Rodrigo Rodriguez
06/22/2024
Applied Data Science
DSC680 (SUMMER)

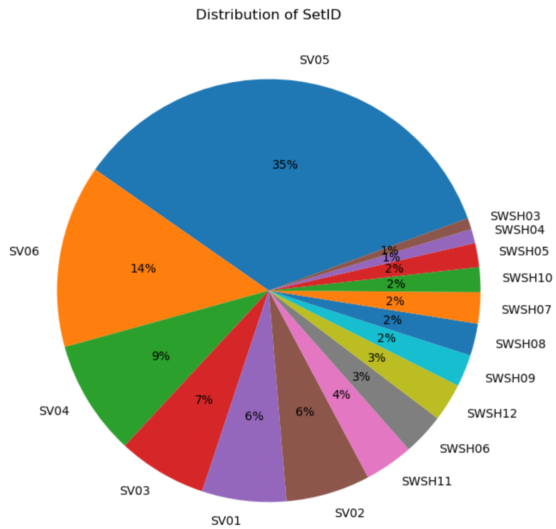
Project Whitepaper Draft: Analyzing TCG Box Sales to Predict Future Sale Prices

Business Problem TCG (Trading Card Game) Booster Boxes are primary assets in TCG marketplaces, similar to stocks. They are phased out after a year, creating scarcity and driving value. Card shop owners face challenges due to price fluctuations, potentially selling below MSRP. This project aims to analyze TCG Booster Box sales data to predict future sale values, helping shop owners and collectors make informed decisions, optimize pricing strategies, manage inventory, and anticipate market trends.

Background and History Trading Card Games have a history dating back to the early 20th century. Modern TCGs began in the 1990s with games like Magic: The Gathering and Pokémon. Booster Boxes, containing random assortments of cards, became primary products due to their potential for valuable rare cards. Factors influencing Booster Box value include rarity count, set size, scarcity, and demand. TCGplayer.com, established in 2008, provides sales data for market trend analysis. Historically, card shop owners relied on intuition, but data analytics now offers a scientific approach to pricing and inventory decisions.

Data Explanation The TCG marketplace is niche, with macroeconomic variables having less influence. The dataset combines attribute variables and sale data. Attribute data is from Bulbagarden, and historical sale data is from TCGplayer. The dataset primarily uses daily sales records for the most recent four months, aggregated to weekly data to reduce noise. Only basic data cleaning was performed, such as changing data types for certain variables.

Methods Data exploration revealed insights such as sales distribution by set and the positive relationship between price and days since release (DSR). Anomalies in the data, particularly with SetID "SWSH07," were retained for their valuable insights. Two modeling techniques were tested: Time-Series ARIMA and Multi-Variate Random Forest. The dataset was split to ensure sufficient training data and avoid data leakage. The ARIMA model showed performance metrics by SetID, while the Random Forest model focused on attribute variables.



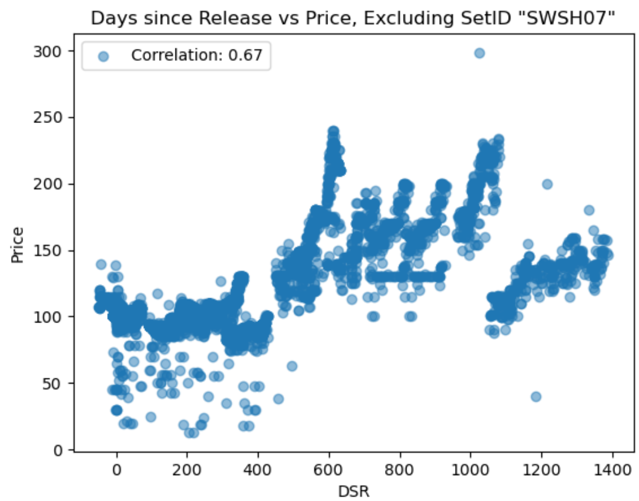


Figure 1

Figure 2

Analysis The ARIMA model's performance was measured by MAE, MSE, and RMSE for each SetID. Anomalies, especially with older sets, negatively impacted the model's performance. The Random Forest model showed a high R2 score, indicating attribute variables explained a large proportion of price variance. However, outliers skewed the performance metrics. The Random Forest model performed well for short-term predictions but poorly for long-term ones, indicating overfitting.

