Prediction of Player Draft Success and Longevity in NFL

Sayed Jaber Hossaini, Lauren James, Daniel Lyczak, Likhit Nayak, Meiyun Xiao

July 28, 2023

1 Introduction & Background

The National Football League (NFL) features 32 teams, each with a 53-man roster. The majority of players enter the NFL through the annual draft. To be eligible for the draft, players must be three years out of high school and invited to the regional and national combine, where their physical and mental abilities are measured. With an average career length of 3.3 years, and the 2023 league minimum salary of $750,000 per year, selecting the right player in the NFL draft has immediate and long term costs and expected payoff for teams. Player selection is predicated on combine data, but due to the uncertainty and importance of choosing a player, teams are invested in additional means and models to best ensure an optimal draft selection.

Recent work on this subject [1] found historical combine performance data is selectively useful; only adequately predicting the draft-efficacy of a player for specific positions, while also calling into question the use of the Wonderlic test, the NFL's standard measure of mental acuity. Still, [1] highlights the growing demand and interest in models to predict success (which they define as draft-ability), despite their historically spurious usefulness. Other studies which looked only at the quarterback position (often the leader of the team's offense) and how the Wonderlic cognitive test predicts success concluded that while not able to reliably predict a player's draft-ability, test scores do have a statistically significant impact on player productivity [2]. The strategic use of collegiate variables provide a useful framework for inclusion in the larger model proposed here while also helping to outline what success measurements can and should be used [2].

Currently, the two main contributors to ending an NFL career include not finding a team to play for (due to a lack of success [2]) and having a major injury. Studies looking at whether combine data can predict future injury conclude that there are combine measurements which have a statistically significant impact on the prediction of a player's physical longevity in the league [3]. This provides validation to our approach of considering the impact of injury on ending an NFL career and using pre-draft testing to predict the relative success of a player drafted.

The new approach proposed here is considering combine data, physical and mental attributes, and collegiate data (both college and player level data) to build comprehensive models to predict a successful player, measured as one with high draft-ability and one with a career extending beyond the average by avoiding injury and maintaining productivity.

2 Problem Definition

Can a player's performance in the NFL scouting combine be used as an accurate indicator of the player's draft-ability and longevity in the NFL?

2.1 Potential Problems

Sports at the professional and collegiate levels are profitable, making free data-collecting difficult. Statistics of players and teams are often behind paywalls, or, when free to access, limited in scope or depth. To help overcome this, we will work to leverage myriad data sources and consolidate those sources.

Consolidation is in itself a problem. Players across their career are not assigned a universal ID or primary key. Additionally, different sources will contain different features, making it difficult to create a consolidation key. To deal with this, we will leverage an iterative consolidation practice wherein the concept of superkeys in entity relationships will be used to ensure accuracy. When accuracy cannot be assured, disparate data sources will not be consolidated for that record.

When a homogenous data file is created, there is still potential for asymmetric and missing data. Two approaches can be taken to reconcile this. First, only records with data will be considered for model building and Second, a K Nearest-Neighbors approach to impute missing values. As each player will only be considered in relation to other players in the same position, a KNN imputation approach can help reconcile incomplete data strategically.

An additional problem is in the disparity between player positioning and their subsequent metrics. For example, a Quarterback will have data on throwing strength, which will be absent in other positions. But, a Quarterback is unlikely to have combine data on the bench press. To help overcome this, we will build models for each position individually and independently.

3 Data Collection

The data collection process included web scraping, API calls, and direct downloads. Web scraping was accomplished using the 'BeautifulSoup' library in Python. API calls were made through two methods; the "nfl_data_py" library in Python and through an API key.

Website	Method
www.pro-football-reference.com	Direct Download
nflcombineresults.com	Web Scraping
pypi.org/project/nfl-data-py/	API
api.collegefootballdata.com/api/docs/	API

Table 1: Data Sources and Acquisition Methods

After the data was collected, when necessary, the individual sources were concatenated. Due to the data collection process, some sources provided data by the year as separate files, necessitating a consolidation of records for a single source of all years.

