**Predicting the Heisman Trophy Winner**



Final Project - IST 707 Summer 2022

Vlad Nivorozhkin

Mike Wright

Nick Videtti

Nick Smith

Khalil Robinson

Kushal Shah

Vlad Nivorozhkin (vnivoroz@syr.edu)

Mike Wright (mwrigh16@syr.edu)

Nick Videtti (nvidetti@syr.edu)

Nick Smith (nsmith26@syr.edu)

Khalil Robinson (krobin08@syr.edu

Kushal Shah (kshah07@syr.edu)

Final Project

Decision Tree, Naive Bayes, Support Vector Machine, and Clustering Algorithms

September 2022

| Predicting The Heisman Trophy Winner |
| --- |

**Introduction**

The objective of our project is to use the skills taught in this class to solve a real data mining problem. Our group decided to see if we could predict the Heisman trophy winner, the most prestigious award in collegiate football. This is awarded to the most outstanding player in college football each season.

College football is part of the National Collegiate Athletic Association (NCAA) and is split into three Divisions. The Heisman trophy is awarded annually to a player in the Division One Football Bowl Subdivision (FBS), which is the premier college football Division. The Heisman trophy was created in 1935 and, since then, there have been 86 winners, consisting of over 30 quarterbacks, over 30 running backs, and even one defensive player (Charles Woodson).

The focus of our project will be on the quarterback (QB), who is the offensive leader and is in charge of creating plays and throwing the football. Well known NFL quarterbacks include Tom Brady, Peyton Manning, Patrick Mahomes and Joe Montana. Past quarterbacks who have won the Heisman trophy include Lamar Jackson and Cam Newton, and they have gone on to flourish in the NFL. The latest winner was Alabama quarterback Bryce Young, who won in 2021.

In our project, we will utilize the historic quarterback Heisman winner statistics to see if we can predict the 2021 Heisman trophy winner (i.e., quarterback Bryce Young) based on all Division One quarterback statistics for 2021. The statistics range from touchdowns to interceptions. Additionally, we will use a variety of data exploration and data mining techniques to see what interesting trends and patterns we may uncover in the data.

The Heisman winner statistics dataset comes from Sports-Reference.com and will be imported into R as an Excel file. It will be referred to in this paper as the Heisman Winner dataset. The 2021 Division One quarterback statistics dataset also comes from Sports-Reference.com and will be imported into R as an Excel file. It will be referred to in this paper as the 2021 Passing Stats dataset.

# 

# **Analysis**

**DATA PREPARATION AND CLEANING**

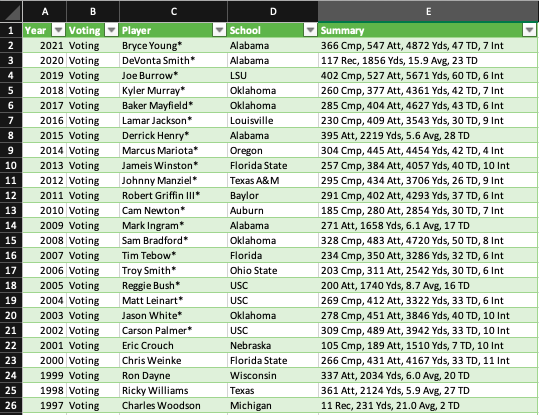
**General Data Processing**

The Heisman Winners and 2021 Passing Stats datasets required extensive preprocessing, but the fully processed data frame will combine both datasets and will consist of the following:

* Year: indicates the year the player statistics correspond to (e.g., 2020, 2021).
* Player: the name of the player (e.g., Joe Burrow, Bailey Zappe)
* School: the name of the school the player attended (e.g., Alabama, Pitt)
* Completions: the number of passing completions (e.g., 230, 317)
* Attempts: the number of passing attempts (e.g., 362, 441)
* Yards: the number of passing yards (e.g., 4435, 1912)
* Touchdowns: the number of passing touchdowns (e.g., 31, 60)
* Interceptions: the number of passing interceptions (e.g., 4, 13)
* HeismanWinner: binary variable set to 1 if the player has won the Heisman, and 0 if the player has not won the Heisman
* CompletionsDisc: discretized Completions column (e.g., (320,413])
* CompletionsDiscLevels: ranking of discretized CompletionsDisc column (e.g., “Very Low”, “High”)
* AttemptsDisc: discretized Attempts column (e.g., (449,568])
* AttemptsDiscLevels: ranking of discretized AttemptsDisc column (e.g., “Medium”, “Very High”)
* YardsDisc: discretized Yards column (e.g., (4875,5972])
* YardsDiscLevels: ranking of discretized YardsDisc column (e.g., “Low”, “Very High”)
* TouchdownsDisc: discretized Touchdowns column (e.g., (38.4,50.2])
* TouchdownsDiscLevels: ranking of discretized TouchdownsDisc column (e.g., “Very Low”, “High”)
* InterceptionsDisc: discretized Interceptions column (e.g., (1.97,7.2])
* InterceptionsDiscLevels: ranking of discretized InterceptionsDisc column (e.g., “Medium”, “High”)

**Heisman Winner Dataset**

1. *Heisman Winner Excel Preview*

**

As mentioned previously, the Heisman Winner data was pasted into and saved as an Excel file from the Sports-Reference website. From there, it was imported into R, with the following image providing a preview of the dataframe:

1. *Original Dataframe*



The data frame contains all Heisman Winners from all positions, such as quarterback and running back. Since our project focuses solely on quarterbacks, we found that the key word under the “Summary” column that only quarterbacks have is “Cmp” (indicating completions). As such, we filtered to just rows with this word, leaving us with the 35 QB rows we need.

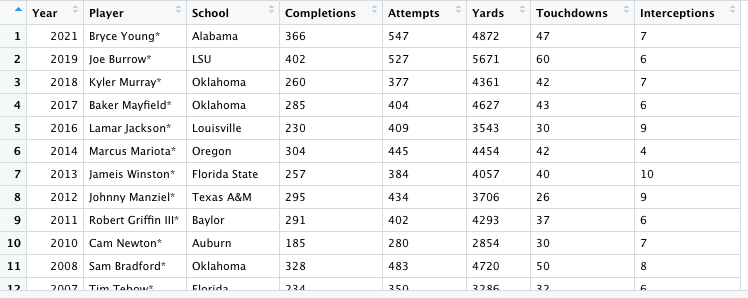
From there, we separate the “Summary” row based on the comma delimiter.

