CMPT 353 Final Project - UFC Fight Data Analysis

Introduction

Even though MMA is known for being unpredictable, we still wanted to see if we could find any patterns or trends that consistently make a difference. For this project, our goal is setting out to explore whether physical traits and fighting styles, things like height, reach, stance, age, and experience affect the outcome of a fight.

We started by refining the original idea into a few key questions:

- Does experience beat youth when it comes to winning fights?
- Does a longer reach give fighters a real edge?
- Are some stances more effective than others?
- Does win patterns differ by weight class?
- How different finish types (KO, submission, decision) relate to fighter stats?
- Is there any difference between different genders?
- How the finished round would influence the result?

Our focus was on core stats like reach, height, age, fight experience, and stance. We also paid special attention to the heavyweight division, since it has the widest weight range (206–265 lbs) and opens the door for big weight mismatches. Then we analyzed more about how gender, finish round and fights per year as extra stats related to the trends.

Data Collection and Cleanup

We used a public dataset of UFC fight results from 2010 to 2024. Cleaning the data was a big part of the project. We dropped unnecessary columns like betting odds and rankings, standardized stance labels, merged decision outcomes into one category, fixed fighter name mismatches using fuzzy matching, and removed data that clearly didn't make sense.

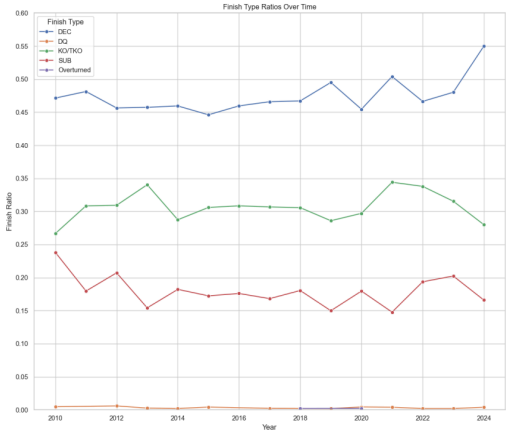
How We Analyzed the Data & Methods

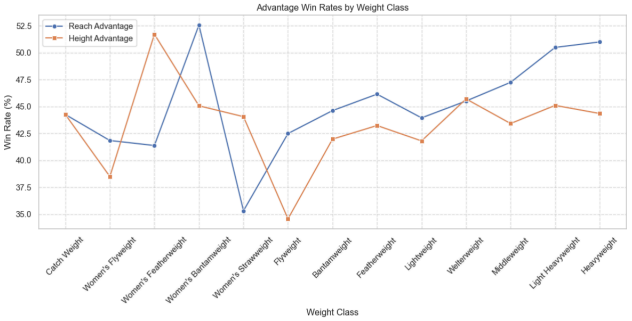
We applied a combination of descriptive statistics, visualizations, and statistical tests. Basic trends were explored with group averages and value counts, while visualizations using Seaborn and Matplotlib helped highlight how things varied over time or by weight class. To test specific hypotheses, we used Ordinary Least Squares (OLS) regression to model the effect of different physical advantages on fight outcomes, and one sample t-tests to check whether winners tended to have measurable advantages.

Moreover, we also expanded our analysis by creating more data analysis of the dataset such as gender differences, distribution of finishing rounds, and the overall trends in the number of fights per year. These analyses enabled us to incorporate richer insights into our predictive modeling approach. By integrating these additional factors, we trained and compared three machine learning models, the Logistic Regression, the Random Forest, and the K-Nearest Neighbors (KNN). We evaluated their performances through metrics such as accuracy, precision, recall, F1-scores, and AUC values, and then found which is a better model to use.

Findings & Results

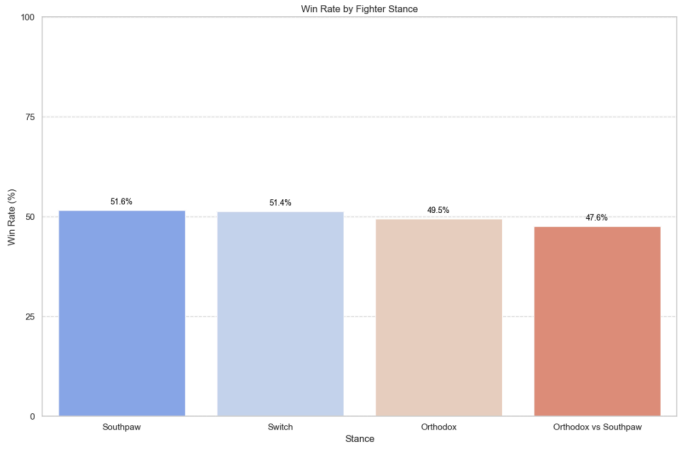
1. When we looked at how fights ended, we found that 49% resulted in a decision, 32% in a KO or TKO, and 18% by submission. Interestingly, decision wins have become more common, increasing by about 7% over the past decade.(left below figure)





