**THESIS PROPOSAL**

**JAVID MURADOV (570933)**

**MSc Business Information Management**

**Introducing Machine learning to MMA: A predictive analysis of UFC fight outcomes**



Author: Javid Muradov

Student number: 570933

Supervisor: Matthijs Wolters

Co-reader: Otto Koppius



**Table of Contents**

**1 Introduction3**

1.1 Preface3

1.2 MMA Overview4

1.3 Sports analytics & The power of predictive analytics6

1.4 Problem Statement & Motivation7

1.5 Research questions & Knowledge gap identification8

1.6 Relevance9

**2 Literature review & Theory10**

2.1 Predictive analytics in sports10

2.2 Predictive analytics in combat sports12

2.3 MMA analytics13

2.4 Literature summary & Contributions15

**3 Methodology & Data16**

3.1 Research strategy16

3.2 Conceptual framework, Modeling & Evaluation18

3.3 Available data sources & features21

3.4 Timeframe22

**Abbreviations23**

**References24**

**1. Introduction**

**1.1 Preface**

In the 21st century, people who advance in the technological and analytical background have united with the athletes and sports organizations in the professional realm of sports. Soccer, basketball, baseball, etc. are using sophisticated data-driven approaches and machine learning algorithms to gain a competitive advantage both on and off the pitch. Advanced wearable tracking systems and ball trackers are utilized to obtain precise on-field data about players and ball movements to excel the tactical analysis and increase team performance. Club managers and coaches, therefore, assess athletes’ effectiveness, build optimal team formations and improve team performance, while scouts and agents have access to a wider range of players to select from. There is also a business perspective that can be boosted using off-field data analytics. For instance, the analysis of fan behavior, social media engagement, ticket pricing and sports betting is held to leverage the business performance (Maryville, n.d.).

Nevertheless, what is even more fascinating about advanced analytics and data-driven modeling is their ability to predict the future. Predictive analytics has been around since 1940 (Van Rijmenam, 2013). However, the scarcity of data points made the predictions unreliable. The intuitive approach of athletes, coaches, and scouts either in the selection or in training processes has always played a decisive role in whether to apply a particular training method, transfer a certain player, substitute the player at a given point in time, etc. Before the emergence of data mining techniques, the decisions were made based on the experience of the decision-makers mentioned above. Same as in all fields of industry, sports has also experienced an abundance of data and numbers in the last decade. However, without the ability to translate these data into insightful information for the stakeholders, these numbers are useless.

Sports analytics is a subfield of data analytics helping to transform these statistics into a piece of information useful for managers, coaches, players, etc. (Athithya, 2019), while sports analysts are the people fulfilling this mission. Nowadays, analysts use predictive analytics to forecast the injury of a certain player, goal probability, potential way of finish of a fight or win/loss situation of an athlete or a team. The recent popularity outburst in the world of Mixed Martial Arts (MMA) made this field specifically urging for predictive analysis. This Master’s thesis will strive to translate currently available data on the most popular MMA promotion named Ultimate Fighting Championship (UFC) fights into fight outcome and bout resolution predictions, as well as provide insightful information for sports professionals, athletes, bookmakers, bettors and fans. Since the decisive result in a fight can occur at any time, prediction of UFC fights outcomes makes it a challenging objective. On the other hand, this is what makes this research rigorous as well.

The global rise in interest rates for MMA has created not only an opportunity but also an intrigue to discover this industry from a data scientific standpoint. More attention to MMA caused more data points to become available to the stakeholders. One of the most prominent MMA promotions, UFC, had a revolutionary impact on the development of this sports discipline. UFC is a global combat sports organization. Until now 28 countries and 157 cities have hosted UFC events while the viewership audience is generated by 165 countries (IMGArena, n.d.). While soccer and basketball associations are still dominating, UFC has already overpassed such prominent associations as NFL (National Football League) and NHL (National Hockey League) by the number of followers in social media sites, such as Facebook and Instagram. As such, more sources of data created an opportunity for more rigorous research of the industry.

Machine learning (ML) being an inalienable part of Artificial Intelligence (AI) is another crucial factor creating a possibility for such research. The large availability of data in such spheres as chemistry, biology, engineering, etc. has substantially boosted the application of ML techniques in a vast range of industries such as healthcare, autonomous driving, and so on (Schmidt et al., 2019). Nowadays, with the digitization of everything, the sports industry has also given a way to ML integration.

The academic interest of this research paper is evoked by the application of the above-mentioned ML algorithms to the popularized UFC bouts which can be used in a number of institutional, athletic, and industrial spheres. The institutional base of the research is largely backed up by the utilization of innovative data science models to such a novel field of analytics which still has a large base for improvement. The athletic basis is represented by the people of sports involved in it directly, such as athletes and coaches, or indirectly, such as scouts and sports analysts. And, the last but not least part of this study might be impactful to the business field representatives associated with UFC organization, bookmaking agencies and bettors themselves.

**1.2 MMA Overview**

*“I’m a martial artist. I don’t train for a fight.”* These words belong to George St Pierre (St-Pierre, G., n.d), one of the greatest Mixed Martial Artists in history. Mixed Martial Arts is often interpreted as a street fight. Nevertheless, most of the time such an expression is used by the people dissociated from sports. The respective naming of the discipline is for a reason. Mixed Martial Arts is a combination of multiple sports disciplines, such as boxing, kickboxing, karate, judo, jiu-jitsu, wrestling, etc. (Britannica, 2016). A number of technique variations inherited in MMA from the aforementioned martial arts both in standing and ground positions are used to gain a victory over an opponent. Therefore, Mixed Martial Arts cannot be referred to as the street fight, where participants are not limited by a set of rules and are not using any specific technique to conquer. This is a professional sport where professional athletes are sharpening each part of their fight strategy to gain a competitive advantage in the octagon.

UFC bouts take place in the 8-sided polygon cage, also referred to as octagon (Mahon and McGowan, 2015). The majority of the fights consist of three 5-minute rounds, while the main events of the evening go a long distance of five 5-minute rounds where the last 2 rounds are called championship rounds. The fight lasts for full rounds in case it is not stopped by the referee beforehand. There are several ways to end an MMA fight, which are submission, knockout, technical knockout, or decision (UFC, 2018). In case of a decision result, a 10-point must system is applied by the referees based on the striking and takedown accuracies, transitions in the grappling, and overall activity of the fighters. As such, 10 points are assigned to the winner of the round, whilst 9 or fewer points are given to the less advantageous fighter. In the event of even distribution of the above-mentioned determinants of the round outcome, 10 points are allocated for both fighters. Finally, each rounds’ scores are summed up to announce a winner of the UFC bout (Combat Museum, n.d.).

The popularity of MMA is overwhelming. MMA is the third most popular sport in the world after soccer and basketball (Forbes, 2018). There are almost half a billion MMA fans (Forbes, 2018) who contribute to such MMA promotions as One Championship and UFC to take the leading positions in sports viewership statistics. According to Tubular Labs, a major data analytics platform officially certified under YouTube Measurement Program, the viewership numbers for the above-mentioned MMA promotions reached up to 9.4 billion views in total (Sportskeeda, 2020).

As regards the economic performance of the industry, the average industry growth for Martial Arts Studios in the US for the years 2014-2019 has grown up to 3.7% (IBISWorld, 2020a). To compare, the respective index for such business spheres as IT Consultancy and Real Estate is set at an annual rate of 3.2% and 2.6%, accordingly (IBISWorld, 2020b, 2020c). One of the most popular MMA promotions, UFC, once bought for $2 million in 2001, has reached a net worth of $7 billion, which is more than two times higher than the most valuable European soccer club of Real Madrid (RealMadrid, 2020), while the revenue for the last financial year only is equal to $700 million (Forbes, 2018).

The history of MMA takes its roots from 648 BCE when it was part of the Olympic games and considered as pankration, the ancient name of Greek combat sport. However, MMA, as we are used to seeing it today, started its evolution in 1993, when the first UFC event took place (SMASchools, 2019). Since then, the promotion and the sports discipline itself has had a huge recognition, while the number of fans and people practicing it has been continuously growing. Ultimate Fighting Championship had tremendous success for the past couple of decades with a steep increase in the number of pay-per-view (PPV) events and yearly sales.

