Rodrigo Rodriguez

06/22/2024

Applied Data Science

DSC680 (SUMMER)

Milestone2: Project Whitepaper Draft:

Analyzing TCG Box Sales to Predict Future Sale Prices

**Business Problem**

TCG (Trading Card Game) Booster Boxes, containing randomly selected individual trading cards, are primary assets in various TCG marketplaces. These marketplaces function similarly to stocks in a stock market. Each Booster Box is part of a unique "Set," available for about a year before being phased out, creating scarcity that drives value. The challenge for TCG card shop owners and collectors is to understand how the price fluctuations of these booster boxes will ultimately impact their business. Card shops that sell TCG products will buy booster boxes from a distributor at wholesale price and sell the boxes at MSRP. However, because of the TCG marketplaces and the price fluctuations, card shop owners may sell their products at a loss below MSRP.

Factors such as card rarity, set size, and other attributes influence the prices of these Booster Boxes over time. This project aims to analyze TCG Booster Box sales data along with unique attributes related to each booster box to predict future sale values. By developing predictive models, we seek to provide insights that can help TCG card shop owners optimize pricing strategies, manage inventory more effectively, and make informed purchasing decisions. For collectors, these insights can guide investment decisions and enhance their ability to anticipate market trends. Ultimately, the goal is to improve market efficiency, benefiting both sellers and buyers in the TCG marketplace.

**Background and History**

Trading Card Games (TCGs) have a rich history dating back to the early 20th century with the introduction of collectible cards in products like cigarette packs and gum. However, the modern era of TCGs began in the 1990s with the release of games like Magic: The Gathering and Pokémon. These games introduced the concept of booster packs, which contain a random assortment of cards, including common, uncommon, and rare cards. Booster Boxes, which house multiple booster packs, became the primary product for both players and collectors due to their potential to yield valuable rare cards.

Several attributes or factors influence the value of Booster Boxes:

1. Rarity Count: The presence of rare and highly sought-after cards can significantly increase the value of a Booster Box.
2. Set Size: Larger sets with more cards generally offer more variety and can affect the perceived value of the Booster Boxes.
3. Scarcity and Demand: As sets are phased out and become unavailable, their limited supply increases their value in the secondary market.

*List (1)*

TCGplayer.com, an online marketplace established in 2008, has become a central hub for buying and selling trading cards. It provides a platform for individual sellers and small businesses to reach a wide audience of collectors and players. The publicly available sales data from TCGplayer offers a wealth of information that can be analyzed to understand market trends and price dynamics.

Historically, card shop owners have relied on experience and intuition to set prices beyond MSRP and make purchasing decisions. However, the advent of data analytics provides an opportunity to bring a more scientific approach to these decisions. By leveraging sales data and predictive modeling, shop owners can gain deeper insights into market behavior, optimize their inventory management, and improve profitability.

The TCG marketplace has some similarities to the stock marketplace, albeit at a much smaller scale. Statistical approaches applied for stock market forecasting should fit well for the purposes of forecasting TCG booster box prices. Most previous studies have applied statistical time-series methodologies, such as ARCH, ARMA, and ARIMA (Kumbure et al., 2022). Other researchers have used ML techniques such as Random Forest and Bayesian networks. However, because most of the mentioned techniques have their own merits and limitations, some researchers tended to enhance the forecasting accuracy of those methods by combining several methods (Kumbure et al., 2022).

**Data Explanation**

The benefit of the TCG marketplace is that it is niche compared to other marketplaces such as stocks, commodities, or real estate. Macroeconomic variables that considerably influence these markets do not have as significant an influence on the TCG market. For this reason, we will exclude economic variables and instead focus on attribute variables. In general, attribute variables are consistent for all TCG booster boxes with unique values for each TCG Set. Examples of these attributes are shown above in List (1).

The dataset used for this research project is a combination of attribute variables and sale data. Attribute data was obtained from Bulbagarden, an online source for TCG product information. Historical Sale data was obtained from TCGplayer. Figure (1) shows a list of variables used for this dataset along with their data types.

A screenshot of a computer

Description automatically generated

*Figure (1)*

**Data Prep/Cleaning**

The dataset primarily uses daily sales records for the most recent four months. Daily transactions have fluctuations that may act as noise for the model. Aggregating the data to weekly will reduce some noise and retain a mean sale price for the week. Aside from the weekly aggregation transformation, the only other cleaning performed was changing datatypes to category for the ‘Generation’ variable and datetime for the date-related variables.

