
UFC
ULTIMATE FIGHTING CHAMPIONSHIP

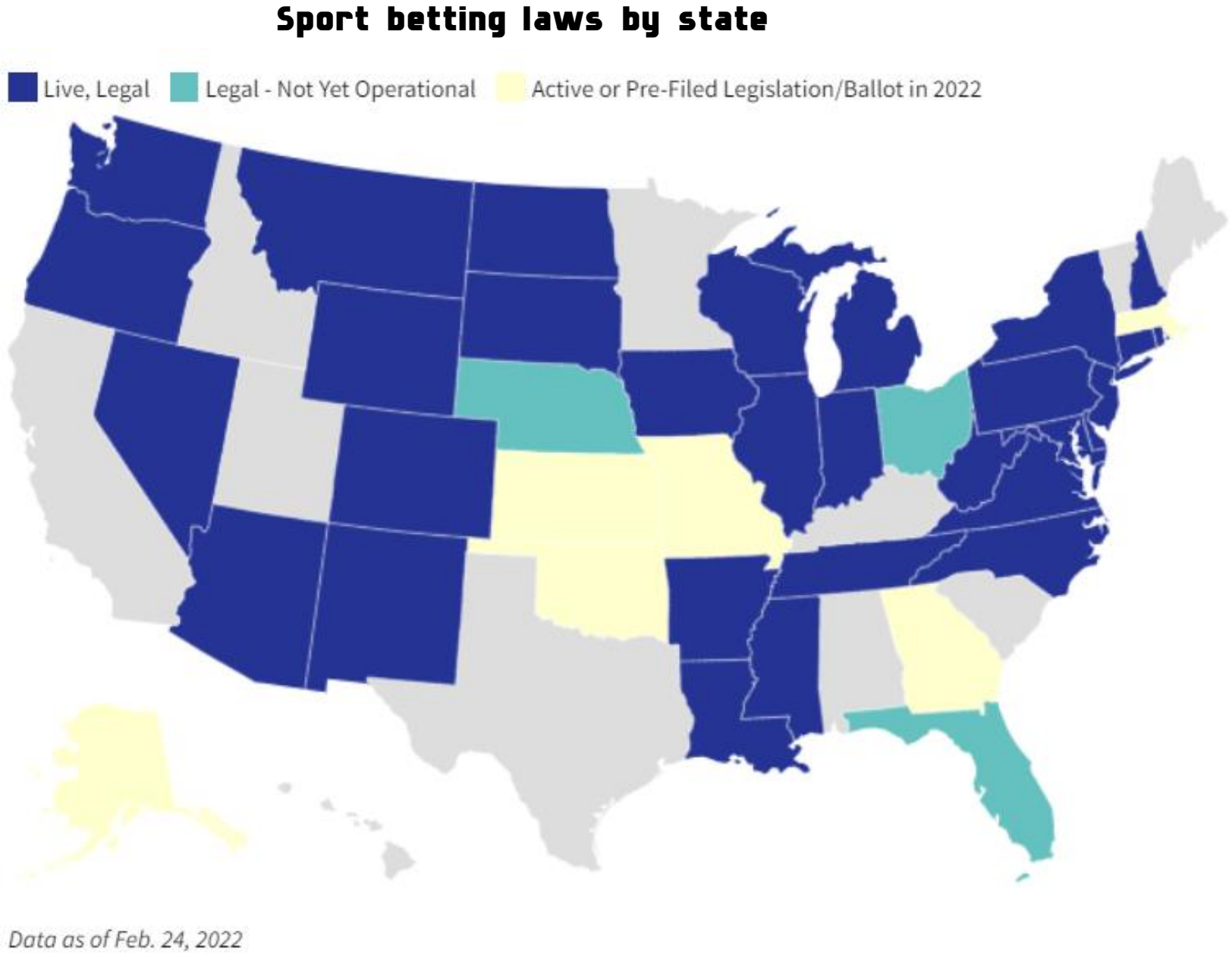
Agenda

- Motivation
- Data Exploration
- Model Results/Comparison
- Model Walkthrough
- Conclusion

Motivation

Why do we care who wins?

- Sports Betting Companies

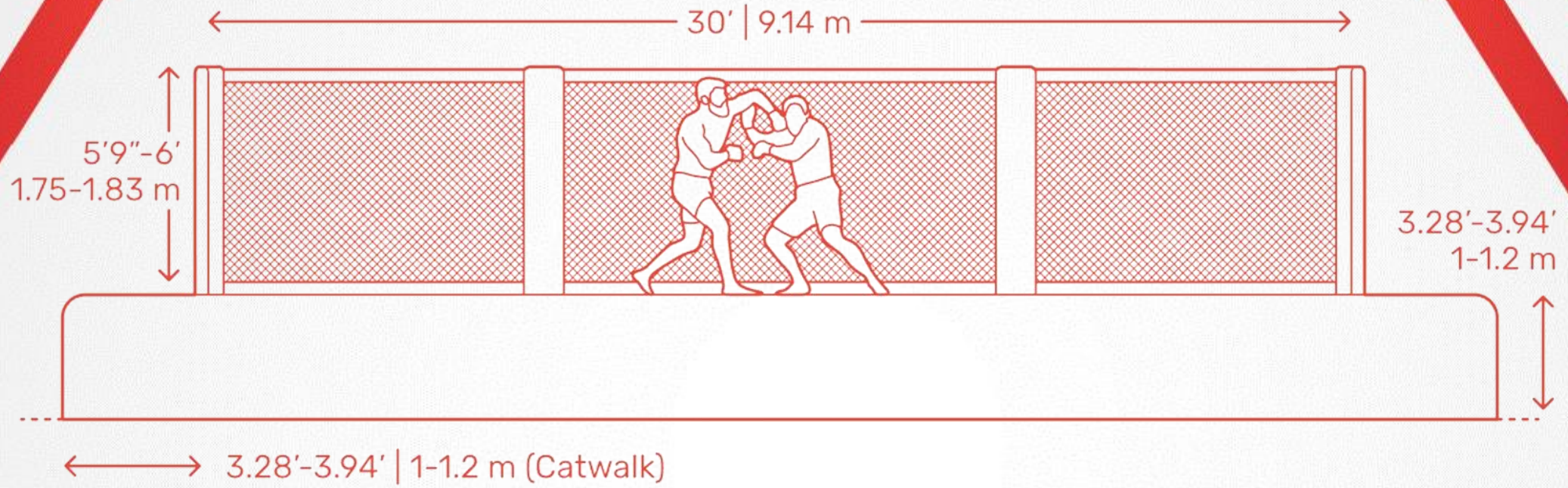


Data Exploration

Source

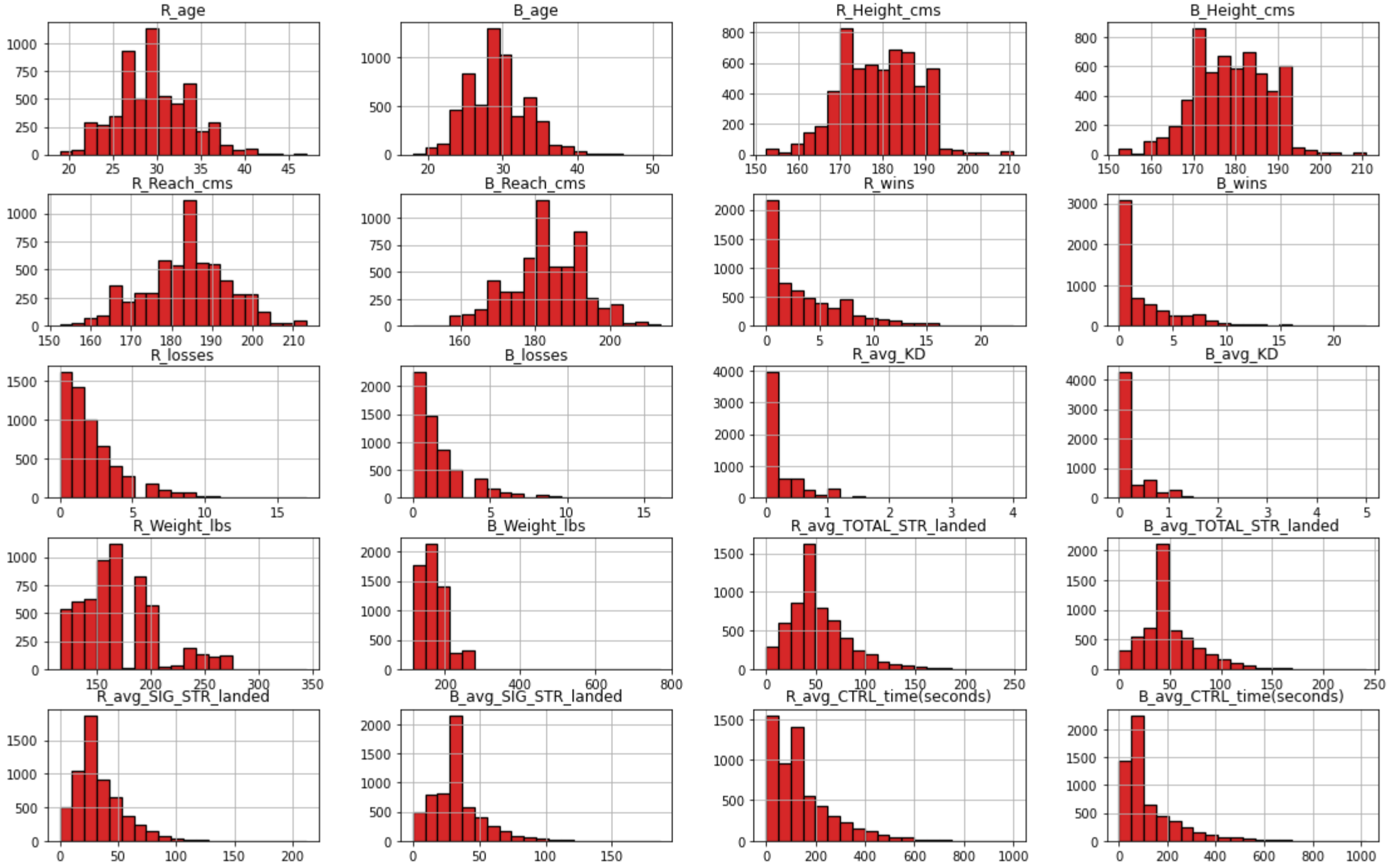
UFC-Fight historical data from 1993 to 2021

