MGT 6203 / Spring 2024 / Team 007 / Project Final Report

UFC Fight Prediction

MGT-6203-Spring-2024-Canvas/Team-7

Alejandro Forero
aeng61@gatech.edu

Andrew Berkowitz
aberkowitz8@gatech.edu

Samuel W. Wang
swang3068@gatech.edu

Juliette Noelle Wong
jwong88@gatech.edu

Abstract

In this project, we apply machine learning techniques to predict fight outcomes and pay-per-view (PPV) sales for the Ultimate Fighting Championship (UFC). After a comprehensive data cleaning and feature selection on a dataset of historical UFC fight statistics, we trained and evaluated logistic regression, random forest, gradient boosting models for the UFC fights dataset, and a linear regression model for PPV sales forecasting. The XGBoost model achieved the best performance with an AUC of 0.938 on the test set, followed by logistic regression with an AUC of 0.873 and random forest with an accuracy of 0.55. The linear regression model for PPV buys yielded somewhat poor results, with an adjusted R^2 of 0.3175. While the specific goals of our models were to answer questions such as *"Who is going to win?"*, *"What strategies should a fighter adopt to win a fight?"*, and *"How can UFC sell more pay-per-views?"*, the overarching purpose of this work is to demonstrate a comprehensive, step-by-step approach to data cleaning, model training, selection, and evaluation, as well as presenting results and measuring model performance.

Table of Contents

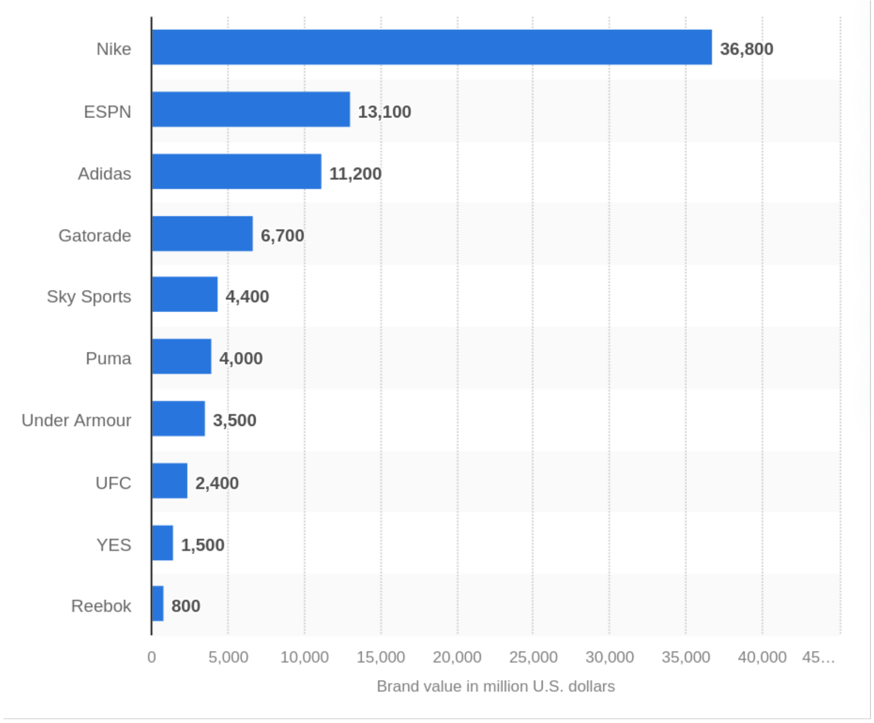
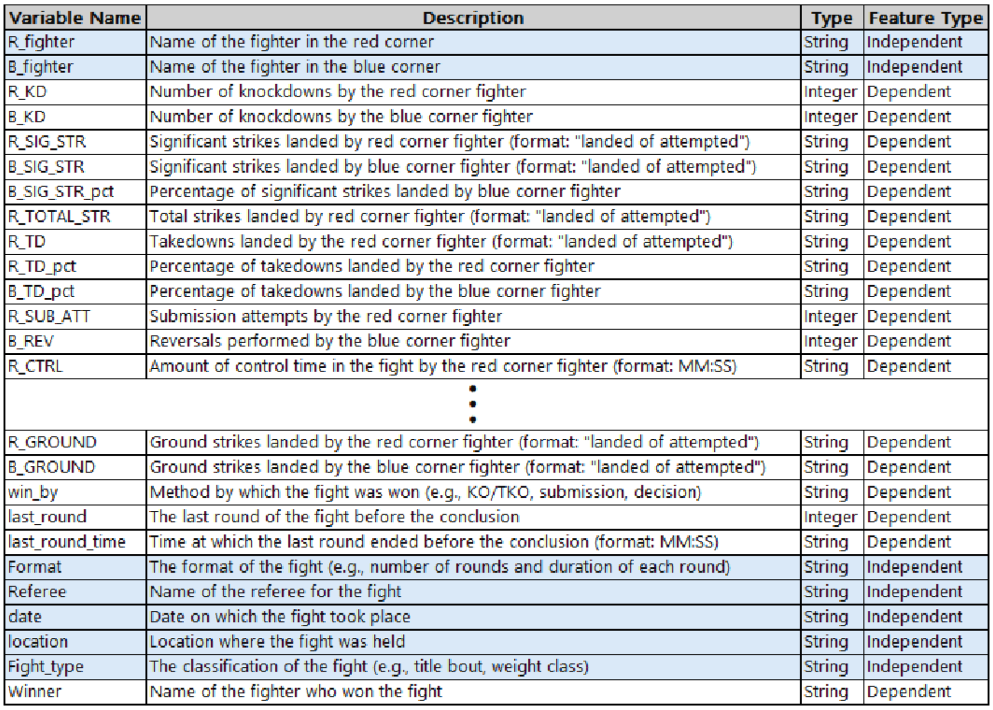


