DATA DOODLES

FIFA Player Wages

Members: Alina Huang, Pipat Gittisupab, Sandra Liz Sunny, Tram Nguyen, Tvisha Ronak Modi

Why This Matters

Goal: Predict the wages of FIFA soccer players using players' attributes: physical attributes, contract details, and technical skills.

Relevance: Help soccer clubs, analysts, and stakeholders make data-driven decisions on players' wages and the recruitment process.

Hypotheses:

- Players in their prime age (20-30) earn higher wages.
- Overall score is the strongest predictor of wages.
- Physical attributes (e.g., height, weight) have less impact than technical skills.

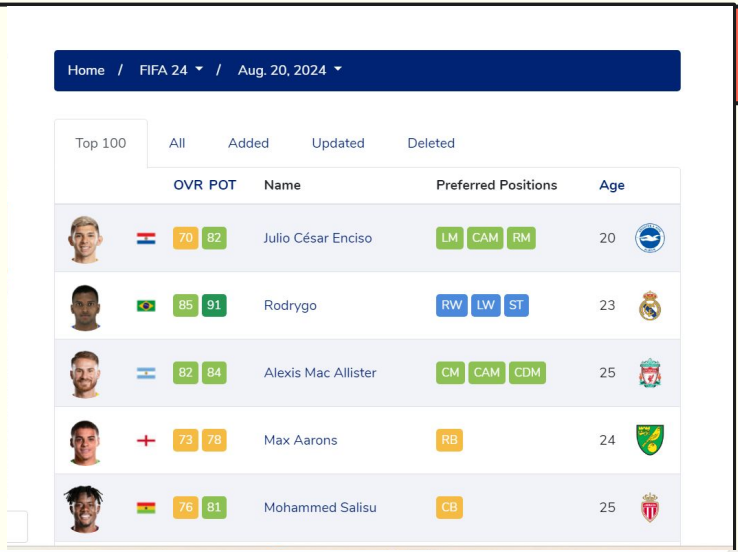
Exploring the Dataset

Source: Web-scraped FIFA data with 56 attributes.

Structure:

- Each row = 18240 players of 2024
- Each column = performance/characteristics

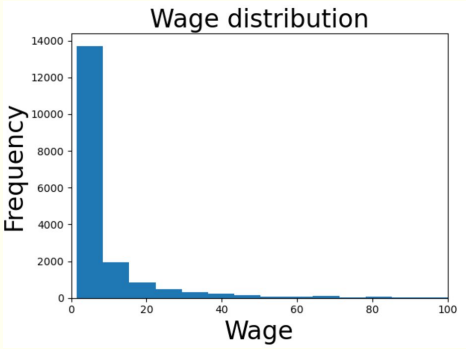
Key Features: Age, height, weight, overall score, technical and physical attributes, wages.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	name	overall_sc	position_s	height	weight	pref_foot	birthdate	age	pref_pos	work_rate	weak_foot	skill_move	value	wage	joined_clu	contract_e	Ball Contro	Dribbling	Marking	Slide Tackl	Stand Tack	Aggressior	Reactions
2	Erling Haal	91	94	195 cm	94 kg	Left	21-Jul-00	24	ST	High / Med	3	3	$157.000.(	$340.00	1-Jul-22	2027	82	79	None	29	47	87	94
3	Kylian Mba	91	94	182 cm	75 kg	Right	Dec. 20, 19	25	STLW	High / Low	4	5	$153.500.(	$225.00	1-Jul-18	2024	92	93	None	32	34	64	93
4	Kevin De B	91	91	181 cm	75 kg	Right	28-Jun-91	33	CMCAM	High / Med	5	4	$103.000.(	$350.00	Aug. 30, 20	2025	92	86	None	53	70	75	92
5	Harry Kane	90	90	188 cm	89 kg	Right	28-Jul-93	31	ST	High / High	5	3	$119.500.(	$230.00	28-Jul-10	2024	87	82	None	38	46	80	93
6	Thibaut Co	90	90	199 cm	96 kg	Left	########	32	GK	Medium / M	3	1	$63.000.0(	$250.00	Aug. 9, 201	2026	23	13	None	16	18	23	88
7	Robert Lew	90	90	185 cm	81 kg	Right	Aug. 21, 19	35	ST	High / Med	4	4	$58.000.0(	$340.00	18-Jul-22	2026	90	86	None	19	42	81	93
8	Karim Benz	90	90	185 cm	81 kg	Right	Dec. 19, 19	36	CFST	Medium / M	4	4	$51.000.0(	$95.00	1-Jul-23	2026	91	87	None	18	24	63	92
9	Lionel Mes	90	90	169 cm	67 kg	Left	24-Jun-87	37	CFCAM	Low / Low	4	4	$41.000.0(	$23.00	16-Jul-23	2025	93	96	None	24	35	44	88
10	RÃºben Dia	89	90	187 cm	82 kg	Right	########	27	CB	Medium / H	4	2	$97.500.0(	$250.00	Sept. 29, 2	2027	75	64	None	87	91	93	89

