Lego Learning: Using Machine Learning to Predict the Future Value of Lego Sets

Alex Racapé & Liam Jachetta
CSCI 3465 Financial Machine Learning

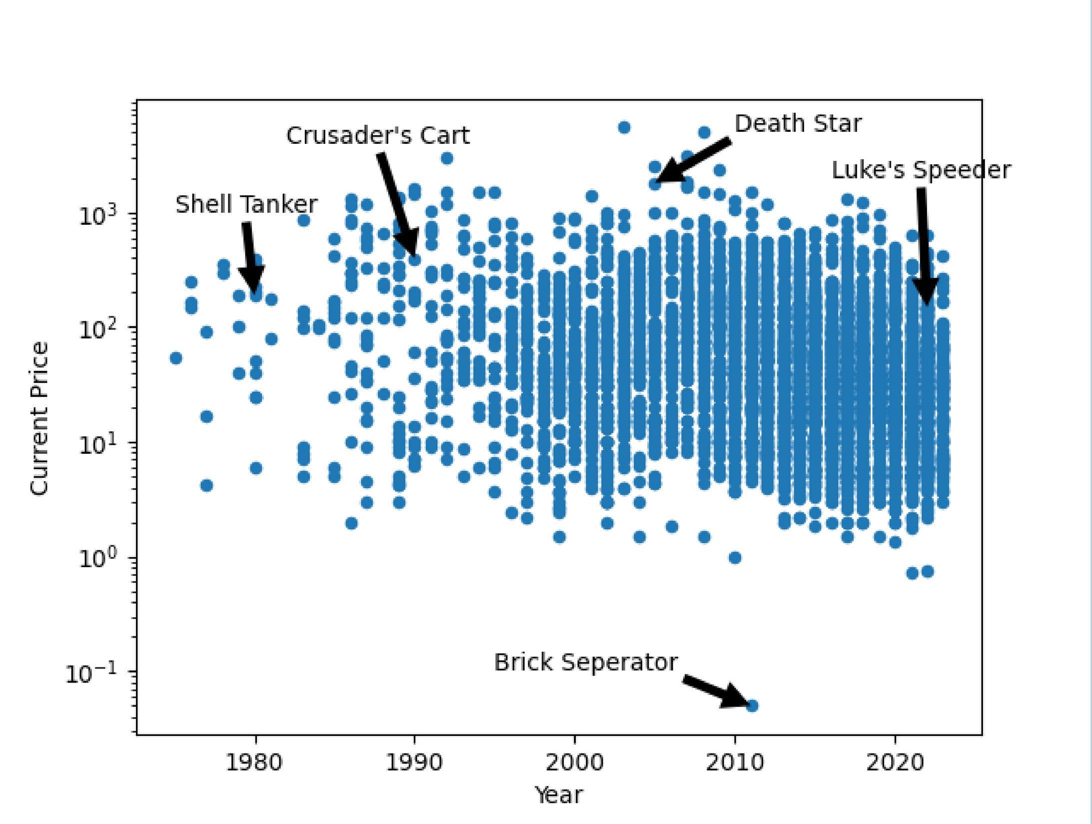
Abstract

- Goal: explore the viability of trading Lego sets as an investment strategy.
- Employed machine learning algorithms to predict future prices of sets and tested trading performance against a custom benchmark
- Collected data by scraping and combining two APIs
- Analyzed which factors contribute most to list price
- Contrasted results from two types of regression learners

- Model was able to consistently outperform an equally weighted portfolio across all sets and the S&P 500
- Number of pieces emerged as the most influential factor in determining overall value
- Future price forecasts revealed that large sets, especially from the Star Wars theme, were predicted to see the largest increase in value.

