Australian Housing Construction Cost Prediction Model:

A Critical Forecast for Contractors, Owner-Occupiers and Investors

Matt Stevens PhD MBA[1] and John Smolders AM FAIB[1]

Western Sydney University Kingswood, NSW 2751 Australia

Matt.Stevens@WesternSydney.edu.au

Abstract

Home ownership is a significant quality-of-life factor and a vehicle for wealth accumulation, while others view it as an investment opportunity. The cost of which consumes a considerable portion of Australian citizens' income. To sustain the housing market's positive dynamics, a cost-prediction model is needed so everyone can plan their lives and proactively adjust to any significant economic surprises. The researchers have endeavoured to create a forecasting methodology using a data and statistical-centric approach that is readily accessible and calculable for most citizens. We are currently publishing this lookahead each quarter This paper developed a multifactor algorithm that uses decades of residential cost input, both international and domestic. The researcher buttressed this forecast by adding a second prediction using machine learning to show the level of agreement between the two. This model aims to be one more consideration for new housing stakeholders that reduces the risk of cost changes. Looking forward, the model will be calibrated to assimilate changes as new trends emerge.

Keywords

Residential Inflation, Construction Cost, Unaffordable Housing, Home Expense, Building Escalation

1. Introduction

The Australian property market is significant. Recent surges in values have boosted the number of Australian millionaires to a global high. This translates to the second-highest median wealth in the world at $US268,000 ($AUS411,000), second only to Luxembourg (Courty 2025). Housing Investor's new loan commitments in the 12 months ending March 2025 were approximately AUD 129 billion. Owner–Occupier home buyers purchased AUD 210 billion in the same period, of which AUD 60.5 billion were to first-time home buyers (ABS 2025).

Prediction is a hazardous business. Any capitalistic market is chaotic, and construction appears to be more so. The probability of error is high. Constructing work involves many uncontrollable factors that affect core dynamics, such as government subsidy impacts, supply chain shocks, labour shortages, and rising design complexity, which have historically been correlated with inflation.

Construction cost prediction models, such as the one from WT Partnership (2024), use scenario-based modelling. These produce rounded predictions, i.e., "4% rise in construction cost". This approach is accepted in the business community. However, distilling individual factors enhances the antifragility or strengthening of built environment stakeholders. Once published, they can be adopted, improved, and recalculated by industry professionals for timely, beneficial decision-making. The chaotic nature of the construction industry indicates that static prediction models increase risk. So, machine learning has to be part of the methodology.

From industry experience, the researcher initially investigated 12 inputs, but upon statistical testing, they dis not. This observation-based starting point of factors is crucial. Without it, any theoretically based application risks an elegant but certain embarrassment. The most famous example is the correlation between ice cream consumption, hot weather, and police arrests. They are positively correlated but logically flawed.

Forecasting is an art and a science. Science is easier to teach than art. Scientists use a documented process structure of enquiry, while Artists use subjective and intuitive determinations internally. Science's methodology can be observed and documented. Art's process is contained within the forecaster and can only be described. However, Tetlock and Gardner (2025) assert that must be present in predicting the future. This paper only focuses on one.

2. Settings and Methods

Our research used historical data to identify significant correlations that predict price changes within a range. This, of course, would benefit construction contractors, property developers, and other stakeholders. Data science was the primary tool of this project. Our process involved transforming historical data into a credible predictive model. The authors engaged research teams from North Dakota State University and Western Sydney University to develop a residential cost prediction process, including inputs and an associated equation. The work presented here is a starting point. The selected data are from the Australian Bureau of Statistics and the Reserve Bank of Australia.

