**Show Me the Money: Athlete Name, Image, Likeness (NIL) Predictions**

**Cody French (cmfrench), Katy Kibbey (kakibbey), Michael Vizzini (mvizzini)**

## Introduction

Effective July 1, 2021, the National Collegiate Athletic Association (NCAA) adopted an interim policy allowing student athletes to monetarily benefit from their Name, Image and Likeness (NIL). Through marketing and promotional endeavors, such as autograph signings and product endorsements, college athletes can receive financial compensation. It is estimated that 72% of commercial student athlete compensation is now coming through social media, as influencers, content creators, and brand endorsers. Athletes are not allowed to be paid directly by their schools- whether in the form of recruiting incentives or play on the field.65% percent of deals are estimated to be in cash and 35% are in-kind deals of some form of trade.

Currently 32 states have implemented NIL legislation for postsecondary athletes and 30 states have NIL legislation for high school athletes. However the NIL landscape is a bit Wild West in nature as the legal landscape is ever evolving. For example, numerous collectives have sprung up, often under questionable non-profit status, including donor-driven that pool booster funds together to pay athletes and marketplace collectives that create meeting places for athletes and businesses to come together to create opportunities.

As NIL has importance to both athletes and businesses (including college athletic departments) to generate new income, this project aims to identify the characteristics which contribute to student athletes NIL valuation and how accurately predict potential NIL value.

This project aims to explore student-athlete NIL value further by using supervised and unsupervised methods to consider individual factors like TikTok engagement, performance stats, and media rankings, as well as university characteristics that may impact brand value, including the athletic program's brand engagement.

In unsupervised learning, latent features for basketball and football based on the PCA and MDS analysis were different which we found very interesting. Within each of the sports the latent features aligned relatively well but between each of the sports the features were quite different. Basketball put a strong emphasis on the height and weight of an athlete whereas football appeared to put a heavier weight on the college outreach and social status of an athlete. In supervised learning, skill emerged as the most crucial feature in both sports, followed by social media followers.

**Related Work**

As NIL is in its infancy, related studies, particularly those employing machine learning, are limited. We drew on these works to frame our project approach , with the goal of shedding more light on what influences the financial prospects of college athletes.

[***There Is No Nil in NIL: Examining the Social Media Value of Student-Athletes' Names, Images, and Likeness***](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3771581)**:** Kunkel et al. conducted two studies before the 2021 NCAA policy change. Study 1 analyzed social media data from male college football and basketball athletes, calculating the monetization value of their accounts. Study 2 extended this to all student-athletes at four NCAA Division universities, including male and female athletes and analyzing social-media (Twitter, Instagram) post engagement. Their research offered methodology for calculating student-athlete brand value and showed that student-athletes' social media has NIL value counter to NCCA arguments at the time that NIL value creation results only from athlete’s association with the program/institution they play for.

Unlike Kunkel et al, our study aims to use machine learning methods for NIL valuation prediction and incorporates more social media platforms (Twitter, TikTok, and Instagram), as well as university athletic department data, such as team revenue and division classification, as features of interest for prediction.

[***Untapped Potential: An Examination of Name, Image, and Likeness Earnings Estimates for Community College Athletes***](https://csri-jiia.org/wp-content/uploads/2022/04/RA_2022_12.pdf)**:** Cocco, et. a. examined the NIL value of community college athletes given variation in state’s inclusion of community college athletes in NIL legislation. Some states were found to directly or indirectly limit inclusion only to athletes attending four-year post-secondary institutions. Researchers examined the potential for community college athletes to monetize their NIL through social media marketing opportunities by addressing two key questions: *What percentage of community college athletes could potentially monetize their NIL in this manner, and what is the earnings potential for them as social media influencers*? Using similar computational methods to Kunkel et. al on 2020 data., researchers found valid evidence of monetization potential for community college athletes studied, warranting the need to ensure community college athletes NIL protections similar to athletes attending four-year post-secondary institutions.

As with the previous reference, our study aims to use machine learning methods for NIL valuation prediction and incorporates more social media platforms (Twitter, TikTok, and Instagram), as well as university athletic department data, such as team revenue and division classification, as features of interest for prediction. Our dataset did not include 2 year colleges, although different college divisions among four-year post-secondary institutions were represented.

[***The Impact of Name, Image, and Likeness Contracts on Student-Athlete College Choice***](https://deliverypdf.ssrn.com/delivery.php?ID=184115094024087069116093119109084006016013058034039018119000022101067022105008016066026061025058112027043069124089081096069014114006023055020030100083112015079077056078048093104067118070004096098127117067115080094094085091011122107007109094008109004&EXT=pdf&INDEX=TRUE)***:*** Owens et.al analyzed data on football programs and recruits for the 2021-2022 recruiting class, employing a two-stage model to predict scholarship offers and recruit decisions. This study found the school value of Name, Image, and Likeness (NIL) deals significantly influences a recruit's choice of school, with a magnitude comparable to other factors like the school's prior success and NFL player placements.

Owens et. al. focus of interest was on the effects of school-level NILs on recruiting, formally modeling the choices for both prospective recruits and colleges/universities. Our project differs in that our focus on interest is on the factors that influence player NIL valuation.

**Data Source**

The explosion of NIL in the past two years has led to the creation of proprietary algorithms by sports third-party platforms for providing NIL data to athletes, coaches, and fans. However, the data released by these companies during this early NIL era is aggregated. Tracking specific types of NIL deals and their sources is challenging because not all NIL activities are reported to universities or third-party platforms, and universal tracking doesn't exist due to privacy regulations and athletes' reluctance to disclose deal details.

To gather our primary data, we relied on On3, a leading third-party platform that offers recruiting data for athletes, including NIL valuations, social media stats, rankings, biometrics, hometown, high school, and commitment (collegiate or professional), as well as composite skill ratings. On3 lacks an API, so we developed our web scraping script using BeautifulSoup and Selenium. Given the layout of the On3 website, we had to navigate to individual recruit webpages, which was time-consuming, as each year for football and basketball contained around 2000 and 300 athletes, respectively. To address this, we processed data for each recruit year and sport separately and then merged them into CSV format for use in our models.

We also utilized a second dataset, which was obtained from the [U.S Department of Education](https://ope.ed.gov/athletics/#/datafile/list), Office of Postsecondary Education, Equity in Athletics Disclosure Act (EADA) Data Analysis Cutting tool. Which provides information about postsecondary institutions varsity athletic teams, athletic department expenses and revenue in xlsx format, which we then repackaged in CSV format for faster processing. The two datasets were merged on the institution name.

