W207 08/2024

Using NFL Combine Data to Predict Player Draft Status

Jenna Farac, Surabhi Gupta,
Aditya Kumar, Seema Vora

Introduction

- Data: NFL Combine Data from 2009-2019 via Kaggle but sourced from https://www.pro-football-reference.com/
  - 18 columns , 3,477 rows
  - Inclusive of player background info (school, position, year,age, height, weight)
  - Most columns include player combine performance
  - Also includes draft status and round they were picked if applicable
- Goal: Use our dataset to train a model that can appropriately predict whether or not a player was drafted to the NFL
- Motivation: Wanting an objective second opinion on whether a player should be drafted or not purely based on the NFL Combine data. Also to assess if the provided combine data is enough to determine if a player will be drafted or not

Use Cases for Our Model

- Prepare athletes for the NFL combine by identifying how their stats stack up against players who have historically been drafted
- Highlight specific players early on - can help teams identify players they may not have noticed but should take a closer look at
- Provide context on an athletes future which is valuable information to leverage in contract negotiations for NIL for example
- Inform draft strategy if taken a step further by predicting the round a player will be drafted - allows coaches to prepare back-up plans for if a player is chosen before their turn

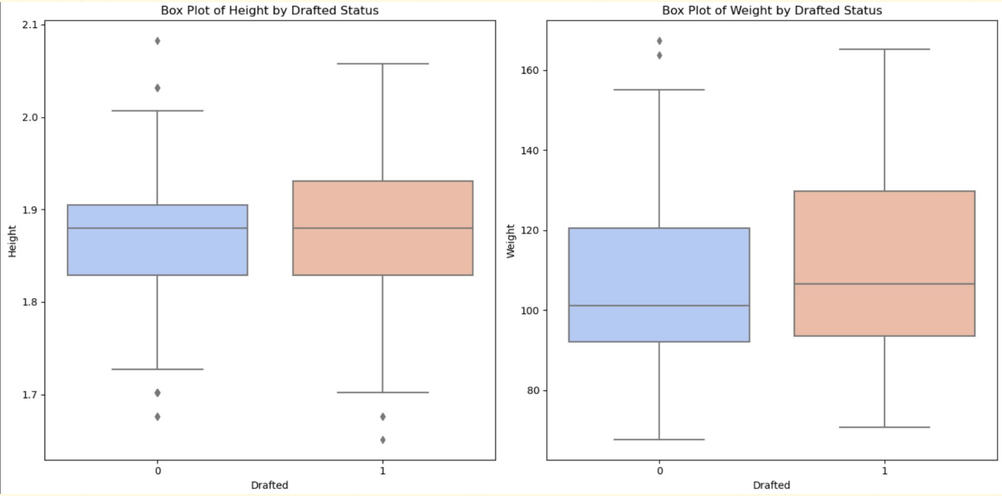
Data Exploration

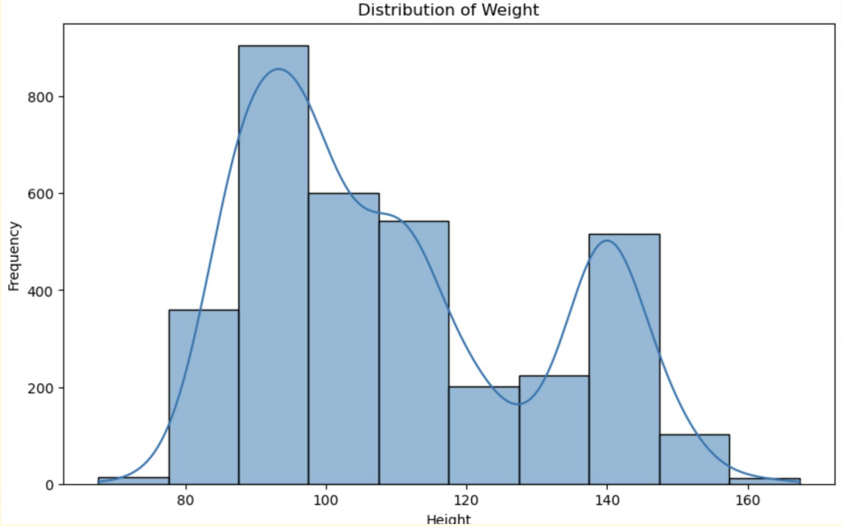
Data Exploration: Data Overview

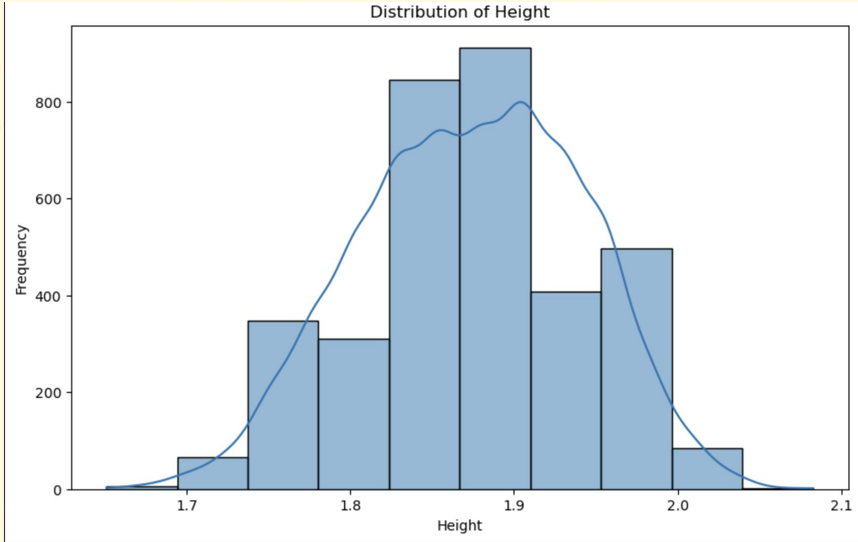
- Dataset has 3477 values
- We will be using "Drafted Column" as our Y-value
- The features below will be used to predict if a player was drafted or not.

	Year	Age	Height	Weight	Sprint_40yd	Vertical_Jump	Bench_Press_Reps	Broad_Jump	Agility_3cone	Shuttle	BMI
count	3477.000000	2927.000000	3477.000000	3477.000000	3303.000000	2780.000000	2572.000000	2749.000000	2260.000000	2337.000000	3477.000000
mean	2013.823699	21.983259	1.873968	109.746393	4.769080	83.392403	20.241058	291.629698	7.237416	4.403843	31.074417
std	3.075616	0.969490	0.067494	20.483780	0.301477	10.678403	6.497600	23.960879	0.410230	0.265224	4.438279
min	2009.000000	18.000000	1.651000	67.585263	4.220000	44.450000	2.000000	198.120000	6.280000	3.810000	21.609798
25%	2011.000000	21.000000	1.828800	92.986436	4.530000	76.200000	15.000000	276.860000	6.940000	4.200000	27.475641
50%	2014.000000	22.000000	1.879600	104.779837	4.690000	83.820000	20.000000	294.640000	7.140000	4.360000	30.122626
75%	2016.000000	23.000000	1.930400	125.645087	4.960000	90.170000	25.000000	307.340000	7.490000	4.560000	34.038647
max	2019.000000	28.000000	2.082800	167.375585	6.000000	114.300000	49.000000	373.380000	9.040000	5.560000	44.680097

