LONGITUDINAL IMPACT REPORTS AND PREDICTIONS:

UMD TERRAPINS BASEBALL

Group 4

TEAM

ISHA TYAGI

ZIDONG LIU

PIYALI BEDAGKAR

GNAPIKA KOMARAGIRI

MISSION STATEMENT

01 IDENTIFYING CORRELATED FEATURES

The mission is to meticulously analyze historical game data to identify features that exhibit significant correlations with game outcomes.

02 OPTIMIZING FEATURES TO IMPROVE IMPACTS AND OUTCOMES

Our mission is to leverage the understanding gained from correlated features to optimize key aspects of game preparation and execution.

03 UNDERSTANDING FUTURE GAME IMPROVEMENT

Our ultimate mission is to gain a comprehensive understanding of which features hold the most promise for improving future games.

OBJECTIVES

- To identify patterns and correlations between game outcomes and various features.
- Conduct Exploratory Data Analysis (EDA) reports.
- Generate statistical analyses matching the Year worksheet.
- Develop analytical models with reports

WORK FLOW

Analyze the UMTerps Baseball dataset to understand its structure and uncover insights through descriptive statistics and visualizations.

EDA

Develop predictive models to forecast game outcomes and optimize performance strategies based on historical data and key predictors.

ANALYTICAL MODELS

STAT REPORTS

Generate reports to analyze historical game data, extract key performance indicators, and identify trends and patterns in team performance.

01

EDA – EXPLORATORY DATA ANALYSIS

1) Win/Loss Percentage by Day of the Week
2) Game results by the Location

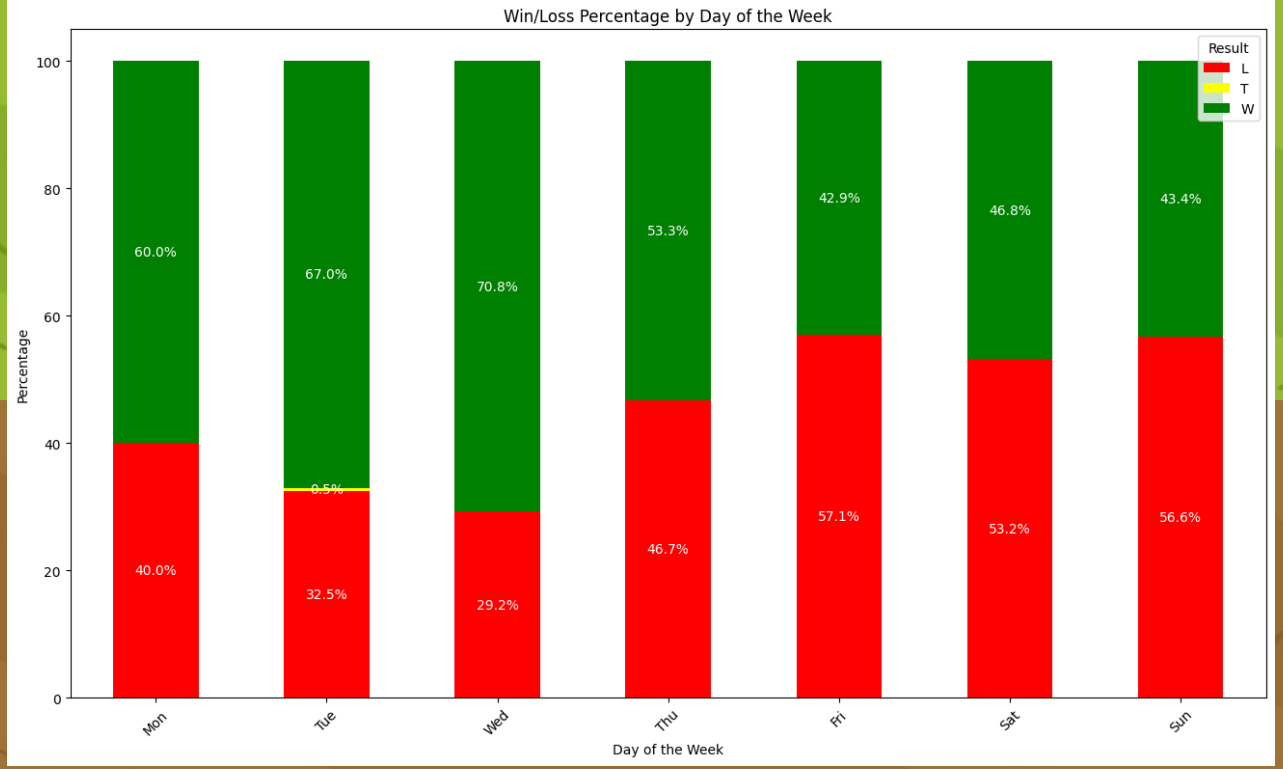
1) WIN/LOSS PERCENTAGE BY DAY OF THE WEEK

