**Ultimate Fighting Championship (UFC) Result Prediction Machine**

Created By **– Sri Charan Bodduna**

**Abstract:**

Ultimate Fighting Championship (UFC) Mixed Martial Arts (MMA) is the fastest growing sport in the world. It has a huge fan base covering many countries. UFC MMA is a little similar to boxing but has unique rules compared to other fighting sports. In this research work, our aim is to provide a brief understanding about competitive analysis and prediction models on UFC fighters. We perform competitive analysis based on recent performance of the player. With our solution, we benefit winning strategies for players and business stakeholders. As we have a large number of variables in the data, we first reduce them with different approaches and do analysis using various models like Decision Trees, K-NN Classifier, Support Vector Machines and Random Forest algorithms to find the best model out of these with higher accuracy. Our model predicts the win or loss results of a match based on various parameters like player performance, characteristics and statistics. Our work is about developing a model which categorises players having chances of being top player of the year. This can also help in futuristic training strategies and game plans for future fights. To accomplish this, we are using Python language, Jupyter Notebooks to build our models. This article also aims to provide reports about the results.

**Keywords:**

Ultimate Fighting Championship (UFC), Mixed Martial Arts (MMA), Machine Learning (ML), Decision Tree (DT), K-Nearest Neighbour(K-NN) and Support Vector Machines (SVM)

**Introduction/Background:**

The Ultimate Fighting Championship is an American mixed martial arts promotion company based in Las Vegas, Nevada. It is the largest MMA promotion in the world as of 2022. With a TV deal and expansion in Australia, Asia, Europe, and new markets within the United States, the UFC has achieved greater mainstream media coverage. It earned US$609 million in 2015, and its next domestic media rights agreement with ESPN was valued at $1.5 billion over a five-year term. In April 2023, Endeavor Group Holdings announced that UFC would merge with the wrestling promotion WWE to form a new public company majority-owned by Endeavor, with Vince McMahon serving as an executive chairman of the new entity and White remaining as UFC president. The merger is expected to be completed in the second half of 2023.

Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behaviour. Artificial intelligence systems are used to perform complex tasks based on mathematics and statistics in a way that is like how humans solve problems. These predictive modelling has its applications in diversified sectors from health care to space technology that helps in decision making on futuristic problems. A system can reach to the point where humans cannot go in finding out crucial points about applications. ML is divided into distinct kinds such as Supervised Machine Learning, Unsupervised Machine Learning and Reinforcement Learning. Supervised learning (SL) is used solving problems where the available data consists of labeled examples, meaning that each data point contains features (covariates) and an associated label. Unsupervised learning allows the system to identify patterns within data sets on its own and predict that are neither classified nor labelled. Reinforcement Learning is a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.

Diagram

Description automatically generated with medium confidence

We use various approaches for feature selection and exploratory data analysis.  We use different ML algorithms like decision tree, K-NN, support vector machines (SVM) etc. for development the model. As we need to train and test the model, we divide the data into training data set and test dataset. We train the machine learning algorithm with training data and develop a model. In the article, we use multiple algorithms to find the best model that is suitable for the dataset we use. After we find the model, next step is to validate the finalized model with testing data set. We find the model which results highest accuracy and other parameters which will be described in further steps. Accuracy is defined as percentage ratio between results correctly predicted and total number of observations. For model to be best, the accuracy rate must be great, which means the number of observations correctly predicted must be higher. The other tools which are used for evaluating the performance of the model are R-square or Adjusted R-square and mean square errors for regression analyses and confusion matrix used for clustering or classification algorithms.

We are going to implement the model based on Machine Learning on UFC data set for predicting match results based on key attributes and multiple parameters. We follow proper ML development life cycle to achieve the results, which means we collect the raw data (which is of csv file) into RapidMiner tool, we clean it to remove unnecessary data from the input data set, we next split the data into training and validation sets and apply training set to prediction algorithm. Then we analyse the performance of the developed model using test set and interpret the results.

Diagram

Description automatically generated

**Project Objective:**

Every business needs a solution which can be a torch bearer in their decision making. Using machine learning, we provide insights on the data which forms a supportive guide for any intuitive decisions made. With our model, we try to introduce a platform for business stakeholders of the UFC sport which can be beneficial in taking strategic commitments and selections in the game.

This model adds an extra layer of confirmation for affirmations made by businessmen. Our predictive model will be trained using wide range of significant attributes and complex problems and produce a simple answer about result of the match either win or loss.

**Data/Problem Analytics:**

Machine learning models can predict with good results only if we have flawless data for the solution we wanted to achieve. We are using Ultimate UFC dataset which is available publicly on Kaggle website. This data set is created based on the different data sources of UFC like ufcstats.com, bestfightsodds.com and other Kaggle data set named ufc-rankings. This data set has most of the required attributes which are essential for our predictive data modelling.