Data

- Created dataset from scratch using a combination of two API's from BrickLink and Brickset
- Explored existing datasets and found that all were lacking recent and complete data
- Scraped the Brickset API to get features for 15 thousand Lego sets dating back to 1975
    - Got year released, list price, number of pieces, number of Minifigs, and many other features
- Used Bricklink to access the last six months of trades from their online stores
    - Only five thousands sets had recent trades
    - Out of those three thousand also had information on list price



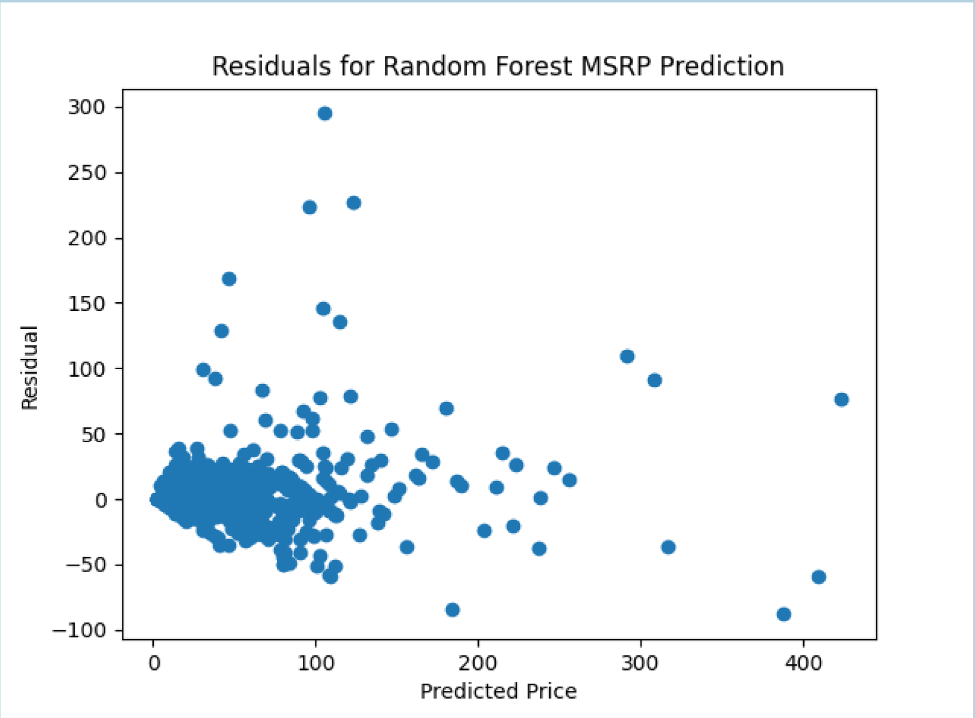
Models

Random Forest

- Used code that we wrote from scratch in a previous lab as the basis for our random forest.
- Random forest is an ensemble learner that is made up of a bunch of random decision trees. Each tree in the ensemble is trained using bootstrapped data, and the ensemble aggregates the results from different trees into a single prediction.
- Marginal benefit of adding trees declined at around twenty.

- Trained the model to predict the current market price
- Used year released, theme, number of pieces, number of minifigures, star rating, number owned, and the list price as features

- Used k-fold cross validation to evaluate the model's predictions compared to the actual current market prices
- Observed an average correlation of 0.78 and RMSE of $81.
- Residuals highlight several outliers and variance at higher prices
    - 4 of the largest outliers were from Lego education sets which include a small number of robotic pieces and motors.