METHOD-

Processing: Utilized Python's pandas library to load and manipulate the dataset. Grouped the data by 'Day' column and counted the occurrences of 'Result' for each day. Reordered the DataFrame to match the correct order of weekdays.

ANALYSIS-

Plotted a stacked bar chart to visualize the win/loss percentage for each day of the week. The chart highlights the distribution of wins, losses, and ties across different days, providing insights into the team's performance patterns throughout the week.



INSIGHTS-

By this graph representation we can say that

- Win percentages are higher on weekdays than weekends.
- Wednesday has the highest win percentage at 70.8%.
- Sunday has the lowest win percentage at 29.2%.

2) GAME RESULTS BY THE LOCATION

METHOD:

Methodology: Utilized pandas library for data loading and manipulation, and matplotlib for visualization. Grouped and counted game results by location ('At') and result ('Result'). Employed pie charts to visualize the distribution of game results for each location type.

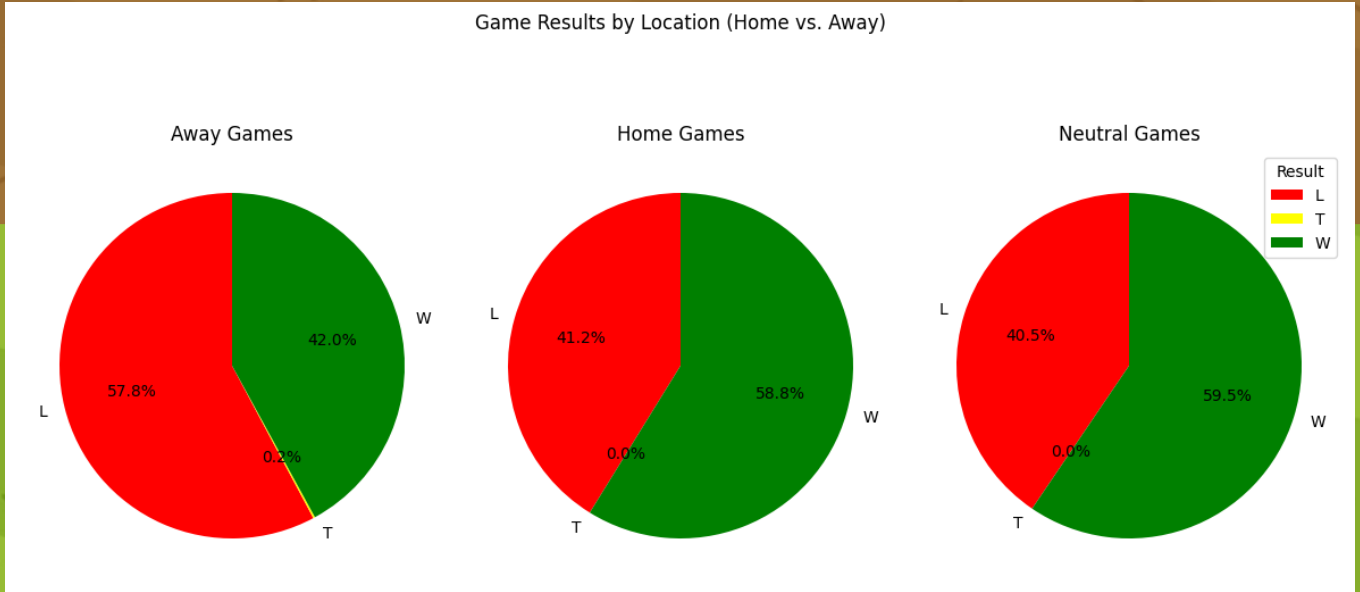
PROCESSING:

Data Processing: Loaded data from an Excel file. Grouped and counted game results by location and result. Utilized pie charts to visually represent the distribution of game results for each location type.