There are several factors amongst the reasons for such instant growth of UFC in the above-mentioned numbers: the UFC organization, deals and sponsorships, social media, sports betting market, etc. (Orland, 2020). The involvement of various parties mentioned above made UFC a powerful business machine generating an immense amount of money. Contracts with huge broadcasting organizations as ESPN valued at $1.5 billion (Forbes, 2018), media platforms, fan engagement, PPVs, ticket sales, commercials, etc. are also representative of large business involvement and implications of this industry. Nevertheless, the most essential “why and how” of UFC's success story is, certainly, the fighters and their contests in the octagon. Apart from being a professional athlete, every fighter himself is a brand. Some of them are even more popular on social media than the UFC organization itself. Khabib Nurmagomedov, Conor McGregor, Ronda Rousey, Anderson Silva and Israel Adesanya are only some of the names, but they have millions of followers and high attention from the public (FightBookMMA, 2020). The performance of the UFC organization substantially depends on the performance of these people.

**1.3 Sports analytics & The power of predictive analytics**

The field that is also worth mentioning and investigating is sports analytics. According to Holman, V. (2018), sports analytics is the practice of application of advanced statistical models and methods on certain data for the purpose of gaining a competitive sports advantage. Analytics was always part of professional sports either in the performance-oriented direction improving results of teams and athletes or in the business-oriented applications such as increasing fan engagement, sponsorship agreements, and formulation of ticket pricing. With the increasing speed of technological advances, more data has become available to the sports associations, organizations, clubs, teams, athletes, and the public, overall. The collision of sports and data caused the emergence of such an innovative and unorthodox field as sports analytics. Sports analytics ranges from proper data management and analytical models applied to the data to the grounded decision-making process. As mentioned above, sports analytics involves a deep understanding of statistical and mathematical principles for exploratory data analysis (EDA) fulfillment, which will be touched upon in this Master’s thesis for discovering the patterns in the features of UFC fighters, their performance, and results.

Nevertheless, one of the most challenging but engaging parts of sports analytics is outcome predictions. Predictive analytics is a wide variety of machine learning models and statistical algorithms applied to historical data in order to determine the likelihood of future outcomes (Edwards, 2019). The science of predictive analytics is helpful in a number of spheres and industries:

1) *Healthcare*: Analytics are used to analyze the patterns of patient data to forecast organism deterioration, probability of recovery, treatment approaches, etc. Moreover, analytics are also utilized to prevent the downtime of healthcare facilities and equipment (Philips, 2020).

2) *Retail*: Analytics can be used in targeted audience optimization of the products based on current consumer behavior, improvement of business processes, and expectations of customers. Data for “product wrapping” purposes (Kumar, 2019).

3) *Telecommunication*: Analytics are used to predict customer churn based on supervised and unsupervised learning of the historical data, optimize marketing strategies and recommendation engines, forecast optimal pricing.

4) *Financial services and banking*: Analysis of data patterns of current and future customers can be implemented. Organizations can improve risk-averse loan management, as well as formulate the impact of new rules and regulations on market trends (Kumar, 2019).

5) *Manufacturing*: Companies are able to calculate the ways of improving internal business processes, avoid machine and equipment malfunction.

These industries are only some of the examples where predictive analytics can be implemented. The final goal of all businesses from using predictions is gaining a competitive advantage. The same approach is executed in sports analytics.

**1.4 Problem Statement & Motivation**

The recent advances in the field of analytics have made its implementation realizable to the area of sports. Along the lines of the business domain, predictive analytics in sports can be utilized in various ways for various purposes. UFC, being an attractive sports domain and comparatively available from the statistical data viewpoint, makes this organization and the organized events tempting to research in terms of data scientific scope.

Whereas business executives are interested in business-oriented data and strategies, athletes and their coaches are eager to learn which game plan to pick for improved performance optimization. On one hand, success is an easily determinable definition, especially when you are performing in one of the most elite MMA promotions in the world, or you are a world champion already. On the other hand, however, success can be hardly measured when the athlete does not know the limits of his best performance level or his objective awareness about getting better from day-to-day. The current study will assist to explore the implications and key performance indicators of this sports discipline based on the obtained data.

From an academic research perspective, several research papers and predictive modeling on the UFC data have uncovered certain trends, explored hidden facts, and only a few of them tested some of the widely used statistical models for fight outcome predictions. Since most of the respective studies have incorporated only an exploratory data analysis, while others applied a separate statistical method or a single predictive model without the possibility of provision of any comparison between the models, the literature on the unified juxtaposition of these models, their performances, as well as the introduction of novel models based on available data points with the involvement of sports betting market is highly sparse in its academic sense. As opposed to these studies, the current Master’s thesis has focused on the implementation of multiple statistical and ML modeling techniques and the comparison of these models among each other, coming up with the best performing algorithm. Hence, the objective of this study is to derive and apply the technique, which will be discussed later in the Conceptual framework, Modeling & Evaluation (chapter 3.2), with a higher prediction quality juxtaposing several models based on the fights’ and fighters’ feature variables, the difference of historical averaging and characteristics methods of two fighters or their rating standings. Moreover, feature engineering, which is considered as one of the most essential determinants of the built models’ performances and has a multiplicity of vectors to be designed, will be embraced by the unorthodox variables allowing the model to predict the fights before the fight has happened in reality.

Furthermore, this research will explore the constructed models from the sports betting perspective. “The dealer always wins” is a widely popular phrase in the field of gambling. The reason is that the house always has a statistical advantage over the risktaker. If the betted amount is constant, the expected return for the bookmaker is higher eventually as the odds are “unfair” for the bettor in most of the cases (Bartels, 2019). The same conception is also applicable to sports betting. Bookmakers are setting their odds by assessing the outcome probabilities of the given match just as this study will do. It has to be highlighted that the omniscient bookie includes a wide variety of factors in their probability predictions ranging from statistical data to sports market data which results in highly precise predictions. Therefore, even though the bettor is able to predict a set of one-off match outcomes better than the bookmaker, it is extremely complicated to compete in the long run. Unalike to other academic studies performed in the UFC arena, this Master’s thesis will strive to assess the result outcomes derived by the built ML models by juxtaposing them with the predicted probabilities of the bookmakers, while the main goal would be to compare the performance and approach the accuracy score represented by the bookies. Since bookmakers display reasonably high accuracy performance in sports predictions, the iterative juxtaposition might assist to increase the predictive accuracy of the built models as well.

Perhaps, one of the most interesting components of the UFC show, which was hardly analyzed within the scope of academic research before, is the ways of finishing a fight. Since the participants of the show come to MMA from different sports disciplines with their own fight style and game plan, each has his/her own favorite technique. For a boxer that might be an uppercut which can result in a knockout, while the grappler can finish a fight with a rear-naked choke. However, the history of MMA, and particularly UFC, has seen vice versa cases as well. The collision of two different schools within a fight and its final result is what makes UFC interesting for the public. Therefore, the current research will not be limited by predicting the resulting outcomes, but also elaborate a more advanced predictive analysis on the ways of finishing a fight with the use of statistical modeling. This novel approach to the fight game from the perspective of potential ways of MMA match resolution will introduce a fundamental contribution to the current academic literature.

**1.5 Research questions & Knowledge gap identification**

Prediction modeling is used in various sports disciplines, such as soccer, baseball, basketball, and so on. Teams and clubs use predictive analytics to forecast win/loss situations beforehand. This practice is also applicable to predict an athlete's injury or future match performance. The respective field has been researched quite thoroughly until now.

However, compared to the aforementioned domains, the UFC organization is comparatively unfamiliar to the stakeholders and has not been explicitly explored from a machine learning perspective yet. Several research papers were performed and non-sizeable articles were written on the topic of UFC fight outcome prediction. Despite this fact, a substantial knowledge gap exists in ways of fight outcome predictions using the ML algorithmic approach. The foremost focus of this research will be aimed at the ability and quality of different machine learning models to predict the results of UFC fight bouts based on a predefined set of variables and fighters’ attributes.

Furthermore, the applied models and results will be aligned with the business-oriented perspective of the UFC as an organization with a large fan base and attention from various industry representatives. The research will try to investigate how the prediction of certain fights will impact the sports betting industry as well as explore its effect on fan engagement and PPV sales. However, the main focus will be directed on the match between the derived predictions and probabilities attained by the bookmakers. Such an alignment will provide a unique specificity to this research paper due to a particular reason of bounded connection between bookies’ result outcomes and predictive sports analytics in existing studies.

