A Comparative Study of Machine Learning Algorithms for Prior Prediction of UFC Fights

Hitkul, Karmanya Aggarwal, Neha Yadav and Maheshwar Dwivedy

Abstract Mixed Martial Arts is a rapidly growing combat sport that has a highly multi-dimensional nature. Due to a large number of possible strategies available to each fighter, and multitude of skills and techniques involved, the potential for upset in any fight is very high. That is the chance of a highly skilled, veteran athlete being defeated by an athlete with significantly less experience is possible. This problem is further exacerbated by the lack of a well-defined, time series database of fighter profiles prior to every fight. In this paper, we attempt to develop an efficient model based on the machine learning algorithms for the prior prediction of UFC fights. The efficacy of various machine learning models based on Perceptron, Random Forests, Decision Trees classifier, Stochastic Gradient Descent (SGD) classifier, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) classifiers is tested on a time series set of a fighter's data before each fight.

Keywords Machine learning algorithms · Mixed martial arts · Classifiers

1 Introduction

Mixed Martial Arts (MMA) is currently one of the fastest growing sports in the world. The UFC or Ultimate Fighting Championship is currently the largest fight promotion in the mixed martial arts world. Between 2013 and 2017, the promotion

Hitkul · K. Aggarwal · N. Yadav (✉) · M. Dwivedy
School of Engineering and Technology, BML Munjal University,
Gurugram 122413, Haryana, India
e-mail: neha.yadav@bmu.edu.in

Hitkul
e-mail: hitkul.bmu.14cse@bmu.edu.in

K. Aggarwal
e-mail: karmanya.aggarwal.14cse@bmu.edu.in

M. Dwivedy
e-mail: maheshwar.dwivedy@bmu.edu.in

N. Yadav et al. (eds.), *Harmony Search and Nature Inspired Optimization Algorithms*,
Advances in Intelligent Systems and Computing 741,
https://doi.org/10.1007/978-981-13-0761-4_7

had presented over 1400 fights and counting, with an event being held bi-monthly and having multiple fights per event.

We attempted to evaluate the accuracy of multiple machine learning algorithms in order to determine which method is best suited to predict fight results given both competitors' records prior to the fight. Though several works have been published that seek to forecast performance of an MMA fighter prior to the fight [1], we attempted to create a dataset that reflects each fighter's statistical record prior to each fight and build a predictive model. Thus, we should ideally be able to predict a fighter's performance. Intuitively, an experienced fighter would most certainly have an advantage over a novice provided the age difference is not large enough to affect athletic performance. We evaluated many different machine learning models, and charted their performance over the dataset. It was found that the Random Forests and SVM gave the best results in terms of prediction accuracy.

For a brief background of a UFC event, the UFC is a fighting promotion. MMA employs various techniques from an ensemble of different martial arts such as Jiu Jitsu, Boxing, Taekwondo, and Wrestling. This allows for a wide variety of strikes and tactics to be employed by the fighters depending on their expertise in each art. A typical UFC event has multiple fights on a particular day—these events take place roughly once every 2 weeks. Each fight typically lasts three rounds of 5 min each. However, major fights will last five rounds. The two fighters are denoted red and blue side, with the better known fighter being allocated the red side. There are multiple ways to win a fight, via Knockout/Technical Knockout wherein the fighter overwhelms his opponent with strikes until he is unable to continue, via submission; wherein a fighter cedes victory, or finally by decision, when the fight reaches the end of the allotted time for the fight and the fighters are judged by a panel of three judges on factors such as damage inflicted, aggression, and ring control. Decision victories are the most common, however these are the hardest to judge, as the judging process tends to be rather opaque [2, 3].

Today, statistical modeling and its applications in the UFC are in its infancy [4–6]. No thoroughly rigorous statistical models have been published till date to predict the UFC fights previously. In this paper, we attempt to correct this imbalance—while there remains insufficient data available to build fighter specific models (with the UFC publishing granular fight data only since 2013 and each fighter fighting less than 10 times every year). We have attempted to build a model to predict which fighter is more likely to emerge victorious. In order to create the dataset, we retrieved each fighter's current statistics and subtracted their per fight statistics in order to create a sort of time-dependent dataset—reflecting what each fighter's statistics were prior to each fight, in terms of strikes, takedowns, styles, etc. In an ideal world, this model can be used to create matchups where both fighters are equally likely to win, as having this sort of equity in winning chance will most likely correlate with more exciting fights, as well as equalizing betting odds for fighters prior to each fight.