INSIGHTS-

By this graph representation we can say that

- Win percentages are higher on weekdays than weekends.
- Wednesday has the highest win percentage at 70.8%.
- Sunday has the lowest win percentage at 29.2%.



02

STAT REPORTS

1) Average Scores Trends
2) Win Rate Calculation
3) Win Rate Over time by location

1) AVERAGE SCORES TRENDS

PROCESSING:

- DataFrame Output: Displayed in console, containing average scores for 'Terps' and 'Oppnt' per year.
- Structure: Indexed by years, with columns for 'Terps' and 'Oppnt' average scores.
- Example Output: Illustrates average scores per year, aiding trend analysis.

Year	Terps	Oppnt
1999	7.041667	7.583333
2000	6.617021	6.808511
2001	6.520000	8.780000
2002	8.603774	6.962264
2003	5.358491	7.528302
2004	5.857143	6.946429
2005	5.701754	6.368421
2006	5.017857	6.321429
2007	5.767857	5.946429
2008	6.500000	6.214286
2009	6.166667	6.462963
2010	4.517857	8.053571
2011	4.375000	5.946429
2012	4.928571	3.803571
2013	5.563636	4.618182
2014	5.333333	4.142857
2015	5.969697	4.196970
2016	5.000000	4.543860
2017	6.114754	4.721311
2018	5.018519	5.796296
2019	5.620690	6.327586
2020	7.400000	4.733333
2021	6.604167	5.312500
2022	9.225806	5.596774
2023	9.174603	6.841270

METHOD:

Visual Representation: Plot depicts trends in 'Terps' and opponents' average scores from 1999 to 2023.

Axes: X-axis represents years, y-axis denotes average scores.

Lines: 'Terps' average scores marked with circle markers, opponents' with 'x' markers.

Title: "Average Scores Trend (1999-2023)" clarifies plot content and time frame.

Axis Labels: "Year" on x-axis, "Average Score" on y-axis.

Legend: Distinguishes between 'Terps' and opponents' scores.

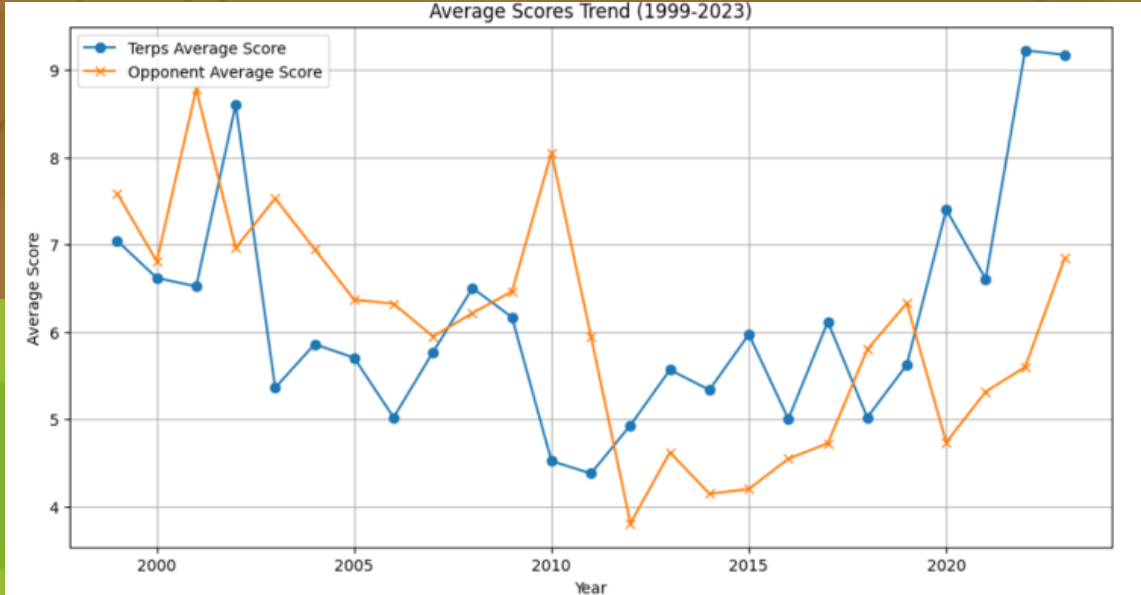
Grid: Enhances readability by providing visual guidance.