Figure 1: Most valuable sports brands worldwide in 2019 (in million USD). Courtesy to [2].

1 Introduction

The Ultimate Fighting Championship (UFC) has become the world's premier mixed martial arts (MMA) organization since its inception in 1993. Over the years, the UFC has experienced remarkable growth, with revenue exceeding $600 million from 2013 to 2016, representing a staggering 335% single-year growth [1]. In 2019, the UFC was ranked among the top 10 most valuable sports business companies globally [2]; see Figure 1. The organization's popularity continues to soar, with its highest live attendance surpassing 57,000 in August 2022 [3] and an average of 447,000 pay-per-view buys in 2018, peaking at 2.4 million buys [1]. As of December 2023, with the UFC's broadcast deal with ESPN, MMA has become the 9th most popular sport regarding consumer interest in the United States [4]. The UFC's success has undeniably played a significant role in propelling MMA to the forefront of the sports world.

The rise of MMA's popularity and the UFC's dominating viewership has naturally led to a growing interest in sports betting. With the vast array of fight statistics available, both individuals and sports betting companies stand to benefit from a better understanding of the factors that drive specific outcomes and how to predict them more accurately. Accurate predictions not only help individuals place winning bets, but also enable companies to set more precise odds. Moreover, by understanding the individual statistics that influence fight outcomes, we can further investigate if these statistics and specific fighters have an impact on the number of pay-per-view (PPV) buys. Gaining insights into how to increase PPV sales is crucial for ensuring the continued growth and success of the UFC. In this project, we adopt a traditional and straightforward approach to tackle these questions, employing various machine learning algorithms and techniques to uncover valuable insights that can benefit both fight predictions and the UFC's business strategies.



Variable Name	Description	Type	Feature Type
R_fighter	Name of the fighter in the red corner	String	Independent
B_fighter	Name of the fighter in the blue corner	String	Independent
R_KD	Number of knockdowns by the red corner fighter	Integer	Dependent
B_KD	Number of knockdowns by the blue corner fighter	Integer	Dependent
R_SIG_STR	Significant strikes landed by red corner fighter (format: "landed of attempted")	String	Dependent
B_SIG_STR	Significant strikes landed by blue corner fighter (format: "landed of attempted")	String	Dependent
B_SIG_STR_pct	Percentage of significant strikes landed by blue corner fighter	String	Dependent
R_TOTAL_STR	Total strikes landed by red corner fighter (format: "landed of attempted")	String	Dependent
R_TD	Takedowns landed by the red corner fighter (format: "landed of attempted")	String	Dependent
R_TD_pct	Percentage of takedowns landed by the red corner fighter	String	Dependent
B_TD_pct	Percentage of takedowns landed by the blue corner fighter	String	Dependent
R_SUB_ATT	Submission attempts by the red corner fighter	Integer	Dependent
B_REV	Reversals performed by the blue corner fighter	Integer	Dependent
R_CTRL	Amount of control time in the fight by the red corner fighter (format: MM:SS)	String	Dependent
	⋮		
R_GROUND	Ground strikes landed by the red corner fighter (format: "landed of attempted")	String	Dependent
B_GROUND	Ground strikes landed by the blue corner fighter (format: "landed of attempted")	String	Dependent
win_by	Method by which the fight was won (e.g., KO/TKO, submission, decision)	String	Dependent
last_round	The last round of the fight before the conclusion	Integer	Dependent
last_round_time	Time at which the last round ended before the conclusion (format: MM:SS)	String	Dependent
Format	The format of the fight (e.g., number of rounds and duration of each round)	String	Independent
Referee	Name of the referee for the fight	String	Independent
date	Date on which the fight took place	String	Independent
location	Location where the fight was held	String	Independent
Fight_type	The classification of the fight (e.g., title bout, weight class)	String	Independent
Winner	Name of the fighter who won the fight	String	Dependent