1. *“Summary” Row Separated*



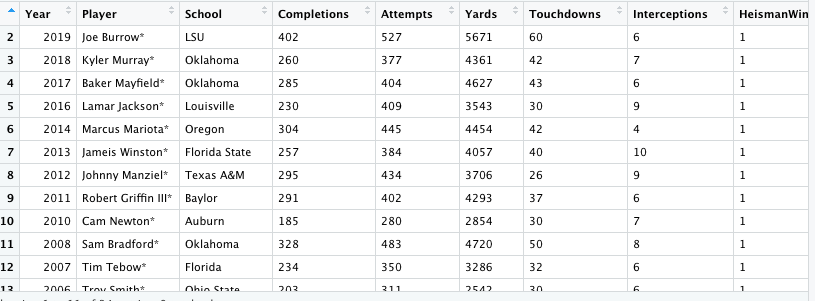
We then remove the space and alphabetical characters between columns “Completions” and “Interceptions”, such as “ Cmp” and “ Att”. This is because we want these columns to be integers for our models. We also remove the “Voting” column, as it does not provide any useful insight.

1. *“Completions” to “Interceptions” Processing and Removal of “Voting” Column*



Finally, we add a “HeismanWinner” column to indicate that all of these players have won the Heisman trophy (i.e., value of 1 for all players). We remove the “Bryce Young” from the dataset as he is the 2021 winner that we want to predict in our models. The following is the final Heisman Winner dataframe that will be combined with the 2021 Passing Stats dataframe:

1. *Final Heisman Winner Dataframe*



**2021 Passing Stats Dataframe**

1. *2021 Passing Stats Excel Preview*

**

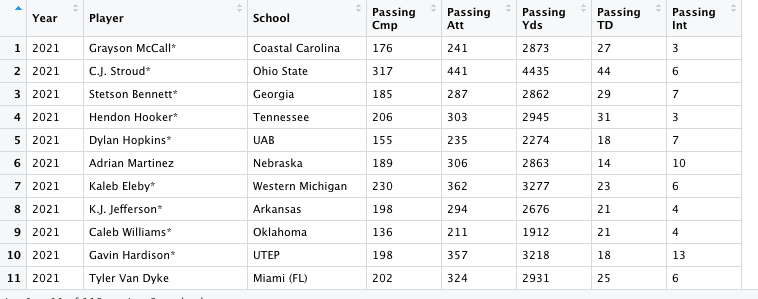
The 2021 Passing Stats data was also pasted into and saved as an Excel file from the Sports-Reference website. From there, it was imported into R, with the following image providing a preview of the dataframe:

1. *Original Dataframe*



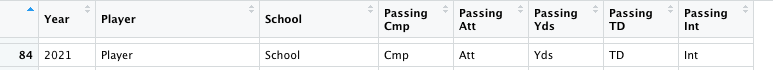
Since we want to combine this dataframe with the Heisman Winner dataframe, we want the columns and formatting to match. As such, to start off, we add a column for “Year” set to 2021 for all rows. We then filter to the columns that match with the data from the Heisman Winner dataframe, specifically “Year”, “Player”, “School”, “Passing Cmp”, “Passing Att”, “Passing Yds”, “Passing TD”, and “Passing Int”.

1. *Adding “Year” and Filter to Proper Columns*

**

We then filter out the needless header rows from the dataframe by filtering out the word “Player” under the column “Player”.

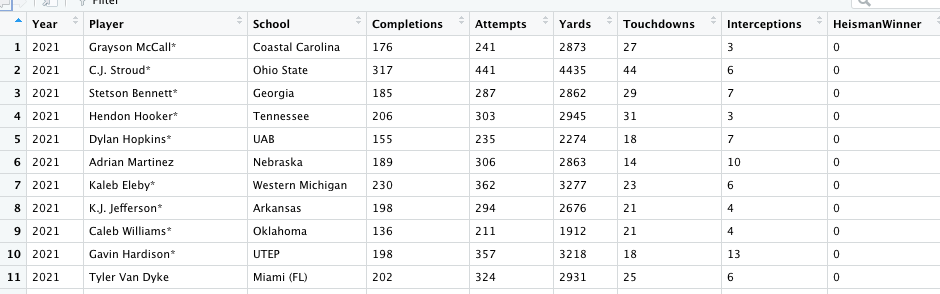
1. *Example Header Row*



From there, we add a “HeismanWinner” column to indicate that none of these players have won the Heisman trophy (i.e., value of 0 for all players). This dataframe includes Bryce Young, as we want our models to predict that Bryce Young was the winner in 2021.

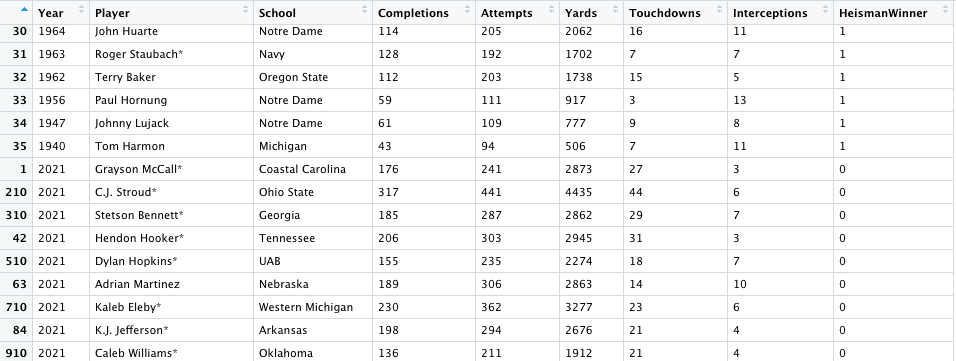
Finally, we rename our columns to match up with the names in the Heisman Winner dataframe (e.g., from “Passing Cmp” to “Completions”), and combine the final dataframe with the Heisman Winner dataframe.

