Sean Steinle

STAT1341

Kenny Pickett, A Dark Horse for the Heisman?

Modeling the Race for the Heisman Trophy

**ABSTRACT:**

Despite its importance to the sport of college football, there have been few known attempts to describe or model the selection criteria of the Heisman trophy. In this paper I explore notable candidates from three different positions over the past seventeen years of voting. My main goals are first to determine if the selection for the Heisman trophy is driven by empirical factors like passing yards or year in school and further to determine if these empirical factors are within the control of a given player—is the trophy “fair”? I will further explore this question by using unsupervised and supervised machine learning techniques. If the Heisman is interpretable, the second goal of this paper is to approximate Kenny Pickett’s chances of winning the 2021 Heisman trophy. I find that neither 1st place votes nor total vote shares are even moderately correlated with basic player statistics and thus modeling techniques proved unfruitful. Current data sources for creating Heisman models may be lacking, and thus more creative models need be explored.

**INTRODUCTION:**

The Heisman Memorial trophy is awarded annually to the most outstanding player in college football. Notably, the Heisman trophy is not awarded to the most valuable player in football—in fact, many Heisman winners like Robert Griffin III and Johnny Manziel do not find success as professional football players. Additionally, like the college football and basketball playoffs, the trophy is awarded based on a selection of a committee. The committee is comprised of sports journalists, previous winners, and even a fan vote.

Due to the peculiar nature of the award, it is obvious to wonder whether the Heisman award is based on empirical factors like passing yards or whether it is a rather opaque selection process that is subject to similar bias and controversy that plagues the college football playoff selection process. To determine the answer to this question I attempt to model the Heisman award over the past 17 years, with the goal of creating a model that productively model players in the 2021 Heisman race like Kenny Pickett of the University of Pittsburgh.

**METHODS:**

The data I use to model the race comes from two main sources. First, I used data from the College Football Database (CFBD) that describes the season stats of a given player as well as how well that player’s team performed. I accessed this data through the Python API made available by CFBD. Second, I gathered Heisman vote data for the past 22 years from Sports Reference in the form of a .csv file for each year of voting. I then manually merged these files together to create one file containing votes for all years. It is worth noting that Sports Reference only tracks the 10 highest scoring candidates in the race in terms of vote-points—a score that is created by weighing 1st votes as 3 points, 2nd votes as 2 points, and 3rd votes as 1 point. However, typically the 10th place finisher has around 25 points which is less than 1% of total votes.

To process the data, I used Python and its assorted data science libraries including Pandas, Scikit Learn, and Matplotlib. All of my code was written in Jupyter Notebook format in order to optimize readability.

Although CFBD’s excellent Python API alleviated much of the data engineering involved in this project, there were still a few tasks I needed to accomplish. More menial tasks included creating .csv files from the Sports Reference tables, reformatting player names from Sports Reference, and merging CFBD and Sports Reference data. To complete the final task of compiling one source of vote, player, and team stats I iterated through the Sports Reference Heisman candidates and queried relevant statistics from the CFBD API. After collecting the statistics in a set of dictionaries, I created a set of Pandas data frames which I later coalesced into a quarterback dataset and a running back and receiver dataset. For quarterbacks, I included passing and rushing statistics; for running backs and receivers I included rushing and receiving statistics. I will further discuss why my research was limited to these three positions in subsequent sections. I did not include fumbles for any position due to data verification issues from CFBD, though this may be a relevant factor in evaluating running backs as candidates for the award. I limited my research to the span of 2004-2020 because CFBD stops tracking individual statistics before 2004. Within these limitations, my dataset is comprised of 89 quarterbacks, 53 running backs, and 12 receivers.

**RESULTS:**

*Exploratory Data Analysis:*

In our exploratory phase, we summarized two possible response variables synchronically and diachronically, five extrapositional predictors, and various position-specific predictors.

Share of 1st Place Votes and Total Points:

Chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generatedChart, histogram

Description automatically generated

In modeling the Heisman race, there are two obvious candidates for a response variable—the share of the 1st place votes or the share of total points. The share of 1st place votes is a better response if one is trying to predict high-end players that could win the award outright, whereas share of total points is a response that is more representative of the entire Heisman field of candidates. The slight difference between these responses can be seen in the scatterplot above. For our purposes, share of 1st place votes caters to our interest more effectively.

Even with basic graphical exploration we find that both responses are not only not normally distributed, but they are very much right-skewed. This is the case because most often a few players get large amounts of votes, and many more players get just a few votes.

Chart, line chart

Description automatically generated

We also explored how the Heisman vote counts changed over time. We found that our representation of the Heisman vote varied by a relatively small amount over the past 17 years. Still, we found it appropriate to control this variance by using vote shares instead of raw vote counts in the rest of our data analysis and modeling.

Year in School:

Another important factor worth considering is the year of the candidates. We found that juniors won a plurality of the 1st place vote for quarterbacks (41.1%) and a majority for running backs (60.0%). The vote count for receivers tends to be quite skewed due to an outlier—Devonta Smith’s 2020 Heisman campaign. He is the only receiver to have won the Heisman in this dataset so many receiver statistics are heavily impacted by his 2020 season.

Team Success:

Chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Team success is another factor worth considering. Although a player’s winning percentage is only weakly correlated with 1st vote share (r = 0.16 for quarterbacks, r = 0.19 for running backs), graphical representations of this statistic give another impression: winning percentage may act as a barrier to winning the Heisman. For example, there are very few losing records in the entire dataset and few players with winning percentages below 60% received serious consideration for the Heisman.

