Prediction Soaring Price by Decision Tree Dealing with Imbalanced Data in Trading Card Game Market

Takeru Shishido[1], Ayahiko Niimi[2]
[1]*Graduate School of Systems Information Science*
[2]*Faculty of Systems Information Science*
Future University Hakodate, Japan

Abstract

The market for used trading card games (TCG) has expanded rapidly in recent years. In the TCG market, the demand for used TCG cards tends to rise in tandem with the release of new products, driving up the prices. Unlike the regularly used market, the used TCG market has some challenges with price increases. We investigated the used market for "Magic: The Gathering," one of the TCGs, to forecast price increases. However, in terms of prediction, there are very few soaring products, almost all of which are classified as "not soaring" and they are of poor precision. Preliminary experiments also revealed that the recall and F1 values were extremely low, resulting in poor prediction performance. Dealing with imbalanced data using three handling techniques, we attempted to increase the recall in this research's predictions of TCG's skyrocketing price. As a result, the handling method improved recall of the methods' ability to predict price increases. Furthermore, we attempted to compare the recall of three methods for dealing with imbalanced data (oversampling, undersampling, and cost-aware learning approach). The outcome demonstrates that the cost-aware learning approach and random sampling were particularly effective handling methods.

Keywords: Data Mining, Imbalanced Data, Decision Tree, Prediction Soaring Price, Trading Card Game

1. Introduction

In recent years, the used sales market for trading card games (TCG) has grown dramatically. TCGs are defined as "games in which players collect cards by trading, buying, etc., and freely select and play cards among them according to the rules" in this research. eBay has acquired "TCGPLAYER," a major TCG ec-selling company [1], and has begun the TCG authenticity guarantee service [2]. The company has garnered a lot of attention by focusing on used TCG sales. There are some differences between the used TCG market, which is the study's target, and the used market of regular products.

First, products are randomly sealed in the first distribution. The products are typically sold by the pack and contain a few cards from a pre-released list in a pack. When sold in the used market, however, each card in a pack is sold separately. Even though the cards are in the same packs, the prices of each card are different on the used market. Furthermore, ascertaining the exact number of cards in the market is difficult due to the randomness of products. As a result, predicting the price difficult given the original price and the quantity of primary distribution.

Second, the cards have several factors that cause soaring. Because the cards are a game item, their performance in the game increases demand for the item. On the other hand, the rarity of the card itself, such as its sales year or illustrator, increases the demand for collectible cards. Therefore, it is necessary to forecast that considers both rarity and performance in the game.

Finally, the price of a card can also rise due to external factors other than the card itself. For example, after the release of a product, prices skyrocket dur to its in-game tournaments. In other cases, price drops are sharply by bans imposed because of rule changes. Due to advanced information, some cards with good synergy with pre-release products are purchased in the futures market. There have been instances where influencers' statements drew attention and caused soaring.

Given the foregoing, the TCG used sales market that this study address is more complex than that of regular commodities. Therefore, when one does not know the game, it is difficult to predict, and pricing is based on intuition and experience.

Finally, we want to use machine learning to attempt to predict the soar of the used TCG market, by focusing on product characteristics.

2. Problem of Imbalanced Data

When new products are released in the TCG market, the demand for old products increases, increasing costs. For example, a card that sold for 2 USD the week before is now in high demand because a new card that matches it was just released. As a result, the price soars to the following week nearly 20 USD. The price increase in the TCG market has been predicted by some research [3], [4].

 | DOI: 10.20533/ijds.2040.2570.2023.0229

Our research sought to predict soaring products using machine learning.

However, the issue is that very few items see price increase. Table 1 details the products that soared in the 2 weeks before and after, and following each of the four launch periods from 2021-02-05 to 2021-0723. The data is skewed, with only about 2% of products soaring at any given time.

Table 1. Breakdown of the number of cards and labels covered in each period

	2021-02-05	2021-04-23	2021-06-11	2021-07-23
not soaring	40,797	42,211	43,520	43,679
soaring	915	546	509	865
Total	41,712	42,757	44,029	44,544

Consequently, most cards' prices in the targeted used market do not increase. In general, data is biased toward a specific class, which is what imbalanced data is. Imbalanced data harms the classification problem. The data is imbalanced because there are so few "soaring" cards when learning as a binary classification problem. As a result, in the case of standard decision tree learning, the majority of cards are classified as "not soaring," resulting in a poorly performing decision tree.

