Case Study: Building a LEGO Investment Strategy Through Data Analytics

By Akshay Prabu
uq6733

Predictive Analysis: Identifying Investment Opportunities

1. **Brief Introduction:**

- Focused on LEGO sets released in 2019 to evaluate their investment potential based on predicted vs actual price.
- Selected key numerical features from the dataset including piece count, number of minifigures, weight, and calculated box volume.
- Trained a regression model (Linear Regression and Random Forest) to predict the retail price of each LEGO set using a 70/30 train-test split.
- Evaluated model performance using R^2 and MAE metrics to ensure reliable prediction accuracy.
- Calculated value potential for each set by subtracting actual price from the predicted price, identifying underpriced and overpriced sets.
- Grouped sets by theme, subtheme (Brickheadz and Juniors), and price tier to extract the top 2 and bottom 2 performers in each group based on value potential.

2. **Linear Regression Model:**

```
Validation R² Score for training: 0.9182
Validation MAE for training: $7.37
Validation R² Score for validation: 0.9199
Validation MAE Score for validation: $11.80
```

```
R² Score (Full Data): 0.9202
MAE (Full Data): $8.71
```

- The model achieved a high R^2 score of 0.9182 on the training set, indicating it explains over 91% of the variance in set prices during training a strong fit with the selected features.
- The training MAE of $7.37 means that, on average, predictions during training were off by around $7.37 from actual prices relatively low considering LEGO sets can range from $10 to $500+.
- On the validation set, the model maintained an even higher R^2 score of 0.9199, showing that it generalizes well and isn't overfitting the training data.
- The validation MAE was $11.80, slightly higher than the training MAE, which is expected in real-world data, but still shows reasonably accurate predictions for unseen data.
- The full dataset R^2 score of 0.9202 and MAE of $8.71 confirm consistent model performance across all 2019 LEGO sets, further validating the model's reliability for predicting price and identifying value opportunities.

3. **Random Forest:**

```
Random Forest R² Score (Full Data): 0.9254
Random Forest MAE (Full Data): $5.33
```

- The Random Forest model achieved an R^2 score of 0.9254 on the full dataset, indicating that it explains over 92% of the variance in LEGO set prices slightly better than the linear regression model.
- The Mean Absolute Error (MAE) was $5.33, meaning the model's predictions were, on average, just $5.33 off from the actual retail price a strong performance considering the wide price range of LEGO sets.
- This lower error compared to Linear Regression (MAE of $8.71) shows that Random Forest handled the non-linear relationships between features and price more effectively.
- The model's high accuracy and low prediction error make it well-suited for identifying sets that are underpriced or overpriced, which is critical for making informed investment decisions.

4. **Comparison of both the models:**

- Both models performed well in predicting LEGO set prices, with high R^2 scores above 0.92, indicating they captured over 92% of the variance in the data.
- Random Forest outperformed Linear Regression with a slightly higher R^2 score (0.9254 vs. 0.9202), showing it modeled the price data more accurately.
- The MAE for Random Forest was significantly lower ($5.33) compared to Linear Regression ($8.71), meaning its average prediction error was smaller and closer to the true price.
- Random Forest is a non-linear, ensemble-based model, which likely allowed it to capture complex relationships and feature interactions (e.g., piece count, weight, volume) that Linear Regression could not.
- While Linear Regression is simpler and easier to interpret, Random Forest offered greater accuracy and consistency, making it the better choice for predicting price and computing value potential in this LEGO investment context.

5. Value potential when grouped by theme, subtheme and price:

[64]:

	Name	Theme	Subtheme	RF Price Tier	US Retail Price ($)	RF Predicted Price	RF Value Potential
0	Dock Side Fire	City	Fire	19.99–29.99	19.99	19.72	-0.27
1	Dock Side Fire	City	Fire	19.99–29.99	19.99	19.72	-0.27
2	Fire Chief Response Truck	City	Fire	19.99–29.99	29.99	29.34	-0.65
3	Fire Chief Response Truck	City	Fire	19.99–29.99	29.99	29.34	-0.65
4	Burger Bar Fire Rescue	City	Fire	34.99–69.99	39.99	42.44	2.45
5	Fire Plane	City	Fire	34.99–69.99	59.99	57.89	-2.10
6	Fire Plane	City	Fire	34.99–69.99	59.99	57.89	-2.10
7	Fire Station	City	Fire	34.99–69.99	69.99	61.29	-8.70
8	Downtown Fire Brigade	City	Fire	74.99–99.99	99.99	103.09	3.10
9	Downtown Fire Brigade	City	Fire	74.99–99.99	99.99	103.09	3.10

- The sets are grouped by Theme, Subtheme, and Price Tier, and within each group, the top 2 and bottom 2 performers are identified based on their value potential (predicted price minus actual retail price).
- In the City - Fire group, Downtown Fire Brigade stands out in the $74.99–$99.99 tier with a positive value potential of +$3.10, indicating it may be slightly underpriced and worth considering for investment.
- Conversely, Fire Station in the $34.99–$69.99 range shows a significant negative value potential (–$8.70), suggesting it may be overvalued relative to its build content and should be approached cautiously.
- Across tiers, Fire Plane appears twice with negative value potential, indicating a pattern of overpricing, while Burger Bar Fire Rescue offers modest positive value, hinting at slightly better price-to-content balance.

6. Value potential when grouped by Theme, Subtheme (Brickheadz and Juniors)

[68]:

	Name	Theme	Subtheme	RF Price Tier	US Retail Price ($)	RF Predicted Price	RF Value Potential
0	Garbage Truck	City	Juniors	19.99–29.99	19.99	20.54	0.55
1	Garbage Truck	City	Juniors	19.99–29.99	19.99	20.54	0.55
2	Garage Centre	City	Juniors	34.99–69.99	49.99	47.49	-2.50
3	Garage Centre	City	Juniors	34.99–69.99	49.99	47.49	-2.50
4	Batman and the Joker Escape	DC Super Heroes	Juniors	34.99–69.99	39.99	38.04	-1.95
5	Batman and the Joker Escape	DC Super Heroes	Juniors	34.99–69.99	39.99	38.04	-1.95
6	Batman vs The Riddler Robbery	DC Super Heroes	Juniors	Below $19.99	9.99	11.04	1.05
7	Batman vs The Riddler Robbery	DC Super Heroes	Juniors	Below $19.99	9.99	11.04	1.05
8	Anna's Canoe Expedition	Disney	Juniors	19.99–29.99	19.99	20.64	0.65
9	Cinderella's Carriage Ride	Disney	Juniors	19.99–29.99	19.99	18.97	-1.02

- The output reflects the result of grouping LEGO sets by Theme, Subtheme (Brickheadz and Juniors), and RF Price Tier, and extracting the top 2 and bottom 2 sets in each group based on their value potential.

- Sets like Garbage Truck (Juniors) in the $19.99–$29.99 tier show a positive value potential (+$0.55), suggesting they are slightly underpriced and may offer a minor investment opportunity.
- In contrast, sets like Garage Centre (City - Juniors) in the $34.99–$69.99 tier are appearing twice with a negative value potential (–$2.50), indicating they are overpriced relative to their predicted value, and may not be good investment targets.
- Interestingly, in the DC Super Heroes - Juniors category, both sets (Batman and the Joker Escape and Batman vs. The Riddler Robbery) show nearly identical predicted prices and values, but only the Riddler set crosses into positive value potential (+$1.05), making it a slightly more favorable pick.
- The table also highlights how subtle pricing gaps (even as small as $1–$2) can impact investment attractiveness when sets are compared within the same theme and price tier using predictive analytics.