1. *Final 2021 Passing Stats Dataframe*



The combined Heisman Winner and 2021 Passing Stats dataframe will be referred to simply as the Combined dataframe going forward.

1. *Combined Dataframe*



**Combined Dataframe**

The final step was to clean and process the Combined dataframe. To do so, we first remove the “\*” character under the “Player” column. We then remove leading and trailing whitespaces from all columns.

In terms of data types, we keep “Player” and “School” as character. However, we convert “Year” and the “Completions” to “Interceptions” columns to integer. We also convert “HeismanWinner” to factor.

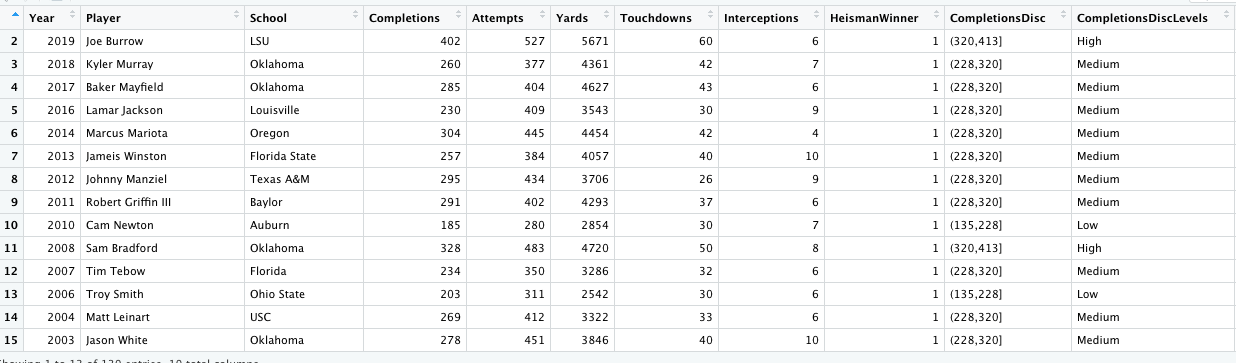
We then discretize the “Completions” to “Interceptions” columns. We found 5 evenly spread clusters to be optimal for usage in our models. The column names will be “CompletionsDisc”, “AttemptsDisc”, ”YardsDisc”, “TouchdownsDisc”, and “InterceptionsDisc" (i.e., taking the original column names and adding "Disc" to them). These columns will be factors.

We also added columns that indicate the ranking of each discretized cluster for easier reference throughout this paper (“Very Low”, “Low”, “Medium”, “High”, “Very High”). For instance, under “Completions”, the cluster for 320-413 completions is labeled as “High”, while the cluster for 228-320 completions is labeled as “Medium”. The labels themselves (e.g., “Very Low”) will not be used in our models. The column names will be “CompletionsDiscLevels”, “AttemptsDiscLevels”, ”YardsDiscLevels”, “TouchdownsDiscLevels”, and “InterceptionsDiscLevels” (i.e. taking the original column names and adding "DiscLevels" to them). These columns will be factors.

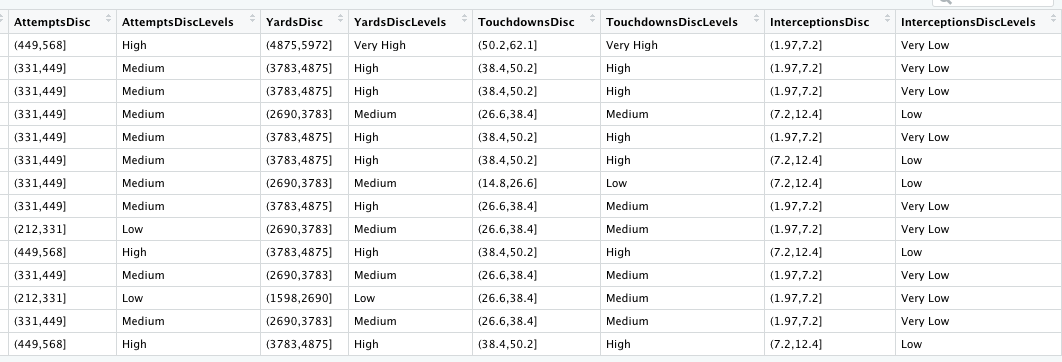
Finally, we overwrite our original two dataframes with the Combined dataframe, but filtered to “HeismanWinner=1” for the Heisman Winner dataframe and “HeismanWinner=0” for the 2021 Passing Stats dataframe.

The final Combined dataframe contains 19 columns and 139 rows, consisting of 34 rows from the Heisman Winner dataframe and 105 from the 2021 Passing Stats dataframe. A preview of the Combined dataframe and data types can be found below:

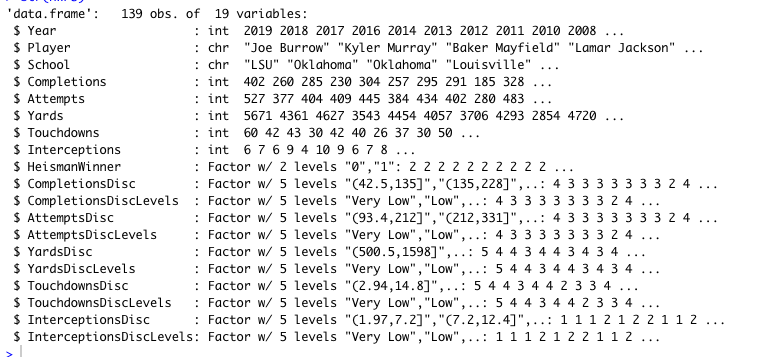
1. *Final Combined Dataframe First 11 Columns*



1. *Final Combined Dataframe Last 8 Columns*



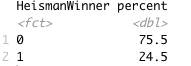
1. *Final Combined Dataframe Data Types*



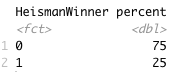
**Training and Test Data Creation**

Our Combined dataframe was split into a training dataset consisting of 92 rows, and a testing dataset consisting of 47 rows (~⅔ split). We used stratified sampling so that the training and testing sets would contain approximately the same proportions of Heisman winners/non-winners as the full Combined dataframe. This is to ensure our models are trained and tested on proportional data.