**Methods**

The initial data exploration phase uncovered two key pieces of insight. First, we can see in Figure (2) that there are 16 different Sets within the sales data, and of these sets, older sets make up a smaller percentage of total sales. This information is important to understand the sales distribution count by Set. TCGplayer only provides the most recent four months of sales, so obtaining older sales data for past Sets is challenging. For future use, obtaining a more robust sales dataset for older Sets is recommended.

A pie chart with numbers and numbers

Description automatically generated

*Figure (2)*

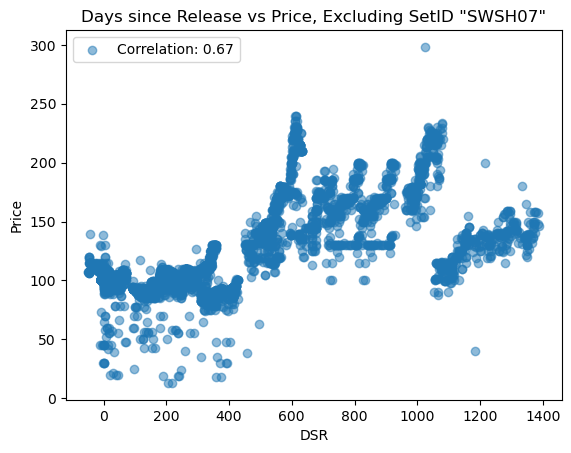
The second piece of insight relates to anomalies in the dataset. When observing the distribution of Price related to DSR (Days Since Release, used to determine the age of the Set), a moderate positive linear relationship exists between the two variables, as seen in Figure (3). This relationship shows that the price also tends to increase as the number of days since release increases. Figure (3) shows an anomaly that causes the correlation to drop to 0.5. However, after further analysis, the source of the anomaly was determined to be a specific SetID, which behaved differently than all others. Figure (4) reveals this SetID to be "SHSW07". After filtering out this SetID from the correlation plot, we see a strong correlation of 0.67 between Price and DSR, as seen in Figure (5).

A graph with numbers and a line of boxes

Description automatically generated with medium confidenceA graph of blue dots

Description automatically generated

*Figure (3) Figure (4)*



*Figure (5)*

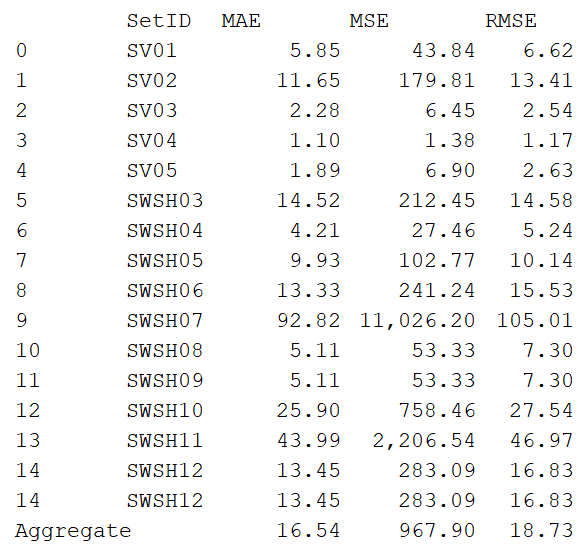
With the discovery of anomalies, I decided to retain the anomalies in the dataset since I believed that SetID “SWSH07” contained vital data in the attribute variables that could explain why the surge in price was occurring.

The next step would be to begin modeling. I will be testing two popular techniques in stock market forecasting. These include a Time-Series ARIMA model and a Multi-Variate Random Forest (Kumbure et al., 2022).

When splitting the data for training and test sets, I used a function to ensure that each subset of data identified by SetID is split into training and test sets based on whether there is enough data. The purpose of this is to train the model for each SetID to see performance by SetID for the ARIMA model. This was done for both the ARIMA and Random Forest. Not only does this help with ensuring that there is sufficient training data, but it also avoids data leakage and keeps the process consistent. The dataset is now ready for modeling.

**Analysis**

A function was used to train the ARIMA model for each SetID, thus there are performance metrics by SetID. List (2) contains the performance metrics MAE, MSE, and RMSE for each SetID and aggregate. Figures (6), (7), and (8) show some of the forecasted data compared to actual data. We can see in the MSE by SetID that older sets account for large errors and indicate the presence of outliers or significant deviations in price. SetID “SWSH07” again show both in List (2) and Figure (8) how the outliers negatively impact the model’s performance.