Rise and Fall (R&F) adjustments embodied in construction contracts are linked to the CPI. They have taken a backward step due to stable working conditions in the industry but were significantly inaccurate about the costs of building during the COVID pandemic. Additionally, the pandemic exacerbated company liquidations and incomplete projects (ASIC 2025). It is critical to create a construction-centric cost predictor to benefit Australians proactively and lessen this chaos

Data used to calculate R&F comes from different sources. The Australian Bureau of Statistics (ABS) publishes its quarterly Producer Price Indexes (PPI), one of which is an input to the House Construction Industry, widely publicised by construction magazines (NSWGov 2021). The PPI is historical and is not predictive. Before 2025, no publicly disclosed housing cost prediction equation for Australia was available, per the researcher's search of Google Scholar and Scopus. This paper presents a well-founded indicator that can be adapted to calculate cost fluctuations today and in the future; however, our formula and process serve as starting points.

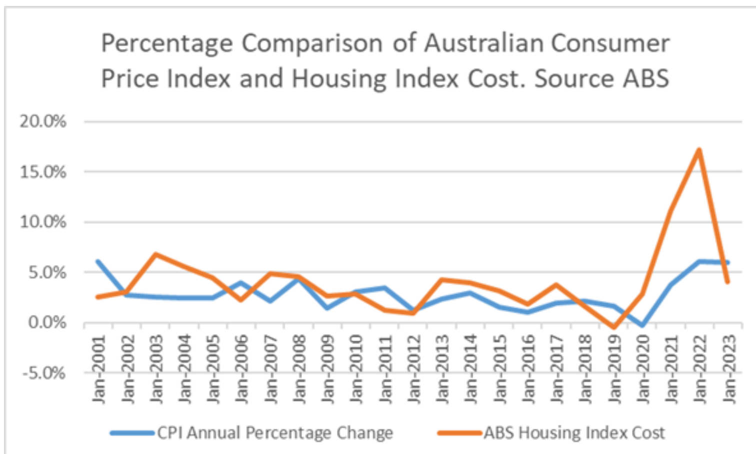

Figure 1. Comparison of the Consumer Price Index and Housing Cost Index (ABS 2025).

All predictive models experience one or more statistical problems. Often, these issues are structural or conditional and require disclosure of their nature and severity.

1. *Low signal/noise ratio:* The "rule" asserts that a minimum 15x events ratio is required for a low signal/noise ratio. This is not possible given the limited number of cases and years.
2. *Intercept:* One needs 96 observations to accurately estimate the intercept accurately, ensuring the overall predicted risk is within a ±0.1 margin of error of the actual risk with 95% confidence. This was not possible.
3. *Static Assumptions:* Another issue is that this model must assume that asset returns and volatilities remain constant over time. In reality, market conditions change, affecting returns and risk levels. For example, the changing nature of contracting includes the increased use of services (less direct labour), "smarter" equipment, innovative materials, and modernising methods.
4. *Single-Period Analysis*: Our design captures a single period (two decades). In reality, the built environment horizons vary, and long-term dynamics may differ from short-term trends. Additionally, there is a timing issue: when does a factor have its most significant impact on cost, such as immediately or a year later?
5. *Market Frictions:* The model doesn't consider changing market frictions, such as liquidity constraints, shifting consumer preferences, or regulatory creep.
6. *Data Reliance:* Historical data drives the model's output. Due to changing economic conditions and unforeseen events, future price changes may not accurately reflect past trends.

The goal of this project was to develop a forecasting model for the price index of housing producers and property investors using time series and machine learning techniques. Since housing prices can be influenced by various factors, such as economic conditions and “animal spirits”, the researchers created a model to not only understand patterns in price history but also account for the impact of external factors that can influence the Housing Producer Price Index. Given the complexities associated with time series forecasting and the influence of external factors, we aim to determine which model is most effective in predicting the Housing Construction Producer Price Index.

- ARIMA Model: An autoregressive integrated moving average model was initially tested without considering external influences, which resulted in limited predictive accuracy (Xue and Hua 2016)
- ARIMAX Model: Extension of the ARIMA model, incorporating various exogenous variables. The aim was to try to capture the influence of these external economic factors on the House Producer Price Index and determine which of the factors had significant predictive power.
- XGBoost is an optimized implementation of gradient boosting designed for speed and performance. It builds models sequentially, where each new model corrects the errors of the previous one (Ramraj 2016)
- Hybrid ARIMAX-XGBoost Model - The model combines ARIMAX to capture linear patterns and uses XGBoost to learn the leftover non-linear patterns from residuals.

