IRJEMS International Research Journal of Economics and Management Studies
Published by Eternal Scientific Publications
ISSN: 2583 – 5238 / Volume 3 Issue 4 April 2024 / Pg. No: 117-124
Paper Id: IRJEMS-V3I4P117, Doi: 10.56472/25835238/IRJEMS-V3I4P117

Original Article

Predicting Construction Price Index Using Deep Learning Method

[1]Bui Anh Tu, [2]Ngo Tri Thu
[1]Faculty of Economics and Management, Thuyloi University, Vietnam
[2]Joint Stock Commercial Bank for Investment and Development of Vietnam

Received Date: 16 March 2024 Revised Date: 24 March 2024 Accepted Date: 01 April 2024 Published Date: 20 April 2024

Abstract: *The Construction Cost Index (CCI) has been widely used to forecast project costs. In recent years, predicting the construction cost index of construction projects has become a significant research topic in the global construction management field. This study introduces various approaches and forecasting models to predict the construction cost index and evaluates the optimal mechanisms of machine learning and deep learning models. The primary purpose of this research is to provide stakeholders in the construction industry with a reliable tool for estimating construction costs for upcoming projects, especially in the current high-inflation environment. The method employed in this study is the Long Short-Term Memory (LSTM) model in the field of deep learning. The results of this study will serve as a useful reference for forecasting construction cost indices in the near future.*

Keywords*: Construction Cost Index, deep learning, Long Short Term Memory.*

I. INTRODUCTION

The price index provides information about changes in the costs of industries caused by a combination of variables such as labor, materials, and equipment prices. The Construction Cost Index (CCI) plays a crucial role in accurately estimating project costs, budgeting during the project planning phase, and concurrently managing and controlling costs throughout the lifecycles of construction projects. In accordance with Circular 13/2021 issued by the Ministry of Construction, the Construction Cost Index (CCI) reflects the fluctuation of construction costs over time. The CCI is utilized for establishing and adjusting project investment totals, bid package estimates and prices, construction project budgets, contract price adjustments, and converting construction investment costs. The Department of Construction calculates the CCI based on the fluctuations of prices for construction materials, labor, and construction machinery and equipment, which are issued through regulations managing construction investment costs. The CCI is published by project type and cost structure (including the building cost index, equipment cost index, and other cost components). Cost factors include the building material price index, construction labor price index, and machinery and equipment price index, determined based on representative project categories and quantities for calculation. It is calculated as an average over the selected time period, excluding compensation costs, support and resettlement costs, construction period interest, and initial working capital for business production (if applicable). The unit of the construction price index is the percentage (%), and the cost structure used to calculate the construction price index must comply with the cost structure regulations for managing construction investment costs, being fixed until the original reference date changes.

Wang and Mei (1998) used ARIMA models to predict the construction price index. Additionally, Cheng, Hoang, and Wu (2013) employed a hybrid method to estimate the construction price index in Taiwan, building a predictive model based on the past 10 years of economic and social indices in Taiwan. However, their research methods heavily rely on input data and training processes, leading to potential inaccuracies in long-term predictions due to unusual fluctuations in economic and social factors.

In the United States, Shahandashti and Ashuri (2013) identified influencing factors on the construction price index, including the consumer price index, total domestic product, building permits, housing starts, money supply, production price index, crude oil prices, and employment levels in the construction industry.

II. RESEARCH METHODOLOGY

A) Determination of Construction Cost Index

The Construction Cost Index for projects is determined by the sum of the products of the weighted averages of construction costs, equipment costs, and other expenses with the respective Building Cost Index, Equipment Cost Index, and Other Cost Components Indices of the selected representative projects.

The Construction Cost Index for projects (1) is calculated using the following formula:

$$\mathbf{I = P_{XD}\, I_{XD} + P_{TB}\, I_{TB} + P_K\, I_K} \qquad (1)$$

In which:

- $\mathbf{P_{XD}, P_{TB}, P_K}$: Weighted averages of construction costs, equipment costs, and other expenses for the selected representative projects; The sum of these weighted averages is equal to 1.
- $\mathbf{I_{XD}}$, $\mathbf{I_{TB}}$, $\mathbf{I_K}$: Indices for the Building Cost, Equipment Cost, and Other Cost Components for the selected representative projects.

a. Categorization of the components in formula (1) is determined as follows:

The indices for the Building Cost, Equipment Cost, and Other Cost Components ($\mathbf{I}_{XD}$, $\mathbf{I}_{TB}$, $\mathbf{I}_K$) are determined according to the instructions provided in the section.

