GA-DSI-123

SUMO Match Outcome Prediction

Ayako Homma

April 17, 2023

Sumo wrestling is a sport that originated in Japan, with a history spanning centuries. It has become an integral part of Japanese culture and tradition, captivating audiences with its unique combination of athleticism and strategy.

相撲

Problem Statement

In this project, I will be analyzing data on sumo wrestlers to predict the outcome of matches. The dataset contains information on wrestlers' physical characteristics such as their height and weight, as well as details on each wrestler's rank and the result of each tournament match.

My goal is to use machine learning algorithms, such as logistic regression, decision trees, and random forests to build a prediction model that will achieve an accuracy score of at least 0.75.

This model will be valuable for both fans and practitioners of sumo wrestling, providing insights into the key factors that contribute to a wrestler's success or failure.

Additionally, a Tableau dashboard will be created to visualize the data and predictions, allowing sumo fans and practitioners to explore and interact with the data in a meaningful way.

Key steps in research process

01. Data Collection

02. Data Cleaning

03. EDA & Feature Engineering

04. Data Modeling & Evaluation

05. Conclusion & Recommendations

01.

DATA COLLECTION

The project's dataset was created by merging two different sets of data

A dataset

Size (226,590 x 29)

The below data are found in both datasets

- Basho (Tournament)
- Wrestler's ID
- Wrestler's Name
- Wrestler's Rank

Banzuke.csv

(Size: 176,000 x 12)

Results.csv

(Size: 226,590 x 13)

Source: Data World

The dataset covers a comprehensive range of wrestlers and tournaments information

Wrestler 1
ID
name
rank
hometown
birth date
height
weight
Previous tournament results

basho
day

match outcome (win or loss)
kimarite
results at the time of the tournament
final record at the end of the tournament

Wrestler 2 (opponent)
ID
name
rank
hometown
birth date
height
weight
Previous tournament results

02.

DATA CLEANING

31 missing values were identified in a fairly clean dataset

31 missing values are identified under weight and height

```
# Check for missing values
df.isnull().sum().sort_values(ascending=False)
```

```
r2_weight    31
r2_height    31
r1_weight    31
r1_height    31
```

These missing values are from two wrestlers: **Takeuchi** and **Miyabiyama**

```
df.loc[df['r1_height'].isnull()]
```

	basho	day	r1_id	r1_rank	r1_shikona
89652	1998.09	13	842	Ms6w	Takeuchi
90425	1998.11	8	842	J11w	Miyabiyama

Missing values were handled based on the available data for each wrestler

Miyabiyama

basho	r1_shikona	r1_height	r1_weight
1998.11	Miyabiyama	NaN	NaN
1999.01	Miyabiyama	NaN	NaN
1999.03	Miyabiyama	187.7	171.0
1999.05	Miyabiyama	187.7	171.0
1999.07	Miyabiyama	187.7	171.0
1999.09	Miyabiyama	187.7	171.0
1999.11	Miyabiyama	187.7	171.0
2000.01	Miyabiyama	187.7	171.0
2000.03	Miyabiyama	188.0	175.5
2000.05	Miyabiyama	188.0	175.5
2000.07	Miyabiyama	188.0	175.5

Miyabiyama's height and weight information were missing in Nov 1998 and Jan 1999. However, his info was recorded in Mar 1999, and remained unchanged until Mar 2003. The data from the Mar 1999 tournament was used to fill in the missing values.

Takeuchi

basho	r1_shikona	r1_height	r1_weight
1998.09	Takeuchi	NaN	NaN

Takeuchi only had one record of his tournament match, which had the missing values. This record was dropped from the dataset.

03.

EDA & FEATURE ENGINEERING

#	Column	Non-Null Count	Dtype
0	basho	226588 non-null	float64
1	day	226588 non-null	int64
2	r1_id	226588 non-null	int64
3	r1_rank	226588 non-null	object
4	r1_shikona	226588 non-null	object
5	r1_result	226588 non-null	object
6	r1_win	226588 non-null	int64
7	kimarite	226588 non-null	object
8	r2_id	226588 non-null	int64
9	r2_rank	226588 non-null	object
10	r2_shikona	226588 non-null	object
11	r2_result	226588 non-null	object
12	r1_heya	226588 non-null	object
13	r1_shusshin	226588 non-null	object
14	r1_birth_date	226588 non-null	object
15	r1_height	226588 non-null	float64
16	r1_weight	226588 non-null	float64
17	r1_prev	226588 non-null	object
18	r1_prev_w	226588 non-null	float64
19	r1_prev_l	226588 non-null	float64
20	r2_heya	226588 non-null	object
21	r2_shusshin	226588 non-null	object
22	r2_birth_date	226588 non-null	object
23	r2_height	226588 non-null	float64
24	r2_weight	226588 non-null	float64
25	r2_prev	226588 non-null	object
26	r2_prev_w	226588 non-null	float64
27	r2_prev_l	226588 non-null	float64