Therefore, the first research question of this study will be formulated as follows:

***1) What is the quality of ML and statistical models in predicting UFC fights outcomes and how do the built models perform in comparison with the UFC bookmakers’ predictions?***

While the outcome predictions are the preeminent aim, this study will also dive into the predictions of the ways of ending the fight based on different features. There are several ways to end an MMA fight which are Decision, KO (knockout)/TKO (technical knockout) and submission. The knowledge gap in this area of predictive analysis is tangible as previously existing literature focuses on the binary target variable of win/loss only. As there is no specific grading or hierarchy among 3 match resolutions, they will be introduced to the eventually built algorithms as nominal categorical target variables. Therefore, the second research question of this Master’s thesis is:

2) ***How useful are ML models in predicting ways to end UFC fights?***

**1.6 Relevance**

***1.6.1 Academic Relevance***

Having reviewed the existing literature on UFC fight predictions, it was found that a sizeable knowledge gap exists in the respective field of sports analytics study. The academic relevance of this paper is attributed to the deficient data scientific research of UFC as one of the most prominent MMA organizations in the world based on the available data sources. This research will dive into the predictive power of data science and its application to UFC fights. The analysis will address various state-of-the-art machine learning algorithms on the up-to-date historical dataset, but which will be applicable to the use for future unseen fights, comparing their performance with each other trying to improve the accuracy of the models. Aside from this, the construction of ML models and their application to predict the ways to finish the UFC fight will introduce a new breath into the previous research literature. The current Master’s thesis will strive to explore the insights of each fight resolution which can create a major basis for future analysis.

Furthermore, having reviewed available research papers, blog posts and articles, a very limited alignment of predictive analytics of UFC fight outcomes with the business application represented by sport betting agencies was created. This research will try to compare the achieved model performances with the probabilities derived by the bookmakers and investigate the impact of sports predictions on the sports betting industry. Such an alignment can create a significant value not only for the academic literature but also for the future betting practice on the UFC fights.

Apart from this, a large contribution to current academic literature and a fundamental basis for future exploration might be introduced by conducting a predictive analysis of the ways to finish UFC fights.

Another unique nature of this research is explained by the fact that the analysis will be fulfilled on the basis of the cross-industry process for data mining (CRISP-DM) methodology. CRISP-DM is a proven technology for organizing and performing data science projects. The research will dive into each phase of the CRISP-DM cycle in order to formulate a clear understanding of the dataset, the created predictive models, and its application in UFC betting.

***1.6.2 Managerial Relevance***

There is a number of manager-relevant insights that this research will provide for the stakeholders. First and foremost, as fighters themselves are at the center of this research, predictive analysis of UFC fights based on historical data will create added value for independent athletes, their managers, and coaches. Fight predictions and variable analysis performed via EDA will assist fighters in analyzing their fight tactics for future fights. It will increase their understanding of the past and current fight tendencies with a possibility to discover new strategies and techniques for improved performance in the future.

For UFC as a huge revenue machine, predictions of fight outcomes might create an opportunity and uncover possible ways of target audience expansion and PPV sales increase. Each UFC event is followed by the main and co-main bouts of the evening which makes a specific UFC event even more attractive to the viewer. Pursuing the above-mentioned goals, UFC can plan engaging events beforehand by predicting the results of close-in fights. This will help to assess the public reaction and evaluate its impact on viewership rates.

The managerial relevance of this research might also be attributed to the sports betting market. The market size of the sports betting industry has reached $203 billion in 2020 (Statista, 2021). The increased Internet accessibility during the last decades caused the popularization of the Internet and mobile gambling (Gainsbury, 2015). The application of AI algorithms and ML models is already transforming the way people bet. More data points available to the public allow the afore-mentioned algorithms to analyze and predict outcomes with higher predictive accuracy (SBD, 2018). Considering that the likelihood of UFC fans using different sports betting platforms is 107% higher than the average consumer (INGArena, n.d.), the use of ML algorithms to predict future outcomes can be even more beneficial for bookmakers specialized in the respective sports discipline. Hence, this Master's thesis will strive to investigate the possibility of potential benefits with betting based on the result of predictive analysis.

**2. Literature review**

**2.1 Predictive analytics in sports**

Having referred to the existing literature, an immense number of research papers, articles, and books have been written until now. These works cover various sports disciplines, correlational and causal studies among different athletes, sports teams, fans, performances, predictions, and other similar topics in the realm of sports analytics.

One of the long-standing papers in the field of predictive analysis has been conducted by Boulier and Stekler (1999). The group tried to predict the sports match outcomes by analyzing the rankings and seedings of teams and players. This study was mostly attributed to National Collegiate Athletic Association (NCAA) basketball tournament, as well as the Grand Slam tennis championship. The results of the study have shown that a significant relationship between the difference in rankings of basketball teams, as well as tennis players and the likelihood of winning, exists.

Soccer, being one of the most frequently subjected disciplines in ML, has undergone a large number of studies from the match outcome predictions perspective. It is practically infeasible to list all of the studies performed in this area of research. Nevertheless, the study deserving attention was performed by Ulmer and Fernandez (2013), where the researchers conducted an analysis on predicting soccer match outcomes in the English Premier League (EPL). Using the historical data from 2002-2003 season until 2011-2012 season results as a training dataset and the following two seasons as a testing dataset, the researchers applied a number of AI and ML models, such as Linear classifier, Naïve Bayes, Support Vector Machine (SVM), Random Forest, and Hidden Markov Model. The best performing model, which was a linear classifier with a stochastic gradient descent algorithm, achieved an error rate of 0.48. The results of the models are quite impressive considering the fact that the baseline accuracy is set at 0.33 (win, draw, lose), however, the model drastically underpredicted on draw results. It has to be highlighted that this study set the stage and was used in other studies by Hessels (2018) aiming to improve the models’ accuracy by using a larger dataset and more advanced features which as a result achieved an accuracy score of 0.60.

A different rigorous source that stands out in the soccer domain is Microsoft Bing’s prediction model constructed during the soccer World Cup 2014 correctly forecasting 15 out of 15 knockout matches along with Germany’s win over Argentina in the final (Soper, 2014). Obviously, the prediction model itself was not revealed by the company. However, such a result creates a basis for future application and improvement of predictive analytics in the field of sports.

Same as in team sports, a large number of studies are carried out in individual sports disciplines. The studies performed in individual sports might be even more adjacent to UFC than team sports in terms of feature engineering and EDA stages of this research. Tennis, known as a king of sports, has undergone several exciting academic research studies. Tennis modeling performed by Sipko (2015) is analyzing the tennis dataset containing over 6,315 Association of Tennis Professionals (ATP) matches played during the years 2013-2014 with specific match details from the predictive analytics point of view. The model built by the author allows estimating not only the results of the test data held out from the model but also for the future match outcomes statistics of which are not known beforehand. In order to construct such a model, several averaging methods were applied and the respective difference between two potential rivals was found. Simple historical average difference, common opponents average, splitting by tennis surface are some of the methods used by the author of the study. However, each of the methods has its advantages and disadvantages. Historical averaging is an agreeably straightforward but way too broad approach in predicting the match outcomes since each player can have a different performance against a certain opponent or on a particular surface. It is intuitively obvious that narrowing down the statistics of the players to common opponents or by surface might be more informative than calculating rough averages. Nonetheless, the latter approaches limit the number of observations to a minimum which leaves a small basis for the model to train on. The aforementioned feature engineering approaches can be useful for the purpose of this study as well.

**2.2 Predictive analytics in combat sports**

MMA, being a universal sport, involves a mastery of a wide variety of sports directions. The review of predictive analytics performed in more popular sports such as soccer and basketball might be beneficial to familiarize with the different ways of application of machine learning models, however, it is more logically rational to explore the respective field from the point of view of MMA constituents which are fight sports such as boxing, wrestling, taekwondo and others.

The sport of boxing is one of the most essential elements of MMA as both fighters start their bout from a standing position engaging in the striking exchanges. One of the predictive modeling approaches used by Warnick et al. (2007) was a logistic regression to test the performance of 400 boxing fights within one month period. The regression model was built based on several predictors, i.e. age, change in weight category, total wins and losses per fighter, as well as the result of the preceding match, possession of the title, which might potentially have a psychological advantage over an underdog opponent, and citizenship, which might have a positive hometown influence on the American fighters as all fights were held in the United States of America. The results of the regression have shown significant regression statistics as well as four significant predictors, including age, preceding result and win/loss statistics. The classification matrix reflected a reasonably high accuracy score of 75.5% predicting 302 out of 400 correctly.