We decided that we would develop 2 different datasets, one for football and one for basketball. We hypothesized that for each sport different features may carry different importance and allow our models to perform better. For each of these sets we used a supervised and unsupervised set of the data, for the supervised data we only included recruits where the NIL evaluation was present as that would be our prediction variable. For the unsupervised set we included all recruits, setting those that did not have an evaluation to $0, with the thought that it could be used in a semi-supervised learning approach.

## Part A – Supervised Learning

### Motivation

### The motivation for supervised learning was to gain insight into the variables that impact student athlete NIL valuation from our dataset and explore the capability of supervised models to predict NIL valuation.

**Data Source and Feature Engineering**

As noted in the Data Source section, we generated separate football and basketball datasets for both supervised and unsupervised learning. For supervised learning, the original basketball dataset contained 137 rows and 24 features. The football dataset was larger with 1263 rows and 24 features. Features can be categorized into roughly three types: athlete-specific (e.g. position, skill rating, height, weight), social media (e.g. Twitter, Instagram, and Tik Tok number of followers) and college related (athletic department revenue and expense of athlete’s committed school, college distance from student’s hometown). Data cleaning including removing rows with missing values. Initially, six(6) features deemed unnecessary (e.g. SPORT, Unnamed:0) or just a duplicate of another feature (e.g. EXP\_YEARS given EXP\_MONTHS) were dropped.

Exploratory data analysis of the remaining features included examining a correlation heatmap for both basketball and football. Analysis of the heatmaps indicated the potential for multicollinearity due to independent variables being highly correlated with each other. This helped us understand which features were redundant or highly correlated with each other, resulting in dropping two features. For example, TOT\_FOL (total social media followers) had a perfect positive relationship with the variables for number of Instagram, Twitter, and Tik Tok followers. Since TOT\_FOL was a sum of the totals for each of the three social media platforms, we decided to drop this feature as well as recruit year which was highly correlated with grade. We felt grade was a more significant feature as it could be indicative of experience. The final total of our dataset features was 16 for both basketball and football.

Our dataset contained both categorical and continuous variables, to prepare this data for linear regression modeling and to ensure that the categorical data didn’t introduce bias into the model., we employed "One-Hot Encoding" for handling the categorical variables (GRADE, POS, Classification\_Code). These variables did not assume any ordinal relationships. We additionally log transformed the continuous data due to evidence of skewness and heteroscedasticity seen in exploratory data analysis. NIL values for football ranged from a minimum of $60,000 to a maximum of $5,200,000 (Shedeur Sanders) with a mean of $157,563, demonstrating a positive(right) skewness. Similarly, basketball NIL values were right-skewed, with a minimum of $60,000, maximum of $6,100,000 (Bronny James). In some instances, continuous variables, such as those for the number of followers per each of the three social media platforms, had zero values which are not necessarily outliers. Initial attempt to use log1p introduced a warning, so this situation was addressed by adding a small constant to the data before applying the logarithm transformation. As two of our chosen supervised learning models, Random Forest Regressor and XGBoost Regressor, are able to handle categorical values, separate datasets were prepared with just log transformation being performed.

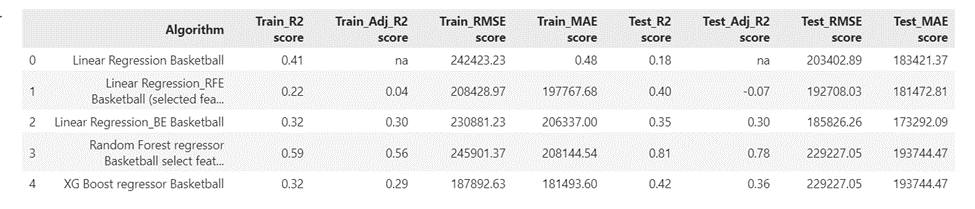
### Supervised Methods

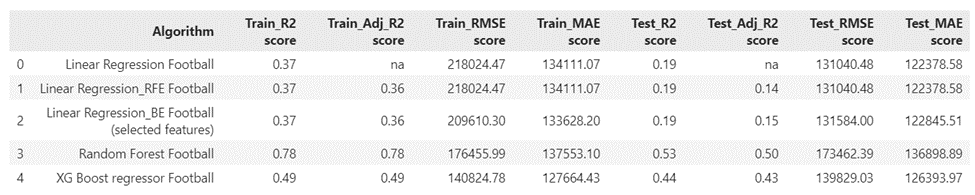
We utilized a diverse set of five models in our attempt to predict NIL valuations, each serving a distinct purpose:

* *Linear Regression with All Features:* This model represents our baseline linear regression approach, where we considered all available features. It provides a straightforward analysis of how well a basic linear model can predict NIL valuations.
* *Linear Regression with Recursive Feature Elimination (RFE):* In this model, we employed Recursive Feature Elimination (RFE) to systematically select a subset of the most relevant features. The goal of RFE is to enhance model interpretability by focusing on the most influential predictors.
* *Linear Regression with Backward Elimination (BE):* Backward Elimination is a stepwise linear regression technique. We started with all features and iteratively removed the least significant ones, aiming to identify the most crucial predictors for predicting NIL valuations.
* *Random Forest Regressor:* The Random Forest Regressor is an ensemble model known for its robustness and ability to capture complex relationships in the data. We chose this model to leverage its capacity to handle intricate patterns and interactions within the dataset.
* *XG Boost Regressor:* XG Boost, a gradient boosting algorithm, was selected for its strong performance and predictive power. This model is known to capture nuances in the data and deliver accurate predictions,

A summary of results for these models is shown in Table 1. Metrics used for evaluating our predictive models included R-squared (R2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Given the presence of outliers in our data and the sensitivity of RMSE to outliers, our main metrics of focus were MAE and R2. Compared to RMSE, MAE is less sensitive to outliers. Essentially we were interested in the error in predicting NIL values (dollar units) of our models compared to the actual NIL values. We also initially calculated a naïve MAE baseline to determine whether our models were actually adding predictive value and outperforming the naive baseline. It was calculated by taking the median of the target variable. 

