## MGT 6203 Group Project Proposal

### **TEAM INFORMATION (1 point)**

#### Team #: 7

**Team Members:** name + gtid + [Insert background information: Name, professional background, education background, previous analytics related projects you have worked on]

- 1. Andrew Berkowitz [aberkowitz8]: Undergrad in business currently working as a senior data analyst.
- 2. Alejandro Forero Eng [aeng61]: Undergrad in Electrical Engineering, master's in electronic and computers, and MBA, currently working as Director of Technology.
- 3. Juliette Noelle Wong [jwong88]: Undergrad in math & statistics, past projects include analyzing the effect of age on defensive ability at different positions for a Major League Baseball team. Currently working as a software quality engineer/data analyst.
- 4. Nicholas P Wilson [nwilson86]: Undergrad in Economics, currently working as a forecasting analyst in the energy industry.
- 5. Samuel Wade Wang [swang3068]: Was a graduate student majoring in pure math. During years in OMSA, worked on a <u>flu spread simulation project</u> in Simulation course; currently also working on course projects from DVA, CDA, DL this semester.

### **OBJECTIVE/PROBLEM (5 points)**

**Project Title: UFC Fight Prediction** 

#### **Background Information on chosen project topic:**

The Ultimate Fighting Championship (UFC) is the most well-known organization in the growing sport of mixed martial arts. As the sport and organization grow in popularity, there will be more people betting on fight outcomes, which increases the amount of people watching the events. We would like to identify key features that predict the winner of the fight. Additionally, we would like to explore if there are fighters and fight history that affect how many PPVs (Pay Per View) an event sells and create a model to predict the number.

# Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

What are the major factors in determining which fighter wins, and how can we use them to predict the results of future bouts?

#### State your Primary Research Question (RQ):

What factors are the strongest determinants of which opponent wins a UFC match?

#### Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

- 1. Factor analysis on UFC fight data for PPV (Pay Per View Buys).
- 2. Given a fighter's past fight statistics, can we predict who will win the match?
- 3. What would be the best strategy to maximize the odds of winning a fight?

Business Justification: (Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)

- 1. The UFC could benefit from a model showing largely objective measures determining fight outcomes. This affects fighter income, betting odds, and perceived fairness of the brand/league.
- 2. Predictive portions would allow UFC and other entities to better understand what drives PPV sales.
- 3. Sports analytics in general are also influential in sports betting industries.

### **DATASET/PLAN FOR DATA (4 points)**

#### Data Sources (links, attachments, etc.):

- Primary Datasource: UFC fight data from 1993 to 2023 https://www.kaggle.com/datasets/akshaysinghim/ufc-fight-data-1993-to-2023?resource=download
- Secondary Datasource: UFC PPV Sales
  https://www.kaggle.com/datasets/daverosenman/ufc-ppv-sales

# Data Description (describe each of your data sources, include screenshots of a few rows of data):

Our primary data source contains fight data of all the UFC fights between March 1994 and August of 2023 (until UFC 292). There are 7272 rows of data and 82 columns. Each row is a match containing statistics for each fighter regarding striking, clinching, and grappling.