Figure 2: A view of the 41 feature variables of the UFC dataset ufg_data_tlll_UFC_292.csv. This image shows some of them.

2 Data Cleaning and Feature Selection

Our team worked together to research the topic to fill in any knowledge gaps, learn about what others have achieved, and plan out a process for how to approach the problems we are trying to solve with the data available to us.

Our dataset came from two sources from Kaggle, and the corresponding links are given as follows:

Description	File Name	Source
UFC fight data from 1993 to 2023	ufc_data_till_UFC_292.csv	🔗 Link to the dataset
UFC PPV Sales	ufc_ppv_buys.csv	🔗 Link to the dataset

We cleaned out data into usable formats to answer our questions. The method we used to clean the dataset was to create separate datasets that would be easily joined together if needed. The datatypes of many columns in the main dataset, ufc_data_till_UFC_292.csv, were character, so during our preprocessing, we had to convert them into usable datatypes. Our secondary dataset, ufc_ppv_buys.csv, needed some cleaning to fix inaccurate data and data that was incompatible with the primary dataset. A brief view of the first dataset is given in Figure 2.

Beyond converting data types, we utilized string processing, data format conversions, and other data cleaning techniques to get the dataset ready for training the models. Besides these data cleaning works, additional feature engineering were performed as follows.

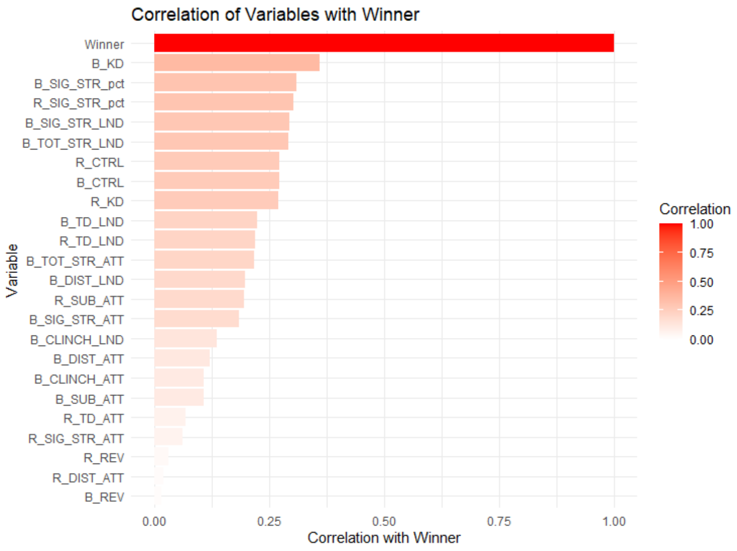


Figure 3: Correlation of feature variables with winner after selecting them using p-value technique.

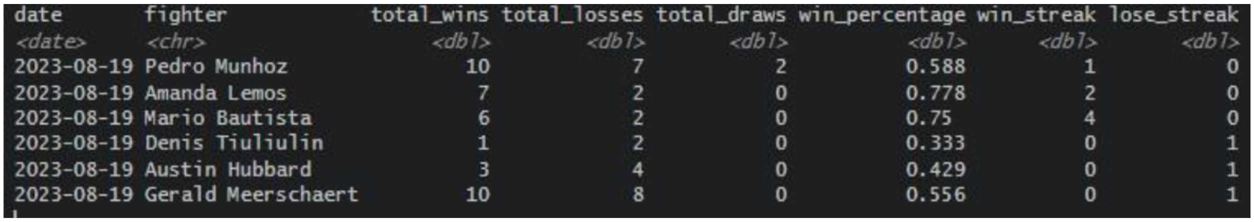
For the main dataset, we initially transformed each categorical feature into dummy variables, resulting in an expansion of the feature space by over 900 dimensions. This led to the "curse of dimensionality," which could result in overfitting, difficulties in interpreting the model, and convergence issues. Notably, this expansion did not account for the more than 2,000 fighters in the dataset. Furthermore, each fighter's performance was considered an autoregressive variable, as it changes over time. So, what appeared to be a straightforward dataset evolved into one characterized by both high dimensionality and autoregressive properties.