The weighted averages of construction costs, equipment costs, and other expenses ($\mathbf{P_{XD}, P_{TB}, P_K}$) are determined as follows:

The weighted averages of construction costs ($\mathbf{P_{XD}}$), equipment costs ($\mathbf{P_{TB}}$), and other expenses ($\mathbf{P_K}$) are determined by the arithmetic mean of the respective weightings of construction costs, equipment costs, and other expenses for the representative projects within each project category.

The weighting for construction costs, equipment costs, and other expenses for each representative project is calculated as the ratio of the construction costs, equipment costs, and other expenses of that representative project to the total of these costs for the project. The formula is determined as follows:

$$\mathbf{P_{XDi}} = \frac{\mathbf{G_{XDi}}}{\mathbf{G_{XDCTi}}} \qquad (2)$$

$$\mathbf{P_{TBi}} = \frac{\mathbf{G_{TBi}}}{\mathbf{G_{XDCTi}}} \qquad (3)$$

$$\boldsymbol{P_{Ki}} = \frac{G_{Ki}}{G_{XDCTi}} \qquad (4)$$

- $\mathbf{P_{XDi}, P_{TBi}, P_{Ki}}$: The weightings of construction costs, equipment costs, and other expenses relative to the total costs of the i-th representative project.
- $\mathbf{G_{XDi}, G_{TBi}, G_{Ki}}$**:** The construction costs, equipment costs, and other expenses of the i-th representative project.
- $\mathbf{G_{XDCTi}}$: The total construction costs, equipment costs, and other expenses of the i-th representative project. The data regarding construction costs, equipment costs, and other expenses for the selected representative projects are derived from collected statistical data.

B) Long Short-Term Memory Model

a. Definition of LSTM Network

Long Short-Term Memory (LSTM) networks, a type of artificial neural network belonging to Deep Learning, are specifically known for their ability to learn long-term dependencies. Introduced by Hochreiter and Schmidhuber in 1997, LSTM has since been enhanced and widely adopted by many in the field. Due to its exceptional performance across various tasks, LSTM has become increasingly popular.

LSTM is designed to address the issue of long-term dependencies in neural networks. Its inherent capability to retain information over extended periods is a default feature, requiring no additional training to achieve memorization. This means that the LSTM model's internal structure is inherently capable of remembering without the need for human intervention.

Every Recurrent Neural Network (RNN) takes the form of a sequence of repeated neural network modules. In a standard RNN, these modules have a simple structure, often a $tanh$ layer.

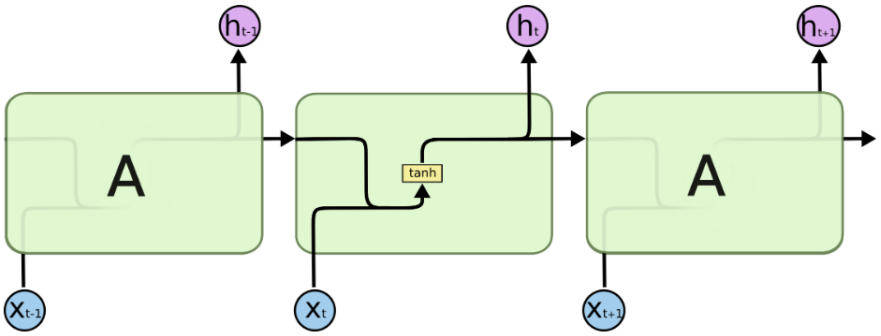


Figure 1: Recurrent Neural Network (RNN) Structure

LSTM also has a sequential structure, but its modules have a different architecture compared to the standard RNN. Instead of having only one neural network layer, they consist of up to 4 layers interacting with each other in a very unique way.

