression models in construction problems because of its predictive ability. SVR computes the distance between all the data points using vector algebra, and a hyperplane is formed between the closest points. This algorithm is well known for high-dimensional datasets, as the optimization of SVR does not depend on the input space's dimensionality. [20, 83]
Random Forest (RF)	Random Forest is an ensemble of decision tree models. This algorithm creates a forest using a number of trees, which makes it much more powerful than a decision tree model. [74]
Gradient Boosting Regressor (GBR)	The Gradient boosting model uses an ensemble of weak prediction models to produce a more robust predictive model. Constructing a series of weak models (such as decision tree and random forest), GBR optimizes loss function in each phase. GBR is able to solve a wide range of problems due to many hyperparameters that help the model be more flexible. [84, 66, 85]
Ridge Regression (RR)	Ridge Regression is used to analyze multiple regression when there is multicollinearity in the data. [52]
K-Nearest Neighbor (KNN)	KNN predicts the test sample based on k nearest samples that are closest to the test sample. [86]

point. It is therefore possible to set different values for each feature and compare their results. Users can determine which changes would be optimal by changing the value of different features. It is easy to compare models, because the tool also provides users with the ability to examine the results of available models. Depending on the number of available projects, models can have different prediction scores; the more projects that are available, the better the prediction score. Some models work far better when the number of samples increases. For example, as the number of projects increases, the neural network model is likely to perform better than other conventional models (although, according to the available data used for this study, it is not optimal.

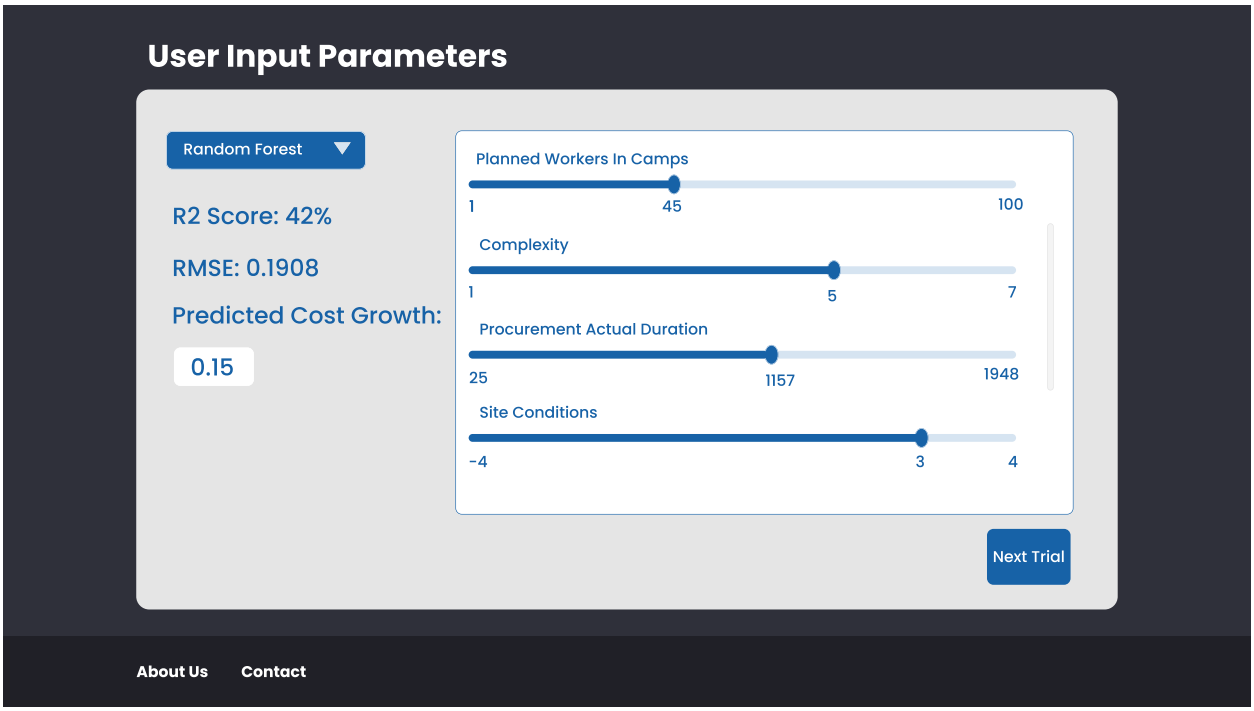

Figure 5.5: User Input Parameter. Users can change values for different features and see the cost growth interactively. Moreover, users can press "Next Trial" to consider another project and do the same for five projects. Finally, results can be compared in figure 5.6. Users can also change the model and observe the corresponding R^2 score and RMSE.

