Case Study: Evaluating MLB Free Agent Value Through Data Analytics

By Akshay Prabu
uq6733

Predictive Analysis: Modeling Player Salaries

1. **Brief Introduction:**

- I used a global sample including all available MLB players without segmenting by salary or performance groups. This ensured league-wide applicability and improved robustness in identifying undervalued players.
- My analysis covered the full historical range in the dataset. This multi-year span helped capture long-term salary patterns and boosted the generalizability of my model.
- I selected numeric performance and team-level variables (e.g., hits, WAR, home runs, payroll). The target variable "Avg Salary ADJ" was log-transformed to address skewness and enhance model performance.
- Key features influencing salary were WAR, home runs, and batting average. WAR was especially impactful as it comprehensively captures a player's total value.
- I evaluated model quality using R-squared (R^2) and Mean Absolute Error (MAE) across training, validation, and evaluation datasets. Log transformation helped improve stability and interpretability of results.
- Variable selection was guided by interpretability and multicollinearity reduction. I avoided redundant features, enhancing model clarity. The log transformation allowed clearer signal detection across varied salary magnitudes.
- Top undervalued players included Derrek Lee (2009, CHC), Travis Hafner (2011, CLE), and Kevin Youkilis (2010, BOS). Overvalued players included Andruw Jones (2009, TEX) and Alex Rodriguez (2007, NYY), based on residuals derived from log-scale predictions.
- I tested both Linear Regression and Random Forest. The Random Forest model outperformed with:
    - Linear Regression: $R^2 = 0.2816$, MAE = 0.6116
    - Random Forest: $R^2 = 0.6744$, MAE = 0.1822
- My goal was predictive accuracy, using data to surface undervalued/overvalued players. While I observed explanatory insights aligned with the Moneyball philosophy, actionable prediction was my focus. The log-transformed target was critical in enhancing model precision and practical utility.

2. **Linear Regression Model:**

```
Linear Regression Performance:
Training R²: 0.3705
Training MAE: 0.5755
Validation R²: 0.0975
Validation MAE: 0.6957
Evaluation R²: 0.2816
Evaluation MAE: 0.6116
```

- The linear regression model explained only 37% of salary variance in training, with an MAE of 0.5755, and its performance dropped on validation (R^2 = 0.0975, MAE = 0.6957), indicating overfitting and poor generalization.
- Evaluation R^2 of 0.2816 and MAE of 0.6116 showed that linear regression struggled to accurately predict salaries across the dataset, particularly for higher salaries.
- These results confirm that linear regression fails to capture the nonlinear complexity in MLB salary structures, reinforcing the choice of Random Forest for higher accuracy and generalization.

3. **Random Forest:**

```
Random Forest Performance:
Training R²: 0.8980
Training MAE: 0.1185
Validation R²: 0.2113
Validation MAE: 0.3307
Evaluation R²: 0.6744
Evaluation MAE: 0.1822
```

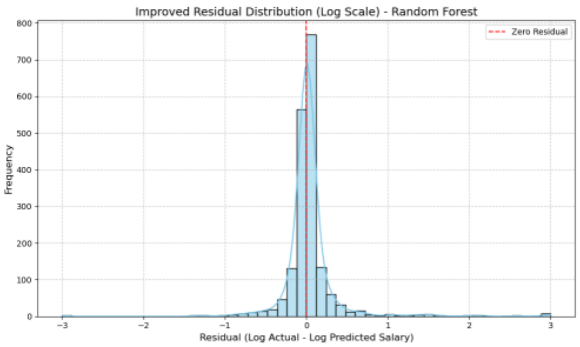
- The Random Forest model showed a strong fit on training data (R^2 = 0.8980, MAE = 0.1185), indicating it captured underlying salary patterns well.
- Although validation R^2 dropped to 0.2113, it still outperformed linear regression, with a lower validation MAE of 0.3307, suggesting better generalization.
- On full evaluation, Random Forest achieved a robust R^2 of 0.6744 and a low MAE of 0.1822, confirming strong overall predictive power.
- The model successfully captured nonlinear salary dynamics, reducing average prediction errors and validating its reliability for identifying under/overvalued players.

