Predictive Analysis Report

Evaluating MLB Free Agent Value Through Data Analytics

1 | Problem Setup

Building on our descriptive exploration of free-agent batter salaries (1998–2013), our goal is to predict fair contract values for 2013 free agents and identify undervalued or overvalued players. Specifically, we:

- **Segment** players into **Low**, **Mid**, and **High** salary tiers using tertiles (33rd/66th percentiles) of 1998–2012 salaries.
- **Train** separate Random Forest regression models for each segment, using batting metrics (HR, OBP, SLG, WAR) and Age as predictors.
- **Evaluate** on 2013 hold-outs via a true year-based split when ≥5 test cases exist; otherwise, use 5-fold cross-validation (CV) on the training segment.
- **Combine** segment-specific predictions to assess overall 2013 accuracy.

2 | Data Preparation & Modelling

Data

- 2,400+ seasons of batting stats and inflation-adjusted free-agent salaries.
- Dropped missing values in core fields (e.g., WAR, OBP).

Segmentation

- **Low**: ≤ 1.29M;∗∗Mid∗∗:≤1.29*M*;∗∗*Mid*∗∗:≤4.47M; **High**: > $4.47M (based on 1998–2012 tertiles).
- **2013 hold-outs**: Low (23 players), Mid (26 players), High (39 players).

Features

- HR (home runs), OBP (on-base percentage), SLG (slugging), WAR (wins above replacement), Age.

Modelling Procedure

- For each segment:
    - **If ≥5 hold-outs**: Train on pre-2013 data, test on 2013.
    - **If <5 hold-outs**: Use 5-fold CV on training data, then refit on the full segment for feature importance.

- **Combined prediction**: Aggregate segment-specific predictions for 2013 and compute overall R^2 and RMSE.

3 | Key Results

Segment Performance

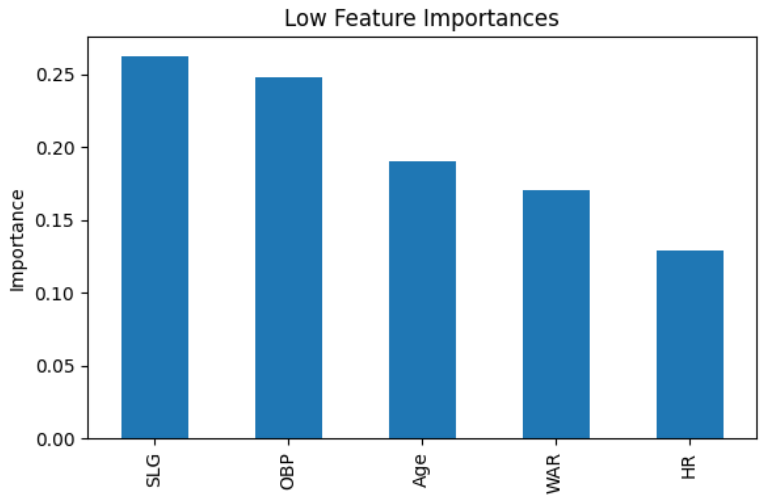
Segment	Method	R^2	RMSE
Low	5-fold CV	–0.38	$396,261
Mid	5-fold CV	–0.17	$907,868
High	Hold-out	0.04	$4,766,681

Overall (Quantile-Segmented Model)