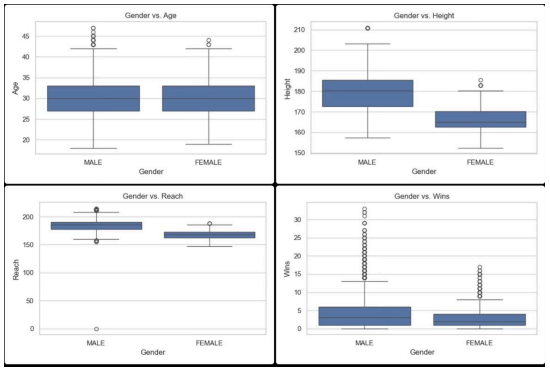
2. On the question of physical advantages, the results were surprising. Fighters with longer reach only won 45.6% of the time, and taller fighters had a win rate of just 43%. These numbers were even lower in the flyweight division. In heavyweight, it was closer to even, but still not dominant. This suggests that simply having longer limbs doesn't guarantee success, there may be trade offs in speed, strength, or technique. (right above figure)

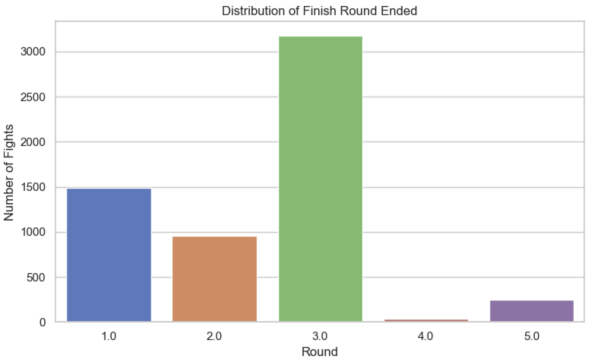
3. When looking at the stance (below figure), Southpaw fighters had the highest win rate at 51.6%. In matchups between Orthodox and Southpaw fighters, Orthodox fighters only won 47% of the time. This suggests there might be a strategic challenge in that pairing.



4. We explored whether gender affected their performance by analyzing four factors: age, height, reach, and overall win percentage(the left below figure). In terms of age, there was no significant difference. However, in terms of height, male fighters are generally about 10 cm taller than female fighters, and similarly, male fighters have

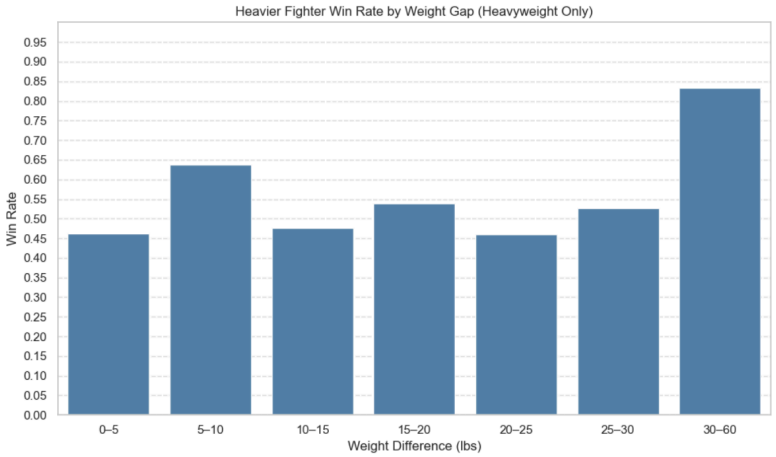
significantly longer hands (180cm to 190cm), while female fighters have mostly 160-170 cm hands. This can greatly improve their range of attack and their ability to maintain a safe distance from their opponents. In terms of total victories, male fighters get more victories, which may also be due to the relatively small data sample of female fighters' fights. In conclusion, these apparent physical advantages in height and tentacles suggest that male fighters may find it easier to dominate attacking communication and control the pace of a fight, which may increase their chances of winning.





