MLB Analytics Project

By: Rohan Kalani

Agenda

1. Question and background
2. Model results and statistics
3. Testing model performance
4. Takeaways

.

Can a model be built to predict a pitcher’s earned runs in a game?

Description of statistics used

1. Earned Runs

- The number of runs a pitcher gives up in a game

2. Innings per game

- Number of innings a pitcher throws in a game (out of 9)

3. Walks per game

- Number of times in a game a pitcher allows a batter to reach first base by throwing four balls outside the strike zone

4. Strikeouts per game

- number of times in a game a pitcher records an out by striking out a batter

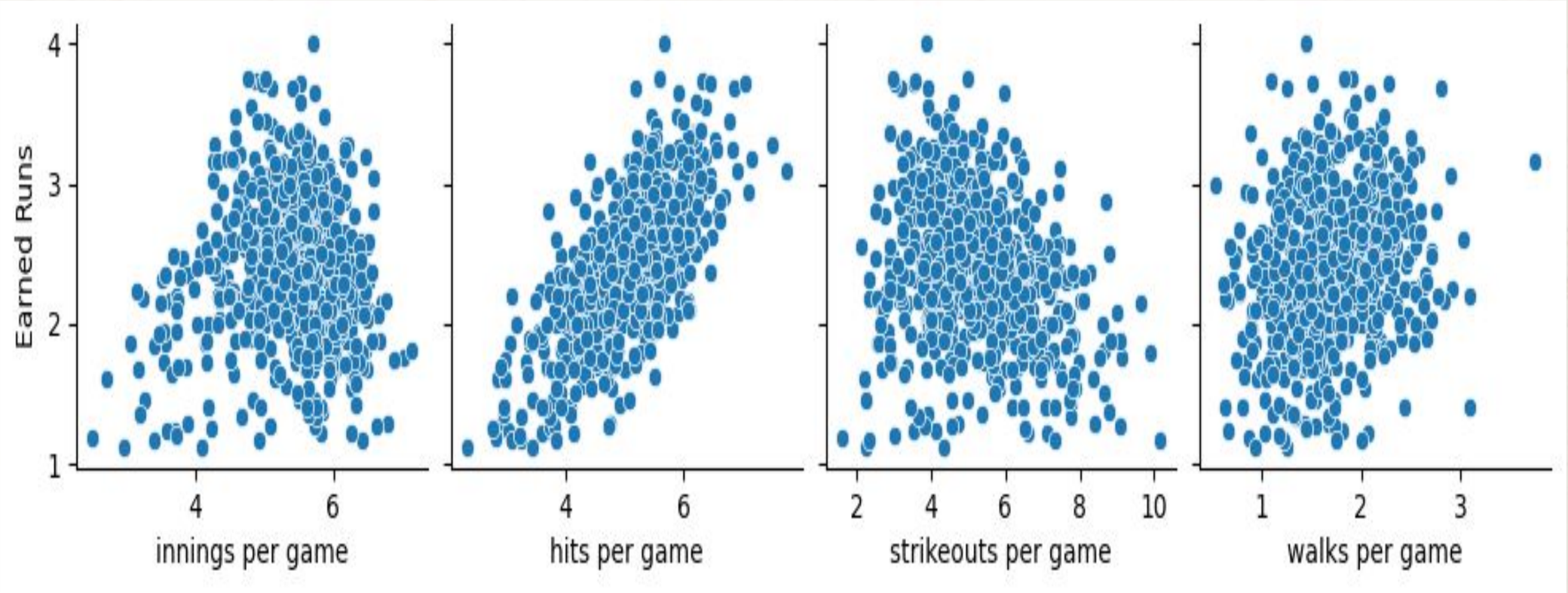
5. Hits per game

- number of times in a game a pitcher allows a batter to make contact with the ball and get on base

The Dataset

	last_name, first_name	player_id	year	hit	strikeout	walk	innings pitched	games	Earned Runs	innings per game	walks per game	strikeouts per game	hits per game
0	Colon, Bartolo	112526	2017	192	89	35	143.0	28	3.678571	5.107143	1.250000	3.178571	6.857143
1	Sabathia, CC	282332	2017	139	120	50	148.2	27	2.259259	5.488889	1.851852	4.444444	5.148148
2	Dickey, R.A.	285079	2017	193	136	67	190.0	31	2.903226	6.129032	2.161290	4.387097	6.225806
3	Lackey, John	407793	2017	165	149	53	170.2	31	2.806452	5.490323	1.709677	4.806452	5.322581
4	Wainwright, Adam	425794	2017	140	96	45	123.1	24	2.916667	5.129167	1.875000	4.000000	5.833333

