Price Trend Prediction and Discount Optimization for Video Games in Online Stores Using XGBoost and Time-Series Analysis: A Data Mining Approach for Metaverse-Driven Market Insights

Siti Sarah Maidin[1,*], Norzariyah Yahya[2]

[1]Faculty of Data Science and Information Technology (FDSIT), INTI International University, Nilai, Malaysia

[2]Kulliyyah of Information and Communication Technology (KICT), International Islamic University Malaysia (IIUM)

ABSTRACT

This research explores the application of data mining techniques, specifically XGBoost, to predict game pricing trends and optimize discount strategies within the digital gaming market. Game prices are influenced by various factors, including production costs, market demand, and promotional strategies. This study analyzes historical pricing data from multiple online stores to identify key pricing patterns and factors that influence price changes over time. The model developed in this study predicts game prices by incorporating features such as retail price, discount percentages, past price trends (lags), and other time-based features. The findings reveal that retail price and recent price trends (e.g., 7-day rolling averages) are the most influential features in predicting future prices. Additionally, discount strategies significantly impact game sales, with certain discount ranges showing higher effectiveness in driving consumer purchases. The model also demonstrates variability in prediction accuracy, particularly at higher price points, highlighting the challenges of capturing complex price fluctuations in a dynamic digital marketplace. The significance of this study extends to the Metaverse market, where pricing and the use of digital assets like non-fungible tokens (NFTs) play a critical role. The model's application could aid in optimizing pricing strategies within virtual economies, enhancing both the consumer experience and retailer profitability. Future work includes integrating additional features such as user reviews and exploring its application to Metaverse game platforms. The practical implications of this research are significant for online game retailers looking to leverage data-driven insights for more effective pricing and promotional strategies.

Keywords Price Prediction, Discount Optimization, Digital Gaming Market, Data Mining, Metaverse Market

Submitted: 15 February 2025
Accepted: 20 April 2025
Published: 20 November 2025

Corresponding author
Siti Sarah Maidin,
sitisarah.maidin@newinti.edu.my

Additional Information and Declarations can be found on page 329

DOI: 10.47738/ijrm.v2i4.41

INTRODUCTION

The rise of the metaverse and virtual economies is changing the fundamentals of digital game pricing by integrating innovative pricing models and discount strategies that vary significantly across online platforms. In the current digital

How to cite this article: S. S. Maidin and N. Yahya, "Price Trend Prediction and Discount Optimization for Video Games in Online Stores Using XGBoost and Time-Series Analysis: A Data Mining Approach for Metaverse-Driven Market Insights," *Int. J. Res. Metav.*, vol. 2, no. 4, pp. 333-353, 2025.

age, game developers and distributors are not only relying on traditional price-setting but also on dynamic mechanisms that account for consumer behavior in virtual environments [1], [2]. As digital economies evolve, virtual game environments are increasingly relying on concepts from the metaverse, where digital currencies and tokenized assets influence pricing strategies and consumer perception of value. This transformation has driven game companies to reexamine their pricing mechanisms, tailoring discount structures to enhance consumer engagement and maximize revenue.

Empirical studies have shown that value-informed pricing strategies play a crucial role in determining the price of virtual digital products, such as in-game accessories and downloadable content [3]. In highly interactive environments like Chinese MMORPGs, pricing is set based on consumer value perception and competitive dynamics, with game prices fluctuating according to intrinsic player demand and market competition. Concurrently, research comparing digital channels has demonstrated that discount effects are not uniform; for example, the response to price discounts varies across PC, app, and mobile website channels, with each channel exhibiting distinct consumer purchasing behaviors in reaction to discount offers [4]. This channel-specific fluctuation implies that game pricing strategies must be optimized for each platform to capture the digital economy's inherent heterogeneity.

Furthermore, large-scale experimental evidence underscores that discounting strategies, such as quantity discounts, can influence consumer behavior and overall revenue outcomes in digital gaming markets. A significant study highlighted a randomized pricing experiment involving over 14 million online game users, demonstrating that quantity discounts profoundly affected purchase volumes, thus reinforcing the importance of strategically calibrated discount policies in virtual economies [5]. These findings suggest that both game pricing fluctuations and the design of discount schemes are intricately tied to consumer behavior, competitive market forces, and the technological evolution underpinning the metaverse.

The objective of this research is to predict game price trends and optimize discounts using data mining techniques. By applying advanced algorithms like XGBoost and time-series analysis, the study aims to uncover patterns in pricing data and provide accurate forecasts for future prices. This predictive model will help in understanding how video game prices fluctuate over time, taking into account seasonal variations, store-specific strategies, and other influencing factors. The significance of this research lies in its ability to offer valuable insights into the pricing trends within the Metaverse market. As the virtual economy grows, understanding the dynamics of pricing across different platforms can provide both consumers and store owners with strategic advantages. Consumers can make more informed decisions on when to purchase games based on predicted price reductions, while store owners can optimize their discount strategies to maximize sales and maintain competitive pricing. This paper primarily focuses on price prediction and discount optimization for video games across major online stores. The scope includes analyzing historical price data from well-known platforms, identifying key

factors influencing price changes, and developing models to predict future trends. By examining the discounting strategies of various stores, this research aims to provide a comprehensive approach to optimizing pricing in the digital gaming industry, with an eye towards the growing Metaverse market.

Literature Review

Overview of Price Trend Analysis in E-Commerce

Recent studies on price trend analysis in e-commerce have highlighted the growing importance of applying advanced statistical and machine learning techniques to forecast the evolution of prices over time. These studies focus on capturing the volatility inherent in online pricing strategies, while also addressing the underlying macroeconomic and market-specific factors that drive price changes. Understanding these factors is crucial for businesses as they seek to optimize pricing decisions and respond effectively to market fluctuations.

One of the most prominent areas of research in this field is dynamic pricing. Dynamic pricing involves adjusting prices in real-time based on various factors such as demand fluctuations, inventory levels, and competitive pressures. A bibliometric analysis of dynamic pricing research shows a marked increase in academic interest since the early 2000s, with a particularly sharp rise in publications in 2021 [6]. This surge in research reflects the growing recognition of dynamic pricing as a key component of e-commerce strategies, enabling retailers to remain competitive by responding swiftly to changes in the marketplace. The bibliometric approach used in these studies offers a comprehensive mapping of the evolving research landscape and identifies emerging challenges in the dynamic pricing domain.

Alongside dynamic pricing studies, various forecasting models have been developed to predict how product prices evolve over time. One such model, proposed by [7], integrates Autoregressive Integrated Moving Average (ARIMA) with Google Trends data to predict future price trends on e-commerce platforms. This approach highlights the power of time-series forecasting techniques in predicting price changes, while also emphasizing the value of incorporating external digital signals—such as online search trends—into pricing models to enhance their accuracy. Similarly, [8] developed a forecasting method based on Gaussian processes, which incorporates factors like seller reputation and sales volume to predict price dispersion and future price movement in the context of Chinese cross-border e-commerce. These methodologies provide strong frameworks for understanding price trends and offer valuable insights into the future of pricing strategies.

In addition to market-specific dynamics, macroeconomic factors also play a significant role in shaping e-commerce pricing trends. For example, research by [9] on exchange rate pass-through effects in Brazilian e-commerce demonstrates how fluctuations in foreign exchange rates can alter the pricing structures of online retailers. These fluctuations influence the cost of imported goods, which, in turn, impacts the final price paid by consumers. This

macroeconomic perspective adds an extra layer of complexity to dynamic pricing models, as it underscores the importance of considering broader economic conditions when developing pricing strategies for e-commerce platforms.