A B	C	0	E		6	н			. K		M		N	O P	Q			5	T U	V	- 10	×		Z	AA	AB	AC	AD	AL	H	AG	AH	Al	Al	AX	AL.	AM
1 R_fighter B_fighter R_KI	00	1,10	R_SIG_STR.	B_SIG_STR	R_\$10_\$TR_8	SIG_STR_	R_TOTAL_S	TB_TOTAL	ST R_TD	8_70	R_TD_s	et 8_TD	pet I	LSUB_ATT B_SUB_	ATT R_REV	B_REV	R_CT	IL B_C	TRL R_HEAD	B_HEAD	R_BODY	B_BODY	R_LEG	B_LEG	R_DISTAN	DE B_DISTANO	E R_CUNOH	B_CLINOI	R_GROUN	D B_GROUN	D win by	last_round	last_round_t1	Format	Referee	date	location
2 Aljamain Ste Sean O'Malli	0		1 17 of 35	25 of 35	48%	71%	24 of 46	26 of 36	0 of 2	O of O		0%		0	0	0	0	0:34	0:18 3 of 12	18 of 25	3 of 10	6-of 9	11 of 13	1 of 1	26 of 33	10 of 16	1 of 2	1 of 1	D of O	14 of 18	KO/TKO		0.51	5 Rnd (5-5-5	Marc Goddar	19-Aug-7	23 Boston, Ma
Chris Weidm Brad Tayorer	0		0 37 of 92	70 of 115	40%	60%	39 of 94	71 of 116	0 of 8	O of O		0%		0	0	0	0	1:21	0:00 30 of 77	21 of 61	5 of 12	8 of 9	2 of 3	41 of 45	35 of 90	68 of 113	2 of 2	2 of 2	0 of 0	0 of 0	Decision - U	A 7	5:00	3 Rnd (5-5-5	Keith Peterso	19-Aug-7	23 Boston, Ma
Neil Magny Ian Garry	0		0.27 of 66	91 of 150	40%	60%	51 of 93	111 of 17:	L D of 2	2 of 2		0%	100%	0	0	0	0	1:50	2:32 10 of 41	32 of 90	7 of 11	16 of 17	30 of 14	43 of 43	25 of 61	84 of 140	2 of 5	4 of 5	D of O	3 of 5	Decision - U	4 7	5:00	3 Rnd (5-5-1	Keith Peterso	19-Aug-7	23 Boston, M.
5 Marion Vera Pedro Munho	0		0 141 of 251	113 of 246	56%	45%	141 of 251	113 of 24	0 of 1	0 of 1		0%	9%	0	0	0	0	0:00	0:00 109 of 21	2 53 of 177	9 of 14	20 of 28	23 of 25	40 of 41	138 of 246	111 of 244	3 of 4	2 of 2	0 of 1	0 of 0	Decision - U	4 7	5:00	3 Rnd (5-5-5	Herb Dean	19-Aug-2	23 Boston, Ma
6 Zhang Welli Amanda Len	1		0 163 of 217	24 of 63	75%	38N	296 of 358	29 of 68	6 of 7	O of O		85%		0	2	0	0	16:07	0:34 106 of 15	29 of 57	37 of 42	3 of 4	20 of 25	2 of 2	50 of 77	23 of 61	17 of 18	1 of 2	96 of 122	0 of 0	Decision - U	i 1	5:00	5 Rnd (5-5-5	Kevin MacDo	19-Aug-7	3 Boston, Ma
7 Da'Mon Blac Mario Bautis	0		0.90 of 174	79 of 137	51%	57%	106 of 192	102 of 16	4 of 7	2 of 33		57%	22%	0	1	0	0	3:00	4:04 48 of 123	48 of 223	22 of 30	21 of 22	20 of 21	10 of 12	52 of 129	51 of 101	37 of 42	24 of 30	1 of 3	4 of 6	Decision - U	4 7	5:00	3 Rnd (5-5-5	Bryan Miner	19-Aug-2	23 Boston, Ma
8 Gregory Rod Denis Tiuliuli	0		0.5 of 6	2 of 7	83%	28%	6 of 8	3 of 8	1 of 2	O of O		50%		0	0	0	0	0:47	0:00 4 of 5	2 of 7	1 of 1	0-of 0	0 of 0	0 of 0	1 of 2	2 of 7	0 of 0	0 of 0	4 of 4	0 of 0	KO/TKO		1:43	3 Rnd (5-5-5	Bryan Miner	19-Aug-2	23 Boston, Ma
9 Kurt Holobau Austin Hubbi	0		0 61 of 119	44 of 327	51%	42%	85 of 144	61 of 125	0 of 0	2 of 6			33%	2	0	0	0	1:08	2:23 46 of 100	25 of 86	10 of 12	17 of 19	5 of 6	2 of 2	50 of 106	31 of 87	9 of 10	9 of 12	2 of 3	4 of 8	Submission		239	3 Rnd + OT I	( John English	19-Aug-7	23 Boston, Ma
O Andre Petros Gerald Meer	1		0.57 of 132	75 of 149	43%	51%	67 of 143	92 of 166	2 of 5	1 of 3		40%	33%	0	0	0	0	2:06	1:54 39 of 114	65 of 135	12 of 12	6-of 9	6 of 6	5 of 5	46 of 117	60 of 131	3 of 4	10 of 10	8 of 11	6 of 8	Decision - Sp	s 2	5:00	3 Rnd (5-5-5	Herb Dean	19-Aug-2	23 Boston, Ma