Year	Terps	Oppnt
1999	7.041667	7.583333
2000	6.617021	6.808511
2001	6.520000	8.780000
2002	8.603774	6.962264
2003	5.358491	7.528302
2004	5.857143	6.946429
2005	5.701754	6.368421
2006	5.017857	6.321429
2007	5.767857	5.946429
2008	6.500000	6.214286
2009	6.166667	6.462963
2010	4.517857	8.053571
2011	4.375000	5.946429
2012	4.928571	3.803571
2013	5.563636	4.618182
2014	5.333333	4.142857
2015	5.969697	4.196970
2016	5.000000	4.543860
2017	6.114754	4.721311
2018	5.018519	5.796296
2019	5.620690	6.327586
2020	7.400000	4.733333
2021	6.604167	5.312500
2022	9.225806	5.596774
2023	9.174603	6.841270

ANALYSIS

The UMTerps Baseball team's performance shows a clear trajectory of improvement over the 25-year span:

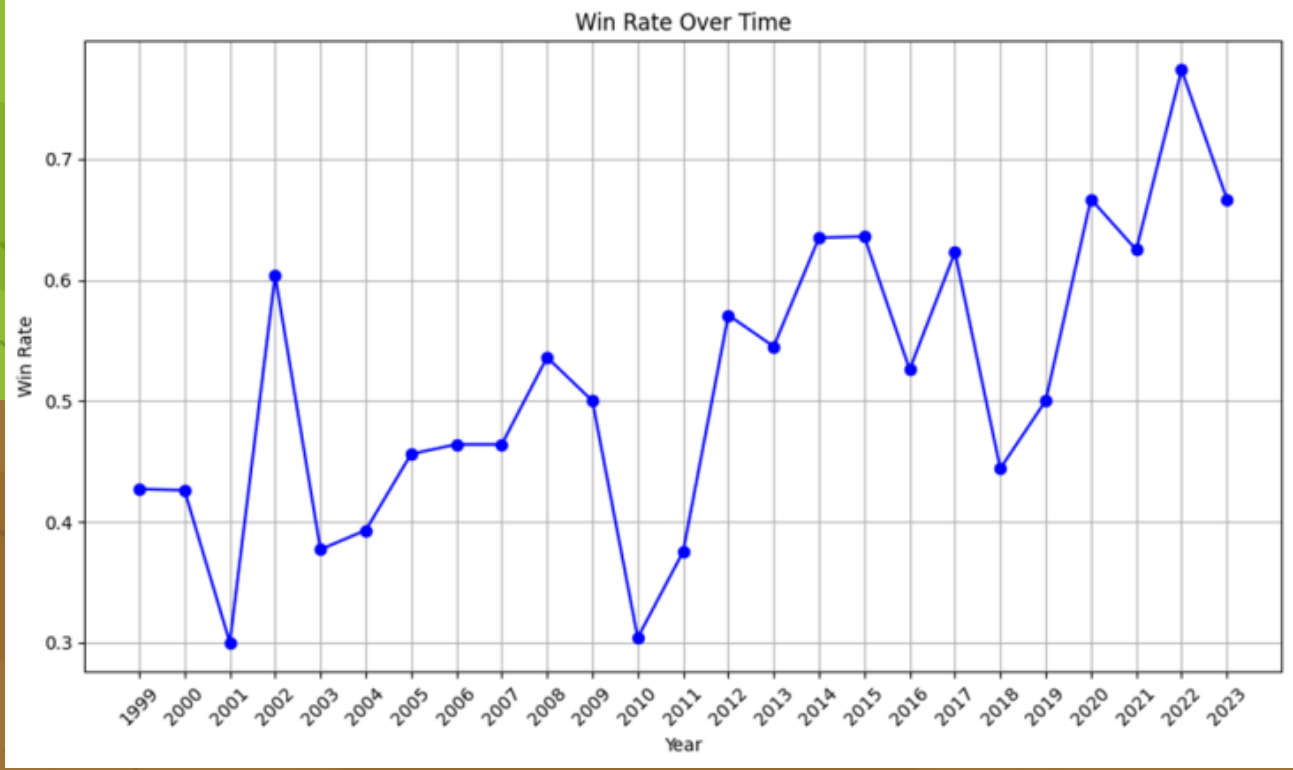
- **Early years**: Characterized by lower win rates and inconsistency.
- **Middle years**: Showed gradual and significant improvements.
- **Recent years**: Demonstrated peak performance with high and consistent win rates, suggesting successful strategies, improved team management, or stronger player development.