The organization of the paper is as follows: brief description of the models used is given in Sect. 2. Section 3 describes about the data exploration and feature manipulation. Statistical models along with the results are given in Sect. 4. Further Sect. 5 continues with results and discussion and finally Sect. 6 concludes the study.

2 Models Used

2.1 Random Forests

Random Forests is an ensemble classification technique consisting of a collection of tree-structured classifiers where random vectors are distributed independently and each tree casts a unit vote for the most popular class for a particular input [7].

2.2 Support Vector Machine (SVM)

SVMs are set of related supervised learning methods used for classification and regression. The input vector is mapped to a higher dimensional space where a maximal separating hyperplane is constructed [8].

2.3 K-Nearest Neighbors (KNN)

KNN is a classification technique that assigns points in our input set to the dominant class amongst its nearest neighbors, as determined by some distance metric [9].

2.4 Decision Tree

Decision trees are sequential models, which logically combine a sequence of simple tests. Each test compares a numeric attribute against a threshold value or a nominal attribute against a set of possible values [10].

2.5 Naive Bayes

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong independence assumptions. An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters necessary for classification [11].

2.6 Perceptron

A Perceptron is composed of several layers of neurons: an input layer, possibly one or several hidden layers and an output layer. Each neuron's input is connected with the output of the previous layer's neurons whereas the neurons of the output layer determine the class of the input feature vector [9].

2.7 Stochastic Gradient Descent (SGD)

SGD, also known as incremental gradient descent, is a stochastic approximation of the gradient descent optimization method for minimizing an objective function that is written as a sum of differentiable functions [12].

3 Data Exploration and Feature Manipulation

Granular fight data is available for UFC fighters by FightMetric LLC. Highly granular data is only available post 2013, thus an assumption has been made that all fighters from that period and beyond start at 0. By collecting and summing statistics per fight, we were able to assemble a tabulation of each fighter's statistics prior to each fight. From this set, we can see that we have a total of 895 columns and one dependent variable. The columns themselves have 13 integer types (Streaks, Previous Wins, etc.), 9 object types (Names, Winner, Winby, etc.) and 873 Float types. The features for data set are represented by Figs. 1, 2 and 3. Some quick observations from the raw dataset-

1. Red side seems to win slightly more than blue (867/1477 = 58.7%).
2. There are more fighters fighting debut fights.
3. Most fights are won by decision, and 2015 had the most fights.
4. The features seek to accommodate different fighter's styles (including both attempted strikes/takedowns versus significant or landed strikes/takedowns in an effort to quantify strike/takedown volume as a meaningful statistic.

We then filled all the Null values in our dataset with 0 values and assigned numeric codes to all categorical values. As one can see from Fig. 1 that the highest correlations are with Round 4 and Round 5 features, since most fights do not have Round 4 and Round 5. To deal with this sparsity, we summed the respective features of each round. Finally, we then attempt to half the number of features again, by taking the ratio of features from red and blue side fighters.