When all data sources were compiled into an individual file, a merging process took place to create a single dataset. The draft information of a given player, specifically year drafted and draft number, was used to form a primary key to match records. Once all records that could be matched this way were, an iterative matching process was created to match the remaining records. A player's full name was stripped of special characters (such as "." and " ' ") and suffixes were removed as their inclusion was not consistent across sources. This cleaned name was then combined with the player's football position and college team, which was then stripped of spacing and set to all lower case. This attribute served as a primary key to match records across sources while being mindful of duplicate keys or logical inconsistencies. Logical inconsistencies, for example, include a player who started college in 2014 but was said to be drafted in 2010. When a player could not be matched with high certainty, data from that source was set to Null. This resulted in a single data file consisting of all previous sources. This data file was used for all subsequent testing.

4 Methods

4.1 Data pre-processing

The dataset consisted of NFL combine statistics for players from 2000 to 2022. The players were grouped into the following positions: Defensive Back (DB), Defensive Line (DL), Quarterback (QB), Wide Receiver (WR), Offensive Line (OL), Linebacker (LB), Running Back (RB), and Tight End (TE). For each player position, samples with any missing or incomplete combine data were removed.

4.2 Unsupervised Clustering

For prediction of draft-ability, we trained three different unsupervised clustering models for each player position: K-means clustering, hierarchical clustering, and Gaussian mixture models (GMM). The models were trained to cluster athletes into two groups ($k = 2$) and were

Player position	No. of drafted players	No. of undrafted players	Total players
DB	142	53	195
DL	174	49	223
QB	65	30	95
WR	118	72	190
OL	196	60	256
LB	92	34	126
RB	70	40	110
TE	70	23	93
Total	**927**	**361**	**1288**

Table 2: The number of drafted and undrafted players with complete data for each player position.

evaluated based on how accurately each group represented the pool of drafted and undrafted players. The initial feature vector consisted of nine measurements. This included age (yrs), height (in), weight (lbs), hand size (in), wingspan (in), 40-yard dash time (sec), vertical leap (in), 20 yard shuttle time (sec), and three cone drill time (sec). Principal component analysis (PCA) was performed on the data to identify the directions of maximum variance. The first three principal directions were selected as the feature vectors for unsupervised clustering and explained 98.1%, 99.5%, 98.0%, 98.9%, 98.0%, 97.3%, 99.0%, and 97.3% of the total variance for DB, DL, QB, WR, OL, LB, RB, and TE respectively. Table 2 shows the number of drafted and undrafted with complete data for each position, consisting of a total of 1288 players.

K-means algorithm clusters the data by minimizing the sum of squared distances between each sample in the data and the mean of its corresponding cluster for a predetermined number of clusters. Hierarchical clustering assigns each sample to its own cluster and proceeds to merge the clusters by minimizing the same distance as the K-means algorithm. The difference between the two algorithms is that the hierarchical method starts with same number of clusters as the number of samples and progressively decreases it till it reaches the given number of clusters. The K-means method, on the other hand, starts directly with the given number of clusters and keeps updating the cluster centers. The GMM model assumes that the samples in the data are generated from a given number of Gaussian distributions. It then attempts to find the parameters associated with those distributions using an expectaion-maximization (EM) algorithm. Once the models were trained to split the data into two different clusters, the Fowlkes-Mallows index (FMI) was used to determine the similarity between the cluster assignments and the ground truth labels. The FMI score, FM, is given by

$$FM = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}}, \tag{1}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. The FMI score ranges from 0 to 1, with a higher value signifying better performance of the clustering algorithm. This score was calculated for each clustering algorithm for each player position.

4.3 Supervised Regression

For prediction of longevity, only players drafted into the NFL were considered. The data was split into training (70 percent) and testing (30 percent) sets for each player position, and the training set will be used for three supervised regression models: multivariate linear regression, multivariate logistic regression, and support vector regression. The ground truth for the regression models would be given by $o, o \in \mathbb{R}, o \geq 0$, denoting the number of years an athlete played in the NFL. All models would be implemented using scikit-learn in Python. The choice of classification and regression models were based on previous success in predicting performance of NFL players [4, 5] and soccer players [6] as well as predicting game outcomes for NFL [7] and NBA [8].

To test the quality across our chosen model structures, the metrics of Mean Squared Error (MSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) were then calculated. The objective function in regression is to minimize the difference between the true values and the model's predicted values. MSE and MAE measure this difference. The coefficient of Determination measures the proportion of total variation the model is able to explain. Unlike MSE and MAE, the coefficient of determination with the maximum value (ceteris paribus) is preferred, but for reasons beyond the scope of this paper, it is not attempted to be maximized.