Result	Year	Win_Rate
0	1999	0.427
1	2000	0.426
2	2001	0.300
3	2002	0.604
4	2003	0.377
5	2004	0.393
6	2005	0.456
7	2006	0.464
8	2007	0.464
9	2008	0.536
10	2009	0.500
11	2010	0.304
12	2011	0.375
13	2012	0.571
14	2013	0.545
15	2014	0.635
16	2015	0.636
17	2016	0.526
18	2017	0.623
19	2018	0.444
20	2019	0.500
21	2020	0.667
22	2021	0.625
23	2022	0.774
24	2023	0.667

2) WIN RATE CALCULATION

- The win rate for each year is computed by adjusting the number of wins and ties (if any) and then dividing by the total number of games played.
- For example, if in a particular year the team won 20 games, lost 10, and tied 5, the adjusted wins would be 20 + 0.5 * 5 = 22.5. If the total number of games is 35, the win rate would be 22.5 / 35 = 0.643, rounded to three decimal places..



ANALYSIS

The UMTerps Baseball team's performance shows a clear trajectory of improvement over the 25-year span:

- **Early years**: Characterized by lower win rates and inconsistency.
- **Middle years**: Showed gradual and significant improvements.
- **Recent years**: Demonstrated peak performance with high and consistent win rates, suggesting successful strategies, improved team management, or stronger player development.

3) WIN RATE OVER TIME BY LOCATION

- The outcome presented in the table compares two sets of data: game results from the 'Game' sheet and season summaries from the 'Year' sheet in an Excel file.
- The resulting DataFrame includes yearly performance metrics, verifying the accuracy and consistency between the two sources.

ANALYSIS

- The table provides a thorough comparison between individual game data and summarized season records, ensuring data accuracy and consistency.
- It allows the university or analysts to verify that their records are correctly maintained and reflects the performance trends of the UMTerps baseball team over the years.