*List (2)*

A graph with a line going up

Description automatically generated

*Figure (6)*

A graph with a line going up

Description automatically generated

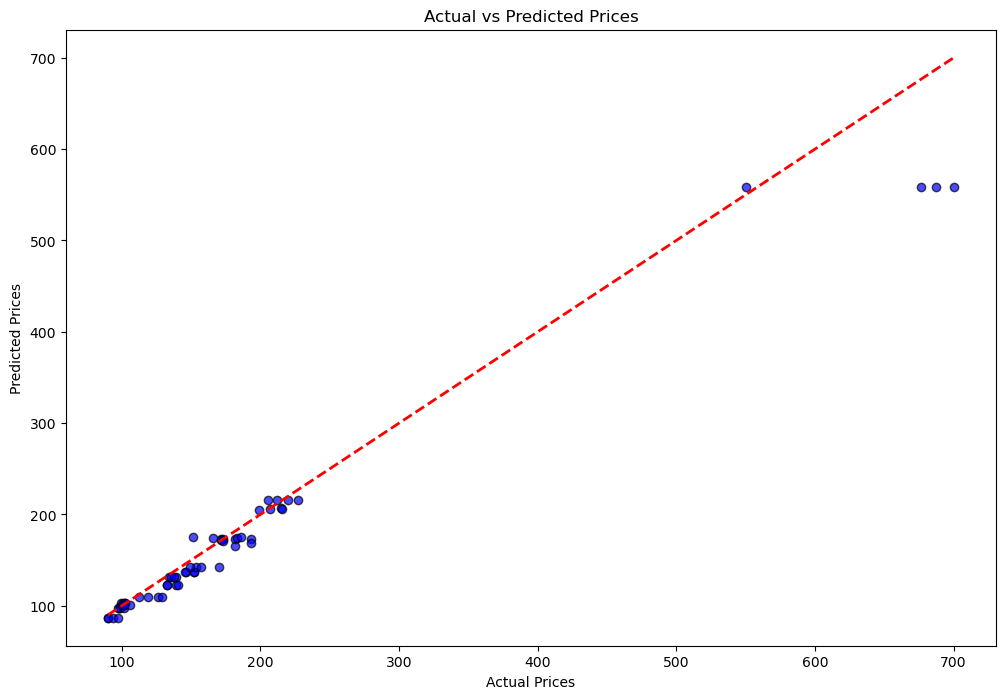
*Figure (7)*

A graph with blue lines

Description automatically generated

*Figure (8)*

The Random Forest model does not show performance by SetID and instead focuses on the attribute variables to influence price. For this model we also make predictions on the test data and result with an MAE of 14.21, an MSE of 948.28 and R2 of 0.95. The high R2 score indicates that the attribute variables explain a large proportion of the variance in price. The MAE is slightly improved than the ARIMA model but not by much. The high MSE indicates that the presence of outliers is negatively affecting the predictions. Figure (9) demonstrates how the outliers are skewing the performance metrics. If these outliers were to be removed an improved MAE could be achieved. The red dotted line indicates a perfect prediction. As observed in Figure (9) most predictions lie close to the red dotted line, while a few outliers shown in the far right skew the performance.



*Figure (9)*

**Conclusion**

The Random Forest model shows better performance than the ARIMA model when measuring MAE, MSE. Additionally, if the outliers were to be removed, the Random Forest model’s performance would further improve.

However, these metrics get much worse for the Random Forest model when making longer-term predictions. I trained the Random Forest model on the entire dataset and attempted to make predictions on the most recent price data (June 20-22). For example, when predicting the price of SetID “SWSH11” for this date range the model predicted the price with 98% accuracy. Likewise, SetID “SV03” price was predicted within 97% accuracy. However, when making predictions a year in advance the price did not reflect expectations. Although long-term predictions cannot be tested and measured for certain, past knowledge of TCG products and what we know about scarcity proves to us that price increases as DSR increases. Figure (10) shows feature importance in the Random Forest model. DSR is not given a high importance. What Figure (10) shows me is that the Random Forest model is good at telling us if a future Set to be release will be in high demand and may predict the sale price in the first 0-4 months of a Set’s life, but beyond that it will not adjust properly to temporal changes; likely indicating overfitting. The model performing well on short-term predictions but poorly on long-term ones can be a sign of overfitting. This means the model has learned the training data patterns very well but does not generalize to unseen future data where patterns may evolve, specifically temporal patterns (Barreñada et al., 2024).

A graph with blue squares

Description automatically generated with medium confidence

*Figure (10)*

ARIMA models deal better with temporal changes than Random Forest models. For this reason, the ARIMA model performs better than the Random Forest Model in long-term predictions and the Random Forest Model performs better in short-term predictions.

