UFC Outcome Prediction through Machine Learning

Manpreet Nanreh
mnanreh@ryerson.ca
Ryerson University
Toronto, Ontario

ABSTRACT

The Ultimate Fighting Championship is the largest Mixed Martial Arts organization in the world. Since UFC has gained so much popularity, it has become evident that techniques other than Statistical Models are being are being applied to predict the outcome of the fights. The dataset we consider in this study consists the outcome from about 5000 between 1993 and 2019. In this paper, the goal is to explore models that will allow to make outcome prediction for fights through the application of Support Vector Machines (SVM) and a Long Short-Term Memory (LSTM). The two models provide classification accuracy of about 70%.

KEYWORDS

ufc, prediction, outcome, mixed martial arts, machine learning, lstm, svm

1 INTRODUCTION

Hand to hand combat has been a source of entertainment for many centuries and to this day Mixed Martial Arts (MMA) has increased a lot in its popularity. During each fight, fighters of various weight class compete with opponents from their respective weight class in order to achieve victory. Also, fighters are allowed to use variety of fighting styles such as Boxing, Wrestling and more. The ring used to present these fights is a pentagon and there are two corners representing each fighter, where commonly used corner colours are red and blue. There are several outcomes to a fight and some of which include: Submission, Knockout, Technical Knockout and Decision via scoreboard[1].

Due to the advancement of Machine Learning techniques and their ability to provide highly accurate results, they are being applied to the domain of sports more often. We attempt to explore two different Machine Learning models and determine whether these models will provide better results compared to the already used models in other studies. The dataset was collected in order to contain each fighter's performance from their previous fights which indicates that fighters who have performed better in their previous fights will have a higher chance of winning. We used SVM and LSTM to make prediction which will provide confidence score for each fighter indicating whether the red or blue corner will win.

2 RELATED WORK

There has been previous study done on UFC outcome prediction that uses the basic Machine Learning algorithm. Therefore, it has already been confirmed that Machine Learning models are successful in provided resonable prediction outcomes. The paper by Hitkul et al.[1], used the following models in their study: Perceptron, Random Forests, Decision Trees classifier, Stochastic Gradient Descent (SGD) classifier, SupportVector Machine (SVM), and K-Nearest Neighbor (KNN). From their study, it was determined that SVM provided the best observed accuracy of 62.8%. The accuracy by their study shows that a fairly good level of success can be achieved by the classic Machine Learning models. The dataset used in the paper[1] was obtained from FightMetric LLC. Since then FightMetric LLC has changed their name to UFC STATS[3]. The dataset contained fight data post 2013 and an assumption was made that all fighters started from that period[1].

Another approach used in the paper by Goddign et al.[2] is to use LSTM in order to predict Football match outcomes. It discusses the effectiveness of LSTM with peepholes, Gated Recurrent Units and different loss functions. This gave the motivation to apply LSTM to the UFC dataset and analyse its performance.

3 DATASET AND FEATURE ENGINEERING

The dataset used in this study was obtained from Kaggle[5]. This dataset is also extracted from UFC STATS[3], therefore adhering to the consistency of data source. It consists of over 5000 fights that have happened since 1993 till 2019. From the raw data, appropriate information was extracted such that it would sufficiently define previous fight records. Each entry consisted of data for each fighter indicated by their corner colour for that fight. The red corner participant had their features indicated by the following format "R_" and the blue corner participant has their features indicated by the following format "B_". After careful feature extraction and feature engineering, 145 features were obtained for each fight, some of which were:

- red and blue fighter's current winning streak
- red and blue fighter's current losing streak
- red and blue fighter's height
- red and blue fighter's takedown percentage

In total, 5144 fight entries were obtained and 144 features were formed with one dependent variable indicating which fighter won the fight.