5.1.5 Partial Dependence Plot (PDP)

Practitioners usually ask about the impacts of a feature across the entire spectrum of values. Therefore, the partial dependence plot can address this demand by illustrating the prediction changes caused by changes in the feature values. A partial dependence plot is one of the most common visualization techniques that help to understand the dependence between the target variable, and one or two input features of interest [87], [58]. A PDP is illustrated as a line chart, where the x-axis depicts feature values, and the y-axis shows the partial dependence. Using model-agnostic approaches by considering the model as a black-box [88], PDP can show the impact of changes in input on output, more specifically, how the model response by increasing or decreasing values of one feature. The partial dependence of a feature is calculated by equation 5.1, where f is the feature for which the partial dependence should be calculated, N is the number of instances in the input matrix x, *pred* is the prediction of an input instance. The average outcome is calculated over

Figure 5.6: Cost Growth Prediction. The figure shows the comparison of cost growth of five data points. Therefore users can understand how changes in cost growth can occur. The comparison can also be examined by changing the models.

all of the input instances while changing the input values of feature f to v [58].

$$pdp_f(v) = \frac{1}{N}\sum_{i}^{N} pred(x_i) \;\; with\, x_{if} = v \tag{5.1}$$

In order to inspect the impact of one feature on the prediction using PDP, we keep features other than the feature of interest unchanged and fix values of all instances of the feature of the interest to some fixed values. So that we can witness the impact of that specific feature on the model prediction. Figure 5.7 shows that more complex projects tend to experience more cost overrun.

5.2 Evaluation

In order to evaluate the tool, after shuffling the observations, 20% of the data is dedicated to testing models; that way, different models can be compared on testing data. The remaining 80% is used to train models. We first need to tune hyperparameters in order to train each model. Hyperparameters are those parameters that need to be set by the designers. However, as the number of features