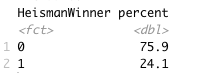
1. *Combined Dataframe Heisman Proportions*

****

1. *Training Dataframe Heisman Proportions*

****

1. *Testing Dataframe Heisman Proportions*

****

In the Combined dataframe, ~75% of rows are non-winners, and ~25% are winners. After using stratified sampling, the training and testing datasets contain ~75% of rows as non-winners, and ~25% as winners. As such, these proportional training and testing datasets will be used to train and test our models.

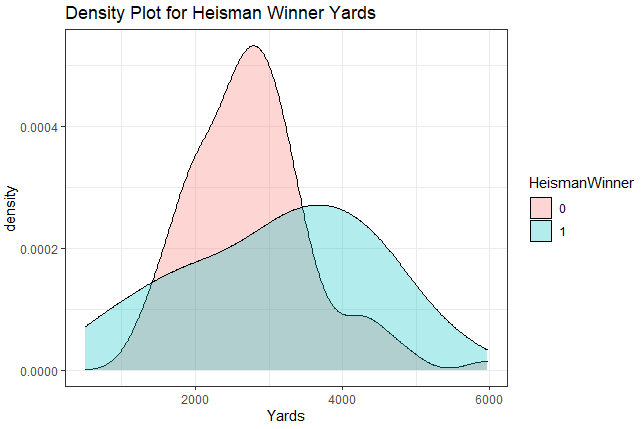
Finally, we remove “Year”, “Player”, and “School” from our training and testing datasets, as these columns would not make sense to use in our models. While we had originally considered using “School” in our models, we found that the models would automatically categorize players as non-winners based on their school. This does not make intuitive sense, as just because a player from Western Kentucky has never won the Heisman, does not mean that a player from that school will never win the Heisman. As such, it would not make sense for, say, our decision trees to automatically categorize players from this school as non-winners.

We also remove “Completions”, “CompletionsDiscLevels”, “Attempts”, “AttemptsDiscLevels”, “Yards”, “YardsDiscLevels”, “Touchdowns”, “TouchdownsDiscLevels”, “Interceptions”, “InterceptionsDiscLevels”. This is because we will use the discretized columns in our models.

**DATA EXPLORATION**

There were two major aspects that jumped out to us when exploring the data. The first was the talent distribution between quarterbacks who won the Heisman award vs those who did not. The plot below helps highlight this difference.

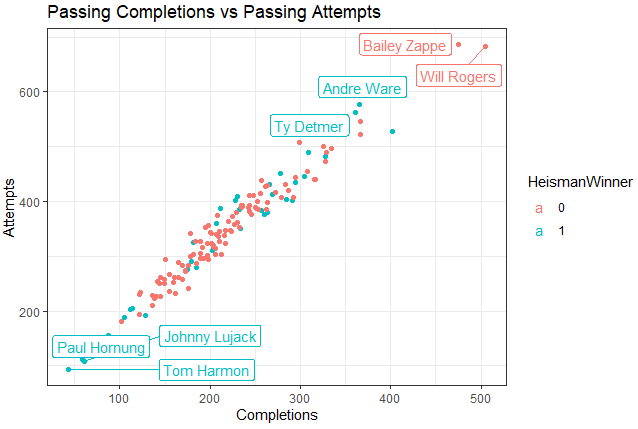
1. *Density Plot for Heisman Winner Yards*



The image above is a density plot of the passing yards thrown by quarterbacks who won the Heisman vs those who did not. There is a clear distinction between Heisman quarterbacks, whose peak was a little under 4000 yards, compared to the non-winners, whose peak was just under 3000 yards. This amounts to an almost 1000 yard difference. Put differently, 4000 yards in a college football season would imply over 300 additional yards per game.

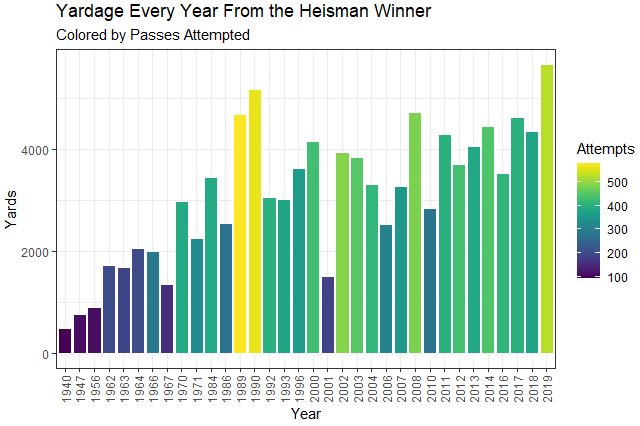
The second aspect that jumped out was the number of completions and attempts. The graph below showcases each quarterback’s completions and passing attempts. The coloring indicates Heisman winners (blue) vs non-winners (red).

1. *Passing Completions vs Passing Attempts*



Interestingly, this graph indicates no clear distinction between Heisman winners vs non-winners. However, there is one grouping on the bottom left which contains a large number of Heisman winners with low completions and attempts. These are not players who played recently, which leads us to our next story: the evolution of passing in college football. The graph below helps highlight the passing game changes in football.

1. *Yardage Every Year From the Heisman Winner*



Over time, college football teams have realized that passing more resulted in a far more effective offense. We can visualize this as the stark difference in passing yards and attempts over time in the graph above. For instance, there is a significant increase in passing yards after the 1980s. Moreover, the change in offensive philosophy is further highlighted by the color gradient. Specifically, the color becomes brighter with time, showcasing that teams are passing the ball more often in general.

**MODELS AND METHODS**

To conduct this analysis, a myriad of models were implemented, including Clustering, Decision Trees, Random Forests, Naive Bayes, and Support Vector Machines. We also used the training and testing datasets created above.

**ANALYSIS GOALS AND PARAMETERS**

The overarching goal of this analysis is to predict the 2021 winner. This is because the 2021 winner is already known to be Bryce Young, so testing for real-world accuracy when building the models will be possible by ensuring that Bryce Young is predicted. Ideally, these models are able to predict many years of Heisman Trophy winners correctly before they are used to predict future winners, but these models will focus on the 2021 winner.

