Predictive Modelling for Cost Estimation in Construction Projects Using Machine Learning Algorithms

MSc Research Project
Data Analytics

Prajwal Shashidhara
Student ID: x22209077

School of Computing
National College of Ireland

Supervisor: Dr. Bharat Agarwal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Prajwal Shashidhara
Student ID:	x22209077
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Dr. Bharat Agarwal
Submission Due Date:	16/09/2024
Project Title:	Predictive Modelling for Cost Estimation in Construction Projects Using Machine Learning Algorithms
Word Count:	6364
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Prajwal Shashidhara
Date:	14th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictive Modelling for Cost Estimation in Construction Projects Using Machine Learning Algorithms

Prajwal Shashidhara
x22209077

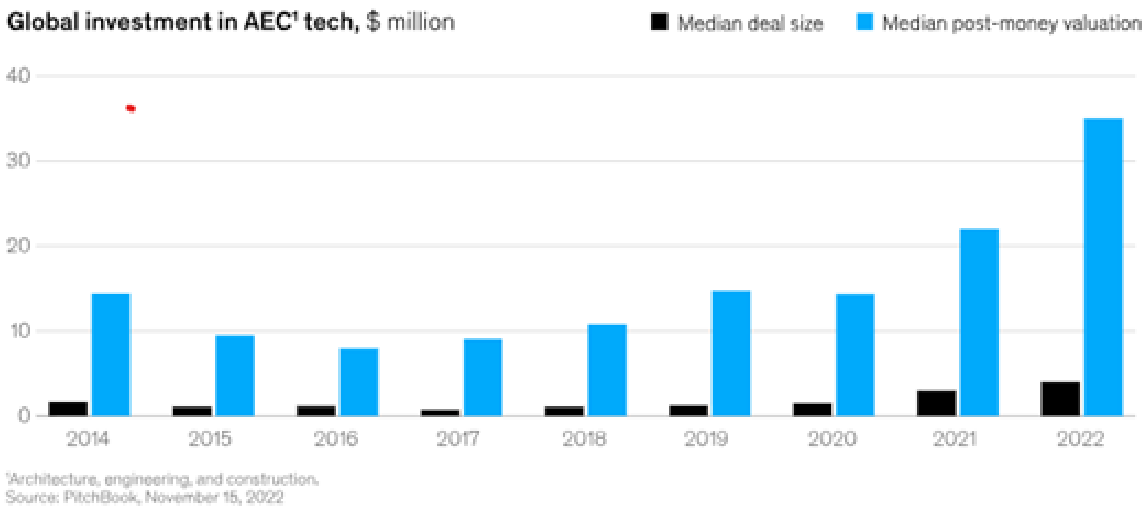
Abstract

The construction industry is growing rapidly around the world, with an expectation to reach $10.5 trillion by 2023. The growth is driven by factors such as the urbanization of the population. The industry necessitates building accurate and reliable methods to forecast costs as compared to traditional methods that fall short and overrun the budget. This study explores the application of predictive modelling using various algorithms for cost estimation in construction projects. The research seeks to determine the best method for early-stage cost prediction by examining models that use statistics, machine learning, and deep learning. According to the study, random forest gives higher accuracy in estimating building costs. Floor space, CPI, lot size, and project time are some of the key factors that affect expenses. Machine learning models, according to the findings, can improve the timeliness and accuracy of cost estimations, which in turn helps with improved resource management and project planning. This study highlighted the use of standard data mining methodologies like CRISP-DM (Cross-Industry Standard Process for Data Mining) and the model explainability method SHAP (Shapely Additive exPlanations) to help in better understanding and decision making.

1 Introduction

1.1 Background

The construction industry all around the world is increasing at a fast pace, with an expectation of reaching a $ 10.5 trillion economy by 2023, as shown in Fig 1.Business Wire (2021). The rise of population and urbanization has given rise to the following market, and it is not going to stop anytime soon. According to McKinsey, the given industry is the largest in the world economy and holds 13% of GDP. Mckinsey highlighted the use of AI (Artificial Intelligence) in the field of construction to provide better solutions throughout the whole project life-cycle, including design, building, energy efficiency, financing, procurement, and construction operations. Most construction projects outrun the budget as they were tackled using traditional methods that rely on historical data and judgements. As the given projects become more advanced and data-driven, cost estimations need to be enhanced to handle the complexity. To address this, predictive modeling offers a better approach that is reliable and may attain maximum accuracy while leveraging larger datasets .


Global investment in AEC[1] tech, $ million

[1] Architecture, engineering, and construction.
Source: PitchBook, November 15, 2022

Figure 1: Investment in Construction Technology McKinsey (2023)

1.2 Motivation

The construction industry faces many challenges while estimating the cost because of complex inputs and variable factors of labour cost, and the following can lead to inappropriate judgement while calculating the prices. Studies have shown that predictive analytics play a major role in predicting cost estimation, reducing the time for managers, and making better decisions for construction managers.Miranda et al. (2022) *Predictive Analytics in Preventing Construction Disasters* (2024) Hence, the given integration of a driven approach would help in better adjustment of project management based on various factors Rafiei and Adeli (2018). The reason why predictive analytics is preferred is because of its ability to analyze large datasets and find hidden patterns in the data, which is not possible with traditional methods. Also, these models can adapt to new datasets and provide real-time results that are more accurate and reliable, helping managers make better decisions for projects.

1.3 Research Objectives

- Analyse the statistical algorithms, machine learning algorithms, and deep learning algorithms for their application in the construction industry, making a benchmark comparative study.
- Find the most important factors that impact the cost of the project and incorporate these features in these models.
- Evaluate the performance of these models on various metrics to identify the most effective approach.

1.4 Research Questions

- Find a real-world dataset and try to find the most compelling features related to predicting the cost of one project and make a generalized model.
- Use these early cost estimation models and human intuition to make better decisions on the project and reduce cost.
- Find the best algorithms that can be used to build the given model and understand the importance of machine learning, deep learning, and statistical analysis.
- Try to understand the accuracy of these models to explain and interpret the given problem better.

1.5 Limitations

- The data should be of high quality and represent the current market conditions to get better results. Hence, data quality and sources from where it is attained are important.
- Making the generalized model for all construction projects is not possible, accounting for the complexity of these projects.
- Machine learning models may tend to overfit the dataset, and hence, domain knowledge is needed to reduce the misinterpretation gap.
- The data should comply with regulations and legal concerns and should be generalized and interpretable.

1.6 Report Structure

- Literature Review: In order to lay the groundwork for the project, the following chapter examines the research that has been conducted in the specified topic.
- Research Technique: Chapter 3 lays out the procedures followed to address the issue at hand by applying the CRISP-DM technique to determine if the data is informative or not, as outlined in Chapter 2.
- Design Specification: The design chapter specifies the new idea and proposed solution to the given challenge, introducing a fresh approach.
- Implementation: The fifth chapter gives a thorough analysis of the research and then describes the process for implementing the data.
- Evaluation: The evaluation results are covered in the sixth chapter, which also compares different models using different criteria for evaluation and examines the outcomes.
- Discussion and Conclusion: The final chapter presents the results, analyzes and discusses the problem's outcomes, lays out the constraints, and delves deeper into each plan's exploration of machine learning and artificial intelligence in the appropriate field, with a primary emphasis on the model's end-to-end deployment.

