Lego Price Analysis

Kinnick Fox

Bellevue University – DSC630

7/25/2023

Milestone 2 - Proposal

Introduction

For this project, I would like to focus on Lego building sets and how price points for these sets are decided. The high prices of certain building sets have always baffled me and have left me wondering how the Lego Group comes to pricing decisions. In this project, I hope to better understand the Lego Group's pricing decisions and build a model capable of predicting a building set's price.

Dataset

Many datasets for Lego are available, however one dataset contained the information I believed to be necessary for this project. The Lego Database from Kaggle contained information of piece count, release year, and theming information. This is all useful for the project but a supplemental dataset, also from Kaggle, titled Lego Sets would need to provide pricing data that will also be crucial to this project. Luckily both datasets contain a product ID that I hope to use to join useful information.

Figure 1 - The Millennium Falcon - Ultimate Collector Series (MSRP: $849)

Model/Evaluate

I believe a decision tree classifier model may be my best bet for this dataset. The combination of float and categorical data would mean two models would be required otherwise. A decision tree classifier would allow me to see the driving factors that determine pricing for sets. Splitting the data into test and train sets will allow me to evaluate the model's accuracy once it has been trained properly. Binning the target variable, price of set, may be necessary to get a decent accuracy rating on the model but that will need to be addressed in the future

Desired Outcome

The desired outcome is a working decision tree classifier model able to accurately predict Lego set prices. I am hypothesizing that piece count, or size of the set, is not necessarily the driving factor. Instead, I am expecting the licensing for the set or recommended age to be more influential of the pricing of the set. I am anticipating licensing to be a driving factor because the company that owns the intellectual property being portrayed by Lego will receive some sort of revenue from sales of sets and I believe recommended age will be a driving factor because the more expensive sets will most likely be made for an adult audience. I am hoping to prove or disprove this hypothesis at the conclusion of this project.

Ethical Implications

I don't foresee any ethical implications that require consideration for this project. No sensitive data is being generated and it is doubtable that the data being used would

be able to generated insights capable of tarnishing the Lego Group's reputation. If that was the case, tarnishing any company's reputation with findings could potentially lead to loss of sales and downsizing of company which would most likely lead to termination of employees that do not fit within the company's new size. Publicizing unbridled findings could cause tremendous second and third order effects that could negatively impact many people's lives. This kind of situation would be made even worse if the findings were discovered to be faulty or in error. Although these considerations need to be taken seriously in any project, they will not be of any consequence for this project in particular.

Contingency Plan

If the Lego Group project hits a dead end, I would like to look into UV exposure and its effect on melanoma. There are plenty of national databases for cancer related data so I believe this should be a solid backup plan that should come together fairly quickly. The challenge for this project would be to find databases with useful data that possess unique identifiers that would be able to be joined on. The desired outcome for this project would be to find a significant jump in melanoma cases once a specific threshold of exposer has been reached or at least to better understand the risks of prolonged UV exposer.

Milestone 3 – Preliminary Analysis

Expectations

After reflecting on the peer review provided by David Pahmer as well as my chosen dataset, I believe my project is still healthy and capable of producing my desired

results. I predict that the dataset contains the minimum features and is robust enough to understand which factors influence price of Lego sets the most.

Visualizations

A scatterplot depicting price and physical product would be a good visualization. This could potentially introduce a trend line that Lego sets could fall above or below to indicate outliers. Histograms for categorical data, specifically theme, may also be beneficial in understanding the dataset. A correlation matrix may assist in removing features from the model prior to modeling.

Data Adjustments

The dataset is not in a usable state as of now. The different counties also have different currency which will need to be normalized. Many of the Lego sets repeat because they are sold in multiple countries. Repeated sets will need to be removed and the remaining reginal exclusive sets will need to be normalized in price. Other various cleaning steps such as dealing with missing values and correcting feature format will also need to occur. Dummy variables will need to be created for a few categorical features as well.

Model and Evaluation

As of now, I see no need to change my initial model and evaluation steps. Some adjustments will most likely be required as the project moves forward but I will not know until an issue arises. In that case I feel that I have laid out enough possible methodologies to pivot to a new model without much issue.

Milestone 4 – Finalizing Results

Data Prepping

Several processing steps needed to take place before the dataset was model ready. Dropping duplicates, normalizing currency within the list_price feature, and dropping unneeded features came first. The target feature (list_price) was selected and the remaining features were separated (piece_count, theme_name, ages, review_difficulty).

Modeling and Revision

The modeling process was difficult and inevitably did not yield a high accuracy model. Due to the presence of categorical features within the separated data, encoding needed to take place. Encoding would allow a decision tree classifier to be fit with the data. Labeling of the target variable was also needed initially due to it being a continues variable. This setup yielded an accuracy that floated between 20% and 30%. To improve accuracy, I attempted removing additional features leaving only piece_count and theme_name, as I hypothesized these features would have the most impact on the target feature. This setup faired only slightly better, hovering around the low 30s for

accuracy percentage. Binning of the target variable ultimately needed to happen to improve accuracy but I believe doing so also diluted the results. Initially, binning occurred from 0 to 800 in increments of 20 which yielded a mid-40% accuracy. Upping the bins to increments of 50 finally yielded a somewhat acceptable accuracy of mid-70%. Unfortunately this led the model to make every end node a 0-50 or 50-100 which is the bulk of the dataset making it a slightly more informed coin toss. This accuracy is even less impressive after finding out that the list_price has a correlation of 86% with piece_count.