Feature engineering process was performed

- Convert categorical features with numeric values
  - shusshin (hometown)
  - wrestler ranks
  - kimarite (winning techniques)
  - heya (organization/clubs)
- Calculations
  - Age
  - Number of wins in
    - the previous tournament
    - the current tournament

TABLEAU

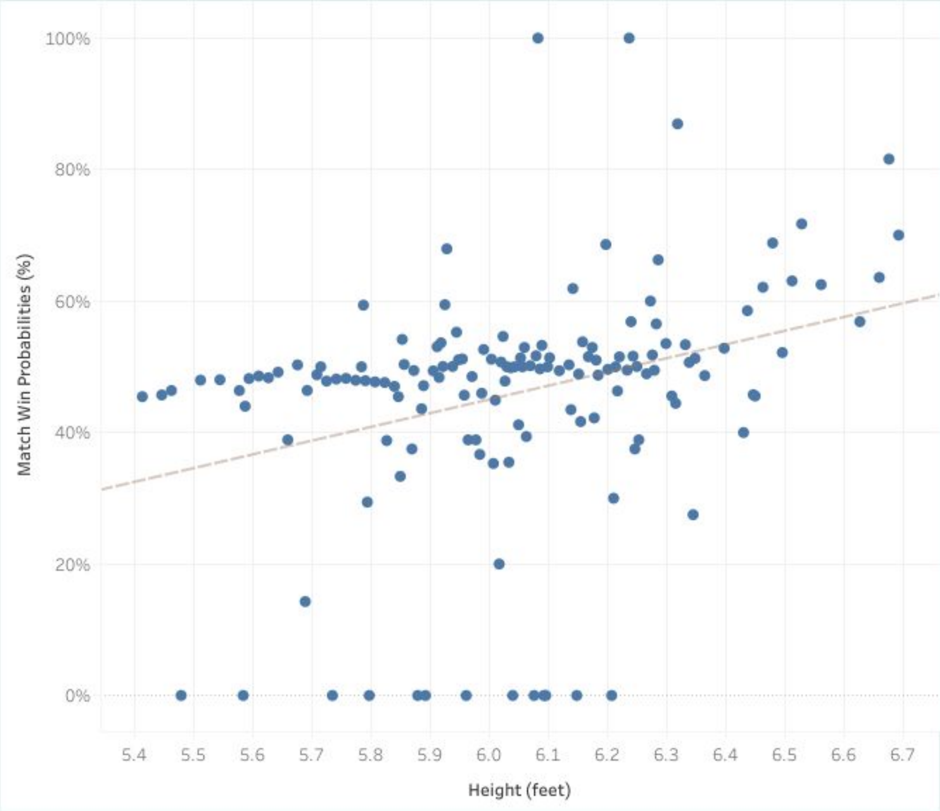
Link

EDA: Sumo Wrestling Outcome Prediction

