Feature Selection for High-Dimensional Data: A Case Study of NFT Valuation

Geun-Cheol Lee [1], Heejung Lee [2], and Hoon-Young Koo [3],*
[1] College of Business, Konkuk University, Seoul 05029, South Korea
[2] School of Interdisciplinary Industrial Studies, Hanyang University, Seoul 04763, South Korea
[3] School of Business, Chungnam National University, Daejeon 34134, South Korea
Email: gclee@konkuk.ac.kr (G.L.); stdream@hanyang.ac.kr (H.L.); koohy@cnu.ac.kr (H.K.)
*Corresponding author

***Abstract*—In this study, we propose hedonic models for valuing Non-Fungible Tokens (NFTs) from the Azuki collection. We first analyze the NFT's metadata and introduce a market volatility-robust dependent variable. Specific information of Azuki attributes is encoded via Term Frequency-Inverse Document Frequency (TF-IDF) to reflect both presence and collection-wide scarcity, yielding hundreds of features for each token. Two hedonic models are considered: a linear model and a squared model. To address high dimensionality, we tailor three variable-selection procedures—forward, backward, and stepwise—and compare them with regularization benchmarks and machine-learning methods. Using actual Azuki transaction data, we evaluate performance on a train-validation partition. The squared model overfits out of sample, while the linear model generalizes better and is adopted as the baseline. Applying variable selection to the linear baseline improves both parsimony and predictive performance. Machine-learning models exhibit very high training fit but notable performance degradation on the validation set, indicating overfitting in this setting. Overall, carefully specified hedonic models combined with principled variable selection offer competitive, interpretable, and more generalizable NFT valuation.**

***Keywords*—Non-Fungible Token (NFT), NFT valuation, hedonic model, variable selection, high-dimensional data, Term Frequency-Inverse Document Frequency (TF-IDF), Azuki**

I. Introduction

Over the past decade, blockchain technology has undergone remarkable development, led by the emergence of Bitcoin as the first widely recognized application. Among the diverse innovations built upon blockchain, one of the most prominent applications is the Non-Fungible Token (NFT). While the concept of NFTs has been discussed from various perspectives, a number of recent studies suggest a broadly accepted definition: NFTs can be regarded as tradeable digital assets, recorded in smart contracts and managed through blockchain technology, that grant ownership rights over physical or digital assets such as videos, images, and artworks [1–3].

During 2021 and 2022, the NFT market experienced an extraordinary surge in trading volume. NFT sales volumes soared from merely 94.9 million USD in 2020 to approximately 24.9 billion USD in 2021, with the number of wallets trading NFTs exploding from around 545,000 to 28.6 million USD [4]. However, beginning in 2022, the market entered a pronounced contraction. According to market research data provided by Statista [5], the downturn following the market peak is evident in both aggregate revenue and per-user metrics. Specifically, global NFT revenue declined from approximately 1,581.3 million USD in 2022 to 611.1 million USD in 2023 and has remained at this lower level since then. A similar downward trajectory is observed in the Average Revenue per User (ARPU), which reached a high of 413 USD in 2021, but dropped to 181.1 USD in 2022 and further to 59.7 USD in 2023. Even so, the NFT ecosystem has shown resilience. Despite the downturn, some high-profile collections have maintained their relevance and formed a stabilizing backbone for the market. Furthermore, despite widespread devaluation—some estimates suggest up to 95% of NFTs may have lost meaningful value—trading activity continues to persist, with weekly volumes still reaching tens of millions of dollars [6].

Despite the downturn, NFT trading remains an active and meaningful segment of the digital asset market. In such a highly volatile and speculation-driven environment, establishing a rational framework for NFT valuation is essential to mitigating risks faced by market participants. According to a recent review paper on NFT [7], the two most highly cited papers in the NFT literature are Dowling's works regarding the determinants of NFT valuation [8, 9], which shows strong scholarly interest in NFT valuation. This growing academic interest is also reflected in a review study on NFT pricing [10]. This study categorized the determinants of NFT prices into three groups: external factors beyond the collection, internal factors within the collection, and inter-collection factors. In addition, the study further classified NFT into five categories: art, gaming, collectibles, utilities, and metaverse.

Manuscript received October 9, 2025; revised October 28, 2025; accepted November 11, 2025; published January 20, 2026.

doi: 10.12720/jait.17.1.141-152

In this study, we aim to propose and empirically validate an NFT valuation model. In particular, we focus on Profile Picture (PFP) NFT projects, which fall under the collectibles category among the five NFT types introduced above. PFP NFT projects include some of the most iconic NFT projects such as CryptoPunks [11] and Bored Ape Yacht Club [12]. As such, they represent the most familiar and widely recognized form of NFTs to market participants. Their importance is further identified by sales volume statistics: according to the all-time rankings on OpenSea [13], the leading NFT marketplace, four of the top five projects by sales volume are PFP NFTs as shown in Table I.

TABLE I. Top Five NFT Collections on OpenSea in Terms of Sales Volume as of September 5, 2025

Collection	Category	Floor Price	Sales Volume	Number of Items
Bored Ape Yacht Club	PFP	9.14 ETH	1.5M ETH	9,998
CryptoPunks	PFP	48.00 ETH	1.3M ETH	9,994
Mutant Ape Yacht Club	PFP	1.3548 ETH	1.1M ETH	19,555
Azuki	PFP	1.78 ETH	824.8K ETH	10,000
Otherdeed for Otherside	Metaverse	0.104 ETH	605.7K ETH	36,263

Table I shows that, except for Otherdeed for Otherside—a metaverse project—all of the listed collections are classified as PFP NFTs. In terms of cumulative trading volume, Bored Ape Yacht Club (BAYC) leads the market; however, the floor price of CryptoPunks is the highest, at approximately 50 ETH. Most collections consist of approximately 10,000 tokens, which has become a common standard for PFP NFTs. Notably, Mutant Ape Yacht Club can be regarded as a derivative of BAYC, reinforcing the central roles of BAYC and CryptoPunks in the NFT market.

Given the importance of PFP NFTs, it is not surprising that much of the existing NFT valuation research has focused on these collections, particularly CryptoPunks and BAYC. Among them, CryptoPunks holds a particularly important place, as it was launched in June 2017 by Larva Labs and is widely regarded as one of the earliest and most valuable NFT projects. Kong and Lin [14] analyzed 23,206 transactions of CryptoPunks conducted between June 2017 and December 2022, while Schaar and Kampakis [15] investigated 11,864 transactions spanning June 2018 to May 2021. Both studies employed hedonic regression models to examine the determinants of CryptoPunks prices, identifying how various attributes and rarity contribute to valuation outcomes. The BAYC, launched by Yuga Labs in April 2021, consists of 10,000 unique digital ape images and several empirical studies have focused on BAYC to explore its valuation mechanisms. Lee *et al.* [16] applied Formal Concept Analysis (FCA) to analyze the effect of rarity on BAYC prices, while Lee *et al.* [17] proposed linear and quadratic hedonic pricing models to evaluate its value. Koo *et al.* [18] employed Structural Equation Modeling (SEM) together with multi-group analysis to investigate causal relationships among valuation determinants and to test differences across investor types. Furthermore, Mekacher *et al.* [19] examined rarity quantification and its impact on market behavior across 410 PFP NFT collections, with CryptoPunks and BAYC serving as the primary case studies. Xiong *et al.* [20] broadened the scope of NFT valuation research by analyzing data from PFP projects such as Sappy Seals and Lazy Lions. Their study proposed a hedonic regression model grounded in rare attributes while also integrating market-level factors. In addition, factor analysis was employed to assess the robustness and improvement of the proposed pricing framework. Recently, studies have begun incorporating visual attributes of NFTs into price prediction models, leveraging transfer learning and deep neural networks to capture image-based rarity [21, 22].

