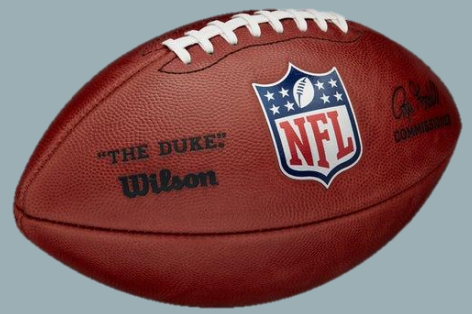
PREDICTING NFL ROOKIE SUCCESS



Amelie Devine

2025-12-17

AGENDA

I. Project Overview
II. Data Introduction
III. Variable Exploration
IV. Modelling Process
    I. Random Forest
    II. Logistic Regression
V. Model Assumptions
VI. Final Visualizations
VII. Conclusions & Limitations

OVERVIEW

- **Goal:** Predict NFL success of college football players during their rookie season
- **Motivation:** Summer experiences made me much more interested in the world of scouting & player evaluation
- **Definition of Success:** on an active NFL roster for > 8 weeks
- **Scope of Inference**: on active NFL roster for at least 1 week in their first season after being drafted, 2015-2024

DATA INTRODUCTION

- **Data Sources:** `cfbfastR` package & `nflreadR` package
  - NFL Roster & Draft information
  - College stats
  - Wanted to include NFL Combine info but couldn't find accurate source

success_v1	**primary_roster_team**	**position_group**	draft_round	dr_av	college_year
weeks_active	**roster_team_wp_prev**	**side_of_ball**	**draft_round_cat**	team	**num_schools**
gsis_id	**changed_teams**	**different_position**	draft_pick	conference	position_college
full_name	**multiple_teams**	**drafted**	draft_team	**div_level**	**pos_skill_flag**
season	position	age	**draft_team_wp_prev**	**comp_level**	

DATA INTRODUCTION

- **Major Decisions**
  - Impute *draft_av* & *age*
  - Drop unmatched observations
  - Creation of binary predictor variables