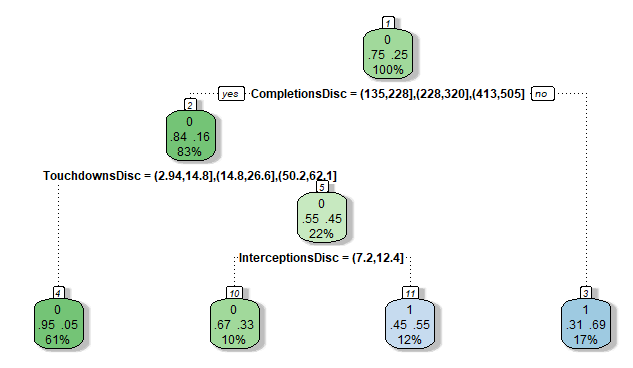
**Results**

**TECHNICAL RESULTS**

**Decision Tree**

In the Decision Tree model below, values of 0 indicate non-winners, while values of 1 indicate Heisman winners. This was built using the rpart function from R’s rpart package. This model predicts the Heisman winner variable as a function of all other variables in the data set, and the “method” parameter was set to “class”.

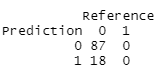
1. *Passing Completions vs Passing Attempts*



The discretized variables for pass completions, touchdown passes, and interceptions were the best indicators in this model for predicting the Heisman winner. In other words, this model can be used to take a quarterback’s completions, touchdowns, and interceptions to indicate if they will win the Heisman trophy.

We then tested the Decision Tree on the 2021 Passing Stats dataframe. Since this dataframe has all players labeled as non-winners, our goal was to simply see which players the Decision Tree would predict as winners. In other words, the Confusion Matrix would not be as insightful, just the player predictions.

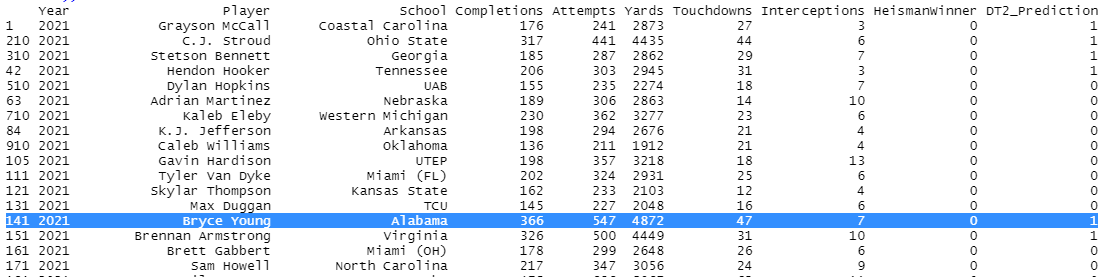
1. *Decision Tree Confusion Matrix*



All players are part of the 2021 Passing Stats dataframe and, as such, are all labeled as non-winners of a Heisman Trophy winner “candidate pool”. This means that no actual Heisman winners are in the testing data, thus there are only players under “0” for the “Reference” columns.

It was found that 18 of the 106 players in the testing data have the completion, touchdown, and interception statistics which make them more likely than not to win the Heisman Trophy. Bryce Young was indeed one of those 18 players, but ideally, this model would pick one player rather than simply indicate who is more likely to win than not. This could be done by finding which player is most likely to win, albeit ties are highly likely in a tree with this few nodes. Below are some results of using the Decision Tree model to predict Heisman Trophy winners. This prediction is found in the “DT2\_Prediction” column on the far right of the data frame. Notice Bryce Young’s record is highlighted.

1. *Decision Tree - Sample of Data Frame with Predictions*

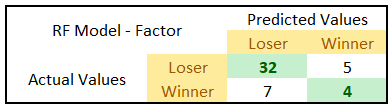
****

**Random Forest**

In addition to a decision tree, we also ran a random forest model to predict the Heisman winner. We utilized two random forest models, one predicting a binary 1 or 0 for whether the quarterback would be a winner and the other predicting a value between 0 and 1 representing the probability of the player winning the Heisman award. This was done by setting the response variable to factor and then to a numeric value. Each random forest utilized the same independent variables: “Attempts”, “Completions”, “Yards”, “Touchdowns”, and “Interceptions”.

When creating the random forest using a factor, we obtained the following confusion matrix:

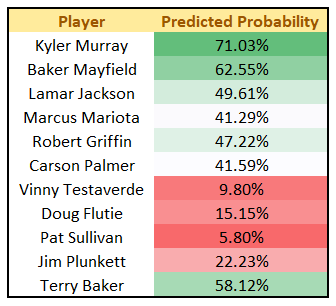
1. *Random Forest - Dependent Variable as a Factor*



The predictions on the testing set show decent results, with an accuracy of ~75% (4 Heisman winners predicted correctly). It inaccurately predicts 7 players to win the Heisman. However, this model also predicts Bryce Young to win the Heisman.

1. *Random Forest - Dependent Variable as a Numeric Value*

While a confusion matrix would not be the best way to accurately measure this random forest, we can look at some of the previous Heisman Winners and the probability the model gave them to win the Heisman.



Quite a few of the players are given over a 50% chance of winning, with Kyler Murray having the highest probability at 71.03%. There seems to skew towards the more recent players, with these players receiving a higher probability of winning the award. This coincides with what we saw earlier in terms of an increasing focus on the passing game. This model gives Bryce Young a 63.52% chance of winning the Heisman, which is quite high.

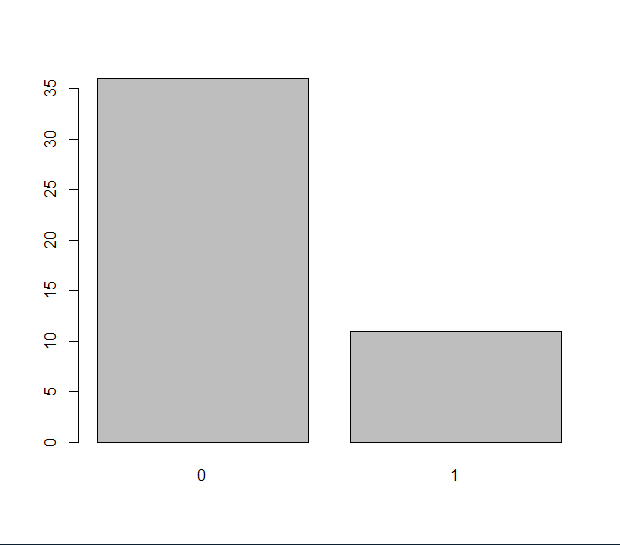
**Naive Bayes**

The Naive Bayes classifier is a machine learning probabilistic classifier derived from Bayes Theorem. The easiest way of describing a Naive Bayes analysis is to identify an outcome (A) of occurring given an instance (B) occurred. In this instance, identifying if a player (A) will likely win the Heisman given his stats (B). For this analysis, two separate packages were used in RStudio: e\_1071 and naiveBayes. The purpose of using various packages was to potentially yield different results.

