Ray Reyes
Professor Roth
Final Project Report

Introduction

The UFC is an organisation that puts on cage matches between mixed martial artists. Predicting the outcome of the fight can be potentially lucrative as there is a big sports betting side to this sport. I wanted to attempt to predict the outcome of the fights using machine learning. I believe that depending on the features of the fighter the outcome of the fight can be predicted effectively. Using a dataset that contains the features of the fighter and the outcome of their matches I trained a random forest model to predict the outcome of the fights. The model was 69% accurate when tested against fighters that were not in the dataset.

Methods

I had to use two separate data sets to make my model possible. First I used a fight outcome dataset, it included individual fight results between two fighters, such as who won along with match up level features like height difference and reach difference. Then I used a second dataset that constrained fighter specific stats like win /loss average strikes and take down averages taken over the course of their career. I used the first data set to train the random forest, as it had labeled outcomes that I could use for the model. The second dataset was useful for actually comparing the two fighters head to head allowing me to use the model to predict future fights. To prepare the training data I calculated the difference in stats between the two fighters for each match up such as the difference in the number of wins, height, or reach. These differences became the feature that I used in the model. I dropped rows with missing values and trained the random forest classifier on the 15 most relevant features. Now that the model is trained on historical outcomes (dataset 1). I can use it to predict new match ups by comparing the two fighters stats(dataset 2), calculating the differences, and feeding that into the model. For example, if I want to simulate a future fight between Paddy Pimblett and Zhang Weili, I pull their stats from the second dataset, compute the feature differences, and input that into the model to get a predicted winner and win probabilities.

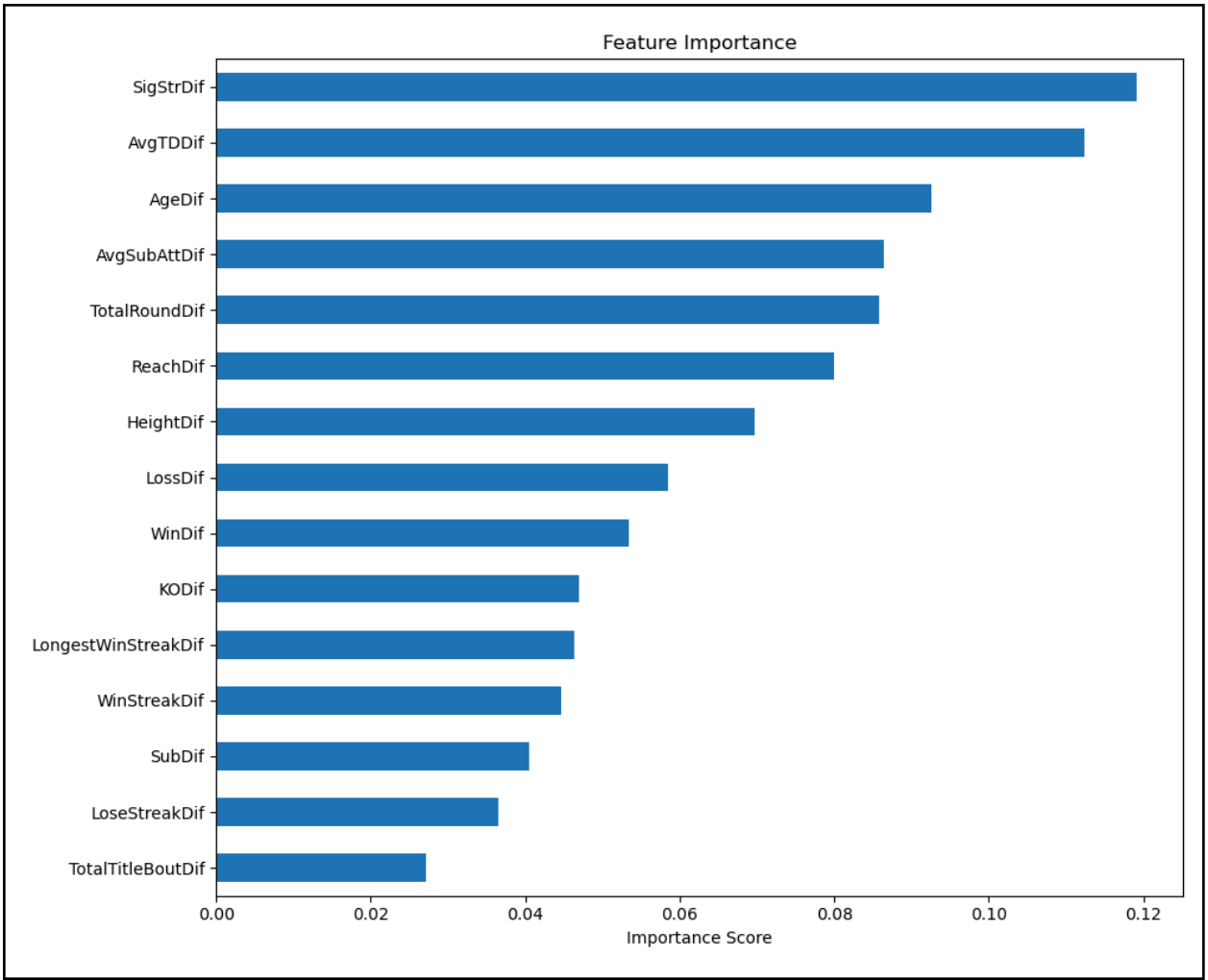
Stats Dataset (Dataset2)