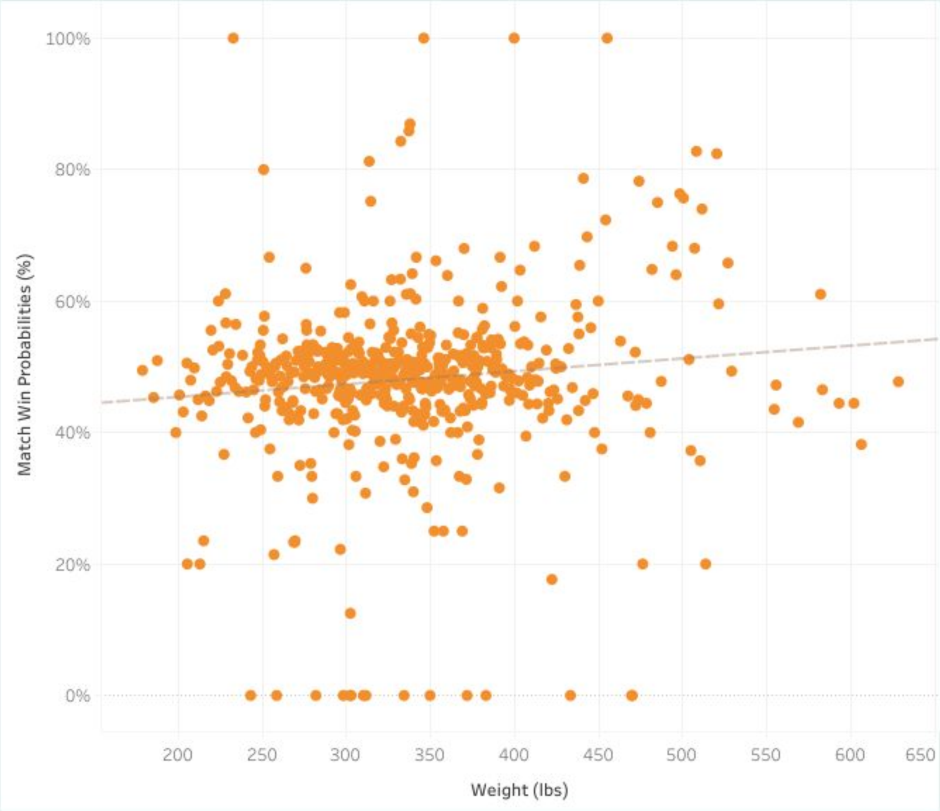
Match Win Probabilities Based On A Wrestler's Height and Weight

Time Period:
Jan 1983 - Mar 2023

Match Win Probabilities By Wrestler's Height



Match Win Probabilities By Wrestler's Weight



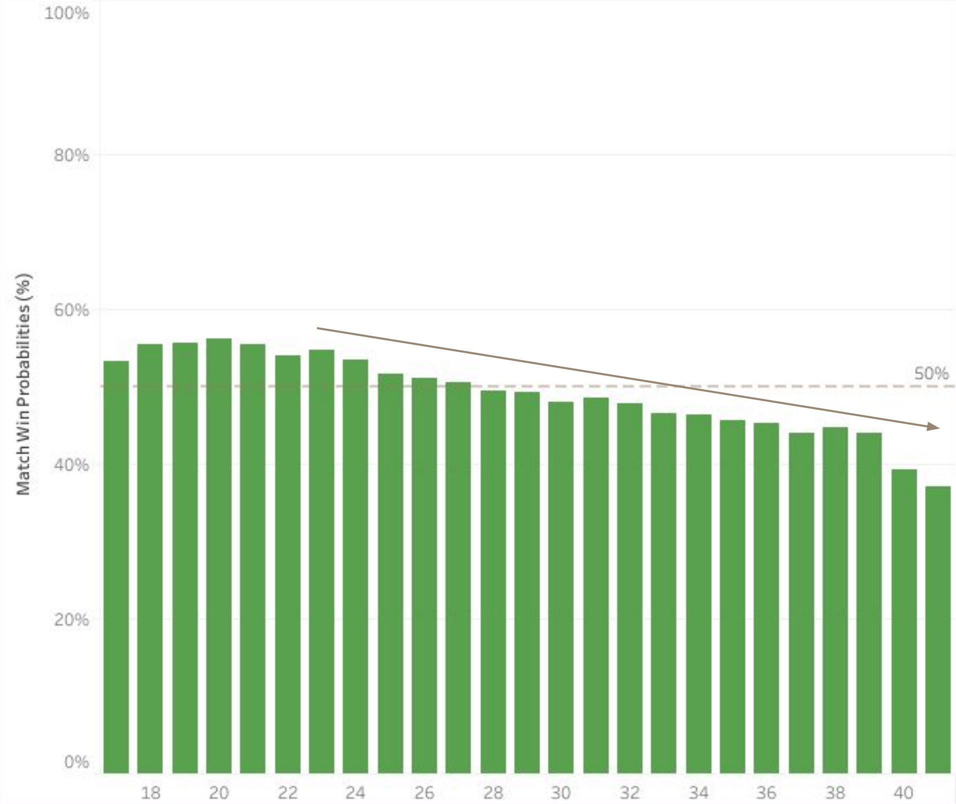
Key Findings: There is a positive correlation between height & weight and match outcome. Also, height has a stronger correlation than weight with match outcomes.