School Factor:

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart

Description automatically generated

One popular theory that explains Heisman voting patterns is that the selection committee is biased towards certain regions, conferences, or schools. Counter to my original hypothesis, I actually found the opposite. Although relatively few schools make up a large share of the Heisman vote, this is likely due to the structure of vote and point allocation rather than a school bias. More rigorous study is required to confirm or deny this finding.

Position Factor:

Finally, one of the most important extrapositional factors to consider is the role that position plays in a Heisman candidacy. We found that the award is dominated by quarterbacks (66.6% of total vote), but that running backs (24.5%) are also routine contenders for the award. Receivers make up almost 5% of total votes, but prior to 2020 it was likely an insignificant amount. The remaining positions make up less than 5% of the vote and have not produced a winner in our dataset.

Position Specific Factors:

Much like the extrapositional factors we examined, position-specific statistics are also lacking in correlation to vote share. There are a few quarterback statistics that are weakly to moderately correlated with 1st vote share (yards per attempt, r = 0.33; passing touchdowns, r = 0.27, completion percentage, r = 0.19), but the remaining statistics have a very weak correlation. Problematically, interceptions are very weakly negatively correlated (r = -0.09) with 1st vote share, despite interceptions being a cornerstone of quarterback evaluation. Additionally, we see no highly correlated rushing statistics, bucking the conventional wisdom that dual-threat quarterbacks can be both highly successful and highly talked about (ex: Lamar Jackson, Kyler Murray, Baker Mayfield).

Running backs have no statistics that are even moderately correlated with the response. What is even more troubling for the prospect of a Heisman model is that the “Year” statistic is more highly correlated with the response than any rushing or receiving statistic for running backs—this essentially means that the trend of running backs decreasing importance is more significant to vote totals than any other statistic.

The receiving category is once again heavily skewed by the presence of an outlier. Thus, in our models we combine running backs and receivers into one model.

*Machine Learning:*

K-Means Clustering:

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

In our only unsupervised learning model, we used k-means clustering (k = 3) to attempt to discover any natural tendencies or prototypes in the data. In our analysis of running backs and receivers, all of our receivers were clustered together and the running backs were clustered without much discrimination in vote share. On the other hand, the average 1st vote share difference for quarterback clusters was noticeable (1st cluster had 23.8%, 2nd has 13.3%, 3rd had 9.2%). The k-value was selected in order to optimize this average vote share difference, although in future research the elbow method may be worth investigating.

Linear Regression:

Chart, scatter chart

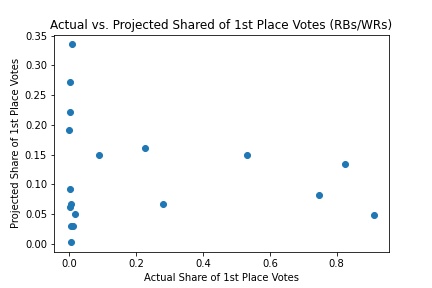
Description automatically generatedChart, scatter chart

Description automatically generated

In our supervised learning efforts, things were even less fruitful. Likely due to a small, non-normal sample, we actually saw a negative R-squared value of -0.04 for the quarterback model and -0.19 for the running back and receiver model. This value being negative tells us that even a fixed estimate for all players would perform better than our models did. Despite troubling results, the fact that the quarterback model outperformed the other skill positions may be an indication that with a larger sample the Heisman could be modeled effectively.

Neural Network:

Chart, scatter chart

Description automatically generated

Our neural network model performed even worse than the linear regression models, with an R-squared value of -0.23 and -1.48 for the quarterback and alternative skill position models respectively. Troublingly, the running back and receiver model’s R-squared value was outside of the [-1,1] bounds of the statistic, which is likely due to an issue in transforming the response before modeling. Even in spite of this error, neural networks are likely not appropriate for this context due to the lack of data available and due to the lack of transparency in feature representation that the model provides.

**DISCUSSION:**

It is clear that from our set of features it is difficult to impossible to effectively model the Heisman award’s selection process. That said, there are a number of reasons this may be the case. First, there are several signs that there is simply insufficient data to model this problem—the receiver position is entirely skewed by one player, running back data is entirely confounded by their decreasing relevance to the sport, and our results from quarterback data are much more expected despite only have 36 more samples than the running back data. Another possible explanation of why our models failed is that the Heisman selection process is able to be modeled, but that the domain and range of the problem are so diverse that any successful model requires magnitudes more data than is currently available. For example, looking through the 17 winners in our dataset reveals a smattering of wildly different prototypical players—rushing quarterbacks, pocket passers, dynamic long-ball receivers, one-cut backs, power backs, and more. The final interpretation of our result is that the Heisman simply does not have an empirical selection process. Though many may be tempted to endorse this statement, too little research has been done on the topic to know with certainty.

Despite the challenges of modeling this process, there are many future directions for research. First, reducing the feature space for any future model is a must. Much of the poor performance in our supervised learning models is likely due to the fact that the models have predictors that are more correlated with each other than they are with the response. However, a more creative direction for research may be in line with the conclusion that the Heisman is not empirical: attempts to model college football “narrative” may produce better results. Examples of this may include bowl game upsets, inclusion in the playoff, or even sentiment analysis and other natural language processing techniques on college football articles.