Name	Wins	Losses	Height (cm)	Reach (cm)	Age	SLpM	Sub Avg	TD Avg
Amanda Ribas	12	5	160.02	167.64	30	4.63	0.7	2.07
Rose Namajunas	13	6	165.10	165.10	31	3.69	0.5	1.38
Karl Williams	10	1	190.50	200.66	34	2.87	0.2	4.75

Justin Tafa	7	4	182.88	187.96	30	4.09	0.0	0.00
Edmen Shahbazyan	13	4	187.96	190.50	26	3.60	0.6	2.24

Outcome Dataset (Dataset 1)

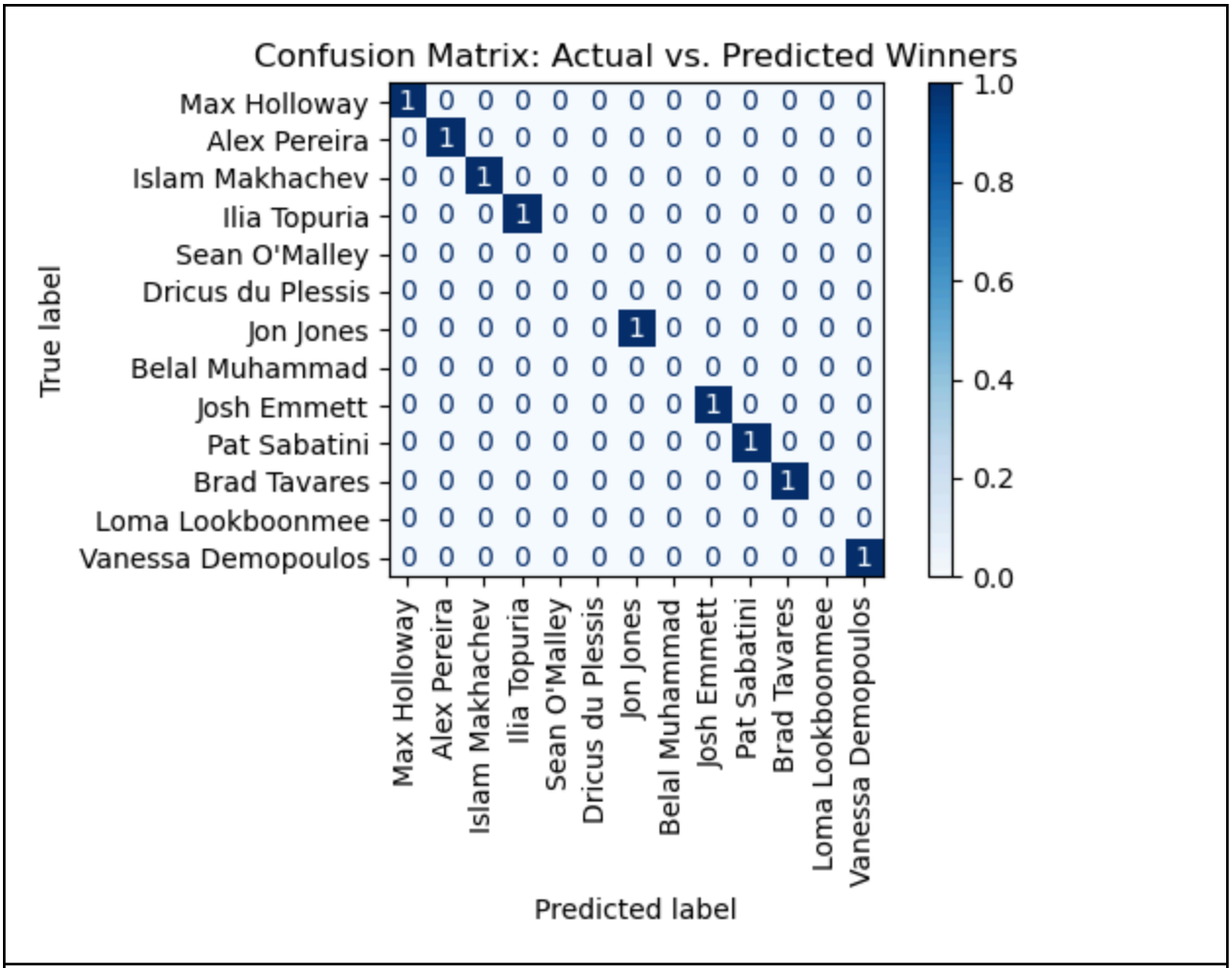
Red Fighter	Blue Fighter	Winner	Height Diff (cm)	Reach Diff (cm)	WinDiff
Alexandre Pantoja	Kai Asakura	Red	7.62	5.08	-12
Shavkat Rakhmonov	Ian Machado Garry	Red	5.08	-7.62	2
Ciryl Gane	Alexander Volkov	Red	7.62	-2.54	3
Bryce Mitchell	Kron Gracie	Red	-2.54	0.00	-6
Nate Landwehr	Dooho Choi	Blue	2.54	-5.08	-1



Feature importance table

Results

I was very happy with the model because it performed better than I thought it would. When I tested my model against the test set it was 58% accurate so it was right most of the time. I also tested the model against fights that were much more recent and were not included in the data that I used to produce the model. The model was 69% accurate which was impressive to me, i created a confusion matrix to represent this.



The confusion matrix of recent fights. Precision: 0.69

Discussion

My project produced a binary result but can be used to give an insight as to what the outcome of a fight could be using the most important features in determining a fight. Something I noticed is that if the fighter has an outlier in one of the important features then it may disproportionately affect the model. For example a fighter like Israel Adesanya who is abnormally tall for his weight class, the model favors him much more than it should. The model is sensitive to outliers. One thing that I could do to improve the model is to figure out a way to add more features which would increase accuracy. As I touched upon earlier, sports betting sites use similar models to determine the odds of a fight and let you make a wager. They have the most advanced models and have people working wound the clock to improve them.

Citations

Data sources

https://www.kaggle.com/datasets/mdabbert/ultimate-ufc-dataset (Credit:mdabbert / Organization:Kaggle)

https://www.kaggle.com/datasets/maksbasher/ufc-complete-dataset-all-events-1996-2024
(Credit: MaksBasher / Organization:Kaggle)

Use of AI
I used AI to help develop and fix code errors.