2 Related Work

2.1 Statistical Models

Chakraborty and Elzarka (2020) proposed a hybrid approach for construction cost estimation using a hybrid approach to estimate construction cost at the design phase and enhance the accuracy of predictions. Value-engineering analysis in construction is used to find the best alternatives as per the design to get optimal results and lower the cost of the project. The study focuses on building a Light Gradient Boosting (LGB) algorithm and a Natural Gradient Boosting (NGB) algorithm for probabilistic estimations of the project. The study used a mixed dataset from six different construction assemblies for the model training, and six different models were built in the research, such as Linear Regression, ANN, Random Forest, XG-Boost, LGB, and NG-Boost. The study used the RS Means assembly book as a reference to build the dataset with the original dataset, and the proposed hybrid approach attained an RMSE of 0.5 and 0.99 as the R2 score. For the model explain-ability, SHAP is used, which is based on game theory, and the features having large SHAP values are considered important features and can be used to make the best decisions.

Fung et al. (2020) used machine learning algorithms to predict the seismic retrofit cost based on the characteristics of the building. The cost provides an idea and estimation of building projects as it contains attributes like size, age, and type of the building. The study used a dataset from FEMA (Federal Emergency Management Agency) 156 and FEMA's software. The dataset has information about the cost and building characteristics available from past data, and the study focused on comparing different models to measure uncertainty. Two different models are used in the research where Standard Linear Regression and Generalized Linear Model (GLM), which highlight the importance of normal distribution of the data, and one model relaxes the same. Out of these models, the GLM model outperformed the previous model, and the inverse of normality decides the cost estimates. The given model was then applied to various federal buildings to understand the retrofit cost and demonstrate results based on that.

Oshodi et al. (2020) Construction economics is directly linked with the economic growth of a country, and defining a relationship between the two can help identify the economic growth and factors responsible for it. However, the given industry is highly fluctuating and requires much attention, and the study highlights the use of predictive modeling to understand these factors and find all the changes responsible. The given review addresses the question of finding the factors responsible for fluctuations in the volume of construction output and whether predictive modeling can be predicted based on these theories. There are various studies backed by the research showcasing the partitioning of the data for model building and evaluation. Different journals were searched, and the theme was identified using the construction output and construction demand keywords. Explanatory models and Predictive Models are built on the given topics, and interest rates are the most common variable for determining the construction output and other factors like population growth, GDP, and Government policies. The study highlights the use of non-linear modeling for building better forecast models and focuses on the factors that can influence construction investments.

Sobieraj and Metelski (2021) Highlighted various factors that are responsible for explaining the performance of a construction project in Poland. The study used some questionnaires, and a total of 197 participants were used in the research, all of whom were construction managers. The analysis shows the relevance of each factor in predicting the project's success. The following is divided into three groups: strong, medium, and weak. The study uses Bayesian Model averaging to find the key factors responsible for the success of the construction project, and the data is categorized to get a clear picture. The given model outperformed basic statistical models like OLS and can address uncertainties.

All these models are statistical models used in the given research that highlight the important factors, such as the base model used in the research. Now, the study will focus on shifting towards a better non-linear predictive model that can be used for construction cost estimation and overcome the issues with these models.

2.2 Machine Learning and Deep Learning Approaches

Castro Miranda et al. (2022) In any construction project, cost management is the one thing that defines whether the project is as per budget or not to define the success of the project. Generally, there are basic statistical methods that deal with multiplying the floor area by cost per meter. The study showcases the use of predictive analytics for cost estimation, highlighting the better accuracy of the models for the given topic. The study highlights a step-by-step approach talking about the goal, aka research questions, then the data collection strategy used, followed by data preparation and EDA, which is the most important step in the model building. Once the EDA is done, the choice of variables and methods decides the truth for the models. Then, these models can be evaluated based on evaluation and model selection, showcasing the results in terms of reporting. The study used a total of 46 studies with data from 1974 to 2022. The predictive models sustain better accuracy than the traditional models, and finding the cost drivers is the main thing in the analysis. The study also focused on the predictive accuracy of the models in elaborating the new predictions. The most common methods used are ANN, Case-Based Reasoning, Multiple Regression Analysis, Boosting Regression Trees, and SVM.

With a standard implementation process available for the machine learning models, research by Ashraf et al. (2023) talks about the use of machine learning and deep learning architectures. The researchers are focusing on finding the best model that can be used for cost modeling that is in an acceptable range and can be used as a reference model for predicting cost. The study highlights that Random Forest and ANN are two adequate ML techniques to be used for the given research.Ebtehaj and Bonakdari (2016). The research used data collected from 220 real projects in the regions of Egypt, and all the data about raw materials was recorded. A80:20 split was taken on the training dataset, and these two models were built into the research. With 128-layer input neurons and 64 neurons in the hidden state to generate the output based on the novel architecture. Random Forest was used as a significant model to find the feature importance, and MDI (Mean Reduction in Impurity) was used to find the features. The models are tuned to understand the importance of hyperparameters. Both models attained a high value of R2 above 99 and a very low MAPE of 0.044, showcasing the supremacy of random forest over ANN.

The deep learning models have shown an application in the given field, and Yun (2022) showcased the implementation of a Neural Network for a multioutput regression model for the construction cost prediction. The study predicted two or more values based on the network to find the sub-construction cost that would sum up to the final construction cost. The main theme of such a project is that the input variable shows a correlation with the output value. All the past studies done in the field are based on finding the construction cost based on various factors. The following research is based on sub-construction costs such as civil engineering, architecture, electricity, etc. The data was used from the Public Production Service (PPS), and a total of 908 cases were taken. Ten input values as influencing factors and 8 output sub-construction cost variables are chosen, and the output Is obtained that is scaled using the Z-scaler to reduce the deviation. The study showcases the type of method used, the hyperparameter, the subfactor, and the techniques used for pre-processing to define the output.

Saeidlou and Ghadiminia (2023) The study uses a deep neural network (DNN) to estimate the building's expenses. DNNs are well-suited to this job because of their reputation for accurately modeling input feature-target non-linear relationships. Several parameters, including the network's architecture and the number of layers and neurons per layer, are fine-tuned to maximize the accuracy of the cost estimate. The data set includes several factors that affect building budgets. Factors such as building area, application type, city hierarchy, unit cost of concrete and formwork, type of structural assembly, amount of superimposed load, and unit cost of concrete and formwork are considered essential factors in construction. These factors offer a thorough foundation for cost assessment and are prevalent in building construction. The data is normalized and sent to Deep Neural Network (DNN), which estimates building expenses by learning from the prepared data. The DNN can capture all the complex relationships between the input variables and the desired cost. A map of 94.67% of the independent variables to the target cost with a mean average percentage error (MAPE) of 11.60% was obtained in the proposed framework. It proves that the framework can outperform traditional and rule-based methods when it comes to producing accurate and dependable cost estimates.

2.3 (XAI)Explainable Artificial Intelligence

With the curse of black box models. All AI models are shifting towards explainable models, with techniques coming like SHAP, ELI5, LIME, and others. These techniques are used to explain the weights and each feature for better explainability and decision-making Lim (2019). To get a better understanding Love et al. (2023) in their paper, they have showcased a narrative review of these techniques to be used in the construction field. The term explainability is all about understanding the functioning of the model, and interpretability refers to the passive property that makes sense for humans in their language. The following can be used in the field of construction as the following topic is not that popular in the field of construction but can be deployed to satisfy the client as well as stakeholder demands. It can also be used in the planning process and fusion of data, and the model can create wonders in the given field.