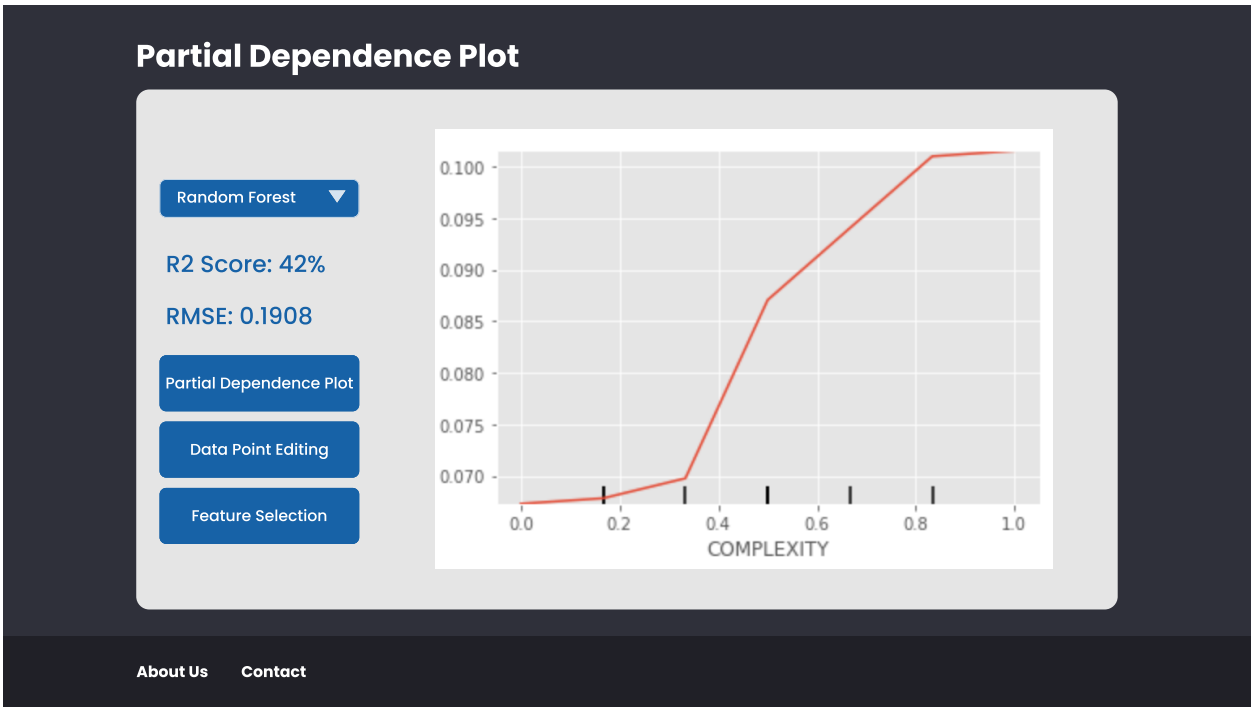

Figure 5.7: Partial Dependence Plot. The x-axis is the feature values, and the y-axis depicts the partial dependence. The vertical line shows the distribution of the data. The figure shows the partial dependence of the feature complexity. As illustrated, as the project's complexity increases, the cost growth is more likely to increase.

is changed based on the user choices, the tool automates the process of tuning models. The tool searches for the best selection of hyperparameters from among the available choices by applying random search and k-fold cross-validation. To start with the random forest algorithm, ten numbers are considered from 50 to 500 as the number of trees for the first hyperparameter. The default value of the minimum samples split (the minimum number of samples needed to split an internal node) is two. Four numbers between two and ten are considered for this parameter. The minimum samples leaf (the minimum number of observations needed to be at a leaf node) is another hyperparameter that needs tuning. The default value for this parameter is one, and five numbers between one and eight are considered. The last parameter that needs tuning is the max feature, which represents the number of features to consider when searching for the best split. Random search tries different combinations of the above parameters, and selects the best combination based on the defined score (which is R^2 score in this tool). A similar process is also done for other algorithms.

In this study, we used k-fold cross-validation to assess the performance of the model. As we only have 139 samples, the training data is divided into three folds to have adequate training data. The model will be trained in the two subsets, and tested on the third subset three times, to help the model's robustness. The optimal estimator is selected based on the highest R^2 score after fitting the randomized search. The best hyperparameter for each model is determined after hyperparameter tuning. Using training data, we applied random search with k-fold cross-validation to determine the optimal model. Users can compare models based on the testing data, and models with the lowest RMSE and Highest R^2 score are provided.

5.3 Summary

In this chapter, we demonstrated an interactive tool to analyze factors and predict project cost growth. Users can select their features of interest to train models in order to predict cost growth. We introduced and applied three different approaches for making interpretable predictive models. The first approach illustrated the permutation feature importance of each feature using the selected model. Although this method can detect the important factors, the contribution of each factor is not considered. Therefore, we can focus more on the important factors' contribution with partial dependence plot and data point editing at the second and third approaches.

Chapter 6

Conclusion and Future Work

6.1 Summary and Concluding Remarks

Cost overruns in the construction industry are common, occurring when the actual cost exceeds the budgeted cost. Data collected by the COAA and CII partnership is used in this study to address the issue.

Researchers are especially interested in improved means of estimating a construction project's costs to help prevent projects from exceeding their budget, thereby increasing their profits and number of projects, as well as enhancing investments and economy. The prediction of cost growth requires identifying the project's contributing factors. This thesis investigated factors impacting cost overruns, using feature selection techniques and designing a tool to predict the project's cost growth based on the selection of different features at various stages of a project.

The first chapter provided an overview of the importance of heavy industrial projects in Alberta, followed by a brief description of the thesis. The second chapter provided a literature review concerning the topics of artificial intelligence and machine learning in project management and, more specifically, in project cost management. Next, different feature selection algorithms are described. Finally, there is a literature review of artificial intelligence and machine learning tools to identify a more interpretable model.

In Chapter 3, machine learning approaches are used to develop a model that can detect the factors with the highest impact on project cost growth. Five machine learning algorithms are developed to predict cost growth and evaluate the predictive ability of extracted features. Data cleaning prepares the data for feature selection, after which 281 variables were selected to apply to the LASSO regression model. LASSO detected 21 significant features to use as an input for the

random forest algorithm. After applying a random forest algorithm, the importance of each feature was calculated through permuting features.