Time-Series Forecasting

Time-series forecasting for price prediction has been widely explored across various industries, including digital advertising, financial markets, commodity pricing, and online auctions. Early studies, such as [10], demonstrate the value of time-series analysis in digital signage advertising by modeling environmental factors and audience attention changes, which influence pricing decisions. Similarly, in financial time-series forecasting, [11] address the challenges posed by low signal-to-noise ratios and the dynamic inter-asset relationships, emphasizing that a successful forecasting framework requires understanding the properties of time-series data such as linearity, stationarity, and volatility. These studies highlight the critical role of time-series analysis in capturing the underlying price dynamics across various sectors.

Classical time-series models, such as ARIMA and exponential smoothing, have long been used to forecast prices in a wide range of markets. For example, [12] successfully applied ARIMA models for forecasting day-ahead electricity market prices, while [13] utilized similar methods for short-term agricultural price indices forecasting. Additionally, [14] demonstrated that Holt's double exponential smoothing method could accurately predict gold bullion prices, illustrating the effectiveness of smoothing techniques in capturing short-term trends. These traditional statistical models, which also include semiparametric regression analysis for dynamic auction price predictions [15], prove to be versatile tools in a variety of forecasting scenarios, emphasizing their continued relevance in modern forecasting tasks.

Advancements in machine learning and deep learning have significantly enhanced time-series forecasting for price prediction. Research [16], [17] employed Long Short-Term Memory (LSTM) networks to capture nonlinear temporal dependencies in stock market data, achieving higher accuracy than traditional models. Similarly, [18] developed a heterogeneous Gated Recurrent Unit (GRU) neural network with an attention mechanism to predict fluctuations in livestock product prices, demonstrating the advantages of deep learning in handling complex, multi-scale price movements. Additionally, [19] highlighted the effectiveness of recurrent neural networks in predicting stock trends, while [20], [21] incorporated convolution-based filtering techniques to isolate latent components in crude oil price series. These innovations reflect the growing potential of deep learning models in price prediction tasks that require handling complex patterns and long-term dependencies.

Beyond these conventional and advanced methods, fuzzy logic has also emerged as a powerful tool for modeling price uncertainty. Research [22] compared fuzzy time-series models to traditional forecasting techniques for composite stock price indices, demonstrating that linguistic-based forecasting provides added flexibility in scenarios where data patterns are ambiguous and

variability is high. This approach, alongside hybrid forecasting techniques, emphasizes the importance of flexibility in modeling uncertain price movements. Studies incorporating exponential smoothing and hybrid methods reveal that using an ensemble of time-series forecasting techniques can be particularly effective in addressing the diverse characteristics of pricing data across different markets.

XGBoost for Price Prediction

XGBoost, an ensemble learning algorithm that leverages gradient boosting methods over decision trees, has gained widespread attention in price prediction due to its exceptional accuracy and computational efficiency in forecasting numerical values. By systematically reducing bias and variance in predictions, XGBoost outperforms many traditional machine learning algorithms, making it an ideal choice for various pricing applications. This robustness in prediction accuracy is particularly useful in industries where timely and precise pricing forecasts are crucial for decision-making [23], [24].

Beyond financial markets, XGBoost has proven its effectiveness in other domains, such as environmental economics and digital asset markets. Research [25] applied an extreme gradient boosting model optimized through the whale optimization algorithm to forecast carbon prices, achieving superior results compared to several benchmark models. Additionally, [2] introduced a hybrid model that first processed carbon price signals before inputting them into an XGBoost framework, leading to a notable reduction in prediction errors. These studies demonstrate that XGBoost can effectively handle complex, high-dimensional datasets and is adaptable across a wide range of pricing scenarios.

Metaverse and Digital Economy Impact

The rise of the metaverse and virtual economies has fundamentally reshaped the landscape of digital product pricing. This transformation introduces new digital assets, revised business models, and innovative pricing mechanisms that integrate traditional economic principles with digital innovation. Virtual economies increasingly rely on digitized currencies, tokenized assets, and dynamic market interactions to determine product value. The integration of digital legal currencies and non-fungible tokens (NFTs) has catalyzed a shift from conventional pricing methodologies towards value-based and opportunity-driven pricing strategies. In this environment, prices are no longer solely determined by production costs or consumer demand but are also influenced by factors like digital scarcity, network effects, and cross-platform interoperability [26], [27].

Metaverse platforms are central to these pricing transformations by creating interoperable ecosystems that support both user-generated content and platform-mediated pricing mechanisms. Research [28] conceptualize metaverse platforms as meta-ecosystems where real-time rendered 3D virtual worlds and digital environments converge, fostering increased consumer engagement and co-creation of goods. This platform-centric view underscores that pricing strategies in digital markets are deeply intertwined with the

metaverse's underlying architecture. In these environments, platform owners and orchestrators have significant influence over market dynamics, shaping how products are priced and exchanged. Moreover, the concentration of power in the hands of major technology companies active in extended reality (XR) raises concerns about pricing transparency and the potential for monopolistic behavior, which could affect competitive pricing strategies.

The pricing of digital products in the metaverse is also influenced by consumer perceptions, digital scarcity, and the provenance of assets. Studies of virtual marketplaces have shown that these factors significantly contribute to how value is ascribed to digital goods. For instance, the pricing of digital game accessories and virtual merchandise is increasingly driven by consumer engagement metrics and perceived exclusivity. This trend is further amplified by the decentralized and borderless nature of metaverse platforms, where digital scarcity and exclusivity are major determinants of value. The emergence of tokenized assets has led to the redesign of traditional discounting and pricing models, where algorithms account for network effects and real-time market signals to optimize pricing strategies in dynamic virtual economies [26]. These developments highlight the need for adaptive pricing strategies capable of responding to rapidly changing market conditions within the metaverse.

Furthermore, the advent of these digital pricing mechanisms introduces a shift in how businesses approach pricing strategy. Unlike traditional e-commerce, where pricing is influenced primarily by cost-plus models or competitive benchmarking, the metaverse demands more dynamic and complex pricing approaches. Platforms must take into account not only the digital scarcity and perceived value of assets but also the governance structures that influence the ownership and distribution of digital products. These new pricing models also emphasize consumer co-creation, where the involvement of users in shaping the product or asset can influence its market value. This represents a significant departure from conventional pricing models and signals the need for ongoing innovation in digital pricing strategies.

Method

The workflow of the proposed Time-Series XGBoost Regression (TS-XGBR) model is illustrated in figure 1. The process begins with data preprocessing, where the dataset containing time, title, storeID, price, retailprice, and savings is loaded, converted to datetime format, and sorted chronologically to preserve temporal order. Missing values are then handled through forward filling or row removal when critical data are absent. Once the dataset is cleaned, feature engineering is performed to extract lag features, autocorrelation coefficients, and the Price Momentum Ratio (PMR), along with temporal attributes such as day of the week, month, and year. These features are compiled into a feature matrix $X$and a target vector y, which are used to train an XGBoost regressor under a time-aware cross-validation scheme. The model is optimized and evaluated iteratively, where performance assessment determines whether retraining is necessary. If the results meet the desired criteria, the final trained model and feature importance metrics are saved for interpretation and

deployment.