Yoo and Kang (2021) For any construction project, the model is first designed using the CAD(Computer-Aided Design) software to get the analysis and quotation for the project. In manufacturing, the cost is calculated based on the CAD drawings, and the following cost is only 5%, but it helps in determining 70% of the cost of the project. Out of these, deep learning models are trained on the CAD dataset to predict the cost but are hard to explain to stakeholders. The research focuses on building the models with XAI for explainability. The study used data collected from 3D CAD, and 34 different categories were considered in the modeling. With the Python open-source libraries, the mesh file is converted into a suitable form, and the data is normalized to be sent to deep learning models. Various baseline models are built, like Vox Net, Point Net, etc. The study proposed deep networks for regression and a convolutional layer to understand the data. The given model 3D Grad-CAM is used to visualize the 3D features and showcase that the cost estimation is lower in our model as compared to the state-of-the-art by 36% in RMSE and 26% in MAPE. The model showcased the visualization of each feature with respect to XAI to determine better manufacturing.

3 Research Methodology

3.1 Introduction

Data mining is a process of extracting patterns from the data that requires a lot of skills and knowledge. The following process requires standardization such that a business problem can be converted into data mining tasks. CRISP-DM, also known as the Industry Standard Process for Data Mining, is one approach that is used in data mining projects irrespective of the industry and technology stack used. The following process is standard and is most widely used as compared to other processes like SEMMA and KDD. *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW* (2021) 42% of the data science projects use the given methodology KDnuggets (2018) as it is easy to apply and shows complexity in terms of scope. The reason these methodologies are required in a data mining project is because the projects require a mix of tools and skills. For better implementation and project management, a process model can be used to understand the phases and the interaction between these phases, as shown in Fig 2.

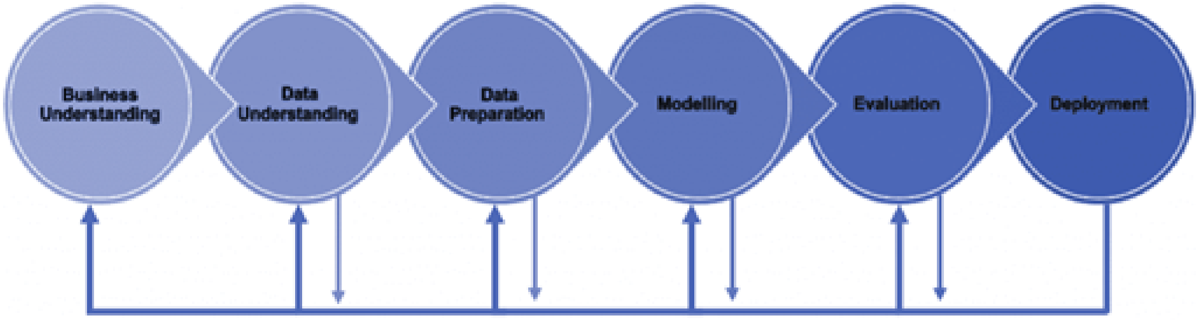


Figure 2: CRISP_DM

There are a total of 6 phases in the given process that are described as follows:

- Business Understanding – Define the business objectives and questions.
- Data Understanding – Find the perfect data to implement the given problem and convert the business problem to a data problem. Also, understand whether the data fulfills the requirement as per business or not.
- Data Preparation – Perform data transformations for better modeling perspectives.
- Modeling – Find, build, and train the model that meets the project requirements
- Evaluation – Find the results and evaluate them on accuracy as well as business metrics.
- Deployment – Launch the model in the working environment and make the decision-making process better for the business.

3.1.1 Business Understanding

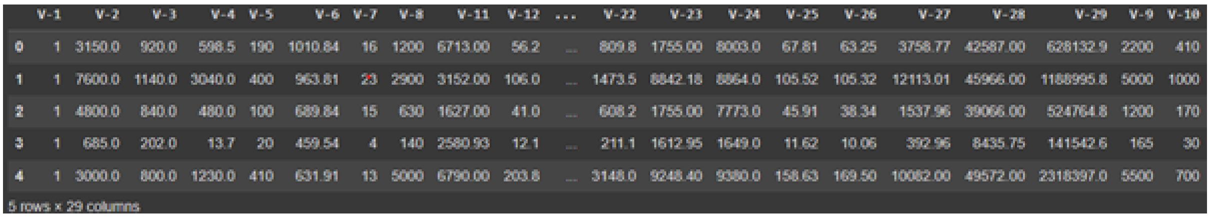
The first step in the data mining process model is business understanding, where one should understand the problem that they are working on. The given phase set the foundation of the whole project. Our research highlights predicting the cost estimation of a project and automating it by learning from past data. The main aim of the research is to predict the prices in the early stages so that the budgets will not overrun. Also, the model will help identify early risk in cost estimation and optimize resource allocation while defining cost. The following will provide a competitive advantage in terms of business domain. The following step also provides the project plan where one can understand the traditional methods used in the research as discussed in the literature and find the data availability and all the machine learning and predictive analytics architecture that has to be used in the research such that the given project can be a success. One of the major things is to make the models interpretable and explainable so that the stakeholders can understand them and make the right decision.

3.1.2 Data Understanding

The next step in the given process is to understand the data. The main objective is to collect the relevant data that can be used in the research. The following phase focuses on collecting the right data, exploring it, and assessing it as per the business problem. The dataset used in the research was taken from the UCI repository Rafiei (2018) The dataset contains a total of 372 entries and 109 columns, out of which, based on the data dictionary, 29 columns are selected to predict the cost of the residential building projects. The dataset contains all the construction costs and sales prices of several projects as well as the economic variables for the real estate single-family apartments in Iran. Once the data is loaded, the first step is to understand the data as well as the variables of the data to ensure that data is in a consistent format and to understand the data dictionary for better model building. One major step to understand here is that the following data depicts the real world, should be complete as per the business problem, and is relevant to the construction industryCholakis (2020). All these things are verified as per the data dictionary.Kreo (2023)

3.1.3 Data Preparation

The third step is the most important in data mining as, most of the time, it goes here, and the following decides the accuracy of the models. The data used in the research looks like this in Fig 3, and most of the data here is continuous data, and one has to understand the data to get better statistical summaries. Based on the reference paper, the given data was subsetted as per the business problem. All the variables starting from index 4 to 31 are chosen as the independent features, and 107 is the dependent feature.