Wrestling is another predominant aspect in the technique base of every mixed martial artist. Its importance can be measured by the fact that several Olympic wrestling and judo medalists in the past such as Henry Cejudo and Ronda Rousey have become UFC world champions by dominating in takedowns and groundwork. Unique research grasping the collegiate wrestling rankings and the respective systems such as PageRank and Elo to predict the future outcomes has been introduced by Bigsby and Ohlmann (2017), whereas the aforementioned rating systems are compared with the existing ranking techniques in order to predict the outcomes of a wrestling match. The baseline of the previous prediction performance was set at 67%. PageRank is a system that allocates the power of an individual wrestler by creating a network where each node represents a wrestler and calculates his rank based on the results of the matches. It has to be noted that the statistically significant increase in the predictions using the PageRank over other tournament seeds can be observed. Elo ranking system, in its place, derives the relative power of wrestlers in the process of ranking calculation. All in all, 429 out of 629 (71%) wrestling matches were predicted accurately as a result of Elo rankings incorporation.

**2.3 MMA analytics**

Unlike soccer, basketball, or even tennis, which is less popular than MMA from the viewership perspective, the latter, being a comparatively young sports discipline, has surprisingly been the subject of a limited number of research. From the data science perspective, only a few academic papers can be referred to in this Master’s thesis. However, this can also be justified by the fact that UFC has emerged only several decades ago as a sport that is visible to the public nowadays while the number of data points is relatively narrow. On the other hand, this creates boundless freedom of action and exploration research to be undertaken. Despite being sparse in terms of academic research papers, a number of blogs and articles in UFC analytics are available on open-source platforms.

To start with, UFC Performance Institute (2018) has released a training guide that envisages an explicit range of steps of the fighter from the training process up to the competition stage. More than 30,000 various performance metrics are used by UFC analysts to explore the win variations in UFC, striking accuracies by weight class, injury-related measures, and many more. The analysis provides an in-depth sight not only on UFC and MMA techniques as a whole but also on ordinary body mechanisms of keeping a good posture and injury prevention. The guide can be beneficial for athletes and their coaches in understanding the athletic edge of UFC.

One of the research papers, presented by Johnson (2012) investigates the predictions of the MMA fight outcomes based on the novel fight variables. Similar to current research, the dataset was obtained from UFC Stats which was previously named as FightMetric. The afore-mentioned dataset contains 3638 MMA fights from various MMA promotions as UFC, Pride FC, StrikeForce, King of the Cage, etc. from 1993 to 2012. Since the target variable is binary, logistic regression was applied to the respective dataset as a method for deriving predictions. The accuracy of the constructed model arrived at 63.4% after carrying out cross-validation which is a reasonably impressive performance. The model evaluation is performed by introducing several statistic measures, such as concordant/discordant index, measuring the frequency of correctly predicted events, rescaled R-squared, showing the fit to the regression line, and Brier score (mean squared error) which seems intuitively useful with discrete-binary outcomes.

The predictions in UFC have also been analyzed by another group of researchers whereas a comparative study of different ML models was constructed for the purpose of prior prediction of UFC fights (Hitkul et al., 2018). The authors use the dataset comprising 1477 fights in the year range of 2013-2017 by initially exploring and wrangling the data. It has to be highlighted that as the study aims to provide the prior predictions, it was decided to manipulate the data in such a way that each fighter has real-time statistics before the respective bout. This was done by subtracting the per fight statistics from the fighters’ current statistics. A wide variety of ML models such as Random Forests, Decision Trees, K-Nearest Neighbors (KNN), SVM, Naïve Bayes, Perceptron, Stochastic Gradient Descent (SGD) were applied to come up with the best predictor. As a result, only SVM was able to step over the line of 60% of prediction accuracy, whilst the Random Forests model was also close to this index.

While only part of the fights dataset covering the period from 2013 to 2017 was used in the previous study, McQuaide (2019) from Stanford University was able to collect and use the entire historical data from 1993 until 2019 for ML prediction purposes. 5144 fights were obtained and analyzed for models’ application purposes. As a result of data cleaning procedures, 3,155 complete data points and 134 features were introduced to the ML models. All in all, four classification models were trained and tested on the respective dataset to arrive at an average test accuracy of 60% with Gradient Boosting being the best. During the iterative testing, it was interestingly explored that fighter age is one of the most important features.

One of the recent projects held in the field of mixed martial arts was presented by Lane and Briffa (2020). The group tried to establish the relationship of skill and vigor represented by the percent significant strikes landed and the number of strikes thrown per second with the type of outcome represented by decision when the fight went the whole distance or defeat when the fight result was determined by submission, KO or TKO. The latest 548 bouts (between February 2019 and March 2020) fought by 599 martial artists were run in the 2 generalized linear mixed-effects models. While the first model was executed to analyze the effect on the binary outcome of decision and defeat, the second one incorporated all 3 resolution methods (decision, submission and KO/TKO). As a result, a statistically significant positive relationship between the above-described indicator of vigor was found in the first model, meaning that a more vigorous fighter predominantly won his fight, even though such a tendency was mostly observed when the fight went the full distance. Another observation displayed that a higher skill set of fighters increases the positive interaction between the vigor and outcome. The second model has supported the relation of vigor and decisive outcome. Fighter sex which was another variable used in this study had no significant effect on the bout outcome.

Since this research paper will tackle the business realms of the UFC organization, a relevant MMA-related analytical research that used a statistical data analysis component in predicting MMA fights was introduced in the blog post named “Betting on UFC Fights – A Statistical Data Analysis” (Singh, 2011). A Random Forests classification was trained on the total data instances of 11,886 fights out of which only 1,390 data points were related to UFC promotion. It has to be highlighted that a reasonably high AUC score of 0.69 was performed by the model. Furthermore, having derived an insight that the fighters older than 32 years of age are more likely to lose with a probability of 62%, it was decided by the author to test this hypothesis using a betting factor which resulted in a positive return on investment of 12.6%.

Aggression and violence, as one of the most crucial factors of the overwhelming popularity, made UFC appealing to Collier and Johnson, 2012 as a methodology for predicting UFC fight outcomes. The authors use a vast number of variables impacting the UFC fight outcomes dividing them into two data samples. One of the samples contains the data about the fights ending by knockout or submission, while another dataset contains the fights going the full distance. According to the researchers, the respective division allows concentrating on the features that are more important for the judges in deriving decisions of the fight outcome. The authors observe the difference between the explanatory variables as a base for the modeling stage. The constructed probit models on both samples of data obtained from FightMetric, which contained 946 bouts for early stoppage fights and 323 fights going the full distance, results in the R-squared of 0.634 and 0.701, respectively. The largest marginal effect of both models, which would increase the winning probability of one fighter over the other substantially, is the number of knockdowns. As such, by an increase in the number of knockdowns by one in the early stoppage fights, the probability of winning would increase by 16.6 percent, whilst for the fights ended by referee decisions, the increase would settle at 9.5 percent, which seems instinctive for both datasets. Among fighter descriptive characteristics, the only significant one is height, while the age and weight of the fighter are statistically insignificant. Furthermore, despite the fact that head jabs and ground punches are positively correlated with the winning probability and are statistically significant, but have low marginal effects of 0.1 and 0.6, accordingly.

A correlational study held by Kraus and Chen (2013) investigated the relationship between professional fighters’ smile expressions during a face-to-face event and an actual fight result. Data for 152 unique UFC fighters along with the face-to-face photographs of the fighters before the fight was obtained and analyzed. The central prediction held that the fighters who tend to smile at the pre-fight events are less aggressive and more inferior in their performances than the ones who do not, thus, losing their fights. The study also showed that the occasions of dominant victories, such as knockouts and submissions, are positively correlated with the absence of smile intensity.

Mahon and McGowan (2015) analyzed the demand rates for UFC events and their determinants by conducting an empirical regression analysis on PPV sales. Several determinants of the high PPV sales were identified during the study. One of the highly impactful reasons for the boosted PPV buy rates is the fighters themselves. The authors analyzed the influence of the fighter identity in main and co-main events primarily. Unsurprisingly, Conor McGregor took place on top of the list with the largest draw in UFC history. Incrementally, Brock Lesnar and Ronda Rousey have settled afterward with the sale rate over 400,000. The particular reason for the circumstance is the celebrity of popular fighters participating in the UFC events.