EDA: Sumo Wrestling Outcome Prediction

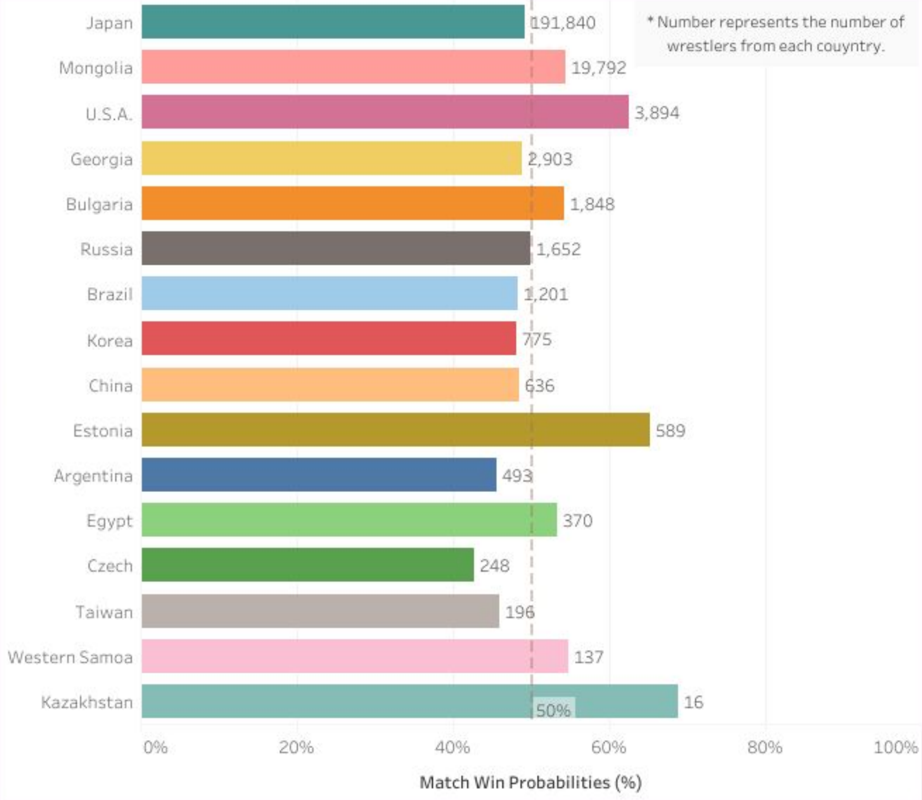
Match Win Probabilities Based On A Wrestler's Demographics
(Jan 1983 - Mar 2023)

Time Period:
Jan 1983 - Mar 2023

Match Win Probabilities By Age



Match Win Probabilities By Home Country



Key Findings: The green chart by age shows that younger wrestlers have a higher chance of winning, especially up to the age of 24, with a slight decline in win probabilities as the wrestler ages. Meanwhile, the chart by home country show..

EDA: Sumo Wrestling Outcome Prediction

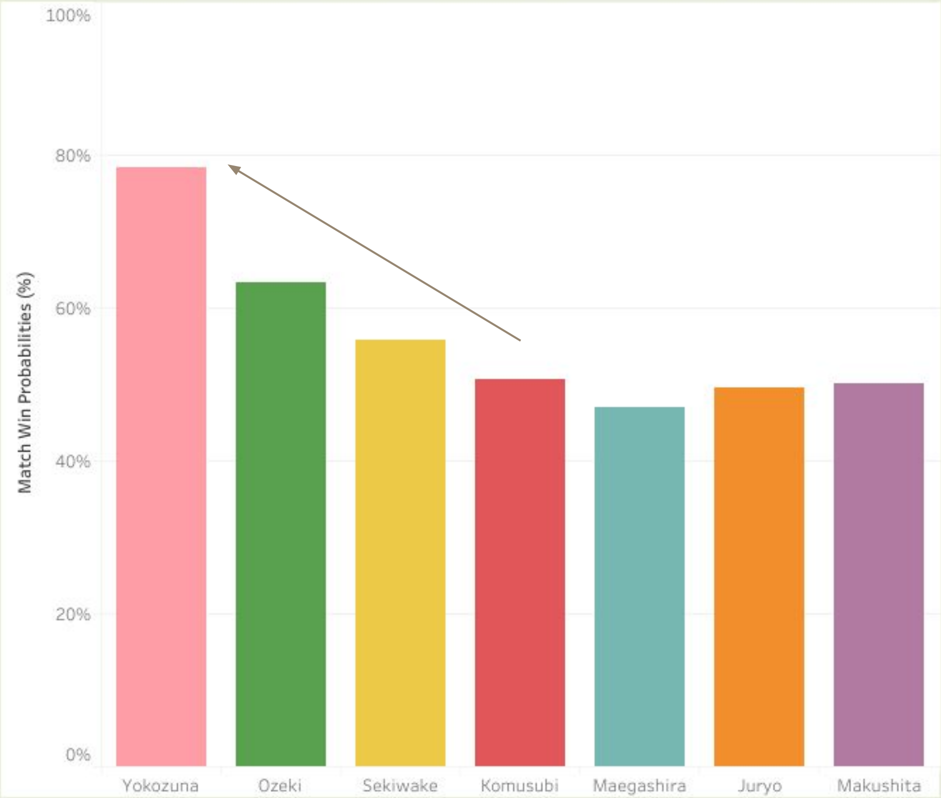
Match Win Probabilities Based On A Wrestler's Rank In The Previous & Current Tournament