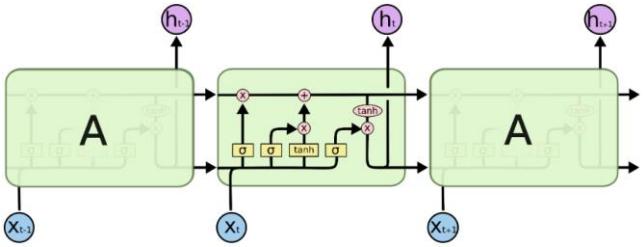


Figure 2: LSTM Network Structure

Every line in the figure transfers a vector from one node's output to another node's input. The yellow boxes are utilized for learning throughout each neural network structure, and the pink shapes indicate operations such as vector additions. Crossing lines signify concatenation while branching lines indicate that its content is copied and passed to different locations.

b. The core idea of LSTM

The cell state, represented by the horizontal line along the highest point of the picture, is crucial to LSTM. A conveyor belt of sorts describes the cell state. It barely slightly interacts linearly while passing via each chain link (network node). This allows information to flow through it without being easily changed.

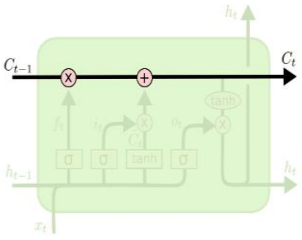


Figure 3: LSTM Horizontal Link along the Highest Point

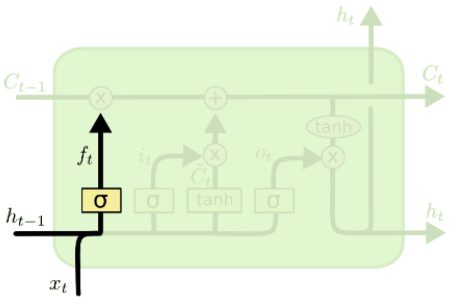
When necessary, the LSTM can add or subtract information from the cell state, all under the careful control of gate groups. These gates, which consist of a multiplication operation and a sigmoid layer, serve as filters though through which data is passed.

The sigmoid layer outputs a number in the range [0, 1], describing how much information can be allowed to pass. When the output is 0, it means no information can pass through, and when it is 1, it means all information can pass through it. An LSTM consists of three such gates to maintain and control the cell state.

c. Core Idea of LSTM

Choosing what data to discard from the cell state is the first stage in the LSTM process. This decision is made by the sigmoid layer, called the 'forget gate layer.' It takes inputs h_{t-1} and x_t and outputs a number between 0 and 1 for each number in the cell state C_{t-1}. Đ A output of 1 means to keep the entire content, while an output of 0 means to discard all the information. For example, in a language model predicting the next word based on all the previous words, the cell state

might contain information about the gender of a character, helping the model use the correct pronouns. However, when switching to a new character, we might want to forget the gender information from the previous one, as it's no longer relevant to the new context.

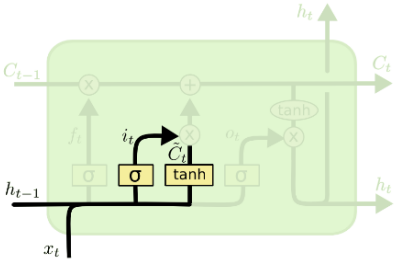


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Figure 4: First Stage in the LSTM Process

Choosing the new data to be stored in the cell state is the next stage. There are two components to this. 'Input gate layer' is a sigmoid layer that we use to determine which variables to update first. Second, a $tanh$ layer creates a vector of new candidate values, $\tilde{C}_t$, that could be added to the state. In the next step, we combine these two values to create an update for the state.

For instance, in the language model example, we might want to add the gender of the new character to the cell state and replace the gender information of the previous character.

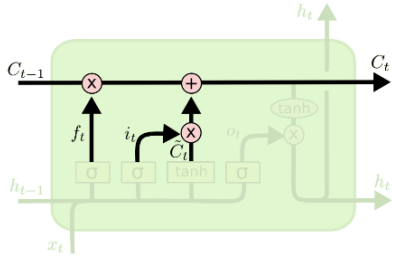


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Figure 5: Second Stage in the LSTM Process

The previous cell state C_(t-1) has to be updated to the new state C_t at this point. We've already chosen what to do in the previous steps; all that's left to do is carry it out.

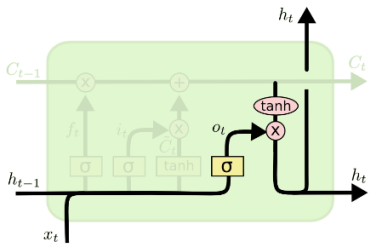
We will multiply the old state by f_t to forget the information we decided to discard earlier. Then, we add $i_t * \tilde{C}_t$. The new state obtained depends on how we decided to update each value in the state. In the language model task, as we determined in the previous steps, entails removing details regarding the gender of the previous characters and providing data about the gender of the new character.