(by RAJEEV WARRIER, Kaggle)

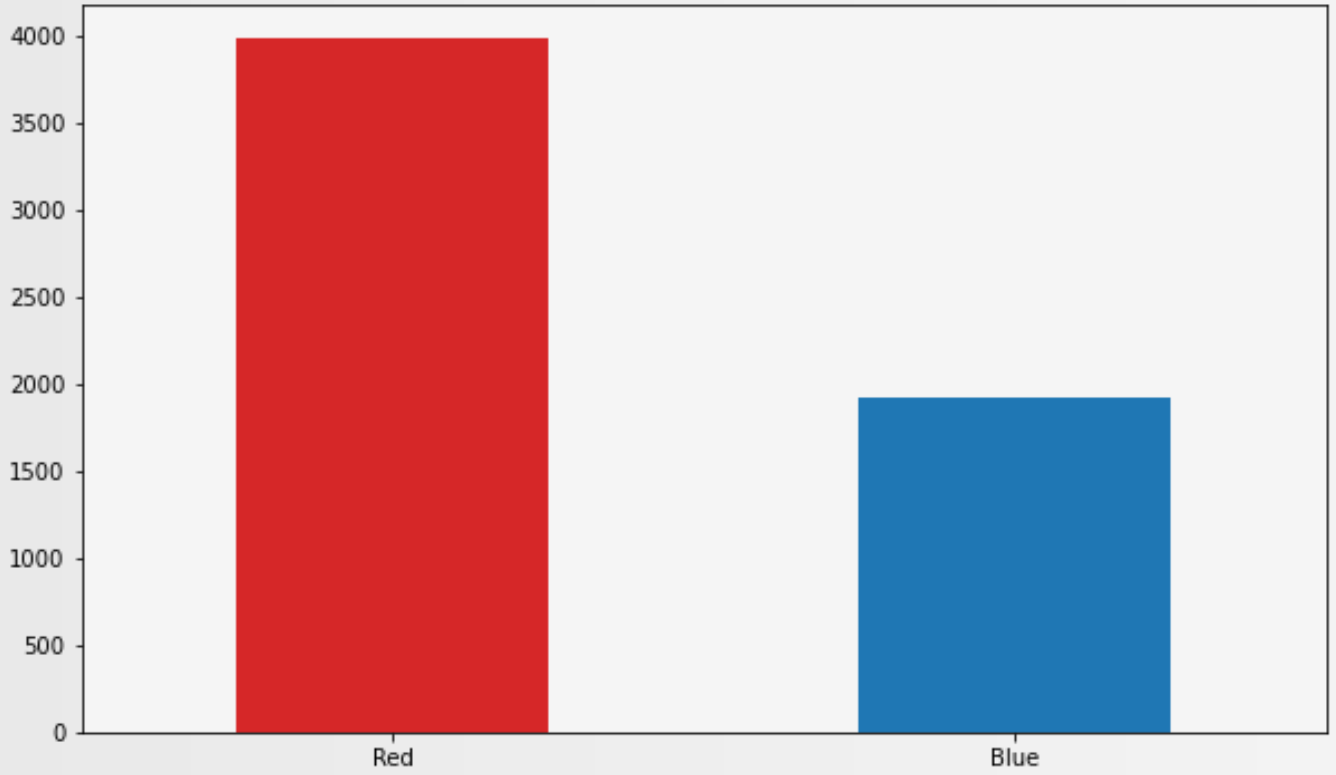


- Our preprocessed data set contains 5902 rows and 99 features
- Each row represents a single fight between two fighters
- The fighters are represented as Red or Blue
- We start with 99 features relating to fighter's physical and other characteristics

- Age
- height
- Weight
- Reach
- Stance/Style
- Strikes Landed in career
- Strikes Received in career
- Wins
- Losses
- Draws
- Knockdowns
- Takedowns
- Headshots
- Body shots
- Distance Strikes
- Significant Strikes
- Title Bout
- Control Time
- ETC...

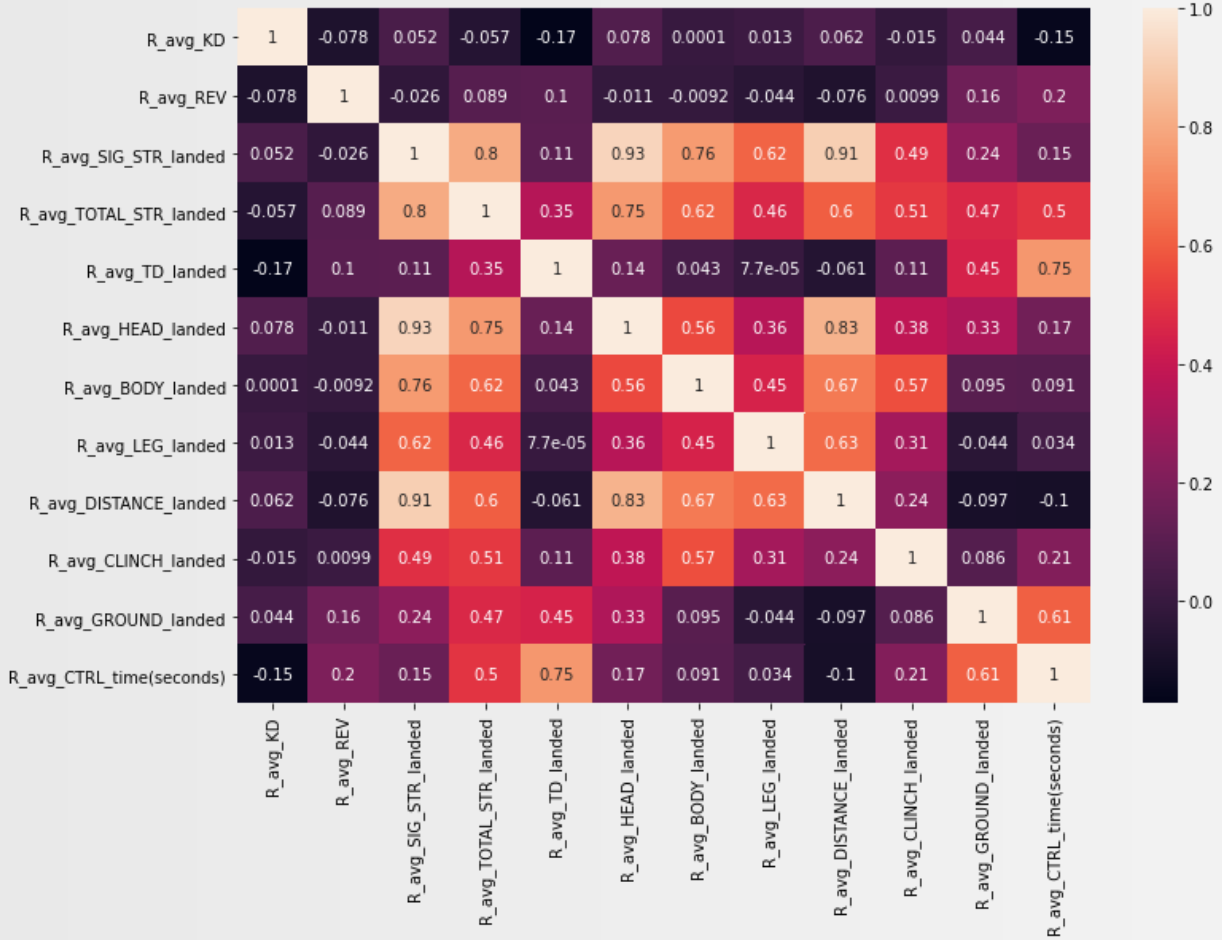

R_age
B_age
R_Height_cms
B_Height_cms
R_Reach_cms
B_Reach_cms
R_wins
B_wins
R_losses
B_losses
R_avg_KD
B_avg_KD
R_Weight_lbs
B_Weight_lbs
R_avg_TOTAL_STR_landed
B_avg_TOTAL_STR_landed
R_avg_SIG_STR_landed
B_avg_SIG_STR_landed
R_avg_CTRL_time(seconds)
B_avg_CTRL_time(seconds)