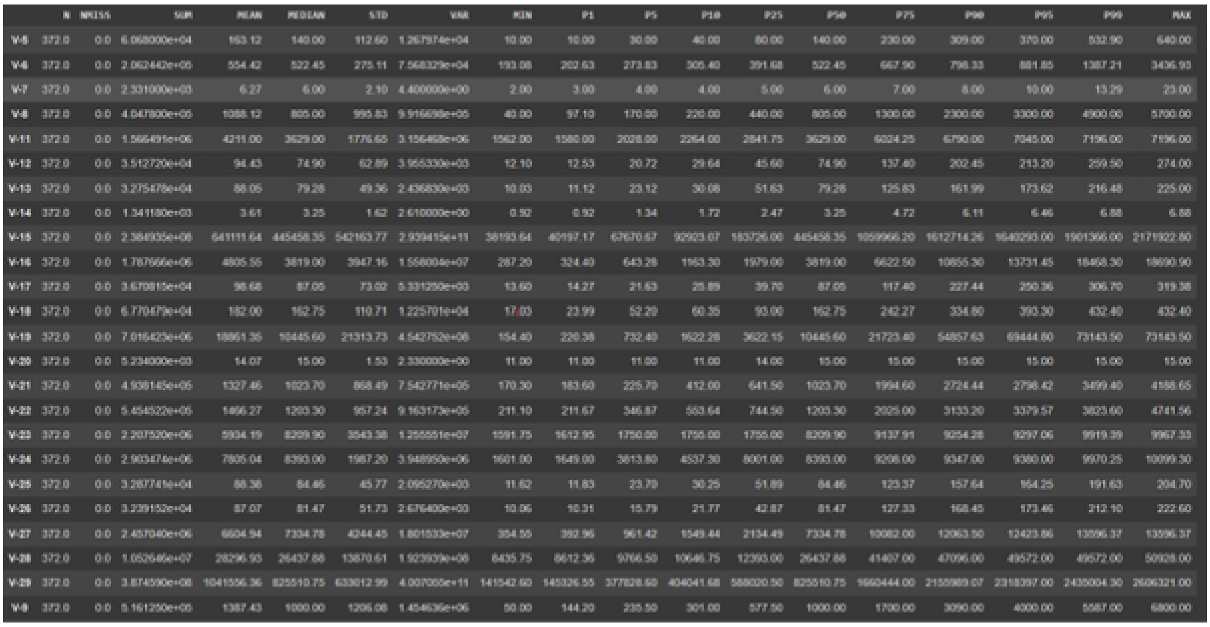


	V-1	V-2	V-3	V-4	V-5	V-6	V-7	V-8	V-11	V-12	...	V-22	V-23	V-24	V-25	V-26	V-27	V-28	V-29	V-9	V-10
0	1	3150.0	920.0	598.5	190	1010.84	16	1200	6713.00	56.2	...	809.8	1755.00	8003.0	67.81	63.25	3758.77	42587.00	628132.9	2200	410
1	1	7600.0	1140.0	3040.0	400	963.81	23	2900	3152.00	106.0	...	1473.5	8842.18	8864.0	105.52	105.32	12113.01	45966.00	1188995.8	5000	1000
2	1	4800.0	840.0	480.0	100	689.84	15	630	1627.00	41.0	...	608.2	1755.00	7773.0	45.91	38.34	1537.96	39066.00	524764.8	1200	170
3	1	685.0	202.0	13.7	20	459.54	4	140	2580.93	12.1	...	211.1	1612.95	1649.0	11.62	10.06	392.96	8435.75	141542.6	165	30
4	1	3000.0	800.0	1230.0	410	631.91	13	5000	6790.00	203.8	...	3148.0	9248.40	9380.0	158.63	169.50	10082.00	49572.00	2318397.0	5500	700

5 rows × 29 columns

Figure 3: Dataset

The next step is to understand the data statistically, as shown in Fig 4, and for that, the summaries are drawn to understand the data with a 360-degree perspective. One can observe that there are no missing and duplicate values in the dataset, but it contains some outliers that need to be cleaned. The percentile method was used to treat the outliers as it provides a robust and flexible approach to deciding the range.



	N	NMISS	SUM	MEAN	MEDIAN	STD	VAR	MIN	P1	P5	P10	P25	P50	P75	P90	P95	P99	MAX
V-5	372.0	0.0	6.068000e+04	163.12	140.00	112.60	1.267974e+04	10.00	10.00	30.00	40.00	80.00	140.00	230.00	309.00	370.00	532.90	640.00
V-6	372.0	0.0	2.062442e+05	554.42	522.45	275.11	7.568329e+04	193.08	202.63	273.83	305.40	391.68	522.45	667.90	798.33	881.85	1387.21	3436.93
V-7	372.0	0.0	2.331000e+03	6.27	6.00	2.10	4.400000e+00	2.00	3.00	4.00	4.00	5.00	6.00	7.00	8.00	10.00	13.29	23.00
V-8	372.0	0.0	4.047800e+05	1088.12	805.00	995.83	9.916698e+05	40.00	97.10	170.00	220.00	440.00	805.00	1300.00	2300.00	3300.00	4900.00	5700.00
V-11	372.0	0.0	1.566491e+06	4211.00	3629.00	1776.65	3.156468e+06	1562.00	1580.00	2028.00	2264.00	2841.75	3629.00	6024.25	6790.00	7045.00	7196.00	7196.00
V-12	372.0	0.0	3.512720e+04	94.43	74.90	62.89	3.955330e+03	12.10	12.53	20.72	29.64	45.60	74.90	137.40	202.45	213.20	259.50	274.00
V-13	372.0	0.0	3.275478e+04	88.05	79.28	49.36	2.436830e+03	10.03	11.12	23.12	30.08	51.63	79.28	125.83	161.99	173.62	216.48	225.00
V-14	372.0	0.0	1.341180e+03	3.61	3.25	1.62	2.610000e+00	0.92	0.92	1.34	1.72	2.47	3.25	4.72	6.11	6.46	6.88	6.88
V-15	372.0	0.0	2.384935e+08	641111.64	445458.35	542163.77	2.939415e+11	38193.64	40197.17	67670.67	92923.07	183726.00	445458.35	1059966.20	1612714.26	1640293.00	1901366.00	2171922.80
V-16	372.0	0.0	1.787666e+06	4805.55	3819.00	3947.16	1.558004e+07	287.20	324.40	643.28	1163.30	1979.00	3819.00	6622.50	10855.30	13731.45	18468.30	18690.90
V-17	372.0	0.0	3.670815e+04	98.68	87.05	73.02	5.331250e+03	13.60	14.27	21.63	25.89	39.70	87.05	117.40	227.44	250.36	306.70	319.38
V-18	372.0	0.0	6.770479e+04	182.00	162.75	110.71	1.225701e+04	17.03	23.99	52.20	60.35	93.00	162.75	242.27	334.80	393.30	432.40	432.40
V-19	372.0	0.0	7.016423e+06	18861.35	10445.60	21313.73	4.542752e+08	154.40	220.38	732.40	1622.28	3622.15	10445.60	21723.40	54857.63	69444.80	73143.50	73143.50
V-20	372.0	0.0	5.234000e+03	14.07	15.00	1.53	2.330000e+00	11.00	11.00	11.00	11.00	14.00	15.00	15.00	15.00	15.00	15.00	15.00
V-21	372.0	0.0	4.938145e+05	1327.46	1023.70	868.49	7.542771e+05	170.30	183.60	225.70	412.00	641.50	1023.70	1994.60	2724.44	2798.42	3499.40	4188.65
V-22	372.0	0.0	5.454522e+05	1466.27	1203.30	957.24	9.163173e+05	211.10	211.67	346.87	553.64	744.50	1203.30	2025.00	3133.20	3379.57	3823.60	4741.56
V-23	372.0	0.0	2.207520e+06	5934.19	8209.90	3543.38	1.255551e+07	1591.75	1612.95	1750.00	1755.00	1755.00	8209.90	9137.91	9254.28	9297.06	9919.39	9967.33
V-24	372.0	0.0	2.903474e+06	7805.04	8393.00	1987.20	3.948950e+06	1601.00	1649.00	3813.80	4537.30	8001.00	8393.00	9208.00	9347.00	9380.00	9970.25	10099.30
V-25	372.0	0.0	3.287741e+04	88.38	84.46	45.77	2.095270e+03	11.62	11.83	23.70	30.25	51.89	84.46	123.37	157.64	164.25	191.63	204.70
V-26	372.0	0.0	3.239152e+04	87.07	81.47	51.73	2.676400e+03	10.06	10.31	15.79	21.77	42.87	81.47	127.33	168.45	173.46	212.10	222.60
V-27	372.0	0.0	2.457040e+06	6604.94	7334.78	4244.45	1.801533e+07	354.55	392.96	961.42	1549.44	2134.49	7334.78	10082.00	12063.50	12423.86	13596.37	13596.37
V-28	372.0	0.0	1.052646e+07	28296.93	26437.88	13870.61	1.923939e+08	8435.75	8612.36	9766.50	10646.75	12393.00	26437.88	41407.00	47096.00	49572.00	49572.00	50928.00
V-29	372.0	0.0	3.874590e+08	1041556.36	825510.75	633012.99	4.007055e+11	141542.60	145326.55	377828.60	404041.68	588020.50	825510.75	1660444.00	2155989.07	2318397.00	2435004.30	2606321.00
V-9	372.0	0.0	5.161250e+05	1387.43	1000.00	1206.08	1.454636e+06	50.00	144.20	235.50	301.00	577.50	1000.00	1700.00	3090.00	4000.00	5587.00	6800.00