In this study, we propose an NFT valuation model focusing on the Azuki project. To the best of our knowledge, no existing work has systematically investigated valuation mechanisms for Azuki NFTs. The Azuki collection, launched in January 2022 by Chiru Labs, consists of 10,000 Japanese anime-inspired avatar NFTs [2, 23]. Aside from BAYC and CryptoPunks, Azuki ranks highest in all-time sales volume and remains a top performer in the NFT market, and it continues to rank among the top performers in NFT sales [24]. Similar to other PFP NFTs, Azuki items are generated from combinations of multiple attributes; specifically, the collection is characterized by 12 attributes—Type, Special, Clothing, Offhand, Hair, Headgear, Face, Neck, Eyes, Mouth, Ear, and Background. Table II presents examples of some of the most expensive Azuki NFTs ever traded [25]. In the table, the values of each attribute, referred to as traits [19], thereby allowing us to identify which attribute values (i.e., traits) contribute to higher valuations

TABLE II. Selected Items of the Highest-Priced Azuki NFTs

Image and Attribute		Azuki #9605	Azuki #5172	Azuki #4666
Profile Picture				
Attribute	Type	Spirit	Spirit	Spirit
	Special	Fireflies	N/A	Fireflies
	Clothing	N/A	Azuki track jacket	White qipao with fur
	Offhand	Golden shuriken	Hand wrap	Golden umbrella
	Hair	Spirit fluffy	Spirit spiky	Spirit ponytail
	Headgear	N/A	Ikz baseball cap	N/A
	Face	N/A	N/A	N/A
	Neck	Golden headphones	N/A	Golden headphones
	Eyes	Chill	White	Striking
	Mouth	Pout	Pout	Grin
	Ear	Small hoop	N/A	N/A
	Background	Cool gray	Red	Dark purple

As shown in Table II, a common feature among these highest-priced items is the Spirit type, which is one of the rarest and most valued attributes in the Azuki collection. Each Azuki item is identified by a unique number, called token ID, which serves as its identifier within the collection. Note that not every item possesses values for all attributes. For instance, token ID #9605 does not contain entries for Clothing, Headgear, or Face, which are denoted as "N/A" in the table. Despite such missing attributes, these items exhibit combinations of highly valued traits—most notably the Spirit type and golden accessories—that substantially increase their market prices. While these features reflect the general characteristics of PFP NFTs, Azuki also exhibits distinctive features. To show a comparative perspective, Table III summarizes the key characteristics of representative PFP NFT projects—CryptoPunks, BAYC, and Azuki.

TABLE III. KEY CHARACTERISTICS OF CRYPTOPUNKS, BAYC, AND AZUKI NFT COLLECTIONS

Category	CryptoPunks	BAYC	Azuki
# of items	9,904	9,998	10,000
# of attributes	2	7	12
Attributes (# of traits)	1.Accessory (95) 2.Type (5)	1.Background (8) 2.Clothes (43) 3.Earring (6) 4.Eyes (24) 5.Fur (19) 6.Hat (36) 7.Mouth (33)	1.Background (8) 2.Clothing (98) 3.Ear (32) 4.Eyes (27) 5.Hair (123) 6.Face (19) 7.Headgear (36) 8.Mouth (33) 9.Neck (15) 10.Offhand (53) 11.Special (9) 12.Type (4)
Remarks	Can have up to 7 accessories	Some attributes allow missing values	

Table III provides a comparative summary of the key characteristics of three representative PFP NFT collections. Although all three collections consist of approximately 10,000 items, they differ considerably in terms of attribute structure and diversity. CryptoPunks is defined by only two attribute categories, whereas BAYC incorporates seven. Azuki further extends this complexity by incorporating 12 attribute categories with a significantly larger number of possible values, such as 123 distinct hair traits and 98 clothing options. This richer attribute space results in a high-dimensional dataset, making Azuki notably more complex and distinctive compared to the other two leading PFP NFTs.

In this study, we propose a valuation model that considers the distinctive characteristics of Azuki NFTs. As noted in Table I, Azuki is one of the most actively traded NFT collections, yet it has not been systematically studied in prior literature. Moreover, compared with other PFP NFT collections, Azuki exhibits a much higher degree of feature dimensionality, which makes a variable-selection process particularly important. Therefore, the present study distinguishes itself from prior NFT valuation research by (i) being the first to develop and validate a hedonic pricing framework for Azuki, and (ii) addressing the challenges of high-dimensional NFT metadata through systematic feature-selection procedures. The remainder of this paper is organized as follows. Section II provides an exploratory analysis of the Azuki NFT dataset. Section III introduces the hedonic modeling framework, including the specification of the dependent variable and the construction of independent variables. Section IV presents three alternative methods for variable selection to address the high-dimensional structure of the dataset. Section V evaluates the performance of the proposed valuation model, comparing it against benchmark approaches through a series of empirical experiments. Finally, Section VI concludes with a summary of findings, practical implications, and directions for future research.

II. EXPLORATORY DATA ANALYSIS OF THE AZUKI

In this section, we conduct an exploratory statistical analysis of the Azuki NFT dataset in order to better understand its structural characteristics. Specifically, we focus on frequency-based analyses of the 12 attributes that define each Azuki token. Table IV summarizes the initial statistics for each attribute, including the number of unique values and the Gini indices to measure the homogeneity of each attribute. The Gini index measures distributional inequality, where values close to 1 indicate a highly uneven distribution dominated by a small number of traits, while values close to 0 suggest a more balanced distribution.

TABLE IV. UNIQUE VALUE COUNTS AND GINI INDICES OF AZUKI NFT ATTRIBUTES

Attribute	Unique Value	Unique Values[1]	Gini Index	Gini Index[1]
Type	4	4	0.669	0.669
Hair	124	123	0.241	0.235
Clothing	99	98	0.32	0.314
Eyes	27	27	0.28	0.28
Mouth	30	30	0.301	0.301
Offhand	54	53	0.656	0.533
Background	8	8	0.284	0.284
Neck	16	15	0.79	0.368
Headgear	37	36	0.723	0.292
Ear	33	32	0.839	0.292
Face	20	19	0.706	0.252
Special	10	9	0.844	0.127

Note: [1] "N/A" is excluded.

Table IV reports the number of unique values and the Gini indices for each of the 12 attributes. For each attribute, results are reported both including and excluding missing values (N/A). A total of 461 distinct trait values are identified when including N/A categories, and 454 when they are excluded. Among these, Hair (123) and Clothing (98) account for the largest numbers of unique traits, thereby playing a dominant role in driving the high-dimensional nature of the Azuki dataset. As the table shows, with the exception of Type, Eyes, Mouth, and Background, most attributes contain a substantial proportion of missing values. Notably, attributes such as Headgear, Ear, Face, and Special exhibit large differences in their Gini indices depending on whether N/A values are included. This indicates that, although these attributes appear highly uneven when N/A is included, the remaining non-missing traits are relatively evenly distributed across items. In the next table, we further investigate these characteristics by presenting the most and least frequent values for each attribute.