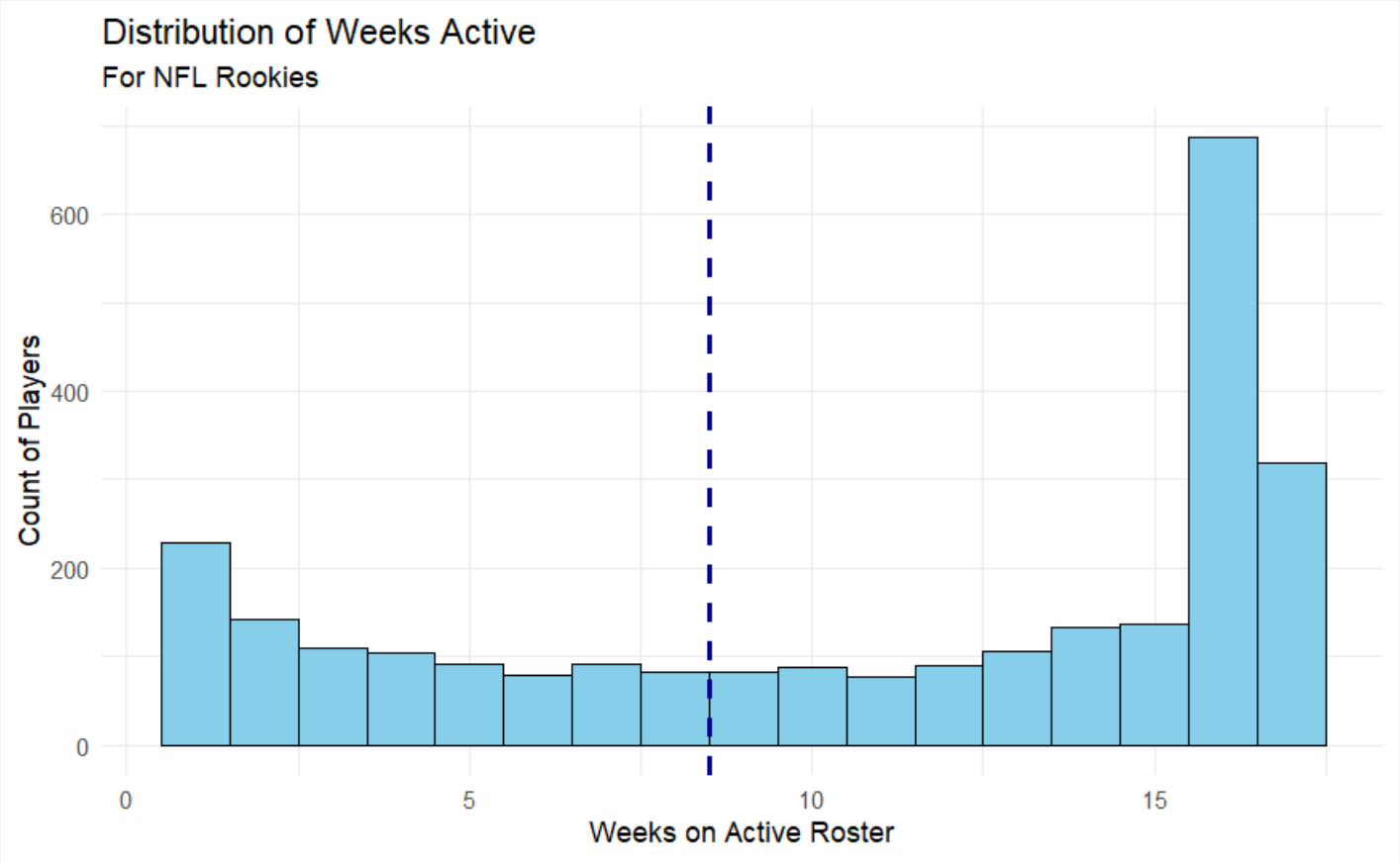
DATA INTRODUCTION

- **Creation:** pos_skill_flag

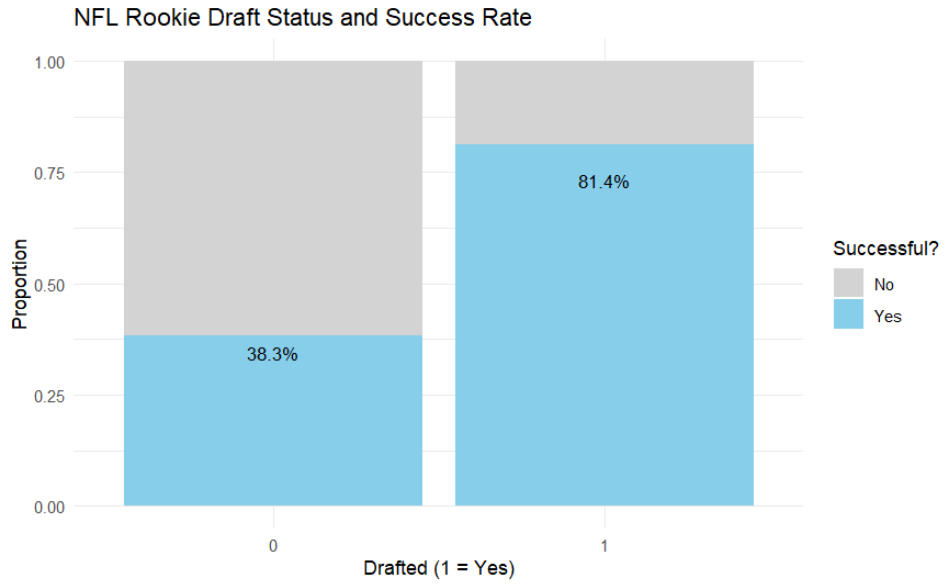
	position	rushing_car	rushing_yds	rushing_td	rushing_ypc	rushing_long	receiving_rec	receiving_yds	receivin
1	FB	21.684211	98.17763	1.2302632	3.425000	12.296053	4.341615	42.34161	0.
2	RB	50.272476	251.58713	2.3461866	4.329591	26.383894	8.946948	73.74145	0.
3	TE	4.420863	16.41367	0.4064748	3.746043	7.820144	10.704197	122.34826	1.
4	WR	4.914526	26.95385	0.2299358	4.818218	12.147673	17.647696	234.42602	1.