While none of the models we attempted outperformed the simple baseline, the Random Forest Regressor was the best performing model. For both Random Forest and XGBoost, an initial model was run and then hyperparameter tuning (Randomized CV) was performed using a 10-fold cross validation over different combinations of parameters.





### *Analysis of Best Performing Model:* Random Forest Regressor

This analysis focuses on the Random Forest Regressor (RF) model as the “best” performing model to predict NIL value. On the basketball dataset, the model's R2 explains approximately 78% of the variance in the target variable. The small size of the basketball dataset likely caused the test R2 to be higher than the training R2. Noise in training data due to outliers among social media features. Athletes can vary significantly in how they use social media platforms like Instagram, Twitter, and TikTok. Some athletes may have a strong social media presence with a large following and high engagement, while others may have a smaller or less active online presence. The RF Basketball model appeared to perform reasonably well based on 5 k-fold cross-validation metrics. It exhibited a moderate level of predictive power (R2 while maintaining relatively low mean absolute error (MAE) and root mean squared error (RMSE), indicating accurate predictions on average across different cross-validation folds.

On the football dataset, the training R2 score (0.77) is higher than the test R2 score (0.51). While this suggests some degree of overfitting, the model still exhibited a reasonable level of generalization since the test R2 score is not significantly lower than the training R2 score. The Random Forest CV Football model, when subjected to k-fold cross-validation, shows consistent performance metrics across different subsets of the data. The mean R2 CV score (0.50) is quite close to the test R2 score of the Random Forest model (0.51). A screenshot of a computer screen

Description automatically generated

### Feature Ablation Analysis

In our analysis, we conducted a feature ablation study using a Random Forest Regressor to assess the impact of individual features on the predictive performance of the model. This process involved systematically removing one feature at a time and re-evaluating the model's performance. The objective was to gain insights into the relative importance of each feature and how it contributes to the accuracy of our predictions. See Appendices C & D for data tables.

Basketball dataset insights:

* The feature 'SKILL' exhibited the most significant impact on the Mean Absolute Error (MAE) when removed, with an increase of approximately 0.067 in MAE. However, its influence on R-squared was relatively small, with a decrease of approximately -3.5e-06.
* Other features that notably affected MAE included 'INSTA\_LONG' and 'EXP\_MEN,' with MAE changes of approximately 0.023 and 0.0207, respectively.
* The majority of features had relatively minor impacts on both MAE and R-squared, with changes close to zero.
* Notably, removing features such as 'GRADE,' 'AGE,' 'NUMOFF,' 'POS,' 'HEIGHT\_IN,' and 'WEIGHT\_LBS' had minimal impact on both MAE and R-squared.

These insights suggest that 'SKILL' and 'INSTA\_LONG' are key contributors to model performance, primarily affecting MAE. However, most other features have limited influence on the model's performance, as indicated by small changes in both MAE and R-squared.

Football dataset insights:

* The Ablation MAE values for feature removals are relatively consistent, with small fluctuations around the 147,034 mark. This suggests that the removal of individual features has a relatively stable impact on the Mean Absolute Error (MAE) of the random forest regressor model.
* The Ablation R-squared values consistently hover around -1.218, indicating that the removal of features does not significantly affect the goodness of fit of the model. This suggests that the model retains its predictive power even after feature removal.
* While most features have a minor impact on MAE and R-squared, "AGE" stands out with the highest MAE Change of 0.025, suggesting that its removal has a slightly more noticeable impact on model performance. However, even this change is relatively small in absolute terms.
* The "SKILL" feature exhibits a negative MAE Change, indicating that its removal results in a slight improvement in model performance. This feature seems to have a negligible impact on the model's predictive accuracy.

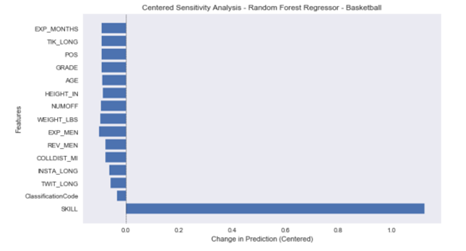
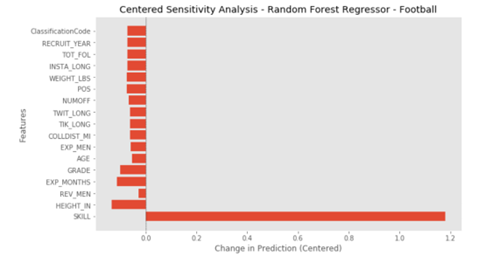
#### Sensitivity Analysis

We conducted a sensitivity analysis on our Random Forest Regressor model to gain insights into the influence of specific features on predictions. We began by training the initial model using our training data and calculated the mean feature values for our dataset. Subsequently, we created a central data point by using the mean feature values and made predictions for this central point using RF1 as our reference. We focused on all the independent features in the dataset. These features were systematically perturbed by a predefined value (0.5), and predictions were made for the perturbed data points.. We systematically examined the impact of each independent feature in our dataset by perturbing them with a predefined value (0.5). This perturbation allowed us to observe how changes in these features affected our model's predictions.

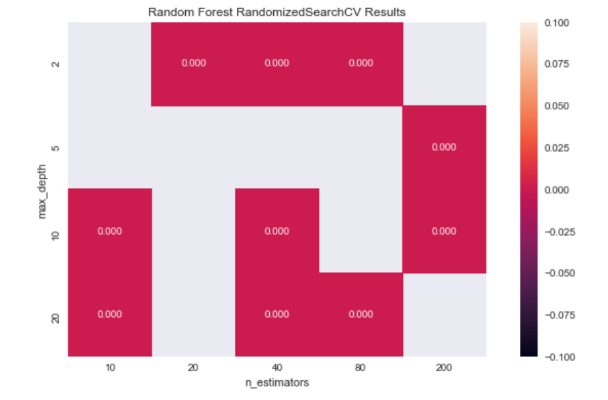
Our analysis revealed noteworthy insights for both football and basketball datasets:

* ‘Skill' Feature Significance: Across both sports, the 'skill' feature exhibited the most substantial impact on predictions when perturbed. This underscores the significance of an athlete's skill level in attracting attention from fans and brands. However, it's essential to note that the nature of the 'skill' data, obtained from On3, is somewhat opaque due to its proprietary creation process. It appears to be a composite rating based on various factors, including recruiting media rankings, coach and ranking committee feedback, and draft prospects' evaluations.
* Basketball's 'Height' Influence: In the basketball dataset, the 'height' feature followed 'skill' in terms of impact on predictions. Height is a critical attribute in basketball, affecting an athlete's performance and appeal/notoriety.
* Football's Program Expenses. For football, the expenses associated with an athlete's college football program emerged as a second significant influencing factor. This suggests that the financial resources allocated to a program can have a notable impact on player performance and recognition.
* Program Classification and Revenue: Interestingly, in basketball, the revenue of the athlete's college basketball program had a relatively lower impact on predictions. Similarly, for football, the classification of the program (with a prevalence of NCAA Division 1 representation) appeared to be less influential in our model's predictions.