Correlations



Regression Model

```
Predictmodel = sm.OLS(y_train, x_train).fit()
print(Predictmodel.summary())
```

```
                                 OLS Regression Results
==============================================================================
Dep. Variable:            Earned Runs   R-squared (uncentered):              0.986
Model:                            OLS   Adj. R-squared (uncentered):         0.986
Method:                 Least Squares   F-statistic:                     1.053e+04
Date:                Thu, 25 Apr 2024   Prob (F-statistic):                   0.00
Time:                        18:04:43   Log-Likelihood:                    -105.80
No. Observations:                 602   AIC:                                 219.6
Df Residuals:                     598   BIC:                                 237.2
Df Model:                           4
Covariance Type:            nonrobust
==========================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
innings per game      -0.2886      0.029     -9.899      0.000      -0.346      -0.231
hits per game          0.6274      0.020     31.271      0.000       0.588       0.667
strikeouts per game    0.0382      0.014      2.720      0.007       0.011       0.066
walks per game         0.3445      0.025     13.971      0.000       0.296       0.393
==============================================================================
Omnibus:                        1.455   Durbin-Watson:                   1.890
Prob(Omnibus):                  0.483   Jarque-Bera (JB):                1.276
Skew:                           0.096   Prob(JB):                        0.528
Kurtosis:                       3.118   Cond. No.                         28.5
==============================================================================
```

- 98.6% of the variability in ERA can be explained by the predictors included in the model
- All predictors are statistically significant

Other model statistics

Weight	Feature
1.7294 ± 0.4112	hits per game
0.4408 ± 0.0784	innings per game
0.1601 ± 0.0627	walks per game
0.0166 ± 0.0184	strikeouts per game

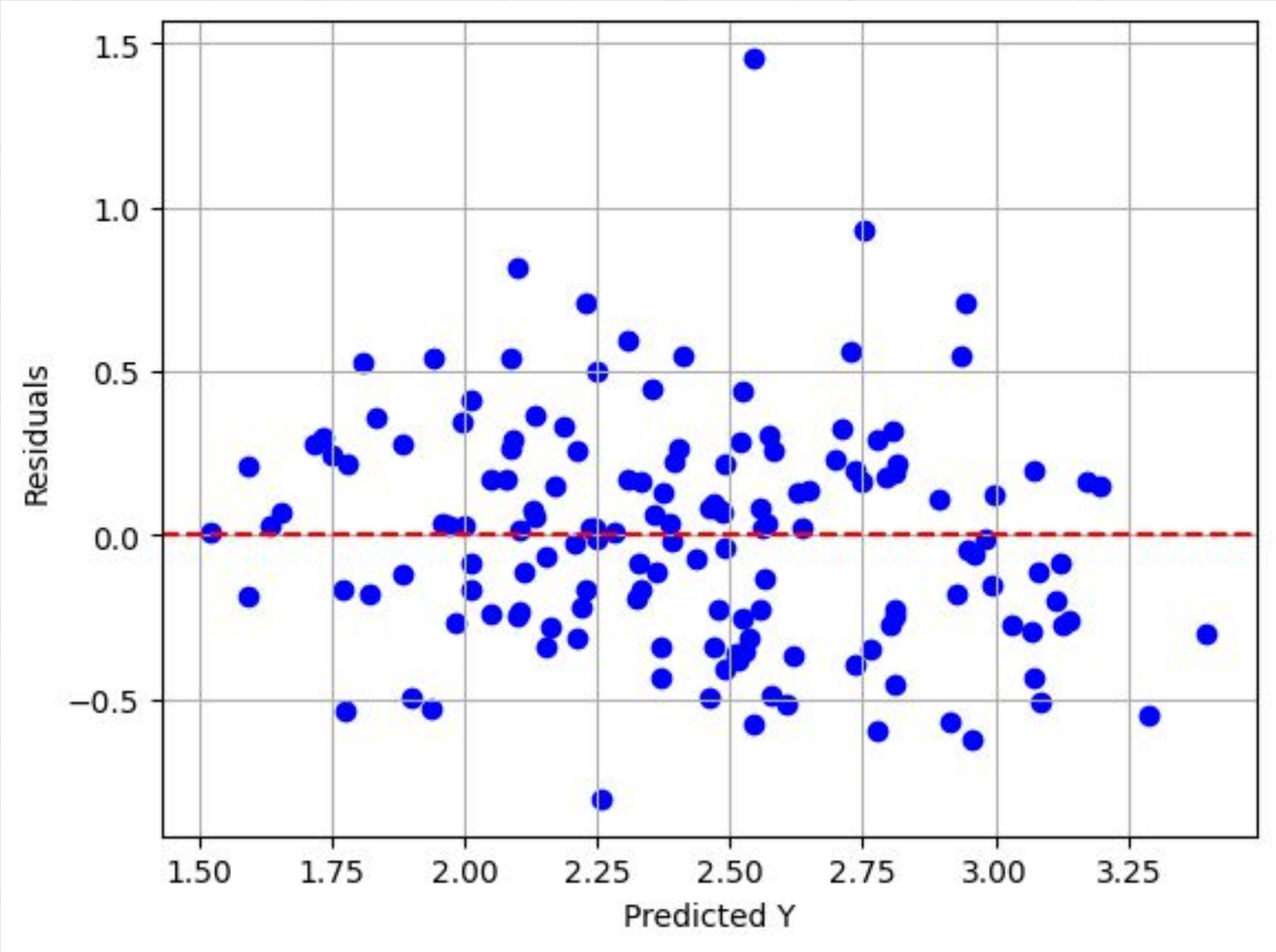
- Permutation Importance
  - Hits are the most important feature in the model
    - shuffling the hits feature results in an increase in the ERA by approximately 1.7294 runs.

```
[191] mse = mean_squared_error(y_test, y_pred)
      print("Root Mean Squared Error:", mse)
```

```
Root Mean Squared Error: 0.11808734678846378
```

