Sports Analytics Final Project

Vineeth Madhavaram, Kevin Wu, Vincent Pi

Introduction

Our project goal is to predict the draft round of College Football Quarterbacks based on some of their senior season statistics such as yards thrown, touchdowns, interceptions and yards rushed. We aim to address uncertainties in the draft process by proposing a statistical foundation that could lead to a more meritocratic system, potentially reducing biases related to external factors like media exposure or school prestige. We also had a goal of determining which variables are most important in determining a QB's draft stock.

This project serves as a decision support tool for teams, enhancing their drafting strategy by offering an objective answer, reducing reliance on human bias, and allowing for a more data-driven approach. This not only helps in making informed choices during drafts but also aids in efficient allocation of scouting resources. The analysis of key performance indicators through this project can reflect which metrics are most valued by teams in quarterback selection, thus offering insights into player success factors and potentially predicting long-term professional success, and potentially paving the way on future work for understanding athlete development and career trajectories.

We used a few models: a generalized linear model, generalized additive model, and multinomial logistic regression model. In this scenario, we are using a GLM and GAM specifically tailored for a binary outcome. Logistic regression is used when the dependent variable is binary and the goal is to find the probability of the dependent variable being a success (in this case, being drafted in the first round). Our multinomial logistic regression model offers a direct and more nuanced approach to handling the categorical nature of draft rounds. It provides probabilities for each round, offering a detailed profile of potential draft outcomes for each player based on their metrics.

Exploratory Data Analysis & Data Summary

Data

The data set that we used for this project comes from "sports-reference.com" under the passing statistics for the different college football seasons. We then exported the data into a blank file in R and created a csv file of the data.

Our cfb_data dataset was taken from all 2023 college football quarterbacks stats and includes 107 observations with 17 variables. We filtered the original dataset to only include the top 60 ranked quarterbacks in order to streamline the prospects. The following were the different variables: rank, player name, school, conference games, passing completion, passing attempts, passing completion percentage, passing yards, passing yards per attempt, passing yards per game, adjusted passing yard per attempt ((Yards + 20TD - 45Int)/Att), touchdown, interceptions, passing efficiency rating ((8.4Yds + 330Td - 200Int + 100Cmp)/Att), rush attempts, rush yards, rush yards per attempt, and rushing touchdowns.

Similarly, our QBData came from "sports-reference.com" where we pulled different NFL quarterback's college football statistics. We looked at the college football seasons of 2018, 2019, 2020, 2021, and 2022, pulling out the quarterbacks that were drafted and a few undrafted players. The following variables was included in our data: rank, player name, school, conference games, passing completion, passing attempts, passing completion percentage, passing yards, passing yards per attempt, passing yards per game, adjusted passing yard per attempt ((Yards + 20TD - 45Int)/Att), touchdown, interceptions, passing efficiency rating ((8.4Yds

$+ 330$Td - 200$Int + 100$Cmp)/Att), rush attempts, rush yards, rush yards per attempt, rush touchdowns, and draft round (rounds 1-7 and 8 meaning "undrafted").

Histogram of Passing Yards Per Game for 2023 College Quarterbacks and Drafted Quarterbacks:

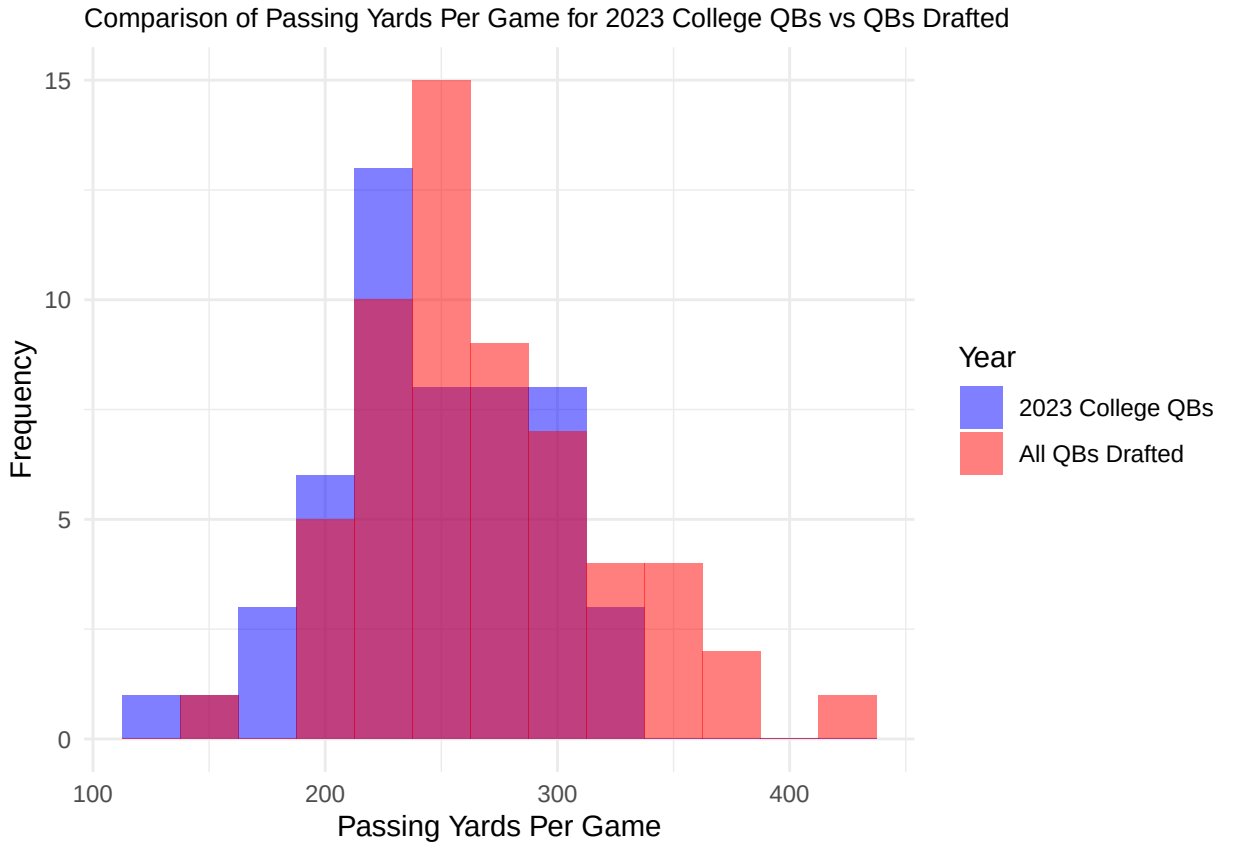

Figure 1: Histogram Comparing Passing Yards Per Game for 2023 College QBs vs QBs Drafted

EDA of Histogram for QBs Drafted Versus 2023 College QBs Passing Yards:

Based on the provided histogram, we observe a comparison of passing yards between 2023 college quarterbacks and all quarterbacks drafted. Both histograms seem to be unimodal, normally distributed, and with similar medians. This could be explained through the filtering of the sample of 2023 college quarterbacks to only the top 60 ranked quarterbacks. This also means that the stats are more likely to resemble the quarterbacks that were drafted from 2018 to 2021. However, the quarterbacks that were drafted seem to be shifted to right, meaning they had more individuals with greater passing yards per game than the quarterbacks that played in 2023. When looking at possible outliers, the college quarterbacks that were drafted have had outliers at over 400 yards per game. Furthermore, the difference in median seem to indicate that the QBs that were drafted had higher passing yards compared to the distribution of 2023 college quarterbacks.