The sensitivity analysis shed light on the critical features that drive predictions in both football and basketball. While 'skill' and 'height' played pivotal roles, we acknowledge that the 'skill' feature's proprietary nature adds complexity to its interpretation. Skill as a feature is a bit subjective as it seems to be compiled based on other recruiting media rankings, coach and ranking committee feedback, and draft prospects (e.g. NBA, NFL).



1. **Identified Tradeoffs**

In Random Forest, n\_estimators and max\_depth are two important hyper-parameters. n\_estimators is the number of trees t to build. max\_depth represents the depth of each tree. We conducted a Randomized CVCV grid search on thesetwo dimensions.However, we saw a correlation 0 between the max\_depth (10,20) and n\_estimators (40,80)hyperparameters in a correlation heatmap, which indicated that there is no linear relationship between these two hyperparameters in our model's performance on the given football and basketball dataset. Changing hanging the value of one hyperparameter (max\_depth) did not have a consistent and predictable impact on the other (n\_estimators) based on linear correlation alone. We considered these hyperparameters independently when tuning the model, using max\_depth 10 based on the size of our datasets and n\_estimators at 40. No significant change was seen between n\_estimators at 40 versus 80.

#### Failure/Error Analysis

In the course of our error analysis for predicting NIL values using a Random Forest model, we identified several specific instances where the predictions failed to meet the desired accuracy threshold. We noticed that the model struggled to accurately predict values in the presence of significant outliers in the test data. Outliers can introduce noise into the predictions, leading to deviations from the expected results. Future improvements might involve implementing more robust outlier detection and handling techniques, such as robust regression models or outlier exclusion strategies. In the future it would be advantageous to set boundaries and remove the samples with more extreme Nil valuations to improve this issue. Scatterplot and histogram visualization of predictions compared to actuals can be found in Appendix: E.

## Part B- Unsupervised Learning

***Motivation***

After scraping and cleaning our data, we had just over 23 features for each athlete. Due to the high dimensionality of this data, we needed to perform two crucial steps in order to retrieve new insights.

1. Perform feature engineering to reduce the dimensions of our dataset down to 2 or 3 dimensions such that cluster plotting would be interpretable
2. Utilize unsupervised learning techniques to plot clusters and score them using scoring techniques

The goal of this section was to provide some insight beyond the traditional exploratory data analysis (EDA) approaches and potentially give the supervised learning section additional data points to train, validate, and test on.

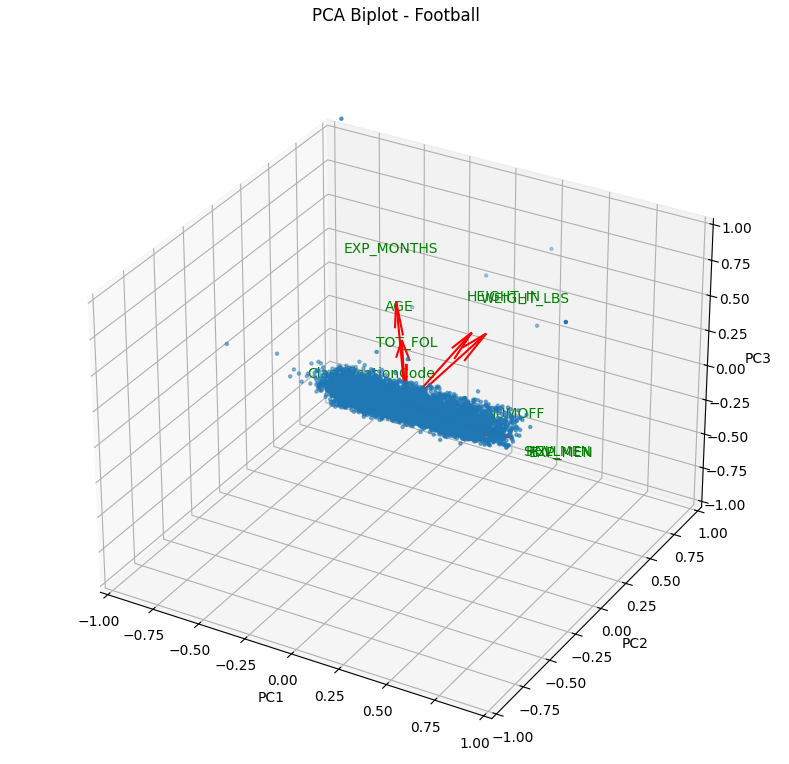
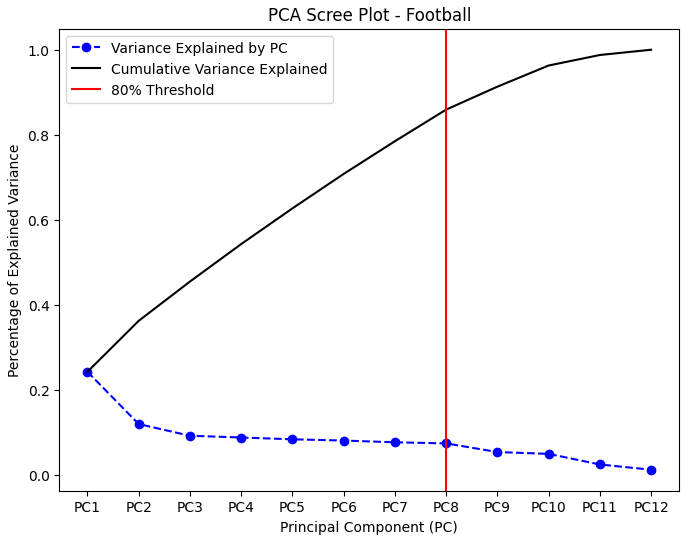
***Data Source***

One of the outputs from the web scraping and cleaning section was a dataset that contained all of the features as well as our target variable “*NILVAL\_LONG\_USD*” (Appendix B).. However, not every athlete had this label (ex: some athletes do not yet have an NIL evaluation whereas some do). We thought it might be interesting to use all of these athletes in the analysis for the unsupervised section and based on how these athletes cluster, potentially assign those athletes without an NIL evaluation to a cluster of athletes that *have* one (sort of a semi-supervised learning approach).