(Jan 1983 - Mar 2023)

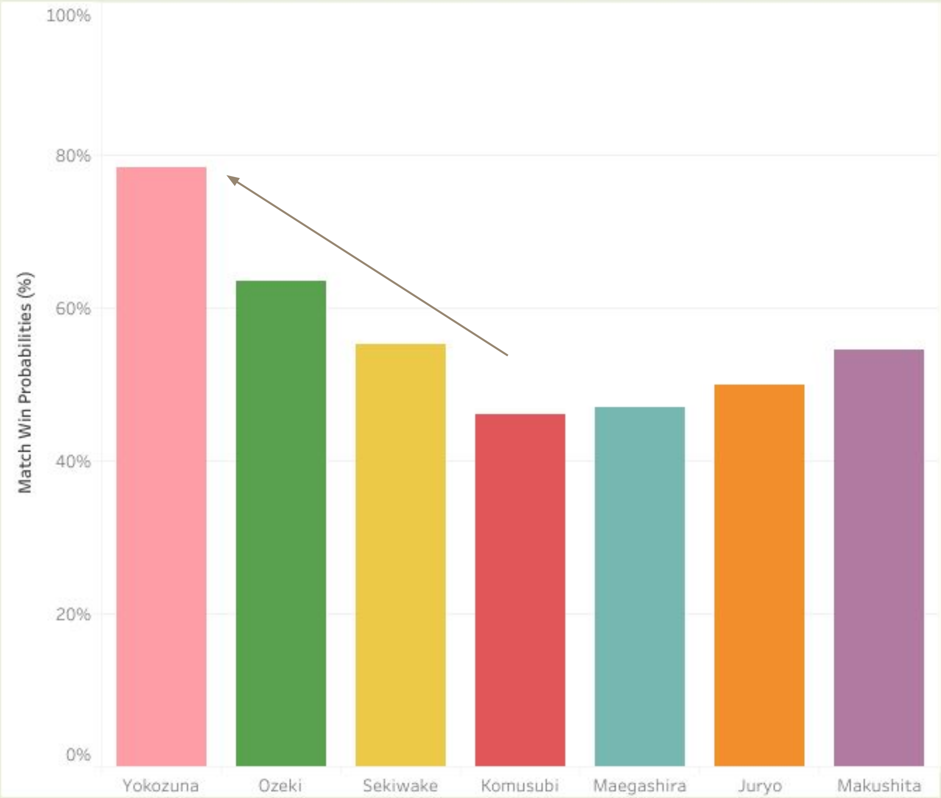
Sumo Ranks

Yokozuna
Ozeki
Sekiwake
Komusubi
Maegashira
Juryo
Makushita

Match Win Probabilities By Wrestler Rank In The Previous Tournament



Match Win Probabilities By Wrestler Rank In The Current Tournament

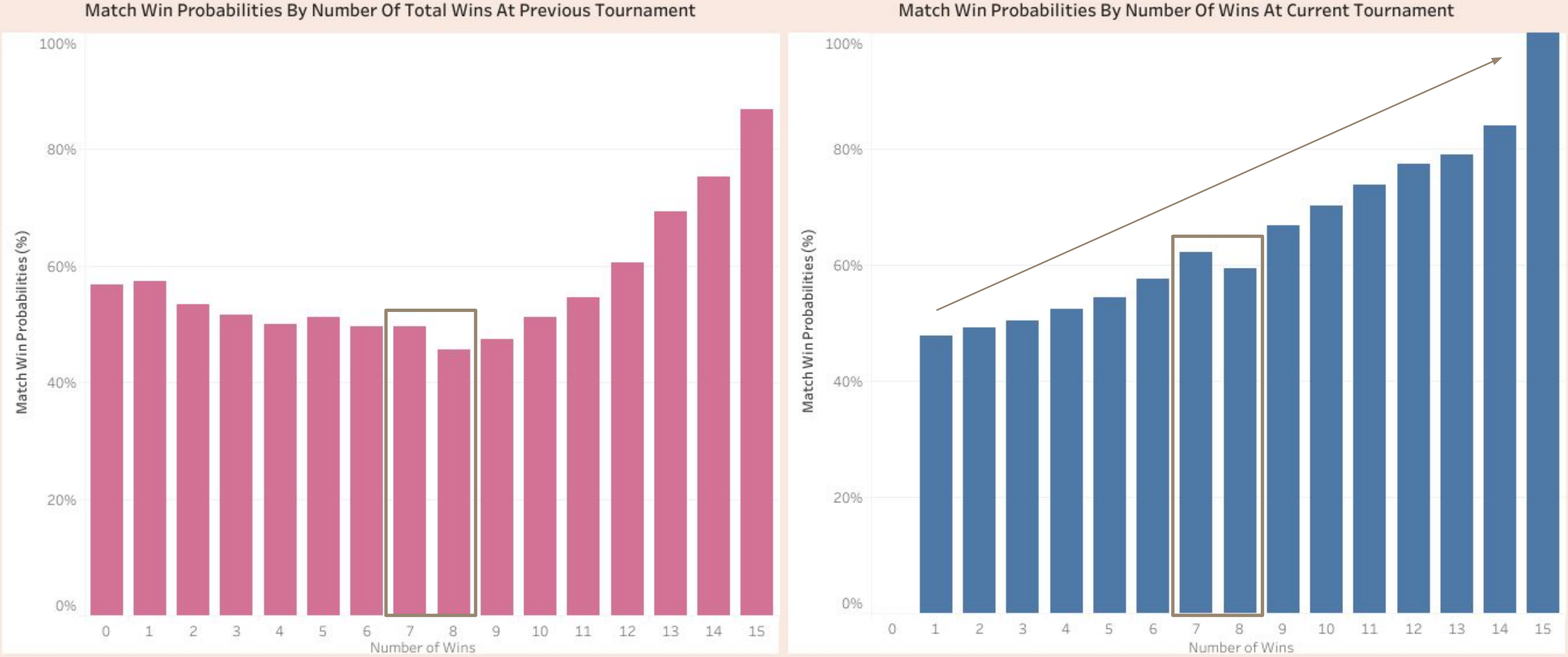


Key Findings: There are correlations between ranks and match outcomes in both previous and current tournaments, with higher-ranked wrestlers having a greater chance of winning matches. That's especially true for Yokozuna, Ozeki and Sekiwake ranks.

EDA: Sumo Wrestling Outcome Prediction

Match Win Probabilities Based On A Wrestler's Number Of Wins

Time Period:
Jan 1983 - Mar 2023



Key Findings: The current tournament data shows strong correlation between number of matches and match oucomes, as the number of win increases the winning probabilities also increase. Also, both charts show a higher probability of winning matches with 7 wins compared to 8 wins. In sumo wrestling, winning 8 matches is an important achievement for maintaining or getting promoted to a higher rank. Therefore, wrestlers who have already won 7 matches may have a higher sense of motivation or urgency to win their 8th match, leading to a slightly higher win probability than those with 8 wins.

05.

DATA MODELING & EVALUATION

What is PyCaret?

- PyCaret is an open-source machine learning library in Python that automates machine learning workflows with **minimal coding** required.
- With PyCaret, data scientists can spend less time coding and more time analyzing data.

Minimal coding required

Install PyCaret

```
!pip install pycaret
import pycaret
```

Setup

```
from pycaret.classification import *
setup(data = df, target = 'r1_win', train_size = 0.8, session_id=123)
```

Compare Models

```
best = compare_models()
```

Analyze Model

```
p = plot_model(best, plot = 'auc')
plot_model(best, plot = 'confusion_matrix')
```

Classification models using existing numeric data and feature-engineered data are developed

01.

Models using existing numeric data only

02.

Models incorporating new features

This process allows me to see how these features impacted the performance of the models and whether they enhance the predictive ability of match outcomes

The best performing model only achieved an accuracy score of 0.57

01.