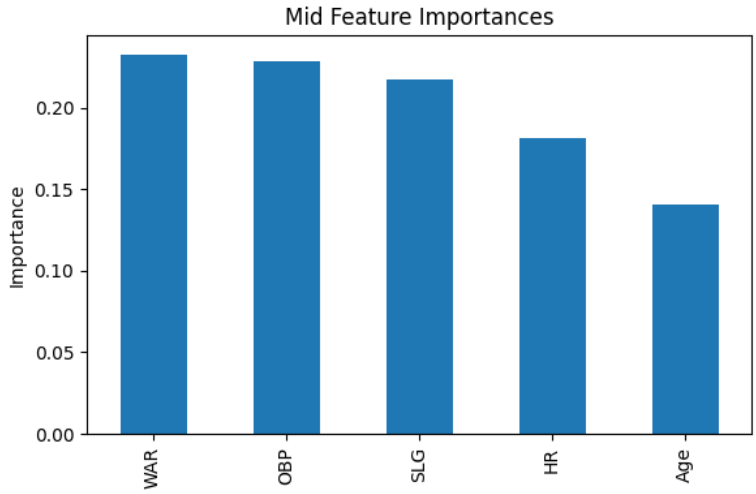
- **R^2 = 0.62**, **RMSE = $3,217,795**

Figures

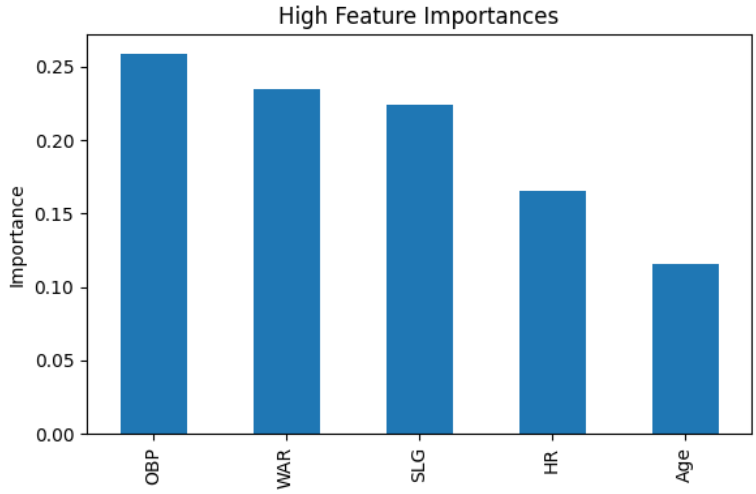
Low Feature Importance



Mid Feature Importance

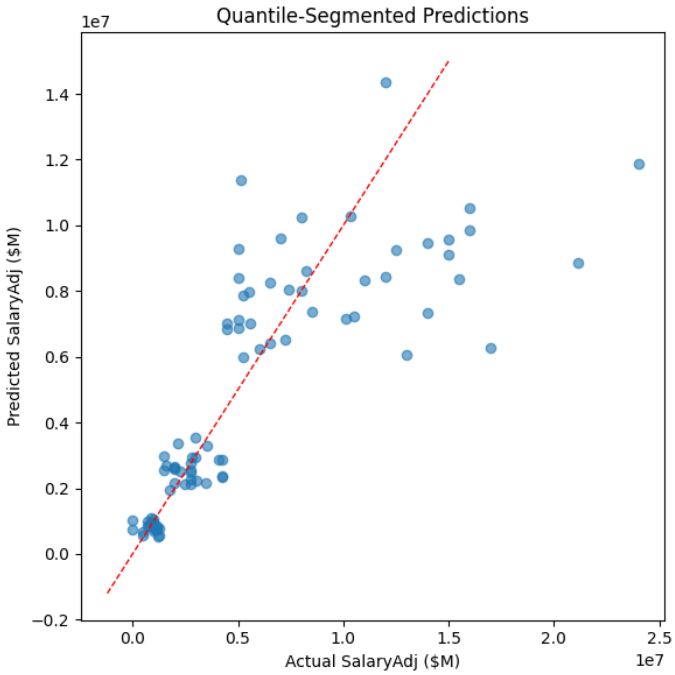


High Feature Importance



- Feature importances for Low/Mid/High segments show OBP, SLG, and WAR as top predictors; Age and HR are less impactful.

Predicted vs Actual Salaries



- Predicted vs. actual 2013 salaries (quantile-segmented model) show broad alignment ($R^2 = 0.62$) but heteroskedasticity (larger errors for high salaries).

4 | Interpretation & Insights

- **Segmented modelling** outperforms a global model (global $R^2 \approx 0.08$), capturing nonlinear salary dynamics across tiers.
- **Low/Mid tiers**: Negative R^2 indicates models perform worse than predicting the mean salary, likely due to limited variance and small sample sizes.
- **High tier**: Minimal predictive power ($R^2 = 0.04$), but still better than no segmentation.
- **Combined $R^2 = 0.62$** suggests tiered modeling is practically useful for mid/high-value free agents.
- **Key drivers**: OBP and SLG dominate, reflecting teams' emphasis on on-base and power metrics.

5 | Recommendations

1. **Improve Low/Mid models**:
    - Augment training data (e.g., minor-league stats, partial seasons).

- Incorporate additional features (postseason performance, positional flexibility).

2. **Blend models**:
    - Ensemble Random Forests with linear regression to balance bias-variance trade-offs.
3. **Diagnose residuals**:
    - Analyse under/over-prediction patterns by position, team market size, or player age.
4. **Enrich data**:
    - Integrate contract length, defensive metrics (e.g., DRS), and injury history.