Table V presents the most and least frequent values for each attribute, along with their corresponding counts. Several patterns can be observed. First, for fundamental traits such as Type and Background, the distributions are highly skewed: the majority of tokens are categorized as Human (9,018) and share Off White backgrounds, while rare categories such as Spirit (97) or Dark Purple backgrounds (463) are extremely scarce. Second, attributes with a wide variety of categories, such as Hair and Clothing, reveal a long-tailed distribution. While common traits include Maroon Bun or Light Kimono, some traits appear fewer than ten times, such as Black Blonde Half Bun or Golden Cat Kigurumi. Third, accessory-related attributes such as Offhand, Neck, Headgear, Ear, and Face contain a large proportion of missing values, reflecting that many tokens do not feature these traits. Within the non-missing subset, however, rare traits such as Golden Zanbato (11), Golden Headphones (35), and Red Bean (14) emerge as highly distinctive identifiers of rarity. Finally, the Special attribute, which is present in fewer than 10% of items, is dominated by N/A values. Nonetheless, when present, traits such as Fireflies (88) or Lightning (48) represent exceptionally rare and visually distinctive features that often drive premium valuations. As you can see, while some traits are broadly shared and form the visual identity of the collection, others are extremely rare and serve as critical drivers of scarcity and valuation.

TABLE V. MOST AND LEAST FREQUENT VALUES OF EACH ATTRIBUTES

Attribute	Most Frequent	2nd Most Frequent	Least Frequent	2nd Least Frequent
Type	Human (9018)	Blue (444)	Spirit (97)	Red (441)
Hair	Maroon Bun (150)	Brown Spiky (142)	N/A (8)	Black Blonde Half Bun (12)
Clothing	Light Kimono (311)	Maroon Yukata (221)	Golden Cat Kigurumi (5)	N/A (5)
Eyes	Closed (1551)	Determined (742)	Lightning (46)	Fire (57)
Mouth	Relaxed (834)	Closed (812)	Surprised (74)	Sleep Bubble (88)
Offhand	nan (3165)	Katana (439)	Golden Zanbato (11)	Golden Sheathed Katana (12)
Background	Off White D (1990)	Off White C (1962)	Dark Purple (463)	Cool Gray (483)
Neck	N/A (7746)	Chill Headphones (400)	Golden Headphones (35)	Sloth Headphones (36)
Headgear	N/A (6512)	IKZ Baseball Cap (265)	Red Panda Beanie (28)	Frog Beanie (30)
Ear	N/A (8181)	Corded Earbuds (164)	Red Bean (14)	Blue Bean (16)
Face	N/A (6790)	Red Stripes Face Paint (290)	Heart Eye Patch (65)	Lipstick Kiss (72)
Special	N/A (9371)	Fireflies (88)	Lightning (48)	Water (49)

Building on the preceding descriptive analysis of Azuki attributes, we now turn to a more detailed examination of the distribution of individual traits. As noted above, the Azuki collection comprises a total of 454 unique trait values across its twelve attributes. To better capture the rarity structure of these traits, we analyze the frequency distribution of all unique values. Fig. 1 presents a pie chart that groups trait frequencies.

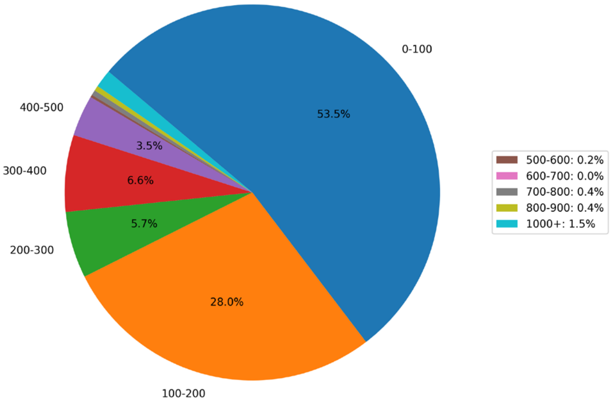

Fig. 1. Pie chart of attribute frequency distributions in the Azuki NFT collection. Frequencies of traits are grouped into bins of width 100, with all frequencies above 1000 combined into a single category ("1000+").

Fig. 1 illustrates the frequency distribution of the 454 unique Azuki trait values, grouped into bins of width 100. The chart clearly demonstrates that traits with frequencies below 200 account for more than 80% of all values, while traits with fewer than 100 occurrences alone make up over 50%. Given that the Azuki collection consists of 10,000 items, any trait appearing fewer than 100 times would be regarded as rare. However, more than half of all traits fall into this category. This finding suggests that relying solely on raw frequency as an indicator of rarity may be insufficient for accurate valuation. Instead, it becomes necessary to consider additional contextual factors, such as the attribute category to which a trait belongs and its interactions with other traits.

TABLE VI. LEAST AND MOST FREQUENT TRAITS OF AZUKI NFTS

Least Frequent Traits			Most Frequent Traits		
Trait	Freq.	Attribute	Trait	Freq.	Attribute
Golden Cat Kigurumi	5	Clothing	Human	9018	Type
Golden Sloth Kigurumi	6	Clothing	Off White D	1990	Background
Golden Red Panda Kigurumi	6	Clothing	Off White C	1962	Background
Golden Frog Kigurumi	10	Clothing	Off White A	1814	Background
Golden Zanbato	11	Offhand	Off White B	1758	Background
Spirit Bob	12	Hair	Closed	1551	Eyes
Spirit Long	12	Hair	Red	1006	Background
Black Blonde Half Bun	12	Hair	Relaxed	834	Mouth
Golden Katana	12	Offhand	Closed	812	Mouth
Golden Monkey King Staff	12	Offhand	Determined	742	Eyes

Table VI presents a comparative summary of the least and most frequent trait values within the Azuki collection. Among the least frequent traits, a notable pattern is the predominance of values containing the prefix "Golden". These rare values are concentrated within a limited set of attributes, particularly Clothing, Hair, and Offhand. On the other hand, the most frequent traits are dominated by fundamental features. The Human type stands out with an exceptionally large count of 9,018. In addition, several values included in the Background attribute are also among the most frequent.

To incorporate both attribute categories and individual trait frequencies, we calculate the rarity score of each Azuki token based on its constituent traits [19]. The rarity score of a specific trait r_t is defined as:

$$r_t = \frac{1}{(f_t/10000)} \tag{1}$$

where f_t is frequency of trait t out of 10,000, i.e., the denominator represents the proportion of tokens in the collection that contain the trait. The rarity score of token k, R_k, is then obtained by summing the rarity scores of all traits that the token possesses:

$$R_k = \sum_{t \in T_k} r_t \tag{2}$$

where T_k is the trait set of token k. A larger value of R_k indicates a higher level of rarity. Note that, in the calculation of rarity scores, missing values were also taken into account. For each attribute, the absence of a trait was treated as a valid category and incorporated into the computation. The next figure presents the histogram of rarity scores across all 10,000 Azuki tokens.

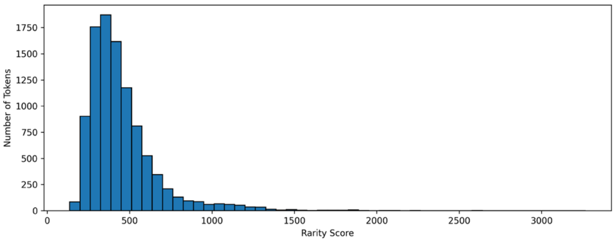

Fig. 2. Histogram of rarity scores across all 10,000 Azuki tokens.

As you can see from Fig. 2, the rarity has a highly skewed distribution, where the majority of tokens fall within the lower score. Only a small fraction of tokens achieve rarity scores above 1,000, and extremely rare cases exceed 2,000 or even 3,000. In the following figure, we compare two Azuki tokens that represent the extremes of the distribution: the one with the highest rarity score and the one with the lowest rarity score. Interestingly, the token with the highest rarity score corresponds to Azuki #9605, which was already introduced earlier in Table II as one of the historically most expensive Azuki NFTs ever traded. The following table provides a comparison of two tokens with the highest and lowest rarity scores.