An initial multivariate linear regression model with all numeric features for each position was created. This served as a benchmark to see what additional methodologies and model structures improve accuracy in predicting career longevity.

To help improve model accuracy, a lasso regression was then run for each position. Lasso works by implementing a penalty to shrink and reduce a coefficient's value towards zero. Lasso aids in feature selection by nullifying (setting a coefficient equal to zero) features whose inclusion in the model does not contribute to the improvement of the model. The alpha used for each lasso regression was calculated by doing a 10-fold cross validation. The resulting coefficients for that regression were recorded, and only those with an absolute value greater than zero were put into a new regression model. The full list of features with their corresponding coefficients can be found in Table 3. A new multivariate linear regression, with only the features identified in the lasso regression (uniquely identified for each position), was run, and the accuracy measures were recorded.

Feature	DB	DL	LB	OL	QB	RB	TE	WR
Class	0.0	-0.042	0.0	-0.117	0.0	0.0	0.0	0.0
Round	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Pick	-0.004	0.0	0.008	-0.001	0.003	-0.002	-0.001	0.001
Age	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Hght	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Wght	-0.025	0.004	0.0	-0.024	0.017	-0.007	0.008	0.017
BMI	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Arm len	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Hands	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Span	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
x40Yd	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
x20Yd	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
x10Yd	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bench	-0.004	0.0	0.0	0.0	0.0	0.068	0.0	0.017
Vert	0.0	0.0	0.0	0.0	0.0	0.022	0.0	0.0
Jump	0.0	0.0	0.051	0.004	-0.009	0.0	0.0	0.0
Shuttle	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
x3_Cone	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
wAV	-0.007	-0.014	0.025	0.0	0.0	0.0	0.0	0.016
DrAV	0.0	-0.006	0.0	-0.019	-0.005	0.0	0.0	0.0
HOF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 3: Resulting Coefficient Values From Lasso Regression (rounded to 4 decimal places)

Finally, Support Vector Regression (SVR) models were built for each position. Unlike traditional regression, SVR identifies a hyperplane instead of a line to best fit the data. Using SVR allows for better modeling of non-linear relationships between variables, allowing for more flexibility. SVR models allow us to specify a kernel, where the kernel determines how to transform the data. We built two additional models for each position using SVR, each with a unique kernel; a Gaussian Kernel Radial Basis Function (RBF) and a Polynomial Kernel. Only features previously identified in the corresponding positions' lasso regression were inputted into the SVR models.

5 Results

5.1 Unsupervised Clustering

The FMI scores of all the unsupervised clustering algorithms for each player position is compared in Figure 1. The K-means and hierarchical clustering algorithms perform best for offensive line (OL) position with FMI scores of 0.647 and 0.718 respectively. The GMM

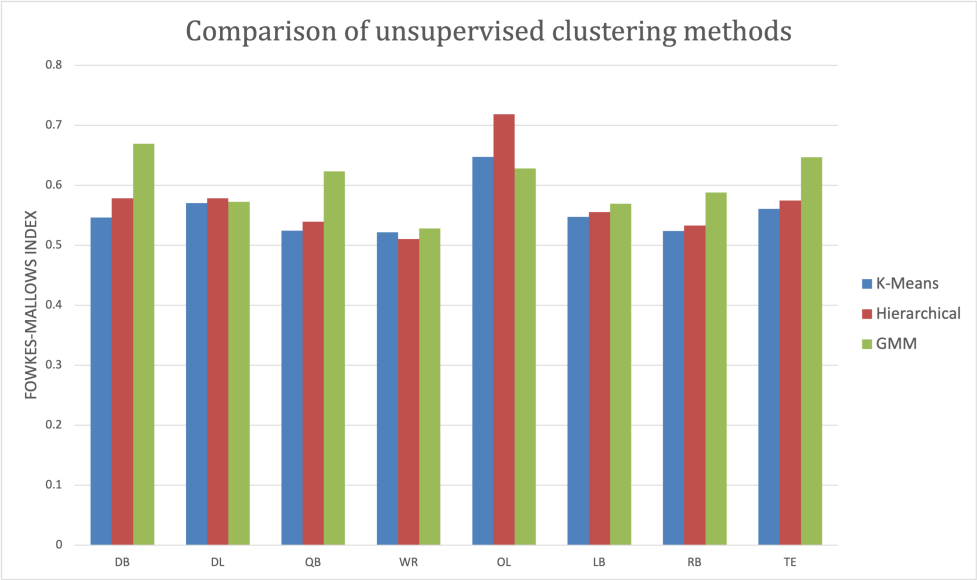

Figure 1: Graph showing comparison of Fowlkes-Mallows index of each unsupervised clustering algorithm for each player position.