To address these complexities, we adopted a dual-strategy approach. Initially, we sought to characterize the fighters without directly using their names as model parameters. Instead, we employed the Cumulative Sum (CUSUM) technique to aggregate statistical data for each fighter based on the lengths of their past matches. This method significantly reduced the number of variables and mitigated the issues associated with incorporating time as a feature.

Secondly, aiming to enhance model interpretability, we implemented Recursive Feature Selection (RFS), Variance Inflation Factor (VIF), and p-value assessments to identify the most significant predictors of victory and PPV buys. Unfortunately, the recursive feature selection process, after running for two days, converged on only 15 variables, hence we had to discard this approach. The latter was more convenient, since it led us to only 23 features strongly correlated with winner prediction as in Figure 3.

Since our dataset only contained a column for the winner, we used the names of the fighters in each corner (red/blue) to determine the loser. From there, we could build a new dataset capturing details like the fighter, winner, loser, and the date of the fight.

Using the fight results, we hypothesize fighter's win or loss streak could predict things like PPV sales or fight outcomes. By grouping data by each fighter and sorting it by date, we calculated cumulative wins, losses, and draws leading up to a match on any specific date. We then applied the `streak_run` function to pinpoint winning or losing streaks. This enriched dataset gives us a handy snapshot of a fighter's record—wins, losses,



```
date       fighter            total_wins total_losses total_draws win_percentage win_streak lose_streak
<date>     <chr>                   <dbl>        <dbl>       <dbl>          <dbl>      <dbl>       <dbl>
2023-08-19 Pedro Munhoz               10            7           2          0.588          1           0
2023-08-19 Amanda Lemos                7            2           0          0.778          2           0
2023-08-19 Mario Bautista              6            2           0          0.75           4           0
2023-08-19 Denis Tiuliulin             1            2           0          0.333          0           1
2023-08-19 Austin Hubbard              3            4           0          0.429          0           1
2023-08-19 Gerald Meerschaert         10            8           0          0.556          0           1
```

Figure 4: A portion of our calculated aggregated fighter record data.

draws, win percentage, and current streaks—right before they head into a match. A portion of this dataset is given in Figure 4.

Lastly, we cleaned our secondary dataset of the pay-per-view data. We first had to combine the year, month, and date columns into one date column to easily match with our other datasets. While attempting to join this dataset with our main dataset, we noticed and manually fixed some inaccuracies found in the dataset with the UFC fight dates. Finally, we standardized the fighters in the dataset with the fighters from our main dataset because they sometimes used a fighter's nickname instead of the fighter's last name.

3 Exploratory Data Analysis

Before answering the question *"Who will win the fight?"*, we first wanted to have some insight on what the data are saying prior to running any models. Some factors we thought would be important in determining the outcome would be win/loss streaks, strikes landed, and control time. We looked at those metrics in relation to the winner and found that the strikes and control advantage aligned with our hypothesis, but the streaks did not. Some preliminary visualizations are given in Figure 5.

Similarly, we wanted to conduct exploratory data analysis before forecasting and conducting factor analysis on PPV buys. We hypothesized that the date, win streak of fighters, the continent the match took place in (because of time zone differences), and whether the main event was a title match would affect PPV buys. See Figure 6.

4 Model Selection and Results

We decided to use different machine learning algorithms such as Logistic Regression, Random Forest, and Gradient Boosting for the UFC fights dataset to analyze factors. Additionally, we applied a linear regression model for the UFC PPV dataset.

4.1 Who will win the fight?

4.1.1 Logistic Regression Model

For our primary question of *"Who will win the fight?"*, a logistic regression model was an easy choice since we will be looking for a probability to be the winner. After a few iterations of the model, we finalized with 23 variables that were statistically significant to 0.05. Using the final logistic regression model on our test set of data, a threshold cutoff of 0.7 provided the highest area under the curve of 0.8641. This works for correctly guessing a winner, but what happens if we need to take into consideration the risk tolerance of a better? We calculated threshold predictions that minimize loss, which was a threshold of 0.13 but this