2.1 Data Preparation

The most reliable data for this project was found from three online sources.

- Reserve Bank of Australia (RBA)
- Australian Bureau of Statistics (ABS)
- U.S. Federal Reserve Economic Data (FRED)

The dataset utilised for the model consists of quarterly historical data on the Housing Construction Australia Index, as well as various economic and construction-related indicators, ranging from September 1996 to December 2024

The original least squares regression model provided by the client gave us insight into some predictive variables that may help develop a model. The five key indicators in the original model provided to us included housing credit, the Australian employment rate, commodity prices, and base metal prices. These variables provided insight into the macroeconomic indicators that may influence the Housing PPI. Our approach took the five indicators into account and expanded on them. The goal was to understand the various factors influencing housing PPI and enhance our model's predictive power. To achieve this, new economic and construction-related data were implemented into the dataset. The latest indicators introduced into our model, which differ from the client's, include balance on goods, number of building approvals, index numbers for ceramic tile, steel, timber board, and joinery, wage growth for employees in the construction industry, and the percentage of unemployed Australians. After obtaining data sheets from the online database, the data was cleaned, transformed and aligned to ensure consistency in our dataset.

2.2 Feature Scaling and Percentage Change Calculation

Once all the necessary data was cleaned, it was merged into a single dataset. Before modelling, all the features of the dataset, except for the date and target variables (Housing PPI) were standardised using z-score normalisation. This is an essential step because it ensures that each indicator plays an equal role in training the model, preventing any single feature with a larger scale from having too much influence and creating a more balanced learning process. Thus, it allows the model to treat all variables uniformly during the training and prediction process. A new variable, Housing PPI percentage change, was calculated to represent the percentage change in Housing PPI from one time period to the next. This metric was introduced to help capture price volatility over time. To identify sudden price changes, a spike indicator was also introduced to flag periods when the percentage change exceeded 3 per cent from its last 3 months. By identifying these spikes, the model can capture volatile periods that indicate economic anomalies. The Spike indicator variable provides the model with an additional feature that can be used to adjust predictions during a period.

2. 3 Time Series Stationarity and Differencing

Ensuring that data is stationary is essential in time series forecasting. Stationary means that the mean or variance of the data remains consistent over time, which allows the model to capture the patterns better and the relationship overall, leading to a better forecast. Data needs to be stationary so that forecasting models may accurately capture the underlying pattern in the data. The augmented Dickey-Fuller (ADF) test was conducted on the housing PPI series to determine if the housing PPI series was stationary or not.

3. Results

Based on the output from the above image, the ADF test result showed a p-value of 0.99. The ADF test fails to reject the null hypothesis, indicating that the housing PPI series is non-stationary. Because the data showed it was not stationary, a transformation through differencing was applied to stabilise the mean and variance, to make the series suitable for modelling. Differencing is a method where the current value of the time series is subtracted from the previous value. This helps make the data easier to analyse because it eliminates trend components, allowing the model to focus on variation. The ADF test was conducted again on the difference series 'ts data diff'. The p-value was 0.046, confirming that the data were now stationary.

Multicollinearity occurs when two or more predictors in a model are highly statistically interrelated. That means the variables contain overlapping information. Having highly correlated variables makes it hard to determine the individual effect of each predictor, which leads to unreliable estimates and reduces the predictive power of the model, as the Pennsylvania State University [2018] explains. To address the issue of multicollinearity in our dataset, a variance inflation factor(VIF) test was conducted to detect multicollinearity among our predictors. The initial test conducted showed signs of multicollinearity among predictors.