algorithm shows the best performance for the defensive back (DB) position with FMI score of 0.669. The GMM algorithm outperforms both K-means and hierarchical clustering for six positions, namely, DB, QB, WR, LB, RB, and TE (Table 4). Hierarchical clustering gives the highest FMI scores for the remaining two positions of DL and OL (Table 4). The OL position shows the highest similarity between the cluster and actual labels, while the WR position shows the lowest similarity (Table 4).

Player Position	**Best FMI Score**	**Best Clustering Algorithm**
DB	0.669	GMM
DL	0.578	Hierarchical
QB	0.623	GMM
WR	0.528	GMM
OL	0.718	Hierarchical
LB	0.569	GMM
RB	0.588	GMM
TE	0.647	GMM

Table 4: The best performing FMI scores and corresponding clustering algorithms for each player position.

5.2 Supervised Learning

The relevant features (coefficients greater than zero) identified through the lasso regression and summarized in Table 3 for each position were used in each subsequent model. The resulting Mean Squared Errors (MSE), Mean Absolute Errors (MAE), and Coefficients of Determination (R2) were calculated. The results from MSE are displayed in Table 5, MAE are Table 6, and coefficients of determination are represented in Table 7.

Position	Full Feature	W/ Lasso	Poly SVR	RBF SVR
DB	0.8772	0.8909	1.2198	1.6967
DL	1.6397	1.1056	1.2841	2.2320
LB	67.4419	1.6957	2.6315	2.0006
OL	2.5242	2.4124	2.5657	4.0862
QB	70.5922	2.1672	3.5001	5.3642
RB	31.2825	1.7188	1.7033	2.5552
TE	37.9172	1.0687	1.6594	3.0214
WR	15.1037	1.4279	1.8125	2.3573

Table 5: Mean Squared Error Scores for Regression Models

Position	Full Feature	W/ Lasso	Poly SVR	RBF SVR
DB	0.6912	0.7252	0.8356	1.0396
DL	1.0936	0.8978	0.9332	1.2689
LB	5.8499	1.0067	1.1940	0.9895
OL	1.3978	1.3521	1.3429	1.7071
QB	6.6826	1.2136	1.6225	1.8071
RB	4.4332	1.0898	1.1605	1.3604
TE	5.0258	0.8870	1.0852	1.4246
WR	3.3559	0.9858	1.1293	1.3096

Table 6: Mean Absolute Error Scores for Regression Models

The accuracy and performance of the models vary by position. However, quite consistently, the linear model with the lasso coefficients performs either the best, or close to it. Most promising are the R-Squared values, which, with lasso, perform around 70 percent, meaning our model does a good job of predicting career longevity. Notably, the Running Back (RB) position has the worst overall model performance. This reflects prior research which identified only certain positions being able to accurately predict a player's longevity in the NFL. Overall, the select feature linear regression better predicts career longevity for non-skill positions (DB, DL, LB, OL), referring to those positions which do not contribute to a team's offensive scoring. This could indicate that combine testing is not the strongest indicator for success in skill positions (QB, RB, TE, WR), especially since most combine drills

Position	Full Feature	W/ Lasso	Poly SVR	RBF SVR
DB	0.7640	0.7603	0.6718	0.5435
DL	0.5772	0.7149	0.6689	0.4245
LB	-9.4202	0.7380	0.5934	0.6909
OL	0.4978	0.5200	0.4895	0.1870
QB	-10.6371	0.6427	0.4230	0.1157
RB	-10.1329	0.3883	0.3938	0.0907
TE	-6.8830	0.7778	0.6550	0.3718
WR	-2.2148	0.6961	0.6142	0.4983

Table 7: R Squared for Regression Models

involve strength and agility, which are at best proxies for attributes vital to skill positions.