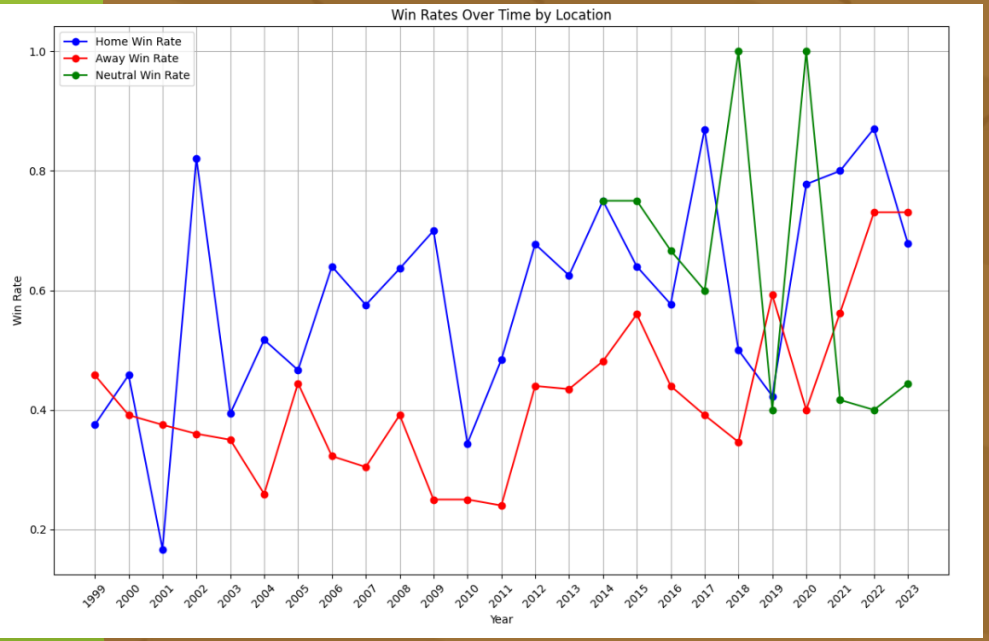
	Year	Overall	WLT	OverallP	Win_Rate	Same
0	1999	20-27-1	20-27-1	0.427	0.427	True
1	2000	20-27	20-27	0.426	0.426	True
2	2001	15-35	15-35	0.3	0.3	True
3	2002	32-21	32-21	0.604	0.604	True
4	2003	20-33	20-33	0.377	0.377	True
5	2004	22-34	22-34	0.393	0.393	True
6	2005	26-31	26-31	0.456	0.456	True
7	2006	26-30	26-30	0.464	0.464	True
8	2007	26-30	26-30	0.464	0.464	True
9	2008	30-26	30-26	0.536	0.536	True
10	2009	27-27	27-27	0.5	0.5	True
11	2010	17-39	17-39	0.304	0.304	True
12	2011	21-35	21-35	0.375	0.375	True
13	2012	32-24	32-24	0.571	0.571	True
14	2013	30-25	30-25	0.545	0.545	True
15	2014	40-23	40-23	0.635	0.635	True
16	2015	42-24	42-24	0.636	0.636	True
17	2016	30-27	30-27	0.526	0.526	True
18	2017	38-23	38-23	0.623	0.623	True
19	2018	24-30	24-30	0.444	0.444	True
20	2019	29-29	29-29	0.5	0.5	True
21	2020	10-5	10-5	0.667	0.667	True
22	2021	30-18	30-18	0.625	0.625	True
23	2022	48-14	48-14	0.774	0.774	True
24	2023	42-21	42-21	0.667	0.667	True

	Year	Home	Home_WLT	Away	Away_WLT	Neutral	Neutral_WLT
0	1999	9-15-0	9-15	11-12-1	11-12-1	0-0-0	NaN
1	2000	11-13	11-13	9-14	9-14	0-0	NaN
2	2001	3-15	3-15	12-20	12-20	0-0	NaN
3	2002	23-5	23-5	9-16	9-16	0-0	NaN
4	2003	13-20	13-20	7-13	7-13	0-0	NaN
5	2004	15-14	15-14	7-20	7-20	0-0	NaN
6	2005	14-16	14-16	12-15	12-15	0-0	NaN
7	2006	16-9	16-9	10-21	10-21	0-0	NaN
8	2007	19-14	19-14	7-16	7-16	0-0	NaN
9	2008	21-12	21-12	9-14	9-14	0-0	NaN
10	2009	21-9	21-9	6-18	6-18	0-0	NaN
11	2010	11-21	11-21	6-18	6-18	0-0	NaN
12	2011	15-16	15-16	6-19	6-19	0-0	NaN
13	2012	21-10	21-10	11-14	11-14	0-0	NaN
14	2013	20-12	20-12	10-13	10-13	0-0	NaN
15	2014	21-7	21-7	13-14	13-14	6-2	6-2
16	2015	16-9	16-9	14-11	14-11	12-4	12-4
17	2016	15-11	15-11	11-14	11-14	4-2	4-2
18	2017	20-3	20-3	9-14	9-14	9-6	9-6
19	2018	13-13	13-13	9-17	9-17	2-0	2-0
20	2019	11-15	11-15	16-11	16-11	2-3	2-3
21	2020	7-2	7-2	2-3	2-3	1-0	1-0
22	2021	16-4	16-4	9-7	9-7	5-7	5-7
23	2022	27-4	27-4	19-7	19-7	2-3	2-3
24	2023	19-9	19-9	19-7	19-7	4-5	4-5

- The output presents a summary of the University of Maryland Terrapins (UMTerps) baseball team's performance from 1999 to 2023, categorized by game locations (home, away, and neutral sites).