Models using existing numeric data only

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
xgboost	Extreme Gradient Boosting	0.5713	0.6090	0.5757	0.5707	0.5732	0.1427	0.1427	17.4200
lightgbm	Light Gradient Boosting Machine	0.5678	0.6045	0.5758	0.5667	0.5712	0.1356	0.1356	2.3570
gbc	Gradient Boosting Classifier	0.5570	0.5877	0.5635	0.5563	0.5598	0.1140	0.1140	22.8280
rf	Random Forest Classifier	0.5519	0.5788	0.5373	0.5534	0.5452	0.1037	0.1038	42.4100
ada	Ada Boost Classifier	0.5478	0.5742	0.5468	0.5479	0.5473	0.0955	0.0955	5.8420
knn	K Neighbors Classifier	0.5441	0.5621	0.5437	0.5441	0.5439	0.0881	0.0881	1.8620
et	Extra Trees Classifier	0.5400	0.5627	0.5264	0.5411	0.5337	0.0800	0.0800	30.2730
qda	Quadratic Discriminant Analysis	0.5334	0.5540	0.5438	0.5327	0.5381	0.0667	0.0667	0.3640
nb	Naive Bayes	0.5315	0.5520	0.5362	0.5312	0.5337	0.0630	0.0630	0.1740
dt	Decision Tree Classifier	0.5237	0.5237	0.5259	0.5235	0.5247	0.0473	0.0473	1.9230
ridge	Ridge Classifier	0.5227	0.0000	0.5230	0.5227	0.5228	0.0454	0.0454	0.1860
lda	Linear Discriminant Analysis	0.5227	0.5348	0.5230	0.5227	0.5228	0.0454	0.0454	0.4890
lr	Logistic Regression	0.5220	0.5341	0.5217	0.5220	0.5218	0.0439	0.0439	1.5560
svm	SVM - Linear Kernel	0.5018	0.0000	0.4077	0.5550	0.3954	0.0036	0.0056	8.7650
dummy	Dummy Classifier	0.5000	0.5000	0.5000	0.2500	0.3333	0.0000	0.0000	0.1320

Baseline = 0.5

Incorporating new features led to a significant improvement in model performance

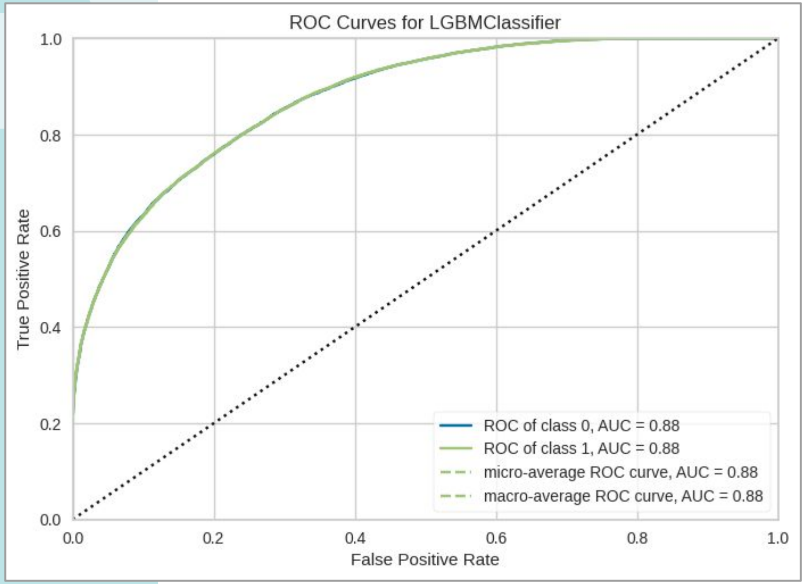
02.