1 Karine Silva Maryna Mon	1		0.30 of 53	16 of 55	56%	29%	52 of 76	25 of 66	1 of 1	O of O		100%		1	0	0	1	0:51	1:41 15 of 31	8 of 41	7 of 9	7 of 12	8 of 13	1 of 2	18 of 40	10 of 47	11 of 12	6 of 8	1 of 1	0 of 0	Submission		4:59	3 Rnd (5-5-5	S John English	19-Aug-7	23 Boston, Ma
2 Andrea Lee Natalia Silva	0		0.43 of 125	70 of 174	34%	40%	43 of 125	70 of 174	0 of 0	0 of 0		-		0	0	0	0	0:00	0:00 12 of 74	34 of 111	11 of 23	17 of 43	20 of 28	19 of 20	43 of 125	70 of 174	0 of 0	0 of 0	0 of 0	0 of 0	Decision - U	i 1	5:00	3 Rnd (5-5-5	Kevin MacDo	19-Aug-7	23 Boston, Ma
3 Brad Katona Cody Gibson	0		0 160 of 311		51%	61%	173 of 327	169 of 27	0 of 2	0 of 1		0%	0%	0	0	0	0	0:00	0:00 116 of 25	4 130 of 22	8 25 of 37	22 of 27	19 of 20	12 of 13	127 of 262	134 of 233	33 of 49	30 of 35	0 of 0	0 ef 0	Decision - U	( )	5:00	3 Rnd + OT (	( Marc Goddar	19-Aug-2	3 Boston, Ma
4 Khalil Rountr Chris Daukau	1		0 16 of 24	8 of 21	66%	38%	16 of 24	14 of 28	0 of 0	0 of 1			0%	0	0	0	0	0:31	0:04 12 of 19	6-of 19	3 of 4	2 of 2	1 of 1	O of O	11 of 17	7 of 19	3 of 4	1 of 2	2 of 3	0 of 0	KO/TKO		2:40	3 Rnd (5-5-5	6 Herb Dean	12-Aug-7	23 Las Vegas,
5 Terrance Mc Mike Breede	0		0 30 of 47	1 of 5	63%	20%	30 of 47	1 of 5	0 of 2	0 of 1		0%	0%	0	0	0	0	0:03	0:00 23 of 36	0 of 4	7 of 11	1 of 1	0 of 0	0 of 0	16 of 30	1 of 5	11 of 14	0 of 0	3 of 3	0 ef 0	KO/TKO	7	1:25	3 Rnd (5-5-5	Chris Tognon	12-Aug-2	3 Las Vegas,
6 AJ Dobson Tafon Nichuk	0		0.67 of 134	53 of 99	50%	53%	98 of 176	59 of 106	2 of 3	O of O		66%		0	0	0	0	4:33	0:00 19 of 75	11 of 47	12 of 18	12 of 15	36 of 41	30 of 37	61 of 119	49 of 92	2 of 5	4 of 7	4 of 10	0 of 0	Decision - U	4 2	5:00	3 Rnd (5-5-5	6 Herb Dean	12-Aug-2	z3 Las Vegas,
7 Cub Swansor Hakeem Daw	0		0.77 of 156	95 of 176	49%	53%	107 of 195	129 of 22	1 of 2	O of O		50%		0	0	0	1	1:25	3:10 41 of 112	41 of 104	27 of 31	29 of 37	9 of 13	25 of 35	59 of 135	79 of 158	18 of 21	16 of 18	D of O	0 of 0	Decision - U	i 7	5:00	3 Rnd (5-5-5	Jason Herzog	12-Aug-7	23 Las Vegas,
3 JP Buys Marcus McG	0		1 4 of 30	7 of 13	13%	53%	4 of 30	7 of 13	0 of 1	O of O		0%		0	0	0	0	0.00	0:05 2 of 22	6-of 11	1 of 3	Outo	1 of 5	1 of 2	4 of 29	7 of 13	0 of 1	O of O	D of O	O of O	KO/TKO		2:19	3 Rnd (5-5-5	Jason Herzog	12-Aug-2	23 Las Vegas,
9 Vicente Lugu Rafael Dos A	0		0.72 of 141	66 of 136	51%	48%	138 of 221	94 of 170	8 of 11	2 of 7		72%	28%	0	1	0	0	12:01	3:54 48 of 115	42 of 108	17 of 19	19 of 23	7 of 7	5 of 5	60 of 127	38 of 105	B of B	26 of 27	4 of 6	2 of 4	Decision - U	4 7	5:00	5 Rnd (5-5-5	Mark Smith	12-Aug-2	23 Las Vegas,
10 Josh Fremd Jamie Picket	0		O 31 of 80	14 of 52	38%	26%	59 of 109	26 of 67	C of 3	2 of 3		0%	66%	2	0	1	0	7:15	3:01 11 of 49	4 of 31	10 of 18	6 of 16	20 of 13	4 of 5	22 of 64	11 of 45	6 of 11	3 of 7	3 of 5	Q of Q	Decision - U	d 7	5:00	3 Red (5-5-5	Mark Smith	12-Aug-2	23 Las Vegas, "