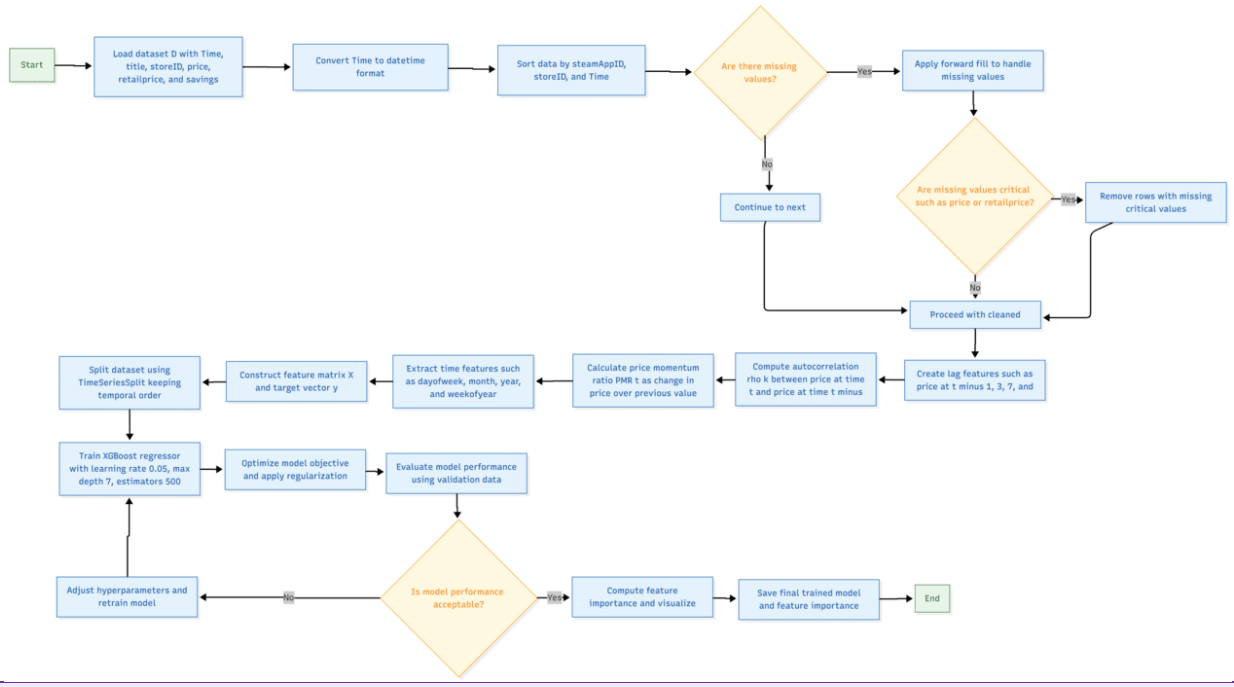

Figure 1 Proposed Time-Series XGBoost Regression (TS-XGBR)

Data Preprocessing

The process of predicting video game price trends across multiple online stores begins with a detailed data preprocessing stage, designed to ensure data integrity and temporal coherence. The dataset consists of essential columns such as *Time*, *title*, *storeID*, *price*, *retailprice*, and *savings*. To facilitate temporal analysis, the *Time* column is converted into a datetime object, enabling chronological operations such as lag creation and time-based resampling. The dataset is then sorted according to *steamAppID*, *storeID*, and *Time* to maintain sequential order, which is vital for modeling time-dependent variables. Missing data are handled using forward filling, a method that propagates the most recent known value to subsequent missing entries, thereby preserving price continuity within each game-store combination. If essential information, particularly in *price* or *retailprice*, remains unavailable, those rows are removed to avoid inconsistencies during model training. This process ensures that the temporal relationships between price observations are not disrupted, forming a reliable foundation for subsequent analysis.

Feature Engineering

Once the data has been cleaned and structured, feature engineering is performed to enhance the model's ability to capture temporal patterns and price dynamics. Lag features are created to provide the model with historical price information, enabling it to recognize both short-term fluctuations and longer-term cycles in the data. These lag features are derived at various intervals, such as 1, 3, 7, and 14 days, to represent different temporal horizons.

To further quantify the relationship between current and past prices, an

autocorrelation coefficient is introduced. This metric measures the degree of similarity between a time series and a lagged version of itself and is expressed as

$$\rho_k = \frac{\sum_{t=k+1}^{T}(p_t - \bar{p})(p_{t-k} - \bar{p})}{\sum_{t=1}^{T}(p_t - \bar{p})^2} \quad (1)$$

$\rho_k$indicates the strength of correlation between the price at time $t$and its value $k$periods earlier. A high $\rho_k$suggests strong temporal persistence, implying that past prices significantly influence future values. Additionally, a distinctive metric known as the Price Momentum Ratio (PMR) is introduced to measure short-term acceleration or deceleration in pricing movements. It is defined as

$$PMR_t = \frac{p_t - p_{t-k}}{p_{t-k}} \quad (2)$$

PMR_trepresents the relative rate of change in price compared to a previous point $k$steps back. This ratio allows the model to capture dynamic price behavior such as discounts, flash sales, or gradual price increases.

Temporal features such as dayofweek, month, year, and weekofyear are extracted to capture seasonal patterns and cyclical effects in price trends, which are common in digital game markets influenced by events like sales periods or holiday seasons.

Model Training

The cleaned and feature-enhanced dataset is then used to train an XGBoost regression model, selected for its efficiency and ability to model nonlinear interactions among temporal features. The algorithm constructs an ensemble of decision trees through gradient boosting, minimizing a regularized objective function that balances accuracy and model complexity. Hyperparameters such as learning rate (0.05), maximum depth (7), and the number of estimators (500) are optimized to achieve stable and accurate results.

To evaluate the model reliably, a TimeSeriesSplit cross-validation strategy is employed, ensuring that the temporal order of the data is preserved. Each validation fold uses earlier data for training and later data for testing, thereby simulating realistic forecasting scenarios and avoiding data leakage.

Model Evaluation

After training, the model's predictive performance is assessed and interpreted through a combination of quantitative metrics and qualitative visualization. Instead of relying solely on general metrics, deeper analysis is conducted using XGBoost's internal feature importance scores, which indicate how much each feature contributes to reducing prediction error. These insights help identify the most influential predictors in price estimation, typically including *retailprice*, *PMR*, and the autocorrelation terms.

Visual assessment complements the numerical evaluation through scatter

plots of predicted versus actual prices, residual plots showing error distributions, and time-series plots comparing observed and predicted price trends. These analyses provide a clear understanding of model accuracy and stability over time, while also revealing potential temporal patterns or deviations that may require model refinement.

Algorithm 1 **Time-Series XGBoost Regression (TS-XGBR)**

Let the dataset be denoted as

$$D = \{(t_i, \text{title}_i, \text{storeID}_i, p_i, r_i, s_i) \mid i = 1,2,\dots,N\}$$

where $t_i$is the timestamp, $p_i$is the price, $r_i$is the retail price, and $s_i$is the savings value.

Step 1: Data Preprocessing

1. Convert $t_i \rightarrow \text{datetime}(t_i)$for all i.
2. Sort $D$by (steamAppID,storeID, t_i).
3. Handle missing values using forward filling:

$$p_i = \begin{cases} p_i, & \text{if } p_i \neq \emptyset \\ p_{i-1}, & \text{if } p_i = \emptyset \end{cases}$$

Remove all tuples where $p_i = \emptyset$or $r_i = \emptyset$.

