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 data set that contained features, fighters, and the outcome of their fights. I used that data set to train the random forest, with the target being to predict the winner. Then I used a separate dataset that contained fighters and their features so the fighters could be better compared to one another. I used the features that the model determined was most important(refer to the table).

Feature importance table
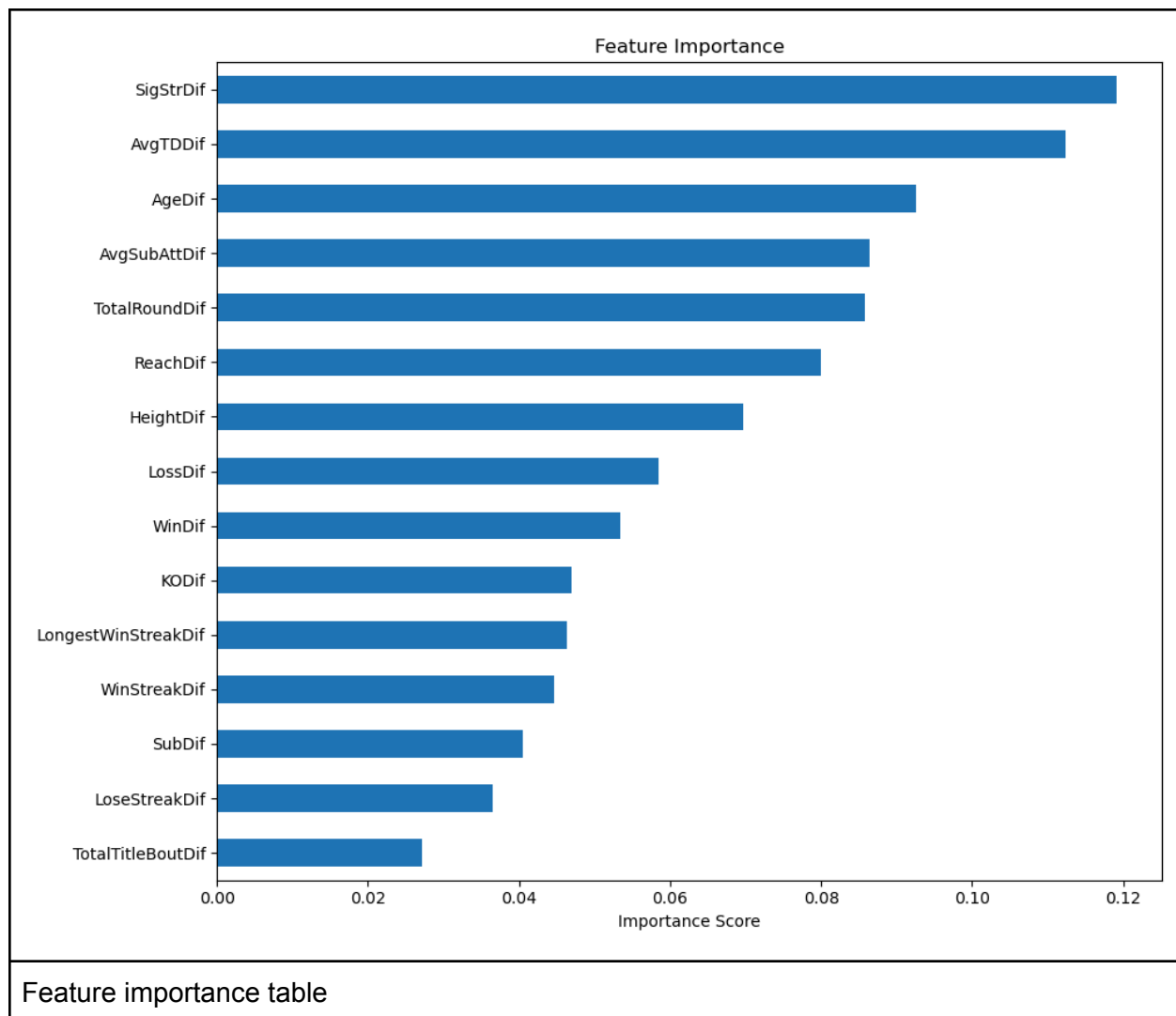
The model takes the difference of the features to determine the odds of who would win the fight. For example it will take the difference in height of the two fighters and the taller fighter will have the advantage in that feature. Both dataset constrained similar data but not the same so I didn't have to clean the dataset but instead exclude the features that did not overlap in both datasets. Choosing a winner is a binary result so I did not try other models because I believe that the random forest would be the best model for the job.