Evaluating the features themselves, it's interesting to find that a player's pick (hierarchy in draft selection) is often relevant and detrimental to longevity. This could be due to increased wear on a player, as it is suspected that players who were most active in high school and college are most likely to be drafted higher (and having a predicted shorter career). Such notion of longevity prior to entry in the NFL warrants additional research. Additionally, the consistent inclusion of weight as a major factor is unexpected. For non-skilled positions like DBs and OLs, weight, which is often considered a merit of their respective positions, actually decreases predicted longevity. Conversely, skill positions, like QBs and TEs, benefit from more weight, perhaps signalling that there is an ideal target weight for those most needing agility.

The overall strong performance of the linear approach to predicting a player's longevity in the league warrants additional investigation, as these results point to an ability to early-on predict whether a drafted player will last in the league, which not only would influence draft selection, but also future casting of a team's roster.

6 Discussion

The trained machine learning models shown in this project serve to assist NFL teams in determining: 1) whether an athlete should get drafted for a particular position, and 2) the length of the athlete's NFL career. The models would be an invaluable resource for NFL scouting and front-office personnel to inform their draft strategy. For predicting the draft-ability of an athlete, the GMM algorithm shows the best overall performance. As shown in Figure 2, the ground truth clusters of all the drafted and undrafted players in the PCA space calculated in section 4.2 is not distinct. This suggests that the selected feature vector consisting only of the NFL combine statistics is not a good predictor of an athlete getting drafted into the NFL.

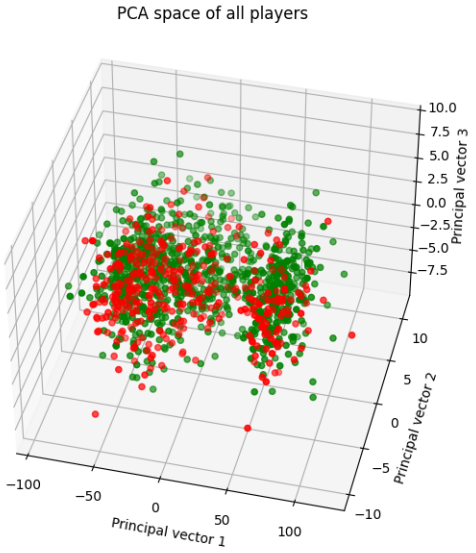

Figure 2: Scatter plot depicting 1288 players in PCA space of the feature vector consisting of height, weight, hand size, wingspan, 40-yard dash time, vertical leap, 20 yard shuttle time, and three cone drill time. The drafted players are depicted in *green* and the undrafted players are depicted in *red*.

Moving forward, the model could be updated to group eligible players into three categories based on the combined metric of draft success and longevity - green, yellow, or red. A green-flagged player would be a top draft pick, a yellow-flagged athlete may warrant additional evaluation, and a red-flagged athlete might be avoided completely. Using the model, NFL organizations can draft efficiently to address any gaps and improve their roster.

Future improvement to the supervised models can be attempted through adding to the feature vector. If more information, such as high school and college game day statistics are included, we can increase the positional accuracy of the models. Additionally, more dimensionality reduction techniques, such as Principal Component Analysis (PCA), can be used to work to improve the models and visualize their features. By better understanding pre-draft characteristics and features which influence a player's career longevity, the better informed a team will be when drafting players. Further, if the data can be changed to isolate why a player's career ended, either through injury or not finding a team, the overall strength and understanding gleaned from the coefficients will be increased. The results of which, if found, would not only inform a team of a player's draft-ability, but also aid in providing preventative care to a player in the hopes of prolonging their career.

7 Conclusion

While the dataset compiled for this research did an adequate job in predicting both the draft-ability and longevity of prospective and future NFL players, there is still room for improvement. The feature vectors heavily skewed towards combine data and player physical attributes. The results of feature selection in the supervised and unsupervised models differed, meaning that which makes a player more draftable does not also help predict their longevity. Further, the individual contributions of any given feature were not overwhelmingly so, meaning there is need for more desperate data, as is so often the case. The results of the supervised models indicate that the amount of past football experience may be negatively correlated with the length of a player's NFL career. Thus arises an interesting new question, valuable for teams and players alike; what is the optimal amount of time on the field for a player interested in a long and successful career in the NFL?