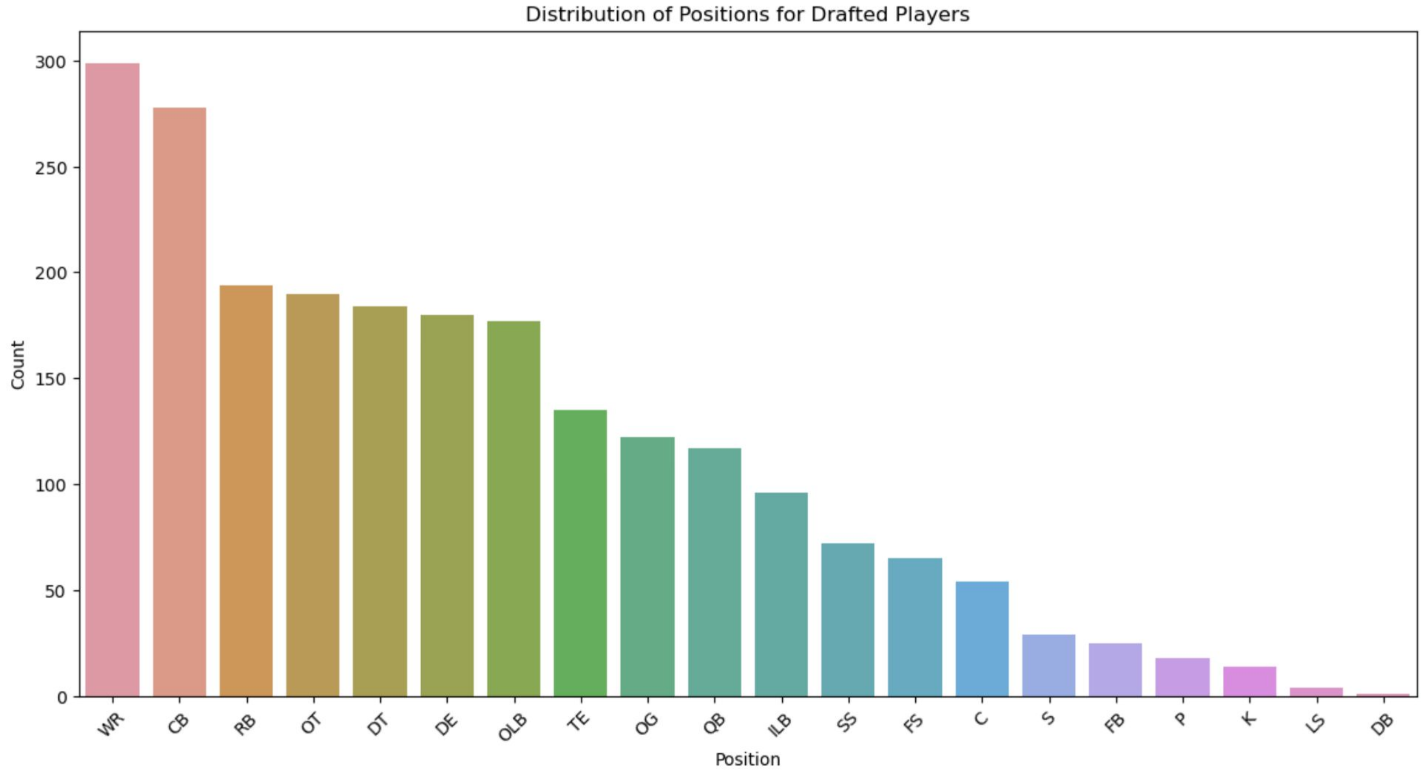
Data Exploration: Height & Weight



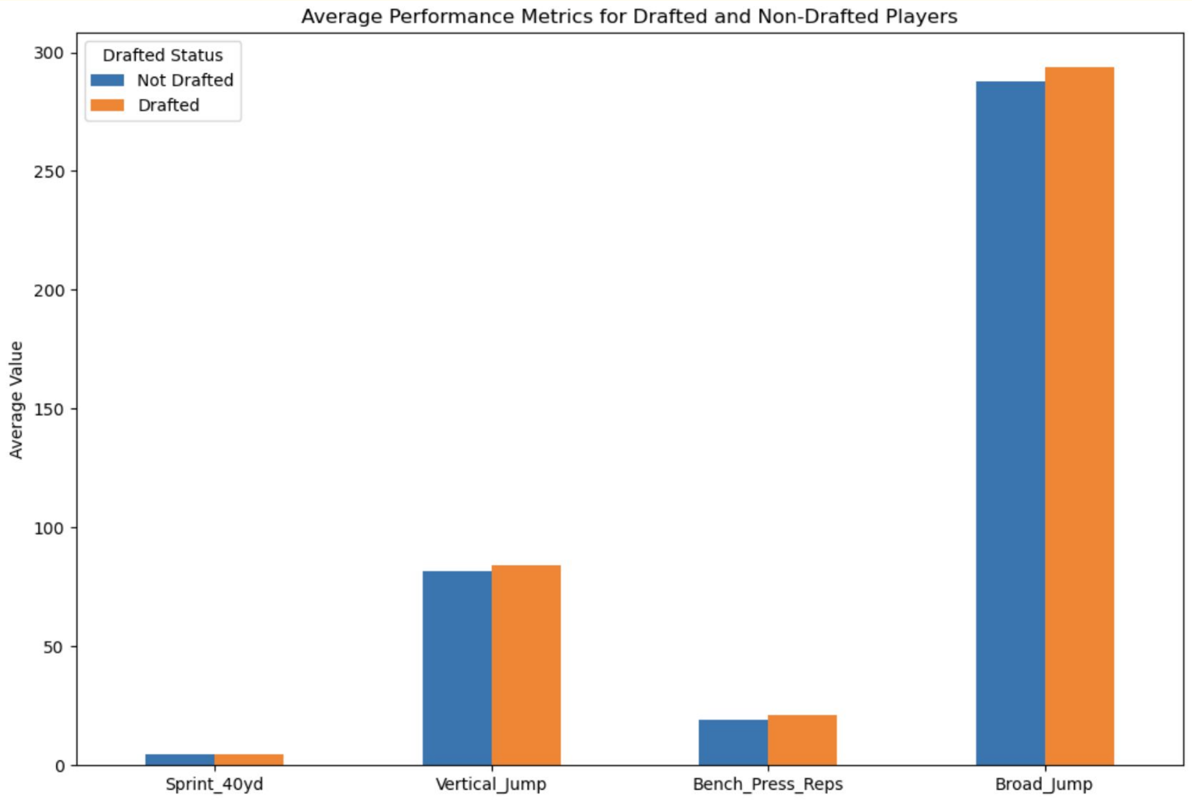




Data Exploration: Player Distribution

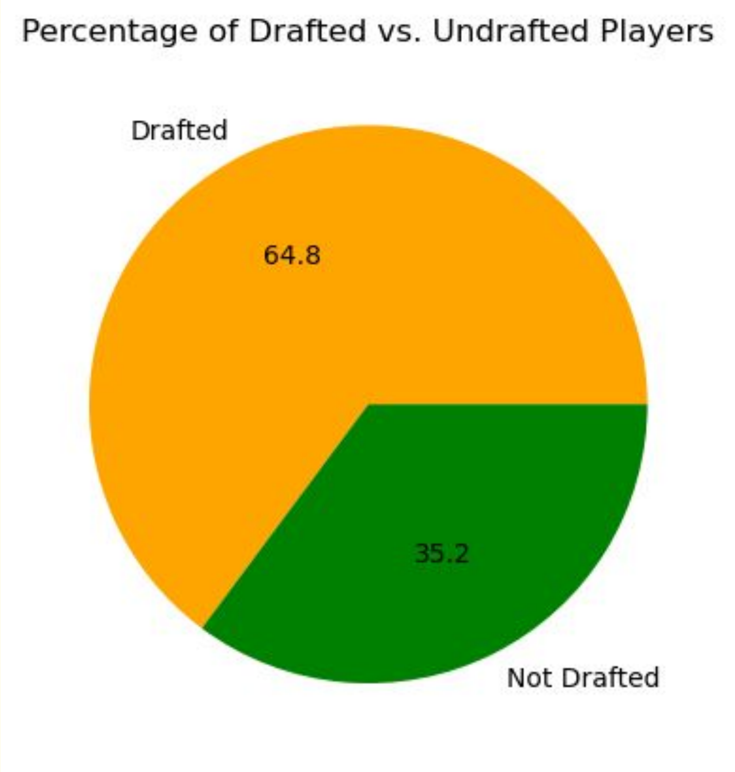


Data Exploration: Average Performance Drafted vs. Not



Data Pre-Processing: Resampling