As shown in Table VII, Azuki #9605, which holds the highest rarity score, is characterized by a distinctive science-fiction-like figure with golden accessories. In contrast, Azuki #8546, with the lowest rarity score, portrays a human figure dressed in a simple and ordinary costume. The disparity in rarity scores is mirrored in their market performance. A particularly noteworthy feature of Azuki #9605 is its Clothing attribute, which is marked as "N/A". That is, the absence of clothing ("nude") significantly increases the trait rarity score, thereby elevating the overall rarity of the token.

TABLE VII. COMPARISON BETWEEN TWO AZUKI TOKEN WITH THE HIGHEST AND LOWEST RARITY SCORES

Image and Value	Azuki # 9605	Azuki #8546
Profile Picture		
Rarity Score	4080.89	153.41
Last Sale Price	420.7 ETH	29 ETH

In this section, we have analyzed the metadata of the Azuki collection, providing descriptive statistics of attributes and trait values, as well as a systematic assessment of rarity scores. These findings serve as a critical foundation for the development of valuation models in subsequent sections.

III. HEDONIC MODELS

In this section, we introduce the hedonic models employed for the valuation of the Azuki collection. As discussed in Section I, applications of hedonic modeling to PFP NFTs are well-documented in the literature [14, 15, 17, 20]. Beyond PFP NFTs, prior research has also extended the use of hedonic models to other categories such as art and metaverse NFTs [26–28]. The fundamental premise of the hedonic pricing model is that the price of a heterogeneous asset is determined not only by external market factors but also by the intrinsic characteristics of the asset itself. Historically, this framework has been widely applied in real estate and traditional art markets [10]. Given that NFTs share the feature of being highly heterogeneous digital assets in one collection, hedonic models naturally provide an appropriate and effective framework for analyzing NFT valuations.

In the remainder of this section, we first describe how the dependent and independent variables of the hedonic model are constructed from the Azuki dataset. Following this, we present the hedonic models that incorporate the corresponding variables. The next subsection begins with a detailed discussion of the dependent variable.

A. Dependent Variable

In this study, the goal of the hedonic model is to evaluate the value of NFTs; therefore, the dependent variable must be the economic value of each token, specifically its price.

However, similar to other cryptocurrencies, NFT prices are subject to extreme volatility [29]. To account for this volatility, it is necessary to incorporate a dependent variable that captures the relative value of each token while neutralizing fluctuations caused by market-wide dynamics. Following the previous studies [16, 17], we adopt the Premium Ratio to measure the values of NFTs. The Premium Ratio measures the relative value of an individual token at the time of transaction, which can offset price fluctuations caused by overall market movements. For a transaction s of token k, the Premium Ratio, m_{sk}, is defined as:

$$m_{sk} = \frac{p_{sk}}{MA(20)_s} \tag{3}$$

where p_{sk} denotes the transaction price of token k at transaction s, and $MA(20)_s$ represents the 20-day moving average price of all Azuki tokens at the date of transaction s. As such, the Premium Ratio reflects the relative market value of the token within a 20-day window. Since multiple transactions may exist for the same token, the final dependent variable is derived by averaging the Premium Ratios across all transactions of a given token. Formally, the Average Premium Ratio (APR) for token k, is defined as:

$$M_k = \sum_{s \in \Theta_k} \frac{m_{sk}}{|\Theta_k|} \tag{4}$$

where Θ_k denotes the set of all transactions for token k. The M_k thus provides a normalized measure of the relative value of each token, accounting for market-wide fluctuations. In this study, the Average Premium Ratio is used as the dependent variable in the hedonic model. Given the substantial variation in values across tokens, the dependent variable is log-transformed prior to model estimation.

B. *Independent Variables*

In a hedonic pricing model, the explanatory variables should represent the intrinsic characteristics of the assets under consideration. For Azuki NFTs, these characteristics correspond to the trait values of each attribute. Since these traits are expressed as categorical string values, it is necessary to transform them into numerical representations suitable for regression modeling. A straightforward approach would be to apply one-hot encoding, which creates binary indicators for the presence of each trait. However, in this study we employ Term Frequency-Inverse Document Frequency (TF-IDF), to better capture the informational content of traits. TF-IDF is widely used in text mining to evaluate the importance of terms within a corpus by balancing their frequency within a document against their overall rarity across the entire collection.

Here, each Azuki token is treated as a "document", and its attribute-trait combinations are regarded as "terms". Using the TF-IDF method, traits that appear frequently across the collection receive lower weights, whereas rare traits are assigned higher weights, thereby emphasizing their contribution to uniqueness. The implementation was carried out using the TfidfVectorizer function from Python's scikit-learn library.

The resulting feature space consists of 461 dimensions, each corresponding to a unique attribute-trait pair identified across the 10,000 tokens. Thus, every Azuki token is represented as a sparse 461-dimensional TF-IDF vector. This representation not only reflects the richness of the Azuki dataset but also highlights the high-dimensional structure of PFP NFTs. Importantly, by leveraging the weighting mechanism inherent in TF-IDF, the independent variables capture both the presence of traits and their relative rarity within the collection, providing a robust numerical foundation for the subsequent hedonic regression analysis.

C. *Linear Model*

With the APR defined in Section III.A and the TF-IDF variables prepared in Section III.B, we now construct hedonic pricing models for the Azuki collection. In this study, we propose two hedonic models. The first is a linear hedonic regression model that includes all 461 TF-IDF variables as first-order terms. The second, which will be introduced in the next subsection, extends the specification by adding squared terms resulting in 922 independent variables in the model.

The linear model assumes that each attribute-trait pair contributes additively and proportionally to the NFT's value. Its mathematical form is given as follows:

$$log(M) = \beta_0 + \sum_{i \in TF} \beta_i d_i + \epsilon \tag{5}$$

where *M* denotes the APR introduced in Eq. (4), β_0 is the intercept, β_i is the coefficient associated with the *i*-th TF-IDF variable, TF represents the set of 461 TF-IDF variables, d_i is the TF-IDF score of the *i*-th trait, and ϵ is the error term capturing unobserved influences. This linear specification provides a straightforward benchmark to assess the explanatory power of trait-level features. By assuming a purely linear relationship, it enables us to evaluate the direct marginal effects of each attribute-trait pair on NFT valuation. This serves as the baseline before introducing nonlinear extensions, which are presented in the next subsection.

D. *Squared Model*

The second hedonic specification is the squared model, which extends the linear framework by incorporating quadratic terms. The motivation for this extension lies in the speculative nature of NFT markets, where certain traits may exert disproportionately large effects on token valuations. To capture such potential nonlinearities, the squared model augments the 451 first-order TF-IDF variables with their squared counterparts, resulting in a total of 902 explanatory variables. Formally, the squared hedonic model is expressed as:

$$log(M) = \beta_0 + \sum_{i \in TF} \beta_i d_i + \sum_{i \in TF} \gamma_i d_i^2 + \epsilon \tag{6}$$

where γ_i denotes the coefficient associated with the squared term of the *i*-th TF-IDF variable, and the remaining notation is identical to that in Eq. (5). By allowing for nonlinear effects, this model provides a richer representation of the relationship between trait characteristics and NFT valuations. In particular, it enables the analysis of whether rare or distinctive traits have an amplified impact on prices, beyond the proportional contribution assumed in the linear model. This extended specification thus serves as a crucial step in examining how high-dimensional trait structures influence the valuation of Azuki NFTs.