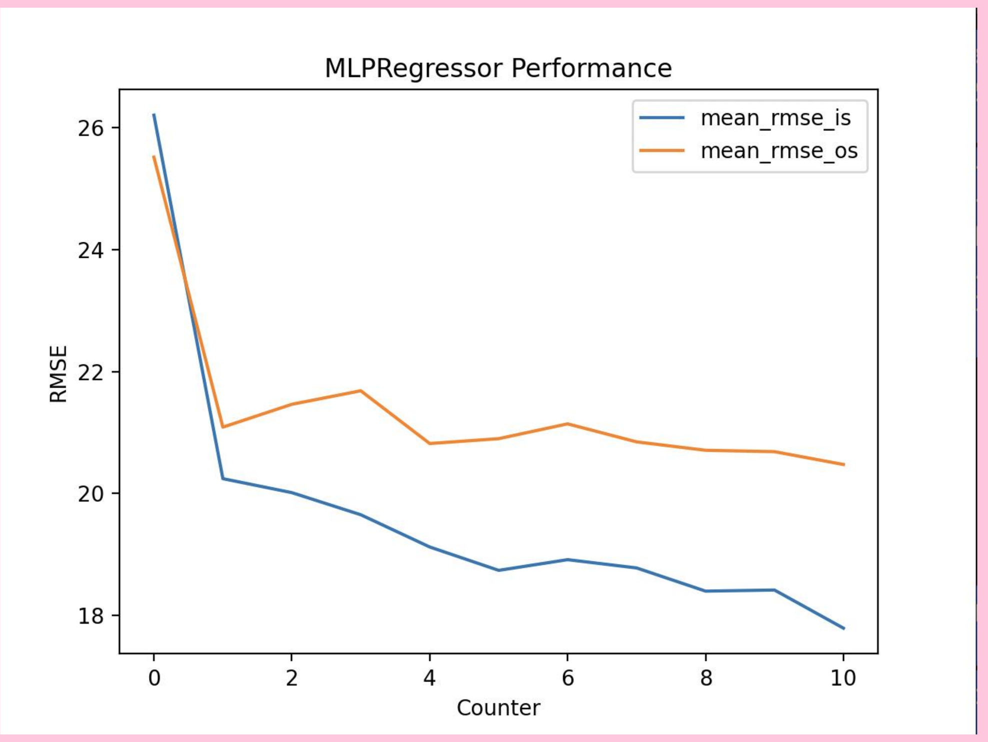


Neural Networks

- Used SKLEARN's MLPRegressor function
- Tested the number of neurons in each layer and the number of iterations to tailor the model
- Used the same features as the random forest

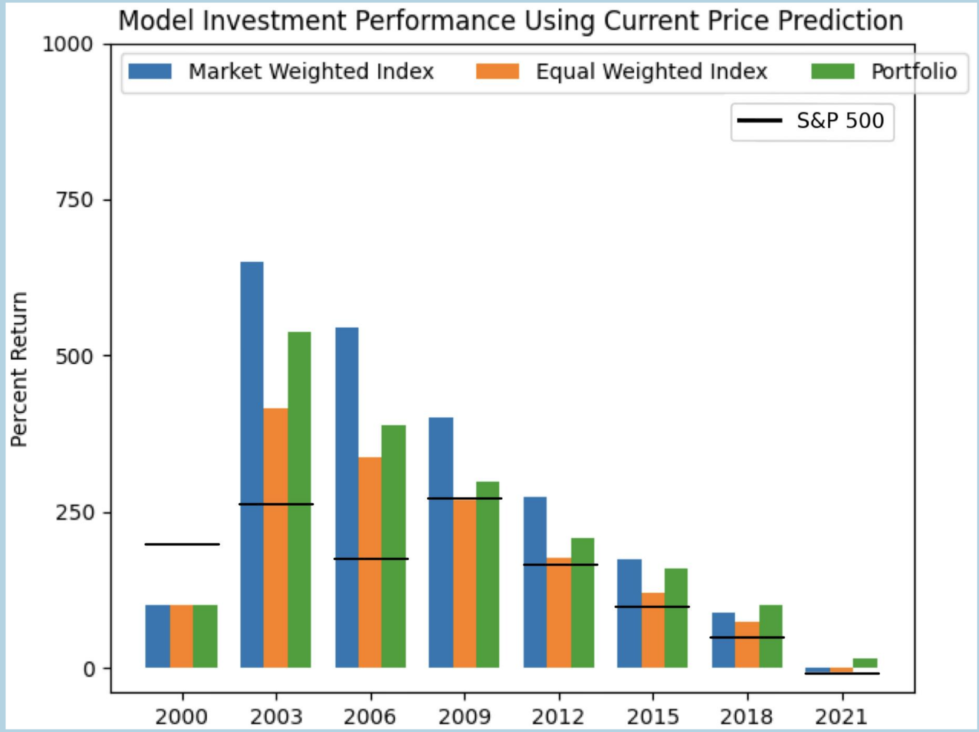
- 4000 iterations was best as it allowed the model to convergewhile not taking an inordinate amount of time to run
- RMSE in and out of sample declined as more neurons were added to the model
- The optimal amount of neurons was (35, 45, 55) as this was the point where out-of-sample data stopped improving and began to plateau

- Used k-fold cross validation to evaluate the network
- Observed a correlation of 0.93 between current price and the predictions.