Figure 4: Statistical Summaries

Fig 5 shows that the Y-variable distribution is not normally distributed, and the data was transformed using the log transformations. The following transformation is used for the statistical machine learning models and is generally not required in a machine learning-based approach. One can understand the change in skewness values from -0.37 to 0.94 based on the transformation.

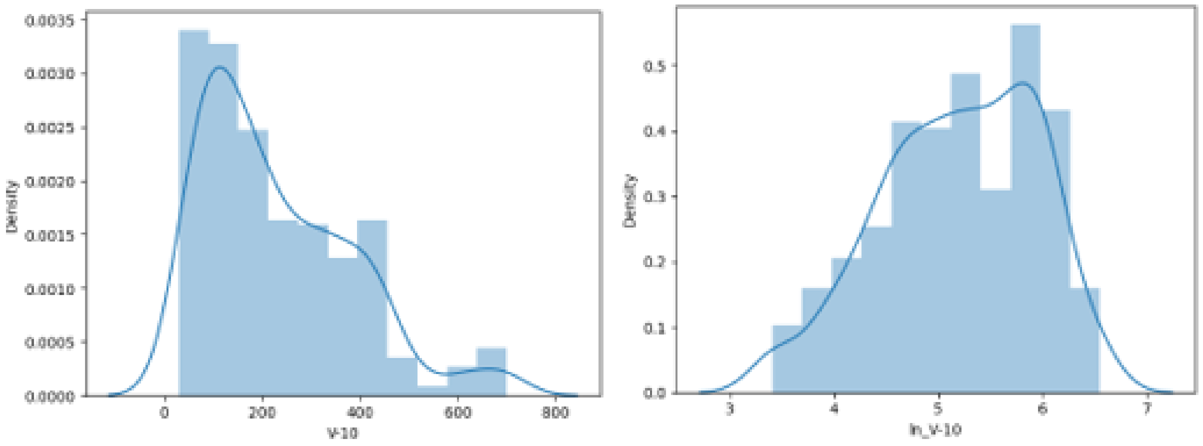


Figure 5: Original V/s Log Transformations

Fig 6 shows that There are many variables available that show a really high correlation with other variables, like V-13 (Wholesale price index (WPI) c of building materials for the base year) is highly correlated with the V-25 (Consumer price index (CPI) in the base year). Hence, based on the high correlation, one can confirm multicollinearity that can affect the model.

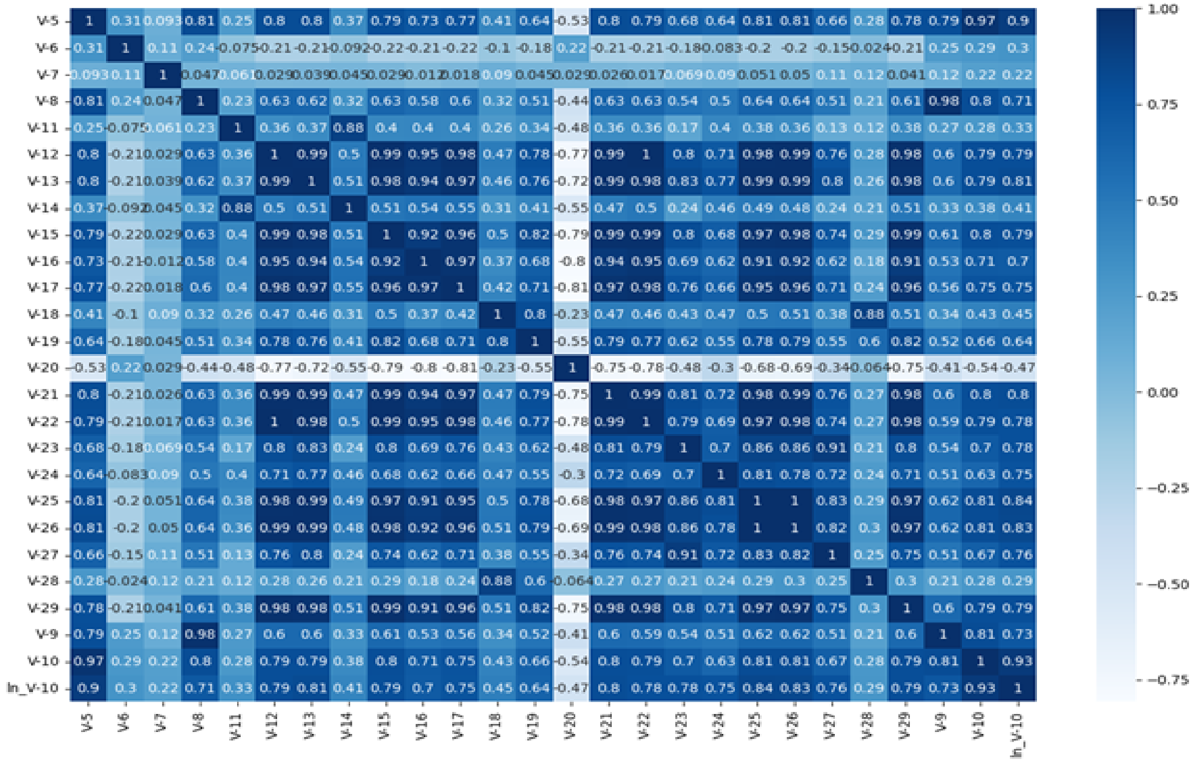


Figure 6: Heatmap

All the distribution plots so the variables also highlight the categorical behaviour like V-19 (The number of loans extended by banks in a time resolution) and V-20 (The interest rate for a loan in a time resolution), etc. one can observe the outliers in the dataset based on the boxplots. Now, the data has been prepared for model building and is made suitable for the formulas of statistical models. For machine learning models, the study used the RFE to confirm insignificant variables from the statistical models. Two different approaches are used in the research: statistical modelling and machine learning models. The main aim is to understand the relationship between the dependent and independent variables to make better predictions. The following techniques are discussed in the next chapter.