EDA of Box Plot of QBs Passing Yards by Draft Round:

When analyzing the box plots of the QBs passing yards per game by draft round, we can see that the median of the QBs passing yards were slightly similar despite the different rounds the quarterbacks were drafted. While they were similar, the later rounds had lower medians compared to QBs drafted in rounds 1 and 2. Another differing aspect of the quarterbacks drafted in early round is that they had slightly larger quartile ranges compared to QBs drafted in the later rounds. Furthermore, in round 4 it was Baily Zappe who had 426 yards per game and in round 7 it was Cole McDonald who threw for 295 yards per game. Overall, the box plots do support the assumption that the more passing yards per game a qb had, the higher they would be drafted. A possible limitation to this visualization and the data is the skew for players in different

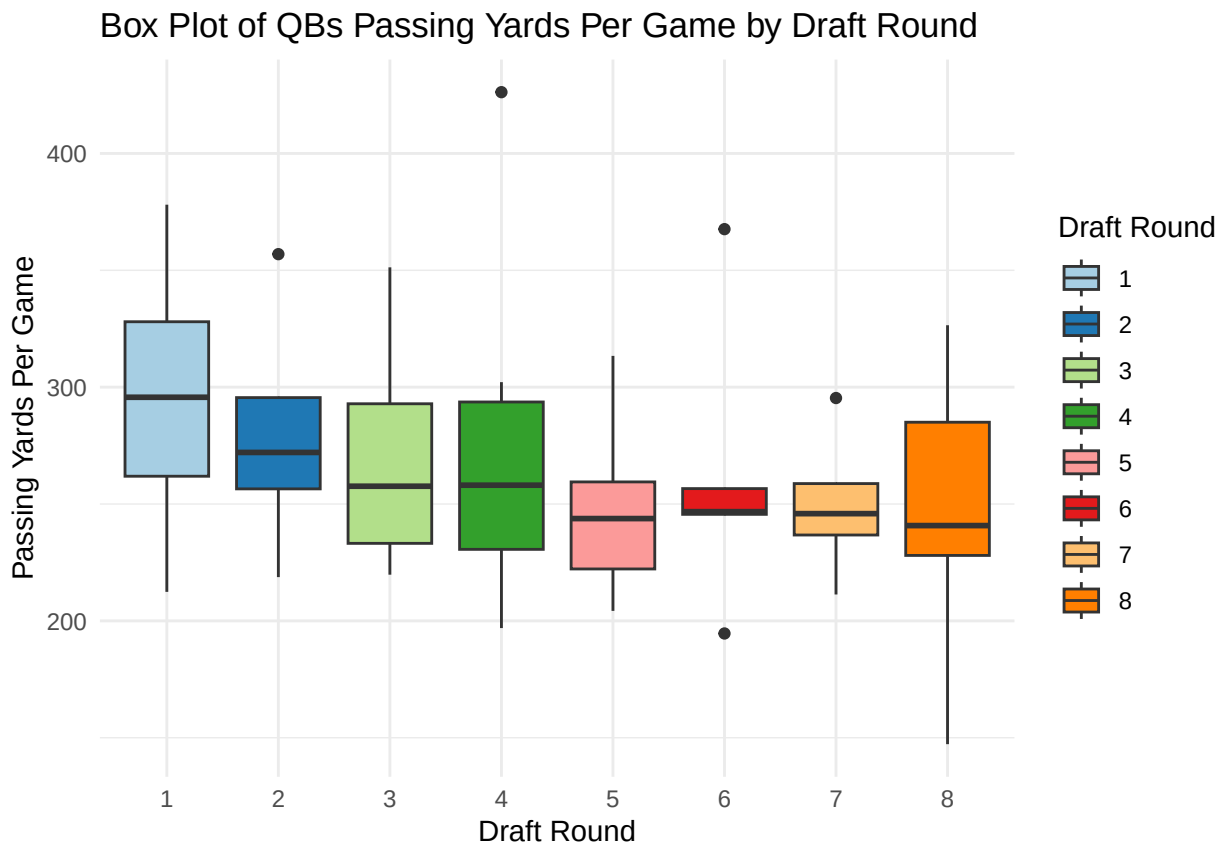

Figure 2: Box Plot of QB's Passing Yards Per Game by Round the QB was Drafted

conferences. A quarterback who played in a "weaker" conference can see his stats boosted by playing against less competitive defenses.

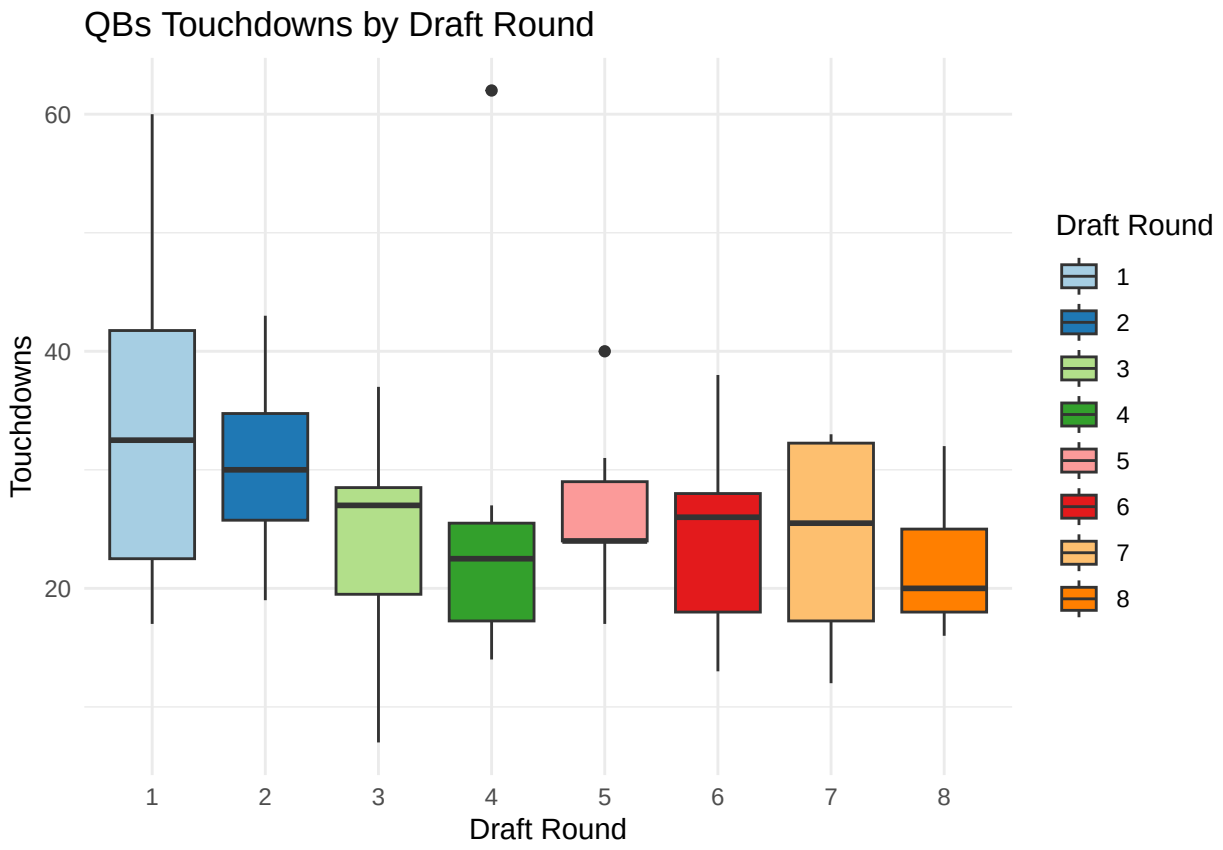

Figure 3: Box Plot of QB's Touchdown by Round the QB was Drafted

EDA of Box Plot of QBs Interceptions by Draft Round:

When analyzing the box plots of the QBs touchdown by draft round we can see that the median of the QBs touchdowns declined the later the quarterback was taken in the draft. While they were similar, the later rounds had lower medians compared to quarterbacks drafted in rounds 1 and 2. Another differing aspect of the quarterbacks drafted in the early rounds is that they had slightly larger quartile ranges compared to those drafted in the later rounds. Furthermore, a few of the rounds had high outliers. In round 4 it was Baily Zappe who had 62 touchdowns and in round 5 it was Clayton Tune who threw for 40 touchdowns. Also, Davis Mills threw only 7 touchdowns in his senior season before getting drafted, playing only 5 games. Overall, the box plots do support the assumption that the more touchdowns a quarterback had, the higher they would be drafted and as the touchdowns thrown descended the later the quarterback is drafted.

EDA of Box Plot of QBs Interceptions by Draft Round:

After analyzing the different box plots of QBs interceptions by draft round, we can notice the slight trend of early round quarterbacks having a lower median of interceptions thrown. A surprising outcome was that the quarterbacks in round 2 had one of the higher median of interceptions thrown but with the smallest range. A common explanation for why some of the quarterbacks drafted in the early rounds have higher interceptions could be due to the better quarterbacks having more passing attempts and a larger role in the passing game. One noticeable difference from the quarterbacks drafted in the early rounds is the range compared to the later rounds. The 1st and 2nd rounds had a much smaller quartile range compared to the later rounds. There are also noticeable outliers with some quarterbacks taken in round 1 with more interceptions. The round 1 outliers were Anthony Richardson who threw 9 interceptions and Jordan Love who threw 17 interceptions.

