Predictive Modeling of IPL Player Salary

Rohit Bhamidipati
Springboard Data Science Career Track
Capstone 2

Problem Statement

- The Indian Premier League (IPL) is the most viewed and fastest growing league in cricket.
- This has driven salaries for and the number of contracted players upwards.
- In turn, this has driven the need to understand the factors that are most impactful in determining the salary of the player.
- **Project Objective:** Develop a Machine Learning model to predict the player salary based on statistical aggregate features.

Summary of Results

- Strongest determining factor in a player’s salary is their previous year’s salary.
- Predicting the next-year salary is almost a one variable problem - depends on the cutoff.
- The trained models had metrics as follow, and indicate decent model performance.
  - Batting:
    - Mean Absolute Error: 0.252
    - R^2: 0.691
  - Bowling:
    - Mean Absolute Error: 0.265
    - R^2: 0.653
- A function was developed to predict player salary.
- As the IPL gets older, and more data becomes available, models will become better.

Data Sources

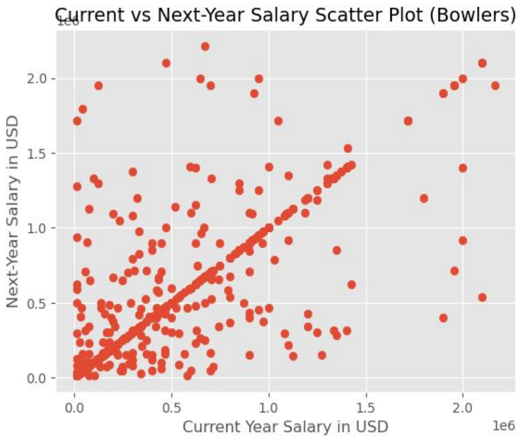
- IPL ball-by-ball data was sourced from Kaggle: IPL ball-by-ball dataset
- General player statistics were sourced from Cricmetric: Player Salary and General Player Statistics

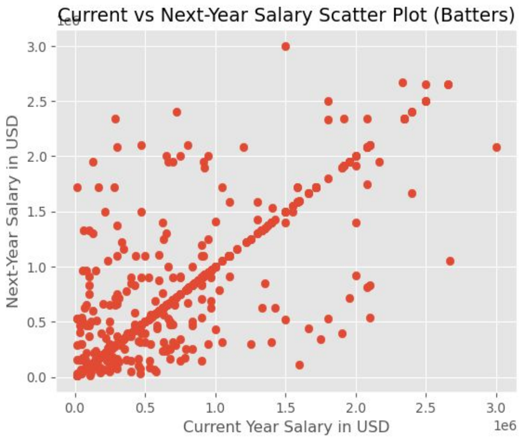
Wrangling and Feature Engineering

- The match data was integrated and filtered with the player data.
- The following features were engineered.
    - Batters: balls faced, total runs, batting average, strike rate, 50s, 100s, 4s, 6s.
    - Bowlers: balls bowled, total runs, total wickets, bowling average, economy, strike rate, 3-wicket games, 5-wicket games, dots, 4s, 6s.
- The following features were also considered.
    - Country, Role, Team.
    - match count, season count, previous year salary.

Exploratory Data Analysis

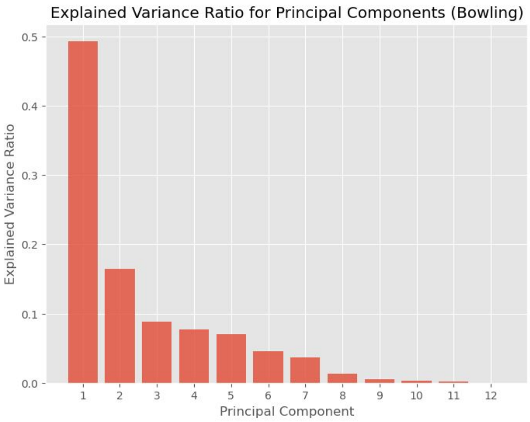
- Initial explorations of the data yielded the insight that most of the variability in the player salary is determined by the previous year's salary.

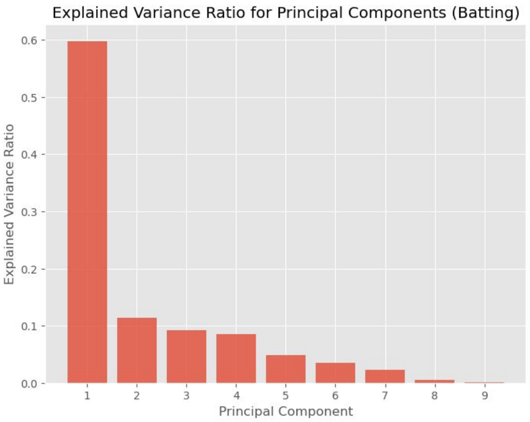




Exploratory Data Analysis (contd.)