CONCLUSION

- The output provides a comprehensive overview of the UMTerps baseball team's performance across different game locations throughout the years.
- It allows stakeholders to identify trends, assess strengths and weaknesses in performance based on location, and make informed decisions to improve the team's overall competitiveness.



- The generated graph shows the win rates over time for UMD baseball matches at home, away, and neutral locations.

BASED ON THE TRENDS OBSERVED:

- **Home win rate**: Expected to remain stable, barring major changes in team dynamics or venue conditions.
- **Away win rate**: Likely to stabilize further, with potential for improvement through strategic adjustments.
- **Neutral win rate**: May fluctuate, but focus on improving performance in neutral venues could lead to consistency.

03

ANALYTICAL MODELS

Model 1 – Logistic Regression
Model 2 – Random Forest Classifier

MODEL 1 – LOGISTIC REGRESSION

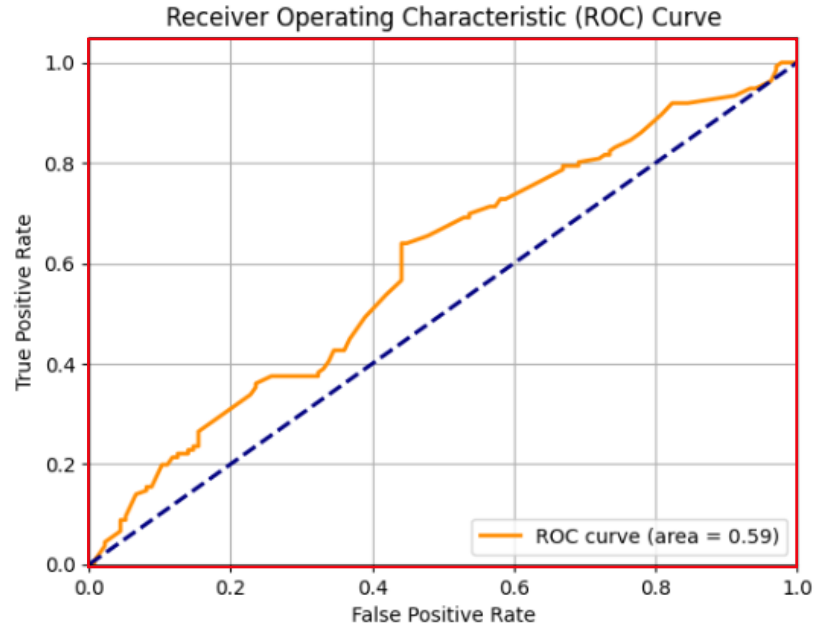
Performed a binary classification task using logistic regression to predict the outcome of baseball games based on certain features.

- **Data Preprocessing**: The 'Result' column is converted to binary, where 'W' (win) is encoded as 1 and 'L' (loss) as 0.
- **Categorical variables** 'Day' and 'At' are encoded using LabelEncoder, which assigns a unique numerical value to each category.
- 'TBA' values in the 'Time' column are replaced with the median hour of the day after converting them to the 24-hour format.
- **Features and Target Variable**: Features (X) include 'Day', 'Time', and 'At'.
- certain features.

```
Model Coefficients: [[0.19031406 0.02293814 0.60343236]]
Model Intercept: [-1.18065173]
Accuracy: 0.5955882352941176
Precision: 0.589041095890411
Recall: 0.6323529411764706
Classification Report:
              precision    recall  f1-score   support

           0       0.60      0.56      0.58       136
           1       0.59      0.63      0.61       136

    accuracy                           0.60       272
   macro avg       0.60      0.60      0.60       272
weighted avg       0.60      0.60      0.60       272
```




```
Predicted outcomes: [0 1 1 0 1]
Predicted probabilities of winning: [0.49837277 0.5229969  0.56449579 0.4950264  0.50410721]
```