During exploratory data analysis, there were nine categorical features found and the following observations we made:

- red and blue fighter names were dropped from each fight because the fight index was enough to define which fight was being discussed
- most fights occur in USA as seen in **Figure 1**
- the year 2014 had the most fights as seen in **Figure 2**
- referee, data and location features was also found to be of no use because it showed no correlation to the actual outcome of the fight
- most fights occur in Lightweight and Welterweight weight class as seen in **Figure 3**
- one hot encoding was made for the weight class feature
- one hot encoding was made for the stance of each fighter

Another observation is that "draws" only occur 1.6% of the time

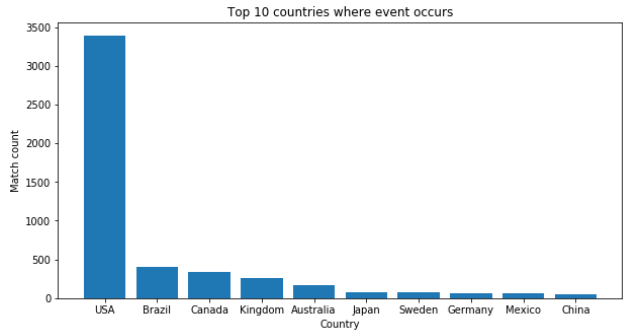

Figure 1: The number of fights occurring by the country.

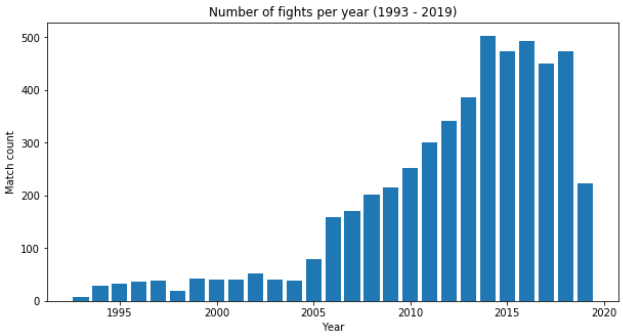

Figure 2: The number of fights happened per year.

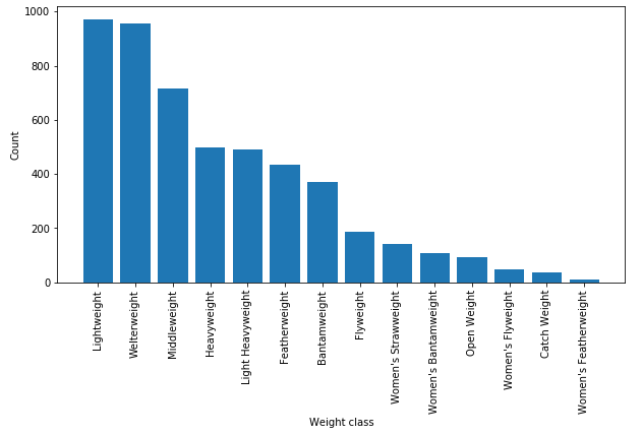

Figure 3: Total number of fights per weight class.

as seen in **Figure 4** and red corner wins more often. Therefore, fights that end up in draw outcome were removed from the dataset. As seen in **Figure 5**, a positive correlation was found between height and reach. In order to fill in Null value or missing values the following decisions were made:

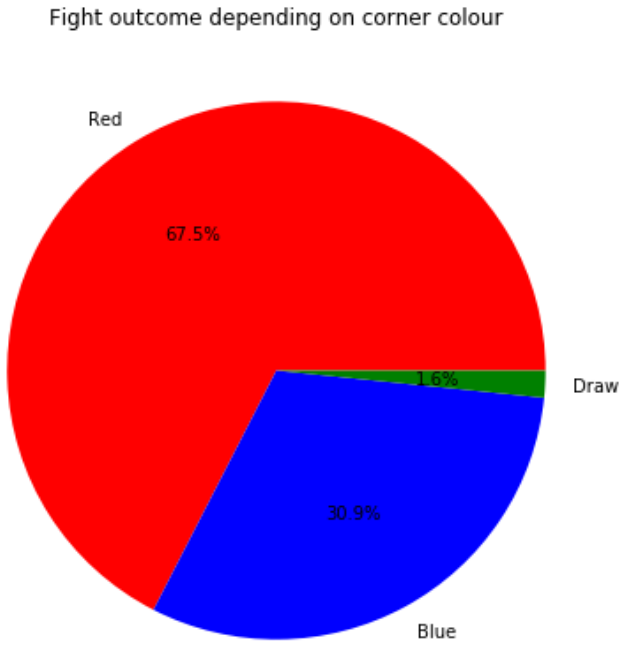

Figure 4: Fight outcomes depending on colour.