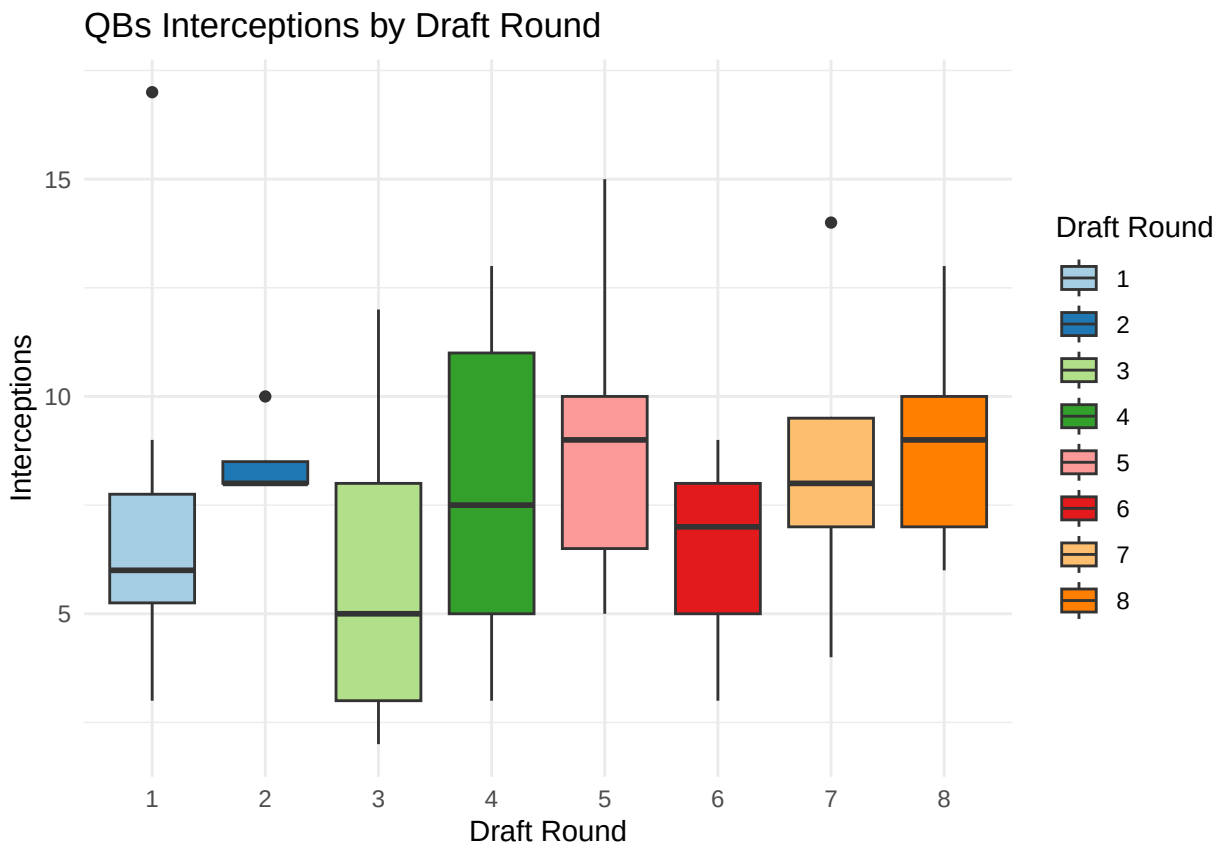

Figure 4: Box Plot of QB's Interception by Round the QB was Drafted

Table 1: Multinomial Estimates 95 Percent CI

	2.5 %.2	97.5 %.2	2.5 %.3	97.5 %.3	2.5 %.4	97.5 %.4	2.5 %.5	97.5 %.5	2.5 %.6	97.5 %.6	2.5 %.7	97.5 %.7	2.5 %.8	97.5 %.8
(Intercept)	-82	-80	-67	-64	-35	-33	-40	-36	171	172	-55	-53	21	24
Yds.G	0	0	0	0	0	0	0	0	0	0	-2	-1	0	0
TD	-1	0	-1	0	-1	0	0	1	0	3	6	11	-1	0
Int	1	6	-1	0	-2	0	-1	1	-12	-3	-5	2	-1	0
R.Avg	-3	1	-2	1	-2	0	-2	0	-14	9	-3	12	-2	0
Pct	0	1	1	2	1	2	1	2	-2	2	8	11	0	1
factor(Conf)American	NaN	NaN	97	105	-14	-14	45	53	NaN	NaN	-9	-9	-16	-16
factor(Conf)Big 12	21	32	50	70	-51	-51	NaN	NaN	95	104	109	114	-3	24
factor(Conf)Big Ten	-29	-29	NaN	NaN	-11	3	-8	3	68	71	-24	-6	-5	6
factor(Conf)CUSA	NaN	NaN	NaN	NaN	92	92	NaN	NaN	0	0	5	5	-13	-13
factor(Conf)Ind	-39	-39	45	54	-16	-16	-6	9	1	1	-18	-10	-28	-28
factor(Conf)MWC	-32	-32	-19	-19	-65	-65	NaN	NaN	NaN	NaN	53	58	-5	7
factor(Conf)Pac-12	-6	-6	25	43	-9	3	-15	-2	59	70	-63	-63	-5	6
factor(Conf)SEC	28	40	39	47	-2	9	-7	6	7	7	-75	-75	-2	6
AY.A	-2	9	-8	0	-10	-1	-8	-1	-34	-19	-52	-30	-6	2

Methods

The first model we looked at was a Multinomial Logistic Regression. In order to directly model a relationship that showcases our question at hand, we need to predict multiple rounds for each of the players. The reason that we choose to model using this statistical modeling technique was primarily to estimate probabilities for the multiple classes that pertains to the different rounds Draft. In order to utilize the model for this relationship we must assume that college players' performances are independent from one another given the EDA, recognize that players cannot be drafted into multiple rounds, construct our model such that we reduce closely related predictor variables, and remove outliers within the data. Additionally, from EDA we see a generally linear relationship between continuous variables albeit not incredibly strong providing some justification for a linear relationship between the response and the predictors.

In choosing the variables that would go into the model, we wanted to select the best model that also reduces closely related predictor variables (multicollinearilty). Through leave-one-out cross-validation measuring classification error, we refined our variable selection.

Likewise, the next models we used GLM and a GAM were specifically tailored for a binary outcome. Logistic regression is used when the dependent variable is binary and the goal is to find the probability of the dependent variable being a success (in this case, being drafted in the first round). As we have slight concerns about the linear relationship between the response and the predictors, we choose to also model a non-parametric GAM. The variables selected for this model were similar to those in the multinomial logistic regression model, but were to address the predicting first-round picks, which may differ in their predictive factors compared to later rounds.

We quantify uncertainty for our estimates of interest through 95% confidence intervals for the glm and multinomial models shown below:

Tables 1 & 2 reveal that a large majority of the confidence intervals encompass zero, spanning both positive and negative estimates. This suggests that the linear relationships between the predictor variables and the response variable may not be statistically significant. This evidence implies that the current model might not adequately capture the underlying relationships, which is important to consider as we move forward.

Results

Our Generalized Linear Model and Generalized Additive Model predicts the binary outcome of whether a player would be drafted in the first round by evaluating quarterback draft prospects based on numerous different quarterback statistics from their senior college season. These quarterback statistics for both models include: yards per game, passing touchdowns, interceptions, average rushing yards, passing completion percentage, the players conference, and average passing yards per attempt.

After fitting the models, we were able to summarize the data and get a percentage indicating the chance a quarterback will be drafted in the first round. The GLM model predicted that the quarterbacks that have the highest chance of getting drafted in the first round were:

Table 2: GLM Estimates 95 Percent CI

	lower_bound	upper_bound
(Intercept)	-93.968	86.722
Yds.G	-0.254	0.248
TD	-1.013	1.186
Int	-2.778	3.014
R.Avg	-3.076	3.718
Pct	-2.415	2.061
factor(Conf)American	-54939.014	54899.248
factor(Conf)Big 12	-43.068	30.368
factor(Conf)Big Ten	-18.961	17.689
factor(Conf)CUSA	-42976.515	42936.174
factor(Conf)Ind	-27.537	21.614
factor(Conf)MWC	-23.612	23.278
factor(Conf)Pac-12	-21.216	19.694
factor(Conf)SEC	-18.557	14.222
AY.A	-9.391	12.070

Table 3: LOOCV Classification Errors for Different Models

GLM	GAM	Multinomial
0.1379	0.1034	0.5862

Jayden Daniels (99%), Bo Nix (76.4%), Haynes King (69.0%), Drake Maye (59.3%), D.J. Uiagalelei (54.8%), and Caleb Williams (51.4%). Some of the quarterbacks with the lowest percentage of getting drafted in the first round were: Rocco Becht, Jacob Zeno, and Quinn Ewers. While the GAM model predicted the highest: Jayden Daniels (98%), Bo Nix (64.4%), Kaidon Salter (57.5%), and J.J. McCarthy (53.8%).