**Assumptions**

We are assuming that product scarcity will lead prices to increase over time. There are market conditions that could result in this not occurring. For example, if the scarcity of the products were to go down as a result of manufactures increasing the amount of TCG products produced. These market conditions are not something that can be accounted for, and thus we make the assumption based on historical context that scarcity and prices will continue to rise. The relationship between Price and DSR and the assumption of increased scarcity over time leads to the belief that the ARIMA model makes better predictions long-term.

**Limitations**

This project had two key limitations. The first is related to the data. As previously stated, TCGplayer only provides the most recent four months’ worth of sales data. This causes the class imbalances and underrepresentation of older Set IDs as seen in Figure (2). This limitation hinders the robustness of the data and creates extra effort to ensure that all SetIDs are adequately represented in the test and training splits.

The second limitation relates to the four-week window for conducting the project. Much was achieved and learned in this first iteration of the project. However, with additional time improvements can be made to both the data and models to improve performance. For example, no hyperparameters were tuned in the Random Forest model. Tuning could find the optimal number of trees and prune accordingly to reduce overfitting.

**Challenges**

As stated under limitations, the main challenge was with the dataset. Although a function was used to manually assign tests and training splits, it would still be better to improve the robustness of the dataset. Aside from the challenges with the data, the limitations of the models themselves also posed a challenge. ARIMA is good at dealing with temporal changes and can adjust for non-linear relationships noise in the data. However, it cannot manage multiple variables. Random Forest can handle multiple variables but can be prone to overfitting or learning trends in the training data that may not be present in new data. The short-term approach would be to use a hybrid approach.

**Future Uses / Additional Applications**

This project could be further developed by implementing recommendations that I will detail below. Not only could this work be used for Card Shops and Collectors but can also be translated to other hobbies with similar niche markets. A similar market could be the Sneaker market. Both temporal variables and attribute variables could be used to forecast the popularity or price of Sneakers.

Aside from applications in niche markets, the work started here could be used to further develop the field of stock market forecasting. While stock markets are much more complex and have an increased amount of macro and microeconomic variables, the work done here could be used to understand at a fundamental level work being done in the field of stock market or commodity forecasting.

**Recommendations**

My recommendation to improve the performance of both models is to focus on increasing the robustness of the data. This could be done by acquiring more historical data from TCGplayer or similar sources. Alternatively, this could be achieved by periodically collecting new data from TCGplayer to amas a large quantity of sales data over time. web scraping tools could be implemented to automate data collection and build a robust size of sales data.

With regards to the models themselves, I would firstly recommend tuning the Random Forest model using k-folds to determine the proper number of trees. Tunning this and other parameters could improve performance and reduce overfitting. Lastly, I would recommend adopting a hybrid approach where short-term predictions are made using the Random Forest model and longer-term predictions are made using the ARIMA model.

**Implementation Plan**

Before implementing I would advise that the improvements laid out in the recommendations section be made. Afterwards, the best approach would be to leverage a web scraping script to automate data collection and have a continuous stream of data fed to the models. Then I would again recommend the hybrid approach of utilizing a combination of both the Random Forest and ARIMA model.

**Ethical Assessment**

When a project involves assets that could be traded for financial gain it is important to make sure that anybody seeking to leverage the model’s output understands that all investing involves risk. Even with a highly accurate and optimized model there is still the chance that an investor could lose money.

The limitations and assumptions of the predictive models must be clearly communicated so as to avoid misleading stakeholders. I do not recommend anybody utilize the model’s results to guide investment decisions with these products. The purpose of this project is strictly educational and is in no way endorsed as financial advice by myself, or any other person(s) or entities discussed in this project.

**Sources**

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems With Applications,*, *197*, 2. https://doi.org/https://www.sciencedirect.com/science/article/pii/S0957417422001452

Barreñada, L., Dhiman, P., Timmerman, D., Boulesteix, A. L., & Van Calster, B. (2024). Understanding random forests and overfitting: A visualization and simulation study. *Academic*. https://doi.org/10.48550/arXiv.2402.18612

Bulbagarden (2024, April 16). *List of Trading Card Game Expansions*. Retrieved June 7, 2024, from https://bulbapedia.bulbagarden.net/wiki/List\_of\_Pok%C3%A9mon\_Trading\_Card\_Game\_expansions#Scarlet\_&\_Violet\_Series

TCGplayer (2024, June 1). *Trading Card Marketplace*. Retrieved June 7, 2024, from https://www.tcgplayer.com/