Sixteen of 21 input features were identified, based on feature importance, via the processes shown in figure 3.3. Two of these 16 variables (project type and project delivery methods) were detected in the first stages of the projects. Contingency percentage and engineering accuracy were considered important factors during the front-end planning phase. The remaining extracted features mostly pertain to the last two phases of the project. In Chapter 4, the feature selection techniques were assessed using 29 features selected by the domain expert, validating the results of the previous chapter.

In Chapter 5, with 16 extracted features and 29 selected by the domain expert (features dedicated to the first three phases of the project), an interactive tool is developed to identify the contribution of each factor on project cost growth. Users can select desired features and train models to see how they can improve models in an interactive user interface. This tool can provide practitioners with better insight into the predictive model. The tool makes the models more interpretable and explanatory by providing interactive data point editing to test hypotheses on how each feature affects prediction. Users can tweak each feature of a data point's values and see how the cost growth of a project can be changed.

Furthermore, the tool plots the permutation importance of each feature, which can help find the most important features that impact cost growth. The tool also supports a partial dependence plot, which helps users better understand the impact of features on overall prediction. Users can go back to the first step and redesign their models based on insights from the visualization.

6.2 Contributions

This thesis has three significant contributions.

1. Theoretical contribution

- Proposed a hybrid approach that can increase the cost performance by considering only the most important features.
- Identified a few features that can achieve the highest possible performance to predict cost growth from a high-dimensional dataset.

2. Practical Contribution

- Presented the analytical techniques and tools that can be used in industry to provide an interactive predictive model for large-scale projects. Need to be tested on future project.

A project is more likely to succeed(in terms of cost growth) if it is executed with careful consideration of these key factors; however it is not possible to generate a universal list of success criteria that suits all projects.

6.3 Limitation

There are several limitations to this study. The performance of machine learning algorithms may increase by training a model with more data. However, one of the main challenges of this study is the high number of features containing a small number of samples. Although the sample size for this study is minimal, the results are still promising compared to previous construction studies. The results would be more accurate with the addition of more samples to the models.

There are additional factors that can also influence the cost growth of projects; however, the effects of such variables on the performance were not investigated in this study due to their limited availability. Data collection occurred over a long period, and the survey instrument's formation was not part of this study. Hence, this study limits its analysis to the features for which data was collected.

6.4 Recommendations for Future Work

Recommendations for future work relevant to identifying factors and predicting cost growth on construction projects could be extended in various directions.

- This study's generalization ability is limited to projects in Alberta; thus, data from other locations could be used in further research. Given more data, we can explore the potential advantages of using deep learning and neural networks, and compare results with the traditional machine learning approaches.

- The analysis of data in this study is limited to only one aspect of project performance. Future studies can use the same approach on other aspects of project performance, such as schedule growth and safety. Although the tool provides good insight of the model to help users make decisions, it could still be enhanced. This tool only predicts and analyzes contributing factors to the cost growth of projects, and needs to be expanded to cover other project performance areas such as schedule, safety, etc.

- This approach was applied to reduce the dataset's dimension for predictive modelling of cost growth in Alberta's construction projects, but can also be used for other areas in need of a predictive model for a high-dimensional dataset.

- An interactive tool designed in this study follows modularity principles. Therefore, more models can be added to the tool in future.

- The functionality of the tool can also be improved. In this study, the partial dependence plot only supports one feature at a time; however, in future study we could plot the PDP for two features to see the interaction between them.

Bibliography

[1] M. Robu, F. Sadeghpour, and G. Jergeas, "Best Practices Impacting the Cost Performance of Heavy Industrial Projects," *Canadian Society for Civil Engineering, CSCE, Fredericton, NB, Canada*, 2018.

[2] ——, "BEST PRACTICES IMPACTING CONSTRUCTION PROJECT SCHEDULE," 2019.

[3] COAA, "Alberta Report # 2 COAA Major Projects Performance Assessment System Project Performance Engineering Productivity Construction Productivity," 2014.

[4] ——, "Alberta Report # 3 COAA Major Projects Performance Assessment System Project Performance Engineering Productivity Construction Productivity," 2019.