Below table describes about each attribute in the data set we used for the application.

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Prefix 'R' | Which tells Red corner player in the match |
| Prefix 'B' | Which tells Blue corner player in the match |
| fighter | Name of the Fighter |
| age | Age of the player |
| date | Date of the fight |
| location | Location of the fight |
| country | Countty of the fight |
| Winner\* | winner of fight either Red or Blue Player |
| title\_bout | binary attribute whether match is title match or not |
| Weight\_class | Weight class of the game(underweight etc) |
| gender | gender of the player |
| no\_of\_rounds | number of rounds in the match |
| current\_lose streak | number of serial loses by player at present |
| current\_win\_streak | number of serial wins by player at present |
| draw | number of draws of the player in UFC |
| Avg\_SIG\_STR | number of significant strikes |
| Avg\_SUB\_ATT | number of submissions attempted |
| TD | number of takedowns |
| rounds fought | number of rounds fought |
| win\_by\_decision\_majority | number of wins by judges majority decision |
| win\_by\_decision\_split | number of wins by judges split decision |
| win\_by\_decision\_unanimous | number of wins by judges unanimous decision |
| win\_by\_KO | number of wins by Knockouts in UFC by player |
| win\_by\_Submission | number of wins by Submission in UFC by player |
| win\_by\_TKO\_Doctor\_Stoppage | number of wins by Doctor Stoppage in UFC by player |
| wins | number of wins of the player in the UFC |
| loses | number of loses of the player in the UFC |
| Stance | stance/style of the player |
| Height\_cms | height of the player in cms |
| Reach\_cms | reach of the player in cms |
| Weight\_lbs | weight of the player in lbs |
| Player\_WeightClass\_Rank | Rank of the player at different wightclass ranks |
| Finish | Win by decision |
| Finish Round Time | Time at fight finshed |
| Finish Round | Fight finished round |
| total\_fight\_time\_secs | Total Fight time in seconds |
| odds | this is related betting not used in our model |
| ev | this is related betting not used in our model |

\* The attribute represented with asterisk is a target variable of the model.

There are a few other attributes in the data which we are not using them for our analysis and prediction those are not included in the above table.

**Descriptive Analysis:**

Before we deep dive into prediction modelling, we need to first understand about attributes present in the data set clearly and take initial conclusions for further explorations and modelling. We also need to pre-process the data before modelling which can be explained in next section. To begin with, we understood total count of fights and count of blue player and red player won the fights along with the percentage share in the data set present. We have total of 4896 observations in the dataset with 2037 wins by blue player and 2859 by red player.

|  |  |  |
| --- | --- | --- |
| **Target** | **Count** | **Share** |
| Blue | 2037 | 42% |
| Red | 2859 | 58% |
| Grand Total | 4896 | 100% |

**Data Cleaning:**

In this step, we perform data cleaning task on data set. It is important that we need right set of data for modelling for more accurate results. Data cleaning process include multiple processes applied on data like data filtering, changing data types, remove duplicates, handling null or NaN values etc.

**Removing unwanted attributes:**

Firstly in this model, we have implemented selecting required attributes. As the data set consists of unnecessary attributes which are not required hence we need to remove them. We have used Select Attributes operator to do this job and included attributes only required.

**Removing missing Values:**

Our data used have few missing values in it we need to handle them carefully so that our model does not produce unambiguous results. We have resolved missing values by applying average of the values present in that column.

**One Hot Encoding:**

Next, we found the categorical variable available in the data set which we need to encode them into numeric values. We found 10 attributes which are of type category using RapidMiner tool, we can ignore winner attribute here since it is our target variable.

Table

Description automatically generated

There are multiple encoding methods available in Machine Learning. We will be using One Hot encoding and Label encoding. One hot encoding is a one of the techniques for converting different categories into numerical values. It works best when there are a smaller number of features in the attributes hence, we use one hot encoding on gender and title\_bout, Weight\_class, R\_Stance and B\_Stance. Other attributes like Location, Date, R\_Fighter and B\_Fighter can be removed since they can lead to abnormal number of attributes generated in the data set. At this point, we have 4896 observations and 72 attributes in the data set.

Graphical user interface, application, table

Description automatically generated

**Train and Test data split:**

As explained earlier, we need training data set for developing the model and testing data set for validating the performance of the model we developed. We kept the ratio of training to test data at 70:30. Which means 70% of input dataset is partitioned as training and 30% as test data.

**Predictive Modelling:**

With above selected attributes from PCA analysis, we are now going to apply them to model.

We will be using multiple ML models and compare their results and find best model for this dataset. Below results are generated based on test data applied to the models developed for evaluating the performance.