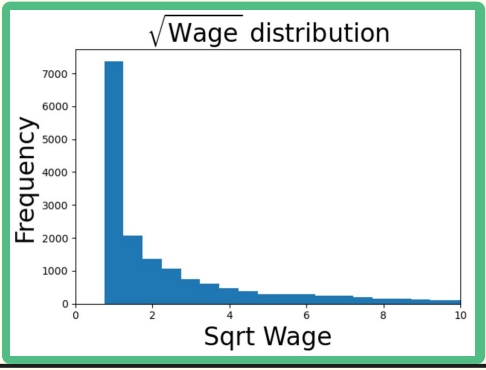
Data Preprocessing - Wage

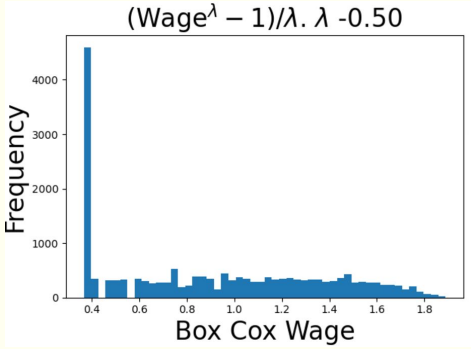
Initial Wage Distribution



Transformed Wage Distributions

The best!





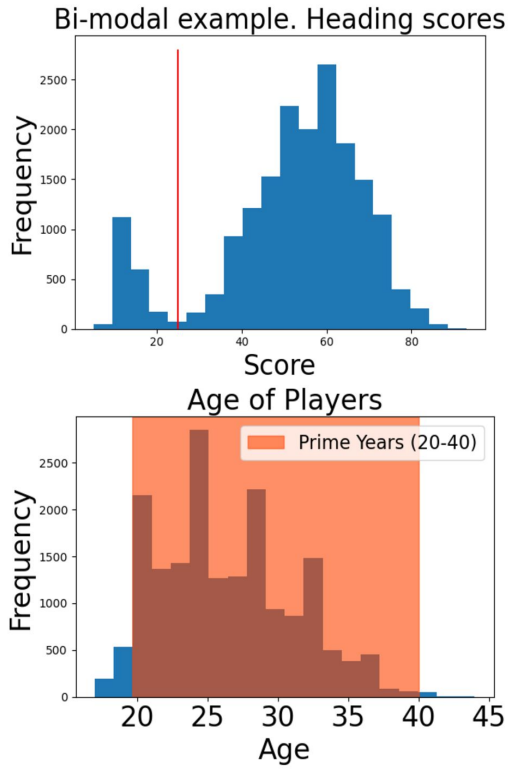
Creating New Features

HighScore = 0, if score < 25
HighScore = 1, otherwise

Then centered the scores with sklearn's Robust Scaler.

Prime = 1, if 20 < age <= 40
Prime = 0, otherwise

Helped the model!



