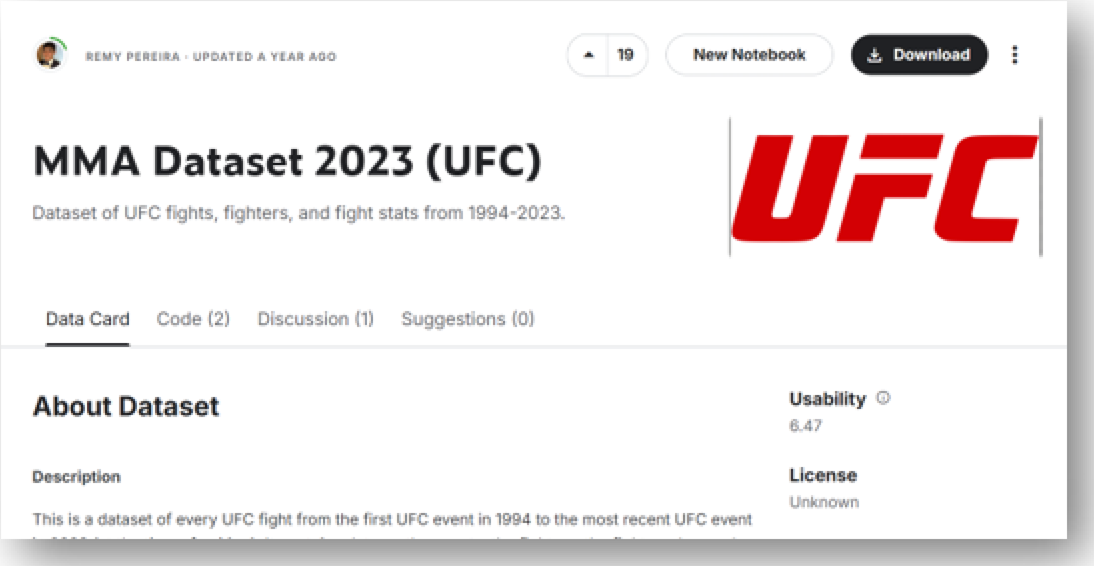
ML Enters the Octagon: Predicting UFC Winners

John Carl Tagbor, Luke White,

Molly McDade, Ann Gray Perdue

Data

The dataset used for our analysis is the MMA Dataset 2023 (UFC). It was sourced from Kaggle and details UFC fights, fighters, and fight stats from 1994-2023.



MMA Dataset 2023 (UFC)

Dataset of UFC fights, fighters, and fight stats from 1994-2023.

About Dataset

Usability

6.47

License

Unknown

Description

This is a dataset of every UFC fight from the first UFC event in 1994 to the most recent UFC event

Variables

Seniority (l_seniority, r_seniority): The difference in age between each fighter and their opponent. A positive value in l_seniority means that the left fighter is that much older than the right fighter. All the difference columns follow suit.

Height difference (l_height_dif_cm, r_height_dif_cm): The difference in height between each fighter and their opponent, measured in centimeters.

Weight difference (l_weight_dif_lb, r_weight_dif_lb): The difference in weight between each fighter and their opponent, measured in pounds.

Reach difference (l_reach_dif_cm, r_reach_dif_cm): The difference in reach (arm span) between each fighter and their opponent, measured in centimeters.

Control time (l_ctrl_time_sec, r_ctrl_time_sec): The amount of time each fighter maintained control during the fight, measured in seconds.

Total fights (l_total_fights, r_total_fights): The total number of professional fights each fighter has had.

Total fights difference (l_total_fights_dif, r_total_fights_dif): The difference in total number of fights between each fighter and their opponent.

Win difference (l_win_difference, r_win_difference): The difference in number of wins between each fighter and their opponent.

TKO Received (l_TKO, r_TKO: The cumulative knockouts a fighter has sustained prior to each event.

Win Streak (l_winst, r_winst: The number of consecutive victories a fighter has achieved up to each fight.

