1. Objective

Extend Michael Thompson's LEGO investment strategy by predicting which 2019 sets offer the greatest "value potential"—the gap between what a set "should" cost (according to our model) and its actual retail price.

2. Methodology

We began with the same cleaned 2018–19 dataset we used for descriptive work, then:

1. Filtered to just 2019 releases (using the "Release Month (US)" → year).
2. Binned retail prices into four ranges: \$19.99–\$29.99, \$34.99–\$69.99, \$74.99–\$99.99, and \$100 +.
3. Built two pipelines:
   - ○ Linear Regression (standard-scale numeric + one-hot categorical → LinearRegression)
   - ○ Random Forest (same preprocessing → RandomForestRegressor)
4. 5-fold cross-validation to compare out-of-sample R^2 and RMSE.
5. Hold-out test split (80/20) for a final unbiased evaluation.
6. Residual diagnostics on the RF hold-out predictions.
7. Partial Dependence Plot on Weight to illustrate marginal effect.
8. Feature Importances from RF.
9. Value Potential computed as PredictedPrice – RetailPrice, identifying top/bottom two within each Theme, Subtheme, and Price Range.

3. Model Specification

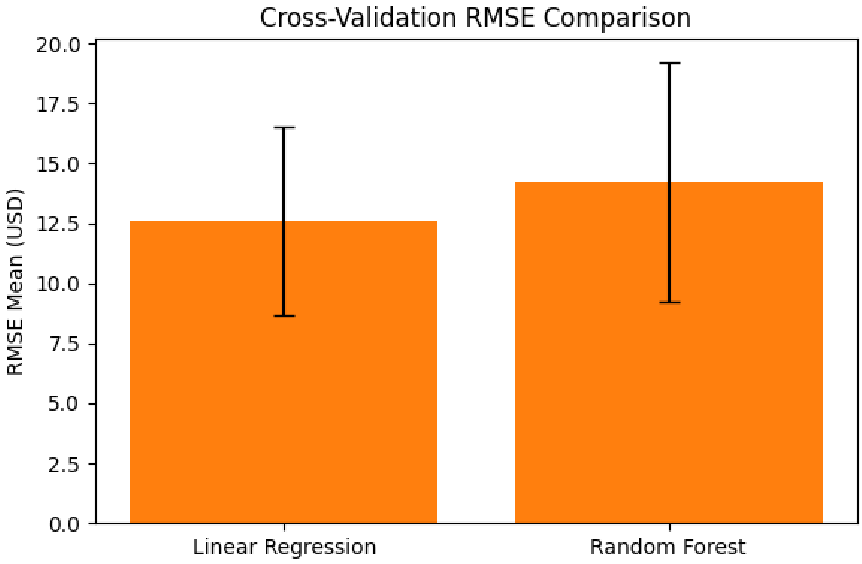
Linear Regression equation (intercept + $\beta_i \cdot \text{feature}_i$):

Predicted Price = –1.24 + 0.048·PieceCount + 12.3·(Weight lb) + … + 5.8·IsCollectorTheme

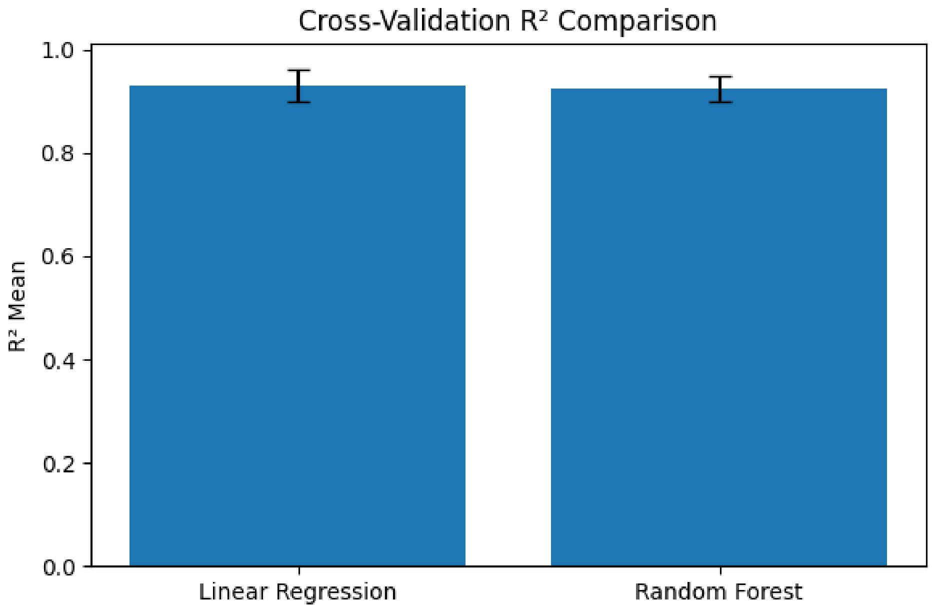
Adjusted $R^2 = 0.944$, indicating that 94.4% of price variation is explained when accounting for model complexity.