IV. Variable Selection

As discussed in the previous section, the inclusion of squared terms in the hedonic model results in a high-dimensional representation of the Azuki dataset. Compared to other prominent PFP NFT projects such as CryptoPunks or BAYC, the Azuki collection exhibits a much larger number of unique attribute-trait combinations, as presented in Table III. High-dimensional datasets of this kind are increasingly common in modern statistical applications, and variable selection has become an essential process to enhance model interpretability and predictive performance [30]. In this section, we introduce three well-known variable selection procedures—forward selection, backward elimination, and stepwise selection—and adapt them to the context of our hedonic models. The following subsections describe these procedures in detail.

A. Forward Selection

The first variable selection method employed in this study is the forward selection procedure. As the name suggests, this approach begins with an intercept-only model that contains no independent variables (i.e., features). Variables are then added one by one, based on their statistical significance and contribution to model fit, thereby gradually expanding the model. At each step, only the variable that provides a certain amount of improvement in explanatory power is included, and this process continues until no further meaningful improvement can be achieved. The detailed procedure is summarized in the below.

```
Forward Selection Procedure
Input:   Dependent variable and feature set
Output:  Selected feature subset
Step 1:  Rank all candidate features by their $p$-values from
         the full model.
Step 2:  Start with an intercept-only model (no features) and
         compute adjusted $R^2$.
Step 3:  Sequentially test features in ranked order:
         Temporarily add one feature at a time to the current
           model.
         If adjusted $R^2$ improves, keep the feature and update
           the model.
         Otherwise, count as a failure. Stop the procedure
           when the number of consecutive failures exceeds a
           threshold.
Step 4:  Return the final set of selected features.
```

In the context of this study, the initial set of candidate features corresponds to the set of independent variables defined in the hedonic models. Specifically, for the linear model in Eq. (5), 461 first-order TF-IDF variables are considered, while for the squared model in Eq. (6), the pool expands to 922 variables, including both first-order and quadratic terms. The forward selection process continues until no improvement in adjusted R^2 is observed for 10 consecutive candidate features, which is set as the stopping threshold.

B. Backward Selection

The second variable selection method is backward selection, which operates in the opposite way of forward selection. Instead of starting from an empty model, the backward procedure begins with the full model that includes all candidate features. At each step, a feature presumed the least important is considered for removal. The model is then refitted without this feature, and if the explanatory power of the model does not decrease, the feature is permanently eliminated. This iterative process continues until no further improvements can be made. The detailed procedure is summarized in the below.

```
Backward Selection Procedure
Input: Dependent variable and feature set
Output:  Selected feature subset
Step 1:  Start with a full model including all candidate
           features.
Step 2:  Fit the model. Compute adjusted $R^2$ and compute $p$-
           values for all features.
Step 3:  Identify the feature with the largest $p$-value.
Step 4:  Temporarily remove that feature and refit the
           model.
         If adjusted $R^2$ does not decrease, remove the feature
           permanently.
         Otherwise, retain the feature and count as a non-
           removal.
Step 5:  Repeat Steps 2–4 until no feature removal occurs for
           a fixed number of consecutive iterations.
Step 6:  Return the final set of selected features.
```

In this study, the termination criterion for the backward selection procedure is set based on consecutive non-removals of a candidate feature. Specifically, if no feature is removed for a fixed number of iterations, the procedure is stopped. For our implementation, this threshold is set to 10 consecutive iterations without removal.

C. Stepwise Selection

The final method, stepwise selection, combines the logic of forward and backward selection into an iterative procedure. The basic structure of this approach is to alternate between adding features and removing features, thereby refining the model in both directions. Detailed steps of this process are outlined below.

```
Stepwise Selection Procedure
Input:   Dependent variable and candidate feature set; initial
           model.
Output:  Final selected feature subset
Step 1:  Fit the initial model, compute adjusted $R^2$.
Step 2:  (Forward step)
         For a randomly selected feature not yet in the model:
                Temporarily add the feature to the current model.
                Refit and compute adjusted $R^2$.
                If adjusted $R^2$ improves, keep the feature, update the model
                Otherwise, Do not add the feature.
        End forward step if no feature addition occurs for a fixed number of consecutive times.
Step 3: (Backward step)
        For a randomly selected feature currently in the model:
                Temporarily remove the feature.
                Refit and compute adjusted $R^2$.
                If adjusted $R^2$ does not decrease, Remove the feature, update the model.
                Otherwise, Keep the feature.
        End backward step if no feature removal occurs for a fixed number of consecutive times.
Step 4: If neither Step 2 nor Step 3 changes the model in the current iteration, Stop; Otherwise go to Step 2
Step 5: Return the final set of selected features.
```

In this study, the initial model for the stepwise procedure is taken from the result of the forward selection process. While forward and backward selection rely on *p*-values from model fitting to determine candidate features for inclusion or removal, the stepwise procedure adopts a randomized choice of candidate features in order to broaden the search space and reduce potential bias toward early-ranked variables. Within the stepwise framework, the forward step terminates when no additional feature is accepted for ten consecutive trials, and similarly, the backward step terminates when no feature can be removed for ten consecutive iterations.

V. Computation Experiments

In this section, we present the computational experiments conducted to evaluate the performance of the proposed hedonic models and variable selection procedures. To this end, we compare our approach against several benchmark methods and report the results.

A. Data

We begin by describing the dataset used in this study and the procedure for constructing the training and validation sets. The transaction data were collected from Dune Analytics (dune.com), a blockchain analytics platform that provides publicly accessible dashboards and query-based data extraction from on-chain sources. The SQL query used to extract the Azuki transaction data is provided in Appendix A. The dataset covers the period from January 12, 2022, to May 24, 2024, comprising a total of 30,114 transactions. For robust estimation, we limited the analysis to tokens with more than one transaction, resulting in 7,143 tokens out of the 10,000 in the Azuki collection being included in the experiments.

The dataset was then partitioned into training and validation sets for fair comparison and to avoid overfitting. The partition has been done with a 70:30 ratio using stratified sampling. Stratification was necessary because the attribute Special has very few unique values and is predominantly represented by N/A entries, as observed earlier in Table IV. Without stratification, simple random sampling could lead to situations where rare trait values are absent in either the training or validation set. By ensuring proportional representation of the Special traits, the stratified partitioning provides a more reliable evaluation framework for the proposed models.

B. Metrics

To assess how well the proposed models and benchmarks explain NFT valuations, we employ several statistical fitness measures. The primary measure is the coefficient of determination, R^2, which captures the proportion of variance in the dependent variable explained by the model. However, since the Azuki dataset is high-dimensional with a large number of independent variables, the adjusted R^2 is used as the essential metric. Unlike the plain R^2, adjusted R^2 penalizes the inclusion of non-informative variables, thereby providing a more reliable indicator of explanatory power in high-dimensional settings. The mathematical formulas of the two metrics are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

$$Adjusted\ R^2 = 1 - \left(\frac{n-1}{n-p-1}\right)\left(1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right) \tag{8}$$

where n denotes the number of observations, p is the number of estimated coefficients, y_i represents the actual value of $\log(M)$, and $\hat{y}_i$ is the predicted value. In addition, we adopt two widely used information criteria: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both AIC and BIC balance model fit against model complexity, but they emphasize different goals. AIC prioritizes predictive accuracy, even for complex models, whereas BIC applies stronger penalties for complexity and thus favors simpler, more parsimonious models [31]. The formulas of these two criteria are as follows:

$$AIC\ =\ n \cdot ln\left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}\right) + 2p \tag{9}$$

$$BIC\ =\ n \cdot ln\left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}\right) + p \cdot ln(n) \tag{10}$$

where the notations are the same as those in Eqs. (7) and (8).