Table 1. Variation Inflation Factor Test for Selected Potential Factors

```
> print(vif_results)
          credit_housing          commodity_price               base_metal        construction_rate
               11.953998                38.258844                 5.842382                 9.397727
             Broad_money          balance_on_good                   permit             ceramic_tile
              144.875763                12.320547                 3.309795                15.861577
    timber_board_joinery                    steel             ready_cemant wage_growth_total_quarterly
               92.272214                37.299431                57.513817                 5.828670
    Average_Unemployment          Spike_Indicator
                9.702215                 2.894623
>
```

VIF Values of over 10 indicated that our variables presented multicollinearity; however, they are considered unreliable. The variables showing signs of multicollinearity underwent a systematic removal process, where the variable with the highest VIF was removed one at a time. The researchers removed broad money, timber board prices, steel prices, ready cement price, and commodity price. This left a) Base Metals, b) Housing Credit, c) Employment, d) Wage Growth, e) Balance of Trade, f) Residential Permits, g) Construction Activity, h) Clay Materials Cost, and Average Unemployment. After this step, the VIF analysis no longer showed signs of multicollinearity. The resulting model operates with a set of independent variables, thereby enhancing its predictive power.

3. 1 Model Development and Analysis

As Mehandzhiyski [2023] notes, the ARIMAX model extends the ARIMA structure by incorporating exogenous variables that account for external factors affecting Housing Construction PPI. In this case, the model can account for various economic trends regarding housing construction costs, such as building permits as a representation of future construction supply, unemployment rates as an influence of demand, and base metal prices, as construction material costs are directly involved. By utilising the auto arima function, an ARIMA(1,0,0) structure was identified, indicating one autoregressive lag. This means the model effectively captures the time series component with a single lag in the AR term without needing additional differencing.

3.2 Residual Analysis

Residuals from the ARIMAX model were assessed using the Ljung-Box test. A p-value of over 0.05 suggests that the residuals don't show any strong autocorrelation. This indicates that the ARIMAX model has successfully captured the time-dependent patterns in the Housing construction PPI data.

3.3 XGBoost Model on ARIMAX Residual

To improve on the ARIMAX model, XGBoost was applied to the residuals to capture non-linear patterns not addressed by ARIMAX. By applying XGBoost to the ARIMAX residual, we leverage XGBoost's ability to detect non-linear relationships, capturing patterns that ARIMAX could not model. Cross-validation determined optimal hyperparameters for XGBoost:

- objective = "reg:squarederror"

- learning rate(eta) = 0.05
- max depth = 6
- column sample by tree = 0.7

From the feature importance plot for XGBoost, base metal prices emerged as the most influential variable, followed by housing credit and unemployment rates, highlighting the relevance of financial conditions in shaping housing demand.

3.4 Combining ARIMAX and XGBoost

By combining the two models, the hybrid approach leverages the linear strength of ARIMAX and the flexibility of XGBoost to capture nonlinear relationships. This results in a model capable of addressing a wider range of patterns in the Housing Construction Producer Price Index.

3.5 Training and Training Sets Split

To evaluate the model's generalisation and predictive power effectively, the dataset was divided into an 80 per cent training set and a 20 per cent test set. By using 80 per cent of the data for training, the model gets an understanding of patterns and trends in the data. The remaining 20 per cent is used for testing, acting as the "future" data the model hasn't seen before. This approach helps us determine if the model can make accurate predictions. This ensures that the model is not just memorising old data but can adapt to new inputs reliably.

- Training Set: RMSE = 0.685, R-Squared = 0.744, MAE = 0.429
- Test Set: RMSE = 0.558, R-Squared = 0.952, MAE = 0.429

The Training Set (80%) - The training metrics suggest that the model was able to learn reasonably well from the training data, being able to capture 74.4% of the variance in the target variable (Housing Construction PPI). The low RMSE and MAE on the training set indicate that the model was able to fit the training data closely without high error.

The Testing Set (20%) - The test metrics are slightly better than the training set, which is a good indicator. It means that the test set suggests that the model generalised well to new unseen data, capturing about 95.2% of the variance in the test set.