[5] M. Robu, *"Impact of best practices on the cost and schedule performance of heavy industrial construction projects" unpublished M. Eng*. thesis, Schulich School of Engineering, Calgary, AB, 2019.

[6] D. K. H. Chua, Y. C. Kog, P. K. Loh, and E. J. Jaselskis, "Model for construction budget performance-neural network approach," *Journal of Construction Engineering and Management*, vol. 123, no. 3, pp. 214–222, 1997.

[7] Y.-R. Wang and G. E. Gibson, "A study of pre-project planning and project success using ANNs and regression models," *Automation in Construction*, vol. 19, no. 3, pp. 341–346, 2010.

[8] U. Gazder, M. S. Islam, and M. Arifuzzaman, "Parametric Modeling of the Cost of Power Plant Projects," in *2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)*, Dec. 2020, pp. 1–5.

[9] D. Haines, *"Factors impacting cost growth on heavy industrial projects in Alberta," unpublished M. Eng*. thesis, Schulich School of Engineering, Calgary, AB, 2020.

[10] V. R. Ambrule and A. N. Bhirud, "Use of artificial neural network for pre design cost estimation of building projects," *The International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 5, no. 2, pp. 173–176, 2017.

[11] M. Juszczyk, A. Leśniak, and K. Zima, "ANN Based Approach for Estimation of Construction Costs of Sports Fields," *Complexity (New York, N.Y.)*, vol. 2018, pp. 1–11, 2018.

[12] V. Arabzadeh, S. Niaki, and V. Arabzadeh, "Construction cost estimation of spherical storage tanks: Artificial neural networks and hybrid regression—GA algorithms," *Journal of Industrial Engineering International*, vol. 14, no. 4, pp. 747–756, 2018.

[13] B. Tong, J. Guo, and S. Fang, "Predicting Budgetary Estimate of Highway Construction Projects in China Based on GRA-LASSO," *Journal of Management in Engineering*, vol. 37, no. 3, p. 04021012, 2021.

[14] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing (Amsterdam)*, vol. 300, pp. 70–79, 2018.

[15] C.-F. Tsai, "Feature selection in bankruptcy prediction," *Knowledge-Based Systems*, vol. 22, no. 2, pp. 120–127, 2009.

[16] P. Langley, "Selection of relevant features in machine learning," in *Proceedings of the AAAI Fall symposium on relevance*, vol. 184, 1994, pp. 245–271.

[17] G. G. Tiruneh and A. R. Fayek, "Feature selection for construction organizational competencies impacting performance," in *2019 IEEE International Conference*, 2019, pp. 1–5.

[18] Y. Zhang and S. Fang, "RSVRs based on Feature Extraction: A Novel Method for Prediction of Construction Projects' Costs," *KSCE Journal of Civil Engineering*, vol. 23, no. 4, pp. 1436–1441, 2019.

[19] Y. Kog, D. Chua, P. Loh, and E. Jaselskis, "Key determinants for construction schedule performance," *International Journal of Project Management*, vol. 17, no. 6, pp. 351–359, 1999.

[20] C. Q. Poh, C. U. Ubeynarayana, and Y. M. Goh, "Safety leading indicators for construction sites: A machine learning approach," *Automation in Construction*, vol. 93, pp. 375–386, 2018.

[21] A. Assefa Tsehayae and A. Robinson Fayek, "Developing and optimizing context-specific fuzzy inference system-based construction labor productivity models," *Journal of Construction Engineering and Management*, vol. 142, no. 7, 2016.

[22] H. Son, C. Kim, and C. Kim, "Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables," *Automation in Construction*, vol. 27, pp. 60–66, 2012.

[23] V. Kumar and S. Minz, "Feature selection: A literature review," *SmartCR*, vol. 4, no. 3, pp. 211–229, 2014.

[24] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," *Journal of Statistical Software*, vol. 36, no. 11, 2010. [Online]. Available: https://doaj.org/article/3b770563968e448aa712b89a989e781d[AccessedNov.19,2020].

[25] K. Liu and N. El-Gohary, "Feature discretization and selection methods for supporting bridge deterioration prediction," 2018, pp. 413–423.

[26] S. Chi, S.-J. Suk, Y. Kang, and S. P. Mulva, "Development of a data mining-based analysis framework for multi-attribute construction project information," *Advanced Engineering Informatics*, vol. 26, no. 3, pp. 574–581, 2012.