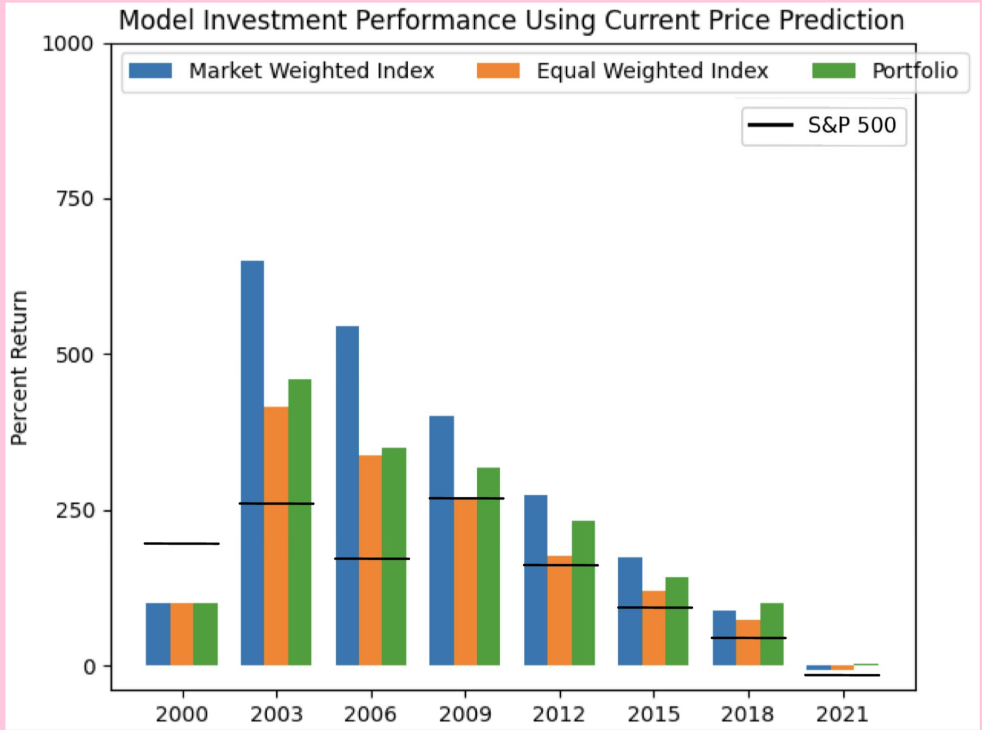


Trading Experiments

- Used roll forward cross validation to test on sets grouped by year
- For each year, the model trained on sets produced in previous years and learned to predict the future market price for all the sets
- Constructed a portfolio that was weighted by the predicted gain for each set.
- To analyze the returns, constructed two benchmarks