Predictor Variable: Fight Winner



- The dependent variable is imbalanced
  - Red-side winner: 3979 67.42%
  - Blue-side winner: 1923 32.58%

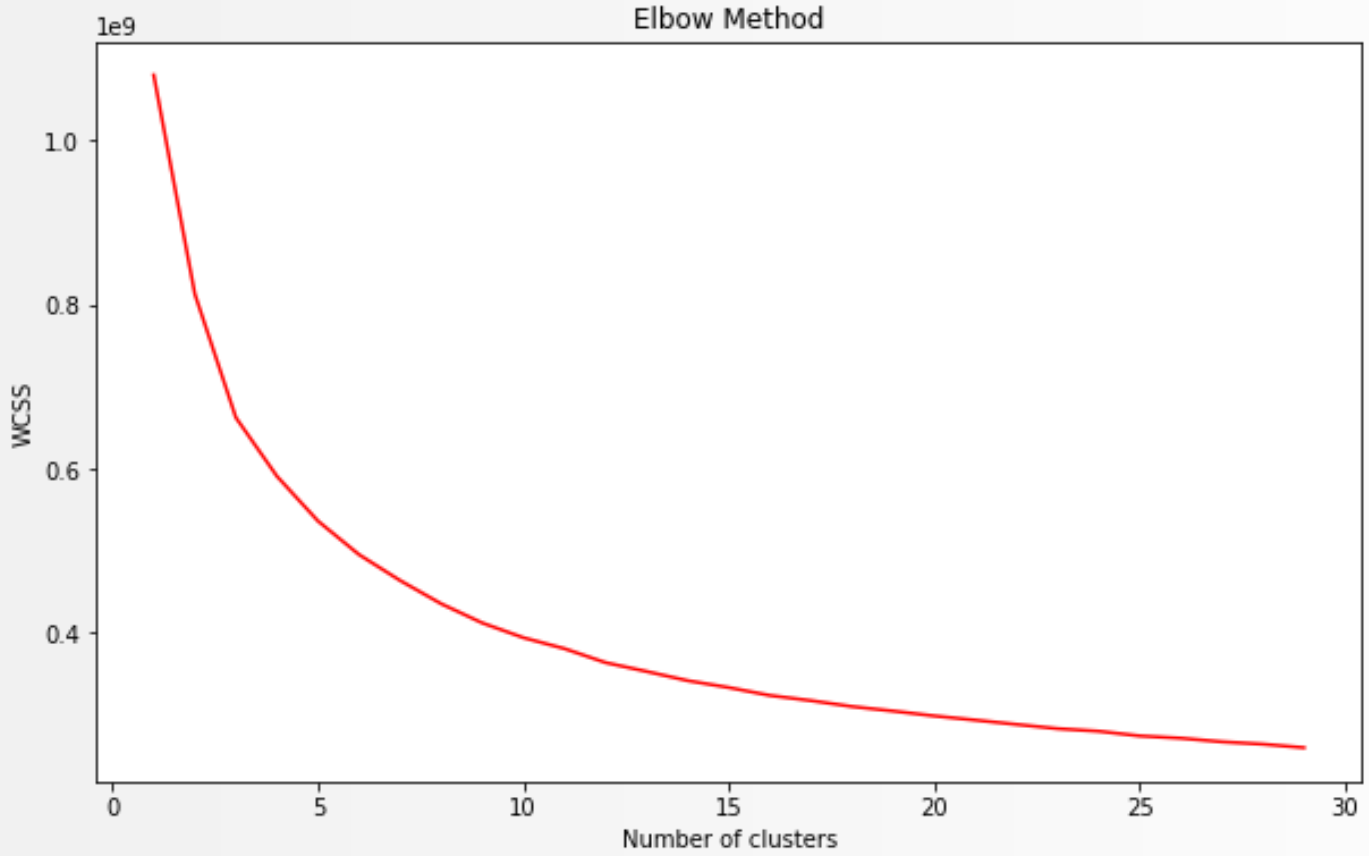
Correlation Matrix



- Correlation matrix used to look at X variable relationships
- Identify highly correlated attributes
- Helped with later Feature Selection

K-means Clustering

- **WCSS - For a given number of clusters K, this value represents a mathematical distance of sample to the cluster center, we want to minimize this value**
- **As seen from the right, this unsupervised method groups data into many clusters instead of our Y label, which groups it into 2 groups…Red or Blue**



Model Results Comparison

	KNN	DT	RF	Naive Bayes	Logistic Regression
Accuracy:	67.02%	60.64%	67.48%	68.94%	66.00%
True Positive Rate:	98.33%	70.65%	89.13%	85.40%	66.80%
True Negative Rate:	1.91%	39.83%	22.43%	34.20%	64.30%

Model Walkthrough

K Nearest Neighbor

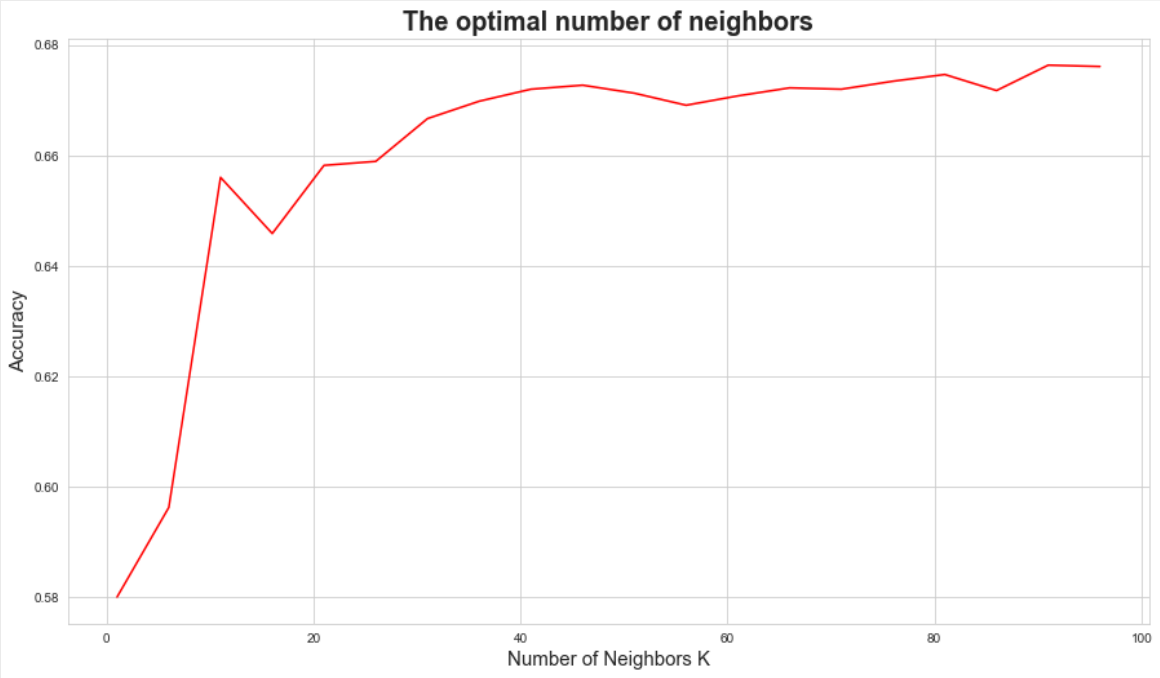
Decision Tree & Random Forest

Logistic Regression (Feature Selection & PCA)