C. Benchmarks

Next, we introduce the benchmark methods used for comparative experiments. The benchmarks are divided into two categories. The first category consists of regularization-based methods, which are widely applied to high-dimensional datasets and inherently perform embedded variable selection. In this study, we include LASSO and Elastic Net as representatives of this class. The second category comprises machine learning approaches, which are among the most widely used techniques for predictive modeling in structured and unstructured data domains. Specifically, we employ Random Forest, Support Vector Regression, XGBoost, and LightGBM as benchmarks. In total, six benchmark methods are considered, providing a broad comparative framework against which the performance of the proposed hedonic models can be evaluated.

Least Absolute Shrinkage and Selection Operator (LASSO) introduces an L_1 regularization term on regression coefficients, which penalizes their absolute magnitudes. This constraint not only prevents overfitting but also performs embedded variable selection by shrinking some coefficients to exactly zero, thereby identifying the most influential predictors among high-dimensional features [32]. Similarly, Elastic Net combines the L_1 penalty of LASSO with the L_2 penalty of ridge regression, allowing it to handle correlated predictors more effectively and maintain stability in cases where multiple variables exhibit multicollinearity [33]. Random Forest (RF) is an ensemble method that builds a large number of decision trees, typically through averaging in regression tasks, which enhances predictive accuracy and reduces the risk of overfitting compared to single-tree methods [34]. Support Vector Regression (SVR) extends the principles of Support Vector Machines (SVM) to continuous outcomes. It constructs an optimal regression hyperplane that fits most observations within a defined margin, while kernel functions enable the modeling of nonlinear relationships between predictors and outcomes [35]. Extreme Gradient Boosting (XGBoost) is a powerful gradient boosting algorithm that iteratively builds an ensemble of weak learners by minimizing a regularized loss function. It incorporates second-order gradient information for optimization, making it computationally efficient and highly scalable for structured data [36]. Finally, Light Gradient Boosting Machine (LightGBM) represents an improved variant of gradient boosting. It uses histogram-based algorithms and leaf-wise tree growth to significantly accelerate training speed, reduce memory usage, and enhance scalability. Its design is particularly suitable for large, high-dimensional datasets [37]. Together, these six benchmark methods provide a strong comparative foundation against which the proposed hedonic models can be assessed in terms of both explanatory power and predictive performance.

D. Preliminary Test

Before conducting the comparative experiments, we first carried out a preliminary test to determine which of the two hedonic specifications introduced earlier—the linear model or the squared model—should be adopted as the baseline framework. The purpose of this step is twofold. First, it provides a consistent foundation for the subsequent variable selection procedures by fixing a single model structure. Second, it ensures fairness in benchmarking, as all alternative methods are evaluated using the same feature set derived from the chosen baseline model. In the following Table VIII, the results of the preliminary test are summarized.

As you can see from the table, on the training dataset, the squared model outperforms the linear specification across all measures, with higher R^2 and adjusted R^2, as well as lower AIC and BIC values, suggesting that the inclusion of quadratic terms substantially improves in-sample explanatory power. However, the validation results reveal a different pattern. While the linear model maintains relatively stable performance with R^2 and adjusted R^2, the squared model exhibits significant overfitting, achieving much lower out-of-sample R^2 and adjusted R^2. Furthermore, the AIC and BIC values of the squared model are also worse on the validation dataset.

These findings indicate that although the squared model can better capture nonlinearities in-sample, its generalization ability is weaker due to the increased complexity and high-dimensionality. Considering the principle of parsimony, and given that the linear model achieves comparable performance with only about half the number of variables, it is adopted as the baseline specification for subsequent variable selection and benchmarking experiments.

TABLE VIII. COMPARISON OF LINEAR AND SQUARED HEDONIC MODELS

Baseline Model	Train Set				Validation Set			
	R^2	Adj. R^2	AIC	BIC	R^2	Adj. R^2	AIC	BIC
Linear Model	0.7243	0.6972	−5677	−2751	0.717	0.6422	−7281	−4735
Squared Model	0.7979	0.7541	−6342	−523	0.558	0.4411	−5437	−374

E. Main Test

In this subsection, we report the results of the main computational experiments designed to evaluate the performance of the proposed hedonic models combined with variable selection procedures, compared against the six benchmark methods introduced earlier. The baseline model for variable selection is the linear specification described in Section V.D. First, the results of applying the proposed variable selection procedures and the six benchmark methods to the training set are summarized in Table IX.

Table IX summarizes the comparative performance on training set of the baseline only, baseline with three variable selection methods, two regularization methods, and six machine learning benchmark models. Across all evaluation metrics, the machine learning methods delivered the strongest performance. In particular, Random Forest achieved R^2 and adjusted R^2 values exceeding 0.9, while also attaining substantially lower AIC and BIC scores compared to the hedonic and regularization-based models. However, such large performance gaps observed on the training set raise concerns about potential overfitting. To further investigate this possibility, we next turn to the validation set results, which provide a more reliable benchmark for assessing the generalization performance of each method.

TABLE IX. COMPARATIVE RESULTS OF THE TESTED MODELS ON TRAINING SET

Method	# of vars.	R^2	Adj. R^2	AIC	BIC
Baseline	461	0.7243	0.6972	−5677	−2751
Forward Selection	136	0.6921	0.6835	−5749	−4856
Backward Selection	432	0.7235	0.6981	−5715	−2958
Stepwise Selection	223	0.7142	0.7008	−5946	−4486
LASSO	353	0.7192	0.6979	−5777	−3477
Elastic Net	363	0.7194	0.6975	−5760	−3394
Random Forest	461	0.9407	0.9346	−27522	−24518
SVR	461	0.8841	0.8724	−24176	−21172
XGBoost	461	0.8768	0.8643	−23869	−20864
LightGBM	461	0.7649	0.741	−20638	−17633

Table X presents the comparative results of all tested models on the validation set. The performance degradation of machine learning methods is particularly notable in the validation set. Their previously strong results on the training set appear to be a consequence of overfitting, as evidenced by the sharp decline in adjusted R^2. The two regularization methods also achieve reasonable results with reduced feature sets, but the proposed variable selection procedures consistently outperform them. Applying variable selection to the baseline model not only enhances performance but also improves interpretability. Among the proposed methods, stepwise selection achieves the highest adjusted R^2, while forward selection delivers the best performance in terms of AIC and BIC. Notably, forward selection accomplishes this using only 136 variables—substantially fewer than the baseline or other approaches—demonstrating that it is the most suitable method for valuing the Azuki collection. The following table illustrates how effectively the model obtained through forward selection has mitigated multicollinearity.

TABLE X. COMPARATIVE RESULTS OF THE TESTED MODELS ON VALIDATION SET

Method	# of vars.	R^2	Adj. R^2	AIC	BIC
Baseline	461	0.717	0.6422	−7280.6	−4734.8
Forward Selection	136	0.7024	0.6823	−7796.8	−7020.0
Backward Selection	432	0.7178	0.6466	−7318.7	−4863.6
Stepwise Selection	223	0.7176	0.6847	−7734.5	−6464.4
LASSO	353	0.7172	0.6613	−7471.5	−5464.3
Elastic Net	363	0.7169	0.6592	−7449.8	−5386.0
Random Forest	461	0.6162	0.5109	−6603.1	−3989.3
SVR	461	0.608	0.5006	−6558.3	−3944.5
XGBoost	461	0.6392	0.5403	−6736.0	−4122.1
LightGBM	461	0.6427	0.5447	−6756.8	−4142.9

Table XI presents the distribution of Variance Inflation Factor (VIF) values for the baseline model and the forward selection model. The results clearly show that multicollinearity has been substantially reduced after variable selection. In the baseline model, all 461 variables exhibit VIF values greater than 10, indicating severe multicollinearity. In contrast, the forward selection model retains 136 variables, among which 133 have VIF values below 5 and only one exceeds 10. This confirms that the forward selection procedure effectively eliminates redundant predictors. In the following figure, actual and predicted log (APR) values on the validation set using the hedonic model obtained through forward selection are demonstrated.