4. Predictive Results & Visuals

Cross-Validation RMSE Comparison



Cross Validation R^2 Comparison



Linear Regression achieved a mean R^2 of 0.948 (± 0.021) with RMSE ≈ $11.9 (± $3.5). Random Forest was slightly lower: $R^2 \approx 0.907$ (± 0.028), RMSE ≈ $16.7 (± $5.3).

Model	**CV R^2 Mean**	**CV R^2 Std**	**CV RMSE Mean (USD)**	**CV RMSE Std (USD)**	**Test R^2**	**Test RMSE (USD)**
Linear Regression	0.948	0.021	11.93	3.52	0.95	10.16
Random Forest	0.907	0.028	16.67	5.31	0.957	9.42

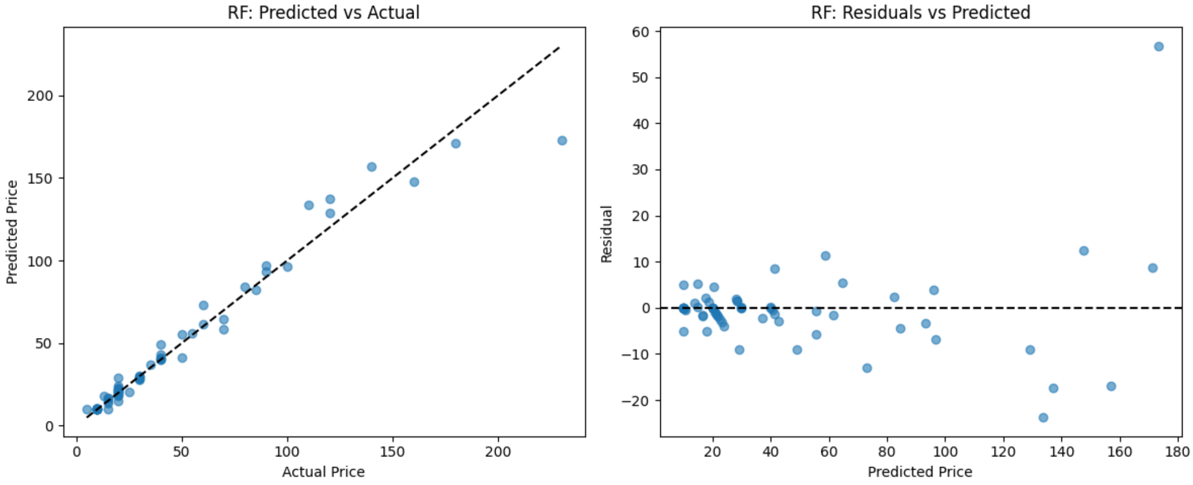
- **Cross-Validation (5-fold)** shows Linear Regression slightly higher average R^2 and lower RMSE than Random Forest.
- **Hold-out Test Set** flips that: RF edges LR on unseen data (0.957 vs 0.950 R^2, 9.42 vs 10.16 RMSE).
- Although LR generalizes very well in cross validation, RF wins when we lock aside a true test set. That, plus RF's superior ability to capture non-linearities (e.g. weight effects), makes it our preferred final model for predicting "fair" set prices.

