Predicting the 2020 NFL MVP Winner
Name: Aiden Marcoux

The Associated Press MVP winner is the "most valuable player" in the NFL. It is the most prestigious award given to the best performing player during the NFL regular season, which consists of 16 games per team. The award is voted upon by a panel of 50 sportswriters, and although the accolade is awarded primarily by a player's season statistics, it is important to note that subjective factors such as the personal opinion of the writers influence the selection. My goal was to use machine learning to predict the MVP winner for the 2020 season, which is currently through 13 of 16 total weeks. My motivation for this analysis stems from the countless opinions observed online about who the winner should be this year. Often fans become very passionate about their opinion, leading to hostile arguments throughout the interweb. My findings can be valuable for three reasons: to offer statistically backed evidence in NFL MVP debates, to evaluate the choices of the 50 sportswriters who vote on the award, and to offer an advantage in sports betting for the selection of MVP. Here was my process:

I first gathered publicly accessible data of the MVP winners by year since the award originated in 1957. This data can be found from https://www.pro-football-reference.com/awards/ap-nfl-mvp-award.htm. I then collected quarterback data from each year since 1957. The reason for just choosing quarterback data and neglecting other positions will be explained later in the report. This is the passing data I gathered from 2019: https://www.pro-football-reference.com/years/2019/passing.htm. I concatenated each year's data into one large data frame. I added a column to this data frame named "MVP" which references the MVP winners data, and is the value "1" if the quarterback won MVP that year, and "0" if they did not. This is the data I used to train my model which will be applied to the 2020 season.

Looking at **Figure 1**, we can observe the distributions of the MVP awards allocated by a player's position. This bar chart is grouped by the "pre-modern era" which consists of the years 2000 and prior, and the "modern era" consisting of years after 2000. We can see that during the pre-modern era many more positions were considered for MVP, with quarterback (QB) consisting of 60% of MVP awards. The modern era shows much less variance in position, considering only quarterbacks and running backs (RB) for the award, with QB's winning 85% of the time. The increase in likelihood that a QB will win MVP in the modern era is likely attributed to the 65.8% increase in pass attempts between the two eras, showing a greater dependence on the quarterback position to win games in this era. These findings demonstrate my rationale to only considering quarterback data in my model, as quarterbacks are the overwhelming favorite to win. Because of this, I only used QB passer data in the modern era to train my model. Current MVP rankings for the 2020 season consist of nearly all QB's, further confirming my decision.

After deciding to strictly analyze passer data in my model, I chose to only consider data from quarterbacks who started in at least 12 of 16 games during the season. I used multiple linear regression to determine a player's "MVP score", which can be classified as the appropriateness for that player to win MVP based on the statistics of past MVP winners. My explained variance score for this model was 0.6572, suggesting my model was fairly appropriate. Polynomial Features and StandardScaler transformations were not appropriate, as they produced lower explained variance scores of 0.6278 and 0.4323, respectively. I then applied this model to the current quarterback stats in 2020, multiplying the appropriate columns 1.23x, which estimates their stats at the conclusion of the regular season. **Figure 2a** shows the resulting MVP scores for the top 10 players, showing that Patrick Mahomes is projected to win MVP with a score of 0.201, followed extremely closely by Aaron Rodgers. **Figure 2b** shows the variables that were accounted for in the model and their regression coefficients, showing that the number of wins has the greatest positive weight on a quarterback's MVP score of 0.0098, while the number of interceptions has the greatest negative weight on a quarterback's MVP score of -0.0013. We also see that surprisingly, the amount of passing yards has little effect in the MVP race, with a coefficient of 0.00013. It was interesting to see that my rankings for the 2020 MVP largely correspond to the current MVP rankings the media shows.

After predicting this year's MVP winner, I sought to use unsupervised learning in the form of k-means to cluster past MVP performances since 2000. The k-means model accounts for team wins, which was previously shown to be the most influencing factor in MVP evaluations, and quarterback rating which is a common metric that is calculated from the other variables in our linear regression (Touchdowns, Interceptions, Yards, Completion %). **Figure 3** highlights two clusters based on these metrics, showing MVP winners with an average MVP performance in green and MVP winners with elite performances (above average) in red. The blue dots show the centroids the clusters were based off. It is apparent that our top two considerations for MVP this year have also won in the past, showing Aaron Rodgers and Patrick Mahomes both with elite MVP performances. These performances occurred in 2011 and 2014 for Rodgers and 2018 for Mahomes.

In conclusion, my model predicts Patrick Mahomes and Aaron Rodgers as the two main favorites to win MVP in 2020. We see since 2000, MVP's were determined most by the amount of wins their team had during the season, followed by the amount of touchdowns they threw. We also categorized past MVP performances into two clusters, "average" and "elite" MVP performance. Overall, one can use my findings as evidence in MVP debates and my findings can be utilized to make more informed sports betting decisions for the 2020 NFL MVP.
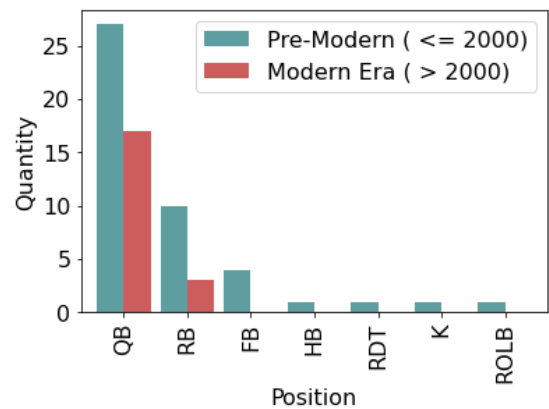
Figure 1: Number of MVP Awards by Position

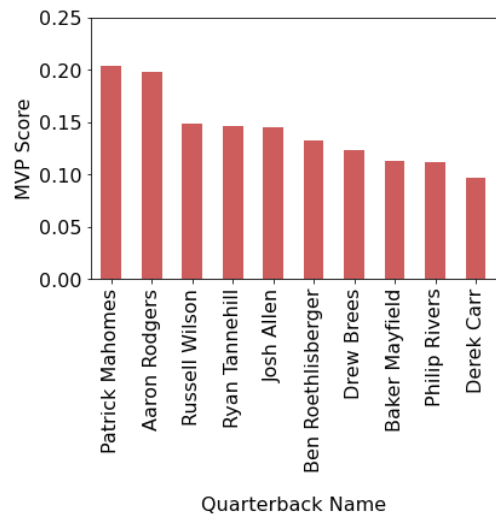Figure 2a: MVP Score Predictions for 2020 NFL Season
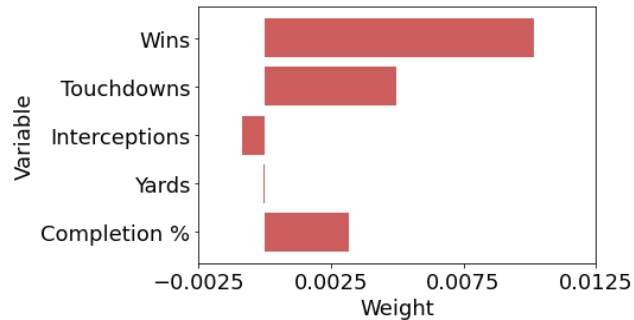

Figure 2b: Linear Regression Coefficients

Figure 3: Distinguishing Performances of QB MVP Winners Since 2000