- PCA revealed that most of the variability is captured by one variable, but the model can be fine-tuned by including more.



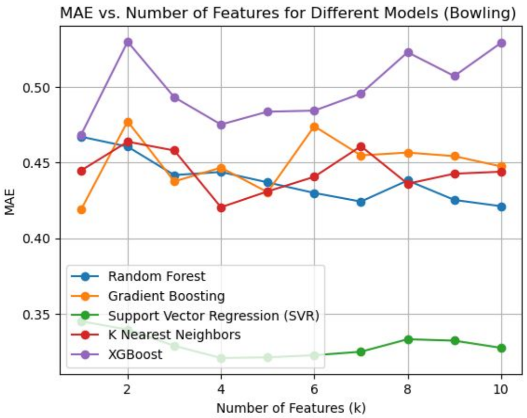


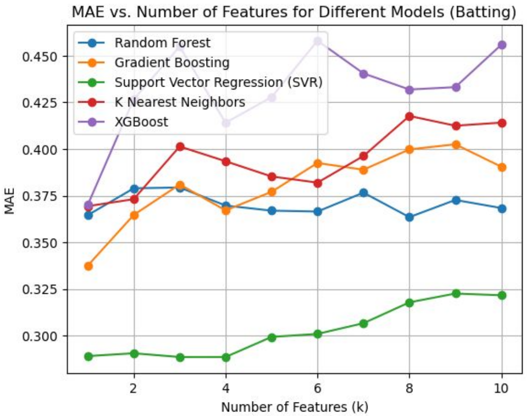
Model Selection: Metric

- Mean Absolute Error was chosen as the metric.
- MAE directly measures the average salary prediction error, which aligns with estimating player salaries accurately.
- MAE is less sensitive to outliers compared to other metrics because it treats all errors equally without squaring them.
- MAE is a linear metric, meaning that errors are considered proportionally.

Model Selection: Features and Models

- Various models were tested for their performance against the number of selected features.





Model Selection: Tuning and Evaluation

- For both bowling and batting, Support Vector Regression models with 4 features were chosen.
- The following optimal hyperparameters were found.
  - Batting: 'C' (Regularization parameter) set to 10, 'epsilon' (Epsilon parameter for margin of error) set to 0.01, and the 'kernel' chosen as 'linear'.
  - Bowling: 'C' (Regularization parameter) set to 1, 'epsilon' (Epsilon parameter for margin of error) set to 0.01, and the 'kernel' chosen as 'linear'.
- These parameter combinations yielded the following metrics.
  - Batting: R^2 score of 0.691, MAE score of 0.252.
  - Bowling: R^2 score of 0.653, and an MAE score of 0.265.
- These metrics indicate that the models have a good level of efficacy in predicting player salary.

Sample Predictions

The following sample predictions were made for some input data in either bowling or batting:

```
sample_bat_data = [
    'Shubman Gill', 'India', 0, 2023, 'Kolkata Knight Riders', 963501.60,
    17, 5, 564, 890, 59.33,
    157.80, 4, 3, 85, 33, 0,
    0
]
```

```
sample_bat_salary = predict_salary(sample_bat_data, 'batting')
```

```
sample_bat_salary
```

```
977323.0334347271
```

```
sample_bowl_data = [
    'Umran Malik', 'India', 0, 2022, 'Sunrisers Hyderabad', 481914.00,
    14, 2, 295, 444,
    22, 20.18, 9.03, 13.40, 1,
    1, 20, 20, 40, 0, 0
]
```

```
sample_bowl_salary = predict_salary(sample_bowl_data, 'bowling')
```

```
sample_bowl_salary
```

```
483044.7507882039
```

Further Questions

- How do player salaries and their determinants vary across different leagues?
- How does player performance in other leagues or international games affect their IPL salary?
- Can the model be refined to make separate predictions for the Mega-auctions (occurring once every four years) and Mini-auctions (occurring every year there is no Mega-auction)?
- How do more advanced models (such as deep learning models or ensemble methods) compare to the current SVR-based approach in terms of predictive power?
- Can NLP models be applied (such as analyzing ESPNcricinfo article sentiment) to account for player "hype" in predicting player salary?

Conclusion

- **Project Objective:** Develop a Machine Learning model to predict the player salary based on statistical aggregate features.
- Most of the variability in the player salary is determined by the previous year's salary.
- Mean Absolute Error was chosen as the metric.
- For both bowling and batting, Support Vector Regression models with 4 features were chosen.
    - Batting: R^2 score of 0.691, MAE score of 0.252.
    - Bowling: R^2 score of 0.653, and an MAE score of 0.265.
- These metrics indicate that the models have a good level of efficacy in predicting player salary.
- Salary for two sample players was predicted. Can be applied to other players provided their IPL statistics.
- More data would reveal more interactions- such as between mega/mini auctions and player salary.