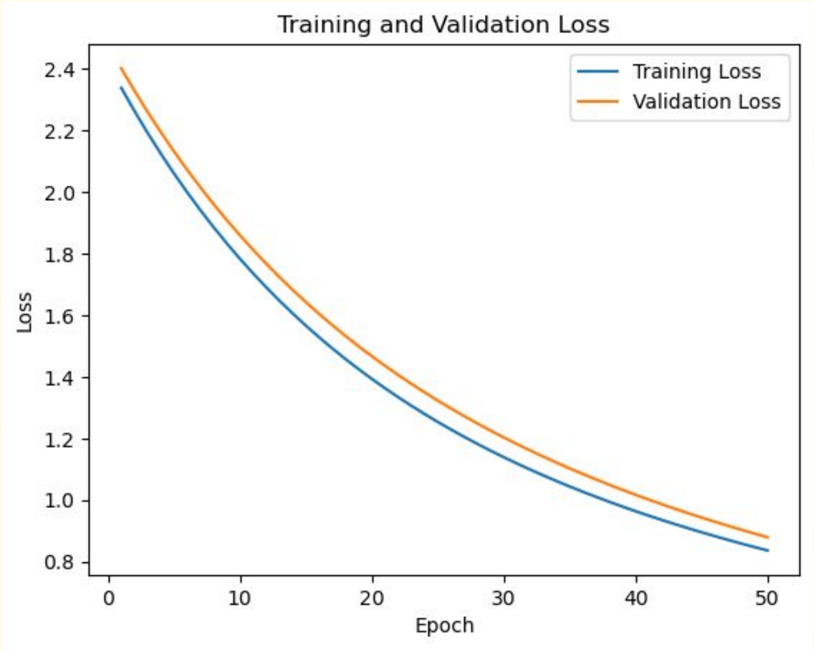
- After initial data exploration we found our data had far more samples of athletes that were drafted than not drafted
- To avoid bias in our model and help improve generalizability, we resampled from the not drafted class until our dataset was 50-50 drafted and not drafted