## *Feature Engineering*

With our data extracted, the first step prior to feature engineering was to handle non-numeric feature columns. Thankfully in our data cleaning stage there was really only one feature that needed to be addressed, “*STARCOLL*.” This feature represents the college that an athlete has committed to, is going to commit to, or is most likely to attend. As many of the feature engineering techniques below require numeric features it was decided that this feature would be dropped as it resulted in a sparse data frame if we One-Hot encoded on this feature. Also, as we retained columns like “*NUMOFF*” (number of offers the athlete received) and “*COLLDIST\_MI*” (distance in miles from the athletes hometown/high school to their *STARCOLL*) so it was determined that pruning this feature would be acceptable for the analysis.

Three feature engineering techniques were utilized in our pipeline to help perform a feature selection/extraction that we could use to reduce our features dataset down to either 2 or 3 dimensions. These techniques were chosen to cover a broad spectrum of underlying assumptions, and together prove helpful for later clustering methods.. Principal Component Analysis (PCA) was initially used to reduce our feature space. PCA assumes that the data are linear in nature and by understanding how well our model does or does not fit we can gain an understanding of whether our data is linear or nonlinear. Figure 3 shows a Scree plot as well as a Biplot for the football dataset (basketball dataset was separately run and these plots are in the Appendix F).



A Scree plot (Figure 3) was used to see how many principal components were needed in order to retain 80% of the variance from the original data (left) and a Biplot was created showing utilizing the first 3 principal components and plotting the original features in this space (right). Note that this was also performed on the basketball dataset.

For football, we needed eight principal components (PCs) to capture 80% of the data's variance (six for basketball), hinting at potential non-linearity in the data. Given the complexity of human data, especially in athlete NIL evaluations, linear assumptions may not apply. However, analyzing the Biplot helps identify features closely aligned with each PC, aiding in feature selection and extraction. We performed cosine similarity between original features and each PC for both sports, generating Table 4 to assist in naming latent features or selecting representative features for interpretation.

| **Sport** | **PC #** | **Feature Selection** | **Feature Extraction** | **Latent Feature** |
| --- | --- | --- | --- | --- |
| Basketball | 1 | COLLDIST\_MI | COLLDIST\_MI, WEIGHT\_LBS, HEIGHT\_IN | Biometric Data |
| Basketball | 2 | REV\_MEN | REV\_MEN, EXP\_MEN, TOT\_FOL | College Outreach |
| Basketball | 3 | SKILL | SKILL, NUMOFF, POS | Ability |
| Football | 1 | EXP\_MEN | EXP\_MEN, SKILL, REV\_MEN | College Outreach/Ability |
| Football | 2 | WEIGHT\_LBS | WEIGHT\_LBS, HEIGHT\_IN, POS | Biometric Data |
| Football | 3 | AGE | AGE, EXP\_MONTHS, TOT\_FOL | Age |

It is interesting to see how basketball and football are slightly different from one another. We can see that basketball places a strong emphasis on biometric data like weight and height as these are closely associated with PC1 whereas football puts a much stronger emphasis on the college’s revenue and spending to retain variation!

From here, the team decided that a nonlinear technique might be interesting to utilize so Multidimensional Scaling (MDS) was chosen. MDS differs from PCA in that this technique does not make a linearity assumption. We retain the ability to see what features from the original feature space contribute most closely with each dimension of MDS and we can see if there are any athletes who might be outliers in our dataset as MDS specializes in moving dissimilar data points far away and similar data points close together. In Figure 4 we show the MDS plot for the basketball as well as the original feature loadings onto each of the MDS features through a linear regression (football dataset was separately run and these plots are in the appendix). A close-up of a graph

Description automatically generated

In Figure 4 athletes greater than 8 units away from the central cluster are named (left) and the feature loadings from the original feature space is plotted onto the MDS dimensions via a heatmap (right).

We can see from the MDS plot that athletes like “Bronny James” (Lebron James’s son) is very dissimilar from the rest of the pack and this is something we would expect. We can also see athlete “Mikey Williams” who was recently tried for gun charges which may influence his social media presence thereby differentiating him from the rest. Table 5 below was generated using the MDS heatmap in order to associate which features might be most similar to each MDS dimension.



| **Sport** | **MDS Dim #** | **Feature Selection** | **Feature Extraction** | **Latent Feature** |
| --- | --- | --- | --- | --- |
| Basketball | 1 | POS | POS, HEIGHT\_IN, WEIGHT\_LBS | Biometric Data |
| Basketball | 2 | SKILL | SKILL, NUMOFF, EXP\_MEN | College Outreach |
| Basketball | 3 | AGE | AGE, TOT\_FOL, EXP\_MONTHS | Social Media Status |
| Football | 1 | POS | POS, WEIGHT\_LBS, EXP\_MONTHS | Biometric Data |
| Football | 2 | AGE | AGE, TOT\_FOL, EXP\_MEN | Social Media Status |
| Football | 3 | EXP\_MONTHS | EXP\_MONTHS, HEIGHT\_IN, AGE | Age |

We can see that even though PCA alone and MDS alone may struggle to summarize the data,, the two methods produce very similar latent features which can at least verify our understanding of reducing the feature space or pruning the original datasets. Finally, we utilized t-Distributed Stochastic Neighbor Embedding (t-SNE) to see if we could augment the feature space to put an emphasis on local clusters while preserving some semblance of distances on a global scale. t-SNE comes with some challenges in that the distortion of how data points are represented may not be true to how these points are actually clustered. For this reason we chose to incorporate some of our labeled data into the t-SNE plot to give us a sense of how athletes without an NIL evaluation stack up to those athletes that do. t-SNE is also our first feature engineering technique in which a key parameter, perplexity, must be evaluated. As we have an incomplete column of labeled data, we are going to have to use visual inspection in order to determine which value of perplexity we should utilize. A group of images of different colored dots

Description automatically generated

Figure 5 shows the matrix of t-SNE plots with varying levels of perplexity with the basketball dataset (football dataset was separately run and these plots are in the appendix).