2.2. Purpose

Based on the preceding discussion, we used a decision tree to predict whether the price of each item will increase in this research, which focuses on the used market for the TCG title "Magic: The Gathering" [5]. We concentrate on the fact that most of the cards do not soar, and that the data is imbalanced. To classify the existence or absence of price increases, we used methods that have been proven effective in dealing with situations involving imbalanced data. We discovered that the methods listed below were appropriate for the problem. We demonstrated the efficacy of this method in predicting the TCGs price increases.

Furthermore, we aimed to improve recall to correctly classify the actual price. We compared approaches with and without imbalanced data handling methods (oversampling, undersampling, cost-aware learning approach) and discussed ways to predict "soaring" products. The goal of this research was to improve recall of soaring prices, and we discussed how to interpret the data while analyzing it with decision trees.

3. Related work

Research on TCG predictions - Many researchers are predicting the used TCG market. Matthew et al. [3] used logistic regression and Support Vector Machine to predict soaring prices in the used TCG market using 13,608 card data from 2012 to 2014 and tournament card data. Dustin et al. [4] use n-grams from text data to analyze their performance in the card game and predict the price of 14,352 cards up to 2015.

The difference between this research and others is that it focuses on the presence or absence of price soaring to predict. This research focuses on products sold in 2022, with 41,236 card data sets, which is more than previous studies. Furthermore, the imbalanced data is a significant difference between this study and previous studies.

3.1. Research on dealing with imbalanced data

In terms of dealing with imbalanced data, there are several approaches. Sampling approaches and cost-aware learning approaches are well-known methodologies for dealing with imbalanced data [6] generally. By adjusting the number of data, the sampling approach is a method for dealing with imbalanced data. There are two types of methods: oversampling, which increases data, and undersampling, which decreases data.

Oversampling is a technique for increasing minority class data. Making random copies of identical data is a common example, but it is prone to overfitting. SMOTE (Synthetic Minority Oversampling Technique) [7] is used to expand the data by combining neighboring data to address this issue.

Undersampling is a technique for the majority of class data. A typical example is the use of random sampling to reduce the data volume. Undersampling works well in many situations, according to Wallace et al. [8]. However, it may delete even critical data.

The problem of imbalanced data is corrected in the cost-aware approach by weighing the objective function during the learning phase. Xu-Ying et al. showed that class imbalance influences cost-sensitive classifiers [9]. This method learns to prioritize minority classes by calculating a loss function based on the amount of data. One issue with this method is that costs are not always uniform, and the optimal cost for each dataset must be determined.

Sampling and cost-aware approaches were used as the main experimental approaches in this research.

Another method is the ensemble approach, which combines multiple learners to improve learning accuracy. To perform combinatorial learning, the ensemble approach employs boosting, bagging, and stacking. In this research, the interpretability of soaring factor analysis is prioritized, and the ensemble approach is not addressed. Before this research, preliminary experiments on imbalanced data analysis with decision trees were conducted. This experiment is detailed below.

4. Preliminary experiment

Before addressing the imbalanced data, we present a pre-experiment for this research in which no imbalanced data handling was performed. In this experiment, we will examine learning accuracy when learning is performed without responding to imbalanced data, and the results will clarify the importance of responding to imbalanced data.

4.1. Method

The experiment used a data set of 44,544 cards for which price data was available as of August 9, 2021. This data set excludes cards that cannot be used according to the game's official rules, as well as special cards, and cards with special forms. Of these data, card information was derived from "MTG JSON" [10]. We also examined the attributes in the "MTG JSON" data and chose the categorical and numerical attributes to be included in the dataset.

We used price data from the "MTG JSON" database to generate used card sales data from two e-commerce sites. This price data was obtained for a total of 200 days excluding days when some data could not be obtained, from January 8, 2021, to August 9, 2021. This was done for 48,249 cards, totaling 9,649,800 used price data items. If the prices of a card differed between sites, the average value was used to determine the value of a card.