Model Evaluation

Model Selection

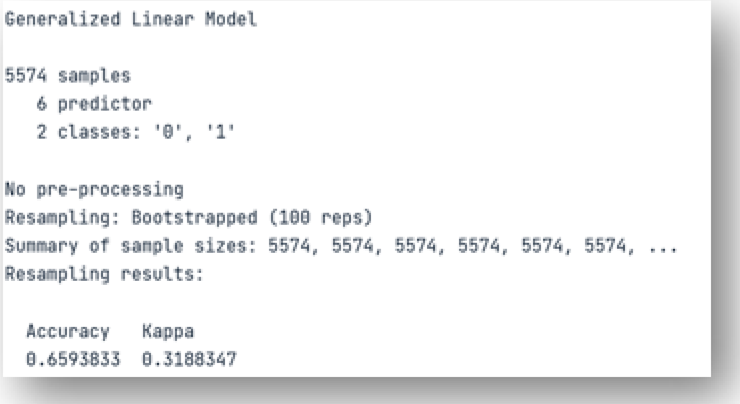
We originally developed two models for our predictions, a logistic regression model and an LDA model. After testing these models out, a couple times with different seeds, we found that all the accuracy metrics went up or down about 3-4%. We decided this probably had to do with the sampling, so we decided to create a bootstrapped model of each.

Bootstrapping

The accuracy of these bootstrapped models are almost identical and have very similar accuracies (within 2%) to the non-bootstrapped models. However, the bootstrapped LDA model had a slightly higher accuracy by 0.006..

Chosen Model

We chose to use the bootstrapped LDA model to predict the winners of December 7th's fights (and Jake Paul vs Mike Tyson from a couple weeks ago). We selected this model because it had the best overall accuracy.

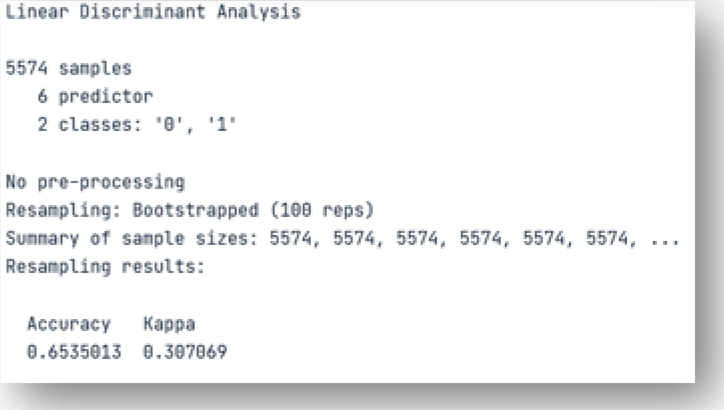


```
Generalized Linear Model

5574 samples
   6 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Bootstrapped (100 reps)
Summary of sample sizes: 5574, 5574, 5574, 5574, 5574, 5574, ...
Resampling results:

  Accuracy   Kappa
  0.6593833  0.3188347
```