[27] D. Chakraborty, H. Elhegazy, H. Elzarka, and L. Gutierrez, "A novel construction cost prediction model using hybrid natural and light gradient boosting," *Advanced Engineering Informatics*, vol. 46, p. 101201, 2020.

[28] Y.-R. Wang, C.-Y. Yu, and H.-H. Chan, "Predicting construction cost and schedule success using artificial neural networks ensemble and support vector machines classification models," *International Journal of Project Management*, vol. 30, no. 4, pp. 470–478, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263786311001256

[29] H. H. Elmousalami, "Comparison of Artificial Intelligence Techniques for Project Conceptual Cost Prediction: A Case Study and Comparative Analysis," *IEEE Transactions on Engineering Management*, vol. 68, no. 1, pp. 183–196, Feb. 2021.

[30] M. Fan and A. Sharma, "Design and implementation of construction cost prediction model based on SVM and LSSVM in industries 4.0," *International Journal of Intelligent Computing and Cybernetics*, 2021.

[31] Y. Zhang, R. E. Minchin Jr, and D. Agdas, "Forecasting completed cost of highway construction projects using LASSO regularized regression," *Journal of Construction Engineering and Management*, vol. 143, no. 10, p. 04017071, 2017.

[32] M. H. Rafiei and H. Adeli, "Novel machine-learning model for estimating construction costs considering economic variables and indexes," *Journal of Construction Engineering and Management*, vol. 144, no. 12, p. 04018106, 2018.

[33] ——, "A novel machine learning model for estimation of sale prices of real estate units," *Journal of Construction Engineering and Management*, vol. 142, no. 2, p. 04015066, 2016.

[34] T. D. Akinosho, L. O. Oyedele, M. Bilal, A. O. Ajayi, M. D. Delgado, O. O. Akinade, and A. A. Ahmed, "Deep learning in the construction industry: A review of present status and future innovations," *Journal of Building Engineering*, p. 101827, 2020.

[35] E. Cantú-Paz, S. Newsam, and C. Kamath, "Feature selection in scientific applications," 2004, pp. 788–793.

[36] S. Aghakhani, R. Alhajj, J. Rokne, and P. Chang, "An Effective LmRMR for Financial Variable Selection and its Applications," in *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, vol. 217. IEEE, 2017, pp. 535–543.

[37] H. Rafieipour, A. Abdollah Zadeh, A. Moradan, and Z. Salekshahrezaee, "Study of Genes Associated With Parkinson Disease Using Feature Selection," *Journal of Bioengineering Research*, vol. 2, no. 4, pp. 1–12, 2020.

[38] Z. Salek, F. M. Madani, and R. Azmi, "Intrusion detection using neural networks trained by differential evaluation algorithm," in *2013 10th International ISC Conference on Information Security and Cryptology (ISCISC)*, Aug. 2013, pp. 1–6.

[39] S. Chowdhury, X. Dong, and X. Li, "Recurrent neural network based feature selection for high dimensional and low sample size micro-array data," in *2019 IEEE International Conference on Big Data (Big Data)*, Dec. 2019, pp. 4823–4828.

[40] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Recent advances and emerging challenges of feature selection in the context of big data," *Knowledge-based systems*, vol. 86, pp. 33–45, 2015.

[41] R. Kavitha and E. Kannan, "An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining," in *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*. IEEE, 2016, pp. 1–5.

[42] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[43] S. Aghakhani, *"Effective Data Analysis Framework for Financial Variable Selection and Missing Data Discovery," unpublished Ph.D.* thesis, Faculty of Graduate Studies, University of Calgary, Calgary, AB, 2017.

[44] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *The Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.

[45] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*. Springer, 2008, vol. 207.

[46] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[47] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," vol. 1. Citeseer, 2001, pp. 74–81.

[48] M. Kuhn and K. Johnson, *Applied predictive modeling*. Springer, 2013, vol. 26.

[49] ——, *Feature engineering and selection: A practical approach for predictive models*. CRC Press, 2019.

[50] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[51] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[52] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," in *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, Oct. 2016, pp. 18–20.

[53] V. Fonti and E. Belitser, "Feature selection using LASSO," *VU Amsterdam Research Paper in Business Analytics*, vol. 30, pp. 1–25, 2017.