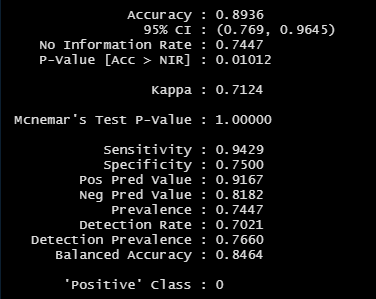
*26. e\_1071 Confusion Matrix*



*27. e\_1071 Barplot*



*28. e\_1071 Accuracy Results*

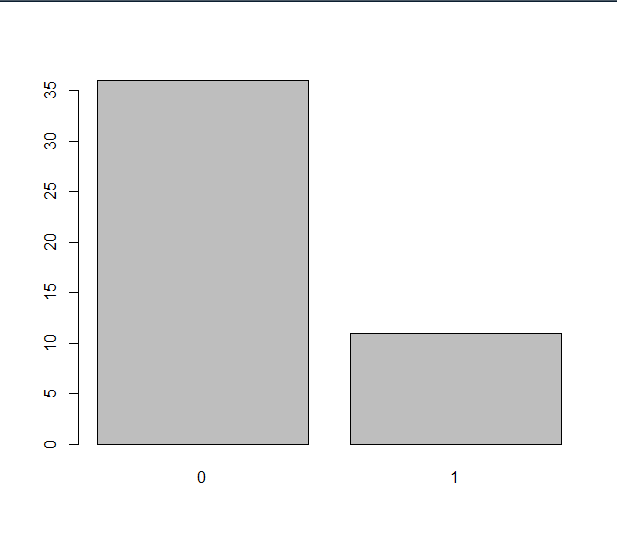
**

The first model was created using the e\_1071 package. When interpreting the results of the confusion matrix, it correctly identified 33 players as non-winners and miscategorized 3 as winners. Furthermore, the model correctly identified 9 players as winners and miscategorized 2 as non-winners. The e\_1071 model has ~89% accuracy when evaluating the model. Furthermore, the model has a .71 Kappa, which can be interpreted as having substantial inter-rater reliability.

*29. naiveBayes Confusion Matrix*



*30. naiveBayes Barplot*



The naiveBayes package model seemingly yielded similar results, identifying 36 non-winners and 11 winners. The winners the model predicted consist of: Joe Burrow, Sam Bradford, Eric Crouch, Andre Ware, Terry Baker, Tom Harmon, Brennan Armstrong, Taulia Tagovailoa, Tanner Mordecai, and Steven Krajewski. However, out of those players, the only non-winners are Brennan Armstrong, Taulia Tagovailoa, Tanner Mordecai, and Steven Krajewski. Based on these findings, it is likely that the e\_1071 model is slightly more accurate than the naiveBayes model.

We then applied the naiveBayes model to the entire Combined dataframe. We predicted 106 players as non-winners, and 33 players as winners.

*31. naiveBayes Confusion Matrix*



**Support Vector Machines**

For the purposes of this analysis, three separate models of the SVM were using the following kernels: polynomial, linear, and radial. The purpose of running the various kernels was to identify what model provided the highest accuracy rate when predicting the results of a Heisman winner.

The radial kernel is found to be the most accurate with ~93% accuracy. It used a cost parameter of 100. From there, the linear kernel had an ~85% accuracy, and the polynomial had an ~83% accuracy. When interpreting the confusion matrix, the model was able to correctly identify 67 players as non-winners and miscategorized 4 as winners. Furthermore, 19 players were correctly identified as winners, but 2 were miscategorized.

*32. SVM Confusion Matrix*



After interpreting the confusion matrix, each of the variables were measured for significance. The results similarly reflect those of our Decision Trees, in that there is more emphasis on completions and touchdowns.

The images below depict each of the variables by winner (1) and non-winner (0). Overall, what the images illustrate is that winners follow a similar formula consisting of: more completions with fewer attempts that produce more yards, and fewer interceptions. In some cases, non-winners may show similar results, but do not follow the formula fully. For instance, they may have more completions, but through more attempts with fewer yards. Or they may simply throw too many interceptions. To become a Heisman winner, the player needs to be optimal when it comes to completions, attempts, and interceptions.