Models & Experiments

Model 1a: Linear Regression With 3 Features

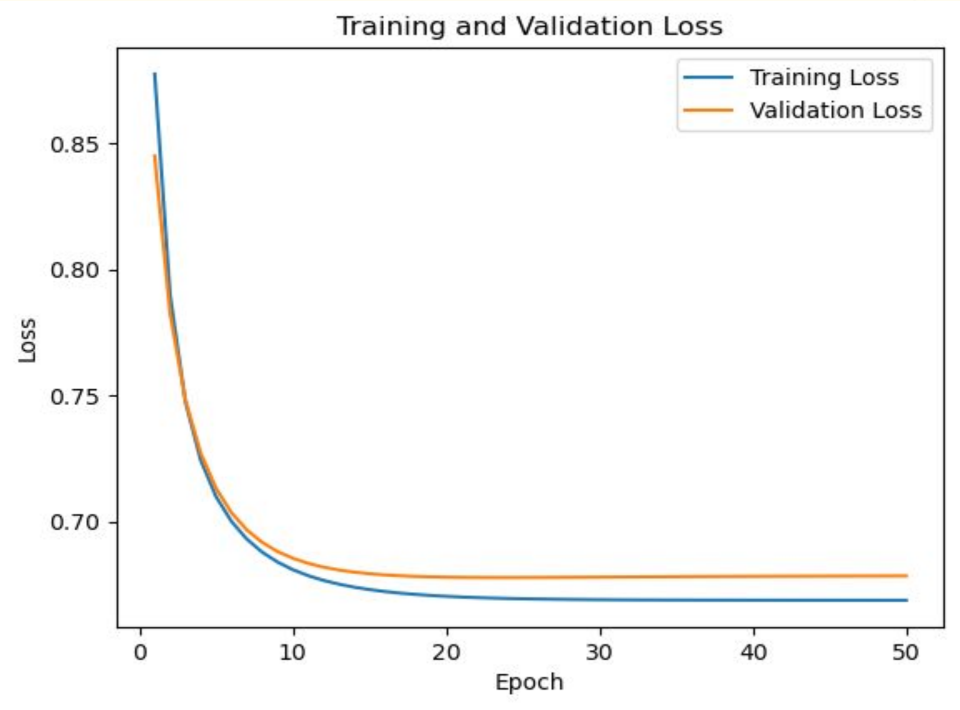
- We decided to start off with a fairly simple model using only 3 columns from the dataset
- Height, Sprint_40yd, and Vertical_jump
- We felt these showcased different components of a strong football player and were not too strongly correlated
- The model consists of a single dense layer and a sigmoid activation
- Utilized stochastic gradient descent as the optimizer and binary cross entropy as the loss function
- Epochs: 50
- Learning rate: 0.001




```
Training Accuracy: 0.5182
Validation Accuracy: 0.4960
Test Accuracy: 0.4974
```

Model 1b: Logistic Regression With 3 Features

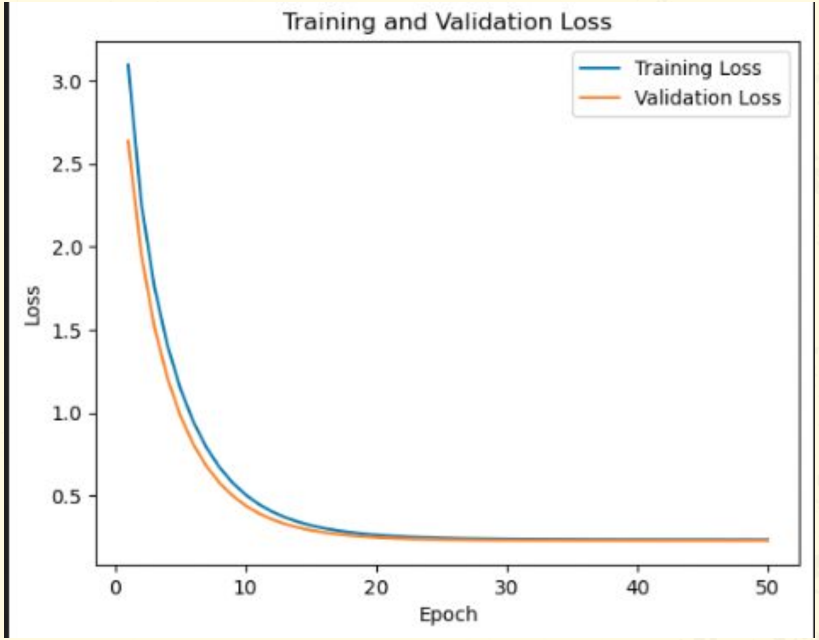
- The model consists of a single dense layer and a sigmoid activation
- Utilized stochastic gradient descent as the optimizer and binary cross entropy as the loss function
- After adjusting the number of epochs and learning rate we decided to use 50 epochs and a learning rate of 0.02




```
Training Accuracy: 0.5777
Validation Accuracy: 0.5721
Test Accuracy: 0.5625
```

Model 2a: Linear Regression With 5 Features

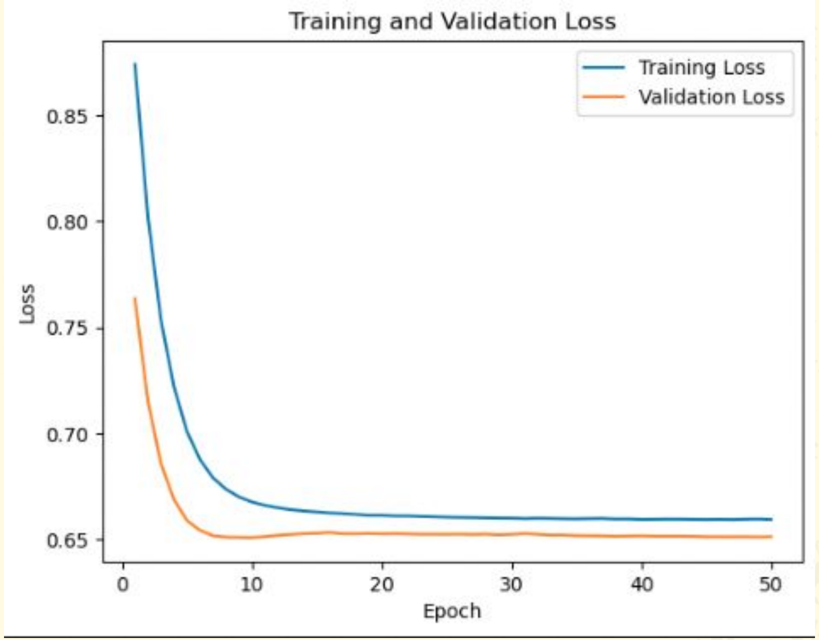
- Height, Weight, BMI, Vertical_Jump, Broad_Jump
- Assumed based on the analysis above, that there was some correlation between these fields.
- In football jumping is a critical skill and height/weight/bmi is a great indicator of overall athlete health and rigor.
- Used SGD optimizer




```
Training Accuracy: 0.6071
Validation Accuracy: 0.4587
Test Accuracy: 0.5963
```