INSIGHTS-

- The ROC curve is plotted with an area under the curve (AUC) of approximately 0.60. This suggests that the model has a fair discrimination ability in distinguishing between positive and negative instances.
- The curve is above the diagonal dashed line, indicating that the model performs better than random guessing.
- The curve leans towards the upper left corner, suggesting that the model has a higher true positive rate compared to the false positive rate across various thresholds.

MODEL 2 - RANDOM FOREST CLASSIFIER

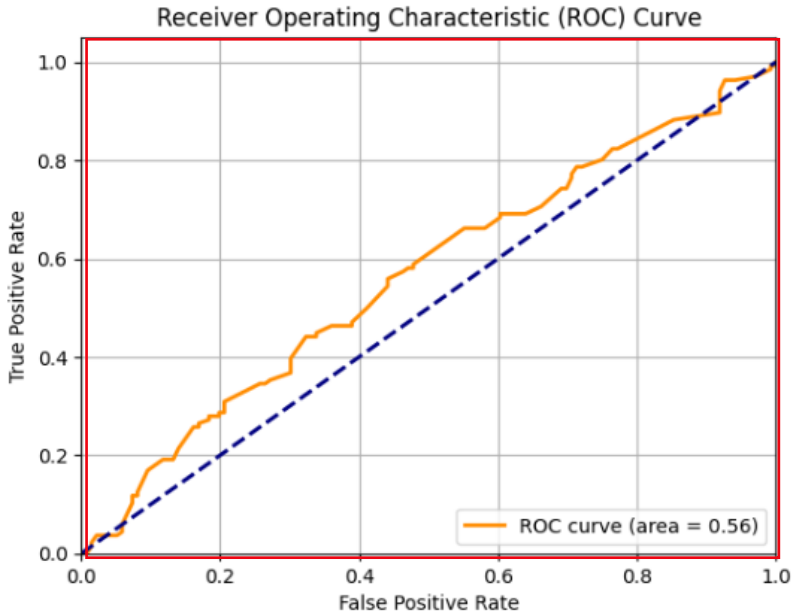
Performed several tasks related to building, training, and evaluating a Random Forest classifier model for predicting the outcomes of baseball games based on certain features.

- **Data Preprocessing**: The 'Result' column is converted to binary, where 'W' is encoded as 1 (Win) and other values as 0 (Loss).
- **Categorical variables** 'Day' and 'At' are encoded using LabelEncoder to convert them into numerical values.
- 'TBA' values in the 'Time' column are replaced with the median hour of the day, which is calculated from the existing time data.
- **Features and Target**: Features (X) and target variable (y) are defined. Features include 'Day', 'Time', and 'At', while the target is 'Result'.

```
Accuracy: 0.5441176470588235
Precision: 0.5483870967741935
Recall: 0.5
Classification Report:
              precision    recall  f1-score   support

           0       0.54      0.59      0.56       136
           1       0.55      0.50      0.52       136

    accuracy                           0.54       272
   macro avg       0.54      0.54      0.54       272
weighted avg       0.54      0.54      0.54       272
```




```
Predicted outcomes: [1 0 0 1 0]
Predicted probabilities of winning: [0.7005293  0.15148779 0.07       0.75141405 0.23242424]
```

INSIGHTS-

- The model seems to perform moderately well, with an accuracy of around 54%.
- The precision and recall are also around 55% and 50%, respectively, indicating a balance between correctly identifying positive cases and minimizing false positives.

FUTURE WORK

ADVANCED PREDICTIVE MODELING

Implement sophisticated techniques like time series analysis to enhance game outcome predictions.

INTEGRATION OF ADDITIONAL DATA SOURCES

Explore incorporating player statistics and weather conditions to refine analytical models.

INTERACTIVE DASHBOARDS

Develop user-friendly dashboards for real-time performance tracking and decision-making.

THANKYOU