3.1.4 Modelling

The study focuses on building various statistical as well as machine learning models, showing a comparison between the basic models like Linear Regression and Machine Learning Linear Regression; for better interpretability, models like Decision Tree, Random Forest, and ANN are used as stated in the literature. There are a lot of other models like:

- **Transparent models** include linear regression, logistic regression, decision tree/CART, and KNN. These models are interpretable and are designed for clear explanations of predictions, making them more transparent and understandable.François Candelon and Martens (2023)
- **Opaque Models** are the black-box models that generate great accuracy and prioritize accuracy over clear explanations of the results. Random Forest, SVM, and MLP are included in these models.Jakovljevic (2017)

Implementation of AI in any business requires explainability and interpretability so that it can be used for decision-making. However, the studies show that more explainable models are coming into the market without compromising accuracy, and the following is crucial for deploying AI-based models. *Interpretable Machine Learning* (2024)

3.1.5 Evaluation

The evaluation step is the major step while building any model. The step is used to get all the details and accuracy of the models on the unseen dataset to get the best model out for implementation. Also, this stage answers all the questions that are stated in the research questions. Metrics like MSE, RMSE, and MAE are used in the research to obtain results.

- MSE, also known as the Mean Squared Error, averages the square of errors, which is the difference between the actual and predicted values. The model is sensitive to outliers and identifies where the model is making mistakes, but is less interpretable because of square units.
- RMSE is the root of the MSE scores and is a highly interpretable metric as it has the same unit as that of output and is sensitive to outliers.
- MAE or Mean Absolute Error is the absolute difference between the predicted and actual value. It is robust to outliers and is interpretable as it represents the average errors having the same unit as the data.

3.1.6 Deployment

Generally, the deployment phase shares the documentation of the final model deployment in production as well as a strategy for the same. Here, one has to share information with stakeholders and share the knowledge. However, our project makes it more comprehensive and explainable by using XAI via SHAP. Shapely Additive Explanations is a way to explain the output of any machine learning-based model. It tells whether the feature has increased or decreased the prediction.

On the other hand, feature importance can tell how important a feature is to a model. However, SHAP is used for individual features. It helps in finding the wrong predictions and understanding the features dominating those results, as well as in better debugging, Providing better human-friendly explanations.

4 Design Specification

- The following architecture is a three-tier architecture where the business owner or the contractors define their problem. Hence, the data scientist understands it and converts it into a business problem. This is the first tier, where the retrieval and storage of the data are taken care of, including historical construction costs, project specifications, and other factors.

- The next layer contains all the machine learning and deep learning architectures for the cost prediction and is responsible for training, evaluating and deploying th models.

- The last layer is the presentation layer, which handles the user interface, shares the input project details and views the cost predictions for better decision making using XAI.

5 Implementation

5.1 Introduction

The chapter throws light on the implementation of CRISP-DM in predicting the residential cost of buildings based on various attributes and finding the best model that can understand the relationship between these variables. The study used basic libraries like NumPy and Pandas for Data Analysis and Seaborn and Plotly for the visualization to understand the distributions of the data and any discrepancy in the data, which is not directly possible by statistical analysis. Also, a comparison between the deep learning models, machine learning models and statistical models is made using libraries like Sci-Kit Learn, SciPy, Stats models and Keras.

5.2 Feature Extraction

For any model, feature extraction is the major step as it impacts the model's performance and the addition of unwanted features. By removing the unwanted features, one can attain better accuracy and reduce the overfitting. Also, the training would be faster as the model does not have to build insignificant relationships. This will also help in making the model more robust. Our study has incorporated RFE, known as Recursive Feature Elimination, which is used to enhance the model performance by selecting the most relevant features. Based on the statistical analysis, a total of 15 features are selected, where the features are ranked based on the model performance and are removed with the least important features. The following is an iterative process. Random Forest is used as the model for finding important features. It provides an inbuilt method for measuring the feature importance and is more robust to overfitting because of its base concept of Ensemble.

5.3 Model Training

Once the data is prepared as per the problem, these four different models are implemented in the research to make it a comparative benchmark between various approaches of AI.

5.3.1 Ordinary Least Squares

The first method is an old statistical method used in the research for estimating the parameters of a linear regression model. The model is based on minimizing the errors, which is the difference between the observed value and the predicted value. The model uses a straight-line equation and finds the best-fit line that minimizes the distance between the actual predicted value and provides the most accurate prediction for the dependent variable. The reason the method is used is to build a relationship between the variables, and it is a straightforward, explainable approach. However, it has certain assumptions, such as the linear relationship between the dependent and independent variable, the normality of the Y- variable, and no multicollinearity of the independent variables. The data was prepared to address all these assumptions, and the first model was built with all the variables, where the data was divided into a ratio of 70:30. The formula was built, stating a high adjusted R2 score of 0.979. However, there are various insignificant variables in the dataset, as confirmed from the summary of the model and all the variables with high p-values are removed, and the second model achieved an Adjusted R2 score of 0.975.

5.3.2 Linear Regression

The linear regression model, when implemented via machine learning, works similarly, but it uses the gradient descent approach to find the best coefficients. It takes one or more independent variables and uses them to predict a continuous target variable. It operates when the relationship between the independent variables and the dependent variable is linear. The line that minimises the discrepancy between the model's predictions and the observed data is known as the best-fit line. Usually, the method of least squares is used to find the best-fit line.

A standard metric for evaluating the performance of the linear regression model is the Mean Squared Error. To measure the discrepancy between the expected and actual data, the Mean Squared Error (MSE) is often utilised as the cost function. This cost function should be minimised. Finding the best coefficients and intercepting through iteratively altering them to minimise the cost function is the goal of gradient descent, an optimisation process used in a machine learning-based approach.

5.3.3 Random Forest

Random Forest is the most used algorithm in machine learning that combines multiple weak learners, especially decision trees, to reach the final result. The model is used for both the classification and the regression analysis and is an extension of ensemble learning known as bagging. The model utilizes bagging and takes random features to create decision trees that are uncorrelated with one another. The random subspace method that depicts the random subset of features is used to build decision trees with low correlation. For the prediction, the model aggregates the results from these trees by averaging in terms of regression. The model in the research is tuned based on the number of estimators, max depth, min samples split, and min samples leaf to design the model in a way that won't over fit.

5.3.4 ANN

The last model used in the research is Artificial Neural Network. The brain architecture inspires the model and consists of an interconnected group of neurons, the basic building blocks of ANN. The architecture consists of the Input Layer, Hidden Layer and the output layer, where the hidden layer is responsible for all the computations. The model learns through a concept of backpropagation, where the weights and biases are adjusted based on the error that is calculated at the output. The given process is iterative, and the model attains higher accuracy in complex problems and is flexible. Our study used ANN, which had 64 neurons at the start, followed by a layer of neurons 32 and 64, and Relu, which added non-linearity in the architecture to attain the results. The optimizer used to compile the model is MSE, which calculates Adam and loss to update the weights.

6 Results

6.1 Results

For regression analysis, all these models are evaluated using different metrics like MSE, RMSE, and MAE to understand the model that attained the minimum errors and to finalize the best model. Tables 1–6 describe all these evaluation metrics.