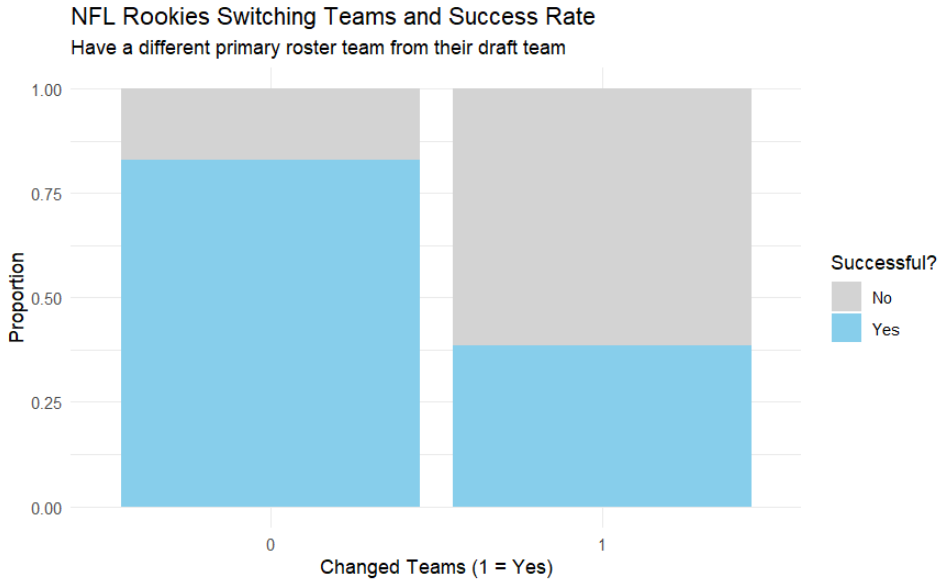
athlete_id	position	rushing_car_flag	rushing_yds_flag	rushing_td_flag	rushing_ypc_flag	rushing_long_flag	receiving
109907	RB	below	below	below	above	above	below
147386	WR	at	at	at	at	at	below
2026567	WR	below	below	below	above	above	at
3116652	WR	at	at	at	at	at	below
3116723	WR	at	at	at	at	at	below
3116742	WR	below	below	below	below	below	above
3116757	RB	above	above	below	above	above	above