TABLE XI. VARIANCE INFLATION FACTOR (VIF) DISTRIBUTIONS BETWEEN BASELINE AND FORWARD SELECTION MODEL

Range of VIFs	Baseline	Forward Selection
VIF ≤ 5	0	133
5 < VIF ≤ 10	0	2
10 < VIF	461	1

In Fig. 3, the blue dots represent individual Azuki tokens, while the red dashed line indicates the 45-degree line where predicted values would perfectly match actual outcomes. The results show that most observations cluster around the diagonal, suggesting that the model achieves a generally good predictive fit. However, deviations become more pronounced for tokens with higher log (APR) values, reflecting the difficulty of fully capturing the extreme volatility associated with rare and highly priced traits.

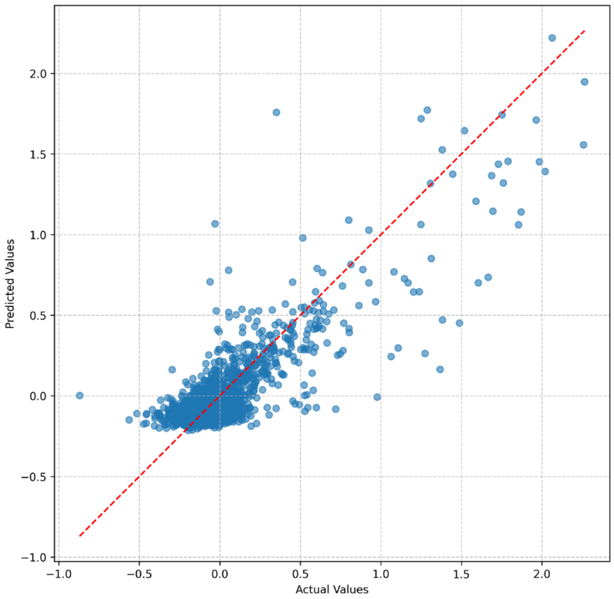

Fig. 3. Scatter plot of actual versus predicted values for the validation set using the forward selection model.

Fig. 4 illustrates the number of traits selected for each attribute through the forward selection procedure. For comparison, the total number of traits within each attribute is also shown, allowing the proportion of selected traits to be easily identified. The exact selection ratio is displayed to the right of each orange bar. As shown in the figure, "Type" and "Special" attributes have all their traits selected (100%), indicating their dominant importance in determining NFT value. The "Offhand" attribute, although only 40% of its traits were selected, contributes the largest number of traits overall, suggesting its strong explanatory power across multiple trait combinations. In contrast, "Hair" and "Clothing" exhibit very low selection ratios (3% and 7%, respectively), implying that despite their large number of traits, these attributes play a relatively minor role in valuation.

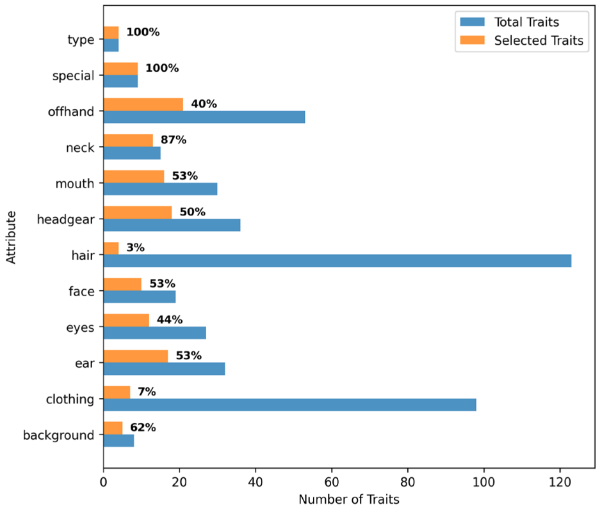

Fig. 4. Comparison of the number of total and selected traits across attributes in the Azuki dataset. The percentages indicate the proportion of traits selected within each attribute by the forward selection procedure.

VI. Conclusion

This study proposed and empirically validated hedonic valuation models for the Azuki NFT as a high-dimensional case study. We first conducted an extensive exploratory analysis of Azuki's metadata, documenting 12 attributes and 454 distinct trait values. Building on these insights, we specified NFT value via a volatility-robust dependent variable and constructed trait-level features with TF-IDF to reflect both presence and collection-wide scarcity. We then introduced two hedonic specifications (linear and squared) and three tailored variable-selection procedures (forward, backward, and stepwise). A preliminary comparison showed that the squared model overfit out of sample, whereas the linear model generalized better and was therefore adopted as the baseline. In the main experiments, applying variable selection to the linear baseline consistently improved performance and parsimony relative to both the unselected baseline and regularization benchmarks. Among the proposed procedures, stepwise selection achieved the highest adjusted R^2 on the validation set, while forward selection delivered the best AIC/BIC with only 136 variables, demonstrating a favorable accuracy-complexity trade-off. By contrast, machine-learning benchmarks attained very high training fit but experienced marked degradation on the validation set—clear evidence of overfitting in this setting. Overall, these results indicate that carefully crafted hedonic models, paired with principled variable selection, can provide competitive and interpretable NFT valuation with stronger generalization.

Several avenues remain open for future research. First, while this study focused on Azuki as a representative case, the proposed hedonic valuation framework can be extended and validated across other NFT collections. Such cross-collection analyses would help assess the generalizability and robustness of the framework. Second, in this study the intrinsic value of NFTs was extracted from attribute-trait names using TF-IDF, a widely adopted text vectorization technique. Future work could explore alternative feature engineering approaches, including more advanced text vectorization or word embedding methods, which may better capture semantic similarities across traits and thus improve explanatory power. Third, given that PFP NFTs inherently convey uniqueness through visual representation, future research could enhance the proposed framework by integrating image-derived features, thereby providing a more holistic understanding of NFT valuation dynamics. Finally, unlike prior studies on BAYC, the quadratic terms in the Azuki dataset did not yield meaningful gains and even resulted in overfitting. A comparative investigation across collections is needed to clarify under what structural conditions nonlinear terms contribute to valuation accuracy. Such work may also shed light on the interplay between attribute richness, market behavior, and model specification.

Appendix A: SQL Query for Extracting Azuki Transaction Data

The Azuki transaction dataset used in this study was obtained from Dune Analytics (dune.com). The following SQL query was used to extract all transaction records associated with the official Azuki smart contract deployed on the Ethereum blockchain. The contract address corresponds to the verified Azuki collection (0xED5AF388653567Af2F388E6224dC7C4b3241C544).

SQL Query for Extracting Azuki Transaction Data

```
SELECT *
FROM nft.trades
WHERE
  blockchain = 'ethereum'
  AND                    nft_contract_address                    =
0xED5AF388653567Af2F388E6224dC7C4b3241C544;
```

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

GL and HL developed the methods; HL and GL analyzed the data; HK and GL wrote the paper; GL and HK conducted the experiments; all authors had approved the final version.