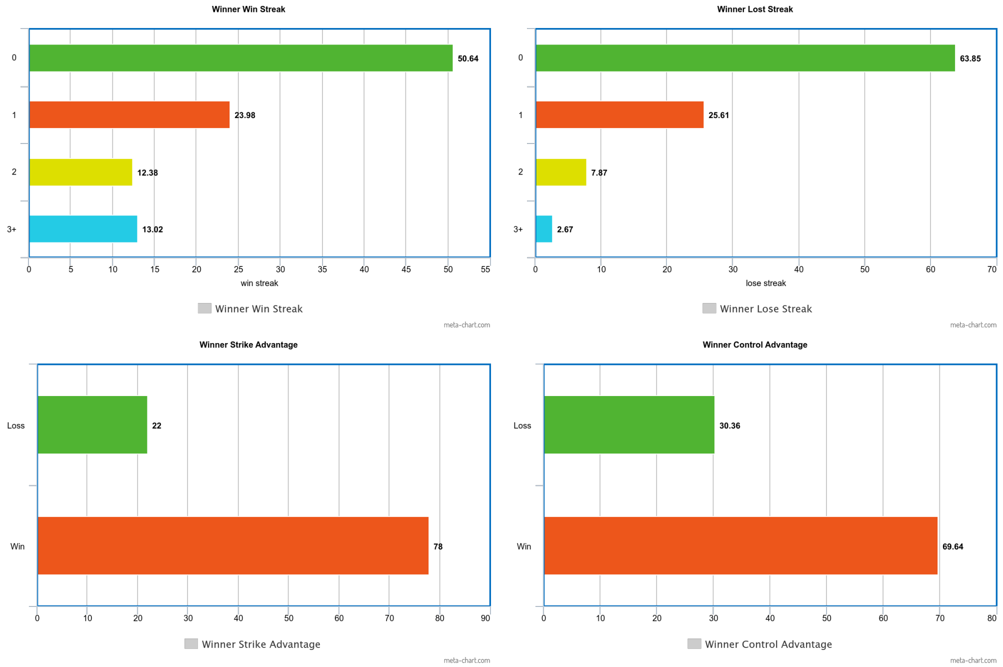


Figure 5: **Top left**. Percentages of win streaks among winners. **Top right**. Percentages of lose streaks among winners. **Bottom left**. Percentages of strike advantage among winners. **Bottom right**. Percentages of percentages of control advantage among winner.

threshold translated to a 0.7318 area under the curve. With the options for using different thresholds available to minimize loss or maximize wins, we can use the model by inputting aggregated data for a matchup and have a probability of a winner. The plots of different types of ROC curves are given in Figure 7.

4.1.2 Gradient Boosting Method

We specifically employed the XGBoost library, an optimized distributed gradient boosting framework, which offers several advantages over regular gradient boosting implementations. It incorporates regularization techniques to control model complexity and prevent overfitting. Additionally, XGBoost supports parallel processing, allowing for faster training times on large datasets (the latter is not our case).

To train our XGBoost model, we first converted our preprocessed training data into the DMatrix format, which is optimized for XGBoost. We then defined the model parameters, specifying the objective function as binary logistic regression and setting the evaluation metric to the area under the ROC curve (AUC). Through experimentation, we tuned hyperparameters such as the learning rate (0.3), maximum tree depth (4), and subsampling ratios to strike a balance between model complexity and generalization performance.

After training the XGBoost model for an optimal number of rounds, we evaluated its performance on the test set. The model achieved a pretty good AUC score (0.938), surpassing the performance of our logistic regression model. This improvement can be attributed to XGBoost's ability to capture complex non-linear relationships and interactions among the features. See Figure 8.

