Unpacking Value: NLP in Trading Card Economics

Introduction

Magic: The Gathering (MTG) is a popular collectible card game with a dynamic secondary market, where card prices fluctuate based on various factors. This study aims to predict MTG card prices by analyzing their in-game utility and attributes. The primary data source is a comprehensive dataset from Scryfall, offering detailed insights into each card's attributes and market prices.

In MTG, each card has various attributes, each potentially influencing its market value in distinct ways. Key elements of a card include its rules text, rarity, type, power, toughness, and mana cost. The rules text, detailing the card's abilities and mechanics, is believed to have a significant impact on its price, as it directly affects the card’s playability in the game. Rarity is another crucial factor, as cards that are less common often command higher prices due to their scarcity and potential power in gameplay. Conversely, attributes like flavor text and artist name, while contributing to the aesthetic and thematic appeal of the cards, are hypothesized to have a minimal direct impact on their market prices. This study aims to create a model that can guess the price of MTG cards.

The dataset obtained from Scryfall for this study was initially very extensive, encompassing a broad range of attributes. However, to focus on the most relevant factors for predicting card prices, significant data refinement was necessary. The final dataset was streamlined to include the following columns:

1. **Card Name:** The unique identifier of each card.
2. **Card Cost:** The cost required to play the card in game.
3. **Card Type:** Different card types do different things in the game.
4. **Card Text:** The rules text explaining the card's abilities.
5. **Power, Toughness, and Loyalty:** These values are unique to certain card types, but they all indicate a card’s strength.
6. **Reserved:** A true or false value indicating whether the card is on the MTG reserved list, a list of cards that are promised not to be reprinted, thus affecting their potential value.
7. **Booster:** A true or false value indicating whether the card is available in booster packs. Packs of 15 random cards, booster packs are the main way cards enter the market.
8. **EDHRec Rank:** A score based on the frequency of the card's inclusion in online deck lists, reflecting its popularity and utility in the game.
9. **Legalities in Standard, Pioneer, Modern, and Commander:** A true or false value indicating whether the card is legal in each format (formats are ways to play MTG).
10. **Number of Printings:** The frequency of the card's reprinting, affecting its availability.
11. **Number of Colors:** Magic cards can have up to five colors.
12. **Rarity:** A key determinant of a card's scarcity and collectability.
13. **Date of Printing:** The date the card was most recently printed.

14. **Price:** The target variable: the price (in USD) of the card on November 25, 2023 when the data was exported.

This refined dataset strikes a balance between comprehensiveness and focus. Each of the variables can be encoded numerically with the notable exception of name and rules text. Name can be dropped because it serves merely as a unique identifier; card text, however, is hypothesized to carry a large amount of information about the card's price.

Methodology

The data preparation process involves cleaning, transforming, and one-hot encoding of categorical variables like type and rarity. Card text needs to be analyzed using Natural Language Processing (NLP) techniques. Because card text is hypothesized to be important, I tried a variety of techniques in an attempt to capture the information therein. These NLP techniques are essential for extracting meaningful information from the card text, which is a rich source of data but presents challenges due to its unstructured nature.

Word Counting: This method involves counting the number of words in the card text. It is a straightforward approach that provides a basic measure of the complexity or verbosity of the card text. The simplicity of word counting is a significant advantage, making it computationally efficient. However, its major limitation is that it fails to capture the semantic content or the context of the words used, which can be crucial in understanding the card's utility and hence its market value.

Phrase Counting: Similar to word counting, this method involves counting specific phrases, such as "draw a card." It is effective for identifying specific, repeated game mechanics that could impact a card's value. While it provides more context than individual word counting, it is still limited in its ability to capture the full semantic richness of the card text.

Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF is a more advanced technique that evaluates how important a word is to a document. This method is beneficial for identifying unique or rare terms that could signify a card's special abilities or features. The main advantage of TF-IDF is its ability to highlight the words that are distinctive to a particular card. However, it can sometimes overlook the importance of common but relevant words in the card text.

Bag of Words (BoW): The BoW model converts text data into numerical feature vectors, representing the occurrence of words within the document. It is useful for capturing the presence of certain terms or phrases. Like all the methods discussed above, it does not account for the order or structure of words, which can be a significant drawback in understanding the contextual meaning of the text.

Fine Tuning ChatGPT: This method is expensive both computationally and in the sense that OpenAI charges money for model fine tuning. Tuning the model used in this project cost me

$15.25. Still, I wanted to see how state of the art techniques stacked up against more traditional methods. This method preserves most of the semantic meaning, including meaning encoded in word order. However, large language models (LLMs) like ChatGPT don't always produce consistent results.