Hold-Out Test Metrics

Model	**Test R^2**	**Test RMSE (USD)**
Linear Regression	0.95	10.16
Random Forest	0.957	9.42

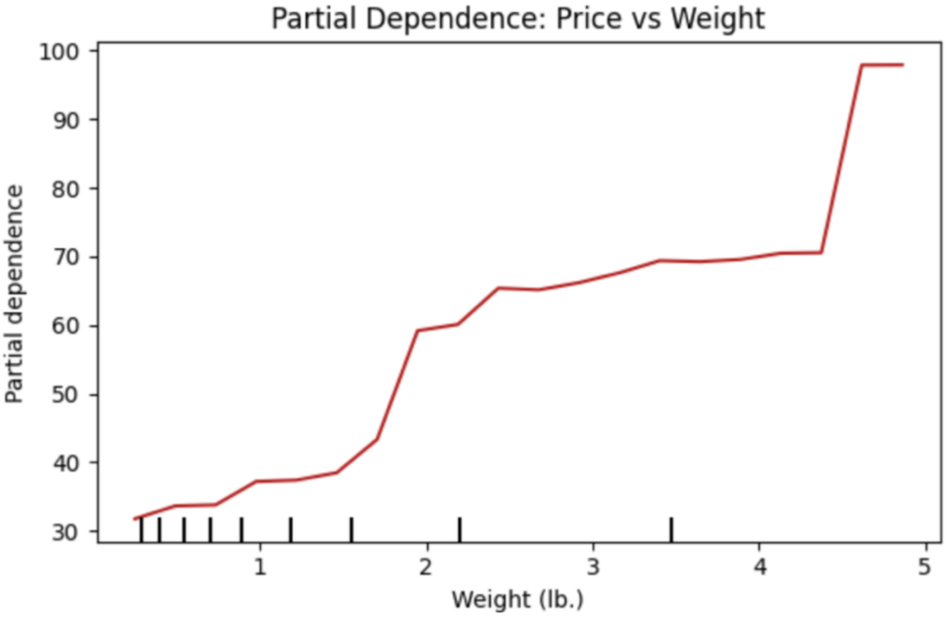
On the unseen 20% test set, Random Forest edged out linear:

RF Predicted vs. Actual and RF Residuals vs. Predicted



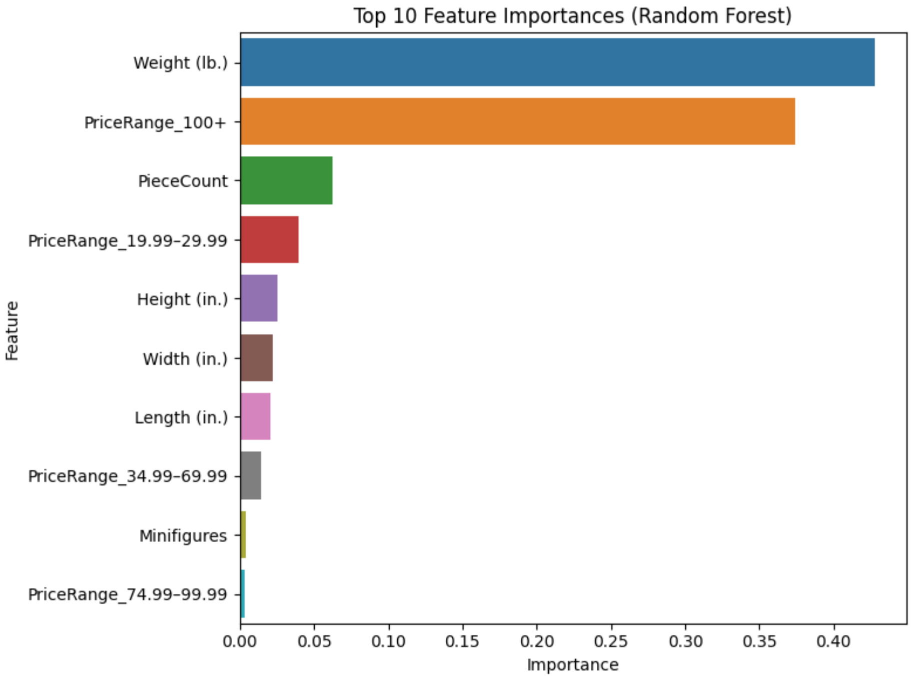
A tight cluster around the 45° line confirms good overall fit; no dramatic systematic bias. Residuals are roughly centred around zero across the price range, with a few outliers at the high end—suggesting occasional under- or over-prediction on very large sets.