Furthermore, one more research was performed by Kuhn and Crigger (2013) in their book FightNomics. The respective study carries out an exploratory analysis based on the statistics of MMA. The author aims to show the hidden numbers in striking finishes and submissions, jabs and power shots, standing and ground positions, and so on. However, the aforesaid study does not review the fights from the ML modeling domain.

**2.3 Literature summary & Contributions**

As it can be observed from the large basis of research papers and studies performed in sports analytics, the existing knowledge about predictive power and the use of big data is overwhelming in certain circumstances. Football, basketball, tennis, boxing, etc. have undergone a large number of various rigorous analyses and have been explored fairly deep, reaching a reasonably high quality of predictability, which cannot be stated in terms of MMA being a comparatively young domain from popularity and development perspective.

Notwithstanding the challenges of the limited dataset and lack of academic knowledge in this sphere, Johnson (2012), who was one of the first to apply the regression method to UFC, reached quite impressive results for that time and created a firm fundament for future work.

The latest profound MMA-related academic research papers introduced by Hitkul et al. (2018) and McQuiade (2019) are unique representations of ML performance and its quality on the present-time UFC dataset. They presented a comparison of different state-of-the-art methodologies to derive the overall quality of predictions.

Other non-academic research studies, such as Singh’s (2011), performed in this domain are also helpful by providing up-to-date datasets and insights on the UFC fight and fighter features. However, certain gaps that are still unexplored by the above-discussed studies open up the opportunity for future improvement and the incorporation of such improvement to different spheres. Therefore, the contributions to the existing knowledge about predictive analysis in UFC are formulated as follows:

* Introduction of an up-to-date historical dataset (1993-2021) to ML algorithms with a goal of potential improvement of predictability performance within the scope of this study.
* Comparison of the derived ML predictions with the bookmakers UFC fight outcome probabilities.
* Introduction of ML algorithms to a new target variable represented by way of UFC fight resolution (decision, KO/TKO, submission) and assessment of the quality of such predictions.

**3. Methodology & Data**

**3.1 Research strategy**

One of the main goals of this research paper is to analyze the predictive potential of machine learning models on historical data of UFC fights from 1994 until 2021. As mentioned before, the research will follow the CRISP-DM methodology and its principles. CRISP-DM cycle is the leader for data mining projects in terms of distinctive practicality and optimized sequence of phases.

There are 6 main phases of the CRISP-DM cycle depicted below:



Figure 1. CRISP-DM cycle (Smartvision, n.d.)

*The business understanding* phase refers to understanding the nature of UFC as an organization. This step is extremely important as the main goals, fundamental requirements, and future accomplishments of the research are established herein (Provost and Fawcett, 2013). Moreover, the respective phase is always reversed back to upon the deployment of the predictive model in order to assess the evolutional effect of the created model and its application to the dataset. The following phase of the CRISP-DM cycle is *Data understanding*. Data understanding refers to the raw data collection and initial cognizance of the dataset (Provost and Fawcett, 2013). The research will strive to understand the key attributes of the scraped data, such as the number of observations, its structure, and format. EDA process helps to create working chemistry between the data and the researcher. The next step of the cycle is referred to as *Data preparation*. Analytical tools used in predictive modeling are powerful but they require the data to be fed in a certain format. Therefore, data clearing and its final arrangement for the further steps take place in this phase (Provost and Fawcett, 2013). *Modeling* is the fourth phase of the CRISP-DM cycle. In the modeling stage, a predictive model is built. The type of the model is determined in accordance with the dataset, number of observations, and format of dependent variables. Some of the modeling techniques used in data mining projects are regression, classification, neural nets, and ensembles. The dataset is then divided into train and test (holdout) sets. Train data allows the model to recognize the data and fit the model, while the testing dataset is used to evaluate the accuracy and generalizability of the built model (Smartvision, n.d.). The fifth stage of the CRISP-DM cycle is referred to as *Evaluation*. Herein, the model and its impact on business success criteria are evaluated. Before moving on with the selected model, the accuracy and generalizability of the model are assessed. In the *Deployment* stage, the results of the model evaluation from the previous step are taken into consideration and the deployment strategy is worked out (Smartvision, n.d.). In this research, the deployment of the model will be touched abstractly describing its potential effect on the business constituent of the UFC organization.

**3.2 Conceptual framework, Modeling & Evaluation**

The available literature reveals that though sports analytics is a well-demanded field of study for researchers, the number of studies in MMA is in short supply. However, it does not mean that the demand for them is also low. As mentioned above, the popularity of UFC all over the world is increasing considerably, thus generating a substantial relevance of this study.

Python programming language will be used throughout the whole process ranging from scraping certain parts of the data from UFCstats.com to a final stage of performance results of the built models. Python is a universal and flexible programming language allowing to access a wide variety of useful libraries, such as pandas, NumPy, scikit-learn, matplotlib, etc. The aforementioned libraries are helpful in processing and transforming the data, creating the highest-quality visualization options, as well as building a large base of ML models (Luashchuk, 2019). Moreover, Python contains built-in UFC scrapers which will be used to extract selected fight details.

Having referred to the research questions defined above, the current study will engage in 2 main analytical standpoints. Even though both research questions will incorporate ML modeling aspect in them, each of them requires a predictive approach since a difference between the type of both independent and dependent variables exists.

The constructed ML models will train and test on the definite dataset containing the historical data on UFC fights as well as the statistics on the bouts and fighter characteristics. The target variable of the dataset will predict either the win or loss of a certain fighter in a specific bout for the first research question. As it was already discussed, the constructed models can be applied not only to the data of past fights but also to the fights that have not even taken place.

While the second research question will use the same dataset, the nominal outcome variable will be predicted herein. The respective inputs are Decision, KO/TKO and submission. Since UFCStats.com does not provide any differentiation between KO/TKO resolutions, this outcome will be analyzed as a single unit. Likewise, even though the aforementioned source differentiates the decisive outcomes among 3 various methods (unanimous, majority and split), the current research will unite them under a single output of decision.

The models to be used in this study are versatile and use different approaches ranging from regression to ensemble modeling. Since the target of the study is a binary discrete variable for the first research question, the logistic regression is a must-have algorithm that needs to be tested (Brownlee, 2016). Logistic regression, also named as sigmoid function, can take any value between 0 and 1 which is a go-to method for the binary classification problems as in the case of UFC bout result prediction. It has to be mentioned that UFC fights can also be ended in a draw but considering the extremely minuscule percentage of fights reaching a tie, this result outcome will be neglected with regard to the scope of this research. However, due to the inherent functionality of logistic regression to determine only binary outcomes, this model will not be applied within the scope of the second research question.

One of the most intuitive models to be used in this study is the Decision trees. The respective model creates a reversed tree algorithm of yes/no conditions leading to an eventually derived prediction. The algorithm assigns the classifiers from top to the bottom based on the homogeneity or in other words purity of the variables. The homogeneity of the features is calculated using the Entropy and Information Gain (also derived based on entropy) putting the purer and, thus more important variables upper in the tree, while highly disordered variables are placed farther from the tree root (Galarnyk, 2019). Notwithstanding the fact that Decision trees are specialized in binary variables, Python built-in Sklearn documentation allows working with categorical predictions which will be applied to forecast match resolutions.

Random Forest, which is also considered as a supervised decision tree algorithm, refers to the ensemble of decision trees applied to the same sample of data. Ensemble modeling, being highly robust, less prone to overfitting and bias, is probably one of the most relevant algorithms for both binary and nominal targets presented in the research questions. The idea of the model is that the combination of learning algorithms elevates the final performance of predictions. Random forest inherits additional randomness to the algorithm, looking for the most important variables amongst a random set of features, eventually finding the average performance from a wide multiplicity of decision trees. This technique makes the model more diverse and, potentially, more accurate. It has to be mentioned that the popularity of this ML model is for a particular reason of its applicability in both regression and classification problems (Donges, 2020).