[54] F. Sharifi, E. Mohammed, T. Crump, and B. H. Far, "A cluster-based machine learning model for large healthcare data analysis," in *International Conference on Big Data Innovations and*

Applications. Springer, 2019, pp. 92–106.

[55] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[56] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! Criticism for interpretability," *Advances in Neural Information Processing Systems*, vol. 29, pp. 2280–2288, 2016.

[57] X. Zhao, Y. Wu, D. L. Lee, and W. Cui, "iforest: Interpreting random forests via visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 407–416, 2018.

[58] J. Krause, A. Perer, and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," 2016, pp. 5686–5697.

[59] Y. Zhou and G. Hooker, "Interpreting models via single tree approximation," *arXiv preprint arXiv:1610.09036*, 2016.

[60] V. Schetinin, J. E. Fieldsend, D. Partridge, T. J. Coats, W. J. Krzanowski, R. M. Everson, T. C. Bailey, and A. Hernandez, "Confident Interpretation of Bayesian Decision Tree Ensembles for Clinical Applications," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 3, pp. 312–319, May 2007.

[61] P. Tamagnini, J. Krause, A. Dasgupta, and E. Bertini, "Interpreting black-box classifiers using instance-level visual explanations," 2017, pp. 1–6.

[62] J. Krause, A. Perer, and E. Bertini, "Using visual analytics to interpret predictive machine learning models," *arXiv preprint arXiv:1606.05685*, 2016.

[63] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, and J. Wilson, "The What-If Tool: Interactive Probing of Machine Learning Models," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 1–1, 2019.

[64] F. Cheng, Y. Ming, and H. Qu, "DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models," *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[65] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[66] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001, publisher: JSTOR.

[67] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[68] P. Cortez and M. J. Embrechts, "Opening black box data mining models using sensitivity analysis." IEEE, 2011, pp. 341–348.

[69] ——, "Using sensitivity analysis and visualization techniques to open black box data mining models," *Information Sciences*, vol. 225, pp. 1–17, 2013.

[70] J. Zhang, Y. Wang, P. Molino, L. Li, and D. S. Ebert, "Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 364–373, Jan. 2019.

[71] J. Krause, A. Perer, and E. Bertini, "INFUSE: Interactive Feature Selection for Predictive Modeling of High Dimensional Data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1614–1623, Dec. 2014.

[72] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner, "Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets." IEEE, 2003, pp. 105–112.

[73] COAA, "Alberta Report # 1 COAA Major Projects Benchmarking Summary," 2009.

[74] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[75] W. Alawad, M. Zohdy, and D. Debnath, "Tuning hyperparameters of decision tree classifiers using computationally efficient schemes," in *In 2018 IEEE First International Conference on Artifical Intelligence and Knowledge Engineering (AIKE)*, 2018, pp. 168–69.

[76] M. Kuhn and K. Johnson, "Measuring performance in regression models," in *Applied Predictive Modeling*, Springer, 2013, pp. 95–100.

[77] C. J. Willmott, "Some comments on the evaluation of model performance," *Bulletin of the American Meteorological Society*, vol. 63, no. 11, pp. 1309–1313, 1982.

[78] N. Tajziyehchi, M. Moshirpour, G. Jergeas, and F. Sadeghpour, "A Predictive Model of Cost Growth in Construction Projects Using Feature Selection," in *2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Dec. 2020, pp. 142–147.

[79] K. Dogan and O. D. Incel, "Mobile device identification via user behavior analysis," 2019, pp. 32–46.

[80] H. H. Elmousalami, "Artificial Intelligence and Parametric Construction Cost Estimate Modeling: State-of-the-Art Review," *Journal of Construction Engineering and Management*, vol. 146, no. 1, p. 3119008, 2020.

[81] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225 – 2236, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865510000954

[82] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[83] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," 1997, pp. 155–161.

[84] J. Cai, K. Xu, Y. Zhu, F. Hu, and L. Li, "Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest," *Applied Energy*, vol. 262, p. 114566, 2020.

[85] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[86] Y. Xiao, Y. Ma, and H. Ding, "Air traffic flow prediction based on k nearest neighbor regression." IEEE, 2018, pp. 1265–1269.

[87] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[88] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.