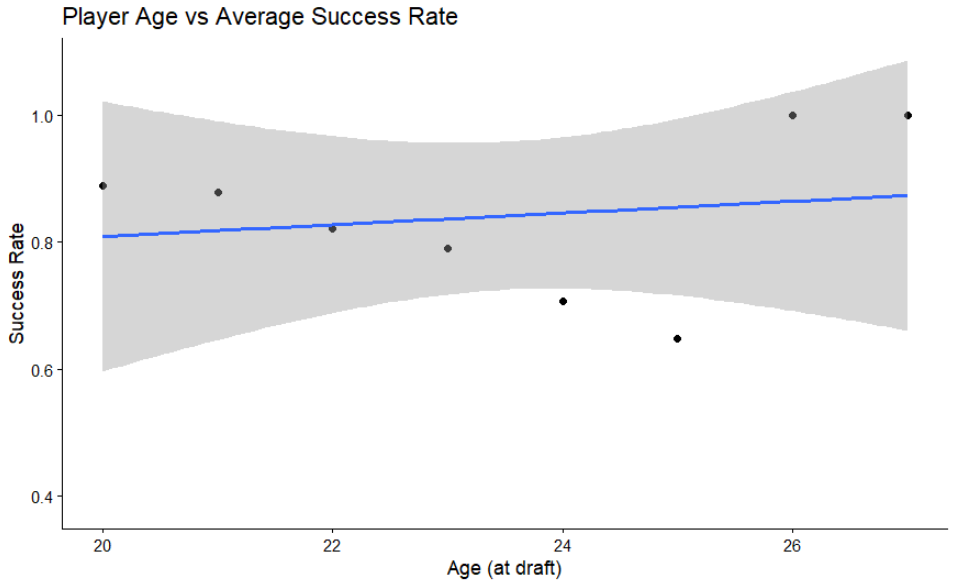
VARIABLE EXPLORATION

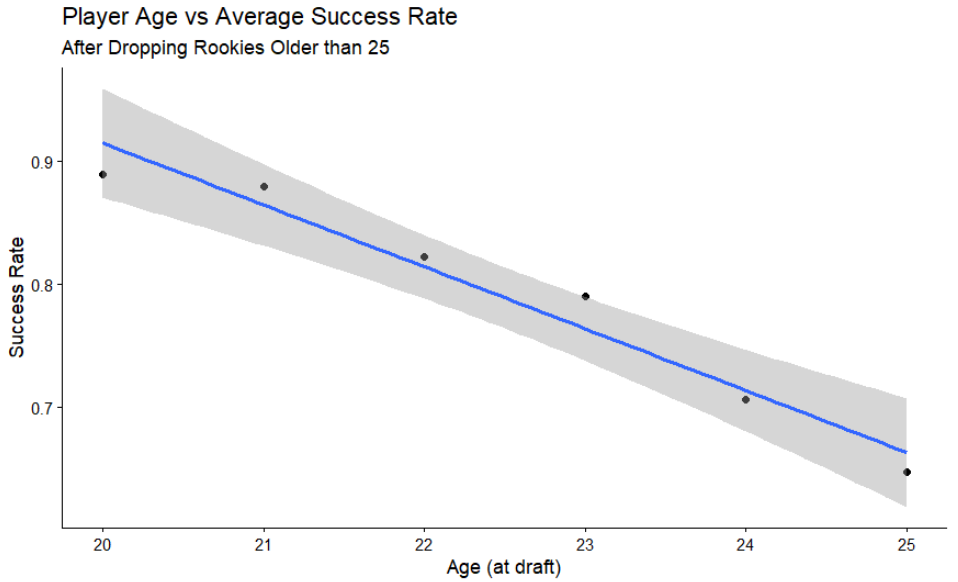


VARIABLE EXPLORATION









VARIABLE EXPLORATION

MODELLING PROCESS

- 80/20 Train/Test split
- Random Forest to select most important variables
  - Used Cramer's V statistic to deal with multicollinearity first
- Use top 8 variables for logistic regression model
- Select best model
- Make predictions for 2025

RANDOM FOREST

- Started with:
  - mtry = default (4)
  - ntrees = 501
- Adjusted each separately to maximize F-score
  - mtry = 4
  - ntrees = 501

FINAL RANDOM FOREST MODEL

VARIABLES

- draft_round
- dr_av
- age
- side_of_ball
- num_schools
- different_position
- multiple_teams
- comp_level
- div_level
- pos_skill_flag
- primary_roster_team
- roster_team_wp_prev

ASSESSMENT METRICS

- Accuracy*: 74.24% (> NIR 64.96%)
- F1 Score: 0.8201

FINAL RANDOM FOREST MODEL

VARIABLES

- **draft_round**
- **dr_av**
- **age**
- **side_of_ball**
- **num_schools**
- **different_position**
- multiple_teams
- **comp_level**
- div_level
- pos_skill_flag
- **primary_roster_team**
- roster_team_wp_prev

ASSESSMENT METRICS

- Accuracy*: 74.24% (> NIR 64.96%)
- F1 Score: 0.8201

LOGISTIC REGRESSION MODEL

- Top 8 features from RF
- **Selection Methods:**
  - Stepwise
  - Forward
  - Backward
- **Selection Criteria:** AIC

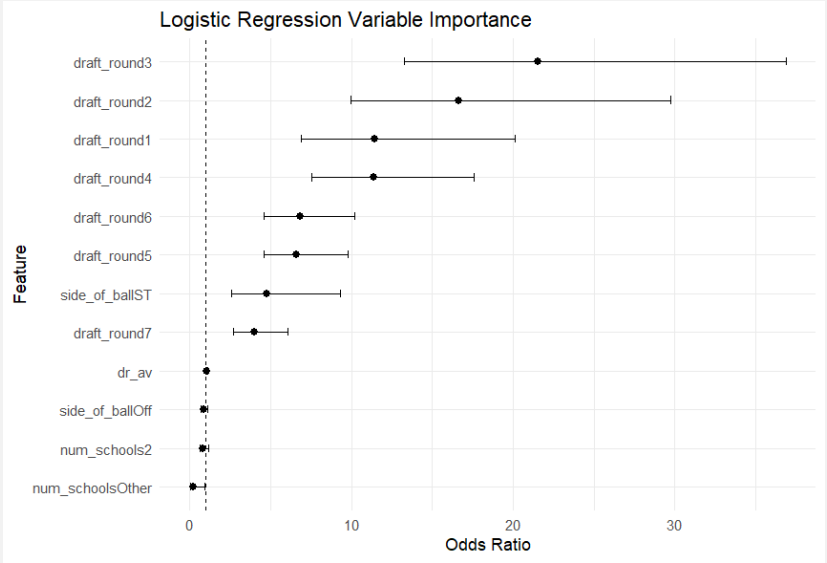
```
Step:  AIC=2180.25
success_v1 ~ draft_round + dr_av + side_of_ball + num_schools

                       Df Deviance    AIC
<none>                      2154.2 2180.2
+ age                   1   2152.5 2180.5
- num_schools           2   2159.6 2181.6
+ different_position    1   2154.2 2182.2
+ comp_level            3   2152.2 2184.2
+ primary_roster_team  31   2113.8 2201.8
- side_of_ball          2   2184.4 2206.4
- dr_av                 1   2200.5 2224.5
- draft_round           7   2666.1 2678.1
```