Funding

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5A2A03083440).

References

[1] I. Alon, V. P. G. Bretas, and V. Katrih, "Predictors of NFT prices: An automated machine learning approach," *JGIM*, vol. 31, no. 1, pp. 1–18, Jan. 2023. doi: 10.4018/JGIM.317097

[2] K. Ko, T. Jeong, J. Woo, and J. W.-K. Hong, "Survey on blockchain-based non-fungible tokens: History, technologies, standards, and open challenges," *International Journal of Network Management*, vol. 34, no. 1, e2245, 2024. doi: 10.1002/nem.2245

[3] R. Peres, M. Schreier, D. A. Schweidel, and A. Sorescu, "Blockchain meets marketing: Opportunities, threats, and avenues for future research," *International Journal of Research in Marketing*, vol. 40, no. 1, pp. 1–11, Mar. 2023. doi: 10.1016/j.ijresmar.2022.08.001

[4] E. Howcroft. (Jan. 11, 2022). NFT sales hit $25 billion in 2021, but growth shows signs of slowing. *Reuters*. [Online]. Available: https://www.reuters.com/markets/europe/nft-sales-hit-25-billion-2021-growth-shows-signs-slowing-2022-01-10/

[5] NFT-Worldwide | Statista Market Forecast. Statista. [Online]. Available: http://frontend.xmo.prod.aws.statista.com/outlook/fmo/digital-assets/nft/worldwide

[6] K. Oleaga. NFTs Are Down Bad, But Not "Worthless". NFT now. [Online]. Available: https://nftnow.com/news/nfts-are-down-bad-but-not-worthless/

[7] H. Taherdoost, "Non-Fungible Tokens (NFT): A systematic review," *Information*, vol. 14, no. 1, Jan. 2023. doi: 10.3390/info14010026

[8] M. Dowling, "Is non-fungible token pricing driven by cryptocurrencies?" *Finance Research Letters*, vol. 44, 102097, Jan. 2022. doi: 10.1016/j.frl.2021.102097

[9] M. Dowling, "Fertile LAND: Pricing non-fungible tokens," *Finance Research Letters*, vol. 44, 102096, Jan. 2022. doi: 10.1016/j.frl.2021.102096

[10] R. Kräussl and A. Tugnetti, "Non-Fungible Tokens (NFTs): A review of pricing determinants, applications and opportunities," *Journal of Economic Surveys*, vol. 38, no. 2, pp. 555–574, 2024. doi: 10.1111/joes.12597

[11] CryptoPunks. [Online]. Available: https://www.cryptopunks.app/

[12] Bored Ape Yacht Club—Welcome to the BAYC Clubhouse. [Online]. Available: https://boredapeyachtclub.com/

[13] Top NFTs | OpenSea. [Online]. Available: https://opensea.io/stats

[14] D.-R. Kong and T.-C. Lin, "Alternative investments in the fintech era: The risk and return of Non-Fungible Token (NFT)," *Social Science Research Network*, 2021. doi: 10.2139/ssrn.3914085

[15] L. Schaar and S. Kampakis, "Non-fungible tokens as an alternative investment: Evidence from CryptoPunks," *The JBBA*, Jan. 2022. doi: 10.31585/jbba-5-1-(2)2022

[16] H. Lee, G.-C. Lee, and H.-Y. Koo, "Exploring the relationship between rarity and price of profile picture NFT: A formal concept analysis on the BAYC NFT collection," *Blockchain: Research and Applications*, vol. 5, no. 2, 100191, Jun. 2024. doi: 10.1016/j.bcra.2024.100191

[17] G.-C. Lee, H.-Y. Koo, and H. Lee, "Quadratic regression models for profile picture NFT valuation," *IEEE Access*, vol. 13, pp. 114029–114037, 2025. doi: 10.1109/ACCESS.2025.3584222

[18] H.-Y. Koo, H. Lee, and G.-C. Lee, "Investor impact in the NFT market: A comparison between whales and small investor," *Journal of the Korean Operations Research and Management Science Society*, vol. 50, no. 3, pp. 17–29, 2025. doi: 10.7737/JKORMS.2025.50.3.017

[19] A. Mekacher *et al.*, "Heterogeneous rarity patterns drive price dynamics in NFT collections," *Sci. Rep.*, vol. 12, no. 1, 13890, Aug. 2022. doi: 10.1038/s41598-022-17922-5

[20] W. Xiong, Y. Wang, W. Li, Y. Zhang, J. Zhang, and H. Guo, "An advanced pricing mechanism for Non-Fungible Tokens (NFTs) based on rarity and market dynamics," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 6, pp. 7671–7684, Feb. 2024. doi: 10.1109/TCSS.2024.3430846

[21] B. Seyhan and E. Sefer, "NFT primary sale price and secondary sale prediction via deep learning," in *Proc. the Fourth ACM International Conference on AI in Finance*, 2023, pp. 116–123. doi: 10.1145/3604237.3626896

[22] M. Pala and E. Sefer, "NFT price and sales characteristics prediction by transfer learning of visual attributes," *The Journal of Finance and Data Science*, vol. 10, 100148, Dec. 2024. doi: 10.1016/j.jfds.2024.100148

[23] Azuki. [Online]. Available: https://www.azuki.com/

[24] CryptoPunk and Azuki Lead Pack of Top NFT Sales in June 2025. [Online]. Available: https://blockchainreporter.net/cryptopunk-and-azuki-lead-pack-of-top-nft-sales-in-june-2025/

[25] NFTmetria. The most expensive sales of Azuki NFTs: An anime dream for investors. [Online]. Available: https://nftmetria.com/most-expensive-nft/sales-azuki/

[26] G. Fridgen, R. Kräussl, O. Papageorgiou, and A. Tugnetti, "Pricing dynamics and herding behaviour of NFTs," *European Financial Management*, vol. 31, no. 2, pp. 670–710, 2025. doi: 10.1111/eufm.12506

[27] M. Goldberg, P. Kugler, and F. Schär, "Land valuation in the metaverse: Location matters," *J. Econ. Geogr.*, vol. 24, no. 5, pp. 729–758, Sep. 2024. doi: 10.1093/jeg/lbae027

[28] F. Horky, C. Rachel, and J. Fidrmuc, "Price determinants of non-fungible tokens in the digital art market," *Finance Research Letters*, vol. 48, 103007, Aug. 2022. doi: 10.1016/j.frl.2022.103007

[29] M. Jiang and Y. Xia, "What drives the volatility of Non-Fungible Tokens (NFTs): Macroeconomic fundamentals or investor attention?" *Applied Economics Letters*, vol. 31, no. 16, pp. 1439–1448, Sep. 2024. doi: 10.1080/13504851.2023.2187034

[30] J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, vol. 20, no. 1, pp. 101–148, 2010.

[31] K. P. Burnham and D. R. Anderson, "Multimodel inference: understanding AIC and BIC in model selection," *Sociological Methods & Research*, vol. 33, no. 2, pp. 261–304, Nov. 2004. doi: 10.1177/0049124104268644

[32] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Royal Statistical Society. Journal. Series B: Methodological,* vol. 58, no. 1, pp. 267–288, Jan. 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x

[33] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005. doi: 10.1111/j.1467-9868.2005.00503.x

[34] L. Breiman, "Random forests," *Mach. Learn*., vol. 45, no. 1, pp. 5–32, 2001.

[35] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Advances in Neural Information Processing Systems*, vol. 9, pp. 155–161, 1997.

[36] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.

[37] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, 30, 2017.