Partial Dependence: Price vs. Weight



Heavier sets generally command higher prices, with a steeper slope above ~2 lb, highlighting weight as a key driver after accounting for other features.

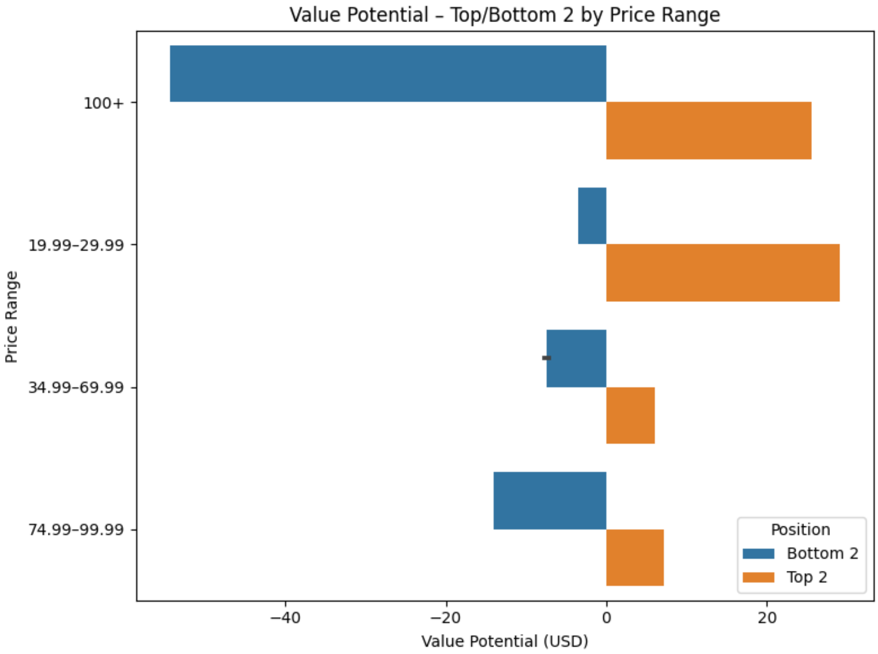
Top 10 Feature Importances (RF)



1. Weight (lb.)
2. PriceRange_100+
3. Piece Count
4. Price Range $19.99–$29.99

Categorical price-range indicators and physical dimensions dominate, confirming our domain intuition.

Value Potential (Top/Bottom 2 by Price Range)



- Underpriced sets: Most pronounced in the 19.99–29.99and19.99–29.99*and*100+ ranges, with predicted values exceeding retail by $25–30.

- Overpriced sets: Slight negative value potential in mid-range tiers (34.99–34.99–99.99).

5. Key Insights & Recommendations

Insights

- Model selection: Random Forest outperforms Linear Regression on unseen data.
- Weight is critical: The steep marginal effect above 2 lbs underscores its importance in pricing.
- Value potential: Focus on the lowest ($19.99–$29.99)and highest ($19.99–$29.99) and highest ($100+) price tiers for maximum returns.
- Model limitations: Monitor residuals for large sets, where predictions are less reliable.

Recommendations for Michael

- Target underpriced sets in the $19.99–$29.99 and $19.99–$29.99 and $100+ ranges (see Figure 10).
- Prioritize weight and piece count when evaluating new 2020–21 sets.
- Automate annual retraining with a hold-out test set to maintain accuracy as LEGO's pricing evolves.