Step 2: Feature Engineering

For each unique pair $(g, s)\in$ (game, store), and for each time index t:

- **Lag Features:**

$$L_k(p_t) = p_{t-k}, k \in \{1,3,7,14\}$$

- **Autocorrelation Coefficient:**

$$\rho_k = \frac{\sum_{t=k+1}^{T}(p_t - \bar{p})(p_{t-k} - \bar{p})}{\sum_{t=1}^{T}(p_t - \bar{p})^2}$$

where $\bar{p} = \frac{1}{T}\sum_{t=1}^{T} p_t$.

- **Price Momentum Ratio (PMR):**

$$PMR_t = \frac{p_t - p_{t-k}}{p_{t-k}}, k = 1,3,7$$

Temporal Features Extraction:

$$\text{TimeFeatures}(t) = \{\text{dayofweek}(t), \text{month}(t), \text{year}(t), \text{weekofyear}(t)\}$$

Construct the feature matrix:

$$X = [L_k(p_t), \rho_k, PMR_t, \text{TimeFeatures}(t), r_t, s_t]$$

and the target vector:

$$y = [p_t]$$

Step 3: Model Training

Split dataset Dinto temporally ordered folds:

$$\{(X_{train}^{(i)}, y_{train}^{(i)}), (X_{test}^{(i)}, y_{test}^{(i)})\}, i = 1,2,\dots,n_{splits}$$

using TimeSeriesSplit such that:

$$\max(t_{train}^{(i)}) < \min(t_{test}^{(i)})$$

Train the **XGBoost regression model** $f_\theta$to minimize the regularized objective:

$$\min_\theta \left[\sum_{i=1}^{N} l(y_i, f_\theta(X_i)) + \sum_k \Omega(f_k)\right]$$

where

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2$$

and $l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$.

Step 4: Model Evaluation

After training, compute feature importance I_j for each feature $x_j \in X$:

$$I_j = \frac{1}{K}\sum_{k=1}^{K} \Delta L_{k,j}$$

where $\Delta L_{k,j}$ represents the average loss reduction contributed by feature j in tree k.

Evaluate prediction quality by comparing:

$$\hat{y}_t = f_\theta(X_t)$$

with the observed price y_t, and visualize:

$$\{(t, y_t), (t, \hat{y}_t), (t, y_t - \hat{y}_t)\}$$

as time series, scatter plots, and residual plots.

Step 5: Model Deployment

Store the final optimized model:

$$M^* = f_{\theta^*}(X)$$

where θ^* is the parameter set that minimizes validation loss.
Export both the model and feature importance scores for interpretability and future inference.

Result and Discussion

Descriptive Statistics

The dataset used in this analysis comprises 73,000 observations, each corresponding to a price record for a video game across five different online stores over a two-year period. The dataset contains essential columns such as Time, steamAppID, storeID, price, retailprice, and savings, which track the price fluctuations and discounts applied to the games over time. The descriptive statistics of the dataset provide valuable insights into the pricing patterns and distribution of values. On average, the price of a video game in this dataset was $39.52, with retail prices ranging from $20.19 to $69.54. The savings, which represent the discount applied to the retail price, had a mean value of 6.46%, with the maximum savings reaching up to 75.59%. Notably, the majority of the games had no discount applied, as 75% of the records showed savings of 0%, indicating that most games were sold at their full retail price. This distribution suggests that the dataset captures a wide variety of pricing behaviors, from full-price sales to heavily discounted games.

Figure 2 displays the distribution of game prices in the dataset. The histogram represents the frequency of game prices, with a smooth curve (KDE) overlaid to show the price distribution more clearly. The prices are concentrated around certain values, with noticeable peaks at various price ranges, such as around $20, $40, and $60. This suggests that these are common price points for the games in the dataset.

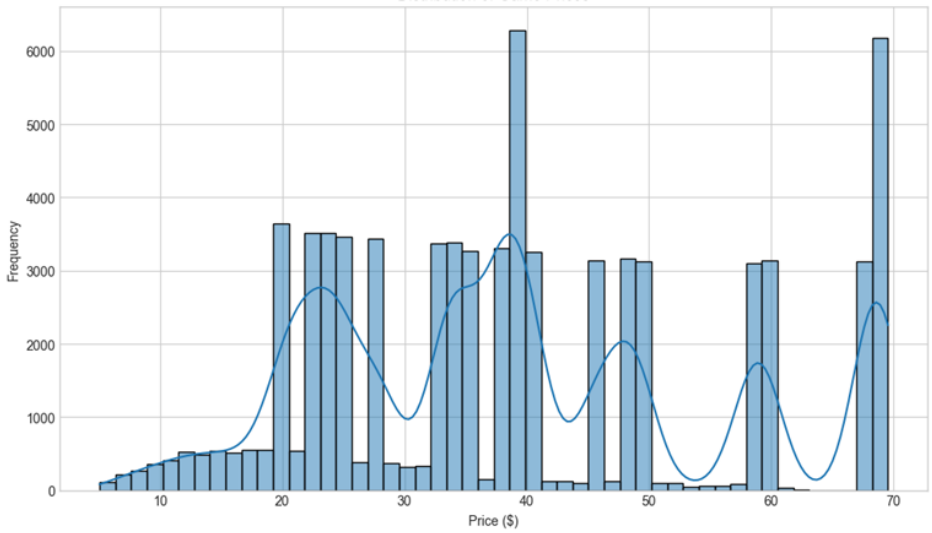

Figure 2 Distribution of Game Prices

The KDE curve indicates that the distribution of game prices follows a multimodal pattern, with several price ranges exhibiting higher frequencies, which could correspond to specific pricing strategies or market trends in the digital gaming industry. The spread of prices is quite broad, extending from around $5 to just under $70, reflecting a wide variety of games and their corresponding pricing structures. Figure 3 shows the distribution of retail prices for the games. Similar to figure 2, the histogram illustrates the frequency of retail prices across different price intervals, and the smooth curve (KDE) helps to visualize the general price trend. The distribution of retail prices has clear peaks at certain values, with the most prominent being around $30 and $50, indicating that these are the most common retail prices for the games in the dataset.