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Figure 6: Third Stage in the LSTM Process

Lastly, we must choose the result we wish to produce. Although it will undergo additional filtering, the output value will be determined by the cell state. To determine which portion of the cell state we want to output, we first run a sigmoid layer. Then, we pass the cell state through a $tanh$ đ function to scale the values to the range [-1, 1] and multiply it by the output of the sigmoid gate to get the desired output value.



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Figure 7: Internal Operations of the LSTM Core

With the example of a language model, we can consider the subject to provide information about the tense of the following verb. For instance, if the output of the subject is singular or plural, we can infer how the form of the verb following it should be.

III. RESULTS

A) Input Dataset

The dataset used for predicting the Construction Cost Index (CCI) is collected from various sources, including the General Statistics Office of Vietnam, the State Bank of Vietnam, the Department of Construction of Ho Chi Minh City, and several other sources.

After the data collection process, the dataset was created with a size of 120 rows and 23 columns, equivalent to 2520 data points. Each row represents monthly data, and each column contains the values of individual variables (including 22 independent variables and 01 dependent variables). However, for ease of presentation in the text, only a few variables with significant influence on the construction cost index are included.
The variables in the dataset are shown in Table 1.

Table 1: Explanation of the Variables in the Dataset

No.	Symbol	Meaning
1	Month	Month of data collection (Data from January 2012 to December 2021)
2	Cpi	Consumer Price Index compared to the base month of January 2012
3	Gold	World gold price compared to the base month of January 2012 (USD)
4	Usd	Exchange rate of the US Dollar against the Vietnamese Dong
5	Xscn_hcm	Industrial production index of Ho Chi Minh City
6	Exp	Export price index compared to the year 2010
7	Imp	Import price index compared to the year 2010
8	Vni	VN-Index
9	Gdp	Gross Domestic Product
10	Vđt_xd	Total social investment capital in the Construction sector at constant Prices 2010
11	Vn_ir	Average interest rate (Medium and long term)
12	Oil	World crude oil price (USD)
13	CCI	Construction Cost Index - Target variable (dependent variable) to be predicted

The correlation matrix is a table that illustrates the correlation coefficients between variables when there are more than 2 variables in the dataset. Each cell in the table displays the correlation between two variables. The correlation matrix is often used before or after conducting Exploratory Data Analysis (EDA) to examine the correlations between factors and to diagnose multicollinearity issues in multivariate regression models. However, diagnosing multicollinearity is also somewhat subjective, as variables may exhibit multicollinearity even in the absence of a strong correlation.

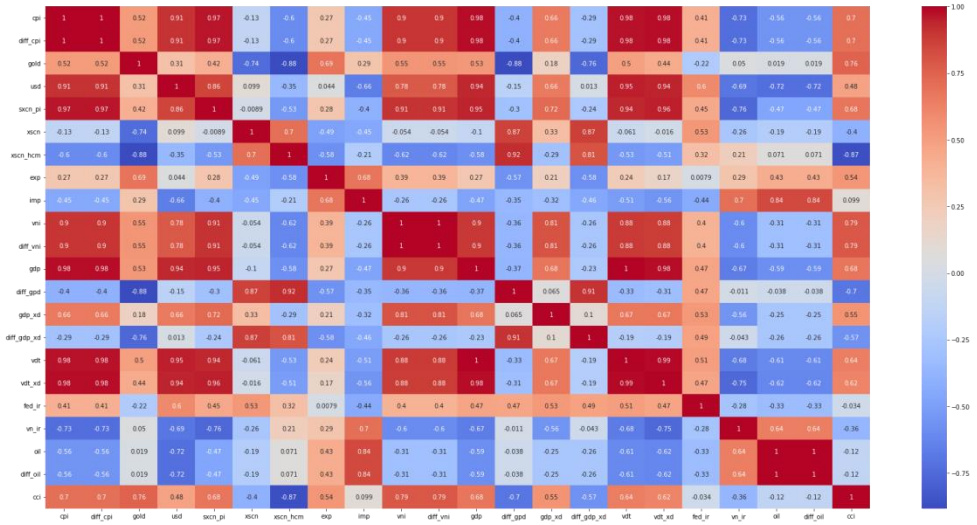


Figure 8: Correlation Matrix of Variables

Figure 4 represents the correlation matrix of variables in the dataset, where red indicates a positive correlation, blue indicates a negative correlation and darker shades represent stronger correlations between variables.