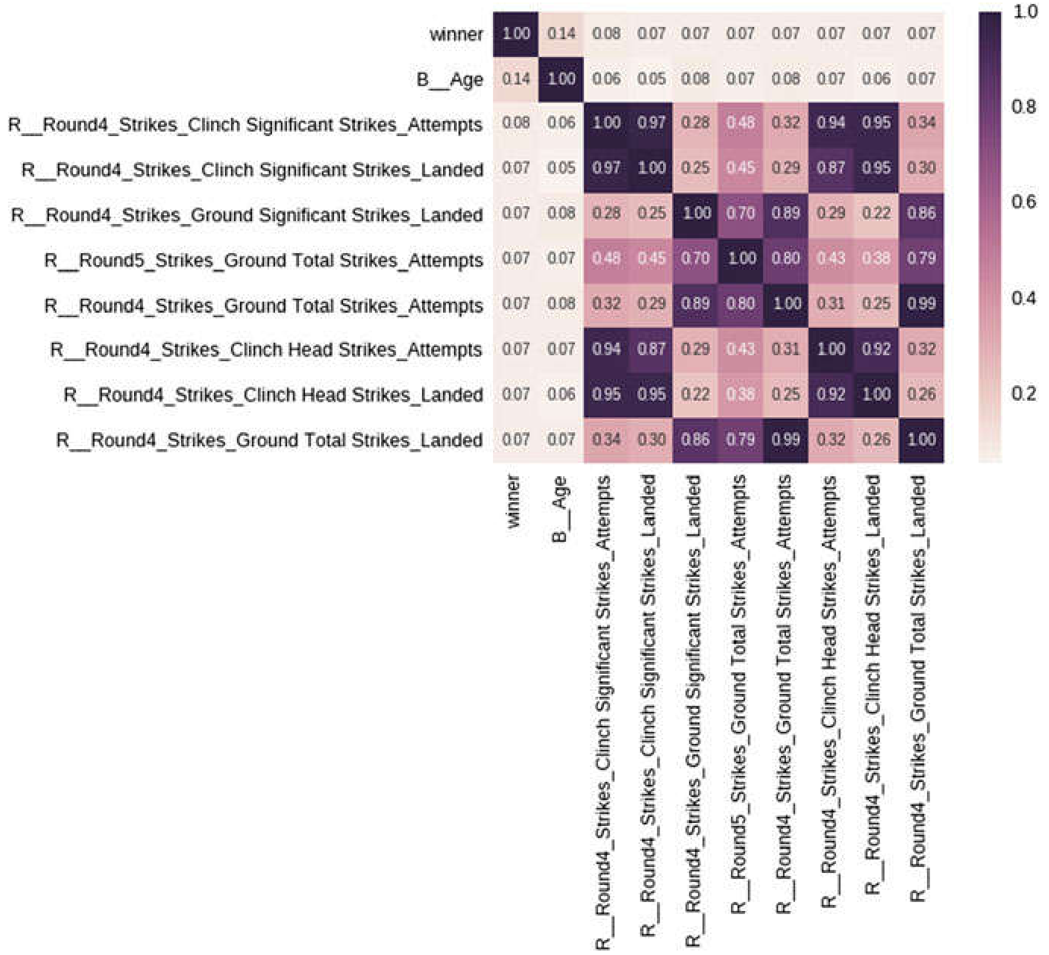

Fig. 1 A heatmap of the highest 10 correlations with our target variable

4 Modeling

Performance of multiple machine learning models on this dataset is then evaluated and explored by a variety of statistical methods described in Sect. 2. Table 1 describes the performance of our chosen models on the raw dataset. Table 2 describes the performance of the same models after we summed respective round features and Table 3 describes the performance of the models post taking the ratio of red and blue side fighters' respective features (Figs. 4, 5 and 6).

5 Results and Discussion

From Fig. 7, it is evident that random Forests and SVM showed the most consistent results against the dataset. Models like Naive Bayes and simple decision trees showed