Our multinomial logistic regression model categorizes quarterbacks into their predicted draft rounds, providing a different perspective compared to the binary outcomes of the GLM. The variables or quarterback statistics that we included in our multinomial logistic regression model was: passing touchdowns, interceptions, passing yards per game, average rushing yards, completion percentage, quarterback's conference, and average passing yards per attempt. Figure 5 shows the implied relationships of Yards per Game to Round Drafted where we see a generally positive relationship across the variable with the total of Yds being increasingly associated with the first round.

The results of our multinominal logistic regression model were that it predicted quarterbacks Jayden Daniels, Bo Nix, Caleb Williams, Michael Penix Jr, Drake Maye, and D.J. Uiagalelel to be first round quarterbacks.

The table below showcases the LOOCV Classification Errors for the different models. The GAM model performed the best with a misclassifcation error of 10.3%. It is important to note that the GLM and GAM models are predicting binary variables whereas the multinomial model is predicting probabilities for each round. Though the model provides some insight on the relationships within the data, given the lack of confidence in the model's reliability through the high error and the uncertainty estimates, it is not entirely clear as to how important some factors impacts draft round.

Discussions

The results of our model to predict the draft round of college football quarterbacks had some pretty accurate results. As the 2024 NFL draft concluded the past week, the quarterbacks Caleb Williams, Jayden Daniels, Drake Maye, Micheal Penix Jr, J.J McCarthy and Bo Nix were all selected in the first round. From our multinomial logistic regression model, we were able to predict all but one of the first round quarterbacks, the missing one being McCarthy. This difference could be due to the model factoring only certain quarterback's

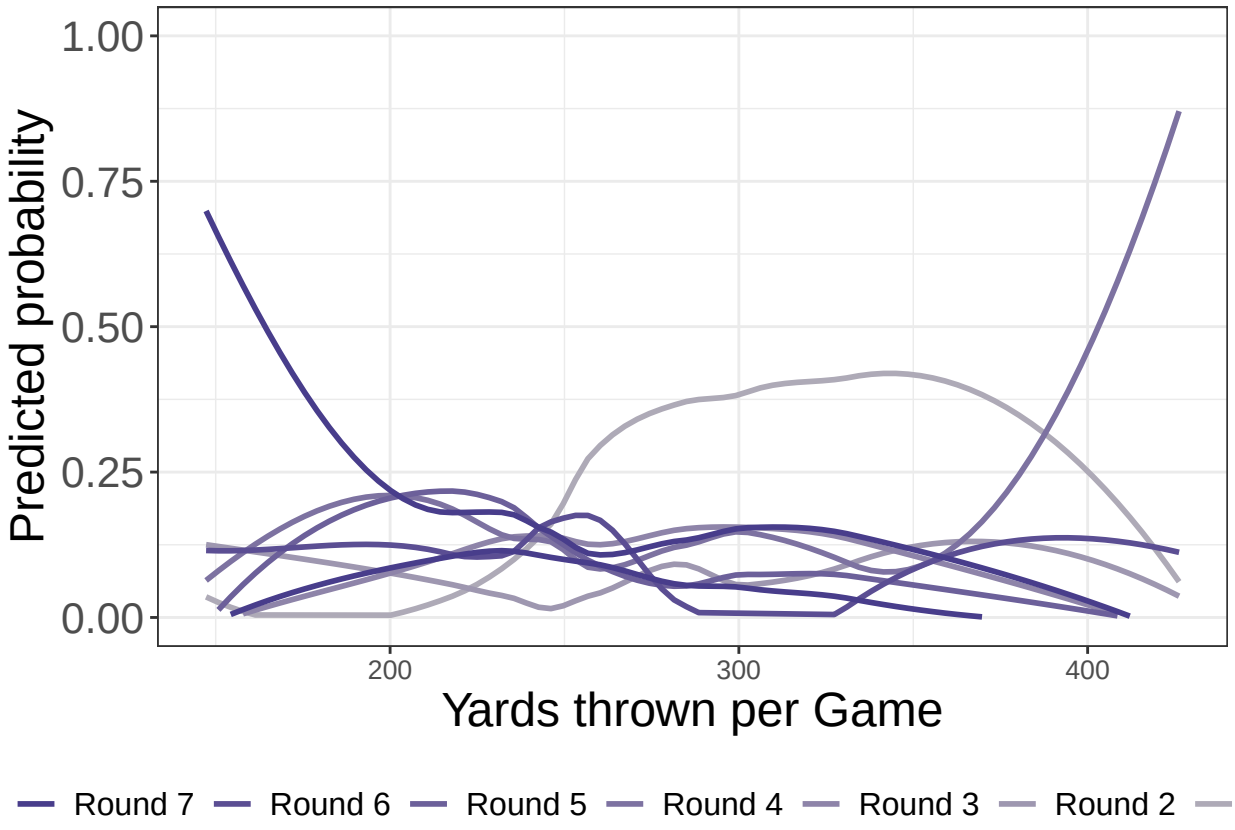

Figure 5: Implied relationships of Yards per Game to Round Drafted

statistics while McCarthy plays in a run heavy offense at Michigan. Similarly, our GLM was able to predict some of the top first round draft picks.

The model developed to predict which round college football quarterbacks will be drafted into the NFL offers significant insights, yet, like all models, it is constrained by certain limitations that could impact its accuracy and generalizability. One major limitation is selection bias. The model assumes a uniform interest in drafting quarterbacks across all NFL teams, though not every team will be looking to draft a quarterback each year, which can skew the results, potentially inflating the number of quarterbacks predicted to be drafted. This is a critical consideration since team needs significantly vary and influence drafting decisions. Another limitation pertains to different conferences. Good players in weaker conferences might be undervalued due to the perceived lower level of competition, which could skew the predictive accuracy of the model. Additionally, the size of the training data (only 56 players) limits the model's ability to learn and generalize. Compounded by a lack of data on many undrafted players, we might not provide a robust statistical basis to predict outcomes accurately across the wider population of college quarterbacks.

To address these limitations and enhance the model's predictive power, several steps could be undertaken: Incorporating a larger dataset that includes more players, including undrafted players, would help in developing a more accurate and robust model. A larger dataset would allow the model to better learn the nuances and variability inherent in draft outcomes and improve its generalization capabilities to unseen data. Also, extending this modeling approach to other positions such as running backs, wide receivers, and defensive ends could provide a comprehensive view of the draft dynamics across all key football positions. Specialized models could capture these nuances between positions more effectively. By expanding the scope and depth of the data and accounting for specific biases and external factors, future iterations of this project could offer even more precise guidance to NFL teams. This would not only improve draft strategies but also help in nurturing talent more effectively across the league and foster a deeper understanding of the draft process and player evaluation, potentially leading to innovations in how teams approach the draft.