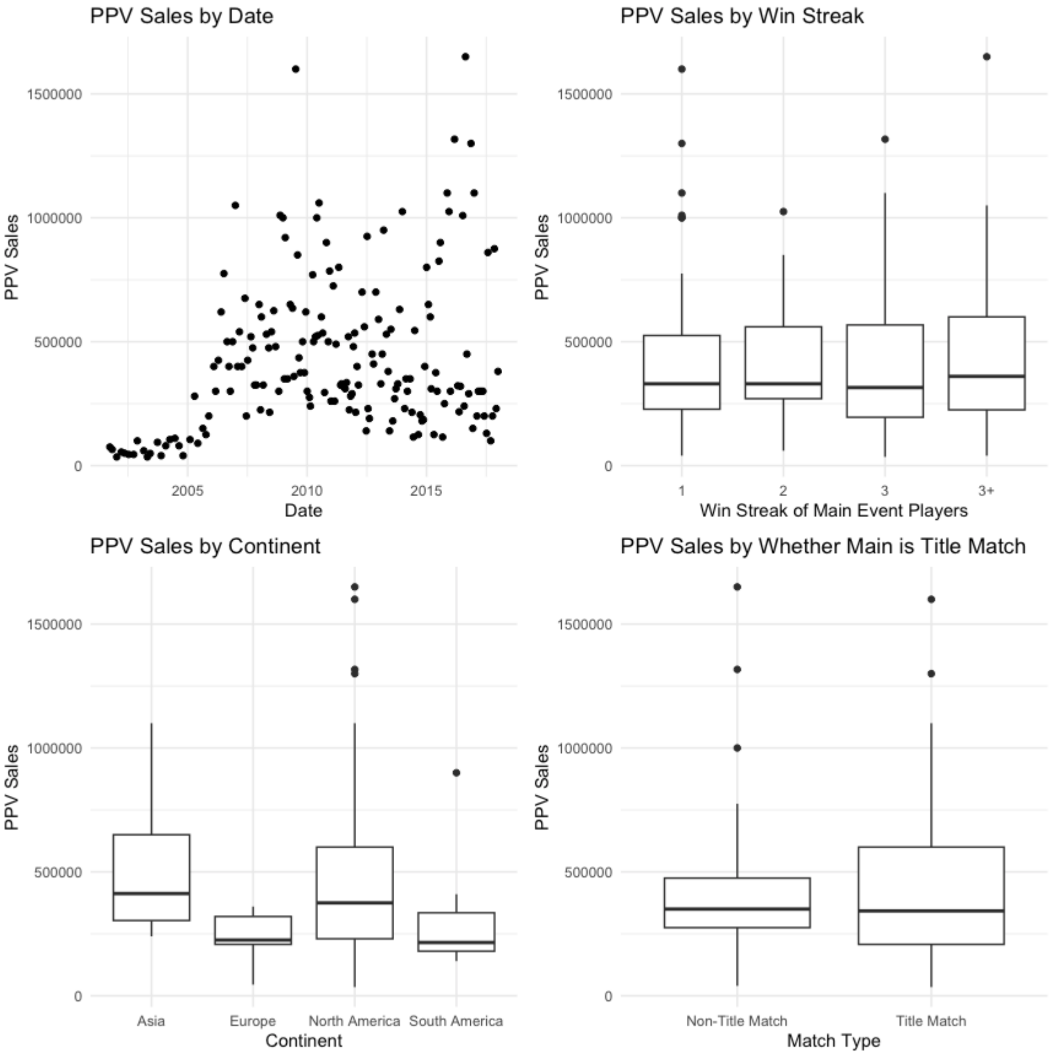


Figure 6: **Top left.** PPV Sales by Date. **Top right.** PPV Sales by Win Streak. **Bottom left.** PPV Sales by Continent. **Bottom right.** PPV Sales by whether the main event was a title match.