Table 1: Training MSE

Models	**MSE**
Linear Regression - Stats	1387.54
Linear Regression - ML	604.37
Random Forest	139.92
ANN	1295.05

Table 2: Testing MSE

Models	MSE
Linear Regression - Stats	1923.91
Linear Regression - ML	849.46
Random Forest	709.93
ANN	1690.45

Table 3: Training RMSE

Models	RMSE
Linear Regression - Stats	37.25
Linear Regression - ML	24.58
Random Forest	11.83
ANN	35.99

Table 4: Testing RMSE

Models	RMSE
Linear Regression - Stats	43.86
Linear Regression - ML	29.15
Random Forest	26.64
ANN	41.12

Table 5: Training MAE

Models	MAE
Linear Regression - Stats	18.05
Linear Regression - ML	18.33
Random Forest	7.68
ANN	28.54

Table 6: Testing MAE

Models	MAE
Linear Regression - Stats	24.2
Linear Regression - ML	20.08
Random Forest	18.87
ANN	32.27

Out of all the models built, the Random Forest model has attained the lowest scores on any matrix and is the best model. Also, one can see the patterns of errors obtained on all the evaluation metrics where the mean squared errors show the maximum values.

6.2 Explainable AI (XAI)

The last phase in the CRISP-DM model is model deployment. However, to have a better understanding, our study is focused more on better results and explanations. As the Random Forest and ANN models are more black box models, their interpretability and explainability are really low, and the study has focused on using the XAI framework SHAP.

Random Forest model is used in the XAI framework SHAP for model explainability on the test dataset. The model predicts an almost accurate value of 221.179 and showcases all the features that increase or decrease the prediction. V_5 (cost at the time of starting the project) is negatively impacting the cost price, and it is obvious that the starting cost is just an estimation and is V_6 (cost in the base year). The factors that have a positive impact are V_3(Lot Area), V_25(Consumer Price Index), V_9(Actual Sales), V_7(Duration), V_2(Floor Area), V_21The average construction cost of buildings by the private sector at the beginning of the construction V_27(Stock Market) and others, as shown in Fig 7. Hence, one can observe that most of the factors are economical and dependent on one another.

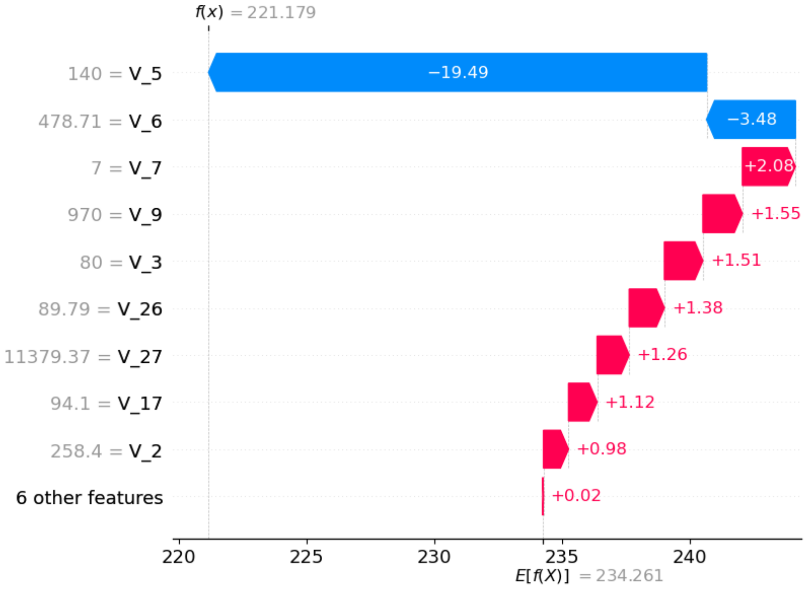


Figure 7: SHAP Values of Features

The next chart in Fig 8 used is the summary plot showcasing the SHAP values for a feature for the whole dataset and highlighting the value of the feature in giving output, as whether it is decreasing or increasing. The blue area denotes negative impact, and the red area denotes positive impact on the dataset. The model has a high value on the right side, showcasing the importance of the given feature.

Fig 9, The last plot used in the research is the dot plot model, where the impact of each variable is shown on the whole dataset. Here, one can observe that feature V-5 is the most impactful feature and is at the top, and V-11 and other features are least important, where the more the red value is, the more the feature importance increases or decreases. All these features can be correlated with the other maps for a better understanding, and based on that, decisions can be made in building the budget for a construction project.

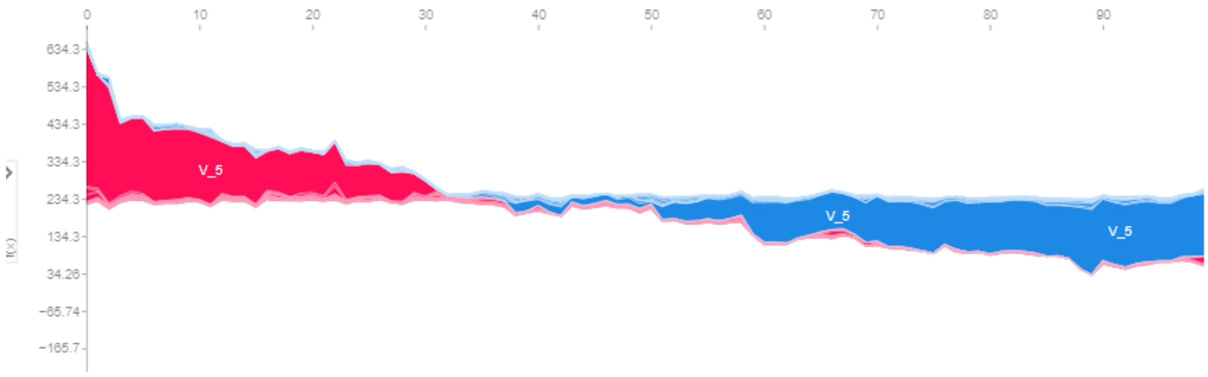


Figure 8: SHAP Summary Plot

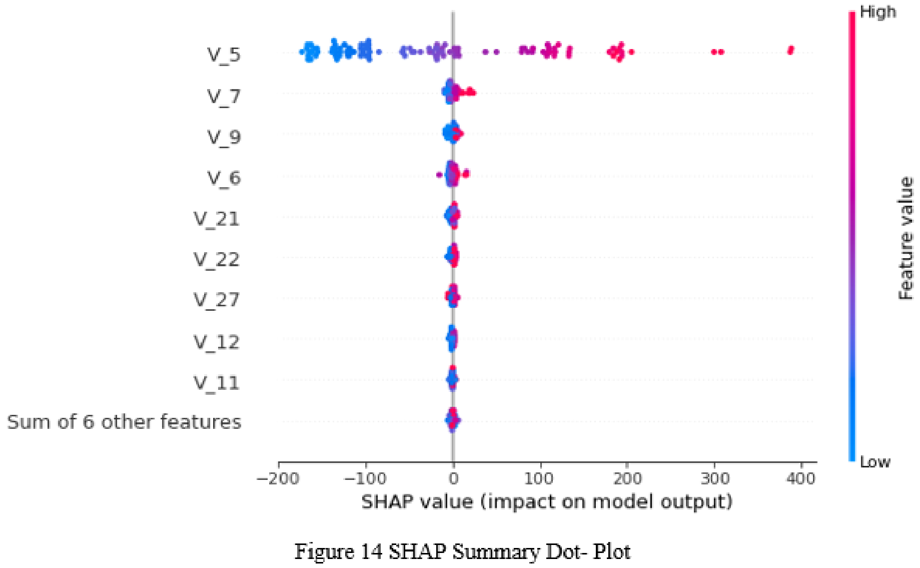


Figure 9: SHAP Summary Dot- Plot

6.3 Critical Evaluations

The model is evaluated on the evaluation metrics as well as the XAI. The model needs to answer all the research questions to be equally validated on the business problem.