Thus, looking at the correlation chart, we observe that:

- Variables such as vni, cpi, xscn_pi, gdp, and vdt_xd have a high level of correlation with the dependent variable CCI.
- Variables such as oil, imp, fed_ir have a low level of correlation with the dependent variable CCI.

B) The Forecasting Results of the LSTM Model

The study has constructed nine machine learning and deep learning models to predict the construction cost index for comparing the effectiveness of the LSTM model with other models. The forecasting capabilities of these models are analyzed based on metrics such as MAE, MSE, RMSE, and R2. The results regarding the predictive performance of the LSTM model compared to other models are presented in Table 2.

Table 2. Forecasting performance of candidate models

No.	Candidate Model	MAE	MSE	RMSE	R2 Square
Base Model					
1	LSTM	0.7710	1.5235	1.2343	0.9536
Comparative Models					
1	Traditional Regression	1.4975	4.1044	2.0259	0.8396
2	Ridge Regression	1.6406	4.8070	2.1925	0.8122
3	Lasso Regression	2.1819	6.4197	2.5337	0.7492
4	Elastic Net Regression	1.5823	3.8701	1.9673	0.8488
5	Random Forest Regressor	0.8164	1.7691	1.3301	0.9309
6	Gradient Boosting Regressor	0.7310	1.6890	1.2996	0.9340
7	LightGBM Regressor	1.2414	3.1854	1.7848	0.8755
8	XGBooost Regressor	0.9016	2.0497	1.4317	0.9199

After training and evaluating the performance of nine different models, as shown in the table above, the study chose the LSTM model for the current task because it has the highest R2 value. A high R-squared value indicates that the model can better understand the underlying relationships in the data, thus having the potential to make better predictions with new data.

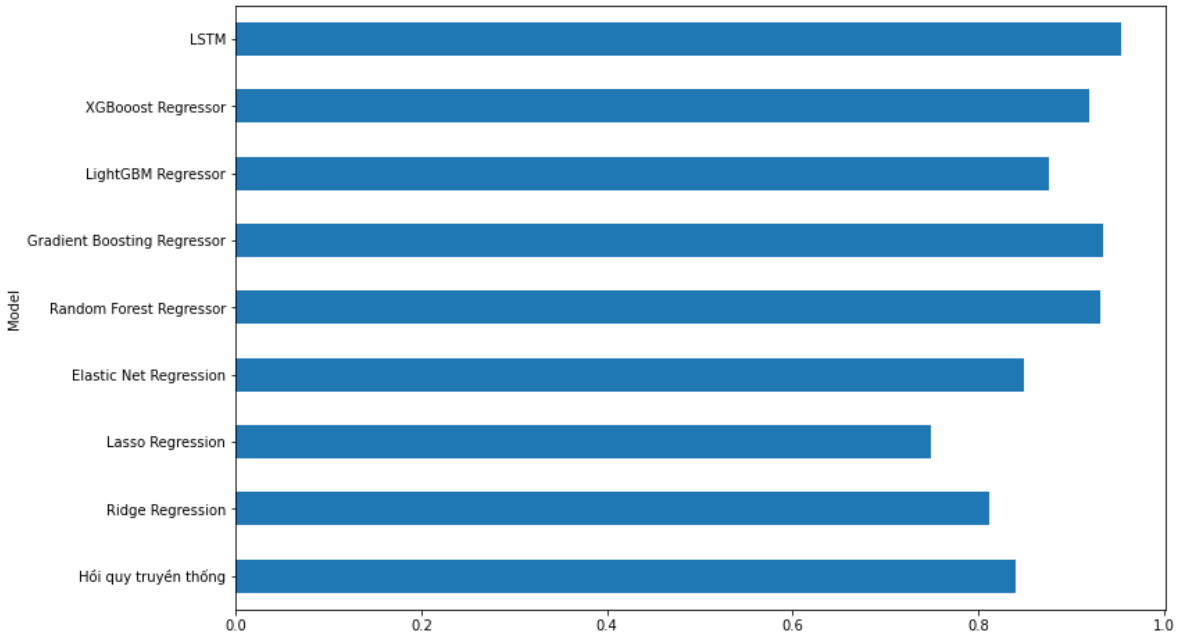


Figure 9: Comparison of Results of Nine Models Based on the R^2 Index

The LSTM model has a slightly higher R-squared value compared to the other models, indicating that it can capture the underlying patterns in the data more effectively, making it the best choice among the nine evaluated models.