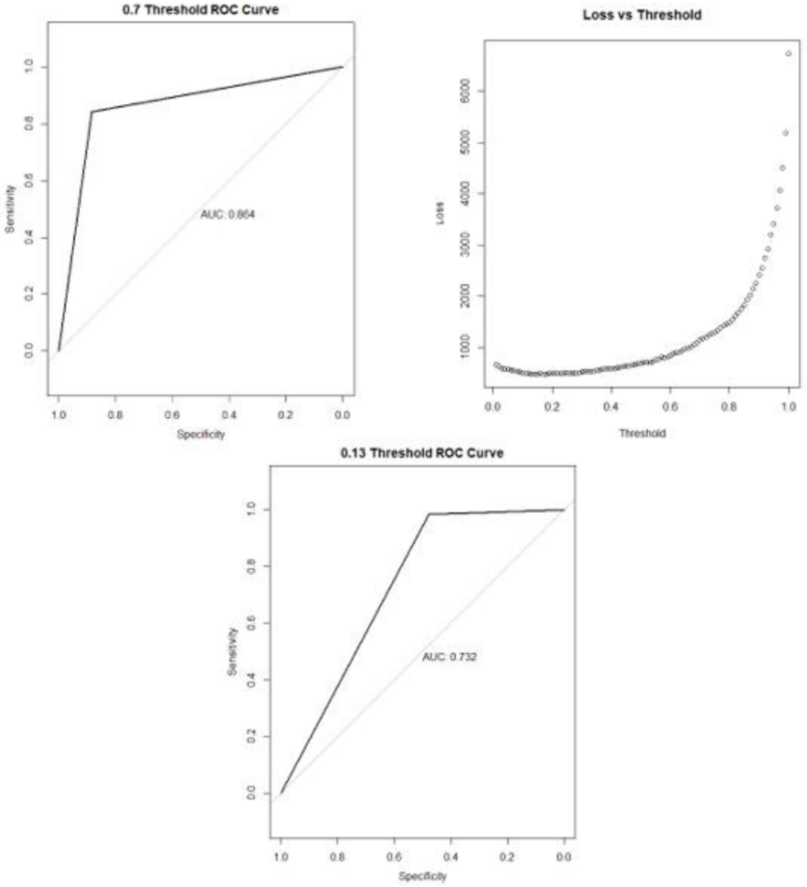


Figure 7: **Top left.** ROC curve with 0.7 threshold. **Top right.** Losses (incorrect) of model by threshold. **Bottom**. ROC curve with 0.13 threshold.

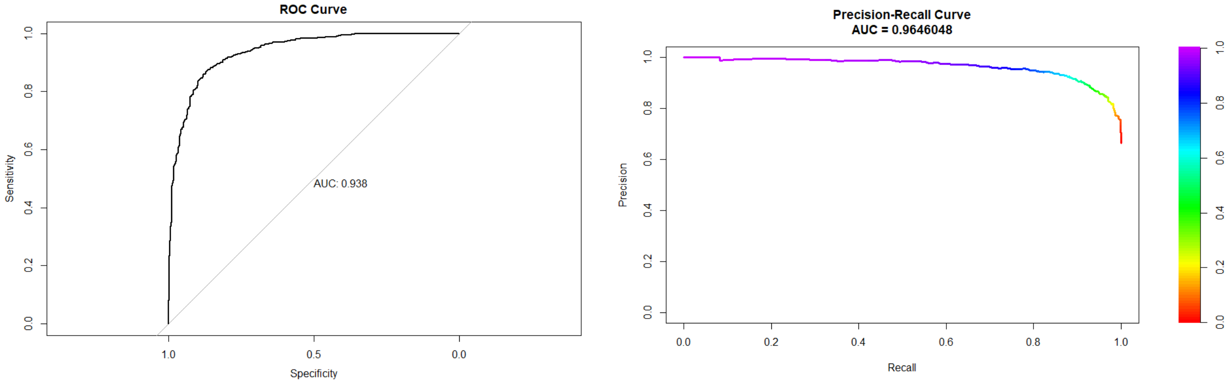


Figure 8: The ROC curve and the PRC curve of our XGBoost model.

4.1.3 Random Forest

We built a random forest model to predict fight outcomes and used a 10-fold cross validation design. The data we modeled upon were given by the fight data joined with aggregated stats.

Our random forest model was built using a grid search among different choices of a few parameters. The parameters we tried is defined by the grid as given as follows:

```
"n_estimators":         [80],
"max_depth":            [4, 6, 8, 12, 15],
"max_features":         ['sqrt', 'log2'],
"min_samples_split":    [2, 10, 30],
"min_samples_leaf":     [1, 3, 6]
```

Sadly, in a 36-minute run, the best model turned out to be highly suboptimal, with an accuracy score of 55.06%, which is not much better than the expected accuracy score of randomly guessing winner or loser by flipping a fair coin.

We suspect that the potential reasons that this initial random forest model didn't do well with this data are the following two:

- Firstly, we know that random forests are built upon decision trees ensembled together on different bootstrapped versions of the data. Since decision trees base each decision thump on one single feature, under the assumption that each fight is determined by the relative difference of stats from both sides, and that each feature in our data (no matter it being an aggregated statistic or not) can only demonstrate the stat of one side, it is highly like that each thump could not see any relative differences in fighter stats.

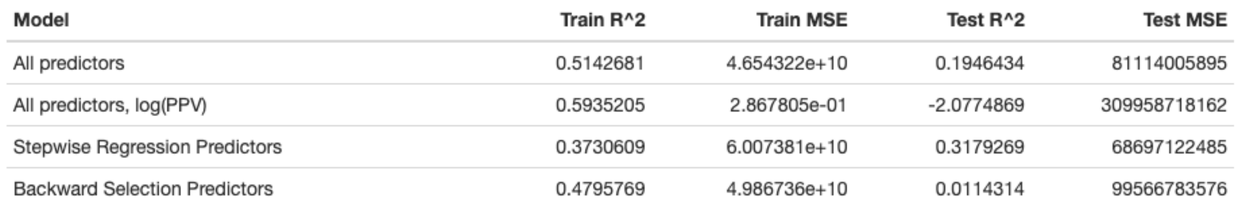
- Secondly, random forest's `max_features` parameter can be crucial in the performance of the model. In our grid search above, we only used a random subset of data when building each decision trees. Therefore, for random forest to perform better, we either have to throw away non-relevant features in the training data, or we should build the model with the limit on `max_features` turned off.