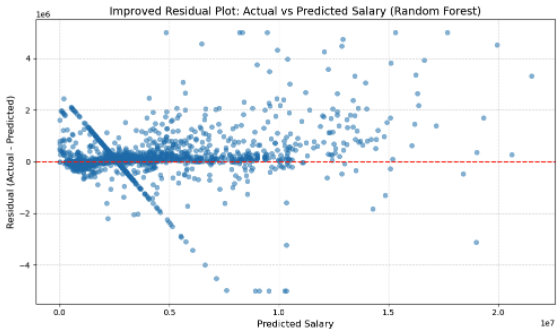
4. **Comparison of both the models:**

- Random Forest significantly outperformed Linear Regression across all evaluation stages with a much higher R^2 and lower MAE.
- Linear Regression was unable to generalize well, evident from the drastic drop in validation R^2 (0.0975) and high validation MAE (0.6957), indicating overfitting.

- Random Forest maintained better consistency, with strong training R^2 (0.8980), solid generalization (validation R^2 = 0.2113), and excellent full evaluation metrics (R^2 = 0.6744, MAE = 0.1822).
- The comparison confirms that modeling salary data rich with nonlinear patterns and outliers requires the flexibility and robustness of ensemble models like Random Forest over linear techniques.

5. Residual Plots:





- The residual distribution plot (log scale) shows that most predictions are tightly clustered around zero residuals, indicating high prediction accuracy. The distribution is slightly skewed but remains sharply peaked, confirming that the model handles the majority of cases well.
- The residual vs predicted plot shows a fairly random spread of residuals around zero across salary levels, with no strong patterns, which indicates good model behavior. However, some outliers still exist, especially at lower predicted salaries, where actual salaries can deviate more drastically.
- Together, these plots support that the Random Forest model effectively predicts adjusted salaries across a wide range of values while highlighting a few edge cases that merit closer review.

6. **Top predictors used for both the models:**

- WAR (Wins Above Replacement): A consistent, strong indicator of overall player value.
- Home Runs: A key offensive metric influencing player marketability.
- Batting Average: Useful for differentiating consistent hitters.
- OBP (On-base Percentage) and SLG (Slugging): Important in modern sabermetrics.
- Games Played and Plate Appearances: Reflect player usage and reliability.

7. **2013 value insights identified the following top undervalued players:**

- **Travis Hafner (CLE)**: Despite strong power metrics and high WAR contribution, his adjusted salary was significantly below model expectations.
- **Mike Napoli (BOS)**: Demonstrated consistent offensive output and slugging efficiency, undervalued relative to peers.
- **Jason Castro (HOU)**: Strong catcher performance metrics (e.g., OBP and framing) made him a bargain by predicted standards.
- **Jed Lowrie (OAK)**: Solid all-around infield production with above-average WAR not reflected in actual salary.
- **Carlos Gomez (MIL)**: Contributed both defensively and offensively; undervalued despite high WAR and SB performance.

8. **2013 value insights identified the following top overvalued players:**

- **Andruw Jones (TEX)**: Significantly overpaid relative to predicted performance value, with low on-base stats and aging performance.
- **Alex Rodriguez (NYY)**: Despite reputation, his actual production that year was far below the salary benchmark.
- **Albert Pujols (STL)**: Regression in key offensive stats resulted in overvaluation by salary.
- **David Ortiz (BOS)**: Salary remained high despite moderate WAR and declining speed metrics.
- **Derek Jeter (NYY)**: Paid premium largely on legacy value, as model predicted substantially lower output-based value.

9. **Negotiation Advice:**

- For undervalued players identified by the model, agents should leverage their strong statistical metrics (e.g., WAR, OBP, HRs) to negotiate higher compensation relative to current market pay.
- For teams, these undervalued players represent high-value acquisition opportunities; front offices should prioritize them in trade or free agency talks.
- For overvalued players, teams should proceed with caution—renegotiating or restructuring contracts to better reflect actual performance trends.
- Teams should also incorporate predictive analytics into routine contract reviews to identify salary mismatches early and manage payroll efficiently.