Feature Selection

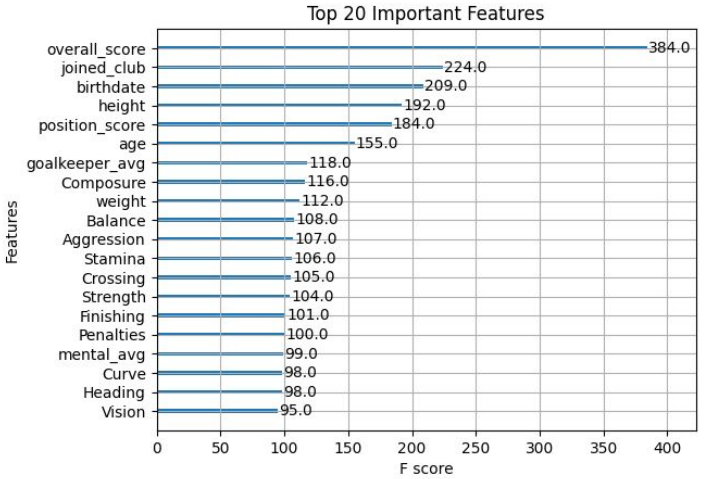
XGBoost after data preprocessing...
MSE: 90.91 → **76.73**
R squared: 0.73 → **0.77**
Dimensionality reduction using XGBoost to select top 20 features from 56.

```
[ ] X = data.drop(['name', 'wage'], axis=1)
    y = data['wage']

    # Split into training and testing sets
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
[ ] # Scale the features
    scaler = StandardScaler()
    X_train_scaled = scaler.fit_transform(X_train)
    X_test_scaled = scaler.transform(X_test)
```

```
[ ] model = xgb.XGBRegressor(random_state=42)
    model.fit(X_train_scaled, y_train)

    # Make predictions
    y_pred = model.predict(X_test_scaled)
```

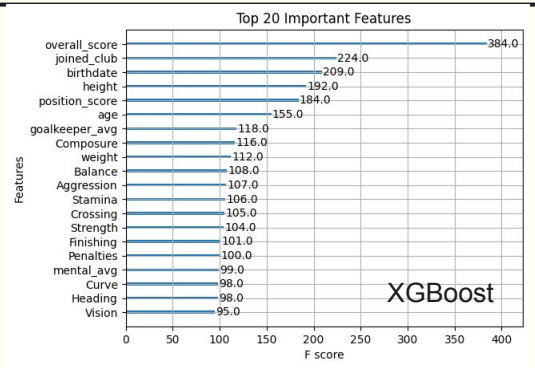
```
[ ] # Evaluate the model
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    mse, r2

    (90.90786482044345, 0.7279767794583916)
```

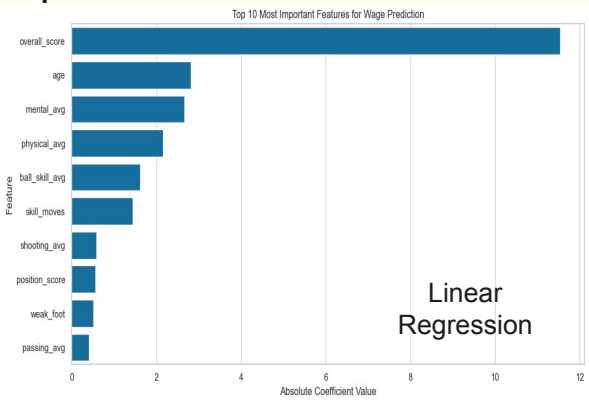


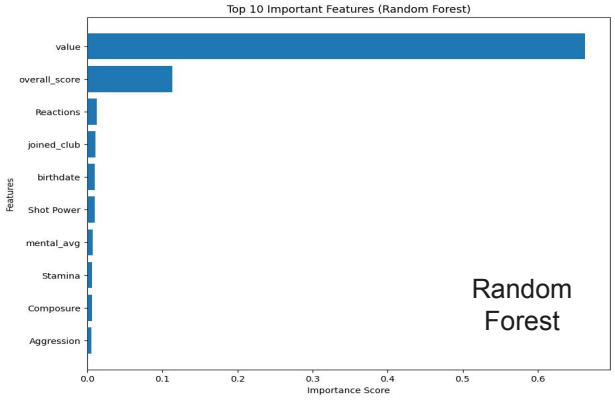
Building Predictive Models

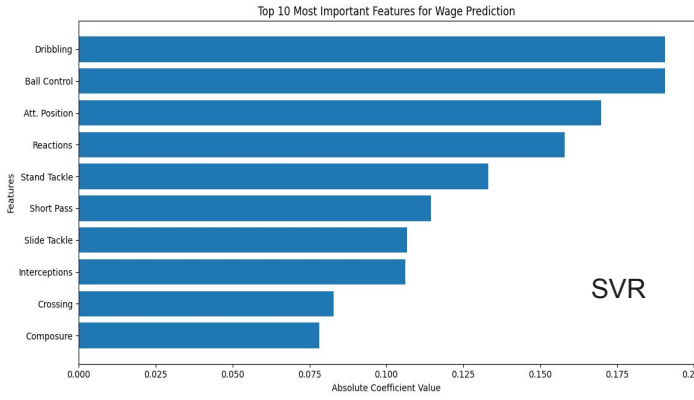
Models used: Linear Regression, Random Forest, XGBoost and SVR.