Model 2b: Logistic Regression with 5 Features

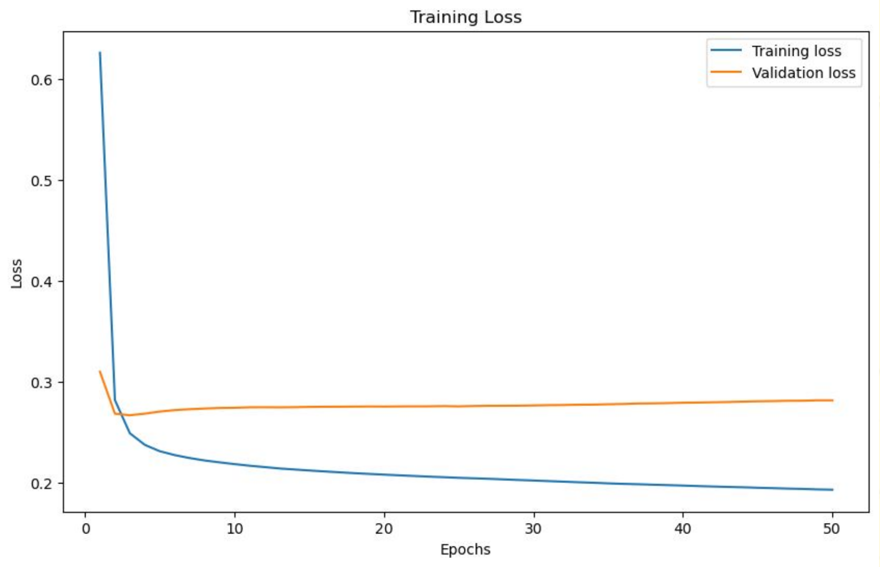
- Height, Weight, BMI, Vertical_Jump, Broad_Jump
- Activation: ‘sigmoid’
- SGD optimizer
- Batch Size: 8
- Learning Rate: 0.01




```
Training Accuracy: 0.6000
Validation Accuracy: 0.6300
Test Accuracy: 0.5933
```

Model 3a: Logistic Regression with 5 Features

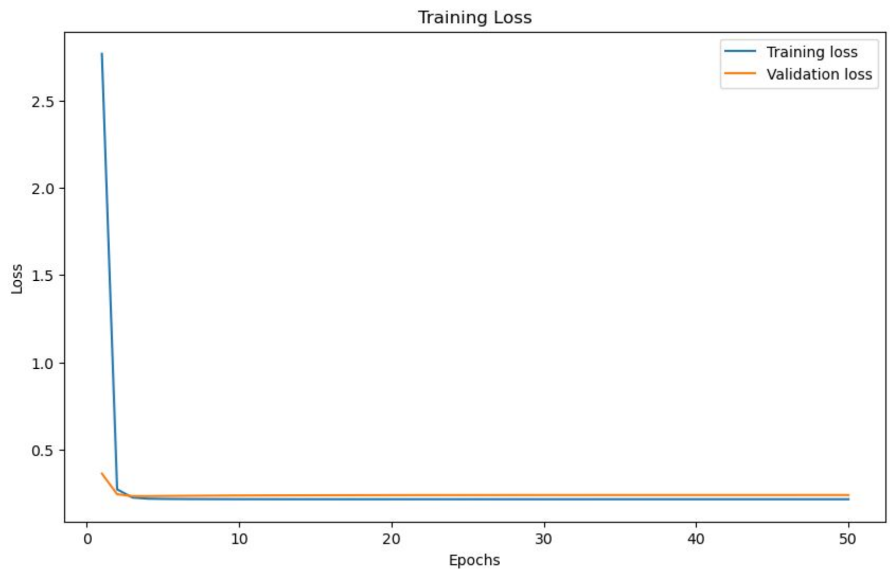
- Features: 'Age', 'Bench_Press_Reps', 'Agility_3cone', 'BMI', 'Shuttle'
- Two Dense layers were added where the activation function is of type Relu.
- Optimizer is Adam.
- Overfitting




```
Training Accuracy: 0.7094
Validation Accuracy: 0.6047
Test Accuracy: 0.6287
```