Sports Analytics Final Project

Introduction

```
# IMPORT TEST DATA
cfb_data = read.csv("cfb_data.csv")
cfb_data <- cfb_data %>%
  mutate(Yds.G = Yds/G)
colnames(cfb_data)[17] = "R.Avg"
cfb_data = subset(cfb_data, cfb_data$Rk<60)

cfb_data = subset(cfb_data, cfb_data$Conf!= "Sun Belt")
cfb_data = subset(cfb_data, cfb_data$Conf!= "MAC")
cfb_data <- dplyr::select(cfb_data,-Player.additional)


cfb_data <- cfb_data %>%
  rename(R.Att = `Att.1`,     # Rename Att.1 to R.Att
         R.Yrds = `Yds.1`,    # Rename Yds.1 to R.Yrds
         R.TD = `TD.1`)       # Rename TD.1 to R.TD

# IMPORT TRAINING DATA
QB2018 = read.csv("2018QBs.csv")
QB2018$season <- 2018
QB2019 = read.csv("2019QBs.csv")
QB2019$season <- 2019
QB2020 = read.csv("cfb_data_2020.csv")
QB2020$season <- 2020
QB2021  = read.csv("cfb_data_2021.csv")
QB2021$Rk <- 0
QB2021 <- QB2021[c("Rk", setdiff(names(QB2021), "Rk"))]
QB2021$season <- 2021
QB2022 = read.csv("2022QBs.csv")
QB2022$season <- 2022



#Make final data
QBData <- rbind(QB2018, QB2019,QB2020,QB2021,QB2022)

QBData <- QBData %>%
  mutate(Yds.G = Yds/G)
QBData <- dplyr::select(QBData, -Rk)

QBData <- QBData %>%
  rename(R.Att = `Att.1`,     # Rename Att.1 to R.Att
         R.Yrds = `Yds.1`,    # Rename Yds.1 to R.Yrds
         R.Avg = Avg,         # Rename Avg to R.Avg
```

```
         R.TD = `TD.1`)       # Rename TD.1 to R.TD

QBData[is.na(QBData)] <- 8


QBData$First = ifelse(QBData$RoundDrafted=="1", 1,0)

QBData = subset(QBData, QBData$Conf!= "MAC")
QBData$First = ifelse(QBData$RoundDrafted=="1", 1,0)
```

Exploratory Data Analysis & Data Summary

Data

```
library(ggplot2)
ggplot() +
  geom_histogram(data = cfb_data, aes(x = Yds.G, fill = "2023"),
                 alpha = 0.5, binwidth = 25, position = "identity") +
  geom_histogram(data = QBData, aes(x = Yds.G, fill = "All"),
                 alpha = 0.5, binwidth = 25, position = "identity") +
  scale_fill_manual(values = c("2023" = "blue", "All" = "red"),
                    name = "Year",
                    labels = c("2023" = "2023 College QBs", "All" = "All QBs Drafted")) +
  labs(title = "Comparison of Passing Yards Per Game for 2023 College QBs vs QBs Drafted",
       x = "Passing Yards Per Game",
       y = "Frequency") +
  theme_minimal() +
  theme(plot.title = element_text(size = 10))
```

```
ggplot(QBData, aes(x = factor(RoundDrafted), y = Yds.G, fill = factor(RoundDrafted))) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Paired") +  # This gives different colors for each round
  labs(title = "Box Plot of QBs Passing Yards Per Game by Draft Round",
       x = "Draft Round",
       y = "Passing Yards Per Game",
       fill = "Draft Round") +
  theme_minimal()
```

```
ggplot(QBData, aes(x = factor(RoundDrafted), y = TD, fill = factor(RoundDrafted))) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Paired") +
  labs(title = "QBs Touchdowns by Draft Round",
       x = "Draft Round",
       y = "Touchdowns",
       fill = "Draft Round") +
  theme_minimal()
```

```
ggplot(QBData, aes(x = factor(RoundDrafted), y = Int, fill = factor(RoundDrafted))) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Paired") +
  labs(title = "QBs Interceptions by Draft Round",
       x = "Draft Round",
       y = "Interceptions",
       fill = "Draft Round") +
```

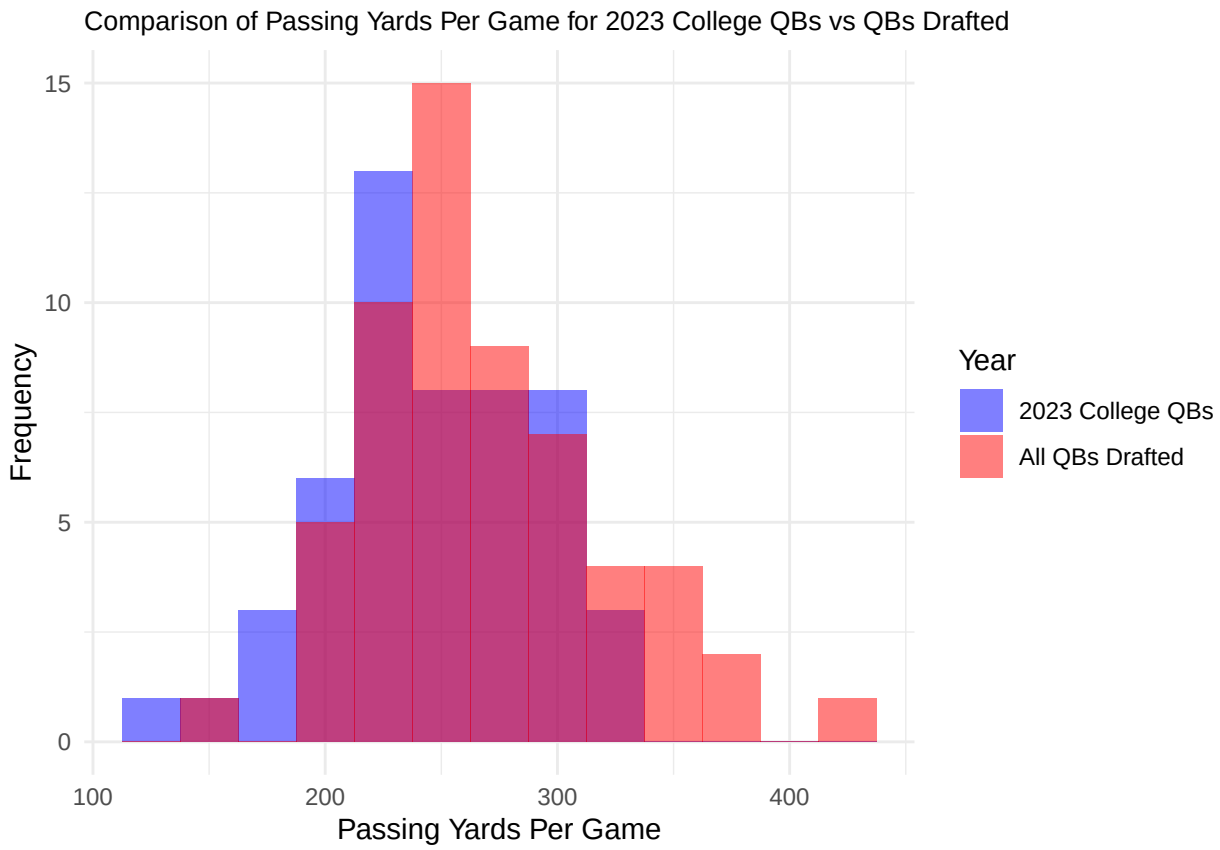

Figure 1: Histogram Comparing Passing Yards Per Game for 2023 College QBs vs QBs Drafted