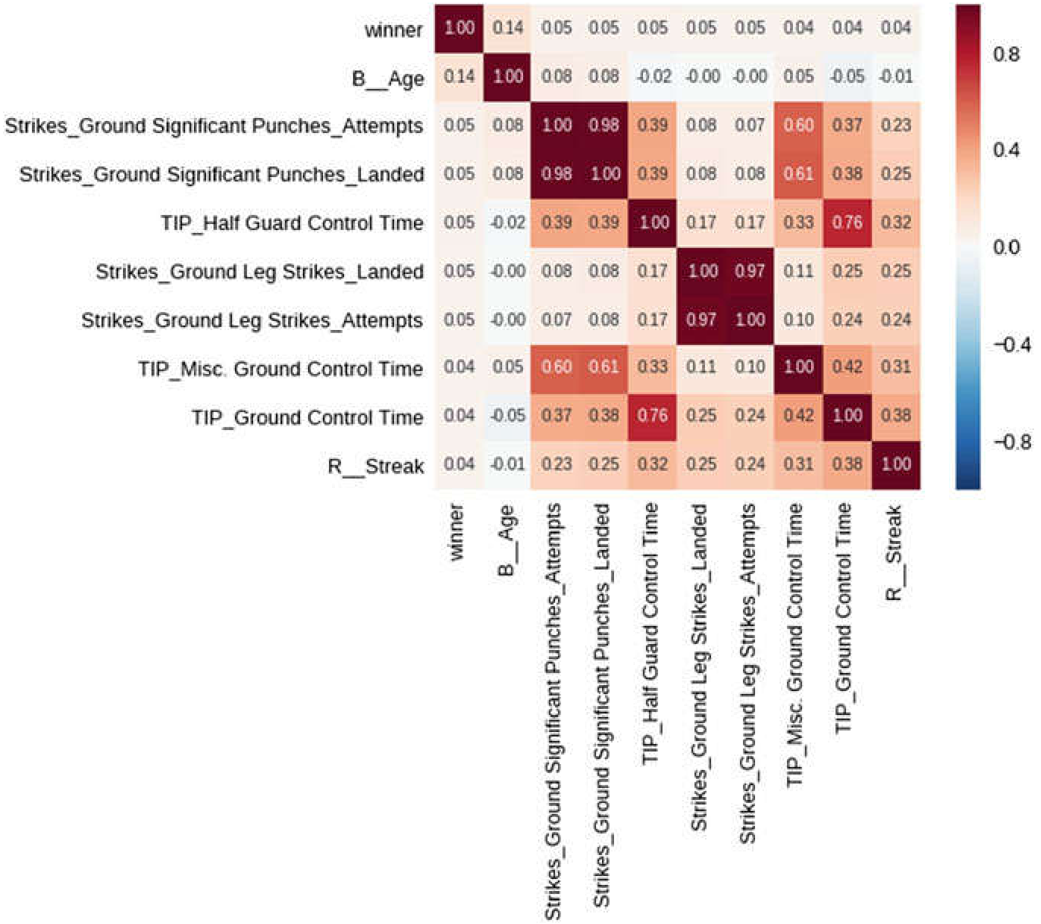

Fig. 2 A heatmap of linear correlations between our target variable, post feature reduction by summing rounds

Table 1 Prediction accuracy of our machine learning models on the data set before any feature manipulation

Model	Prediction accuracy
KNN	0.554054054054
Decision tree	0.533783783784
SGD classifier	0.530405405405
Random forests	0.581081081081
SVM	0.628378378378
Bayes	0.35472972973
Perceptron	0.537162162162

very poor results does not show good result. The dataset itself has much room for improvement, and the assumption that all fighters start from 0 in 2013 coupled with the rise in debut fights for new fighters means that our dataset is very sparse. However, from simply examining the dataset, one can easily see that factors such as fighter age are very relevant to the eventual winner of the fight. Moreover, the Red Side Fighter