Another ML algorithm, applicable to the current study of binary and nominal classification problems, is the eXtreme Gradient Boosting (XGBoost) model. XGBoost is another ensemble model that uses a number of decision trees as a basis for the entire model construction. The main difference of the XGBoost classifier from the previously mentioned Random forest is in how trees are built and how the overall results are derived. As regards the tree construction process, while Random forests use the “bagging” method, creating different bags and train on randomly assigned data points, while the XGBoost algorithm uses “boosting”, which chooses the weak learners and converts them into strong ones building one tree at a time. The difference in results derivative refers to the way of calculating the performance of the model, in which Random forest computes the result at the end of the model learning, whereas the XGBoost algorithm calculates in the course of the learning process. The model is fast and reduces both bias and variance errors (Glen, 2019).

Support Vector Machine (SVM) algorithm is another popular algorithm that will be used in this study for the purpose of UFC fight outcome and match resolution ways predictions based on the existing feature variables. The primary objective of the SVM model is to create a baseline that will split the high-dimensional space into positive and negative hyperplanes, thus deriving the resulting outcome (Gandhi, 2018).

Apart from the learning and statistical algorithms, the research will dive into the assessment of the built models. First and foremost, the accuracy of the predictive models will be calculated. Performance accuracy is a simple measurement of the frequency of correct predictions out of the entire data points. The respective performance rate is often referred to as error rate and computed in accordance with the confusion matrix consisting of two dimensions, actual and predicted values. Having juxtaposed two dimensions, 4 possible outcomes can be derived: True positives (TP), True negatives (TN), False positives (FP), and False negatives (FN). The accuracy of the model depends on the number of the first 2 derivatives of the matrix (DeepAI, n.d.). The accuracy is a universal performance calculator which can be used in the evaluation of both binary and nominal dependent variables.

The Area under the Curve (AUC) is another machine learning evaluation metric that is inherent to binary classification problems only and, therefore, will not be applied in the second research question. The AUC also generates its score based on the confusion matrix deriving the Sensitivity and Specificity of the built model. Sensitivity refers to the TP rate which refers to the proportion of actual positive class predicted correctly, while specificity is the proportion of the actual negative class predicted accurately (TN rate). By changing the threshold values of the predicted data points, a number of confusion matrices can be generated to compare various measurements. Instead, the AUC-Receiver Operator Characteristic (ROC) curve is plotted in order to visualize the results of changing thresholds with Sensitivity (TP rate) on the y-axis and 1- Specificity (FP rate) on the x-axis (Bhandari, 2020). Therefore, the AUC-ROC curve depicts the relative trade-offs that a classifier makes between TP and FP.

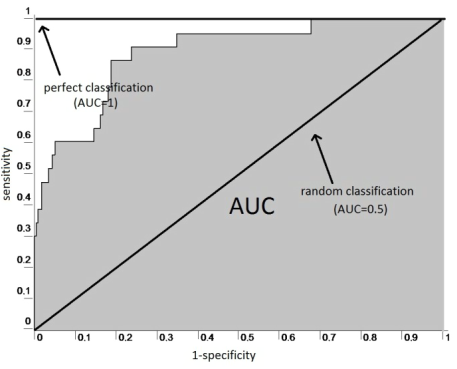


Figure 2. AUC-ROC curve (Belo, 2020)

Another evaluation score used for binary classification that will be used in this study is the Brier score. The aforesaid score is extremely useful in working with finite values and binary classifications. Brier score uses the probabilistic value of the prediction, the actual outcome and the number of forecasting instances. Due to being a cost function with limited boundaries between 0 and 1, the lower the Brier score, the better the prediction of the model. By its impact, the respective evaluation function is frequently equated to mean squared error.

It has to be highlighted that since the target variable is a binary variable having only two result outcomes, the baseline of the model could be set at 50% denoting a random guess by the predictor. However, since the existing research works has already set the stage for future analyses and the objective of this study is to improve currently available accuracy performance, the baseline will be increased up to 63.4% based on the research performed by Johnson (2012) as being one the most similar works in this sphere with the highest predictability. Any result higher than the latter baseline will improve the prediction confidence while the accuracy index equal to or less than the baseline will be incoherent with the purposes and aims of this study. As regards the prediction of the nominal target variable within the scope of the second research question, the baseline will be set at 33.3% denoting a random guess by the predictor since no previous analysis is conducted on this type of prediction.

**3.3 Available data sources & features**

Most of the existing UFC-related research papers and studies have used FightMetric as a detailed database for conducting their research. Recently, the analogy of FightMetric named UFCstats.com was released. The open-source platform traces back to 1994 starting from UFC 2 event and contains all the events incrementally adding each and every fight statistics to the database. The respective repository might be interesting not only to every UFC fan but is also relevant for the purpose of data analytical study.

Accordingly, the data point relevant for this study is obtained from Kaggle, an open-source community platform containing a large variety of datasets on different topics, and uses UFCstats.com as a primary source for it. The profound UFC dataset obtained from Kaggle contains every bout-related detail, as well as the average performance of each fighter at the moment of his/her fight, and is, therefore, applicable to a more in-depth analysis and data wrangling for increased performance accuracy of the models (Warrier, 2021).

As current research will also dive into the comparison of built models and bookmakers predictions, the dataset from another open-source platform, named GitHub, was obtained (Dabbert, 2020). The respective dataset contains betting odds for all fights from 2010 to 2020 and will be useful for the discussed match between the built models and bookies fight predictions.

Since the dataset has no information on the ways the fights were finished, which is a significant target variable for the purposes of the second research question, this information will be scraped from UFCstats.com using Python built-in packages.

Subsequently, the respective datasets will be processed, matched, and merged in one final dataset which will undergo both the EDA and modeling stages.

The variables to be used during the feature engineering procedures will be divided into 3 essential categories, fighter data, bout statistics and details, and the preprocessed differences between particular fighter characteristics. The individual statistics on each UFC fighter, incorporated into the Kaggle dataset, such as age, weight, height, and stance will be utilized in this study as one of the most important features. As some research projects described above have shown, such static characteristics of fighters as age (above or below 32 years of age) can have a significant effect on the predictive modeling and even have a positive ROI for bettors. It also has to be noted that the data is organized incrementally, such that it contains all the details of the fighter information prior to the future fight adding up all previous statistics.

The aforementioned Kaggle dataset consists of fight-related fighter details ranging from average striking accuracy to average landed takedowns, as well as current win or lose streaks, averages of body, and leg strikes, successful moves in the clinch, and grappling, etc. These features might also be crucially important in deriving the difference between the fighters and exploring their strong or weak sides.

The last but potentially not least component of the data will be the differences that will be obtained during data preprocessing. For instance, a difference of age will be calculated to simplify the complexity of the model and potentially increase its predictability. Selected features will undergo a historically incremental averaging difference method to ensure the better performance of the built algorithm.

**3.4 Timeframe**

The below-depicted graph illustrates the general plan for the research to be conducted. However, it has to be noted that the respective planning timetable is subject to change.

*Table 1. Research Timeframe*

|  |  |
| --- | --- |
| **January-21** | Initial analysis, Literature review |
| **February-21** | Literature review, Research questions, Primary modeling |
| **March-21** | Data collection and preparation, Feature engineering |
| **April-21** | Variable analysis, ML modeling, MMA betting analysis |
| **May-21** | MMA betting analysis, Results, Discussion |
| **June-21** | Discussion, Conclusion, Hand-in thesis (10 June) |

|  |  |
| --- | --- |
| **Abbreviation** | **Explanation** |
| AI | Artificial Intelligence |
| ATP | Association of Tennis Professionals |
| AUC | Area under the Curve |
| CRISP-DM | Cross-industry process for data mining |
| EDA | Exploratory Data Analysis |
| EPL | English Premier League |
| FN | False negatives |
| FP | False positives |
| XGBoost | eXtreme Gradient Boosting |
| ML | Machine learning |
| MMA | Mixed Martial Arts |
| NBA | National Basketball Association |
| NCAA | National Collegiate Athletic Association |
| NFL | National Football League |
| NHL | National Hockey League |
| ROC | Receiver Operator Characteristic |
| SVM | Support vector Machine |
| TN | True negatives |
| TP | True positives |
| UFC | Ultimate Fighting Championship |

**References:**

Alpaydin, E., 2010. *Introduction to Machine learning, 2nd edition*. England: The MIT Press. [Accessed 4 February 2021].

Athithya, V., 2019. *Scope of Data Science/Analytics in Sports World*. [online] towardsdatascience.com. Available at:

<https://towardsdatascience.com/scope-of-analytics-in-sports-world-37ed09c39860> [Accessed 4 February 2021].