- Root Mean Squared Error
  - Model's predictions deviate from true earned runs by approximately 0.118 runs

Residual Plot



- Residual plot is randomly scattered which means model can be used for prediction purposes

Evaluating model accuracy using 2024 pitcher data

Hunter Brown

HUNTER
BROWN

Houston Astros • #58 • Starting Pitcher

Follow

HT/WT	**6' 2", 220 lbs**
BIRTHDATE	**8/29/1998 (25)**
BAT/THR	**Right/Right**
BIRTHPLACE	**Detroit, MI**
STATUS	**Active**

News Stats Bio Splits Game Log Bat vs Pitch

Game Log

2024

2024 Regular Season

DATE	OPP	RESULT	IP	H	R	ER	HR	BB	K	GB	FB	P	TBF	GSC	DEC
Sun 4/21	@ WSH	L 6-0	4.0	4	3	3	0	2	6	5	4	84	17	46.0	L(0-4)
Tue 4/16	vs ATL	L 6-2	6.0	5	2	2	1	3	3	10	7	88	23	54.0	L(0-3)
Thu 4/11	@ KC	L 13-3	0.2	11	9	9	1	1	0	6	7	40	14	-7.0	L(0-2)
Fri 4/5	@ TEX	L 10-2	3.0	8	5	5	1	4	3	5	9	80	21	22.0	L(0-1)

Hunter Brown model prediction

```
[193] game_data = pd.DataFrame({'innings per game':[3], 'hits per game': [8], 'strikeouts per game':[3], 'walks per game': [4]})
      predicted_er = lm.predict(game_data)
      print("TEX predicted Earned Runs per game:", predicted_er)
```

```
TEX predicted Earned Runs per game: [[5.56372509]]
```

```
[194] game_data = pd.DataFrame({'innings per game':[.666], 'hits per game': [11], 'strikeouts per game':[0], 'walks per game': [1]})
      predicted_er = lm.predict(game_data)
      print("KC predicted Earned Runs per game:", predicted_er)
```

```
KC predicted Earned Runs per game: [[7.15594413]]
```

```
[195] game_data = pd.DataFrame({'innings per game':[6], 'hits per game': [5], 'strikeouts per game':[3], 'walks per game': [3]})
      predicted_er = lm.predict(game_data)
      print("ATL predicted Earned Runs per game:", predicted_er)
```

```
ATL predicted Earned Runs per game: [[2.4629256]]
```

```
[196] game_data = pd.DataFrame({'innings per game':[4], 'hits per game': [4], 'strikeouts per game':[6], 'walks per game': [2]})
      predicted_er = lm.predict(game_data)
      print("WSH predicted Earned Runs per game:", predicted_er)
```

```
WSH predicted Earned Runs per game: [[2.36097227]]
```

Ronel Blanco

RONEL
BLANCO

Houston Astros • #56 • Starting Pitcher

Follow

HT/WT	**6' 3", 265 lbs**
BIRTHDATE	**8/31/1993 (30)**
BAT/THR	**Right/Right**
BIRTHPLACE	**Santiago, Dominican Republic**
STATUS	**Active**