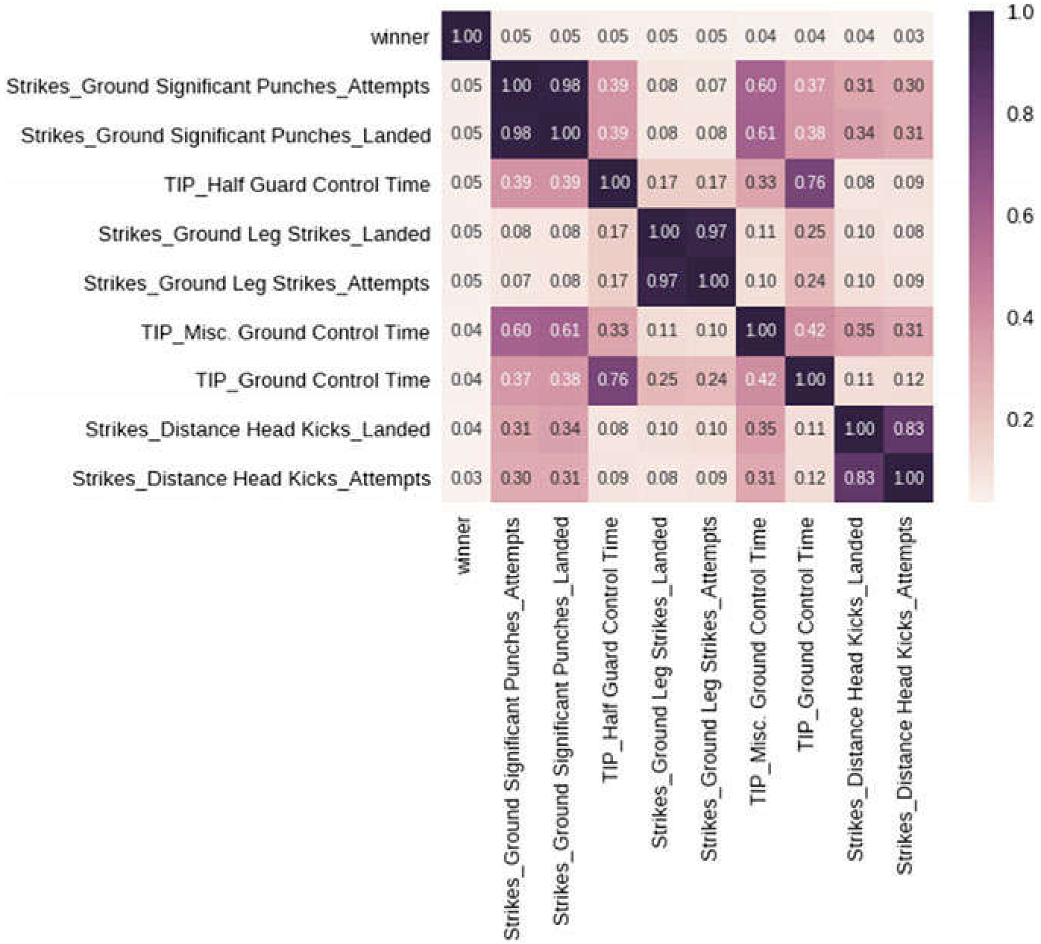

Fig. 3 Correlation matrix heatmap post feature reduction by taking the ratio of features amongst red and blue fighters

Table 2 Prediction accuracy for each of our models upon the dataset with summed features

Model	Prediction accuracy
KNN	0.557432432432
Decision tree	0.516891891892
SGD classifier	0.550675675676
Random forests	0.584459459459
SVM	0.577702702703
Bayes	0.202702702703
Perceptron	0.557432432432

tends to win more frequently. Depending on the model and feature, we exhibit about a 3–6% increase in prediction accuracy from zeroR policy. Our best predictive model is SVM by far—using hyperparameter optimization we were able to get very consistent results with a predictive accuracy of 61% and a best observed accuracy of 62.8%.

Table 3 Prediction accuracy of each model on the data post ratio of features

Model	Prediction accuracy
KNN	0.543918918919
Decision tree	0.503378378378
SGD classifier	0.543918918919
Random forests	0.597972972973
SVM	0.611486486486
Bayes	0.212837837838
Perceptron	0.560810810811