The cards were labeled as "Soaring" for price changes if the percentage of price soaring was 1.5 times or greater than the minimum price. However, preprocessing errors cause the price to soar for low-priced cards, and predicting are easily influenced by results in the ultra-high price range. Therefore, we limited the labeling to cards with a minimum price of 1 USD to 150 USD.

The dataset used for training and testing is broken down in Table 2. We labeled the data based on prices for the two weeks before and after the release dates, yielding in the following amount of data for the experiment.

Table 2. Breakdown of the number of cards and labels in each period

	2021-02-05	2021-04-23	2021-06-11	2021-07-23
not soaring in training	30,611	31,666	32,640	32,761
soaring in training	673	401	381	647
not soaring in test	10,186	10,545	10,880	10,918
soaring in test	242	145	128	218
Total	41,712	42,757	44,029	44,544

Finally, we trained a decision tree to predict whether or not "Soaring" would occur. The dependent variable in learning was the card attributes, and the objective variable was the presence or absence of soaring. Furthermore, the prediction was conducted without dealing with imbalanced data. We assessed the performance of these decision trees by using accuracy, recall, precision, and F1 values in each term. Based on the findings, we assessed whether it is possible to forecast soaring prices without dealing with imbalanced data.

4.2. Result

Table 3 shows the result of the decision trees in each period and their evaluation value.

Table 3. The evaluation value in each period

	2021-02-05	2021-04-23	2021-06-11	2021-07-23
Precision	1	1	1	1
Accuracy	0.981	0.986	0.988	0.980
Recall	0.284	0.017	0	0
F1	0.413	0.033	0	0

The decision tree performed exceptionally well for each term in Table 3, with a precision of 1 and an accuracy of 0.98 for all institutions. However, the values of Recall and F1 were particularly low, implying that the decision trees constructed were extremely inaccurate. The low value of recall indicates that many cards are classified as "not soaring" in the constructed decision tree. Therefore, the soaring prediction failed, implying that an imbalanced data response was required. From the finding of the preceding preliminary experiment, we provided the proposed dealing with imbalanced data in this research.

5. Method

Our research suggested using a decision tree-based model to predict card price increases in the presence of imbalanced data. This is because interpretability must be considered to evaluate the factors causing card price increases, which is the goal of this research. In this research, we will investigate how decision trees respond to imbalanced data. Following the definition of accuracy in this research, we sought to improve recall. To improve the performance of the price increase prediction, it is critical to enhancing the percentage of correct responses to commodities whose prices have increased.

We also applied three of the representative methods to imbalanced three methods, namely, SMOTE oversampling, random undersampling, and the cost-aware approach, in which weights are used to make and compare predictions.

Furthermore, the 2 weeks preceding and following the launch of new products are included in the concept of "soaring" in this research. Because many of these products have a history of skyrocketing prices both before and after the introduction of a new products, they were specifically chosen as the focus of the price increase prediction.

5.1. Dataset and experiments

Creation of card datasets - The experiment used a dataset of 43,535 cards for which price data was available by 2022-06-18. The dataset excludes cards that cannot be used according to the game's official rules, special cards, and cards with special forms such as foil.

The "MTG JSON" [10], a compilation of all card information for "Magic: The Gathering," where this dataset's card information is kept. Data from "MTG JSON" was retrieved to create the dataset. In addition, we analyzed the attributes in the "MTG JSON" data and chose a few of them. Furthermore, only category and numerical attributes. The Figure 1 shows an example of attributes with the actual card, and Table 4 lists the attributes that were used.

Figure 1. Example between the attributes and the card [12]

Categorical attributes were dummy coded among these attributes. There were numerical qualities such as "X" and "1+*" that were described and characterized as changing depending on the situation. We added new attributes (e.g., power X) to some of the numerical attributes to indicate that they are variable. Furthermore, each attribute was 01-normalized to align the numerical attribute's range of values.