Model 3b: Logistic Regression with 5 Features

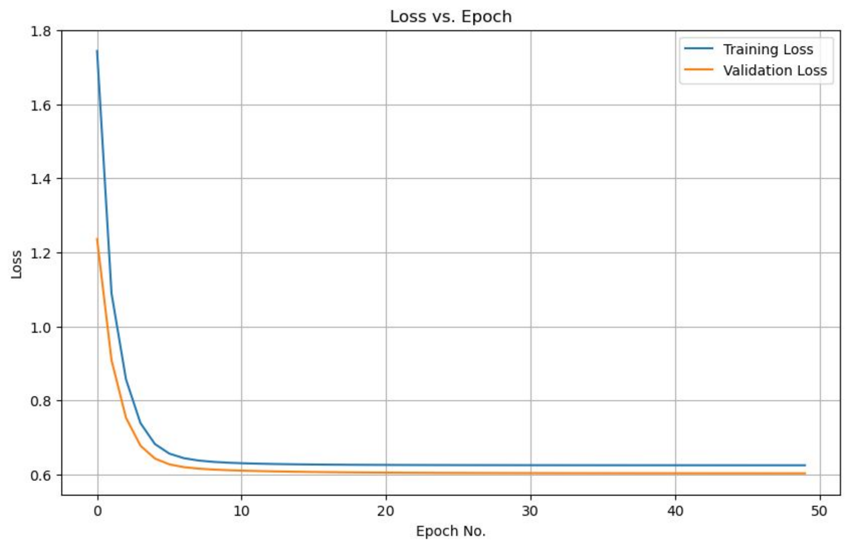
- Features: 'Age', 'Bench_Press_Reps', 'Agility_3cone', 'BMI', 'Shuttle'
- One Dense Layer were added with optimizer as SGD.
- The train and test accuracy are similar suggesting decent performance and limited overfitting




```
Training Accuracy: 0.6332
Validation Accuracy: 0.5980
Test Accuracy: 0.6107
```

Model 4a: Logistic Regression with all Features

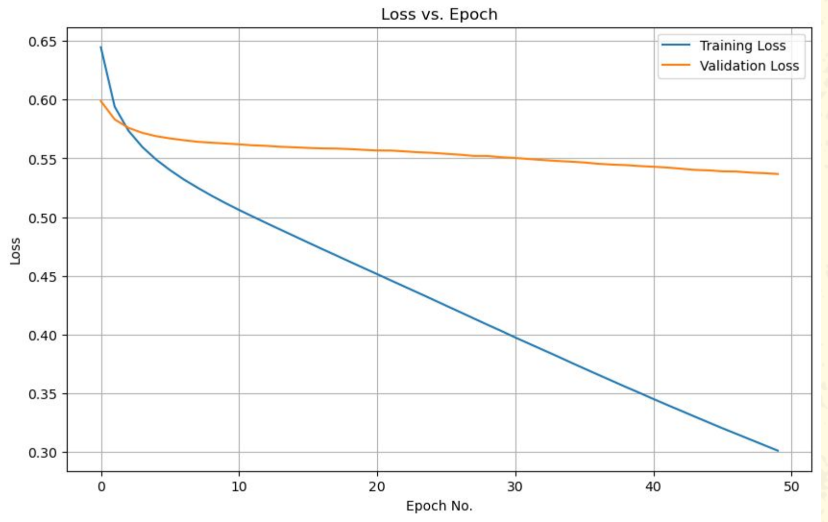
- Features: 'Year', 'Age', 'School', 'Height', 'Weight', 'Sprint_40yd','Vertical_Jump', 'Bench_Press_Reps', 'Broad_Jump', 'Agility_3cone','Shuttle', 'BMI', 'Player_Type', 'Position_Type', 'Position', 'Name','ID'
- One Dense Layer was added with optimizer as SGD.
- This model performed slightly better than all other models despite having a lot of features.




```
Training Accuracy: 0.6395
Validation Accuracy: 0.6830
Test Accuracy: 0.6297
```

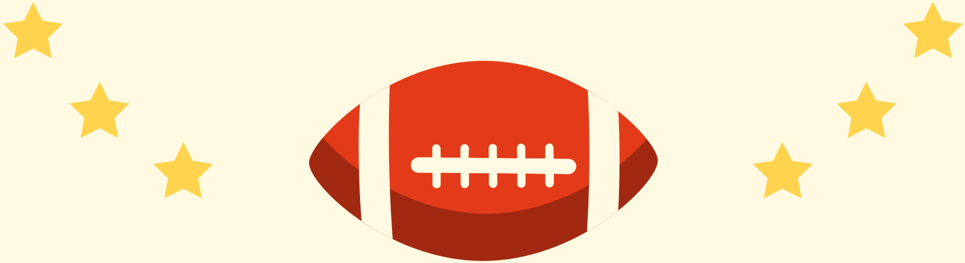
Model 4b: Simple Neural Network with all Features