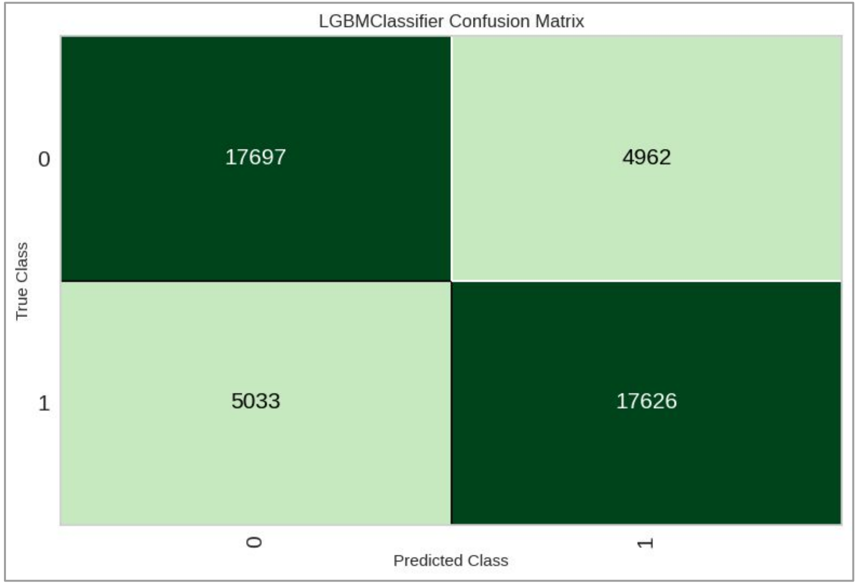
Models incorporating new features

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.7799	0.8785	0.7795	0.7802	0.7798	0.5598	0.5599	4.9240
xgboost	Extreme Gradient Boosting	0.7796	0.8781	0.7805	0.7791	0.7798	0.5592	0.5593	48.9350
gbc	Gradient Boosting Classifier	0.7731	0.8730	0.7777	0.7706	0.7741	0.5462	0.5462	65.6110
lda	Linear Discriminant Analysis	0.7720	0.8677	0.7695	0.7734	0.7714	0.5440	0.5440	1.8290
ridge	Ridge Classifier	0.7719	0.0000	0.7695	0.7733	0.7714	0.5439	0.5439	0.6030
rf	Random Forest Classifier	0.7686	0.8678	0.7623	0.7721	0.7671	0.5372	0.5373	53.2930
et	Extra Trees Classifier	0.7672	0.8656	0.7616	0.7702	0.7659	0.5344	0.5344	45.1910
ada	Ada Boost Classifier	0.7631	0.8642	0.7598	0.7658	0.7621	0.5262	0.5272	13.9880
qda	Quadratic Discriminant Analysis	0.7610	0.8454	0.7624	0.7602	0.7613	0.5219	0.5219	1.1990
lr	Logistic Regression	0.7606	0.8430	0.7537	0.7643	0.7589	0.5212	0.5213	24.1150
nb	Naive Bayes	0.7434	0.8282	0.7405	0.7447	0.7426	0.4867	0.4867	0.5010
dt	Decision Tree Classifier	0.7078	0.7078	0.7087	0.7074	0.7081	0.4156	0.4156	3.8840
svm	SVM - Linear Kernel	0.6626	0.0000	0.5944	0.7376	0.6220	0.3251	0.3668	17.4540
knn	K Neighbors Classifier	0.5625	0.5877	0.5630	0.5625	0.5627	0.1250	0.1250	46.1080
dummy	Dummy Classifier	0.5000	0.5000	0.5000	0.2500	0.3333	0.0000	0.0000	0.3660

Baseline = 0.5

Light Gradient Boosting Machine performed the best





	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.7799	0.8785	0.7795	0.7802	0.7798	0.5598	0.5599	4.9240

CONCLUSION & RECOMMENDATIONS

Conclusion

- Through extensive data cleaning, exploratory data analysis, and feature engineering, I have identified key factors that contribute to a wrestler's success or failure in sumo wrestling matches.
- My analysis has shown that the number of wins in previous and current tournaments, wrestler rank, and age are important predictors of match outcomes.
- I have developed several machine learning models to predict match outcomes, with Light Gradient Boosting Machine and Extreme Gradient Boosting achieving the highest accuracy score of 0.78.
- My Tableau dashboard provides an interactive platform for fans and practitioners to explore and visualize the data and predictions.

Recommendations

- To further improve my model's performance, I recommend collecting more information on sumo wrestlers, including:
  - Physical characteristics (such as muscle mass, grip strength, flexibility, endurance),
  - Injury history
  - The number of wins based on the wrestler's rank
- Based on user feedback and needs, I will expand/update the Tableau dashboard to provide more features and insights.
- I also recommend developing an app on Streamlit that allows sumo fans to predict different sumo wrestling games.
- My findings can provide practitioners and coaches with valuable insights into the key factors that contribute to their wrestlers' success or failure, allowing them to make more informed decisions when developing training and coaching strategies.

THANK YOU!

Any questions?