Creation of price dataset **-** Similar to the card dataset, the price data was derived from the "MTG JSON". The price information was obtained from the "card market" [11], an e-commerce site. Between 2021-09-10 and 2022-06-18, price data could be obtained for a total of 250 days, excluding days when some data could not be obtained. We obtained 10,883,750 price data for 43,535 cards.

Table 4. Attributes of the dataset used in this research

attribute	Attribute Explanation
colors	The color of the card (R:Red, W:White, B:Black, U:Blue, G:Green)
textColor	The color in the card's text except for colors (same as above)
convertedManaCost hasAlternativeDeckLimit	Cost converted to a value for use in game (0Whether or not there is a limit on the number of cards~16)
layout	The type of card layout (normal, aftermath, split, flip, leveler, saga, transform, adventure, modal dfc, class, meld)
loyalty	The value of loyalty
power	Power (-1~15)
toughness	Toughness (-1~17)
types	Card types in-game (Enchantment, Creature, Land, Instant, Sorcery, Artifact, Planeswalker, Tribal)
hasContentWarning	Whether or not to ban for having sensitive content
hasFoil	Existence of a foil version
hasNonFoil	Existence of a non-foil version
isAlternative	Whether it is a special version
isFullArt	Whether it is a full art version
isPromo	Whether it is a promotional card
isReprint	Whether or not same name card was reprinted
isReserved	Whether it is on the list of prohibited reprints
isStarter	Whether or not the product is in a starter deck
isTextless	Whether or not the card does not have a text box
year	Year of the release of the enclosed product (1993~2021)
printings_num	Number of times the same name card is printed (1~192)
borderColor	The color of the card border (black, white, borderless)
frameVersion	The type of card layout (1993, 1997, 2003, 2015, future)
rarity	The inclusion rate of cards in products (common, uncommon, rare, mythic, special)
setCode	Products with enclosed cards (287 types)

5.2. Labeling of soaring

The labels indicate whether or not a price increase occurred in the price datasets. We used the price rise label for 2 weeks before and 2 weeks after the product's release data because we are interested in the price increase of currently available products before and after their release date. To compare predicting performance across multiple periods, we labeled the prices of four products released during the period (products released on 2021-09-24, 2021-11-19, 2022-02-18, and 2021-04-29) based on the prices 2 weeks before and after the release dates, for total of 4 weeks. The ratio of change in the price of each card was calculated using Equation (1).

$$\begin{aligned} &\mathit{The\ ratio\ of\ change} \\ &\quad = \frac{\mathit{Value}}{\mathit{Minimum\ value}} \end{aligned} \tag{1}$$

If the ratio was 1.5 times or greater than the price, the price of a card was labeled as "soaring". The labeling was restricted to cards with a maximum value of 150 USD or less because of the results in the ultra-high price of the category has a major impact on the results.

Table 5 displays the size of the dataset used for training. In this research, we used the dataset labeled by price.

5.3. Dealing with imbalanced data

In this research, we applied each of the three techniques to the three types of imbalanced data (oversampling, undersampling, and cost-aware approach).

SMOTE was used to oversampling of the training data for the minority class in the "soaring" class in the Python library "imbalanced-learn" [13], and the training data in the minority class was increased. The "soaring" class's data was expanded, and it was discovered that it has the same number as the "not soaring" class. Random sampling was used to achieve undersampling, training data was reduced to aid in learning, and the number of "not soaring" class data was equal to that of the "soaring" class. The number of data is shown in Table 5 for both classes, for the cost-aware approach; however, we set the weights of each class as follows. The weights of each class were calculated using Equation (2).

Table 5. Breakdown of the number of cards and labels in each period

	2021-09-24	2021-11-19	2022-02-18	2022-04-29
not soaring in training	31,336	18,229	33,972	32,365
soaring in training	3,182	16,410	663	2,272
not soaring in test	7,834	4,557	8,493	8,091
soaring in test	796	4,103	166	568
total	43,148	43,299	43,294	43,296

$$w = \frac{n_s}{n_c \times n_i} \tag{2}$$

w weights
n_s number of total samples
n_c number of samples in each class
n_l number of occurrences of the label

Using the Decision Tree Classifier in “scikit-learn” which is a Python library, a decision tree was formed after applying each method [14].