Perplexity is an attribute that defines the number of nearest neighbors that the manifold learning algorithm uses. The sklearn [documentation](https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html) (See References) recommends trying perplexity values from 5 to 50 which for the basketball dataset, which appears to be correct. For the football dataset we can see that bigger values of perplexity are required to see more distinct clusters which might be expected as this dataset is far bigger. A perplexity of around 400 produces two distinct clusters of athletes.

Through techniques like PCA, MDS, and t-SNE we can see that reducing our dataset down to 2 or 3 dimensions should retain some latent features and result in a grouping of our data which can provide additional insights.

***Unsupervised Learning Methods and Evaluation***

With our datasets reduced using a variety of techniques, it is now possible to cluster our data and get meaningful insights as to *how* the data was clustered.

As a natural first step in clustering we started with KMeans clustering. KMeans, however, requires a hyperparameter, *n\_clusters*, to be defined prior to running this algorithm. To get a better sense of the number of clusters we should use we generated an Elbow plot and a dendrogram via the agglomerative clustering toolkit in scipy. Both of these outputs should be able to give us a general sense of the appropriate number of clusters that we need before running KMeans. Figure 6 below shows the elbow plot, dendrogram, and KMeans result utilizing t-SNE output when using the ideal number of clusters for the basketball dataset (3 for both basketball and football). Note that the football dataset was separately run and these plots can be found in the appendix

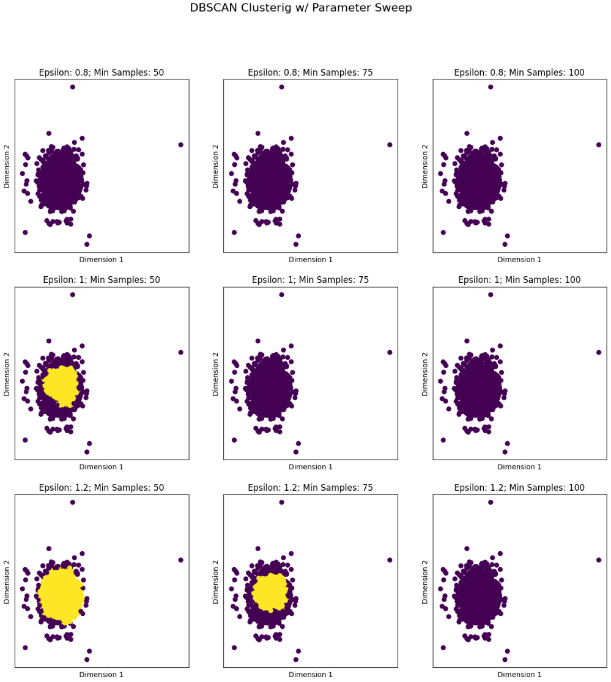
The elbow plot and dendrogram suggested approximately three clusters for both basketball and football datasets. Running the KMeans algorithm multiple times, we obtained average silhouette scores of 0.38 and 0.37 for basketball and football, respectively. These scores, slightly positive, indicate that KMeans with PCA output could separate clusters slightly better than allowing them to overlap but struggled due to the dispersed nature of the data. A diagram of a diagram

Description automatically generated with medium confidenceA chart with different colored dots

Description automatically generatedA graph of a line

Description automatically generated

Given the data's non-globular, spread-out distribution, we decided to proceed with the next clustering algorithm: DBSCAN.

Density Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering technique that does not require the user to define the number of clusters in advance. Instead, DBSCAN determines the number of clusters and assigns data points to said clusters automatically while also identifying outliers and assigning them as such. DBSCAN requires the user to set a few parameters in order for the algorithm to work, most notably the ‘*epsilon*’ and ‘*min\_samples*’ parameters. ‘*epsilon*’ defines the maximum distance from one point to another to still be considered in the neighborhood of that original point and ‘*min\_samples*’ is a parameter that determines if a point will be considered as a core point to a cluster (whereas edge points would be considered boundary points). 

In order to score the ‘best’ parameters for each of the dataset’s parameter sweeps we utilized Density Based Cluster Validation (DBCV) as opposed to a silhouette score from KMeans. DBCV is a scoring technique that compliments DBSCAN and is more appropriate than silhouette score as DBCV does not require clusters to be uniform in size and allows for unique shapes as well as has a way to handle noise data within the scoring. We utilized the DBCV library created by Christopher Jenness and Paolo Galeone [X]. The DBCV scores for the DBSCAN output when run on the MDS feature space were -0.62 and -0.51 for basketball and football, respectively. Going through the sweep of parameters and using the DBCV score to determine which parameters were the best it was found that for basketball an epsilon of 1 and a min sample of 50 and for football an epsilon of 0.8 and a min sample of 50.