Table 3: Forecast Results of the LSTM Model

Month	Fact	Forecast	Error	Deviation
Feb - 2020	121.78	121.232	0.548	0.450%
Mar - 2020	122.11	121.650	0.460	0.377%
Apr - 2020	121.94	121.979	0.039	0.032%
May - 2020	121.85	121.810	0.040	0.033%
Jun - 2020	121.89	121.720	0.170	0.140%
Jul - 2020	121.97	121.760	0.210	0.172%
Aug - 2020	121.66	121.840	0.180	0.148%
Sep - 2020	121.94	121.530	0.410	0.336%
Oct - 2020	121.96	121.810	0.150	0.123%
Nov - 2020	121.95	121.830	0.120	0.099%
Dec - 2020	121.95	121.820	0.130	0.107%
Jan - 2021	125.22	121.820	3.400	2.716%
Feb - 2021	126.38	125.109	1.271	1.006%
Mar - 2021	126.91	126.287	0.623	0.491%
Apr - 2021	129.57	126.827	2.743	2.117%
May - 2021	132.3	129.552	2.748	2.077%
Jun - 2021	134.03	132.368	1.662	1.240%
Jul - 2021	133.98	134.161	0.181	0.135%
Aug - 2021	133.91	134.109	0.199	0.148%
Sep - 2021	133.87	134.036	0.166	0.124%
Oct - 2021	134.66	133.995	0.665	0.494%
Nov - 2021	136.31	134.814	1.496	1.097%
Dec - 2021	136.65	136.528	0.122	0.089%

The predictions of the LSTM model for the civil construction price index in Ho Chi Minh City show relatively low discrepancies and errors, below 0.6% compared to the actual index published by the Ho Chi Minh City Department of Construction. It can be concluded that the LSTM model provides relatively accurate predictions, and the input factors influencing the construction price index are closely aligned with reality. Moreover, the LSTM model can be applied to predict other indices in the future.

C) A Typical Study in Ho Chi Minh City

The research has applied the forecasted construction price index from the LSTM model in the calculation for the completion of construction work of the structural part of the building at the 9 View Apartment project in District 9, Ho Chi Minh City.

The total estimated budget for the project is around 82 billion VND, with an expected construction investment period of approximately one year (2021). According to the conventional calculation method outlined in Circular 13/2021/TT-BXD, the construction price index used to determine cost contingency due to price fluctuations is calculated according to the following formula:

$$G_{DP2} = \sum_{t=1}^{T} (V_t - L_{Vayt})[(I_{XDCTbq} \pm \Delta I_{XDCT})^t - 1]$$

Where the average sliding rate (I_{XDCTbq}) is calculated in the following table:

Table 4: Calculation of the Average Sliding Rate

Indicator	Construction Price Index					
	Base Year: 2015					
	2015	2016	2017	2018	2019	2020
Construction Price Index for Civil and High-rise Buildings as published by the Ho Chi Minh City Department of Construction	100	96.94	98.98	100.74	100.97	103.38
Sliding Rate Factor (Index of the current year divided by the previous year)				1.0178	1.0023	1.0239
Average Annual Sliding Rate I_{XDCTbq}						1.0146

In this problem, it is assumed that the actual price sliding rate compared to the calculated average sliding rate, as published, is 0% per year. This assumption is made due to the absence of specific grounds to accurately choose this figure.

Table 5: Contingency Cost Calculation Based on the Sliding Factor

Content	Project Implementation Schedule
Year of Implementation	2021
Project Implementation Costs According to Progress without Sliding Factor (billion VND)	79.730.000.000
Accumulated Sliding Factor (G_{DP2})	1.167.591.388

Also, with the same data as presented above, but the contingency cost calculation due to the sliding factor in this case will take the results from the predicted construction price index model (2021) for computation.

The model was built to predict the construction price index over 11 consecutive years (from 2012 to 2021). However, only the results for the year 2021 are considered for the calculation. Since the construction project is planned to be implemented within 1 year (2021), the base year for calculating the contingency cost based on the sliding factor will be 2020.