- **RQ1** The study used a real-world dataset collected from the UCI repository and, with the help of XAI, found the best features that can predict the cost of a household project. The study may have opted for feature importance, but that only tells about the important features, not the ones that affect the predictions. The given model can be used to make better decisions.

- **RQ2** The research showcases the important features that affect the cost, and most of them are economic factors or factors related to the construction site that the contractor and the architect can use to predict the cost and maintain budgets.

- **RQ3** Study highlighted that the Random Forest model is the best model for the given research. The literature highlights the same model, and also the study showcases a benchmark comparison between machine learning, deep learning, and statistical analysis and showcases which type is better for the given problem.

- **RQ4** Different metrics are used to evaluate the model where Random Forest tops the charts. The common theme of all the metrics is to showcase the summed error from the actual value

7 Conclusion and Future Work

The study showcased the implementation of various modelling techniques for construction cost estimation. The main aim of the study was to identify the most effective approach that can be used to predict the cost of an attention project in the early stages so that perfect planning and budgeting can be made for successful project completion

- Different models like statistical, machine learning and deep learning models are built and evaluated to get the prediction accuracy. Random Forest demonstrated the superior accuracy performance because of their ensemble nature and ability to capture the patterns in small chunks.
- The study showcased the feature importance and highlighted factors like floor area, consumer price index, lot area, duration taken to complete a project, and the average construction cost, has a high impact on the cost and also cost at the time of starting the project, is negatively impacting the prediction.
- Machine learning models can improve the accuracy and timeliness of cost estimates, which in turn can help avoid budget overruns and make better use of available resources. All parties involved in the construction business, including project managers, contractors, and stakeholders, should consider this.

The study gave promising results based on the dataset provided and showcased a standard way of implementing the project; however, several areas can be used in the future to enhance the accuracy in real time.

- The data quality and the granularity can be improved in the future, where more focus can be given to the comprehensive data sets that include real-time data sets from construction sites to improve the model accuracy.
- Different techniques like CAD outputs, building information, and IoT data can be used to predict costs in real-time and maintain budgets.
- Complex models show higher accuracy but sometimes compromise on the black box nature to make the model better and understandable. Different XAI techniques can be used.
- New adaptive learning models can be built that can add to the changing environment and the conditions of the new data set in future online learning algorithms can be built to make the model up to date with the latest construction practices and economic conditions.
- Economic factors and environmental sectors can also be considered to define the cost estimation models to be more holistic and better.

References

Ashraf, A., Rady, M. and Mahfouz, S. (2023). Price prediction of residential buildings using random forest and artificial neural network, *HBRC Journal* **20**: 23–41.

Business Wire (2021). Global construction industry report 2021, https://www.businesswire.com/news/home/20210111005587/en/Global-Construction-Industry-Report-2021-10.5-Trillion-Growth-Opportunities-by-2023---ResearchAndMarkets.com. Accessed: 2024-08-02.

Castro Miranda, S. L., Del Rey Castillo, E., Gonzalez, V. and Adafin, J. (2022). Predictive analytics for early-stage construction costs estimation, *Buildings*, Vol. 12, MDPI.

Chakraborty, D. and Elzarka, H. (2020). A novel construction cost prediction model using hybrid natural and light gradient boosting, *Advanced Engineering Informatics* .

Cholakis, P. (2020). Characteristics of a reliable construction cost estimate, https://www.linkedin.com/pulse/characteristics-reliable-construction-cost-estimate-peter-cholakis/. Retrieved August 6, 2024.

Ebtehaj, I. and Bonakdari, H. (2016). A support vector regression-firefly algorithm-based model for limiting velocity prediction in sewer pipes, *Water Science Technology* **73**.

François Candelon, T. E. and Martens, D. (2023). Ai can be both accurate and transparent harvard business review, https://hbr.org/2023/05/ai-can-be-both-accurate-and-transparent.

Fung, J., Sattar, S., Butry, D. and Mccabe, S. (2020). A predictive modeling approach to estimating seismic retrofit costs, *Earthquake Spectra* **36**: 875529301989171.

Interpretable Machine Learning (2024). https://emergingindiagroup.com/interpretable-machine-learning-making-black-box-models-transparent/.

Jakovljevic, P. (2017). Opaque vs. transparent ai explained by pegasystemsi.
URL: *https://www3.technologyevaluation.com/research/article/opaque-vs-transparent-ai-explained-by-pegasystems.html*

KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW (2021). https://www.studocu.com/in/document/jawaharlal-nehru-technological-university-hyderabad/computer-science-and-engineering/r4-kdd-crisp-semma/65233033. Retrieved August 5, 2024.

KDnuggets (2018). Poll: Data mining methodology, https://www.kdnuggets.com/polls/2007/data_mining_methodology.htm. Retrieved August 5, 2024.

Kreo (2023). Process of cost estimating in construction, https://www.kreo.net/news-2d-takeoff/process-of-cost-estimating-in-construction. Retrieved August 3, 2024.

Lim, Y. (2019). Ac295 ac295 advanced practical data science, http://yongwhan.github.io/. Retrieved August 5, 2024.

Love, P. E., Fang, W., Matthews, J., Porter, S., Luo, H. and Ding, L. (2023). Explainable artificial intelligence (xai): Precepts, models, and opportunities for research in construction, *Advanced Engineering Informatics* **57**: 102024.
URL: *https://www.sciencedirect.com/science/article/pii/S1474034623001520*

McKinsey (2023). Accelerating growth in construction technology, https://www.mckinsey.com/industries/private-capital/our-insights/from-start-up-to-scale-up-accelerating-growth-in-construction-technology.

Miranda, S., del Rey Castillo, E., Gonzalez, V. and Adafin, J. (2022). Predictive analytics for early-stage construction costs estimation, *Buildings* **12**: 1043.

Oshodi, O., Edwards, D., Lam, K., Olanipekun, A. and Aigbavboa, C. (2020). Construction output modelling: a systematic review, *Engineering Construction Architectural Management* .

Predictive Analytics in Preventing Construction Disasters (2024). https://www.captechu.edu/blog/predictive-analytics-preventing-construction-disasters. Retrieved July 31, 2024.

Rafiei, M. (2018). Residential Building, UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5S896.

Rafiei, M. H. and Adeli, H. (2018). Novel machine-learning model for estimating construction costs considering economic variables and indexes, *Journal of Construction Engineering and Management* **144**.

Saeidlou, S. and Ghadiminia, N. (2023). A construction cost estimation framework using dnn and validation unit, *Building Research Information* **52**: 1–11.

Sobieraj, J. and Metelski, D. (2021). Quantifying critical success factors (csfs) in management of investment-construction projects: Insights from bayesian model averaging, *Buildings* **11**(8).

Yoo, S. and Kang, N. (2021). Explainable artificial intelligence for manufacturing cost estimation and machining feature visualization, *Expert Systems with Applications* **183**: 115430.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417421008472*

Yun, S. (2022). Performance analysis of construction cost prediction using neural network for multioutput regression, *Applied Sciences* **12**: 9592.