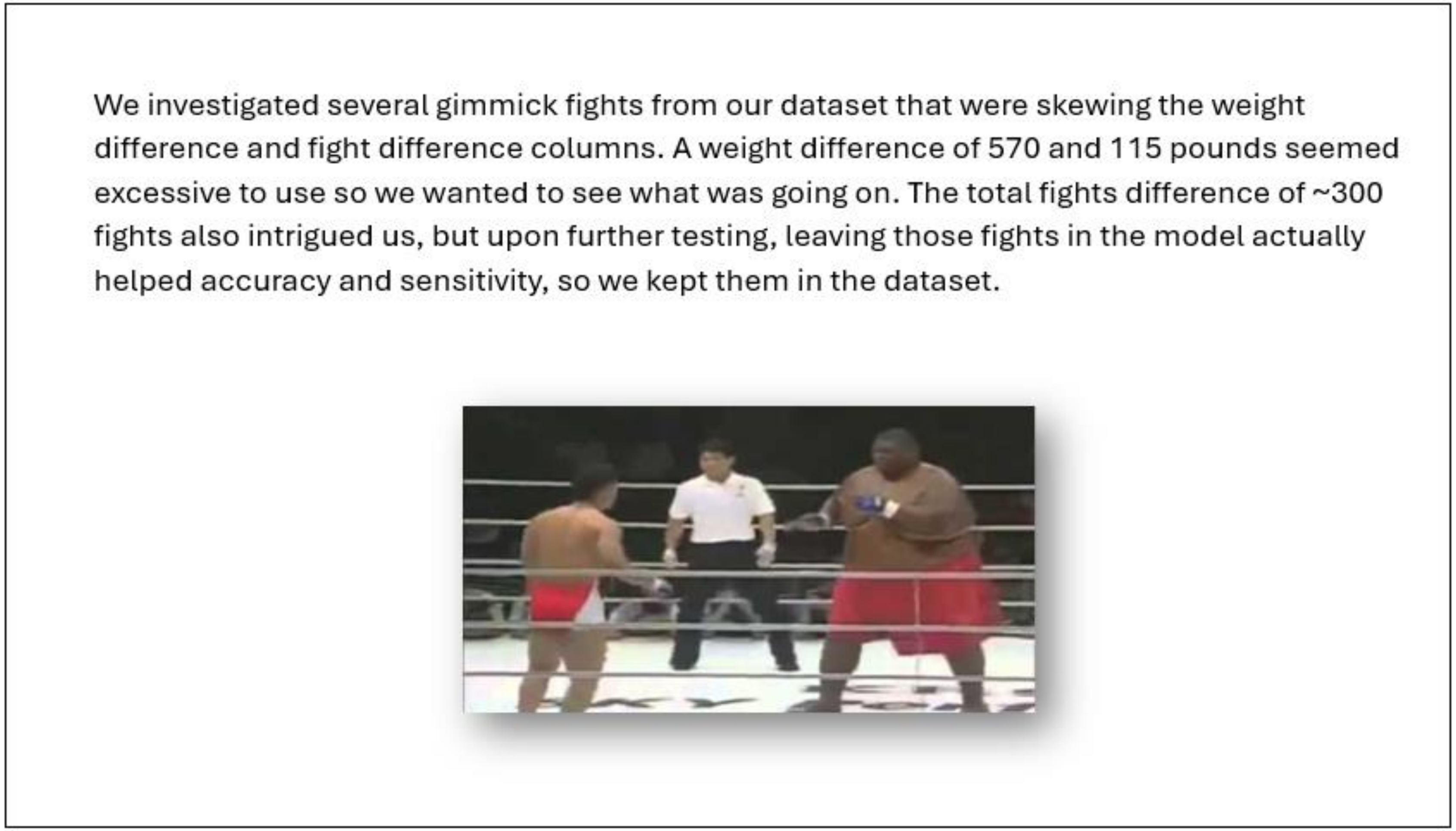
```
Linear Discriminant Analysis

5574 samples
   6 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Bootstrapped (100 reps)
Summary of sample sizes: 5574, 5574, 5574, 5574, 5574, 5574, ...
Resampling results:

  Accuracy   Kappa
  0.6535013  0.307069
```

Outliers



We investigated several gimmick fights from our dataset that were skewing the weight difference and fight difference columns. A weight difference of 570 and 115 pounds seemed excessive to