Finally, we discussed whether the methods could be used to forecast price increases when comparing three different approaches side by side and without handling.

6. Results

Table 6 displays the recall rates for each correspondence method for decision trees for each data period (without handling, oversampling, undersampling, and cost-aware approach). Figures 2, 3, 4, and 5 show the decision trees for each approach in 2022-04-29, where a distinct performance difference was observed.

Table 6. Recall each approach in each period

	2021-09-24	2021-11-19	2022-02-18	2022-04-29
without handling	0.257	0.668	0	0
oversampling	0.548	0.676	0.295	0.603
undersampling	0.735	0.704	0.551	0.723
cost-aware	0.671	0.690	0.545	0.726

7. Discussion

The recall values shown in Table 6 were always higher than those obtained without imbalanced data correspondence. The outcome suggested that an attempt was made to adjust the imbalanced data to predict soaring prices. Therefore, research objective was met.

When the effectiveness of each method was compared, it was discovered that undersampling produced the best results for the majority of the periods. Undersampling performance was comparable to that of the cost-aware approach. Figures 4 and 5 illustrate this conclusion by showing how “soaring” can be classified in the upper layer using both approaches as opposed to the case without handling. As a result, it can be said that both methods performed admirably for the data in this research.

Oversampling outperformed no handling, but it fell short of the performance of undersampling and cost-aware approach. Increasing the minority class data from SMOTE could be to blame. After increasing the data, we examined the data attributes. The variety did not markedly increase because the attributes were of the integer type with a maximum of about a dozen and a bool type. Thus, it is believed that the process of producing intermediate data from neighborhood data was ineffective.

During the time of 2021-11-19, the number of data “soaring” and “not soaring” did not become imbalanced, and the method did not significantly change the prediction. The recall of the predictions was consistent. This could be due to insufficient data both before and after data collection. When compared to other times, the time frame for labeling was extended. As a result, the data was determined to be no longer imbalanced because the number of “soaring” data equaled the number of “not soaring” data. Because of this, the data are thought to be no longer imbalanced for this reason.

Though methods dealing with imbalanced data may be claimed to be meaningful for predicting based on improvements in recall in other periods with imbalanced data.

According to the findings, undersampling and a cost-aware approach are the most appropriate methods for dealing with imbalanced data. Random sampling, on the other hand, is used in the research's undersampling. Random sampling will omit critical information. It may not function properly depending on the data type. Consequently, cost-aware approach is the most adaptable method.

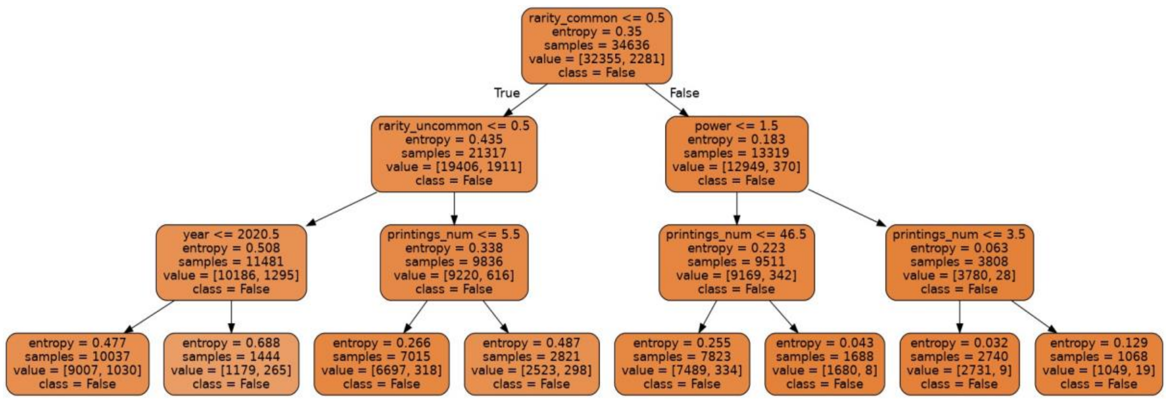

Figure 2. Decision tree without handling (depth = 4)