3.6 Actual vs Hybrid Model Prediction

The predictions and actual values for the test set (data the model has not seen before). The dashed black line represents the actual values, and the blue lines represent the predicted values. This plot tests the model's ability to make predictions on data it has never seen before. It's a test to see if the model has learned patterns rather than memorising past data it has seen. Most of the time, the blue line can be seen following closely with the blue line, especially during periods of spikes. This is a positive indication because it indicates that the model is adapting to gradual change. There is a spike around the period of 2021 and 2022, and there is a bit of separation between the actual line and the predicted line. The exact value (black line) spiked higher than the model predicted and then fell sharply. This is an indication that the model recognised there would be a rise and fall; however, it underestimated the magnitude of the increase and decrease. This discrepancy suggests that while the model is performing well overall, it finds itself struggling with accurately predicting sudden spikes or drops in housing prices.

3.7 Spike Indicator

As a further indicator of change is the implementation of a spike indicator in Housing PPI. There are limitations to this threshold. The equation may be over-sensitive. A 3% change in one quarter may not be considered significant in a highly volatile housing construction market. In some cases, the model may flag too many spikes, leading it to react to noise rather than meaningful change.

3.8 Exponential Smoothing

Subsequently, an 80:20 split was employed, a standard technique in time series analysis. The dataset was partitioned into two subsets: an 80% training set used for model fitting, and a 20% testing set utilised for model performance evaluation. This methodology enhances the subsequent model's ability to discern intricate data patterns and generalise effectively, mitigating the risks of overfitting and underfitting. Exponential smoothing is one of the most effective models in forecasting time series data. Essentially, it is a weighted average of historical data, with weights that diminish exponentially as observations become more distant in time. In essence, it assigns greater emphasis to recent observations, thereby assigning higher weights to them. This approach is well-suited to developing robust forecasting models across a broad spectrum of time series, particularly in the construction industry. Using the ets() function in R, the machine learning algorithm autonomously identified underlying patterns in the data and implemented the optimal exponential smoothing model with the most optimally tuned parameters.

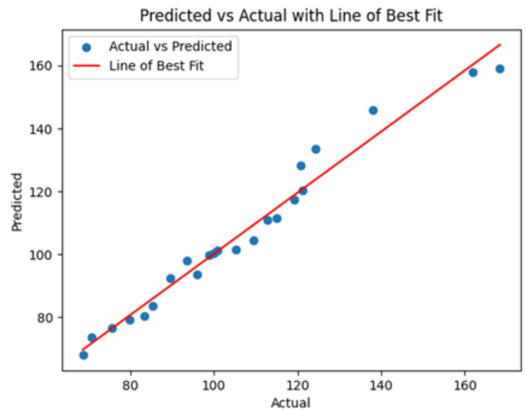

Figure 2. Regression Analysis: Predicted versus Actual with Line of Best Fit

Visualisation: Actual vs. Hybrid Model Predictions

Figure 2 shows the low error nature of the actual versus predicted data. From this starting point, greater creditability and confidence in the early development modal g

In Figure 3, the black dashed line in the plot represents the actual values, and the red solid line represents the model predictions over time. . The visualisation compares the true values (values that occurred and are recorded) with values that the hybrid model predicted by using the training set. As the predicted (red line) closely follows the actual values (dashed line), it indicates that the model is making reliable predictions. Since the lines overlap or are following closely, this suggests that the model is fitting the training data well.