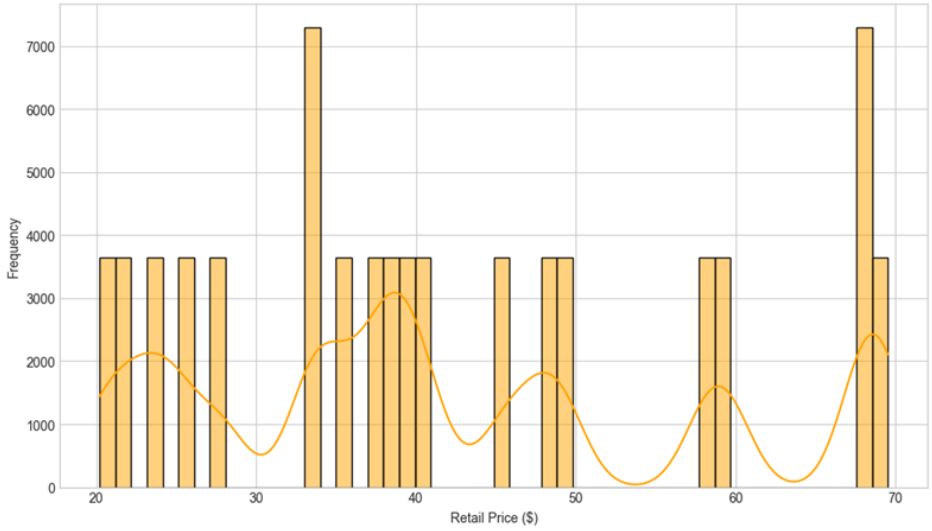

Figure 3 Distribution of Retail Prices

The KDE curve highlights the bimodal nature of retail prices, where there are two primary clusters of prices. This suggests that many games are priced either

in the lower range around $20–$30 or the mid-range around $50–$60. This pattern may reflect the pricing strategy where the majority of games are priced within these two bands, while fewer games are priced at the extremes of the spectrum. Figure 4 illustrates the distribution of discounts (or savings) when a game is on sale, showing the percentage of discount applied to the retail price. The histogram, along with the overlaid smooth curve (KDE), highlights several important trends in the data. The most striking feature of the graph is the significant concentration of values at 0% savings. This is expected because many of the games are likely sold at full price, reflecting that 0% savings accounts for the largest portion of the data.

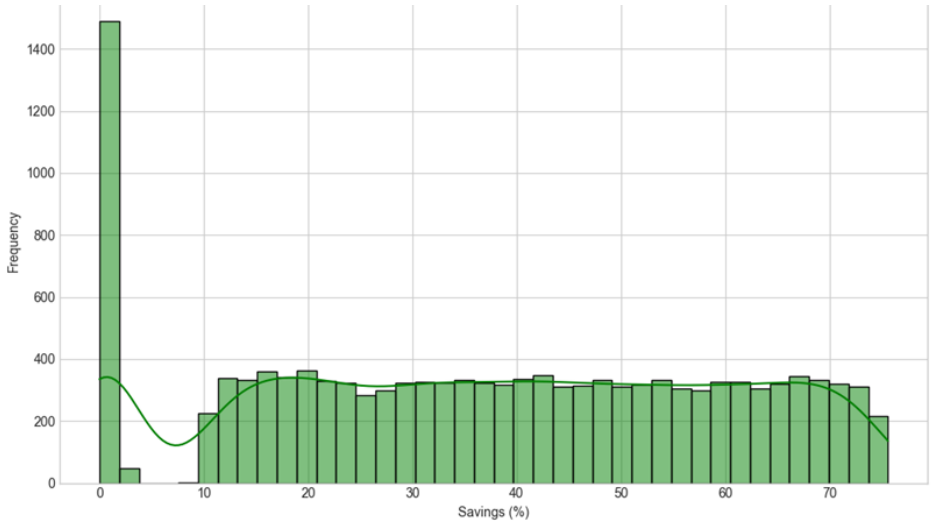

Figure 4 Distribution of Discounts

The bar at 0% savings is extremely tall, indicating that a substantial number of games in the dataset were not discounted at all. For the other discount percentages, the graph shows a much more even distribution, with several discount ranges between 10% to 75%. While these discounts occur less frequently than the 0% discount, the number of games with discounts steadily increases as the savings percentage rises, especially between 10% and 30%. After this point, the frequency of discounts starts to taper off, with fewer games offering higher discounts. The KDE curve illustrates this trend, showing a gentle peak at around 15% to 20% savings, before the distribution becomes flatter as discounts increase. The presence of a long tail in the distribution suggests that some games receive significant discounts, with a few instances of discounts reaching as high as 75%. However, these high discounts are relatively rare, as indicated by the small number of observations towards the right side of the graph. This suggests that while heavy discounts are offered, they are not common and might be reserved for specific sales events or promotional periods. The overall pattern of this distribution indicates a pricing strategy where most games are sold at or near retail price, with moderate discounts applied to a smaller subset of games.

Figure 5 illustrates the average game price over time on a weekly basis. The plot shows fluctuations in the average price of games across the two-year

period. The price trend is characterized by noticeable peaks and valleys, with prices occasionally spiking above $39.8 and then dropping to around $39.2. These fluctuations may suggest periodic pricing adjustments, possibly in response to factors like sales events, promotions, or market conditions.

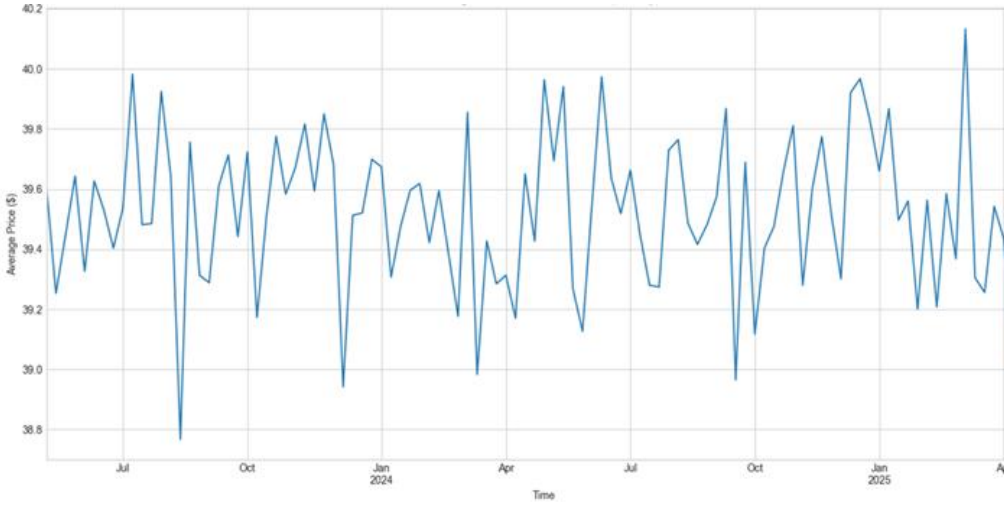

Figure 5 Average Game Price Over Time (Weekly)

The presence of sharp variations indicates that the pricing strategy of the games might change frequently, but there is a general stability around a mean price close to $39.5. The weekly time interval provides insights into short-term trends, capturing the impact of weekly fluctuations in pricing, which could be influenced by factors such as store-specific pricing, seasonal changes, or consumer demand. Figure 6 provides a comparison of the average game price over time for five different stores.