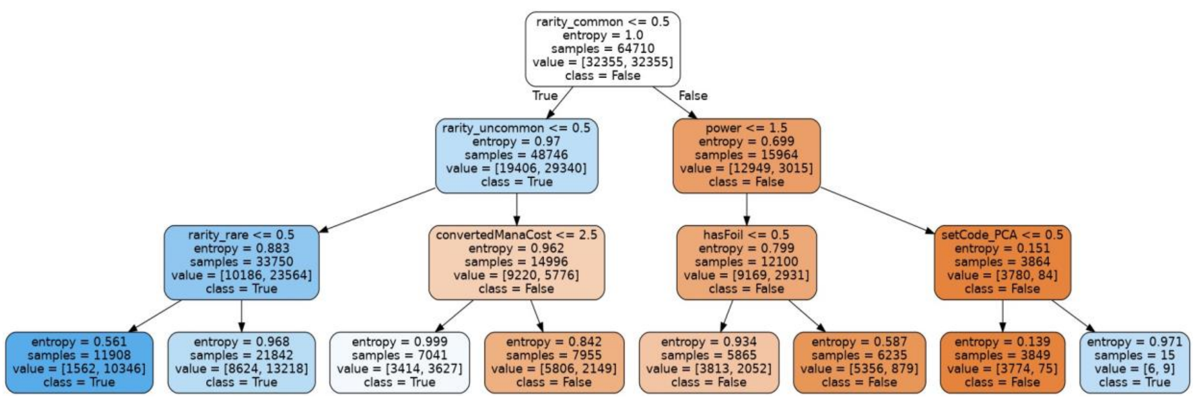

Figure 3. Decision tree with oversampling (depth = 4)

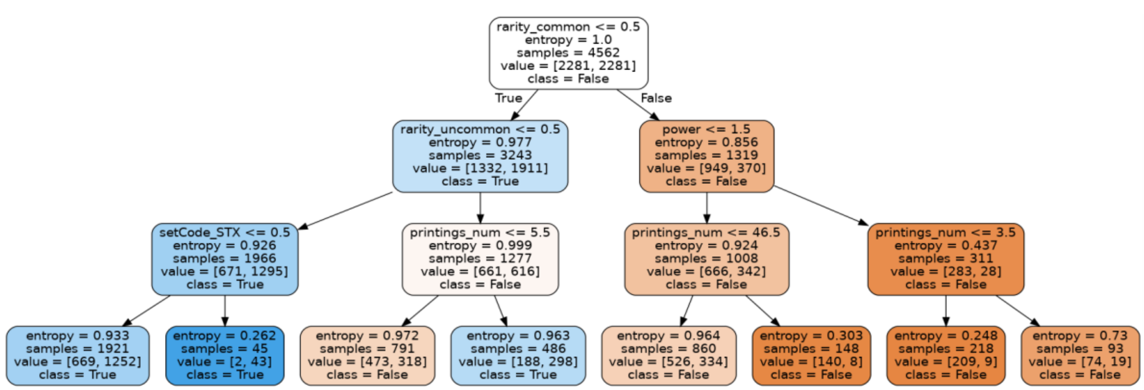

Figure 4. Decision tree undersampling (depth = 4)