Bartels, R., 2019. *Beating the bookies with Machine Learning*. [online] kdnuggets.com. Available at:

< https://www.kdnuggets.com/2019/03/beating-bookies-machine-learning.html> [Accessed 7 March 2021].

Belo, R., 2020. *ROC Curve and Area Under the Curve*. [PowerPoint Presentation] Big Data Management and Analytics course. [Accessed 14 February 2021].

Bhandari, A., 2020. *AUC-ROC Curve in Machine Learning Clearly Explained*. [online] analyticsvidhya.com. Available at:

< https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/#:~:text=The%20Area%20Under%20the%20Curve,the%20positive%20and%20negative%20classes.> [Accessed 14 January 2021].

Bigsby, K. G. and Ohlmann, J., 2017. *Ranking and prediction of collegiate wrestling*. [online] iospress.com. Available at:

< https://content.iospress.com/articles/journal-of-sports-analytics/jsa0024> [Accessed 14 March 2021].

Boulier, L. B. and Stekler, H. O., 1999. *Are sports seedings good predictors?* The United States of America: International Journal of Forecasting. [Accessed 8 February 2021]

Britannica, 2016. *Mixed martial arts*. [online] Britannica. Available at: <https://www.britannica.com/sports/mixed-martial-arts> [Accessed 14 January 2021].

Brownlee, J., 2016. *Logistic regression for Machine Learning* [online] machinelearningmastery.com. Available at:

<https://machinelearningmastery.com/logistic-regression-for-machine-learning/#:~:text=Logistic%20regression%20is%20another%20technique,problems%20with%20two%20class%20values).&text=Techniques%20used%20to%20learn%20the,logistic%20regression%20model%20from%20data.> [Accessed 13 February 2021].

Chakure, A., 2019. *Decision Tree Classification.* [online] medium.com. Available at:

<https://medium.com/swlh/decision-tree-classification-de64fc4d5aac> [Accessed 13 February 2021].

Collier, T. and Johnson, L. A., 2012. *Aggression in Mixed Martial Arts: An Analysis of the Likelihood of Winning a Decision.* [Accessed 12 February 2021]

Combat Museum, n.d. *How Does UFC Scoring Work?* [online] combatmuseum.com. Available at:

< https://combatmuseum.com/how-does-ufc-scoring-work/> [Accessed 10 February 2021].

Dabbert, M., 2020.[online] GitHub.com. Available at:

< https://github.com/shortlikeafox/tiger-millionaire> [Accessed 22 March 2021].

DeepAI, n.d. *Accuracy in Machine Learning.* [online] deepai.com. Available at:

< https://deepai.org/machine-learning-glossary-and-terms/accuracy-error-rate#:~:text=Accuracy%20in%20Machine%20Learning&text=Accuracy%20is%20the%20number%20of,of%20all%20the%20data%20points.&text=Often%2C%20accuracy%20is%20used%20along,true%2Ffalse%20positives%2Fnegatives.> [Accessed 14 February 2021].

Donges, N., 2020. *A complete guide to the random forest algorithm.* [online] builtin.com. Available at:

< https://builtin.com/data-science/random-forest-algorithm> [Accessed 10 February 2021].

Edwards, J., 2019. *What is predictive analytics? Transforming data into future insights.* [online] CIO. Available at:

< https://www.cio.com/article/3273114/what-is-predictive-analytics-transforming-data-into-future-insights.html> [Accessed 17 January 2021].

FightBookMMA, 2020. *10 Most Popular UFC Fighters on Instagram(2020)*. [online] FightBookMMA. Available at:

< https://www.fightbookmma.com/10-most-popular-ufc-fighters-on-instagram-2020/#:~:text=The%2010%20most%20popular%20UFC,Paige%20VanZant%2C%20and%20Israel%20Adesanya.> [Accessed 21 January 2021].

Foote, 2018. *A Brief History of Analytics*. [online] dataversity.net. Available at:

< https://www.dataversity.net/brief-history-analytics/#:~:text=Predictive%20Analytics%20first%20started%20in,concept%20whose%20time%20has%20come.> [Accessed 14 January 2021].

Forbes, 2018. *The New Fight Game: How an MMA Startup wants to capture the Sports’ 450 million global fans*. [online] Forbes. Available at:

< https://www.forbes.com/sites/kurtbadenhausen/2018/12/28/the-new-fight-game-how-an-mma-startup-wants-to-capture-the-sports-450-million-global-fans/?sh=25410d74139e> [Accessed 14 January 2021].

Gainsbury, S., 2015. *Online Gambling Addiction: the Relationship Between Internet Gambling and Disordered Gambling*. [online] Springer.com. Available at:

<https://link.springer.com/article/10.1007/s40429-015-0057-8> [Accessed 21 January 2021].

Galarnyk, M., 2019. *Understanding Decision Trees for Classification (Python).* [online] towardsdatascience.com. Available at:

< https://towardsdatascience.com/understanding-decision-trees-for-classification-python-9663d683c952> [Accessed 13 February 2021].

Gandhi, R., 2018. *Support Vector Machine – Introduction to Machine Learning Algorithms.* [online] towardsdatascience.com. Available at:

< https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47> [Accessed 13 February 2021].

St-Pierre, G., n.d. *Georges St-Pierre quotes.* [online] azquotes.com. Available at:

< https://www.azquotes.com/quote/724207> [Accessed 12 March 2021].

Glen, S., 2019. *Decision Tree vs Random Forest vs Gradient Boosting Machines: Explained Simply.* [online] datasciencecentral.com. Available at:

< https://www.datasciencecentral.com/profiles/blogs/decision-tree-vs-random-forest-vs-boosted-trees-explained#:~:text=Like%20random%20forests%2C%20gradient%20boosting,one%20tree%20at%20a%20time.> [Accessed 14 February 2021].

Goodfellow, I., Bengio, Y. and Courville., A, 2016. *Deep learning: adaptive computation and machine learning*. deeplearningbook.org [Accessed 14 January 2021].

Hessels, J., 2018. *Improving the prediction of soccer match results by means of Machine Learning.* [online]. Available at:

< https://arno.uvt.nl/show.cgi?fid=147179 > [Accessed 16 January 2021].

Holman, V., 2018. *What is Sports Analytics?* [online] AgileSportsAnalytics. Available at:

< https://www.agilesportsanalytics.com/what-is-sports-analytics> [Accessed 16 January 2021].

Hitkul, Aggarwal, K., Yadav, N. and Dwivedy. M, A, 2018. *A Comparative Study of Machine Learning Algorithms for Prior Prediction of UFC Fights*. [Harmony Search and Nature Inspired Optimization Algorithms](https://link.springer.com/book/10.1007/978-981-13-0761-4) (pp 67-76). [Accessed 8 March 2021].

IBISWorld, 2020a. *Martial Arts Studios Industry in the US.* [online] IBISWorld. Available at:

<https://www.ibisworld.com/united-states/market-research-reports/martial-arts-studios-industry> [Accessed 15 January 2021].

IBISWorld, 2020b. *IT Consulting Industry in the US.* [online] IBISWorld. Available at:

< https://www.ibisworld.com/united-states/market-research-reports/it-consulting-industry/> [Accessed 15 January 2021].

IBISWorld, 2020c. *Real Estate Sales&Brokerage.* [online] IBISWorld. Available at:

< https://www.ibisworld.com/united-states/market-research-reports/real-estate-sales-brokerage-industry> [Accessed 15 January 2021].

IBM, 2020. *Machine Learning.* [online] IBM. Available at:

<https://www.ibm.com/cloud/learn/machine-learning> [Accessed 22 January 2021].

IMGArena, n.d. *UFC fan insight*. [online] IMGArena. Available at:

<https://www.imgarena.com/ufc-fan-insight> [Accessed 21 January 2021].

Jee, K., 2020. *The 4 types of Sports Analytics Projects*. [online] playingnumbers.com. Available at:

<https://www.playingnumbers.com/2020/02/the-4-types-of-sports-analytics-projects/> [Accessed 27 January 2021].

Johnson, D. J., 2012. *Predicting Outcomes of Mixed Martial Arts Fights With Novel Fight Variables*. [online] Available at:

< https://getd.libs.uga.edu/pdfs/johnson\_jeremiah\_d\_201208\_ms.pdf> [Accessed 14 January 2021].