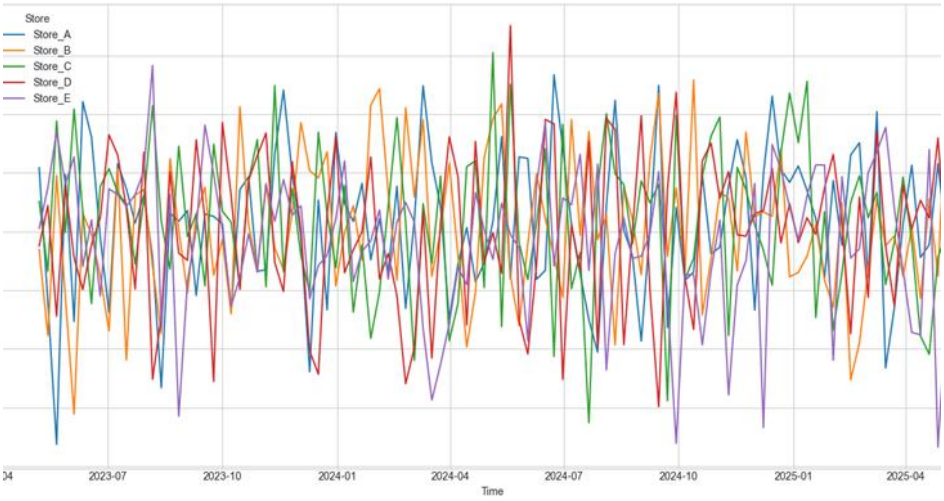

Figure 6 Average Game Price Over Time (Per Store)

The data is visualized with separate lines for each store, with each line showing weekly average price trends over the same two-year period. The graph reveals that each store exhibits distinct price trends, with some stores showing higher average prices than others, particularly Store E, which consistently has prices

above $40. Meanwhile, Store A has the most volatile price fluctuations, experiencing more dramatic ups and downs compared to the other stores. The different stores display variations in their pricing strategies, which could be attributed to factors such as pricing models, promotions, and inventory management. The lines for Store B and Store C generally stay closer to the average price range, while Store D shows a more consistent pricing pattern. This comparative analysis highlights the competitive pricing dynamics among different e-commerce platforms and can offer insights into how price differentiation is managed across multiple retailers.

Data Preprocessing and Feature Engineering

The data preprocessing phase begins with ensuring that the dataset is properly structured for time-series analysis. The Time column, which represents the date of each price observation, is converted to a datetime object, which is essential for handling temporal data correctly. The dataset is then sorted by steamAppID, storeID, and Time to ensure the chronological order is maintained, as time-series models require data to be processed in a temporal sequence. Sorting by steamAppID and storeID ensures that prices and other features are handled correctly within each game/store combination, as lag features will be calculated based on the previous time periods for the same game and store. During this preprocessing stage, missing data is checked, and although no missing values were found in the dataset, a robust strategy is in place for handling missing values in real-world datasets. Missing values in critical columns such as price or retailprice would be dropped, while for less critical data, methods like forward filling could be applied to maintain the continuity of time-series data.

Once the data is cleaned and sorted, feature engineering is carried out to enhance the model's predictive capabilities. The creation of lag features allows the model to take into account the past price and savings data for each game and store, which is crucial for predicting future prices based on historical trends. Several lag periods were chosen, including 1, 3, 7, and 14 days, to capture both short-term and longer-term dependencies in the pricing data. In addition to lag features, time-based features were extracted from the Time column to capture seasonal and cyclical patterns. These features included the day of the week, month, year, day of the year, and week of the year. These time-based features are important because pricing behaviors in digital markets often follow seasonal trends (e.g., price increases during holidays or special events). Furthermore, rolling window features were calculated, such as the 7-day rolling mean and rolling standard deviation of prices, to capture short-term price fluctuations and volatility. These rolling statistics provide the model with insights into recent price trends and price variability, which are important for making accurate short-term predictions.

Model Training and Evaluation

For model training, XGBoost was selected due to its high efficiency and strong performance in regression tasks. The model was trained using a TimeSeriesSplit cross-validation strategy, which is critical for time-series

forecasting tasks. This method ensures that the temporal order of the data is respected by splitting the dataset into training and testing subsets, where each test set represents a future time period. The model is trained on the training set for each fold and evaluated on the corresponding test set. This approach simulates a real-world scenario where the model must predict future prices based on historical data, making it ideal for time-series forecasting tasks.

The XGBoost model was configured with a set of parameters that are optimized for regression tasks. These parameters include a learning rate of 0.05 to control the contribution of each new tree to the final prediction, a maximum tree depth of 7 to prevent overfitting, and 500 estimators (or boosting rounds) to ensure sufficient learning capacity. The objective function was set to reg:squarederror for regression, and RMSE was chosen as the evaluation metric to guide the training process. The performance of the model was evaluated using multiple regression metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). The evaluation results for each fold of the cross-validation were as follows: in the first fold, the model achieved a MAE of 3.93, RMSE of 5.56, and R^2 of 0.60, indicating reasonable performance. The second fold showed a slight decline in performance, with a MAE of 7.90, RMSE of 10.42, and R^2 of 0.44. However, in the third fold, the model's performance improved, achieving a MAE of 4.52, RMSE of 7.24, and R^2 of 0.66. The fourth and fifth folds demonstrated the best performance, with MAE values of 3.70 and 4.22, RMSE values of 6.44 and 7.69, and R^2 values of 0.86 in both cases. The average performance across all folds was a MAE of 4.85, RMSE of 7.47, and R^2 of 0.68, indicating that the model was able to explain 68% of the variance in the price data with relatively low error metrics.

Feature Importance Analysis

After training the model, an important aspect of understanding its decision-making process is analyzing feature importance. This analysis helps to identify which variables contribute the most to the model’s predictions. In the case of this model, the retail price emerged as the most important feature, contributing 53.47% to the prediction of game prices. This result aligns with expectations, as the retail price is likely the most significant factor in determining the final sale price. The rolling mean of prices over 7 days, labeled as price_roll_mean_7, was the second most important feature, with a contribution of 26.47%. This suggests that short-term trends in price data play a substantial role in the model’s predictions. Other lag features, such as price_lag_14, price_lag_7, and price_lag_3, also contributed to the model, with price_lag_14 being the most influential of the lag features at 7.99%. The time-based features, such as dayofweek, month, and year, had lower importance scores, but still played a role in capturing seasonal and cyclical pricing patterns. The savings features, particularly lagged savings values, also contributed to the predictions, though their influence was more modest compared to price-related features.

Figure 7 presented is a feature importance plot generated from the XGBoost model, displaying the relative importance of each feature used in the model for predicting game prices. The bars in the graph represent the relative importance

of the features, with the longer bars indicating greater importance. The features are ranked based on their contribution to the model's predictions. From the graph, it's clear that retailprice is the most influential feature, with the longest bar, accounting for the largest proportion of the model's predictive power. This aligns with expectations, as the retail price is likely the most significant determinant of the final sale price.