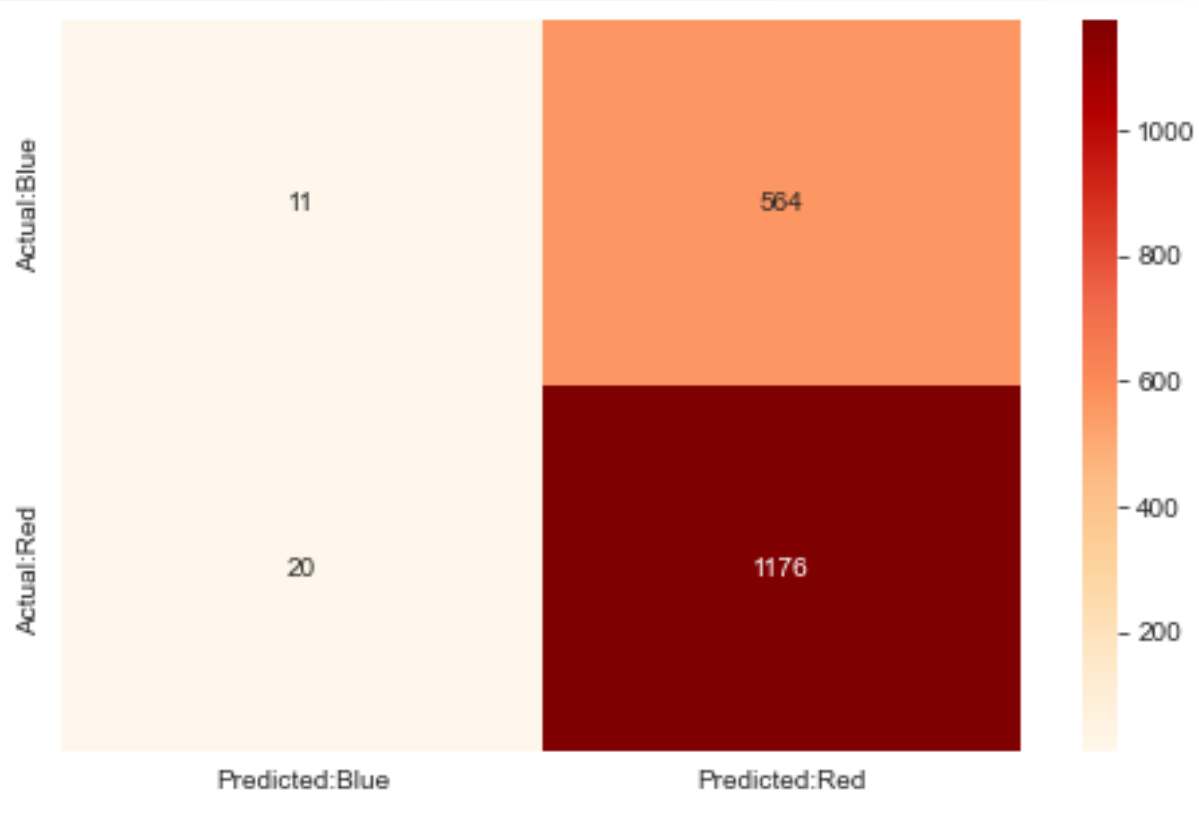
Gaussian Naïve Bayes

K Nearest Neighbor

- Use 10-fold Cross Validation to find the **optimal** number of neighbors
  - Optimal Neighbor: **91**