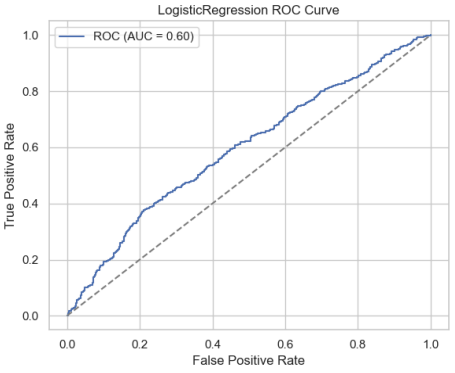
5. As we looked at the finish round plot(the right above figure), we noticed an interesting trend. Most of the matches (3175) ended in the third round. The first round had the second most completions (1492), followed by the second round (953). Only few will end in the later rounds (36 in the fourth round, 250 in the fifth). This pattern emphasizes the importance of the fighters' conditioning and their ability to maintain their performance at the crucial midpoint of the match. Fighters who effectively manage their stamina and adjust their strategy as the battle progresses may improve their chance of winning.

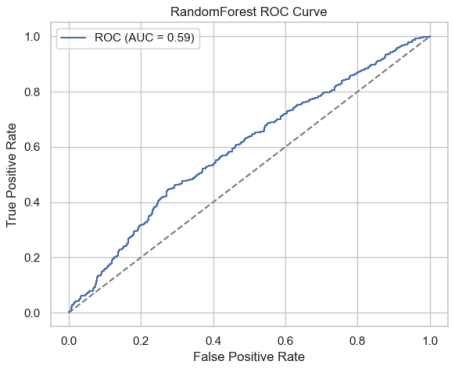
6. We backed these observations up with t-tests. On average, winners did tend to have slightly longer reach ($t = 4.998$, $p < 0.001$) and were a bit taller ($t = 3.00$, $p = 0.0027$). But the more striking results came from age and experience. Younger fighters were significantly more likely to win ($t = -12.85$, $p < 0.001$), and more experienced fighters actually had a slight disadvantage ($t = -2.13$, $p = 0.033$).
We were especially curious about heavyweight fighters, where the potential weight difference can be substantial. Fighters with a small weight edge (5–10 lbs) won about 63% of the time, which is a solid jump. But between 10–25 lbs, win rates bounced around, possibly due to trade offs in endurance or agility. The biggest spike came when the weight gap exceeded 30 lbs here, the heavier fighter won over 80% of the time. However, when we ran a t-test on all heavyweight bouts combined, we didn’t find a statistically significant edge for heavier fighters overall ($t = -1.093$, $p = 0.275$). So it seems that weight only becomes a deciding factor when the difference is large.

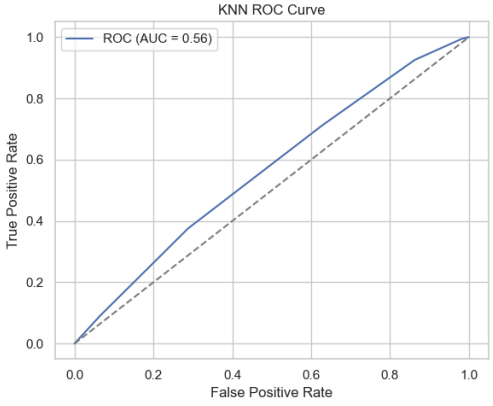


OLS regression results were consistent with this picture. Reach had a small but meaningful effect on win probability (coef = 0.0027, p = 0.002), while height and experience weren't significant. Age, however, had a clear negative relationship with winning (coef = -0.0161, p < 0.001).

7. To better understand what influences fight outcomes, we built three predictive models using key factors from our dataset, including differences in age, height, reach, experience, and fighters' stance matchups. We treated fight outcomes as a simple binary prediction (winner or loser), splitting our dataset into two groups: 80% to train the models and 20% to test how accurate they were.
Here's how each model performed:








```
=== Machine Learning Modeling ===

==== LogisticRegression Classification Report ====
              precision    recall  f1-score   support

           0       0.52      0.28      0.36       496
           1       0.61      0.81      0.70       686

    accuracy                           0.59      1182
   macro avg       0.56      0.55      0.53      1182
weighted avg       0.57      0.59      0.56      1182


==== RandomForest Classification Report ====
              precision    recall  f1-score   support

           0       0.53      0.15      0.23       496
           1       0.60      0.91      0.72       686

    accuracy                           0.59      1182
   macro avg       0.56      0.53      0.47      1182
weighted avg       0.57      0.59      0.51      1182


==== KNN Classification Report ====
              precision    recall  f1-score   support

           0       0.48      0.37      0.42       496
           1       0.61      0.71      0.66       686

    accuracy                           0.57      1182
   macro avg       0.55      0.54      0.54      1182
weighted avg       0.56      0.57      0.56      1182


=== Model AUC Comparison ===
LogisticRegression: AUC = 0.5957
RandomForest: AUC = 0.5936
KNN: AUC = 0.5645
```