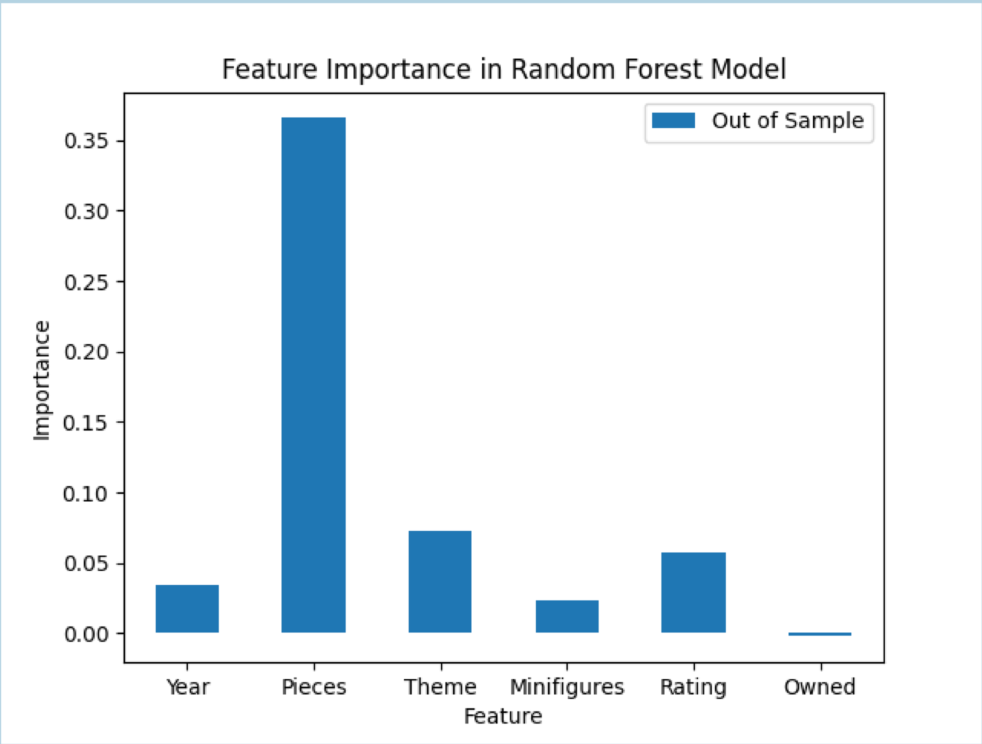


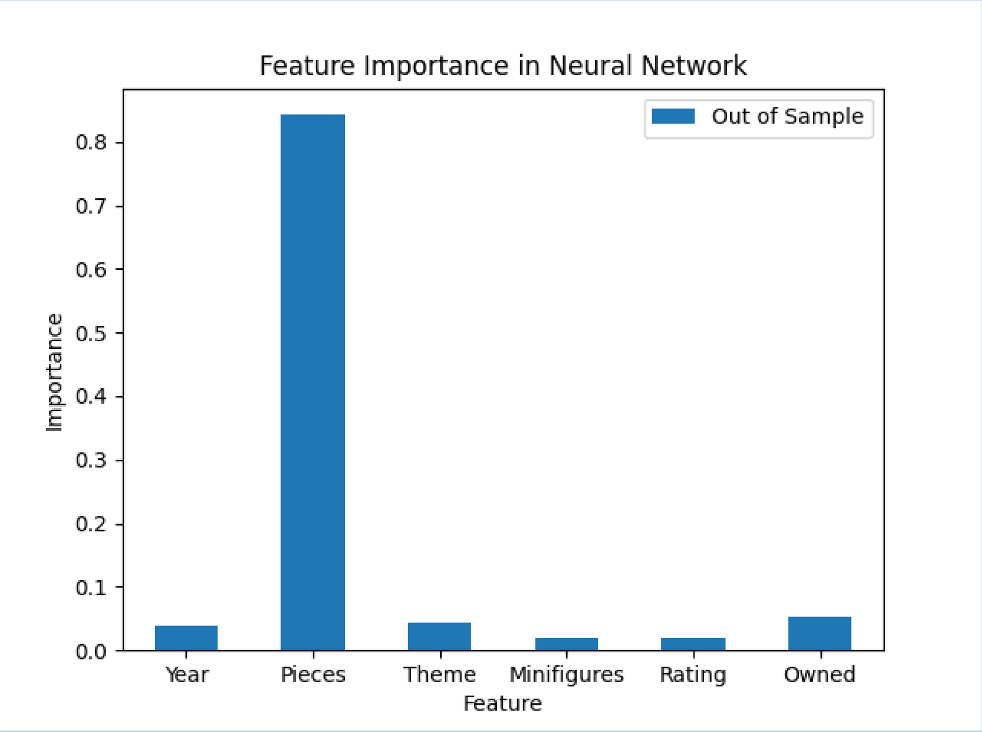
- Equal weighted index invests equally in all sets produced that year
- Market weighted index uses an approximation of 'market cap' to weight each set in the portfolio
    - Number owned attribute on Brickset was used as a proxy for the number of sets produced
    - Slightly biased metric that highlights more popular sets from



List Price Modelling

- Adapted model to predict list price instead of market price
- Achieved .901 and .952 correlation when predicting the list price with random forest and neural network respectively
- Used a permutation test and changes in R^2 to determine a feature's importance
- In a sense, models which features Lego would use to price a set.





Predicting Future Prices

Percent gain
- The model picks ten very cheaply priced sets that it expects to appreciate several hundred percent
- Best pick:
    - **Brick Separator, Orange**
    - Currently priced at 5 cents, with an expected 58,900% increase to $29 dollars by 2028.
    - We do not suggest taking this as investment advice as there were very few sets this cheap in the training data

Total dollar gain
- The model picks ten larger Lego sets
- 6 of the 10 are Lego Star Wars themed
- Best Pick:
    - **Luke Skywalker's Landspeeder**
    - Currently priced at $149.90 with an expected gain of 132% up to $348 by 2028.