**K-NN algorithm:**

KNN is a supervised ML algorithm that can be used for classification or regression tasks. This model is based on the finding observations closest to a given data point are the most similar observations in a data set, and we can therefore classify unforeseen points based on the values of the closest existing points. By choosing K, the user can select the number of nearby observations to use in the algorithm.

Diagram, schematic

Description automatically generated

We got very less accuracy using K-NN model for this data. Below is the confusion matrix of the results. We

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Actual Red** | **Actual Blue** | **class precision** |
| **Predicted Red** | 570 | 401 | 58.70% |
| **Predicted Blue** | 288 | 210 | 42.17% |
| **Class recall** | 66.43% | 34.37% |  |

With, Accuracy: **53.86%**

F-Score: **0.378725608**

Precision: **42.17%**

Recall : **34.37%**

Decision Tree:

A decision tree is one of supervised learning algorithms that is used in machine learning to model and predict outcomes based on input data. Each internal node represents a judgment call or test on a particular feature or attribute, each branch reflects the result of that judgment call, and each leaf node represents the ultimate judgment call or prediction. Decision Tree is simple and can be used for any kind of data.

We now used decision tree model for the data set and genereated the results and interpreted in confusion matrix.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual Red | Actual Blue | class precision |
| predicted Red | 856 | 610 | 58.39% |
| predicted Blue | 2 | 1 | 33.33% |
| class recall | 99.77% | 0.16% |  |

Accuracy: **58.34%**

F1-Score: **0.003184712**

Precision: **33.33%**

Recall: **0.16%**

**Support Vector Machine:**

A supervised machine learning model called a support vector machine (SVM) employs classification techniques to solve two-group classification problems. An SVM model can classify new text after being given sets of labelled training data for each category.

They offer two key advantages over more recent algorithms like neural networks: greater speed and improved performance with a small number of samples (in the thousands). As a result, the approach is excellent for text classification issues, where it's typical to only have access to a dataset with a few thousand tags on each sample.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual Red | Actual Blue | class precision |
| predicted Red | 772 | 495 | 60.93% |
| predicted Blue | 86 | 116 | 57.43% |
| class recall | 89.98% | 18.99% |  |

Accuracy Score: **60.45%**

F-Score: **0.285421539**  
Precision: **57.43%**

Recall: **18.99%**

**Findings/Conclusions:**

We have modelled UFC data using various ML algorithms like K-NN, Decision Tree and Support Vector machines to predict the result of the UFC fight match. Our model has been developed based on training data and validated using test data set. The predicted results are decent when it comes to accuracy with minor variability in other performance parameters like F1-Score and recall.

Accuracy has been improved from K-NN to SVM algorithm with a growth of almost 7%. We found that SVM model have outperformed than other two models with better accuracy and F1-score. F1-Score of the K-NN and SVM is convincing however it is very low with decision tree model. Also, decision tree model resulted very low recall rate of the data hence decision tree performance is worst. We can improve accuracy levels even more by using feature selection algorithms like Principal Component Analysis (PCA) etc. which can be illustrations for future work.

Our model has its benefits of win/loss prediction of the UFC fight in the prospect of player statistics and performance. We also considered wide range of attributes for analysis like age, total wins by player, type of match (Weight Class), Player Stance. We have also encoded and normalized the data set before developing a model. This will keep data aligned across numeric and categorical data. These predictive modelling have covered decent quality of UFC features and player characteristics hence we conclude that our has great advantage.

**Managerial Implications:**

The model we developed can predict the result of the UFC fight which happens in the future by considering various attributes of players characteristics, stats and fight history, game parameters etc. Since we have developed model based on historical data this can help managers and stakeholders in decisioning and predicting results of models. “Humans tend to make mistakes, but machines don’t.” By using this model, UFC stakeholders can take highly precise decisions to earn more profits and promote their business in the right path. It can also provide direction for marketing, TV rights and sponsorship.

**Idea Sharing:**

Developing machine learning model based on UFC data set have many outcomes for sharing ideas. Compiling data on a fighter's height, weight, reach, win/loss record, age, and fighting style can aid in building a dataset that can be used to forecast upcoming matches. You may compile information on previous UFC bouts and their results, including the victor, the loser, the manner of victory, and the round in which the battle concluded.

**References:**

1. “Machine learning: Trends, perspectives, and prospects” by M. I. Jordan And T. M. Mitchell
2. Data is downloaded from Kaggle - https://www.kaggle.com/datasets/mdabbert/ultimate-ufc-dataset?resource=download
3. Ultimate Fighting Championship, U. (2020). UFC stats. Retrieved November 18, 2020, from <http://ufcstats.com/statistics/events/completed>
4. Machine learning applied to a UFC match database by Erick Axel Martinez-Ríos

**Appendix:**

Our model is developed in Rapid Miner tool by downloading the data set from Kaggle website. We have used input data csv file named ufc-master.csv in the modelling. The results are written in UFC\_prediction\_results.csv.