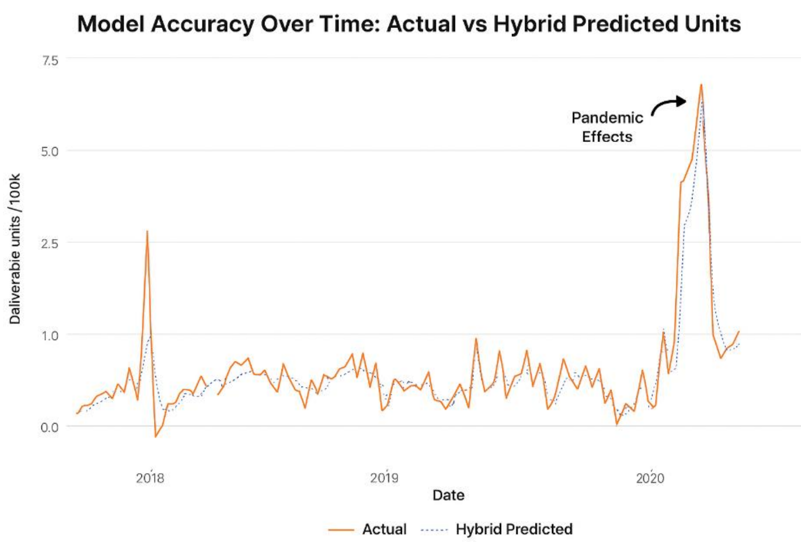

Figure 3. Model Accuracy Over Time: Actual versus Predicted.

Test Set: Actual vs Hybrid Model Prediction

The estimated and the actual values for the test set (data the model has not seen before). See Figure3. The dashed line represents the exact values, and the red lines represent the predicted values. This plot tests the model's ability to predict data it has never seen before. It's a test to see whether the model has learned patterns rather than memorising past data. Most of the time, the red line closely follows the dashed line, especially during periods of spikes. This is a positive sign that the model is adapting to gradual change. There is a spike around 2021 and 2022; there is a bit of separation between the actual and predicted lines. The exact value (dashed line) spiked above the model's prediction and then fell sharply. This indicates that the model recognised there would be a rise and fall; however, it underestimated the magnitudes of the increase and decrease. This discrepancy suggests that while the model is overall performing well, it struggles to predict sudden spikes or drops in housing prices accurately. Therefore, we have created an equation sensitive to volatility: *Spike Indicator = std_dev + (mean_abs_change × 0.2)*

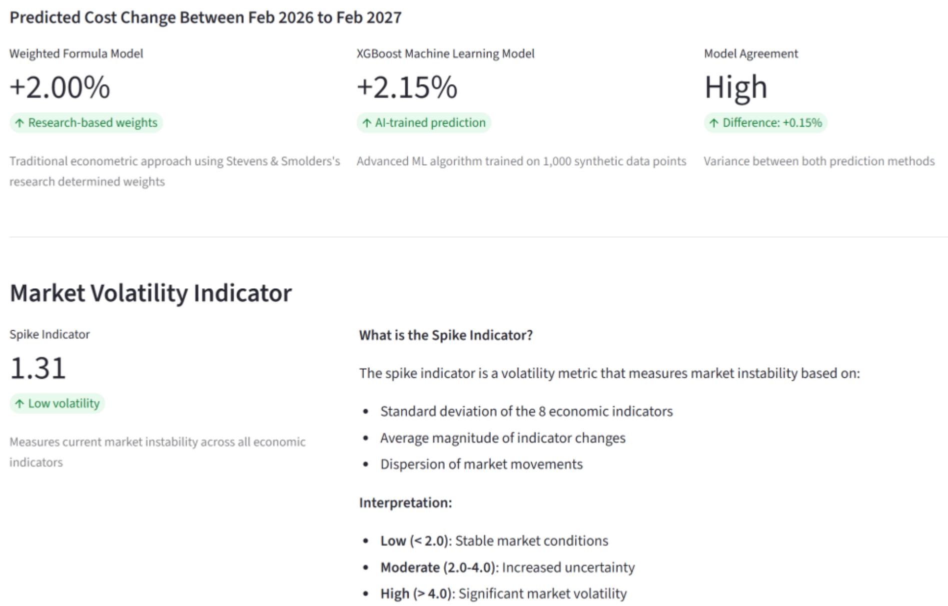

Figure 4. Australian Housing Construction Cost Predictions – Econometric and Machine Learning with Volatility Indicator

Setting in Artificial Intelligence

The researchers implemented the algorithm and its machine learning companion in an Agentic Artificial Intelligence setting. Replit was selected as the appropriate software-as-a-service provider. The equation was set, and database connections were established. See Figure 4.

