Rodrigo Rodriguez

06/26/2024

Applied Data Science

DSC680 (SUMMER)

**Analyzing TCG Box Sales to Predict Future Sale Prices**

**Potential Audience Questions**

1. **Can the demand for unreleased TCG Sets be predicted with this project?**
2. Yes, the demand for unreleased TCG Sets can be predicted by using the attribute variables of the unreleased TCG Set to predict price (demand).

The Random Forest model can be used to predict the price using attribute variables.

1. **Can you explain what factors are most important in predicting the price of TCG sets**
2. The feature importance plot for the Random Forest model tells us that the top three most important features are the following.
3. **Rare1** – A variable that counts the quantity of moderately rare cards in a Set.
4. **Top20-30R** – A variable that counts the quantity of popular characters featured in the Set.
5. **Rare4** – A variable that counts the quantity of highly rare cards in a Set.

What this tells us is that Sets with many rare cards and with many popular characters will have a positive effect on price increase.

1. **How did you determine the specific attributes to include in your predictive models for TCG Booster Box prices?**
2. The attribute variables were chosen due to contextual assumptions that these attributes were good indicators of price. I had a hypothesis that the variables included would be good indicators of future price. Although this hypothesis ended up being correct, that does not mean that there are other attributes not included here that are better predictors of price.
3. **What challenges did you face while collecting and cleaning the data, and how did you address them?**
4. The main challenge when collecting the data was having to scrape web pages on TCGplayer.com for recent sales. Initially I was going to take the approach of creating a python script to scrape the HTML for different URLs for historical price data. However, I hit a roadblock when I discovered that only a small quantity of historical data loads on page load. This means that a button needs to be clicked on the page to have the HTML load additional data. This button would have to be clicked many times to load the full four months history of sales data. Building a click bot that would do the necessary clicks and then scrape the data would take longer than manually grabbing the data one time for this project, so I decided to manually collect the data one time. For this model to be deployed and have a constant feed of recent sales data, a click bot would need to be built to collect the data.
5. **Can you explain why the Random Forest model performed better for short-term predictions but not for long-term predictions?**
6. The Random Forest model performed better for short-term because the model placed a higher importance on attribute variables rather than temporal variables, like ‘DSR’. Attribute variables do not change after a Set releases, only temporal variables. As time passes and DSR increases, the Random Forest model’s predictions tend to decrease in accuracy. For long-term predictions the ARIMA model performs better since it deals better with temporal variables.
7. **What are the main limitations of using ARIMA models for predicting TCG Booster Box prices?**
8. While ARIMA does better for long-term predictions in our project, the main limitation of the model is the lack of supporting multiple variables. In the short-term attribute variables are more important. ARIMA only accounts for time and relies on a robust sales history for accurate predictions.
9. **How do you address the issue of data scarcity, especially for older sets, in your analysis?**
10. The way that I addressed it was using a function to manually assign test and training splits. This was done to ensure that each Set ID had adequate representation in both training and test sets. A better way of addressing this issue is increasing the robustness of the data to include a wider range of historical sales; although, this may be challenging.
11. **How do you plan to continuously update and improve your models as new sales data becomes available?**
12. I plan to create a click bot that can take the actions required on the web to load the historical sales data in the HTML and then can collect that data. This is a scalable approach that creates a steady stream of data for future use.
13. **Can your predictive models be adapted for other types of trading cards or collectible markets?**
14. This type of model can be applied to other collectible markets or trading cards. The important distinction is that the attribute variables will need to change to variables that are relevant to the other markets.
15. **What are the ethical considerations of using predictive models for investment decisions in the TCG market?**
16. Most importantly, it must be well communicated that the model only makes predictions, and no prediction is certain. A shorter way of saying this is that all investing involves risk. Additionally, this model and project were built for educational purposes and is not optimized for deployment in any context. Any individuals looking to build on this work will have to communicate similar ethical considerations.