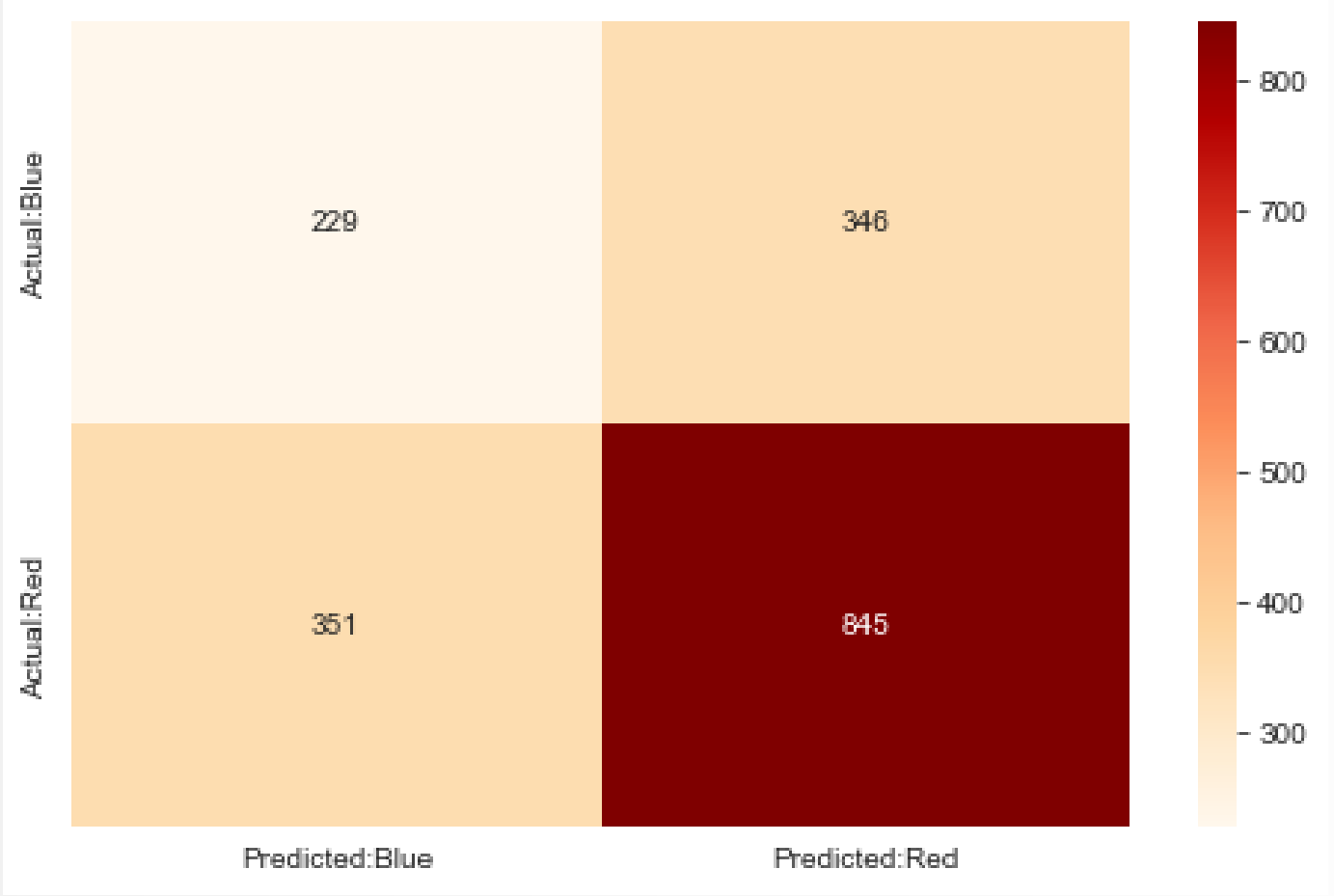


- Accuracy: 67.02%
- True Positive Rate: 98.33%
- True Negative Rate: 1.91%

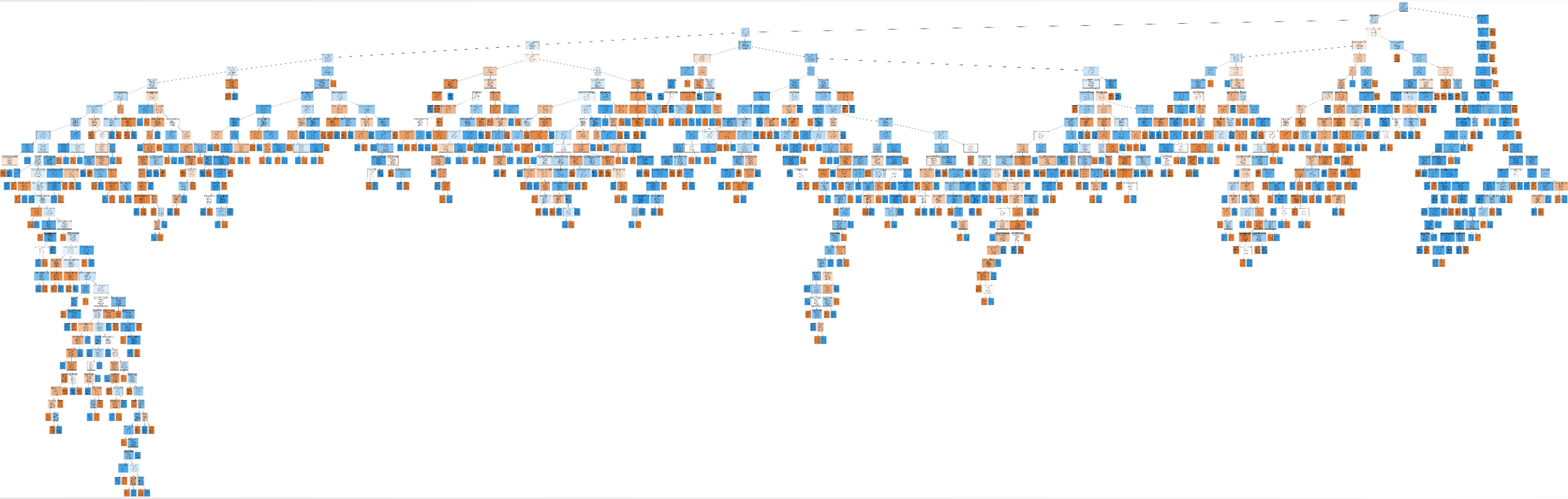


Decision Tree Classification

- Accuracy: 60.64%
- True Positive Rate: 70.65%
- True Negative Rate: 39.83%

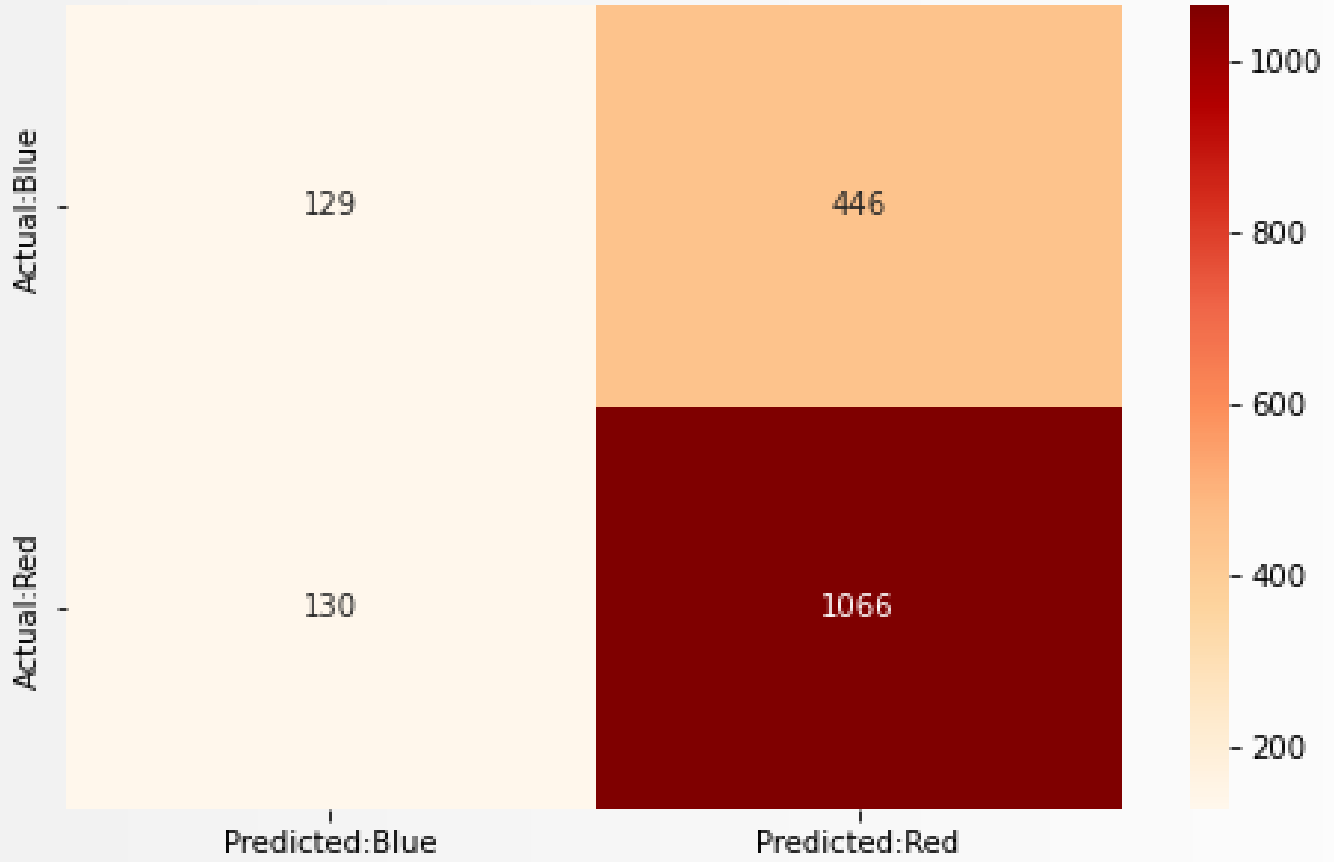


Decision Tree Graph



Random Forest Classification

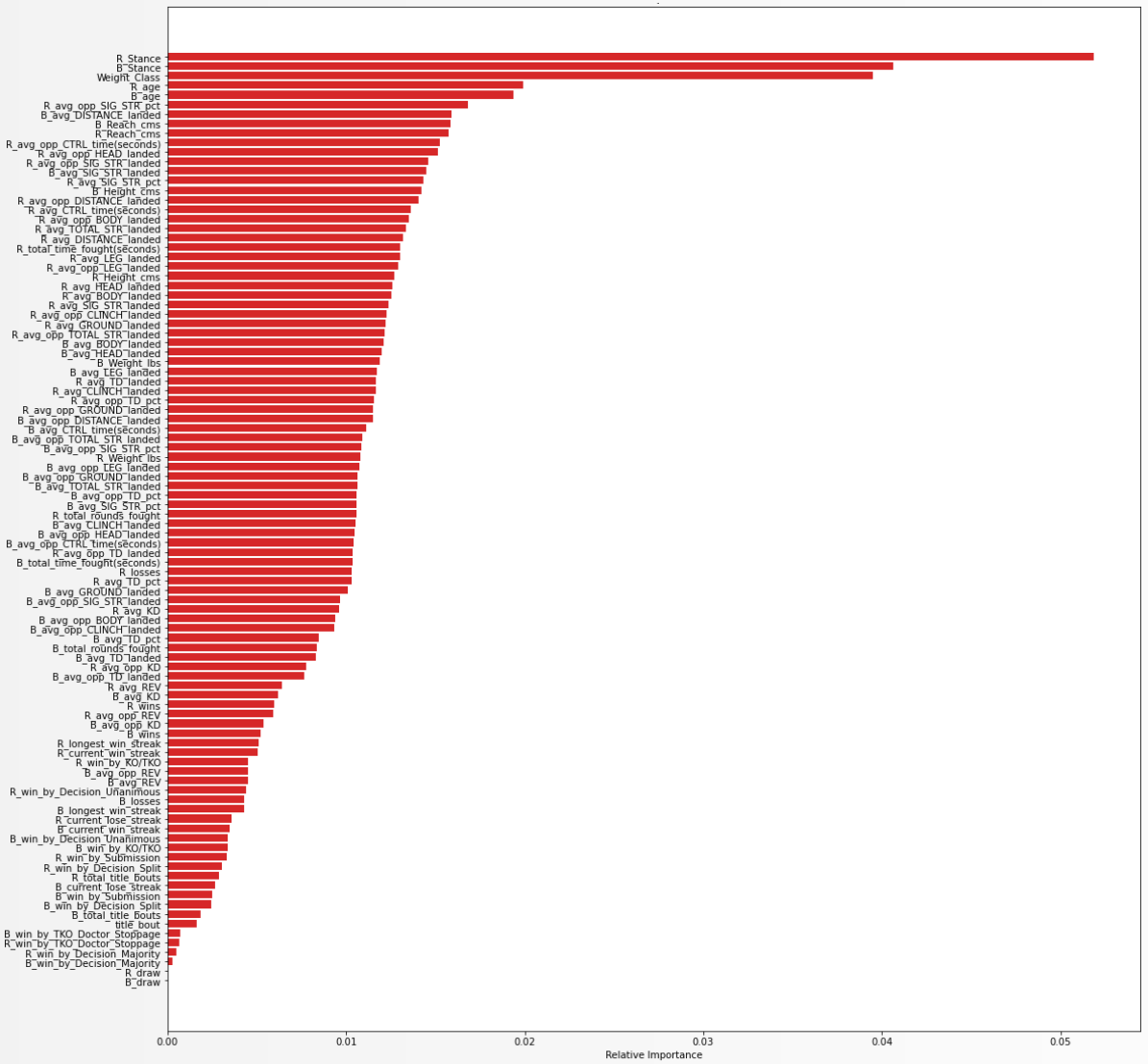
- Created from 100 small decision trees to improve the performance
  - Accuracy: 67.47%
  - True Positive Rate: 89.13%
  - True Negative Rate: 22.43%



Random Forest Classification

- Top 10 important features:
  - Red & Blue Fighter's **Stance**
  - **Weight Class**
  - Red & Blue Fighter's **Age**
  - Red Fighter's Average **Significant Strike** by opponent
  - Blue Fighter's **Distance Strikes** landed
  - Red & Blue Fighter's **Reach**(arm span) in Centimeters
  - Red Fighter's **Control Time**