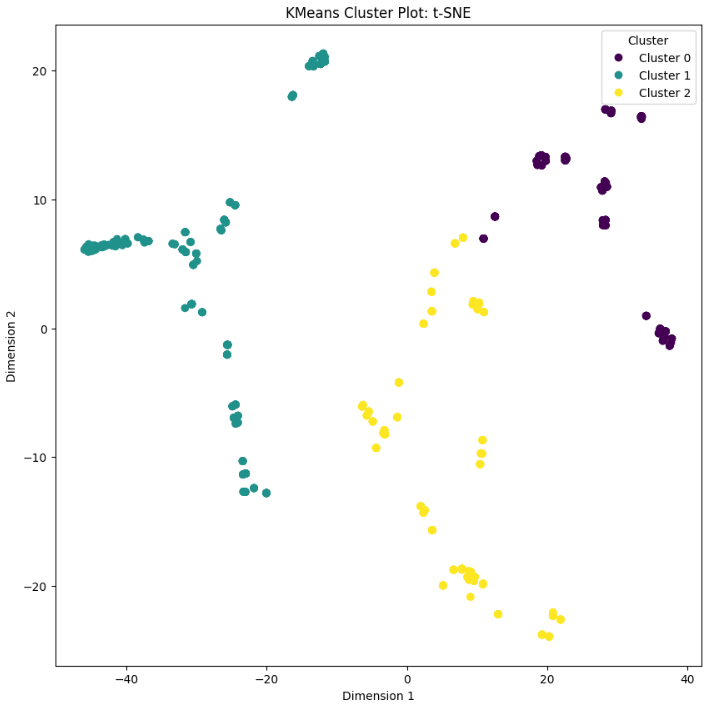
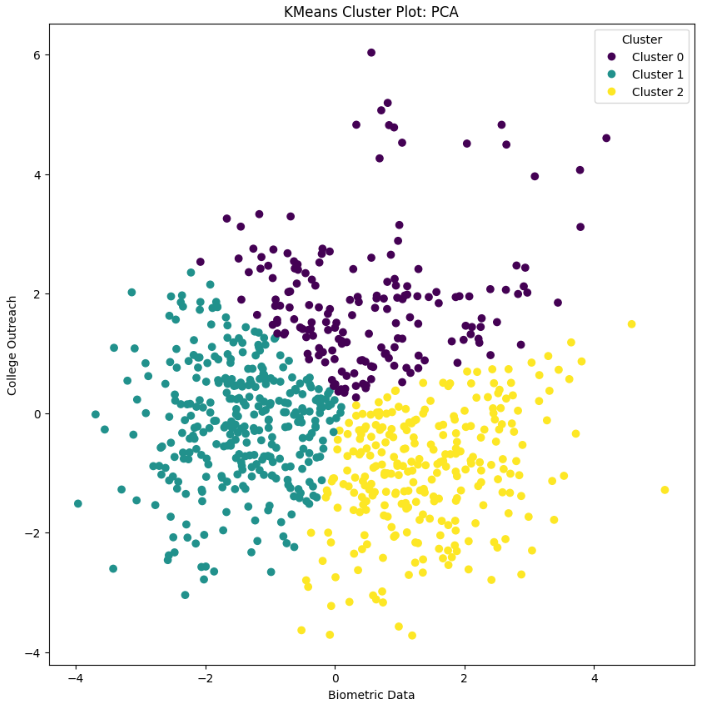
With these two unsupervised models in mind, we determined that the KMeans output run on PCA and t-SNE feature representations were our strongest figures for uncovering additional insights on our data. Table 6 below summarizes our two models (KMeans & DBSCAN) run on the 3 feature representations (PCA, MDS, and t-SNE) with the silhouette scores listed as “basketball/football”. Note that DBSCAN on the football dataset took roughly 26 hours to go through the sweep and score on and this extremely long training time was taken into account beyond the silhouette score for determining which model to use



|  | **PCA** | **MDS** | **t-SNE** |
| --- | --- | --- | --- |
| **KMeans** | 0.38/0.37 | 0.35/0.35 | 0.53/0.53 |
| **DBSCAN** | -0.57/NaN | -0.62/-0.51 | -inf/NaN |

KMeans on the PCA output not only had the highest silhouette scores for the basketball and football datasets but the model trained and ran in the seconds-minutes range whereas DBSCAN took more than 1 day to run and score. Note that “-inf” on the DBSCAN + t-SNE for the basketball dataset just signifies that DBSCAN found that all datapoints were noise.

Showing the best model (KMeans) with the PCA and t-SNE outputs in a bit more detail we can surmise some interesting findings as shown in Figure 8.



From the KMeans PCA plot we can understand how our latent features (and by extension, our original features) are being mapped in this space. For the basketball dataset we can see how biometric data on the x axis and college outreach on the y axis help to locate our cluster centers and change how we group these athletes. From the KMeans t-SNE plot we can see how cluster 1 athletes are more dissimilar to cluster 0 and cluster 2 athletes, but we lose the ability to help us draw in what this differentiation can be attributed to as we have non-linearly distorted space with the t-SNE representation. We think that these two plots together are helpful in communicating some interesting groupings beyond what can be inferred from EDA.

As we can see, the KMeans model is quite sensitive to the feature representation that we choose. KMeans paints dramatically different pictures based on whether PCA or t-SNE (shown above) or MDS is given as the input. KMeans is also sensitive to the starting condition that it is given so we have run these models at least 10 times to produce an average silhouette score and to make sure that cluster representations are shown with repeatability (a link to the Github Repository is attached in appendix and all of these plots were generated in the “unsupervised.ipynb” notebook). KMeans is also sensitive to the number of clusters that you give it. From our elbow plots and dendrograms we know that 2-4 clusters could be reasonable guesses. Silhouette scores of 2, 3, and 4 clusters for the PCA and t-SNE output with “basketball/football” are summarized in Table 7.

|  | **PCA** | **t-SNE** |
| --- | --- | --- |
| **2 Clusters** | 0.37/0.41 | 0.54/0.58 |
| **3 Clusters** | 0.38/0.37 | 0.53/0.53 |
| **4 Clusters** | 0.35/0.37 | 0.52/0.59 |

As we can see, all of the silhouette scores are within a relatively small band given the feature representation and this tells us that KMeans is not very sensitive to the number of clusters that we choose. We believe that 3 clusters, regardless of the silhouette scores, is the correct number as this was indicated by the elbow plots and dendrograms but it is good to also verify that the number of clusters does not significantly disrupt these scores as they are changed.

## Discussion

### Supervised Learning

Supervised learning proved more challenging than anticipated. Linear regression with feature selection techniques like recursive feature elimination and backward elimination underperformed in predicting athlete performance, suggesting non-linear data relationships. Skill emerged as the most crucial feature in both sports, followed by social media followers. Integrating data from our unsupervised modeling into supervised learning was a goal, but time constraints limited this.

Given more resources and time, we could enrich the dataset, explore advanced machine learning models for non-linear patterns, include women athletes and those from smaller schools, and conduct rigorous cross-validation, feature engineering, and deeper exploratory analysis for improved results

### Unsupervised Learning

In performing the unsupervised learning we learned, were surprised, and encountered some challenges, and thought about future iterations of this project involving:

In our unsupervised learning analysis, we uncovered intriguing differences in latent features between basketball and football through PCA and MDS analysis. While the latent features aligned well within each sport, they differed significantly between the two, with basketball emphasizing height and weight and football emphasizing college outreach and social status.

Additionally, we were surprised by the complexity of both datasets, as feature engineering techniques (PCA, MDS, t-SNE, UMAP) revealed more intricate patterns than anticipated, potentially involving unobserved variables like notoriety and name recognition. We also faced the challenge of interpreting clustering outputs, where pairing t-SNE and UMAP with PCA and MDS helped provide more tangible insights.