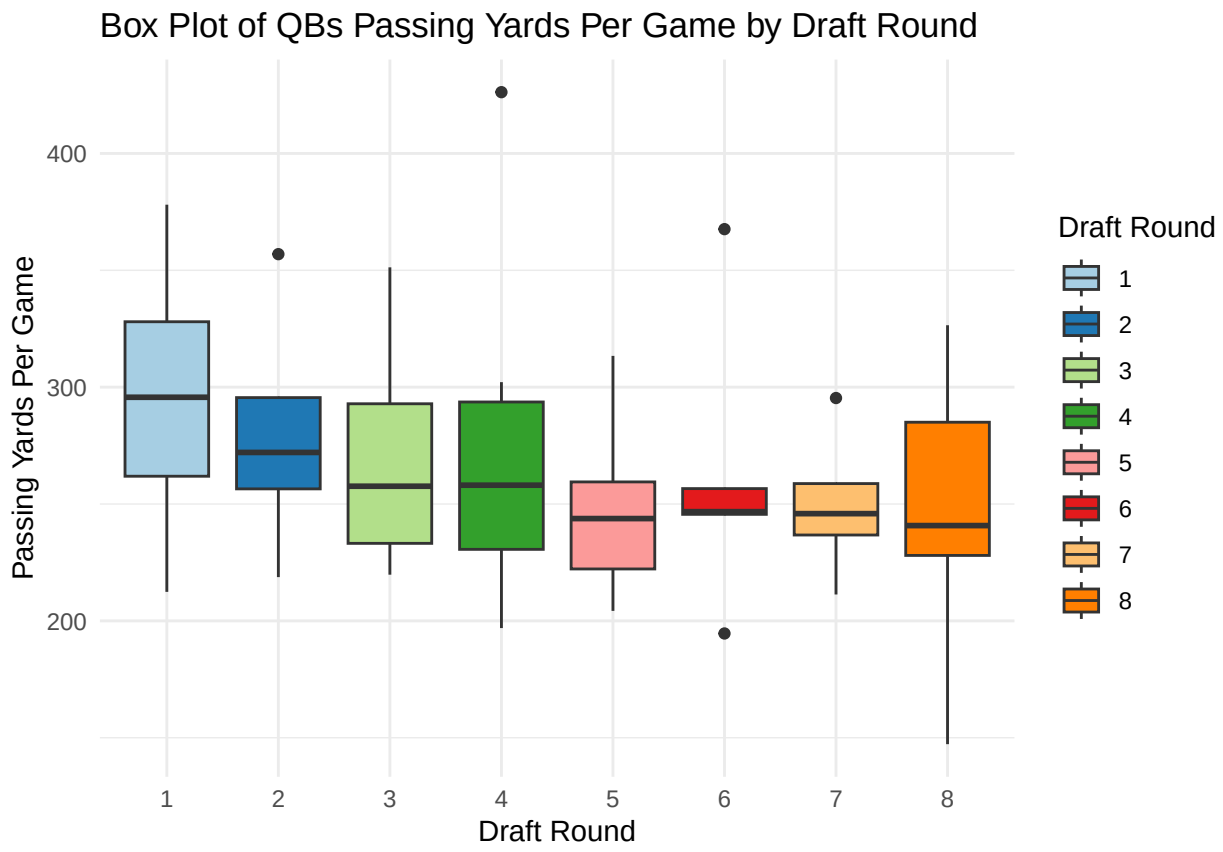

Figure 2: Box Plot of QB's Passing Yards Per Game by Round the QB was Drafted

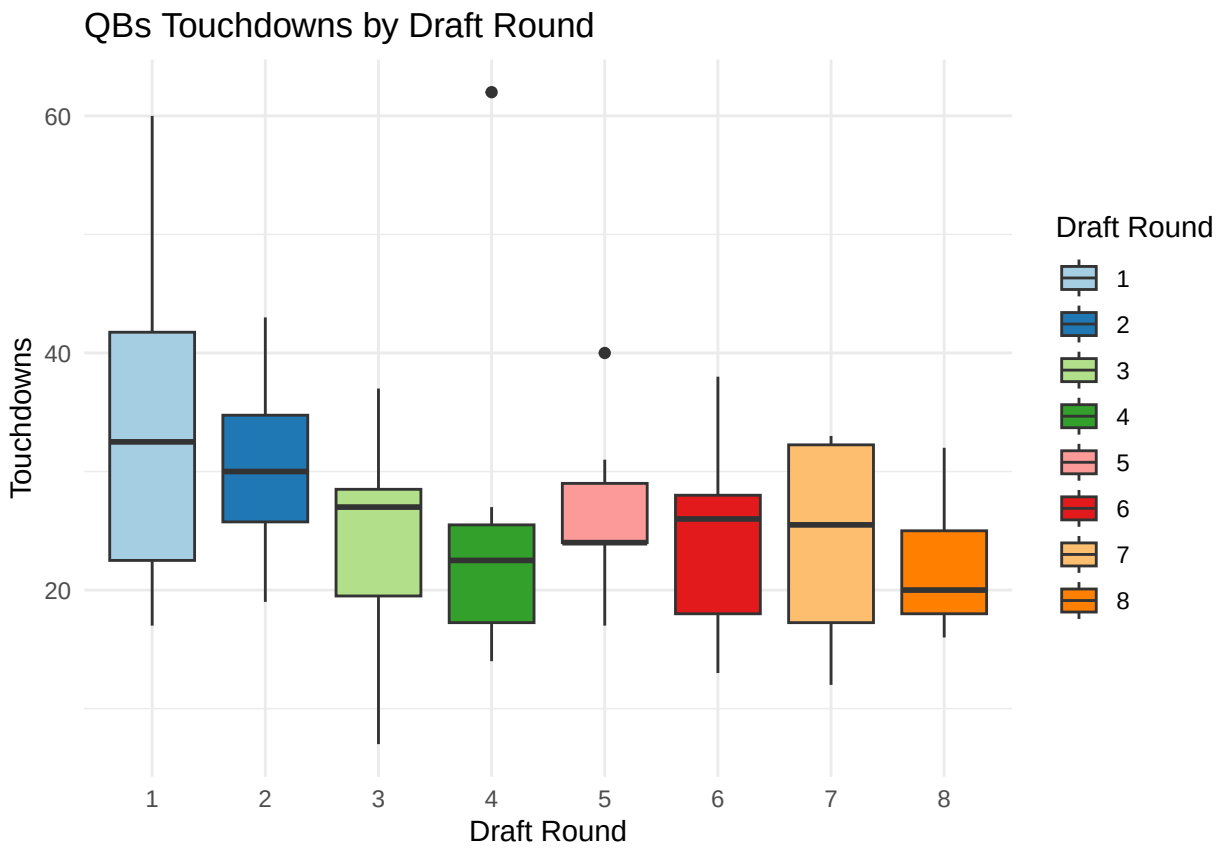

Figure 3: Box Plot of QB's Touchdown by Round the QB was Drafted

```
  theme_minimal()
```