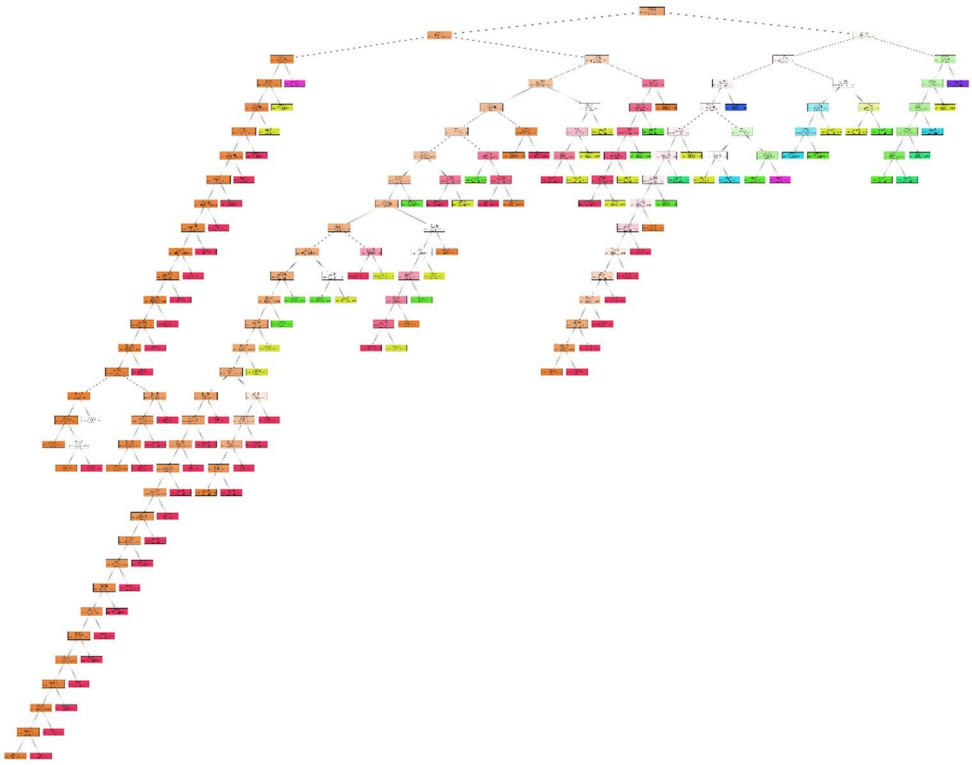

Figure 3 - Final Decision Tree Classifier Visualization

A linear model was then used to possibly determine the variables with the highest regression coefficient with the hypothesis this to most likely be piece count. Some additional data prepping occurred in the form of creating dummy variables for categorical data so that it would fit into a linear regression. Results showed that my hypothesis was wrong and the top six variables with the highest positive regression coefficient can be seen in the figure below.

SERIOUS PLAY®	355.944262
BOOST	31.579483
MINDSTORMS®	29.080914
Challenging	20.626457
Average	11.971038
piece_count	0.080798

Table 1 - Top Regression Coefficients

Conclusion

Unfortunately, this dataset does not contain the ample data to create a working prediction model with acceptable parameters. I believe this outcome is most likely due to the lack of sets priced above $200, and the overabundance of sets priced below $50. Some missing features that would have been helpful in an analysis are sales of a given set and price to produce a given set. These features may have allowed me to find the ample price point for Lego or better understand pricing decisions. Linear regression did shed some light on the features with the highest regression coefficient which ended up proving my initial hypothesis wrong. To break down these features and perform some

speculation, the feature with the highest regression coefficient was the Serious Play theme. This is a small line of Lego sets designed for corporate team building with a focus on tapping into the imagination of employees. The target demographic of this set is a good indicator as to why it may be highly influential of sale price. The Mindstorms and Boost themes are all made up of robotics sets, containing various servos, control boards, and motors. These sets are most likely more expensive to manufacture which might explain their influence on sale price. Challenging and Average are the two highest difficulty ratings from the four-category rating system which may indicate they are created with an older demographic, with more disposable income, in mind. Piece count is the last item on within the positive regression coefficient table and has the smallest (nearly negligible) regression coefficient. This implies a weak relationship between piece count and sale price within this model.

References

Brick Economy (images and pricing figures). Retrieved from https://www.brickeconomy.com/sets/top/most-expensive-lego-sets

Rachael Tatman *Lego Database* (2017). Retrieved from Kaggle at https://www.kaggle.com/datasets/rtatman/lego-database

Mattieterzolo *Lego Sets* (2018). Retrieved from Kaggle https://www.kaggle.com/datasets/mterzolo/lego-sets

The Lego Group *Lego Themes*. Retrieved from Lego https://www.lego.com/en-us/themes