*33. SVM Completions*

****

*34. SVM Attempts*

**

*35. SVM Yards*

**

*36. SVM Touchdowns*

**

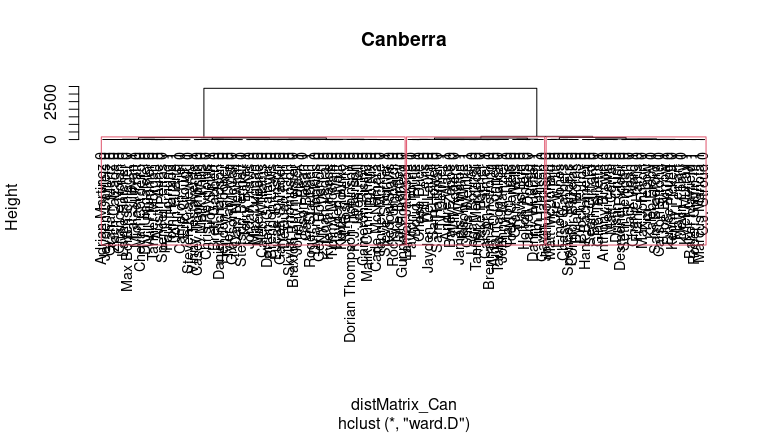
*37. SVM Interceptions*

**

**Clustering**

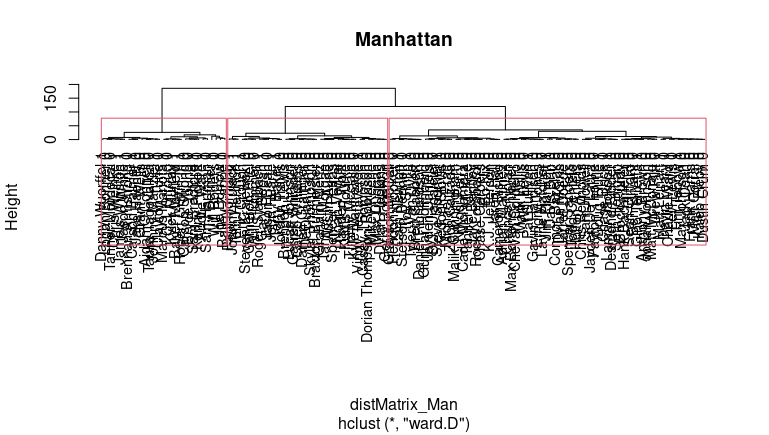
For clustering, hierarchical was chosen for this analysis. We tested different distance measures to see if they were able to properly cluster winners based on their stats. All non-numeric columns were removed from the dataset, which left completions, attempts, touchdowns, yards and interceptions. All rows were normalized using z-score to improve clustering results. The names of the players and Heisman winner status were changed to row names to identify the player in the dendrogram. These distance measures are being evaluated by their ability to cluster the majority of winners together.

*38. Canberra Cluster*



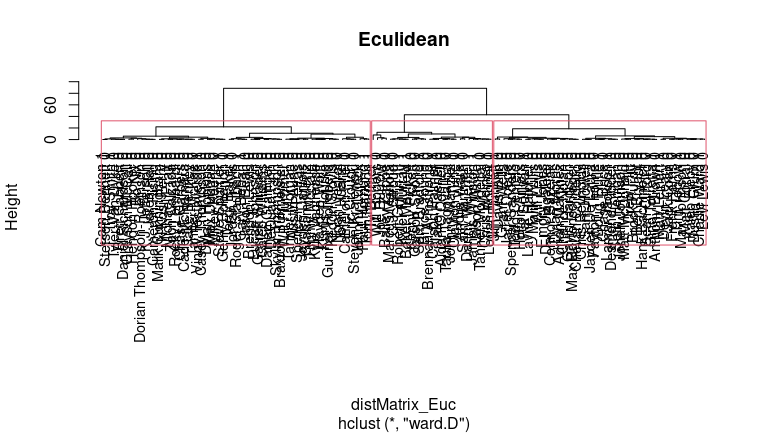
The above graph shows the dendogram for the canberra distance measure. After analyzing the results, hierarchical clustering with canberra distance does not properly cluster the data effectively.

*39. Manhattan Cluster*



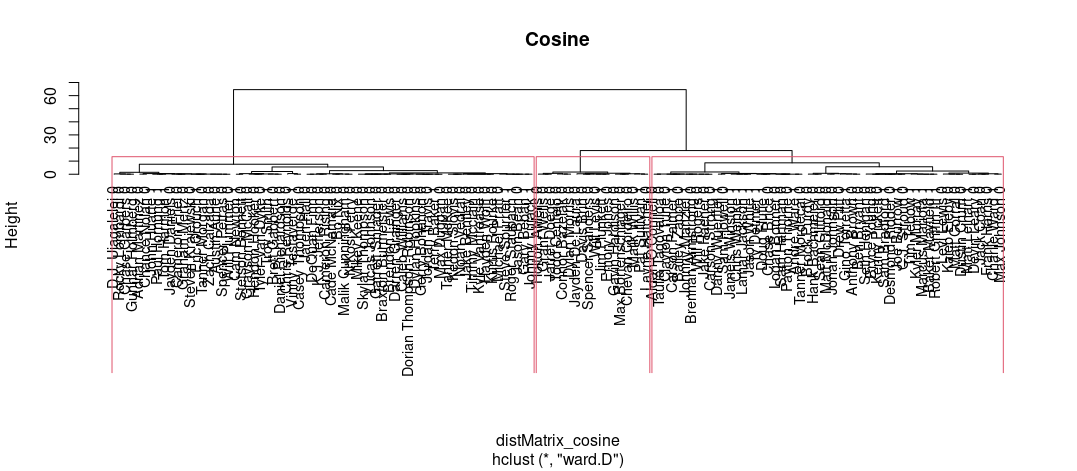
The above graph is the dendrogram for manhattan distance measure.

*40. Euclidean Cluster*



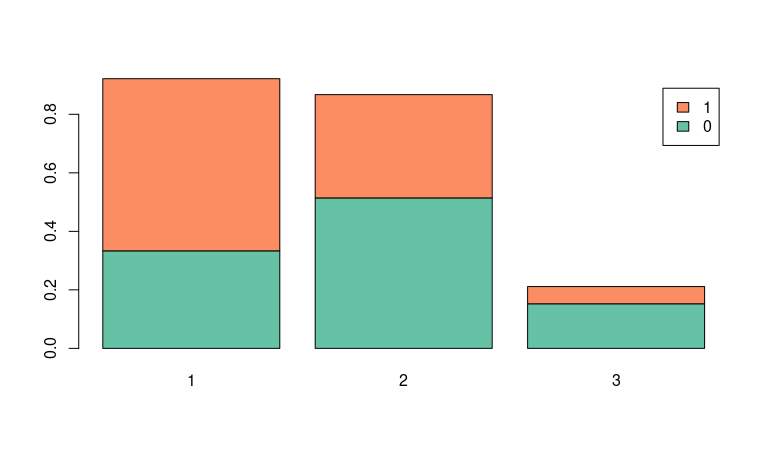
The above graph shows the dendrogram for eculidean distance measure. This distance measure performed better than canberra and manhattan because it was able to group more winners in one cluster.