Table 6: Calculating the Average Price Slippage from Model Forecast Results

Indicator	Construction Price Index						
	Base Year (2015)						
	2015	2016	2017	2018	2019	2020	2021*
The Construction Price Index for civil residential high-rise buildings, as announced by the Ho Chi Minh City Department of Construction	100	96.94	98.98	100.74	100.97	103.38	110.96
The sliding factor (price index of that year divided by the index of the previous year)					1.0023	1.0239	1.0733
The average annual sliding rate I_{XDCTbq}							1.0332

(*): The results predicted by the LSTM model are as follows

Table 7: Contingency costs for the price slide factor (based on the model's results)

Content	Project Implementation Schedule
Year of Implementation	2021
Cost of Project Implementation According to Schedule Without Price Slide (billion VND)	79.730.000.000
Accumulated Price Slide (G_{DP2})	2.643.668.909

The result of calculating the contingency cost for the price slide according to the model: 2,643,668,909 VND.

Table 8: Results of contingency cost calculation using different methods

Calculation Method	Contingency Cost
Circular 13/2021/TT-BXD	1.167.591.388
LSTM Model	2.643.668.909

From the above results, we can see that the contingency cost calculated based on the model's results and the conventional method instructed by the Ministry of Construction shows a significant difference. This difference is due to the annual price slide rate calculated according to the Ministry of Construction's guidance not accounting for the forecasted fluctuations of various factors such as fees and prices in the local and international areas compared to the average annual price slide rate.

IV. CONCLUSION

The data collection process has identified 7 socio-economic indicators to serve as input data for predicting the civil construction price index (CCI) in Ho Chi Minh City. These indicators include the Consumer Price Index (CPI), Gross Domestic Product (GDP), Basic Lending Rate (BLR), Exchange Rate, Total Import-Export Turnover, Total Construction Investment Capital, and Stock Price Index (VN Index). Based on the correlation matrix, the correlation between these factors and the CCI has been assessed, providing insights into the influence and importance of different socio-economic indicators on the CCI in Ho Chi Minh City.

The study has developed an idea and a short-term prediction model for the volatility of the civil construction price index in Ho Chi Minh City. The foundation has been laid for constructing an open-type model that can predict various economic and social indicators beyond the CCI. This model's architecture facilitates easy and accurate prediction of other indices based on the knowledge gained.

By creating a yearly dataset for a larger model, the study aims to have more comprehensive and diverse data, considering additional factors for predicting the construction price index. This approach will likely lead to predictions that closely align with actual values and are more reliable.

V. REFERENCES

[1] C.H. Wang, Y.H. Mei, "Construction Management and Economics," 1998, pp. 147-157.
[2] Dong, J., Chen, Y., & Guan, G., "Cost Index Predictions for Construction Engineering Based on LSTM Neural Networks," Advances in Civil Engineering, 2020.
[3] YasserElfahham, "Estimation and prediction of construction cost index using neural networks, time series, and regression," Alexandria Engineering Journal, June 2019.
[4] Phong Thanh Nguyen, Quyen Le Hoang, "Critical Factors Affecting Construction Price Index: An Integrated Fuzzy Logic," Munich Personal RePEc Archive, May 2020.
[5] Gouvêa, R. R., & Schettini, B. P., "Empirical estimates for the Brazilian total imports equation using quarterly national accounts data," EconomiA, pp. 250-271, 2015.
[6] Akintoye, A., Bowen, P., & Hardcastle, C., "Macro-economic leading indicators of construction contract prices," Construction Management & Economics, pp. 159-175, 1998.
[7] Thomas Ng, S., Cheung, S. O., Martin Skitmore, R., Lam, K. C., & Wong, "Prediction of tender price index directional changes," Construction Management & Economics, pp. 843-852, 2000.
[8] Wang, C.H., & Mei, Y.H., "Model for forecasting construction cost indices in Taiwan," Construction Management & Economics, pp. 147-157, 1998.
[9] Shahandashti, S., & Ashuri, B., "Forecasting engineering news-record construction cost index using multivariate time series models," Journal of Construction Engineering and Management, pp. 1237-1243, 2013.
[10] "Machine Learning Cơ Bản," [Online]. Available: https://machinelearningcoban.com/2016/12/26/introduce/.
[11] "Phamdinhkhanh Blog," [Online]. Available: https://phamdinhkhanh.github.io/deepai-book/ch_ml/index_RandomForest.html.
[12] "Understanding LSTM Networks - Colah's Blog," [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.
[13] "Model Selection - Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Model_selection.