- for each fight, the stance of the fighter was filled with the mode of that corner's fighting stance
- for each fight, the height of the fighter was filled with the mean of that corner's height
- for each fight, the reach of the fighter was filled with the height of that fighter
- for the rest of the feature, they were filled with the mean of the respective feature

Upon analyzing the spread of data for all the features, the following features were removed because they in included only single type of entry 90% or more of the time:

- blue draw
- red draw
- red win by majority decision
- red win by doctor stoppage
- blue win by majority decision
- blue win by doctor stoppage

The final cleaned data was split into 80% training set and 20% test set.

4 METHOD

4.1 Support Vector Machine (SVM)

SVM is a supervised learning model which can be used for regression and classification tasks. During the analysis, "rbf" and "linear" kernel were used to fit the training set. The model was tuned on the following parameters:

- max iterations: 10, 100, 1000, 10000, 100000
- regularization parameter (C): 0.1, 1, 10, 100, 1000

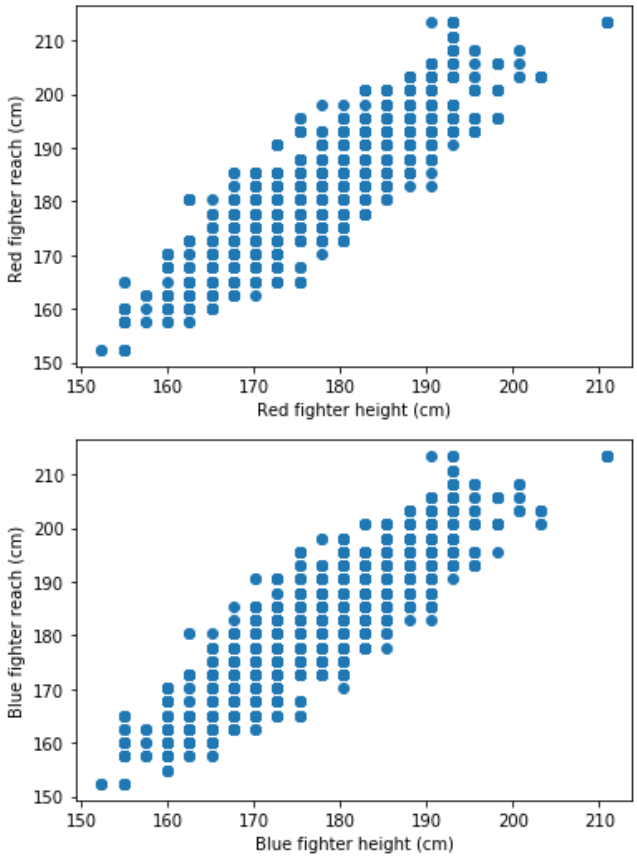

Figure 5: The correlation between height and reach.

4.2 Long Short Term Memory

LSTM is a form of RNN and in this case it is used to predict UFC outcome. The full neural network has three layers. The input the neural network is of 154 dimension which gets fed into the LSTM. The LSTM has a dropout rate of 20% and it returns a two-dimensional vector which gets fed into softmax to obtain the final prediction. Adam was used as the optimizer and the categorical crossentropy was used as loss function. The neural network was setup such that with each iteration it would validate on the validation set which was 10% of the training set. The training was monitored upon the loss occurring during cross validation on validation set and once the validation set loss starts increasing the training was stopped early. The architecture of the neural network can be seen in **Figure 6**. The training was done for 100 epochs and with the batch size of 256.

5 RESULTS & DISCUSSION

5.1 SVM

After tuning the SVM, it was determined the "linear" kernel was found to outperform "rbf" kernel and the optimal maximum iterations were 10000 and optimal regularization parameter (C) was found to be 10. From **Figure 7**, it can be seen that SVM allowed to obtain 65% accuracy on the test set.