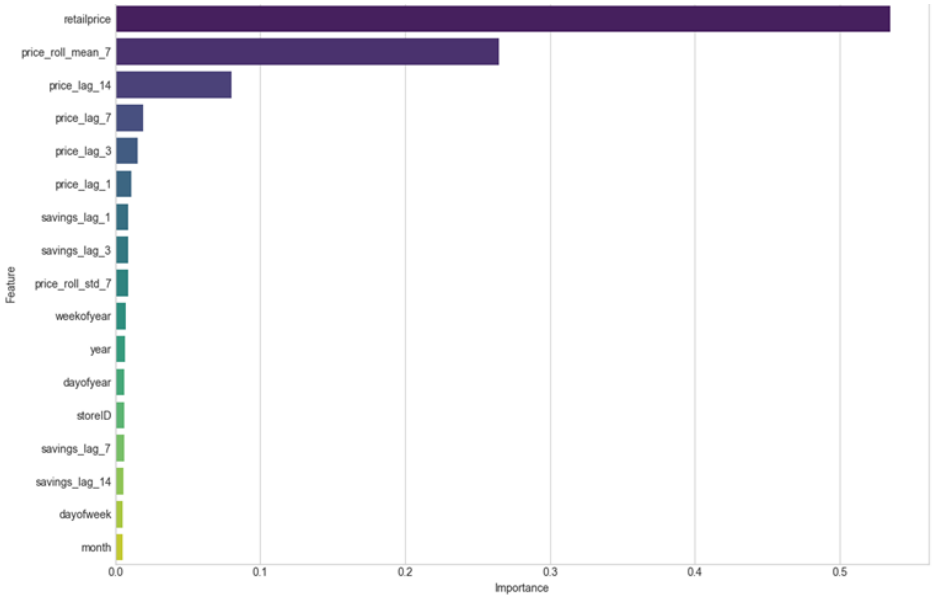

Figure 7 Feature Importance Bar

The rolling mean of price over 7 days (price_roll_mean_7) comes in second, reflecting the importance of short-term price trends in predicting future prices. The lag features also play a significant role, with price_lag_14 (the price from 14 days ago) being the third most important feature, followed by price_lag_7, price_lag_3, and price_lag_1, which capture increasingly shorter time spans of historical data. These lag features allow the model to incorporate the temporal dependencies of game prices. Other important features include savings-related lag features, such as savings_lag_1, savings_lag_3, and savings_lag_7, which reflect the impact of previous discounts on current price trends. The rolling standard deviation of price over 7 days (price_roll_std_7) also contributes significantly to the model, as it helps capture price volatility and fluctuations. Time-related features like weekofyear, year, and dayofyear also show meaningful contributions, helping the model account for seasonal trends and other cyclical effects in pricing. However, features like storeID, savings_lag_14, and month have a relatively lower importance, indicating that while they contribute to the predictions, their impact is not as strong as the price and savings-related features.

Final Evaluation and Visualizations

The final evaluation was based on the results from the last fold of the cross-validation process. A detailed comparison of the predicted prices and actual prices was conducted using various visualizations. Figure 8 compares the actual prices and predicted prices of games over time, specifically for the last

fold of the test set in the cross-validation process. The blue line represents the actual prices of the games, while the orange line represents the predicted prices generated by the model. From the graph, we can observe that the model's predictions generally track the actual prices quite well, especially in the mid-range of the price spectrum.

Figure 8 Comparison of Actual vs Predicted Prices

There are fluctuations where the predicted prices (orange) deviate from the actual prices (blue), but the overall trend is similar. These deviations are most noticeable during periods of high volatility, where both the actual and predicted prices exhibit sharp peaks and valleys. This suggests that while the model captures the overall pricing trend, it may struggle with accurately predicting extreme fluctuations or sudden spikes in prices. The shaded regions on the graph emphasize the areas of deviation between the actual and predicted prices. These areas indicate where the model's predictions are not as accurate, particularly during the periods of sharp price changes. However, the model remains relatively effective in predicting the general price trajectory, especially during more stable periods. The results show that the model is capable of providing a reasonable estimate of game prices, though there is room for improvement, particularly in handling price volatility or sudden market shifts.

The results indicate that the XGBoost model performed well in predicting video game prices, with the best performance observed in the later folds of the cross-validation. The model's ability to capture price trends, along with its reliance on the most relevant features, suggests that it can be effectively used for predicting prices in dynamic markets. The feature importance analysis provided valuable insights into the factors driving price predictions, particularly the significant role of retail prices, lag features, and short-term price trends. Overall, the model demonstrated strong predictive power and could be useful for pricing strategies in online retail, particularly in dynamic and evolving environments such as the metaverse.

Conclusion

The analysis of game prices reveals that they fluctuate according to specific trends, with periodic price adjustments reflecting market conditions, consumer

behavior, and sales events. The exploration of discount strategies demonstrates that discounts are a key factor in driving sales, with certain discount percentages being more effective in increasing consumer interest. The price prediction model, developed using data mining techniques, showed a reasonable ability to forecast game prices, though there were challenges in predicting prices at the extreme ends of the spectrum. The results highlight that game pricing is influenced by both external market factors and internal strategies, such as discounts and seasonal promotions. The ability to predict game pricing accurately is of significant value for the Metaverse market, where virtual goods and services, including games and in-game assets, are increasingly traded. Price prediction models can help anticipate price fluctuations in this emerging market, allowing game developers, retailers, and platform owners to optimize their pricing strategies. Furthermore, such models can also help forecast when sales or discounts should be applied to maximize revenue or drive user engagement. As virtual economies in the Metaverse continue to grow, understanding price dynamics will be essential for stakeholders to stay competitive and profitable. This research contributes to the field of digital economics by showcasing how data mining techniques, specifically XGBoost, can be used to improve the understanding of pricing behaviors in virtual economies. By analyzing and predicting price trends in the digital game market, the study sheds light on the complexities of virtual product pricing, which differs from traditional physical product pricing due to factors like digital scarcity, tokenized assets, and real-time market interactions. This approach enhances our ability to navigate and optimize pricing models within the virtual economy, which is becoming an increasingly significant component of the global digital marketplace.

Further research could expand the scope of the current study by integrating additional features, such as user reviews, game ratings, or social media sentiment, which may influence game pricing and sales trends. Additionally, applying this model to specific Metaverse game platforms could offer more detailed insights into how virtual economies operate in these environments. The incorporation of these features would likely improve the model's accuracy, particularly in predicting the pricing of digital goods and services in a rapidly evolving market. The findings from this research offer practical insights for online game retailers and other stakeholders in the digital marketplace. By leveraging price prediction models, these retailers can optimize their pricing strategies to increase sales, attract customers, and enhance their overall business performance. Additionally, understanding the impact of discounting strategies on game prices can allow retailers to fine-tune their promotional activities and better meet consumer expectations. Ultimately, the application of these models can lead to a more efficient pricing system, benefiting both retailers and consumers in the growing digital economy.

Declarations

Author Contributions

Author Contributions: Conceptualization, S.S.M. and N.Y.; Methodology,

S.S.M. and N.Y.; Software, S.S.M.; Validation, N.Y.; Formal Analysis, S.S.M.; Investigation, S.S.M.; Resources, N.Y.; Data Curation, S.S.M.; Writing—Original Draft Preparation, S.S.M.; Writing—Review and Editing, N.Y.; Visualization, S.S.M. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Y. Yuan and Y. Yang, "Embracing the Metaverse: Mechanism and Logic of a New Digital Economy," *Metaverse*, vol. 2022, no. 1, pp. 1–12, 2022, doi: 10.54517/met.v3i2.1814.

[2] C. Zhang, Y. Zhao, and H. Zhao, "A Novel Hybrid Price Prediction Model for Multimodal Carbon Emission Trading Market Based on CEEMDAN Algorithm and Window-Based XGBoost Approach," *Mathematics*, vol. 10, no. 21, pp. 1–16, 2022, doi: 10.3390/math10214072.

[3] Q. Ren, K. Rong, C. Lu, G. Liu, and M. Ross, "Value-Informed Pricing for Virtual Digital Products: Evidence From Chinese MMORPG Industry," *Int. J. Mark. Res.*, vol. 2018, no. 1, pp. 1–12, 2018, doi: 10.1177/1470785318799909.

[4] H. Liu, S. Zheng, and D. Li, "PC vs App vs Mweb: Price Discounts' Effect on Customer Purchases Across Digital Channels in China," *Nankai Bus. Rev. Int.*, vol. 2023, no. 1, pp. 1–12, 2023, doi: 10.1108/nbri-06-2022-0064.