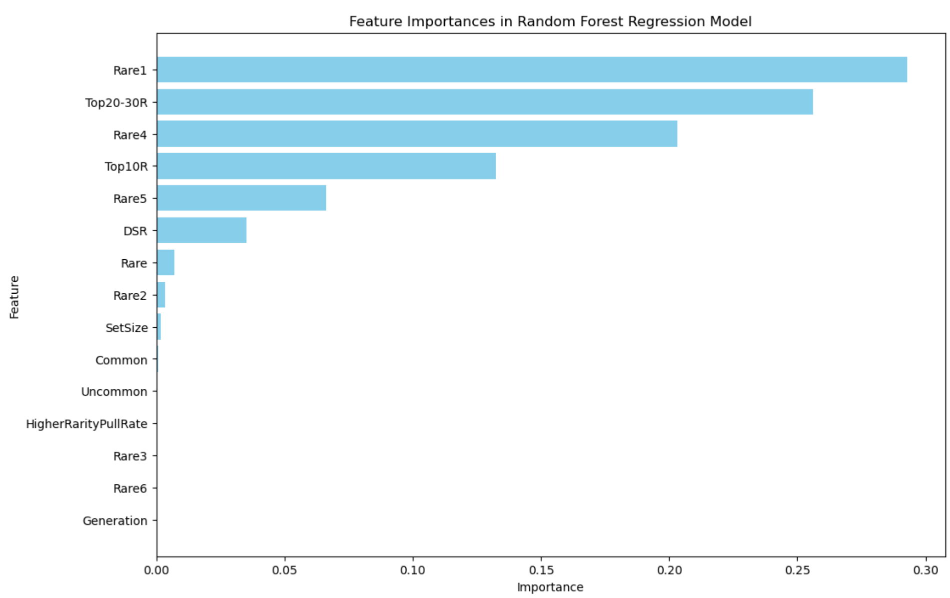


Figure 3

Conclusion The Random Forest model outperformed the ARIMA model in short-term predictions, but ARIMA was better for long-term predictions due to its ability to handle temporal changes.

Assumptions The assumption that product scarcity leads to price increases was made, though it may not hold under certain market macroeconomic conditions.

Data limitations The dataset only contains the recent four months' worth of sales data. The lack of historical data hindered performance. The project's four-week timeframe was also a key challenge. Future work should focus on improving data robustness and model tuning.

Challenges Primary challenges included lack of historical data and the short time period to conduct the project.

Future Uses and Recommendations The project can be further developed for other niche markets like the sneaker market. Recommendations include acquiring more historical data, using web scraping tools for data collection, and adopting a hybrid modeling approach with tuned Random Forest for short-term and ARIMA for long-term predictions.

Implementation Plan Implement improvements from recommendations, automate data collection with web scraping, and use a hybrid approach for modeling.

Ethical Assessment Clearly communicate model limitations and assumptions to avoid misleading stakeholders. The project is for educational purposes and not financial advice.

Potential Audience Questions

1. Can the demand for unreleased TCG Sets be predicted with this project?
2. Can you explain what factors are most important in predicting the price of TCG sets
3. How did you determine the specific attributes to include in your predictive models for TCG Booster Box prices?
4. What challenges did you face while collecting and cleaning the data, and how did you address them?
5. Can you explain why the Random Forest model performed better for short-term predictions but not for long-term predictions?
6. What are the main limitations of using ARIMA models for predicting TCG Booster Box prices?
7. How do you address the issue of data scarcity, especially for older sets, in your analysis?
8. How do you plan to continuously update and improve your models as new sales data becomes available?
9. Can your predictive models be adapted for other types of trading cards or collectible markets?
10. What are the ethical considerations of using predictive models for investment decisions in the TCG market?

Sources/Appendix

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/