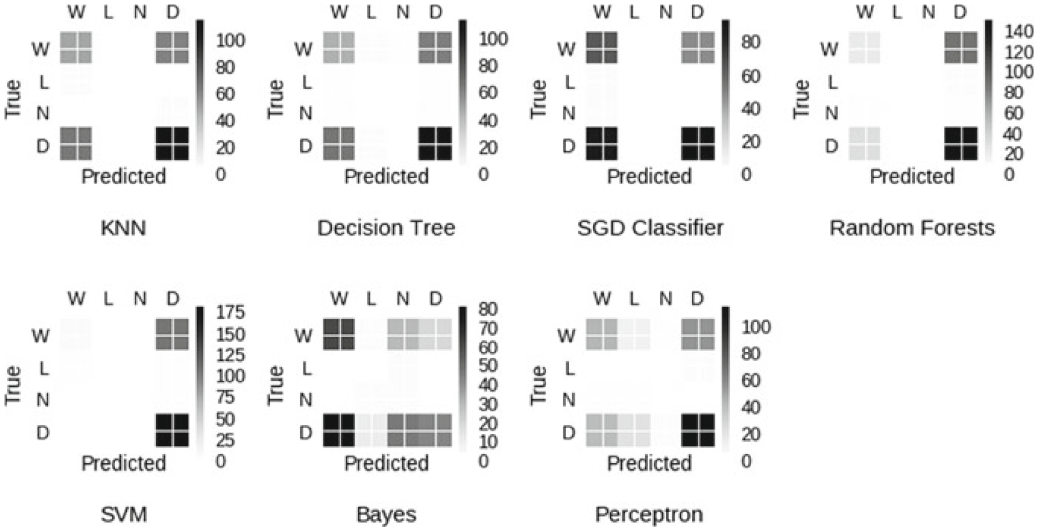

Fig. 4 Confusion matrices for each model on the dataset

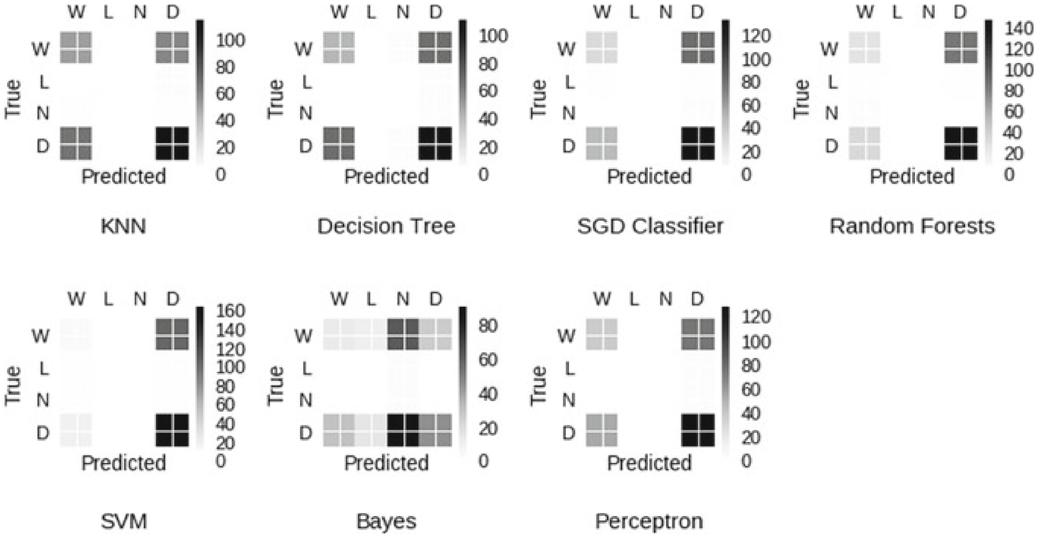

Fig. 5 Confusion matrix for each predictor post feature reduction by summing

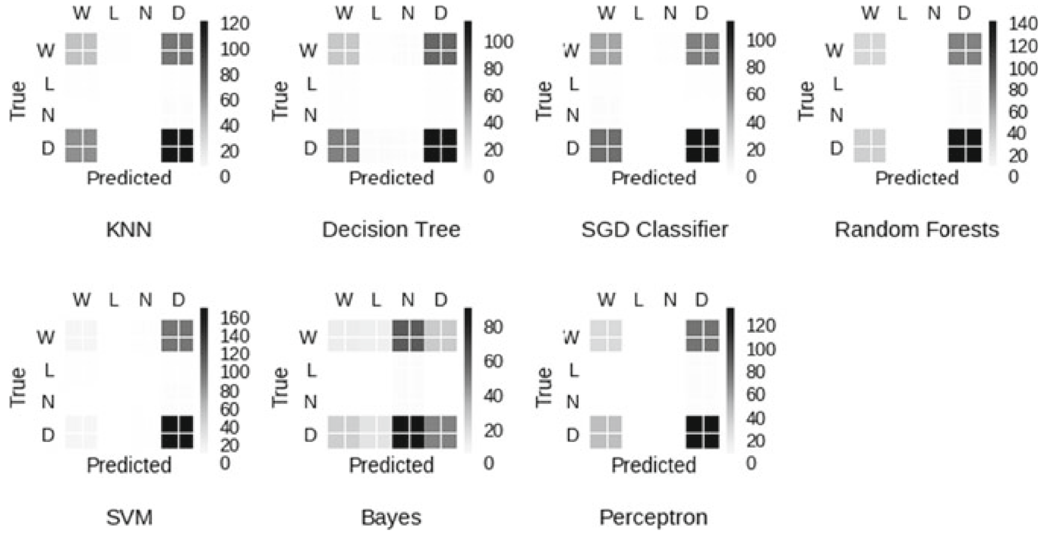

Fig. 6 Confusion matrix for each predictor after all the feature manipulations

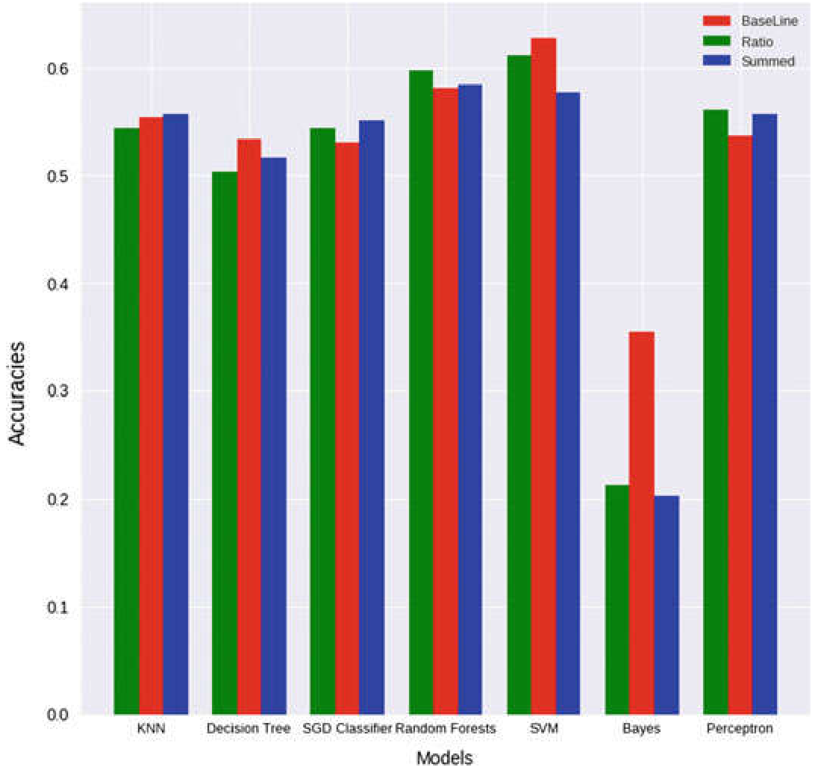

Fig. 7 A bar graph of prediction accuracy of each model over all three sets of data instances, the baseline, the summed rounds and the ratio of features

Moreover, the robustness of SVM can be validated by the drop in prediction accuracy as the features were reduced.

6 Conclusion

In conclusion, SVM proved to be the most resilient of machine learning models for this type of dataset or problem domain, while we did perform some small amount of hyperparameter optimization and feature engineering, it is worth noting that SVM with the RBF kernel performed very well on the dataset straight out of the box. Thus, for sports where a lot of statistical data is not available, it might be a very valuable classifier. In the future, one can also employ some sort of feature selection mechanism to reduce the overfitting in the dataset.

References

1. Johnson, J.D.: Predicting outcomes of mixed martial arts fights with novel fight variables. Master Thesis, University of Georgia, Athens, Georgia (2012)
2. Gift, P.: Performance evaluation and favoritism: evidence from mixed martial arts. J. Sports Econ. (2014). https://doi.org/10.1177/1527002517702422