[5] S. D. Levitt, J. A. List, S. Neckermann, and D. E. Nelson, "Quantity Discounts on a Virtual Good: The Results of a Massive Pricing Experiment at King Digital Entertainment," *Proc. Natl. Acad. Sci.*, vol. 2016, no. 1, pp. 1–12, 2016, doi: 10.1073/pnas.1510501113.

[6] L. Poláček, M. Ulman, P. Cihelka, and E. Šilerová, "Dynamic Pricing in E-Commerce: Bibliometric Analysis," *Acta Inform. Pragensia*, vol. 2024, no. 1, pp. 1–12, 2024, doi: 10.18267/j.aip.227.

[7] S. Carta, A. Medda, A. Pili, D. R. Recupero, and R. Saia, "Forecasting E-Commerce Products Prices by Combining an Autoregressive Integrated Moving Average

(ARIMA) Model and Google Trends Data," *Future Internet*, vol. 11, no. 1, pp. 1–12, 2018, doi: 10.3390/fi11010005.

[8] J. Chen, N. Tournois, and Q. Fu, "Price and Its Forecasting of Chinese Cross-Border E-Commerce," *J. Bus. Ind. Mark.*, vol. 2020, no. 1, pp. 1–12, 2020, doi: 10.1108/jbim-01-2019-0017.

[9] D. P. de Amorim and M. Resende, "Exchange Rate Pass-Through to Brazilian E-Commerce Prices," *Glob. J. Emerg. Mark. Econ.*, vol. 2023, no. 1, pp. 1–12, 2023, doi: 10.1177/09749101221149251.

[10] Y. F. Tan, L.-Y. Ong, M.-C. Leow, and Y.-X. Goh, "Exploring Time-Series Forecasting Models for Dynamic Pricing in Digital Signage Advertising," *Future Internet*, vol. 13, no. 10, pp. 1–12, 2021, doi: 10.3390/fi13100241.

[11] G. Ang and E. Lim, "Temporal Implicit Multimodal Networks for Investment and Risk Management," *ACM Trans. Intell. Syst. Technol.*, vol. 2024, no. 1, pp. 1–12, 2024, doi: 10.1145/3643855.

[12] E. Popovska and G. Georgieva-Tsaneva, "ARIMA Model for Day-Ahead Electricity Market Price Forecasting," *Innov. STEM Educ.*, vol. 2022, no. 1, pp. 1–12, 2022, doi: 10.55630/stem.2022.0418.

[13] M. Wang, "Short-term Forecast of Pig Price Index on an Agricultural Internet Platform," *Agribusiness*, vol. 2019, no. 1, pp. 1–12, 2019, doi: 10.1002/agr.21607.

[14] T. M. Fahrudin, P. A. Riyantoko, K. M. Hindrayani, and I. G. Susrama Diyasa, "Daily Forecasting for Antam's Certified Gold Bullion Prices in 2018-2020 Using Polynomial Regression and Double Exponential Smoothing," *J. Int. Conf. Proc.*, vol. 2021, no. 1, pp. 1–12, 2021, doi: 10.32535/jicp.v3i4.1009.

[15] N. H. Chan and W. Liu, "Modeling and Forecasting Online Auction Prices: A Semiparametric Regression Analysis," *J. Forecast.*, vol. 2016, no. 1, pp. 1–12, 2016, doi: 10.1002/for.2420.

[16] J. A. Acuna-García, "Stock Market Forecasting Using Continuous Wavelet Transform and Long Short-Term Memory Neural Networks," *Int. J. Adv. Res. Comput. Sci.*, vol. 13, no. 6, pp. 1–12, 2022, doi: 10.26483/ijarcs.v13i6.6919.

[17] D. R. Sanjaya, B. Surarso, and T. Tarno, "Stock Price Forecasting on Time Series Data Using the Long Short-Term Memory (LSTM) Model," *Int. J. Curr. Sci. Res. Rev.*, vol. 7, no. 12, pp. 1–12, 2024, doi: 10.47191/ijcsrr/v7-i12-26.

[18] K. Li, N. Shen, Y. Kang, H. Chen, Y. Wang, and S. He, "Livestock Product Price Forecasting Method Based on Heterogeneous GRU Neural Network and Energy Decomposition," *IEEE Access*, vol. 2021, no. 1, pp. 1–12, 2021, doi: 10.1109/access.2021.3128960.

[19] I. S. Shcherbyna, "Improving Stock Price Forecasting Based on Recurrent Neural Networks," *SNSUT*, vol. 2023, no. 1, pp. 1–12, 2023, doi: 10.31673/2518-7678.2023.021313.

[20] C. Gouriéroux, J. Jasiak, and M. Tong, "Convolution-based Filtering and Forecasting: An Application to WTI Crude Oil Prices," *J. Forecast.*, vol. 2021, no. 1, pp. 1–12, 2021, doi: 10.1002/for.2757.

[21] A. M. Wahid, T. Hariguna, and G. Karyono, "Optimization of Recommender Systems for Image-Based Website Themes Using Transfer Learning," *J. Appl. Data Sci.*, vol. 6, no. 2, Art. no. 2, Mar. 2025, doi: 10.47738/jads.v6i2.671.

[22] M. R. Febrino, D. Permana, Syafriandi, and N. Amalita, "Comparison of Forecasting Using Fuzzy Time Series Chen Model and Lee Model to Closing Price of Composite Stock Price Index," *Unp J. Stat. Data Sci.*, vol. 2023, no. 1, pp. 1–12, 2023, doi:

10.24036/ujsds/vol1-iss2/22.

[23] S. Deng, Y. Zhu, X. Huang, S. Duan, and Z. Fu, “High-Frequency Direction Forecasting of the Futures Market Using a Machine-Learning-Based Method,” *Future Internet*, vol. 14, no. 6, pp. 1–12, 2022, doi: 10.3390/fi14060180.

[24] S. Deng, X. Huang, Y. Zhu, and Z. Fu, “A Decision Support System for Trading in Apple Futures Market Using Predictions Fusion,” *IEEE Access*, vol. 2021, no. 1, pp. 1–12, 2021, doi: 10.1109/access.2020.3047138.

[25] Y. Duan, J. Zhang, X. Wang, M. Feng, and L. Ma, “Forecasting Carbon Price Using Signal Processing Technology and Extreme Gradient Boosting Optimized by the Whale Optimization Algorithm,” *Energy Sci. Eng.*, vol. 2024, no. 1, pp. 1–12, 2024, doi: 10.1002/ese3.1655.

[26] S. Labunska, “Prospects for the Development of Market Infrastructure Under Virtual Tokenized Assets Influencing,” *Econ. Anal.*, vol. 2023, no. 3, pp. 1–12, 2023, doi: 10.35774/econa2023.03.130.

[27] A. M. Wahid, T. Turino, K. A. Nugroho, T. S. Maharani, D. Darmono, and F. S. Utomo, “Optimasi Logistic Regression dan Random Forest untuk Deteksi Berita Hoax Berbasis TF-IDF,” *J. Pendidik. Dan Teknol. Indones.*, vol. 4, no. 8, pp. 1–12, 2024, doi: 10.52436/1.jpti.602.

[28] S. Schöbel and J. M. Leimeister, “Metaverse Platform Ecosystems,” *Electron. Mark.*, vol. 2023, no. 1, pp. 1–12, 2023, doi: 10.1007/s12525-023-00623-w.