1) The logistic regression model had an accuracy rate of 59%. For correctly predicting the winners, it had 61% precision (how often it was right) and a high recall of 81% (how many actual winners it found), with an F1-score (overall reliability) of 0.70. The model's AUC score was the best of the three at 0.5957, indicating good balanced performance overall. This model appears to offer stable and reliable predictions, capturing the factors influencing match outcomes effectively.

2) The Random Forest model also achieved an accuracy of 59%, matching the logistic regression model. It performed exceptionally well in finding the winning fighters (91% recall), though slightly lower in precision (60%), with a strong F1-score of 0.72. However, it struggled significantly in predicting the losing side, correctly identifying only 15% of losing fighters. Its AUC score was slightly lower at 0.5936. This means that while Random Forest is excellent at detecting likely winners, it tends to overlook fighters who may lose, potentially making it less balanced for general prediction purposes.

3) The KNN model showed slightly weaker results, with an accuracy of 57% and an AUC of only 0.5645. It performed moderately in identifying winners (precision 61%, recall 71%), but showed weaker performance for losers (precision 48%, recall 37%). Due to these shortcomings, this model isn't as effective for reliably predicting match outcomes compared to the other two.

Overall, among these three models, the **logistic regression model** stands out as the best choice. It combines a good balance between accuracy and precision, and clearly identifies winning fighters more consistently. Its relatively higher AUC score (0.5957) means it's likely to give more reliable predictions about fight outcomes than the other tested methods.

Limitations & Future

1. We didn't include context like short notice fights, injuries, or skill level, which could affect results.
2. For categories like 30+ lb weight gaps, the number of fights was small, which limits how confident we can be.
3. If we had more time, we could expand the analysis to more divisions, and look at performance by round or across fighter careers.

In conclusion, we not only use statistical regression and t test to explain the influence of variables on the results of the competition, but also construct a machine learning prediction model to further improve the prediction of the results of the competition. Then we found that certain physical traits, especially age and reach can have a measurable effect on fight outcomes. But those effects are subtle, and context clearly matters. While MMA remains unpredictable, this analysis helps highlight trends that could support smarter predictions, coaching, or even match planning.

Project Experience Summary

Isaac Jones:

1. Built a data pipeline to clean and unify UFC fight records from 2010–2024 using Pandas, NumPy, and fuzzywuzzy string matching to correct name inconsistencies, unify decision outcomes, and remove invalid records, improving dataset integrity for over 6,500 fights used in downstream statistical analysis.

2. Applied regression and hypothesis testing to evaluate the impact of physical attributes on fight outcomes by using OLS models and t-tests in Python (statsmodels, SciPy) to analyze variables such as height, reach, age, and experience, I found statistically significant relationships including a 0.0027 coefficient for reach and a −0.0161 coefficient for age, supporting nuanced conclusions.

3. Designed and implemented data visualizations to communicate key insights using Seaborn and Matplotlib to display finish type trends over time, stance win rates, and win percentages by weight gap bins, helping illustrate that fighters with a 30+ lb advantage in heavyweight bouts win over 80% of the time.

4. Delivered a comprehensive report summarizing data driven findings targeted at a technically literate audience, balancing statistical explanation with accessible insights, included quantitative evidence, regression outputs, and strategic recommendations to support future match analysis or predictive modeling.

Zekai Li:

1. Based on the UFC fight data (2010–2024), developing additional exploratory data analyses on gender differences, total win methods, the finish round distributions, and annual fights trends, using Pandas and visualization libraries.

2. Built and evaluated predictive models for fight outcomes using Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN) classifiers, incorporating features such as age differences, reach, height, experience, and fighting stance. Conducted detailed comparative assessments using accuracy, precision, recall, F1-scores, and AUC metrics, demonstrating that Logistic Regression provided the most balanced and reliable predictions.

3. Analyze comparative data results and generate clear, informative data graphs to effectively communicate complex analysis results in a way that users can understand. Analyze specific data to draw meaningful conclusions.

4. Written a detailed report, working mainly on some extra data analysis results to ensure that rigorous, clear explanation of the data and actionable insights were effectively communicated, which significantly improved the overall analysis depth and prediction accuracy of the project.