Kampakis, S. and Adamides, A., 2014. *Using Twitter to predict football outcomes*. [online] Arxiv.org. Available at:

<https://arxiv.org/ftp/arxiv/papers/1411/1411.1243.pdf> [Accessed 31 January 2021].

Kraus, W. K. and Chen, D. T., 2013. *A winning smile? Smile Intensity, Physical Dominance, and Fighter Performance.* The United States of America: International Journal of Forecasting.

Kuhn R. and Crigger, K., 2013. *FightNomics.* The United States of America: Graybeard Publishing LLC. [Accessed 20 January 2021].

Kumar, H., 2019. *Top 5 Predictive analytics use cases in the retail industry.* [online] Acuvate. Available at:

< https://acuvate.com/blog/top-5-predictive-analytics-use-cases-retail-industry/#:~:text=Predictive%20analytics%20is%20a%20proactive,and%20gain%20considerable%20market%20share.> [Accessed 17 January 2021].

Lane, S. and Briffa, M., 2020. *Perceived and actual fighting ability: determinants of success by decision, knockout or submission in human combat sports.* [online] towardsdatascience.com. Available at:

<https://doi.org/10.1098/rsbl.2020.0443> [Accessed 14 March 2021].

Luashchuk, A., 2019. *Why I think Python is Perfect for Machine Learning and Artificial Intelligence.* [online] towardsdatascience.com. Available at:

<https://towardsdatascience.com/8-reasons-why-python-is-good-for-artificial-intelligence-and-machine-learning-4a23f6bed2e6> [Accessed 13 February 2021].

Mahon, J. and McGowan, R., 2015. *Demand for the Ultimate Fighting Chmapionship: An Econometric Analysis of PPV Buy Rates*. The United States of America: Academic Star Publishing Company. [Accessed 4 February 2021].

Maryville, n.d. *How to become a Sports Data Analyst*. [online] Maryville. Available at:

<https://online.maryville.edu/online-bachelors-degrees/sport-business-management/careers/sports-data-analyst/#:~:text=Sports%20data%20analysts%20spend%20their,viability%20of%20plays%20and%20tactics> [Accessed 19 January 2021].

McQuaide, M., (2019). *Applying Machine Learning Algorithms to Predict UFC Fight Outcomes*. [online] Stanford.edu. Available at:

< http://cs229.stanford.edu/proj2019aut/data/assignment\_308875\_raw/26426025.pdf> [Accessed 9 March 2021].

Miller, W. T., 2016. *Sports Analytics and Data Science: Winning the game with methods and models*. The United States of America: Pearson Education, Inc. [Accessed 11 January 2021].

Orland, R., 2020. *10 Reasons why UFC is becoming so popular.* [online] Available at:

<https://www.opptrends.com/10-reasons-why-is-ufc-becoming-so-popular/> [Accessed 16 January 2021].

Philips, 2020. *Predictive analytics in healthcare: three real world examples.* [online] Philips. Available at:

< https://www.philips.com/a-w/about/news/archive/features/20200604-predictive-analytics-in-healthcare-three-real-world-examples.html#:~:text=Predictive%20analytics%20in%20healthcare%20can,avoidable%20downtime%20of%20medical%20equipment. > [Accessed 17 January 2021].

Provost, F. and Fawcett, T., 2013. *Data Science for Business*. The United States of America: O’Reilly Media, Inc. [Accessed 14 January 2021].

Robbins, R., T. and Zemanek, E. J., 2017. *UFC pay-per-view buys and the value of the celebrity fighter.* The United States of America: LLC Consulting Publishing Company “Business Perspectives” [Accessed 3 January 2021].

RealMadrid, 2020. *Real Madrid named most valuable European football club by KPMG consultancy firm.* [online] RealMadrid. Available at:

<https://www.realmadrid.com/en/news/2020/05/28/real-madrid-named-most-valuable-european-football-club-by-kpmg-consultancy-firm> [Accessed 16 January 2021].

SBD, 2018. *Sports betting algorithms*. [online] SBD. Available at:

<https://www.sportsbettingdime.com/guides/finance/sports-betting-algorithms/> [Accessed 19 January 2021].

Schmidt, J., Marques, R. G. M., Botti, S., Marques, A. L. M., 2019. *Recent advances and applications of machine learning in solid-state materials science*. [online] Nature. Available at:

<https://www.nature.com/articles/s41524-019-0221-0> [Accessed 21 January 2021].

Singh, V., 2011. *Betting on UFC Fights – A Statistical Data Analysis*. [online] partyondata.com. Available at:

< https://partyondata.com/2011/09/21/betting-on-ufc-fights-a-statistical-data-analysis/> [Accessed 11 March 2021].

Sipko, M., 2015. *Machine Learning for the Prediction of Professional Tennis Matches*. [online] Available at:

< https://www.doc.ic.ac.uk/teaching/distinguished-projects/2015/m.sipko.pdf> [Accessed 21 January 2021].

Smartvision, n.d. *What is the CRISP-DM methodology*? [online] Smartvision. Available at:

<https://www.sv-europe.com/crisp-dm-methodology/> [Accessed 19 January 2021].

SMASchools, 2019. *Ultimate Fighting Champions League (UFC).* [online] SMASchools. Available at:

<https://smaschools.com/why-is-mma-so-popular/#:~:text=The%20fighters%20are%20separated%20into,to%20become%20a%20mainstream%20sport> [Accessed 15 January 2021].

Soper, T., 2014. *Microsoft Bing beats Google in World Cup predictioons*. [online] geekwire.com. Available at:

< https://www.geekwire.com/2014/microsoft-bing-15-16-world-cup/> [Accessed 14 January 2021].

Sportskeeda, 2020. *Mixed Martial Arts, Pro Wrestling post incredible Cross-Platform online viewership statistics in 2019*. [online] Sportskeeda. Available at:

<https://www.sportskeeda.com/mma/news-mixed-martial-arts-pro-wrestling-post-incredible-cross-platform-online-viewership-statistics-2019> [Accessed 14 January 2021].

Statista, 2021. *Key industry data on the sports betting sector worldwide in 2020*. [online] Statista. Available at:

<https://www.statista.com/statistics/1154681/key-data-global-sports-betting-industry/#:~:text=The%20global%20sports%20betting%20industry%20reached%20a%20market%20size%20of,the%20United%20States%20in%202019.> [Accessed 22 January 2021].

Tyagi, N., 2020. *Defining Predictive Modeling in Machine learning.* [online] Medium. Available at:

<https://medium.com/analytics-steps/defining-predictive-modeling-in-machine-learning-887c23b7a278> [Accessed 22 January 2021].

UFC, 2018. *Intro to MMA*. [online] UFC. Available at:

<https://www.ufc.com/intro-mma#:~:text=Mixed%20martial%20arts%20(MMA)%20is,standing%20and%20on%20the%20ground> [Accessed 20 January 2021].

UFC Performannce Institute, 2018. *A Cross-sectional Performance Analysis and Projection of the UFC Athlete*. [online] UFC. Available at:

< https://www.academia.edu/39757604/UFCPI\_Book> [Accessed 24 January 2021].

Ulmer, B. and Fernandez, M., 2013. *Predicting Soccer Match Results in the English Premier League*? [online] Stanford.edu. Available at:

<http://cs229.stanford.edu/proj2014/Ben%20Ulmer,%20Matt%20Fernandez,%20Predicting%20Soccer%20Results%20in%20the%20English%20Premier%20League.pdf> [Accessed 25 January 2021].

Van Rijmenam, M., 2013. *The history of predictive analytics-Infographic*. [online] Datafloq.com. Available at:

<https://datafloq.com/read/history-predictive-analytics-infographic/438> [Accessed 25 January 2021].

Warnick, J. and Warnick, K., 2007. *Specification of variables predictive of victories in the sport of boxing*. [online] sagepub.com. Available at:

< https://journals.sagepub.com/doi/pdf/10.2466/pms.105.1.153-158?casa\_token=uISm75VyL80AAAAA:y2RzHJ\_6GwhpCQsSCLsGsWyB\_absCgNH5cyBfHflDX-H1m9TraNEldMNuVoH5dy9QLjJb2ZxZH\_KhQ> [Accessed 13 March 2021].

Warrier, R., 2021. *UFC-Fight historical data from 1993 to 2021*. [online] Kaggle.com. Available at:

< https://www.kaggle.com/rajeevw/ufcdata> [Accessed 22 March 2021].