The constraints of this research led to a bifurcation of the research questions. A player's longevity in the supervised models was wholly independent of the unsupervised models on player draft-ability. There is reason to suspect that these are not independent events, as the supervised model showed a player's pick order in the draft helps predict longevity. Further, although a player is drafted, there is no guarantee, nor historically any direct correlation, between draft order and success in the NFL. In future studies, these models could be combined to try and assess a draft value to a player based on the likelihood of success in the NFL. The study presented here merely looked at the draft-ability with no further continuation married to that with their career outcomes studied in the supervised models.

The machine learning approaches in this study serve to validate existing research on the advocacy of pre-NFL data on player success. Given the explicit and intrinsic value of this research for the NFL and its players, the work done in this paper, coupled with additional data, can potentially lead to better understanding and quantification of what it means to excel in professional sports; moving from anecdotal to prescriptive statistics.

8 References

[1] Elia Rishis, Kathryn Johnston, and Joseph Baker. “On the predictive validity of the National Football League combine: does it forecast future success?” In: *Journal of Sports Sciences* (2023), pp. 1–15.

[2] JD Pitts and Brent Evans. “Evidence on the importance of cognitive ability tests for NFL quarterbacks: what are the relationships among Wonderlic scores, draft positions and NFL performance outcomes?” In: *Applied Economics* 50.27 (2018), pp. 2957–2966.

[3] Jordan Riley Pollock et al. “Can NFL Combine Results be Used to Estimate NFL Defensive Players Longevity?” In: *Sports Medicine International Open* 5.02 (2021), E59–E64.

[4] Masaru Teramoto, Chad L Cross, and Stuart E Willick. “Predictive value of National Football League scouting combine on future performance of running backs and wide receivers”. In: *The Journal of Strength & Conditioning Research* 30.5 (2016), pp. 1379–1390.

[5] Andrew James Meil. *Predicting Success Using the NFL Scouting Combine*. California State University, Fullerton, 2018.

[6] Konstantinos Apostolou and Christos Tjortjis. “Sports Analytics algorithms for performance prediction”. In: *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE. 2019, pp. 1–4.

[7] Dennis Lock and Dan Nettleton. “Using random forests to estimate win probability before each play of an NFL game”. In: *Journal of Quantitative Analysis in Sports* 10.2 (2014), pp. 197–205.

[8] Roger Poch Alonso and Marina Bagić Babac. “Machine learning approach to predicting a basketball game outcome”. In: *International Journal of Data Science* 7.1 (2022), pp. 60–77.

9 Contribution Table

Final Report Task	Group Member
Finalize Introduction & Background	Daniel
Finalize Problem Definition	Daniel
Finalize Data Collection	Daniel and Lauren
Finalize Methods	Everyone
Finalize Results and Discussion	Everyone
Conclusion	Daniel
K-means Clustering	Lauren
Hierarchical Clustering	Likhit
Gaussian Mixture Models	Likhit
Multivariate Linear Regression	Daniel
Support Vector Regression	Daniel
Finalize References	N/A
Contribution Table	Lauren
Presentation Powerpoint	Daniel and Likhit
Presentation Video	Daniel and Lauren

Midterm Report Task	Group Member
Update Introduction & Background	Daniel
Update Problem Definition	Daniel
Data Collection	Daniel and Lauren
Update Methods	N/A
Update Results and Discussion	Everyone
K-means Clustering	Lauren
Hierarchical Clustering	Likhit and Jaber
Gaussian Mixture Models	Likhit and Jaber
Multivariate Linear Regression	Daniel
Multivariate Logistic Regression	Meiyun
Support Vector Regression	Likhit and Jaber
Update References	N/A
Contribution Table	Lauren

Project Proposal Task	Group Member
Decision of Topic	Everyone
Introduction & Background	Daniel
Problem Definition	Meiyun
Methods	Likhit and Jaber
Potential Results and Discussion	Lauren
Providing References	Everyone
Compiling References	Jaber
Proposed Timeline	Meiyun
Contribution Table	Jaber
Data Collection	Lauren
Presentation Powerpoint	Daniel and Likhit
Presentation Video	Daniel and Lauren