use so we wanted to see what was going on. The total fights difference of ~300 fights also intrigued us, but upon further testing, leaving those fights in the model actually helped accuracy and sensitivity, so we kept them in the dataset.

Model

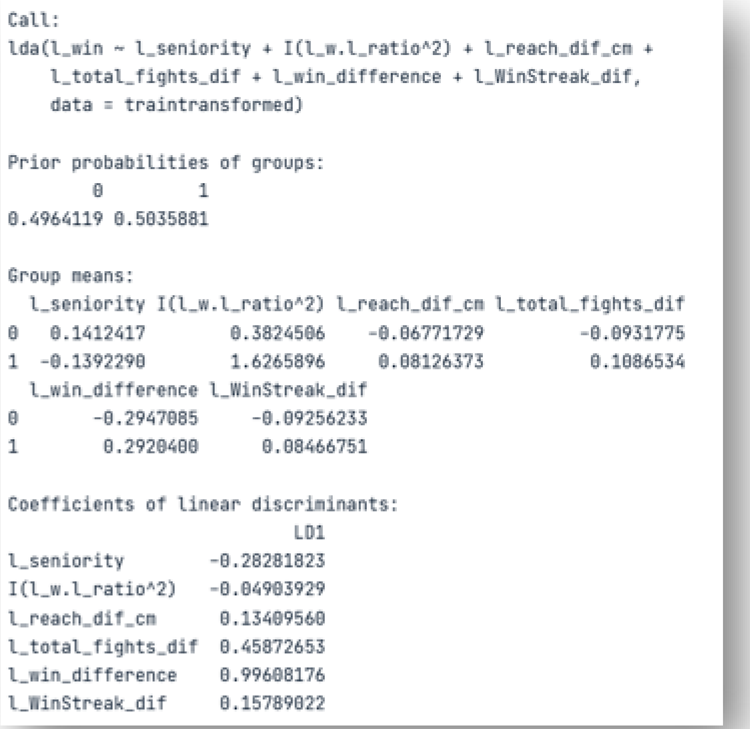
We used a LDA (Linear Discriminant Analysis) model to predict winners in UFC fights based on various fighter attributes. The model uses the following predictors to determine the likelihood of a fighter winning:

- **`l_seniority`:** Fighter's experience level
- **`I(l_w.l_ratio^2)`**: Squared win-loss ratio
- **`l_reach_dif_cm`**: Reach difference in centimeters
- **`l_total_fights_dif`**: Difference in total number of fights
- **`l_win_difference`:** Difference in number of wins
- **`l_WinStreak_dif`**: Difference in winning streak

The group means show the average values of each predictor for losing (0) and winning (1) fighters:

- **`l_seniority`**: Winning fighters tend to have less seniority (-0.1392 vs 0.1412)
- **`I(l_w.l_ratio^2)`**: Winners have a much higher squared win-loss ratio (1.6266 vs 0.3825)
- **`l_reach_dif_cm`**: Winners have a slight reach advantage (0.0813 vs -0.0677)
- **`l_total_fights_dif`:** Winners have fought more fights (0.1087 vs -0.0932)
- **`l_win_difference`**: Winners have more wins (0.2920 vs -0.2947)
- **`l_WinStreak_dif`**: Winners have a better winning streak (0.0847 vs -0.0926)

The difference in number of wins (`l_win_difference`) was the strongest predictor of victory, followed by the difference in total fights. Interestingly, seniority has a negative impact, suggesting that less experienced fighters might have an advantage



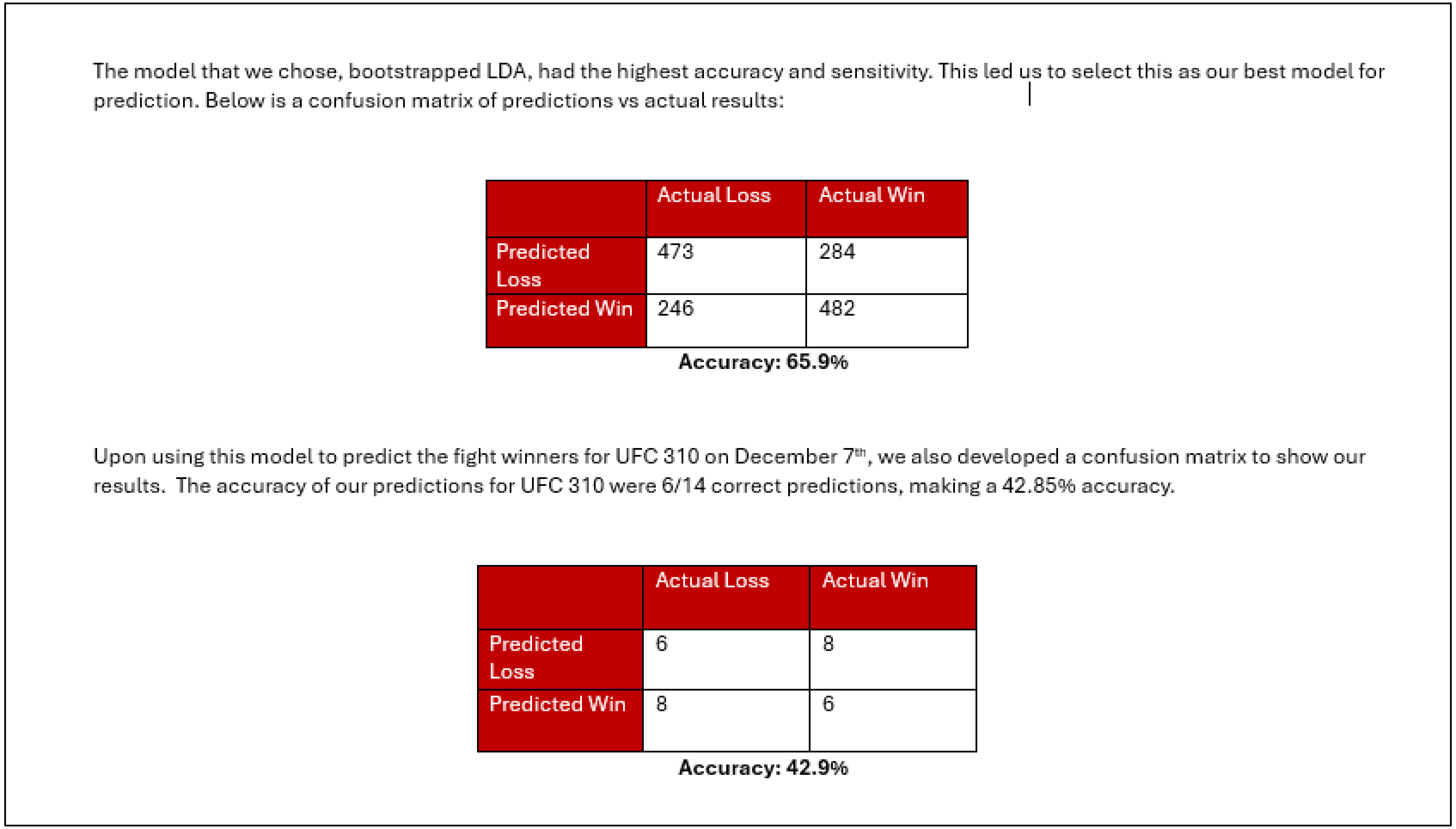
```
Call:
lda(l_win ~ l_seniority + I(l_w.l_ratio^2) + l_reach_dif_cm +
    l_total_fights_dif + l_win_difference + l_WinStreak_dif,
    data = traintransformed)

Prior probabilities of groups:
        0         1 
0.4964119 0.5035881 

Group means:
  l_seniority I(l_w.l_ratio^2) l_reach_dif_cm l_total_fights_dif
0   0.1412417        0.3824506    -0.06771729         -0.0931775
1  -0.1392290        1.6265896     0.08126373          0.1086534
  l_win_difference l_WinStreak_dif
0       -0.2947085     -0.09256233
1        0.2920400      0.08466751

Coefficients of linear discriminants:
                           LD1
l_seniority        -0.28281823
I(l_w.l_ratio^2)   -0.04903929
l_reach_dif_cm      0.13409560
l_total_fights_dif  0.45872653
l_win_difference    0.99608176
l_WinStreak_dif     0.15789022
```

Model Accuracy



The model that we chose, bootstrapped LDA, had the highest accuracy and sensitivity. This led us to select this as our best model for prediction. Below is a confusion matrix of predictions vs actual results:

	Actual Loss	Actual Win
Predicted Loss	473	284
Predicted Win	246	482

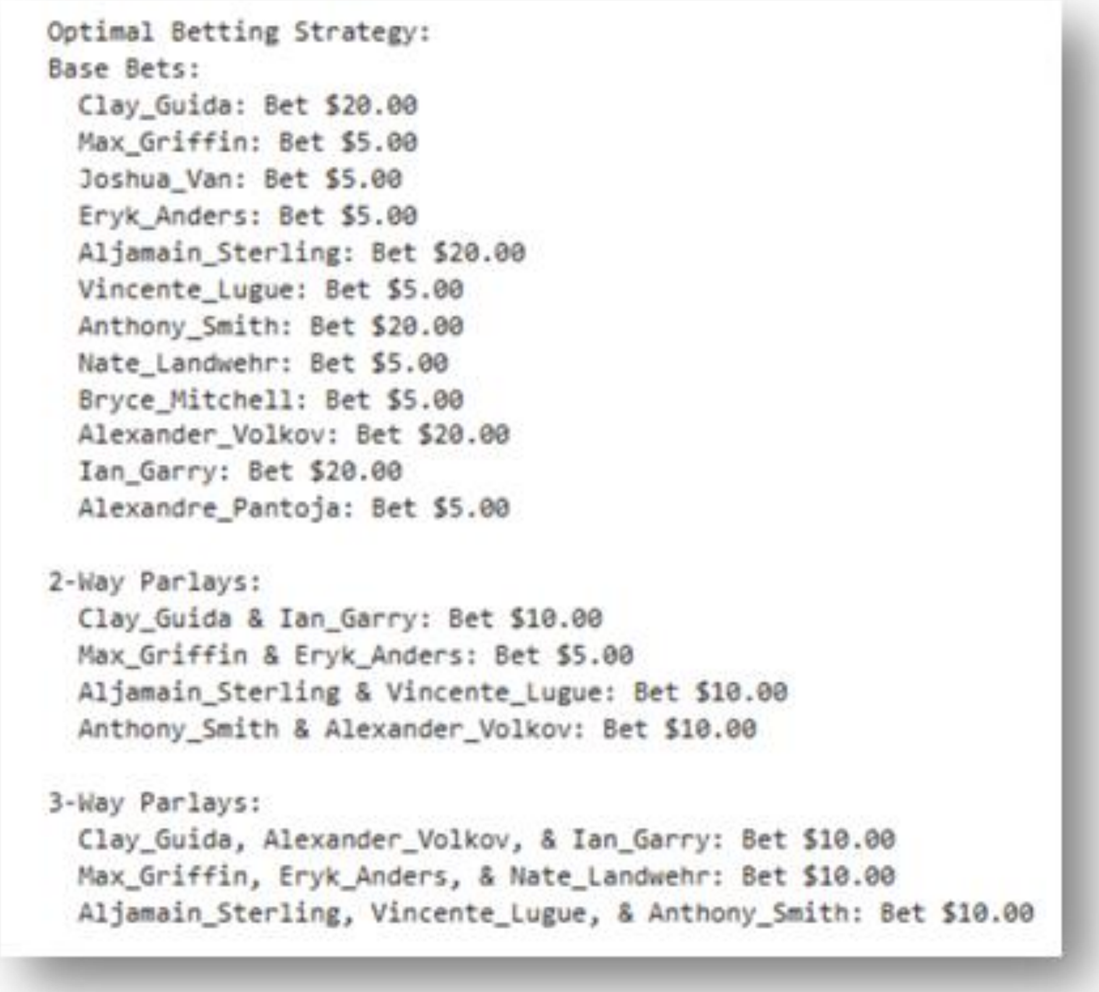
Accuracy: 65.9%

Upon using this model to predict the fight winners for UFC 310 on December 7th, we also developed a confusion matrix to show our results. The accuracy of our predictions for UFC 310 were 6/14 correct predictions, making a 42.85% accuracy.

	Actual Loss	Actual Win
Predicted Loss	6	8
Predicted Win	8	6

Accuracy: 42.9%

UFC Bets



```
Optimal Betting Strategy:
Base Bets:
  Clay_Guida: Bet $20.00
  Max_Griffin: Bet $5.00
  Joshua_Van: Bet $5.00
  Eryk_Anders: Bet $5.00
  Aljamain_Sterling: Bet $20.00
  Vincente_Lugue: Bet $5.00
  Anthony_Smith: Bet $20.00
  Nate_Landwehr: Bet $5.00
  Bryce_Mitchell: Bet $5.00
  Alexander_Volkov: Bet $20.00
  Ian_Garry: Bet $20.00
  Alexandre_Pantoja: Bet $5.00

2-Way Parlays:
  Clay_Guida & Ian_Garry: Bet $10.00
  Max_Griffin & Eryk_Anders: Bet $5.00
  Aljamain_Sterling & Vincente_Lugue: Bet $10.00
  Anthony_Smith & Alexander_Volkov: Bet $10.00

3-Way Parlays:
  Clay_Guida, Alexander_Volkov, & Ian_Garry: Bet $10.00
  Max_Griffin, Eryk_Anders, & Nate_Landwehr: Bet $10.00
  Aljamain_Sterling, Vincente_Lugue, & Anthony_Smith: Bet $10.00
```

Using python and Gurobi, we built an optimization model to maximize returns on bets given the sportsbooks odds for each fighter and our model's accuracy. Constraints were added to make sure every fighter predicted to win from our ML model was bet on, and that the model couldn't spend too much on parlays since they're inherently riskier.