Each of these NLP techniques has its pros and cons. While simpler methods like word and phrase counting are computationally less intensive and easier to implement, they provide a more superficial analysis of the text. In contrast, techniques like TF-IDF and BoW, are more complex, but should offer a deeper, more nuanced understanding of the text at the cost of higher computational complexity. A LLM should be best able to capture the meaning inherent in textual data; however, using a chatbot comes with its own set of complications including price of training the model, and inconsistency in results.

After the data is cleaned and the various NLP techniques are applied, it's time to make a prediction. The study employs a variety of statistical techniques to handle the high-dimensional nature of the Scryfall dataset. Key methods include lasso and ridge regression for variable selection and shrinkage, principal component regression (PCR), and partial least squares (PLS) for handling multicollinearity and reducing dimensionality. These methods were chosen for their efficacy in dealing with high-dimensional datasets and their ability to uncover underlying patterns in the data.

Results

A comprehensive exploration was undertaken to evaluate the effectiveness of various natural language processing (NLP) techniques and predictive models in estimating card prices. This investigation integrated four NLP methods: Word Counting, Phrase Counting, Term Frequency-Inverse Document Frequency (TF-IDF), and Bag of Words (BoW), each paired with four distinct model types: Lasso, Ridge, Partial Least Squares (PLS), and Principal Component Regression (PCR).

The initial findings were not particularly encouraging. Utilizing the mean price as a baseline, which resulted in a Mean Squared Error (MSE) of 171.83, it was observed that among the 16 different combinations of NLP techniques and model types, the pairing of Word Counting with Ridge Regression yielded the most favorable outcome, albeit marginally, with an MSE of 163.33. This performance was only slightly superior to the baseline.

To enhance the predictive accuracy, a modification was introduced by transforming the response variable, price, from a continuous to a categorical metric. Prices were categorized as '0' for values strictly less than $5 and '1' for those greater than or equal to $5. Under this categorical framework, the baseline MSE was computed to be 0.047. Subsequent application of Lasso and Ridge Regression to this redefined dataset revealed that the TF-IDF combined with Lasso Regression emerged as the most effective method, achieving an MSE of 0.040.

Despite these adjustments, the overall performance remained suboptimal. The Language Model (LLM), initially deemed promising due to its ability to interpret context from natural language, reported training and validation losses of 1.03 and 0.47 respectively. However, it failed to deliver accurate price predictions in practical applications. The model frequently produced erroneous non-price outputs, including illogical numerical values with multiple decimal points.

Upon reviewing these outcomes, Professor Antonelli recommended that the underwhelming results could be attributed to the application of linear models to a dataset with inherently nonlinear characteristics. He recommended employing a Random Forest approach for the analysis. Due to project deadlines and the high computational demands of fitting Random Forest using my laptop, I had to limit the scope of this approach. I only had time to fit a single model: Random Forest Regression on the BoW dataset with price as a continuous value. Despite these limitations, preliminary trials demonstrated a marked improvement, achieving an MSE of 141.15. Although still not ideal, Random Forest did significantly better than the linear models.

Discussion

The study's outcomes underscore the necessity for further exploration into non-linear modeling techniques, as preliminary evidence suggests the presence of non-linear dynamics within the dataset. An ideal extension of this research would involve the application of Random Forest algorithms across the entire spectrum of NLP techniques, as well as on the categorically transformed dataset. Additionally, cross-validation of tuning parameters would be a critical step. However, these endeavors were constrained by the limitations of available computational resources and the project deadline.

The Large Language Model (LLM), despite its lackluster performance in this study, holds potential for future applications. The training approach, which originally focused on teaching the model to output a singular numerical value for the price, could be revised. A more effective strategy might involve training the model to generate complete sentences, such as “The estimated price of the card is $0.55.” It is hypothesized that this adjustment could significantly reduce the occurrence of nonsensical outputs. Furthermore, I forgot to include card rarity in the training data. This was an oversight as I believe incorporating this variable could have enhanced the model's accuracy. However, I decided to refrain from retraining the model due to the monetary expense.

This project was a significant learning experience for me. Initially, I thought it would be a straightforward task to solve: there were extensive datasets available online. As an experienced Magic: The Gathering (MTG) player, I am often able to predict the prices of cards by considering aspects like their rules text and rarity. I assumed that since all relevant information about each card was accessible in the dataset, predicting prices would be feasible. However, I underestimated the challenge posed by the unstructured nature of the rules text and the complexities of natural language processing.

I also discovered the importance of exploring both linear and nonlinear models. Initially, I was convinced that models like Lasso, Ridge, PCA, and PLS would be most effective due to the high dimensionality of the data (31 variables before incorporating NLP). What I overlooked was the substantial size of the dataset (over 21,100 observations), which meant that shrinkage was not necessary. This realization opened up the possibility of employing more complex nonlinear models that could yield better results.