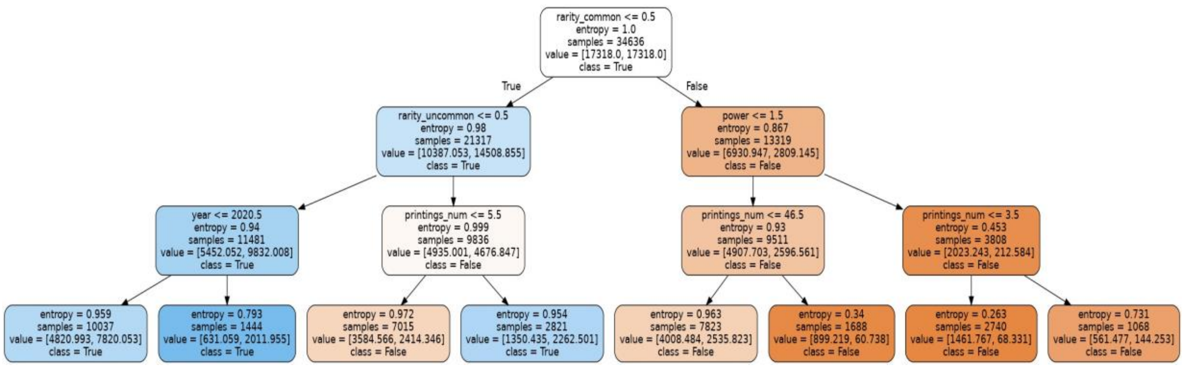

Figure 5. Decision tree with cost-aware approach (depth = 4)

8. Conclusion

In this research, we addressed appropriate methods for the problem of imbalanced data to improve the recall and prediction of the used TCG market. The results showed that each method increased recall and predicted correctly in response to the imbalanced data.

The various methods for dealing with imbalanced data were compared. To improve prediction accuracy, undersampling, and a cost-aware approach were used as effective learning methods. These findings indicate that this research goal was met and that the method for dealing with imbalanced data is effective in predicting the used TCG market's price increases.

In future research, undersampling, and cost-aware approach should be used as well as others that were not used. Although random sampling was used, other methods could also be used. In terms of cost-aware approach, comparisons involving multiple expenses may be considered to determine appropriate cost settings. We will provide a used TCG market predicting system. The used TCG sales market has a very broad price range, with a skewed population in each price range. We are thinking about applying an imbalance response to each price range and using a different model for specific price ranges. To address imbalanced handling, an ensemble approach can be used.

We plan to continue our research, experimentation, and conversations, including those elucidated in this research, even though there are still issues to be resolved in terms of predicting used TCG market price increases. The analytical methods used in this research have the potential to be applied to other fields of used markets, which we intend to investigate.

9. References

[1] Ebay Inc (n.d.). eBay Launches Authentication for Trading Cards. https://www.ebayinc.com/stories/news/ebay-launches-authentication-for-trading-cards/ (Access Date: 24 October 20 22).

[2] Ebay Inc (n.d.). eBay has Entered into an Agreement to Acquire TCGplayer. https://www.ebayinc.com/stories/news/ebay-has-entered-into-an-agreement-to-acquire-tcgplayer/ (Acc -ess Date: 24 October 2022).

[3] Matthew, P., Joseph, P., and Jesse, Z. (2014). Prediction of Price Increase for Magic: The Gathering Cards. Stanford University CS229 Projects, pp. 1-5.

[4] Dustin, F., Benjamin, P., and Neil, S. (2015). Predicting the strength of Magic: The Gathering cards from card mechanics. Stanford University CS229 Projects, pp.1-3.

[5] Magic Wizards. (n.d.). Magic: The Gathering. https://magic.wizards.com/en (Access Date: 24 October 2022).

[6] Rushi, L., and Snehalata, D. (2013). Class Imbalance Problem in Data Mining Review. ArXiv e-prints, pp. 1–6.

[7] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2022). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, vol. 16, pp. 321–357.

[8] Wallace, B., Small, K., Brodley, C., and Trikalinos, T. (2011). Class imbalance, redux", 2011, pp. 754–763.

[9] Xu-ying and Zhi-hua, Z. (2006). The Influence of Class Imbalance on Cost-Sensitive Learning: An Empirical research. Sixth International Conference on Data Mining (ICDM'06). pp. 970-974.

[10] Mtgjson. (n.d.). MTG JSON. https://mtgjson.com/ (Access Date: 24 October 2022).

[11] Cardmarket. (n.d.). Cardmarket. https://www.cardmarket.com/en. (Access Date: 24 October 2022).

[12] Gatherer Wizards (n.d.). Shivan Dragon. https://gatherer.wizards.com/Pages/Card/Details.aspx?multiverseid=469888 gatherer.wizards (Access Date: 24 October 2022).

[13] Imbalanced Learn (n.d.). imbalanced-learn, https://imbalanced-learn.org (Access Date: 24 October 2022).

[14] Scikit-Learn. (n.d.). scikit-learn. https://scikit-learn.org/stable/ (Access Date: 24 October 2022).