*42. Cousine Cluster*



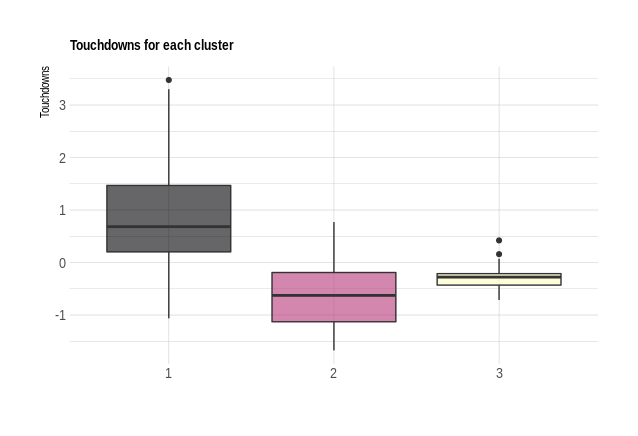
This above graph shows the dendrogram for the cosine distance measure. Hierarchical clustering with cosine distance measure was able to cluster a majority of the winners. Therefore, all further analysis was performed on this model.

*43. Cosine Cluster Breakdown*



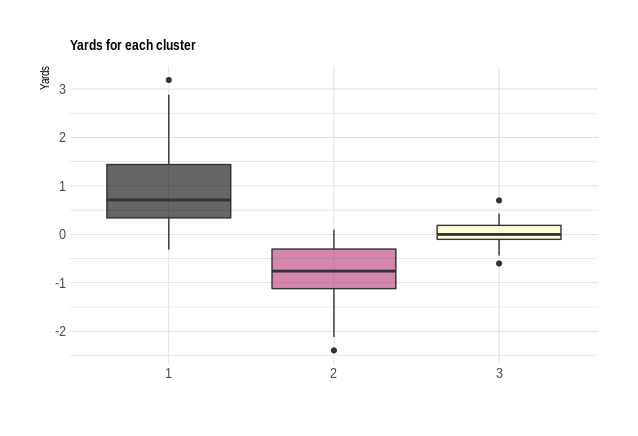
The above graph breaks up each of the three clusters between winners (red) and non-winners (green). Cluster 1 contains about 60% of the winners in the entire Combined dataset. Cluster 2 contains about 35% of the winners. Lastly, cluster 3 contains the remaining 5% of the winners.

*44. Touchdowns for each Cluster*

****

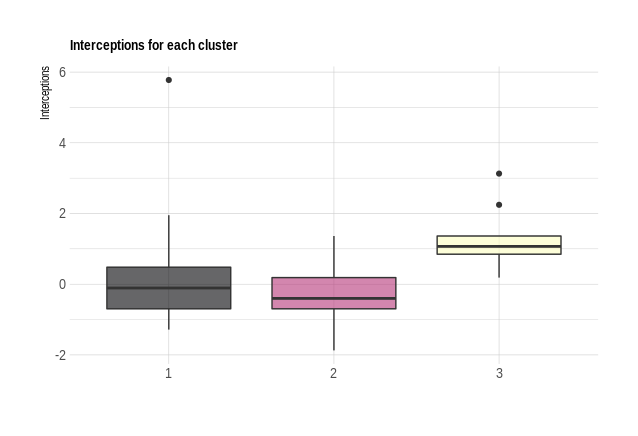
The above boxplot illustrates how touchdowns differ between clusters. Cluster 1 has the most touchdowns on average. Also, the first and third quartiles are above the other two groups, which make intuitive sense given that winners typically have more touchdowns than most other quarterbacks. Cluster 2 and 3 are largely similar when it comes to touchdowns.

*45. Yards for each Cluster*

****

The above boxplot shows the yards based on the stats of quarterbacks in each cluster. Quarterbacks in the first cluster typically have more yards than the other two clusters. Cluster 3 slightly outperforms cluster 2 in yards.

*46. Interceptions for each Cluster*



The above boxplot shows the interceptions for the quarterbacks in each cluster. Cluster 1 and 2 appear to be similar when it comes to interceptions. Cluster 3 has a higher average than the other two clusters. Interceptions is the most impactful of these metrics in terms of distinguishing between clusters.

**Conclusion**

After attempting a handful of methods, Bryce Young was indeed correctly predicted to be more likely than not to win the Heisman Trophy in each case. That being said, while it was found that many methods can predict with decent accuracy, predicting only one distinct Heisman Trophy winner was not always so easy. Some methods predicted which players would win the Heisman Trophy and some gave a probability for each player to win the Heisman Trophy, but only one had the criteria for predicting only one player. For some of those other methods, in certain cases, predicting a single player to win the Heisman Trophy could be done by determining which player is most likely to win the Heisman Trophy, albeit there is a relatively high chance of ties given the small number of factors that go into these predictions.

While making analytical discoveries, it is paramount to remember that correlation does not imply causation. In this context, this means that many factors, and potentially even the most important factor, in predicting Heisman Trophy winners are possibly outside the scope of this study. These reasons could range from class (Freshman, Sophomore, Junior, Senior), location of school, age of player, whether a player had previously utilized a redshirt season, media coverage, popularity, public image, academics, or a myriad of other factors. Without testing and accounting for this statistically, it can not be known how effective the factors used in this study truly are in the grand scheme of things.

It is also important to consider the assumptions that were made in this study. It was assumed that the Heisman Trophy winner was a quarterback, which also limited the training data to only Heisman-winning quarterbacks. As stated at the beginning of this paper, there have been many non-quarterbacks to win the Heisman Trophy (35+ running backs, multiple wide receivers, a defensive back). There is a possibility that completely different findings arise by using the same techniques to predict the Heisman Trophy winner when adding running backs to the candidate pool. Naturally, this would also require the addition of rushing statistics to all of the data used in this study. This might actually also improve the accuracy of quarterback predictions since rushing has become a significant piece of the average quarterback’s production.

So the question is, who will be the next Heisman Trophy Winner for 2022? Although we may not have an answer today, what we do have is a comprehensive data understanding as to what is sought out when nominating/choosing a winner.

Aside from setbacks in the data, there are also setbacks regarding human emotions. For instance, how does the player's popularity with fans and media come into play as it relates to nomination and winning. Is there a correlation between a player getting more headline coverage and their chances of winning the award? Or is it strictly based on the data? To mitigate these setbacks, the team used various methods to answer the question. What we discovered is that there isn’t one model that contains all of the answers.