Our secondary data source contains pay-per-view (PPV) data for specific matches. There are 177 rows and 7 columns of data.

	А	В	С	D	Е	F	G
1	Year	Month	Day	UFC_Num	Opponent	Opponent	PPV
2	2001	9	28	33	Ortiz	Matyushe	75000
3	2001	11	2	34	Couture	Rizzo	65000
4	2002	1	11	35	Pulver	Penn	35000
5	2002	3	22	36	Couture	Barnett	55000
6	2002	5	10	37	Lindland	Bustaman	50000
7	2002	7	13	38	Hughes	Newton	45000
8	2002	9	27	39	Couture	Rodriguez	45000
9	2002	11	22	40	Ortiz	Shamrock	100000
10	2003	2	28	41	Rodriguez	Sylvia	60000
11	2003	4	25	42	Hughes	Sherk	35000
12	2003	6	6	43	Couture	Liddell	49000
13	2003	9	26	44	Couture	Ortiz	94000
14	2003	11	21	45	Hughes	Trigg	40000
15	2004	1	31	46	Couture	Belfort	80000
16	2004	4	2	47	Ortiz	Liddell	105000
17	2004	6	19	48	Shamrock	Kimo	110000
18	2004	8	21	49	Couture	Belfort	80000
19	2004	10	22	50	Ortiz	Cote	40000
20	2005	2	5	51	Ortiz	Belfort	105000

Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables? What variables do you hypothesize beforehand to be most important?)

- Dependent
  - o Winner
  - o PPV
- Independent
  - Knockdowns
  - Significant strikes
  - Total strikes
  - Takedowns
  - Submission attempts
  - o Control time
  - Type of strike (head, body, leg)
  - Striking range (distance, clinch, ground)
  - Winning method
  - o Fight type (title bout, bout, etc.)
- Created Variables
  - o Record
  - Interaction Terms
  - Dummy Variables

For predicting the winner, given the fight goes to decision, we believe the most significant variables will be knockdowns, total strikes, takedowns, and control time. For predicting PPV sales, we believe that a fighter's previous fight outcomes will be most important. If they have been ending fights or on a win streak, they may be a bigger draw.

### **APPROACH/METHODOLOGY (8 points)**

Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))

We will approach the problem in a traditional and methodical manner.

**Integration of dataset:** Starting with a merge of both datasets by date, ensuring that each fight's statistics are paired with its corresponding PPV sales figure. This will enable us to answer our additional PPV research question.

Cleaning and transformation: Data cleaning steps will include splitting some string-type predictors to quantitative values (e.g. R\_SIG\_STR., R\_TD), removing collinear columns (e.g. R\_SIG\_STR\_pct, R\_TD\_pct), decide how to handle missing values (e.g. case of time 0:00 for both fighters). Transformation will involve normalizing numerical variables (e.g., strikes, takedowns) to ensure they are on a similar scale, encoding categorical variables (e.g., winning method, fight type) using one-hot encoding, and aggregating fighter statistics to create comprehensive profiles.

**Features Engineering:** Considering the number of features in our datasets, and how they grow when we transform our data, we may need to apply some reduction dimensionality techniques such as PCA or recursive analysis elimination.

**Model Selection:** For predicting fight outcomes, algorithms like logistic regression, decision trees and random forests could be considered. For PPV sales predictions, linear regression models and random forests considering the continuous nature of the target variable. AIC could be applied to compare and select models.

**Training and Validation:** Models will be trained using a split of the dataset into training and validation sets. Cross-validation, specifically k-fold cross-validation, will be employed to assess model robustness and generalizability, ensuring that performance is consistent across different subsets of the data.

Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an

## analysis that gives the reader insight into the conclusion regarding your objective/problem statement

Prior to our analysis, we expect to be able to identify several features that are statistically significant in predicting the winner of a fight. Features that we think will have the most significance are knock downs, total strikes, control time, and take downs. We then expect to be able to predict the winner of fights with reasonable accuracy. For our secondary question of PPV sales, we expect fighters on a win streak, fighters with a record of knockouts or submissions, and whether the event is a milestone (UFC 100, etc.) are important in PPV sales.

## What business decisions will be impacted by the results of your analysis? What could be some benefits?

Through our analysis, we will better understand what measures are significant in determining fight outcomes. Given those determinants, teams would be able to augment their strategies. This can also help individual bettors have an advantage to win and give the sports betting industry a tool to help calculate betting odds. Uncovering what drives higher PPV sales will further improve the UFC's ability to generate revenue from events.

### PROJECT TIMELINE/PLANNING (2 points)

Project Timeline/Mention key dates you hope to achieve certain milestones by:

Some possible items for data analysis:

- 1. Complete data cleaning: 1 week
- 2. Exploratory data analysis: 3 weeks
- 3. Fitting models to the data, record findings: 3 weeks
- 4. Turning findings into report: 1 week

Some possible items for documents

- 1. Proposal (2/25)
- 2. Progress Report (3/17)
- 3. Final Report (4/21)

Appendix (any preliminary figures or charts that you would like to include):