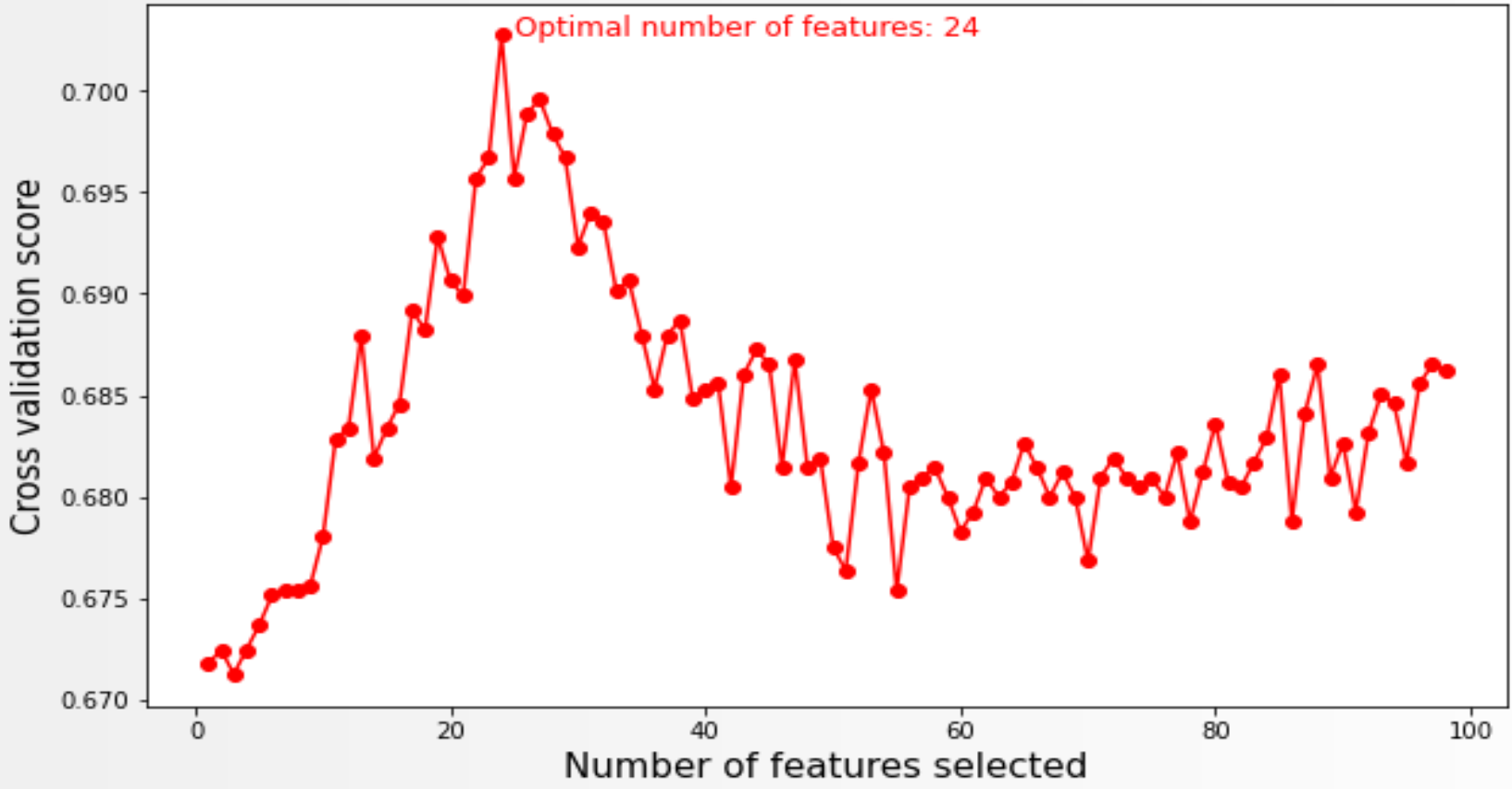
Feature Importance



Dimensionality Reduction

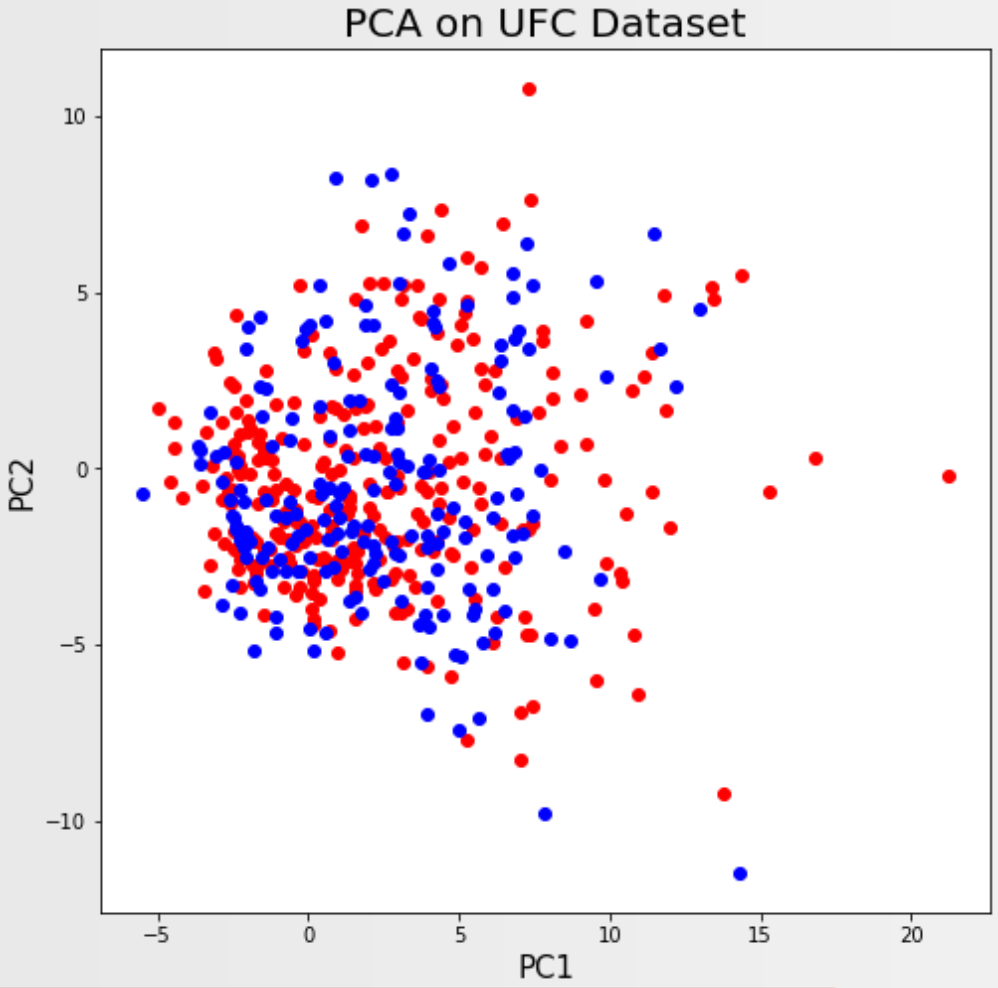
Feature Selection

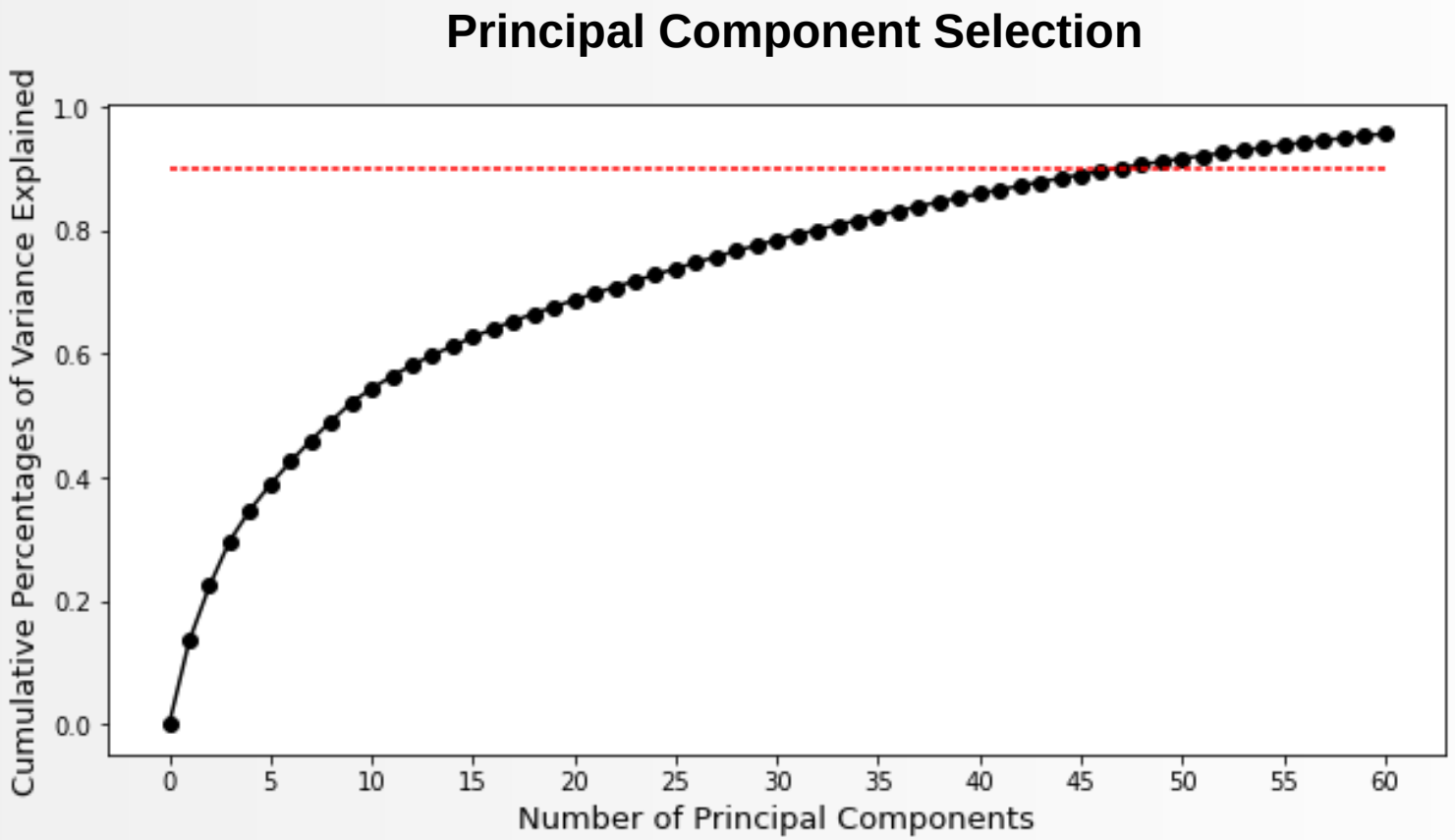
- Optimal number of features: 24
- Average cross-validation score: 0.703
  - Title fight
  - Age
  - Stance
  - Significant strikes
  - Knockdowns
  - Takedowns