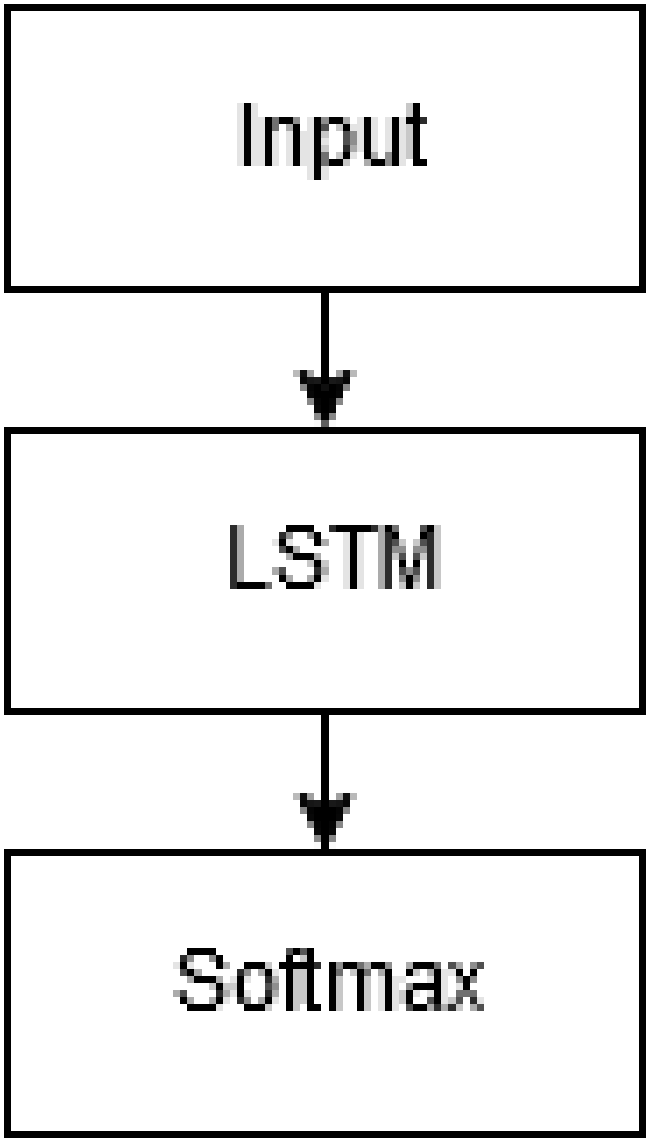

Figure 6: Neural Network Architecture.

	precision	recall	f1-score	support
0	0.75	0.75	0.75	711
1	0.42	0.42	0.42	302
accuracy			0.65	1013
macro avg	0.58	0.58	0.58	1013
weighted avg	0.65	0.65	0.65	1013

Figure 7: SVM classification report.

Table 1: Test set prediction accuracy for different models

Model	Test set accuracy
SVM	65%
LSTM	70%

5.2 LSTM

The LSTM was run on the parameter define in **4.2** with batch size being 256 for 100 epochs and dropout rate being 20%. **Figure 8** shows us that on test set it allowed to achieve 70% accuracy.

Figure 8: LSTM output on test set.

6 CONCLUSION & FUTURE WORK

As it can be seen from **Table 1**, LSTM allowed to obtain better results than SVM and it is much faster to train than SVM. LSTM on average would take 30 seconds to 40 seconds to train while SVM would take 5 minutes to 10 minutes to train. For future work, it would be beneficial to change the architecture of neural to contain more layers in order to increase its depth. There is another paper that uses Machine Learning to predict NFL game outcome[4]. For future work, the techniques from this paper could also be applied to UFC dataset and analyse its accuracy.

REFERENCES

[1] Hitkul, Karmanya Aggarwal, Neha Yadav, and Maheshwar Dwivedy. 2019. A Comparative Study of Machine Learning Algorithms for Prior Prediction of UFC Fights. In *Harmony Search and Nature Inspired Optimization Algorithms*, Neha Yadav, Anupam Yadav, Jagdish Chand Bansal, Kusum Deep, and Joong Hoon Kim (Eds.). Springer Singapore, Singapore, 67–76.

[2] Daniel Pettersson and Robert Nyquist. 2017. *A Sure Bet: Predicting Outcomes of Football Matches.* https://pdfs.semanticscholar.org/e556/af01e86c3414042aa69831ea5fb398e66f94.pdf.

[3] UFC STATS. [n. d.]. *UFC STATS.* www.ufcstats.com/.

[4] Jim Warner. 2010. *Predicting Margin of Victory in NFL Games:Machine Learning vs. the Las Vegas Line.*

[5] Rajeev Warrier. [n. d.]. *UFC-Fight historical data from 1993 to 2019.* https://www.kaggle.com/rajeevw/ufcdata.