FINAL LOGISTIC REGRESSION MODEL

VARIABLES

- draft_round
- dr_av
- side_of_ball
- num_schools

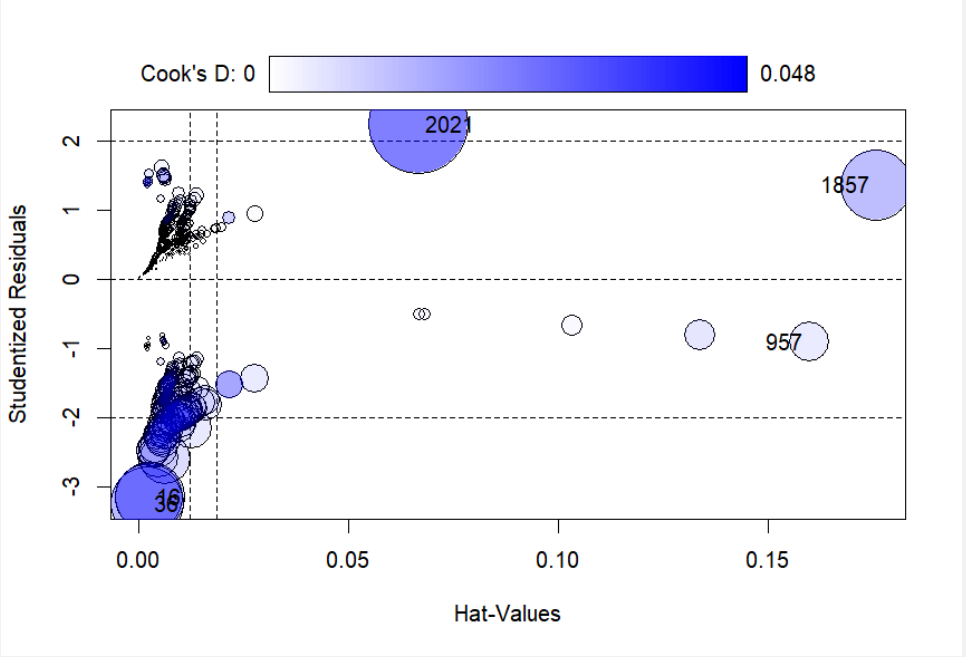


ASSESSMENT METRICS

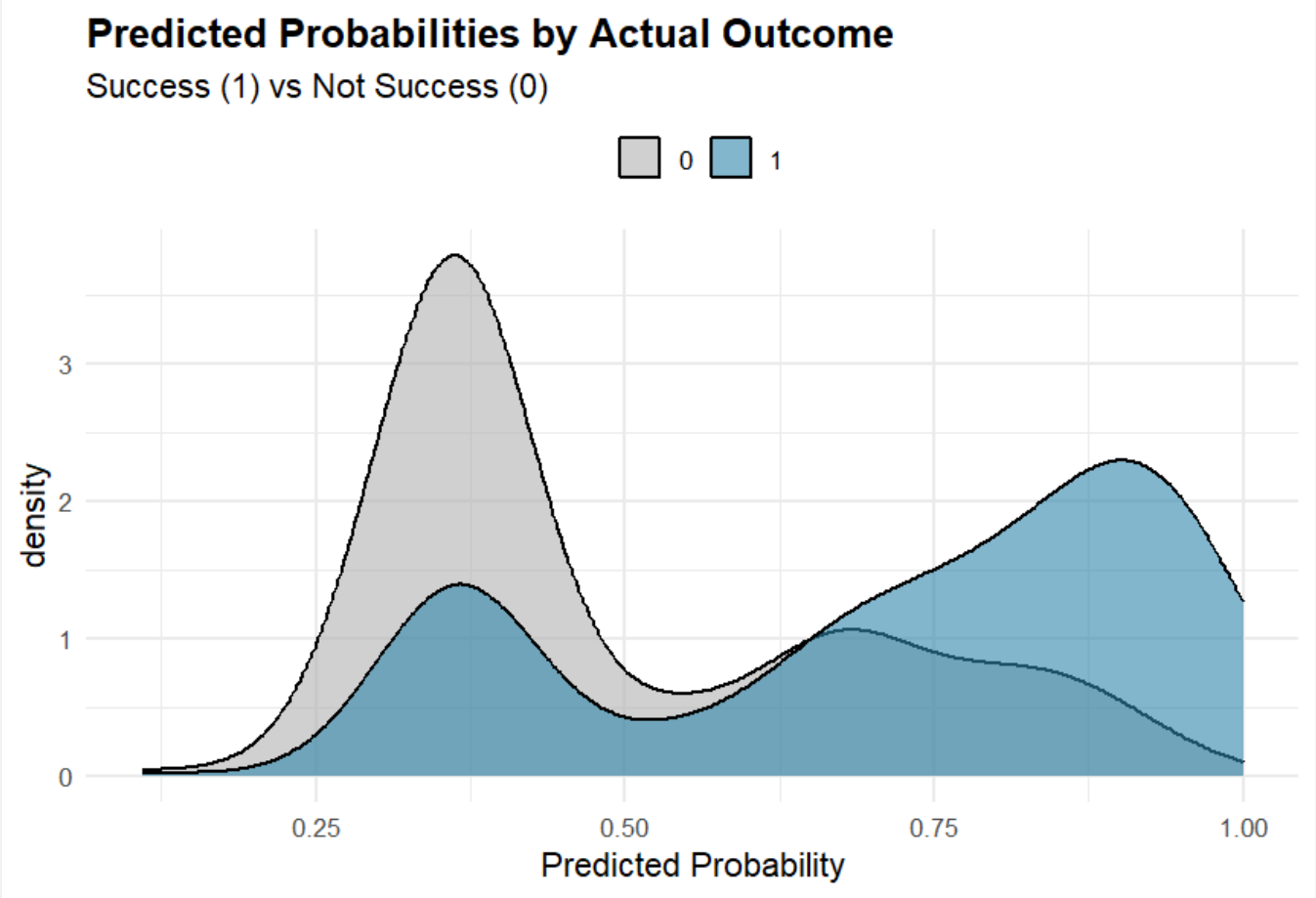
- Accuracy*: 72.16%
- F1 Score: 0.7816
- AIC: 2180

MODEL ASSUMPTIONS

- DV is binary
- Independent Observations
- VIFs all under 2 → no multicollinearity
- Influential Observations: 116 categorized as "influential"
  by Cook's Distance
- Linearity: No truly continuous variables



FINAL VISUALIZATION



FINAL VISUALIZATION EXTENDED

'Cool Case' Player Predictions

PLAYER	PREDICTED PROBABILITY	WEEKS ACTIVE	DRAFT ROUND	SIDE OF BALL	DRAFT AV	# COLLEGES	DRAFT AV IMPUTED?
Tyler Loop	90.8%	14	6	ST	0	1	No
Jalen Milroe	85.3%	4	3	Off	0	1	No
Shedeur Sanders	60.2%	8	5	Off	0	2	No
Darius Cooper	35.3%	10	Undrafted	Off	10	1	Yes
Ben Yurosek	27.2%	9	Undrafted	Off	7	2	Yes

Predictions from Logistic Regression Model

LIMITATIONS

- Data Quality
  - NFL Combine data
  - Missing data when joining
- Influential Observations
- Draft AV tries to account for on-field results
- Possible Interaction Effects

CONCLUSIONS

- VERY difficult to predict college to professional translation
- Important Variables
  - Draft Round
  - Side of Ball
  - Draft AV
  - Number of Colleges Attended

THANK YOU! QUESTIONS?