Evaluation metrics: R^2 and MSE.









Models Comparison

Linear regression	Random forest	SVR	XGBoost
Mean Squared Error: 200.03 R^2 Score: 0.40	Mean Squared Error: 109.73 R^2 Score: 0.67	Mean Squared Error: 99.86 R^2 Score: 0.70	Mean Squared Error: 76.73 R^2 Score: 0.77
Easy interpretability as coefficients correspond to factors	Provides robust predictions by aggregating outputs from multiple decision trees	Effective in High-dimensional spaces, and Robust to Outliers	Ability to handle non-linear relationships
Not effective in this case	Limited interpretability due to ensemble nature	Sensitivity to Hyperparameters	XGBoost performed best

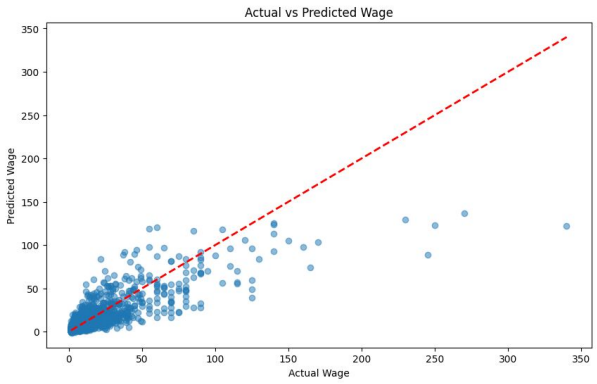
Lessons

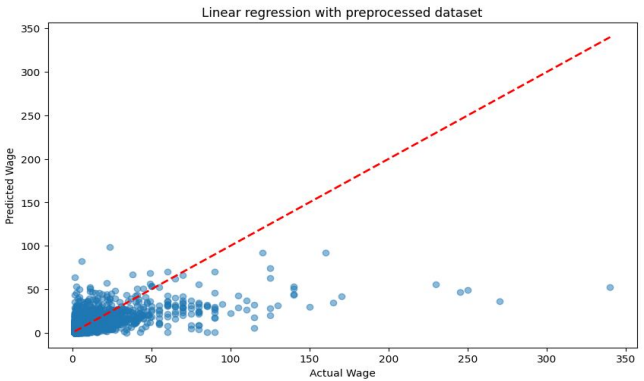
Challenges:

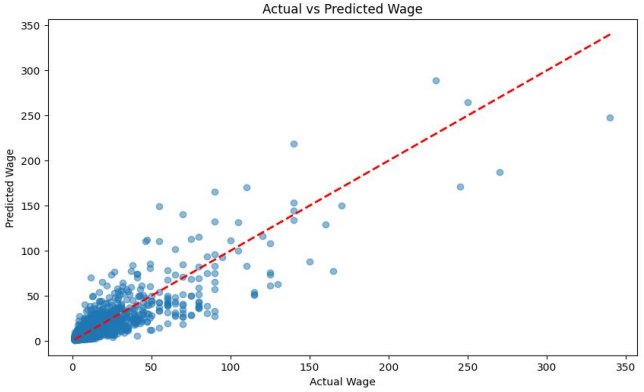
- Handling missing data and skewed distributions.
- Balancing feature relevance and redundancy.
- Dimensionality and redundancy.

Limitations:

- Dataset doesn't include external factors (e.g., sponsorships, team performance).
- Focused on limited seasonal data







Closing Thoughts

Summary:
- Hypotheses validated with data and modeling.
- Preprocessing and feature engineering improved model accuracy.
- XGBoost was the most effective predictive model.

Future Work:
- Enhance prediction accuracy
- Integrate additional data of the players

Applications:
- Wage negotiation
- Player recruitment
- Talent identification