5. Conclusions

A publicly available algorithm for forecasting housing construction costs has not been developed for Australia before. Property ownership is a preferred investment vehicle—it is part of the nation's culture (versus company ownership shares). Our team sought to identify the factors and dynamics of housing construction costs and place them into a predictive model.

Experts are determined by prediction accuracy—not by credentials, experience, or publishing records, but by the ability to project outcomes. Those who can consistently articulate future results provide a highly valuable service. In essence, reality meets theory. Courts, Academies, Industry and others should consider such a qualification requirement. Mastery of a topic's issues, facts, and processes provides present and future insights (Tetlock and Gardner 2025).

The hybrid ARIMAX-XGBoost model offers a complementary solution for forecasting Housing PPI. By combining traditional time-series modelling with modern machine learning, it captures both linear and nonlinear dependencies between target variables and underlying determinants more precisely. The model's robustness is enhanced by incorporating the Spike Indicator, making it more suitable for economic forecasting. Further refinements will be made to anomaly detection, improving confidence intervals, and enhancing the model's stability for recent periods that have been quite volatile.

Acknowledgements

Our thanks go to Thi Kim Truong and Thanh Thong Tran of Western Sydney University's School of Computer, Mathematics and Data Science. Also, David Froslie, Capstone Lead of North Dakota State University's Computer Science Department..

References

Abu Hammad, A., Ali, S. Sweis, G., and Sweis. R. (2010) Statistical analysis on the cost and duration of public building projects. Journal of Management in Engineering, 26(2):105–112, 2010. doi: 10.1061/(asce)0742-597x(2010)26:2(105).

Australian Bureau of Statistics (ABS) (2025) Lending Indicators. Accessed June 25, 2025: https://www.abs.gov.au/statistics/economy/finance/lending-indicators/mar-quarter-2025

Australian Securities and Investment Commission (ASIC) 2025) Bankruptcy Statistics. Accessed June 25, 2025: https://www.asic.gov.au/about-asic/corporate-publications/statistics/insolvency-statistics/

Courty, A. (2025). Property boom prompts surge in Australian millionaires to global high. August 4th 2025. Australian Broadcasting Corporation. https://www.abc.net.au/news/2025-08-04/property-boom-fuels-surge-in-aussie-millionaires-to-global-high/105589352

Elfahham. Y. Estimation and prediction of construction cost index using neural networks, time series, and regression. Alexandria Engineering Journal, 58(2): 499–506, 2019. doi: 10.1016/j.aej.2019.05.002.

Jain, G., & Mallick, B. (2017). A study of time series models ARIMA and ETS. Available at SSRN 2898968.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. Journal of econometrics, 54(1-3), 159-178.https://www.sciencedirect.com/science/article/pii/030440769290104Y.

Mehandzhiyski, V. (2023). What is an ARIMAX model? https://365datascience.com/tutorials/python-tutorials/arimax/, 365 DataScience, 5min read.

New South Wales Government (NSWGov) (2021) General Conditions of Contract Edition 2, Schedule 7 Costs Adjustment Formula.

Pennsylvania State University. (2018) Stat 462: Applied regression analysis. https://online.stat.psu.edu/stat462/node/180/, Eberly College of Science.

Ramraj, S., Uzir, N., Sunil, R., & Banerjee, S. (2016). Experimenting XGBoost algorithm for prediction and classification of different datasets. International Journal of Control Theory and Applications, 9(40), 651-662.

Tetlock, P. & Gardner, D. (2019). Superforecasting: The Art and Science of Prediction. Random House. London

WT Partnership (2024) Australian Construction Market Conditions Report. Accessed September 9, 2025. https://wtpartnership.com.au/wp-content/uploads/2025/05/WT-November-2024-Australian-Construction-Market-Conditions-Report.pdf

Xue, D. M., & Hua, Z. Q. (2016). ARIMA based time series forecasting model. Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering), 9(2), 93-98.