4.2 How can UFC sell more PPVs?

4.2.1 Linear Regression

Finally, to answer how the UFC can sell more PPVs, we employed linear regression. In terms of model selection, we ran the model on a training dataset with 80% of the PPV data and considered the following models: all predictors, all predictors and log transformation on PPV, variables chosen via stepwise regression, and variables chosen via backward selection. We compared model quality with the remaining 20% of the PPV, and ultimately determined that the model with variables chosen via step wise regression was our best fit model, as it had the highest test R^2 and test MSE.

Using this model, some factors that had a positive impact on PPV buys are total wins of both players, hosting the fight in North America, having the main event be a women's event, and having players with strong stats (ex: high reversals, more control and knockdowns, etc.). Conversely, the number of draws and losing streak had a negative impact on PPV buys. However, the resulting model had an R^2 of 0.3857 and an



Model	Train R^2	Train MSE	Test R^2	Test MSE
All predictors	0.5142681	4.654322e+10	0.1946434	81114005895
All predictors, log(PPV)	0.5935205	2.867805e-01	-2.0774869	309958718162
Stepwise Regression Predictors	0.3730609	6.007381e+10	0.3179269	68697122485
Backward Selection Predictors	0.4795769	4.986736e+10	0.0114314	99566783576

Figure 9: Table comparing the different linear regression models considered for PPV sales.

adjusted R^2 of 0.3175 trained on the full dataset, which suggests that most of the variation is not described by the variables in our regression model. Some additional factors that aren't present in our dataset, such as the popularity/reputation of each fighter or the rivalry between the fighters that could have attributed to PPV buys.

5 Conclusions

In this project, we successfully applied machine learning techniques to predict fight outcomes and pay-per-view sales in the UFC. The process began with extensive data cleaning and feature selection on the historical UFC fight statistics dataset. We transformed categorical features into dummy variables, addressed the challenges of high dimensionality and autoregressive properties using techniques like Cumulative Sum (CUSUM) and feature selection methods such as p-value assessments. The dataset was further enriched with information about fighters' win/loss streaks and aggregated fight statistics, which provided valuable insights into their performance over time. Exploratory data analysis revealed that strike and control advantages aligned with our hypotheses, while win/loss streaks did not show a clear relationship with fight outcomes.

Building upon the cleaned and enriched dataset, we trained and evaluated several machine learning models. The XGBoost model demonstrated the best performance, achieving an impressive AUC score of 0.938 on the test set, surpassing the logistic regression (AUC of 0.873) and random forest (accuracy of 0.551) models. The best model from our literature review had a prediction accuracy of 0.864, which aligns with what was achieved on our logistic regression model. This shows how powerful the gradient boosting model is, surpassing the literature model by 0.074. The XGBoost model's success can be attributed to its ability to capture complex relationships and interactions among features, as well as its regularization techniques and parallel processing capabilities.

For predicting PPV sales, factors such as hosting the event in the US and having high caliber players with multiple wins and strong past fight statistics lead to higher PPV sales. However, the resulting linear model didn't have the best fit (R^2 of 0.3857; adjusted R^2 of 0.3175), suggesting that the relationship between fighter statistics and pay-per-view sales may be more complex and require further investigation. As mentioned previously, we believe having additional information such as social media data to determine popularity, whether there is a rivalry, and having the rest of the fights on the card will result in a better fitting model.

References

[1] Statista Research Department, *Ultimate fighting championship (ufc) - statistics & facts*, online, published on Statista, Jan. 2024. [Online]. Available: https://www.statista.com/topics/3376/ultimate-fighting-championship-ufc/#topicOverview.

[2] D. Tighe, *Most valuable sports business brands worldwide in 2019*, online, published on Statista, Feb. 2024. [Online]. Available: https://www.statista.com/statistics/253349/brand-value-of-sports-businesses-worldwide.

[3] Statista Research Department, *Ultimate fighting championship (ufc) events with the highest live attendance as of august 2022*, online, published on statista, Dec. 2022. [Online]. Available: https://www.statista.com/statistics/682422/most-attended-ultimate-fighting-championship-events.

[4] U. Bashir, *Interest in sport types in the u.s. as of december 2023*, online, published on statista, Feb. 2024. [Online]. Available: https://www.statista.com/forecasts/1388963/interest-in-sport-types-in-the-us.