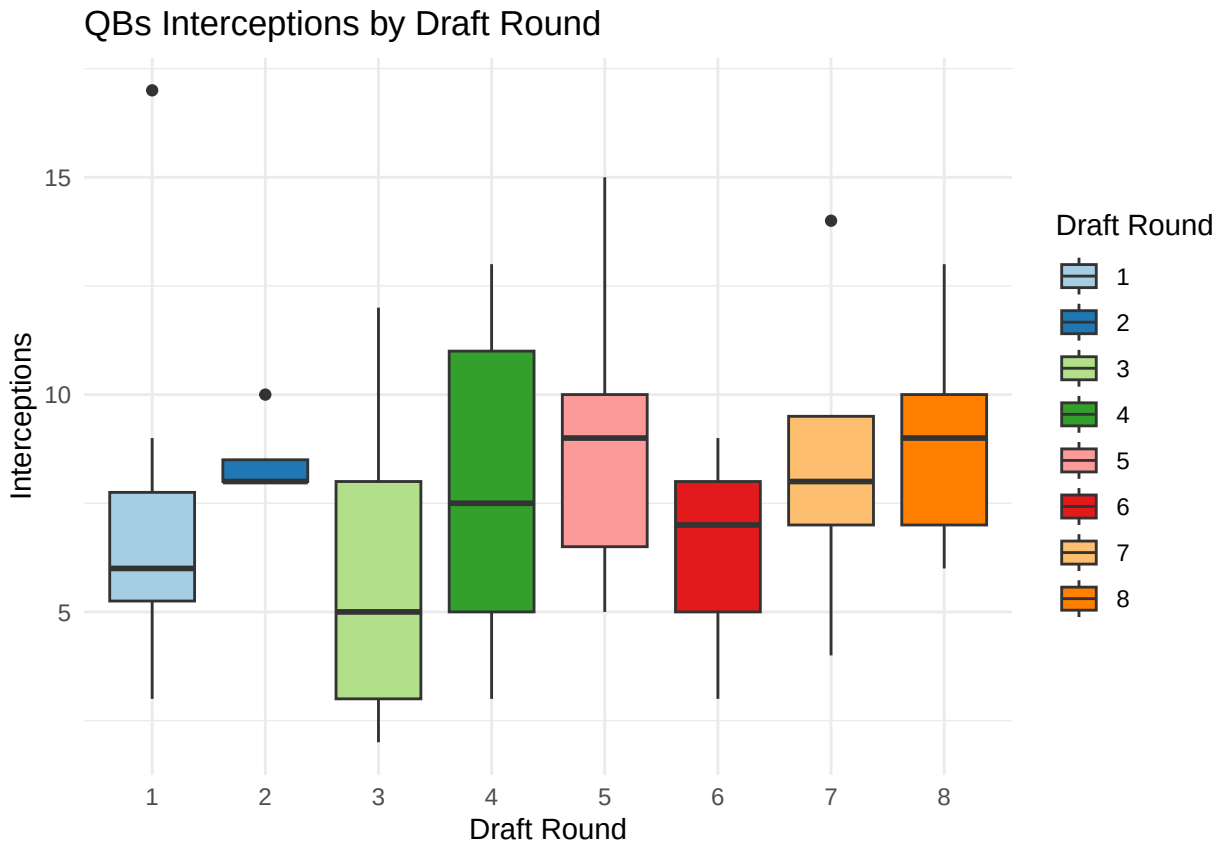

Figure 4: Box Plot of QB's Interception by Round the QB was Drafted

Methods

```
library(nnet)
library(MASS)

QBData$First = ifelse(QBData$RoundDrafted=="1", 1,0)

multi_cv_preds <-
  map_dfr(unique(QBData$Player),
          function(x) {
            # Separate test and training data:
            test_data <- QBData %>% filter(Player == x)
            train_data <- QBData %>% filter(Player != x)
            # Fit multinomial logistic regression model:
            ep_model <- nnet::multinom(RoundDrafted ~ Yds.G + TD + Int + R.Avg
                                       + Pct + factor(Conf) + AY.A,
                                       data = train_data, trace = FALSE)
            # Return dataset of class probabilities:
            preds <- predict(ep_model, newdata = test_data, type = "class")
            return(data.frame(Player = x, PredictedRound = preds))
              })
```

```
multicvsum = 0
for (i in 1:nrow(multi_cv_preds)) {
  # Access each row as df[i, ]
  if (abs(as.numeric(multi_cv_preds[i, "PredictedRound"]) -
          QBData[i, "RoundDrafted"]) == 0) {
    multicvsum <- multicvsum + 1
  }
}
multiacc = multicvsum/nrow(multi_cv_preds)
```

```
gam_cv_preds <-
  map_dfr(unique(QBData$Player),
          function(x) {
            # Separate test and training data:
            test_data <- QBData %>% filter(Player == x)
            train_data <- QBData %>% filter(Player != x)
            # Fit multinomial logistic regression model:
            ep_model <- gam(First ~ Yds.G + TD + Int + R.Avg + Pct +
                              factor(Conf) + AY.A, data = train_data)
            # Return dataset of class probabilities:
            predict(ep_model, newdata = test_data, type = "response")
              })

gam_cv_preds$`1`[gam_cv_preds$`1` >= .5] <- 1
gam_cv_preds$`1`[gam_cv_preds$`1` < .5] <- 0
gamcvsum = 0
for (i in 1:nrow(gam_cv_preds)) {
  # Access each row as df[i, ]
  if (abs(as.numeric(gam_cv_preds[i, "1"]) - QBData[i, "First"]) == 0) {
    gamcvsum <- gamcvsum + 1
  }
}
gamacc = gamcvsum/nrow(gam_cv_preds)
```

```
glm_cv_preds <-
  map_dfr(unique(QBData$Player),
          function(x) {
            # Separate test and training data:
            test_data <- QBData %>% filter(Player == x)
            train_data <- QBData %>% filter(Player != x)
            # Fit multinomial logistic regression model:
            ep_model <- glm(First ~ Yds.G + TD + Int + R.Avg + Pct +
                              factor(Conf) + AY.A, data = train_data)
            # Return dataset of class probabilities:
            predict(ep_model, newdata = test_data, type = "response")
              })

glm_cv_preds$`1`[glm_cv_preds$`1` >= .5] <- 1
glm_cv_preds$`1`[glm_cv_preds$`1` < .5] <- 0

glmcvsum = 0
for (i in 1:nrow(glm_cv_preds)) {
  # Access each row as df[i, ]
```

Table 1: Multinomial Estimates 95 Percent CI

	2.5 %.2	97.5 %.2	2.5 %.3	97.5 %.3	2.5 %.4	97.5 %.4	2.5 %.5	97.5 %.5	2.5 %.6	97.5 %.6	2.5 %.7	97.5 %.7	2.5 %.8	97.5 %.8
(Intercept)	-82	-80	-67	-64	-35	-33	-40	-36	171	172	-55	-53	21	24
Yds.G	0	0	0	0	0	0	0	0	0	0	-2	-1	0	0
TD	-1	0	-1	0	-1	0	0	1	0	3	6	11	-1	0
Int	1	6	-1	0	-2	0	-1	1	-12	-3	-5	2	-1	0
R.Avg	-3	1	-2	1	-2	0	-2	0	-14	9	-3	12	-2	0
Pct	0	1	1	2	1	2	1	2	-2	2	8	11	0	1
factor(Conf)American	NaN	NaN	97	105	-14	-14	45	53	NaN	NaN	-9	-9	-16	-16
factor(Conf)Big 12	21	32	50	70	-51	-51	NaN	NaN	95	104	109	114	-3	24
factor(Conf)Big Ten	-29	-29	NaN	NaN	-11	3	-8	3	68	71	-24	-6	-5	6
factor(Conf)CUSA	NaN	NaN	NaN	NaN	92	92	NaN	NaN	0	0	5	5	-13	-13
factor(Conf)Ind	-39	-39	45	54	-16	-16	-6	9	1	1	-18	-10	-28	-28
factor(Conf)MWC	-32	-32	-19	-19	-65	-65	NaN	NaN	NaN	NaN	53	58	-5	7
factor(Conf)Pac-12	-6	-6	25	43	-9	3	-15	-2	59	70	-63	-63	-5	6
factor(Conf)SEC	28	40	39	47	-2	9	-7	6	7	7	-75	-75	-2	6
AY.A	-2	9	-8	0	-10	-1	-8	-1	-34	-19	-52	-30	-6	2

```
  if (abs(as.numeric(glm_cv_preds[i, "1"]) - QBData[i, "First"]) == 0) {
    glmcvsum <- glmcvsum + 1
  }
}
glmacc = glmcvsum/nrow(glm_cv_preds)
```

```
glm.model = glm(First ~ Yds.G + TD + Int + R.Avg + Pct + factor(Conf)
                + AY.A, data = QBData, family = binomial())

multinom.model = nnet::multinom(RoundDrafted ~ Yds.G + TD + Int + R.Avg
                                + Pct + factor(Conf) + AY.A,
                                data = QBData, trace = FALSE)

gam.model = gam(First ~ Yds.G + TD + Int + R.Avg + Pct + factor(Conf)
                + AY.A, data = QBData)
```

```
glm.results = (predict(glm.model , newdata = cfb_data, type = "response")) %>%
  as_tibble() %>%
  mutate(Player = cfb_data$Player)
colnames(glm.results)[1] <- "Chance of Drafted in First"
glm.results = glm.results[order(-glm.results$"Chance of Drafted in First"),]

gam.results = (predict(gam.model , newdata = cfb_data, type = "response")) %>%
  as_tibble() %>%
  mutate(Player = cfb_data$Player)
colnames(gam.results)[1] <- "Chance of Drafted in First"
gam.results = gam.results[order(-gam.results$"Chance of Drafted in First"),]



multi.results <- predict(multinom.model, newdata = cfb_data, type = "class") %>%
  as_tibble() %>%
  mutate(Player = cfb_data$Player)
colnames(multi.results)[1] <- "PredictedRound"
multi.results$PredictedRound = ifelse(multi.results$PredictedRound==8,"Undrafted",
                                      multi.results$PredictedRound)
multi.results = multi.results[order(multi.results$PredictedRound),]
```

```
kable(round(confint(multinom.model)), caption = "Multinomial Estimates 95 Percent CI") %>%
  kable_styling(latex_options="scale_down")
```

Table 2: GLM Estimates 95 Percent CI

	lower_bound	upper_bound
(Intercept)	-93.968	86.722
Yds.G	-0.254	0.248
TD	-1.013	1.186
Int	-2.778	3.014
R.Avg	-3.076	3.718
Pct	-2.415	2.061
factor(Conf)American	-54939.014	54899.248
factor(Conf)Big 12	-43.068	30.368
factor(Conf)Big Ten	-18.961	17.689
factor(Conf)CUSA	-42976.515	42936.174
factor(Conf)Ind	-27.537	21.614
factor(Conf)MWC	-23.612	23.278
factor(Conf)Pac-12	-21.216	19.694
factor(Conf)SEC	-18.557	14.222
AY.A	-9.391	12.070

```
coef_summary <- summary(glm.model)
se <- sqrt(diag(coef_summary$cov.unscaled * coef_summary$deviance))
lower_bound <- coef(glm.model) - 1.96 * se
upper_bound <- coef(glm.model) + 1.96 * se
confidence_intervals <- cbind(lower_bound, upper_bound)
kable(round(confidence_intervals, 3), caption = "GLM Estimates 95 Percent CI")
```