3. Collier, T., Johnson, A., Ruggiero, J.: Aggression in Mixed Martial Arts: An Analysis of the Likelihood of Winning a Decision. Violence and Aggression in Sporting Contests: Economics, History and Policy, pp. 97–109 (2012)
4. Betting on UFC Fights—A Statistical Data Analysis, https://partyondata.com/2011/09/21/betting-on-ufc-fights-a-statistical-data-analysis, last accessed 12 June 2017
5. Goel, E., Abhilasha, E.: Random forest: a review. Int. J. Adv. Res. Comput. Sci. Software Eng. **7**(1), 251–257 (2017)
6. Bhavsar, H., Panchal, M.H.: A review on support vector machine for data classification. Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) **1**(10), 185–189 (2012)
7. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B.: A review of classification algorithms for EEG-based brain–computer interfaces. J. Neural Eng. **4**(2), R1 (2007)
8. Kotsiantis, S.B.: Decision trees: a recent review. Artif. Intell. Rev. **39**(4), 261–283 (2013)
9. Kaur, G., Oberai, N.: A review article on Naïve Bayes classifier with various smoothing techniques. Int. J. Comput. Sci. Mobile Comput. **3**(10), 864–868 (2014)
10. Lessmann, S., Sung, M., Johnson, J.E.: Alternative methods of predicting competitive events: an application in horserace betting markets. Int. J. Forecast. **26**(3), 518–536 (2010)
11. Lock, D., Nettleton, D.: Using random forests to estimate win probability before each play of an NFL game. J. Quant. Anal. Sports **10**(2), 197–205 (2014)
12. Bottou, L.: Large scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010. Physica-Verlag HD, pp. 177–186 (2010)