Dimensionality Reduction

Principal Component Analysis

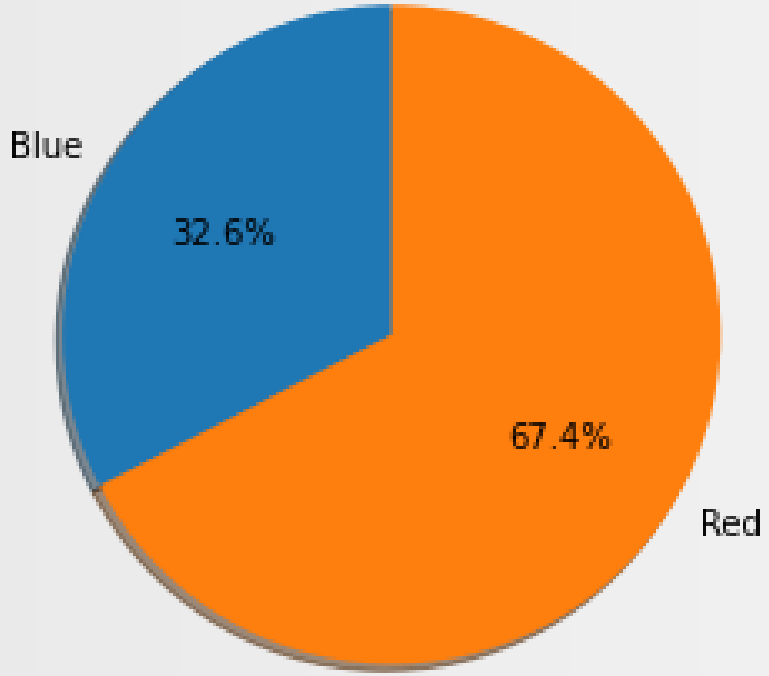




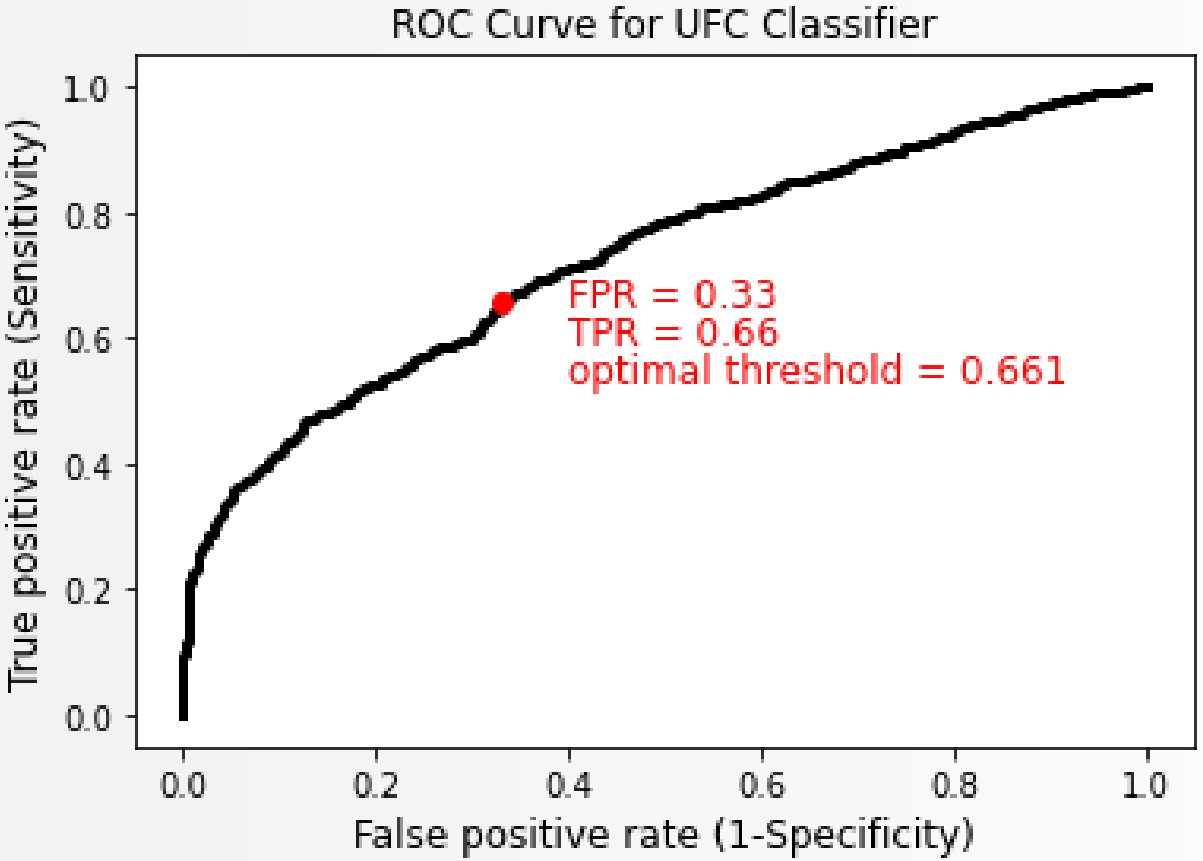
50 principal components ➡ 90% variance

Logistic Regression

Dealing with Imbalanced data



G-means ➡ optimal threshold

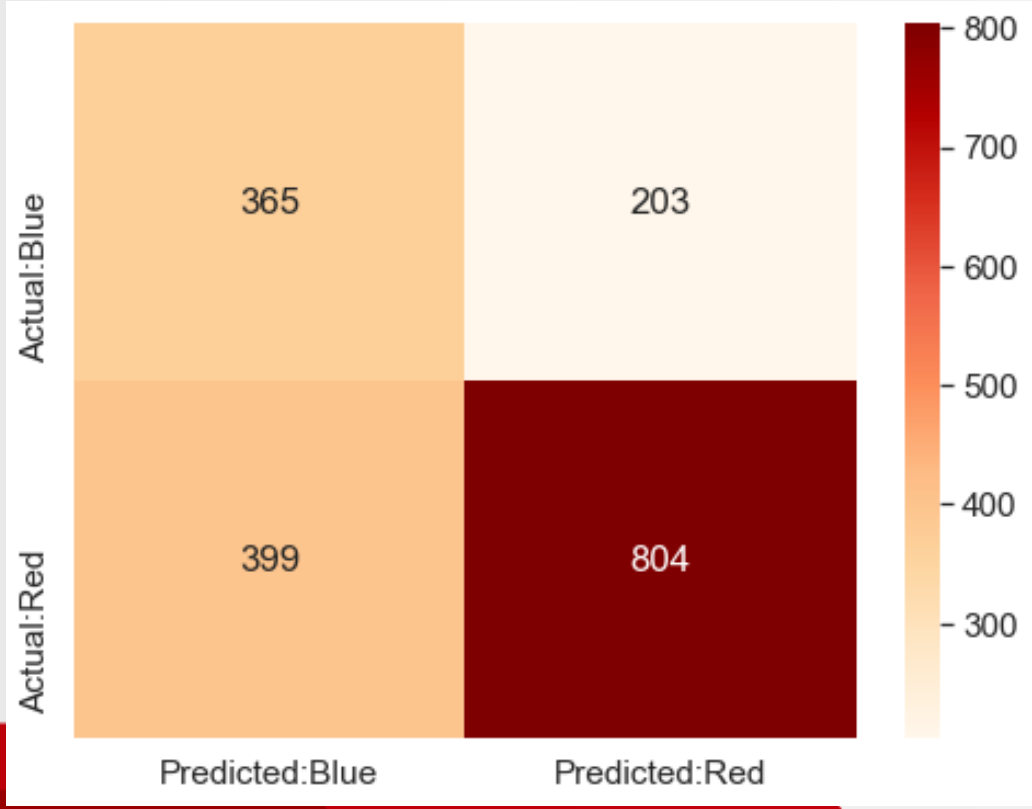


Logistic Regression

Classification Results

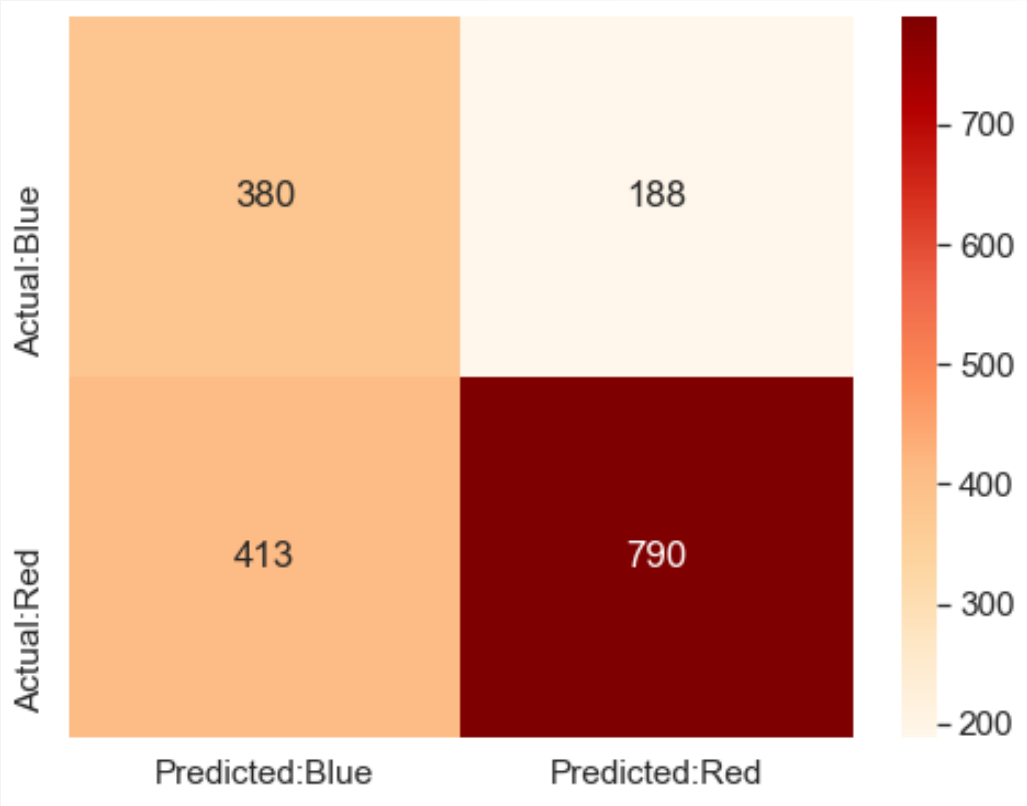
Feature Selection -

- Accuracy: 66.0%
- True Positive Rate: 66.8%
- True Negative Rate: 64.3%



Principal Components -

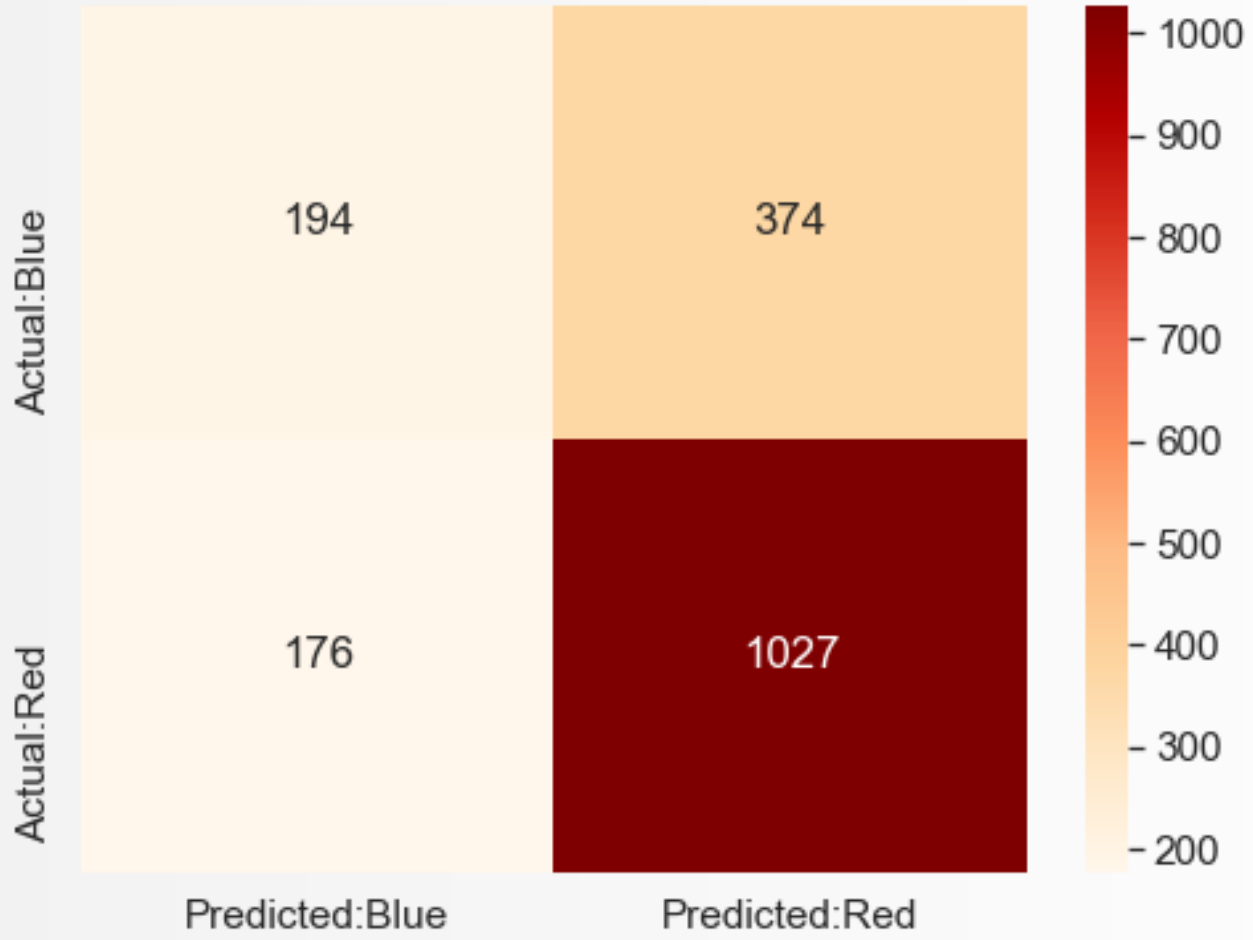
- Accuracy: 66.1%
- True Positive Rate: 65.7%
- True Negative Rate: 66.9%



Gaussian Naive Bayes

- Independent variables
- Normally distributed
- P(Observations | Red)*P(Red)
  VS.
  P(Observations | Blue)*P(Blue)

- Accuracy: 68.94%
- True Positive Rate: 85.4%
- True Negative Rate: 34.2%



Conclusion

- We chose a data set that would be fun and engaging

- Things to Note from EDA:
  - Imbalanced Dataset
  - Large number of features

- Best Model: Logistic Regression
  - Important Features: Age, Stance, Significant Strikes

- Things Learned:
  - Accuracy rate is not high, attributes relating to fighters may not be the only determinants of fight outcomes
  - Using Threshold analysis to seek balance between TPR and FPR helped us get a better model

- Future Works:
  - We may need to further fine tune our models such as Decision Tree for better performance
  - We want to look at other types of features and their effects on fight outcomes, i.e. Judges who scored fights, coach information, Team information. In more recent months, other characteristic like a fighter's fame can often sway decisions (as famous fighters can make the UFC promotion more money).