Results

```
glm.results_QB = (predict(glm.model , newdata = QBData, type = "response")) %>%
  as_tibble() %>%
  mutate(Player = QBData$Player)
colnames(glm.results_QB)[1] <- "Chance of Drafted in First"

glm.results_QB$`Chance of Drafted in First`[glm.results_QB$`Chance of Drafted in First` >= .5] <- 1
glm.results_QB$`Chance of Drafted in First`[glm.results_QB$`Chance of Drafted in First` < .5] <- 0

gam.results_QB = (predict(gam.model , newdata = QBData, type = "response")) %>%
  as_tibble() %>%
  mutate(Player = QBData$Player)
colnames(gam.results_QB)[1] <- "Chance of Drafted in First"
gam.results_QB$`Chance of Drafted in First`[gam.results_QB$`Chance of Drafted in First` >= .5] <- 1
gam.results_QB$`Chance of Drafted in First`[gam.results_QB$`Chance of Drafted in First` < .5] <- 0


multi.results_QB <- predict(multinom.model, newdata = QBData, type = "class") %>%
  as_tibble() %>%
  mutate(Player = QBData$Player)
colnames(multi.results_QB)[1] <- "PredictedRound"
multi.results_QB$PredictedRound = ifelse(multi.results_QB$PredictedRound==8,"Undrafted",
                                         multi.results_QB$PredictedRound)
multi.results_QB = multi.results_QB[order(multi.results_QB$PredictedRound),]
```

```
multi.results_QB$PredictedRound[multi.results_QB$PredictedRound == "Undrafted"] <- 8
```

```
multisum = 0
for (i in 1:nrow(multi.results_QB)) {
  # Access each row as df[i, ]
  if (abs(as.numeric(multi.results_QB[i, "PredictedRound"]) -
          QBData[i, "RoundDrafted"]) > 2) {
    multisum <- multisum + 1
  }
}
multierror = multisum/nrow(multi.results_QB)


glmsum = 0
for (i in 1:nrow(glm.results_QB)) {
  # Access each row as df[i, ]
  if (abs(as.numeric(glm.results_QB[i, "Chance of Drafted in First"]) -
          QBData[i, "First"]) != 0) {
    glmsum <- glmsum + 1
  }
}
glmerror = glmsum/nrow(glm.results_QB)

gamsum = 0
for (i in 1:nrow(gam.results_QB)) {
  # Access each row as df[i, ]
  if (abs(as.numeric(gam.results_QB[i, "Chance of Drafted in First"]) -
          QBData[i, "First"]) != 0) {
    gamsum <- gamsum + 1
  }
}
gamerror = gamsum/nrow(gam.results_QB)

majoritysum = 0
for (i in 1:nrow(gam.results_QB)) {
  # Access each row as df[i, ]
  if (QBData[i, "First"] != 0) {
    majoritysum <- majoritysum + 1
  }
}
majorityvoteerror = majoritysum/nrow(gam.results_QB)
```

```
round_probs <- predict(multinom.model,
                           newdata = QBData, type = "probs") %>%
  as_tibble()

QBData %>%
  # Join the probs:
  bind_cols(round_probs) %>%
  # Only grab a subset of columns
  dplyr::select(Yds.G, "1":"8") %>%
  pivot_longer("1":"8",
               # Name of the column for the outcomes
               names_to = "Round",
```

Table 3: LOOCV Classification Errors for Different Models

GLM	GAM	Multinomial
0.1379	0.1034	0.5862

```
              # Name of the column for the predicted probabilities
              values_to = "pred_prob") %>%
  # Create a score value column to use for the color legend
  mutate(event_value = case_when(
                         Round == "1" ~ 1,
                         Round == "2" ~ 2,
                         Round == "3" ~ 3,
                         Round == "4" ~ 4,
                         Round == "5" ~ 5,
                         Round == "6" ~ 6,
                         TRUE ~ 7)) %>%
  ggplot(aes(x = Yds.G, y = pred_prob, color = event_value,
             group = Round)) +
  geom_smooth(se = FALSE) +
    ylim(0,1) +
  theme_bw() +
  labs(x = "Yards thrown per Game", y = "Predicted probability") +
  scale_color_gradient2(low = "darkorange4", mid = "gray",
                        high = "darkslateblue",
                        breaks = c(1, 2, 3, 4, 5, 6, 7),
                        labels=c("Round 1", "Round 2" ,
                                 "Round 3", "Round 4",
                                 "Round 5", "Round 6",
                                 "Round 7"),
                        guide = guide_legend(title = NULL, ncol = 7,
                                             reverse = TRUE,
                                             override.aes = list(size = 5))) +
  theme(legend.background = element_rect(fill = "white"),
        axis.title = element_text(size = 18),
        axis.text.y = element_text(size = 16),
        axis.text.x = element_text(size = 10),
        legend.position = "bottom",
        strip.background = element_blank(),
        strip.text = element_text(size = 18),
        legend.text = element_text(size = 12))
```

```
results <- data.frame("GLM" = glmerror, "GAM" = gamerror, "Multinomial" = multierror)
kable(round(results,4), caption = "LOOCV Classification Errors for Different Models")
```

Discussions

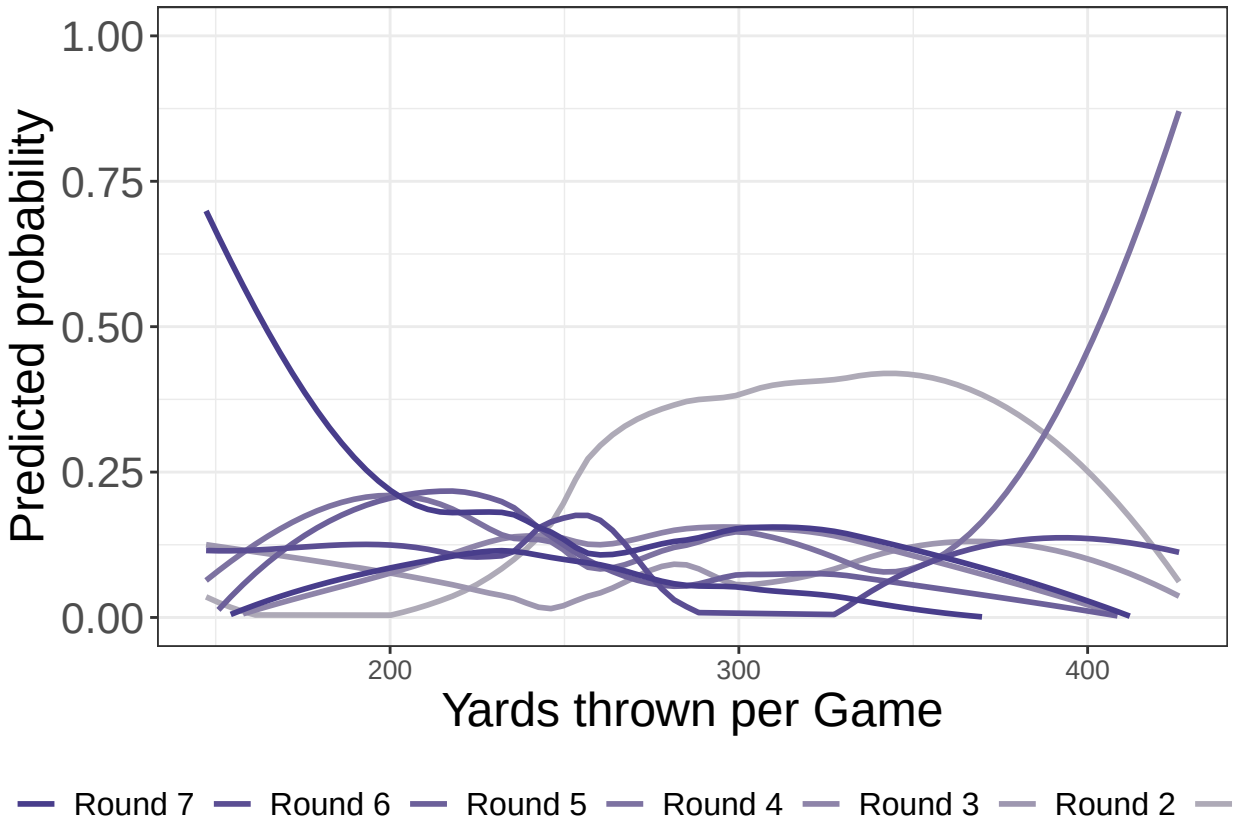

Figure 5: Implied relationships of Yards per Game to Round Drafted