Executive Summary

Members: Jacob Kepes, Ryan Moruzzi, Rudy Perkins, Deepisha Solanki, Sujoy Upadhyay, Mariah Warner, **Project Mentor**: Mohammad Noorandisoot

Research Question/Goal: Can we predict different fan activities, such as attending a game, by considering self-reported fandom and various demographic variables? Our goal is to identify and anticipate the populations engaging in these activities to develop more targeted marketing strategies.

Stakeholders: Streaming Services (Hulu, Roku, Amazon), Networks (Fox, ABC/ESPN), Gambling Companies (DraftKings, FanDuel), and Sports Organizations (NFL, NHL, NBA, etc.)

Data: The data for these analyses come from multiple original surveys used in the writing of *Fans Have More Friends*, a 2022 book by Ben Valenta and David Sikorjak. The surveys were administered by Prodege, an online marketing, consumer polling, and market research company, between February 2021-January 2022. Following data cleaning, the data had N=10,362 respondents/surveys with over 100 variables.

Preprocessing: Our data set includes survey participants demographic information such as Age [S2], Household Income [D4], Gender (Categorical) [S1], Employment Status (Categorical) [D5] , Educational Attainment (Categorical) [D6] , Race/Ethnicity (Categorical) [Hid_Ethnicity_Bucket]. From a self-reported fandom question, [S15], we created a Fan magnitude variable, named fan_magnitude, which is the Euclidean norm of a participant's response. Similarly for a question regarding frequency of viewership [Team6_magnitude].

KPIs: We will measure the accuracy of predicting various fan activities [VL1] against a baseline model of a random coin toss.

Model Selection: Interested in making predictions for fan activities [VL1] questions from the survey, our exploratory data analysis (EDA) suggested that age [S2], income [D4], and Fan Magnitude were most important. Considering several classification algorithms, all models performed similarly well with respect to our outputs of interest, with the Adaboost algorithm performing with a slight advantage overall.

Results: Overall, our model was able to predict almost all outputs with about 64% accuracy or greater. Most interesting was “purchasing multi-game tickets” predicted with 92.76% accuracy and each feature of age, income, and fan magnitude being of almost equal importance. The model’s precision was at least 52% or above, with “purchasing multi-game tickets” at 67%. Our model’s recall was a bit all over, as low as 3% up to about 74%.

Future Directions: Moving forward, it would be interesting to incorporate more demographic information, like race and gender, to test whether our model’s accuracy remains consistent or when our model is restricted to just men or just women. Moreover, we would want to improve the recall of our model. Finally, with the rise of sports betting and time spent online, companies likely collect all sorts of information beyond basic demographics, and modeling approaches with these data could become very specific, which would be of interest to us.