News Stats Bio Splits Game Log Bat vs Pitch

Game Log

2024

2024 Regular Season

DATE	OPP	RESULT	IP	H	R	ER	HR	BB	K	GB	FB	P	TBF	GSC	DEC
Sat 4/20	@ WSH	L 5-4 F/10	6.0	5	2	2	1	3	6	9	8	98	26	57.0	-
Sat 4/13	vs TEX	W 9-2	6.0	5	2	2	0	3	5	6	10	93	24	56.0	-
Sun 4/7	@ TEX	W 3-1	6.0	1	0	0	0	4	4	5	10	90	23	70.0	W(2-0)
Mon 4/1	vs TOR	W 10-0	9.0	0	0	0	0	2	7	12	8	105	29	92.0	W(1-0)

Ronel Blanco model predictions

```
game_data = pd.DataFrame({'innings per game':[9], 'hits per game': [0], 'strikeouts per game':[7], 'walks per game': [2]})
predicted_er = lm.predict(game_data)
print("TOR predicted Earned Runs per game:", predicted_er)
```

```
TOR predicted Earned Runs per game: [[-1.68365901]]
```

```
game_data = pd.DataFrame({'innings per game':[6], 'hits per game': [1], 'strikeouts per game':[4], 'walks per game': [4]})
predicted_er = lm.predict(game_data)
print("TEX predicted Earned Runs per game:", predicted_er)
```

```
TEX predicted Earned Runs per game: [[0.40122386]]
```

```
game_data = pd.DataFrame({'innings per game':[6], 'hits per game': [5], 'strikeouts per game':[5], 'walks per game': [3]})
predicted_er = lm.predict(game_data)
print("TEX predicted Earned Runs per game:", predicted_er)
```

```
TEX predicted Earned Runs per game: [[2.5380203]]
```

```
game_data = pd.DataFrame({'innings per game':[6], 'hits per game': [5], 'strikeouts per game':[6], 'walks per game': [3]})
predicted_er = lm.predict(game_data)
print("WSH predicted Earned Runs per game:", predicted_er)
```

```
WSH predicted Earned Runs per game: [[2.57556765]]
```

Christian Javier

CRISTIAN
JAVIER

Houston Astros • #53 • Starting Pitcher

Follow

HT/WT	**6' 1", 210 lbs**
BIRTHDATE	**3/26/1997 (27)**
BAT/THR	**Right/Right**
BIRTHPLACE	**Santo Domingo, Dominican Repu**
STATUS	**15-Day IL**

ews Stats Bio Splits Game Log

Game Log

2024

2024 Regular Season

DATE	OPP	RESULT	IP	H	R	ER	HR	BB	K	GB	FB	P	TBF	GSC	DEC
Sun 4/14	vs TEX	W 8-5	7.0	5	2	2	1	2	5	3	17	89	27	62.0	W(2-0)
Tue 4/9	@ KC	L 4-3 F/10	5.1	5	3	2	0	3	4	7	11	93	25	49.0	-
Wed 4/3	vs TOR	W 8-0	5.0	1	0	0	0	5	3	3	9	97	20	63.0	W(1-0)

Christian Javier model predictions

```
[209] game_data = pd.DataFrame({'innings per game':[5], 'hits per game': [1], 'strikeouts per game':[3], 'walks per game': [5]})
      predicted_er = lm.predict(game_data)
      print("TOR predicted Earned Runs per game:", predicted_er)
```

```
TOR predicted Earned Runs per game: [[0.99442238]]
```

```
[210] game_data = pd.DataFrame({'innings per game':[5.333], 'hits per game': [5], 'strikeouts per game':[4], 'walks per game': [3]})
      predicted_er = lm.predict(game_data)
      print("KC predicted Earned Runs per game:", predicted_er)
```

```
KC predicted Earned Runs per game: [[2.72594958]]
```

```
[211] game_data = pd.DataFrame({'innings per game':[7], 'hits per game': [5], 'strikeouts per game':[5], 'walks per game': [2]})
      predicted_er = lm.predict(game_data)
      print("TEX predicted Earned Runs per game:", predicted_er)
```

```
TEX predicted Earned Runs per game: [[1.90727443]]
```

Final Takeaways

Accuracy

The model is very accurate as shown by the past three pitchers and their game performances and the R squared

Predictive ability

The model can be altered in the future to predict a pitcher's next game performance. However, opponent data may need to be included