- Features: 'Year', 'Age', 'School', 'Height', 'Weight', 'Sprint_40yd','Vertical_Jump', 'Bench_Press_Reps', 'Broad_Jump', 'Agility_3cone','Shuttle', 'BMI', 'Player_Type', 'Position_Type', 'Position', 'Name', 'ID'
- One Dense Input Layer with Relu activation function and one Output Layer with Softmax activation.
- This model performed the best as compared to all other models.




```
Training Accuracy: 0.8946
Validation Accuracy: 0.7591
Test Accuracy: 0.7531
```

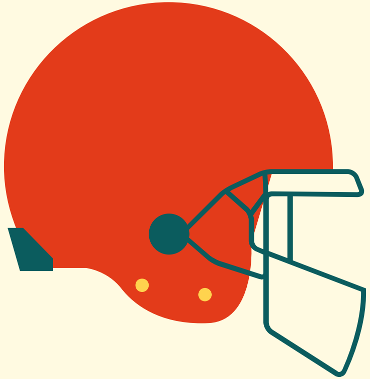
Results



Accuracy of All Models

	Model	Training Accuracy	Validation Accuracy	Test Accuracy
0	AK Neural Network	0.894612	0.759113	0.753141
1	AK Logistic Regression	0.639461	0.683043	0.629712
2	AK CNN	1.000000	0.800317	0.791574
3	AK Random Forest	0.999208	0.866878	0.852180
4	JF Linear 1	0.540016	0.564184	0.538803
5	JF Linear 2	0.579239	0.573693	0.560237
6	JF Random Forest	0.943344	0.749604	0.764967
7	SV Linear 1	0.604082	0.458716	0.599388
8	SV Linear 2	0.569388	0.562691	0.532110
9	SV Logistic 1	0.602041	0.636086	0.590214
10	SG Linear 1	0.633267	0.598802	0.610778
11	SG Logistic 1	0.685371	0.592814	0.622755

Neur IPS Checklist



Accurate Scope?

- Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope
  - Yes, our goal was to use the data to predict draft status based on players combine metrics and that is what we did

Ethics

- Bias and Fairness:
  - Risk of discrimination: Possible bias that could favor certain physical attributes, schools etc. over the objective details
  - Model fairness: The model might have inherent bias towards certain groups of players and this will impact the results on drafting
- Misuse of Data:
  - This dataset has personal information that could be misused to harm the players, such as by impacting the hiring decision
  - Instead everything should be transparent and usage of data must be clear

Potential Negative Impacts

- Player psychological pressure and negative impact
  - Players can be held to a higher more difficult standard as a result of the model.
- Ethical/Privacy Concerns
  - The data may have been collected in a span of a day and players may not want an “off day” shared and used to predict their draft status.
  - Players simply may not want their health data shared
    - BMI
    - Height
    - Weight

Limitations

- Data was limited to only 2009-2019
  - As we see in the Olympics, athletic rigor is often challenged and improved.
  - Data points could be outdated.
- Our model did not account for the school the player went too
  - Certain schools with better football programs have a higher visibility
- Our data is specifically from colleges, instead of the NFL.
  - NFL typically is looking for certain players - which may increase/decrease likelihood of being drafted
- Style of Play
  - You can have good stats but not good team compatibility
- Based on a limited time frame of training
  - The combine is only a few days so if a player was for example sick but has a great college season, our model doesn’t account for historical performance

Conclusion

Model Limitations

- Are we using too old of data? Data includes value from 2009 to 2019 but can we assume that there has been a large change in performance over time with advancements in training/recovery programs
  - Would mean we are training our model that is not reflective of our current reality but we are limited by the combine only occurring once per year

- Does our data capture enough of the context necessary to understand if a player will be drafted or not- does it need to be enhanced by team level data?

- The combine does not include full scrimmages rather just drills such as sprints and jumps and some specific position drills - this may not be sufficient to predict draft status

Next Steps/Looking Ahead

- Rather than only looking at whether or not a player was drafted, we could look at predicting what round a player was picked
- By integrating team data from the previous season, we could identify key context that motivates draft picks
  - Did any players retire?
  - Did they team make any key trades?
  - What players had a particularly bad/good season? AKA what positions are teams in need of
  - Historical draft data - does this team go for a particular type of player
- Could also allow us to predict not only what round a player is picked, but also what team will likely pick what player

Thank You

Appendix

References

- https://www.kaggle.com/datasets/redlineracer/nfl-combine-performance-data-2009-2019/data

Team Contributions

- Jenna Farac: Built graphs for data exploration, built 2 unique models, created slides, team meetings
- Surabhi Gupta: Built graphs and models, contributed to presentation, team meetings
- Aditya Kumar: Built graphs and 4 unique models, contributed to presentation, team meetings
- Seema Vora: Built graphs for data exploration, built 2 unique models, created slides, team meetings