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.

Confusion Matrix: Actual vs. Predicted Winners

The confusion matrix of recent fights. Precision: 0.69
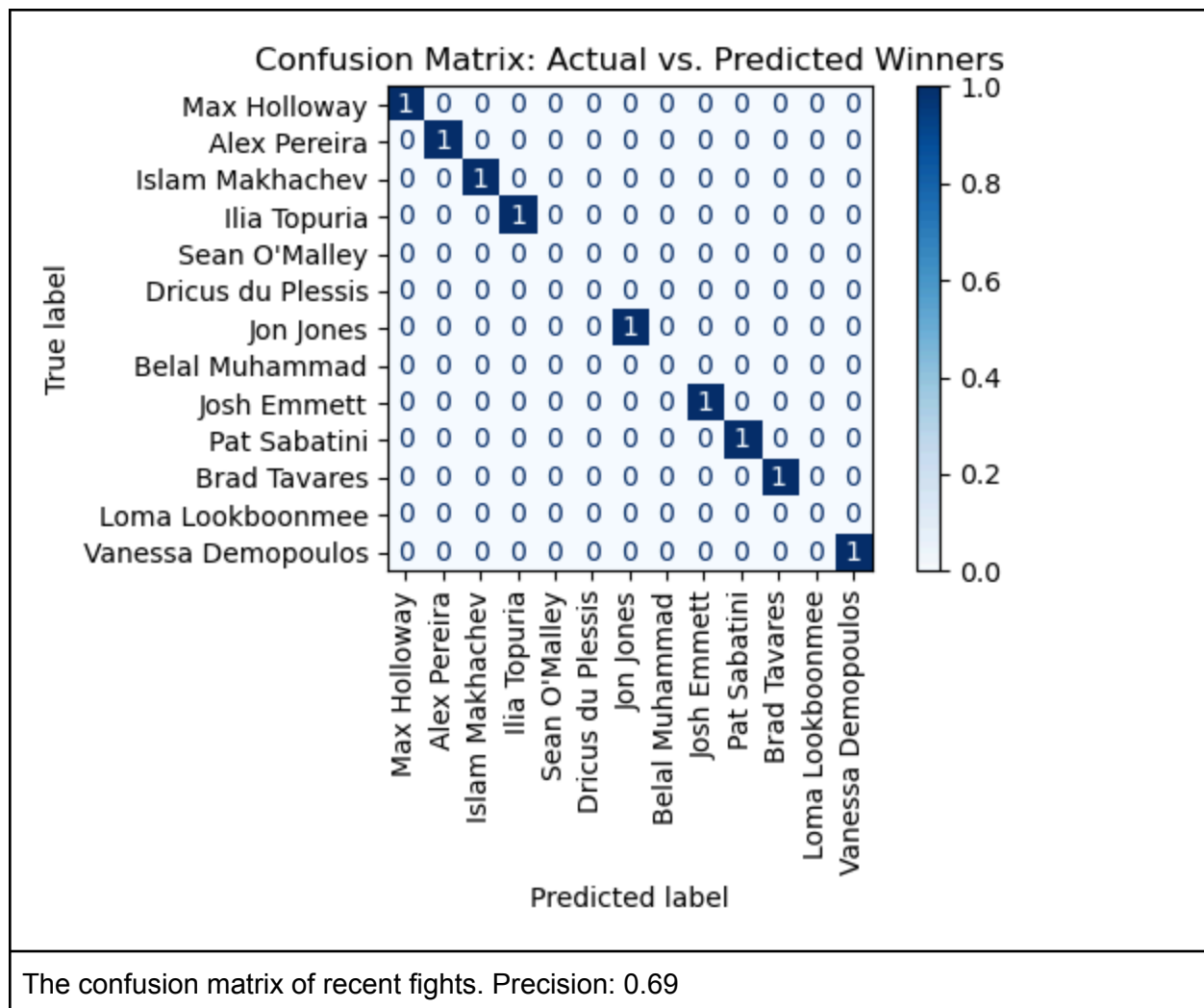
Discussion

My project did not necessarily produce 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.