Expected Returns

The total amount bet across all categories (Base Bets, 2-Way Parlays, and 3-Way Parlays) was $200.00. The total potential payout for all bets combined was $1033.35.

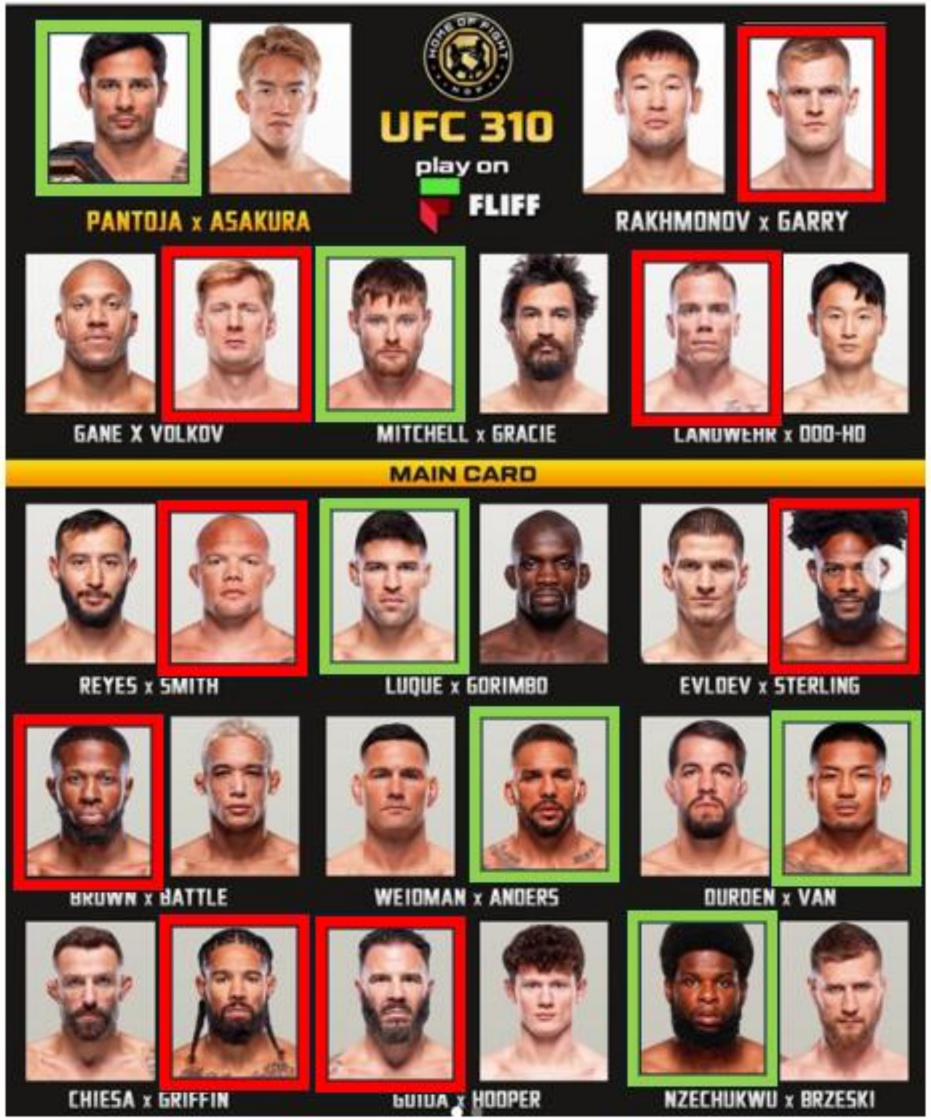
Actual Returns

The total amount won across all categories was actually a net loss of $132.00.

- Total Won: $42.70 (from individual winning bets)
- Total Lost: $175.00 (from losing bets and all parlays)
- Net Loss: -$132.30

310 Predictions

- Model predicted most betting favorites to win
- Predicted 3 Underdogs:
  - Ian Machado Garry
  - Alexander Volkov
  - Clay Guida



310 Results

- ….Our model predicted 6/14 outcomes correctly, so 42.85% accuracy.
- All of our underdogs lost
- All of our parlays lost
- Many of the favorites also lost