For future iterations, we may explore subsets of the football dataset for hyperparameter tuning due to resource constraints and consider incorporating new features to enhance unsupervised learning, potentially enabling better classification in supervised learning and addressing data limitations.

## Ethical Considerations

Our analyses involve student athletes' personal data, raising ethical considerations regarding data usage and privacy. During web scraping from the On3 website, we opted not to store individual athlete social media handles as they were irrelevant for our model.

Bias is a concern in both our supervised and unsupervised models, potentially leading to unequal opportunities based on factors like race or socioeconomic status. While our data lacks direct race or socioeconomic information, variables such as hometown and college enrollment could pose aggregation risks. To address this, we plan to balance our training set further, considering additional stratification variables.

Furthermore, we conducted an Algorithmic Impact Assessment from the Government of Canada, highlighting the importance of understanding automation's potential impact in our models. We believe our model is transparent, with its latent features aiding human decision-making rather than functioning as a "black box."

## Statement of Work

| **Cody French:**  Web Scraping  Data cleaning and preparation  Supervised Model building/testing  Proposal/Report writing | **Katy Kibbey:**  Research  Supervised Model building/testing  Data visualization  Proposal/Report writing | **Michael Vizzini:**  Web Scraping  Unsupervised model building/testing  Data visualization  Proposal/Report writing |
| --- | --- | --- |

**References**

Christopherjenness. *GitHub - Christopherjenness/DBCV: Python Implementation of Density-Based Clustering Validation*. <https://github.com/christopherjenness/DBCV>.

Cocco, Adam R., and Anita M. Moorman. "Untapped Potential: An Examination of Name, Image, and Likeness Earnings Estimates for Community College Athletes." Sport Management Review, January 22, 2021.<https://csri-jiia.org/wp-content/uploads/2022/04/RA_2022_12.pdf>

doi:2024 Industry Top Basketball Recruits. (n.d.). Retrieved from <https://www.on3.com/db/rankings/industry-player/basketball/2024/>.

*Equity in Athletics*. Retrieved from: <https://ope.ed.gov/athletics/#/customdata/search>.

<https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>.

Kunkel, Thilo and Baker, Bradley and Baker, Thomas and Doyle, Jason, There Is No Nil in NIL: Examining the Social Media Value of Student-Athletes’ Names, Images, and Likeness (January 22, 2021). Sport Management Review Available at SSRN: <https://ssrn.com/abstract=3771581> or [http://dx.doi.org/10.2139/ssrn.3771581](https://dx.doi.org/10.2139/ssrn.3771581)

Owens, Mark F. and Rennhoff, Adam D. and Roach, Michael, The Impact of Name, Image, and Likeness Contracts on Student-Athlete College Choice (March 11, 2023). Available at SSRN: <https://ssrn.com/abstract=4385246> or [http://dx.doi.org/10.2139/ssrn.4385246](https://dx.doi.org/10.2139/ssrn.4385246)or <http://dx.doi.org/10.2139/ssrn.3771581>.

Sklearn <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>.

The One Super Player Database | On3.com. (n.d.). Retrieved from <https://www.on3.com/db/>.

# APPENDICES

**APPENDIX A: Project GitHub Link**

# <https://github.com/mvizzini951/MS2_NIL>

# APPENDIX B: Data Set Features

| **Column** | **Description** | **Type** | **Source** |
| --- | --- | --- | --- |
| NAME | Name of recruit | Text | On3 |
| GRADE | Year of recruit (college)  0: Not Enrolled yet  1: Redshirt Freshmen  2: Freshman  3: Redshirt Sophomore  4: Sophomore  5: Redshirt Junior  6: Junior  7: Redshirt Senior  8: Senior | Categorical | On3 |
| AGE | Age of recruit | Numerical | On3 |
| HOTOWN | Hometown of recruit | Text | On3 |
| HISCH | Highschool of recruit | Text | On3 |
| STARCOLL | College that the recruit has enrolled at OR their top choice if they have not yet enrolled | Text | On3 |
| NUMOFF | Number of scholarship offers received | Numerical | On3 |
| POS | Position of athlete in their sport  Basketball:  1: PG  2: SG  3: CG  4: SF  5: PF  6: C  7: ATH/Other  Football:  1: QB  2: EDGE  3: WR  4: S  5: DL  6: OT  7: CB  8: LB  9: RB  10: IOL  11: TE  12: K  13: P  14: LS  15: ATH/Other | Categorical | On3 |
| HEIGHT\_IN | Recruit height in inches | Numerical | On3 |
| WEIGHT\_LBS | Recruit weight in inches | Numerical | On3 |
| SKILL | On3 propriety skill rating | Numerical | On3 |
| COLL\_DIST\_MI | Distance from recruit hometown to enrolled college in miles | Numerical | On3 |
| INSTA\_LONG | Number of Instagram followers | Numerical | On3 |
| TWIT\_LONG | Number of Twitter followers | Numerical | On3 |
| TIK\_LONG | Number of TikTok followers | Numerical | On3 |
| TOT\_FOL | Number of social media followers | Numerical | On3 |
| RECRUIT\_YEAR | Year the athlete will start their college career | Ordinal | On3 |
| EXP\_MONTHS | Number of months at college | Numerical | Calculated |
| EXP\_YEARS | Number of years at college | Numerical | Calculated |
| NILVAL\_LONG\_USD | Name, Image, and Likeness valuation in dollars | Numerical | On3 |
| Classification\_Code | Classification of the colleges athletic program  1: NCAA Division I-FBS  2: NCAA Division I-FCS  3: NCAA Division I without football  4: NCAA Division II with football  5: NCAA Division II without football  6: NCAA Division III with football  7: NCAA Division III without football  8: Other | Categorical | US Department of Education |
| REV\_MEN | Revenue of sport program (men’s) at given college | Numerical | US Department of Education |
| EXP\_MEN | Expenses of sport program (men’s) at given college | Numerical | US Department of Education |

**Appendix C: Correlation Heatmaps**

Feature Correlations

A screenshot of a graph

Description automatically generated

RandomizedCV Hyperparamers- Football

A graph of a football game

Description automatically generated with medium confidence

**Appendix D: Supervised: Feature Ablation Data - Basketball Random Forest Regressor**

| **Feature** | **MAE Change** | **R-squared Change (Scientific Notation)** |
| --- | --- | --- |
| **Removed GRADE** | **0.010** | **-1.100e-09** |
| **Removed AGE** | **0.007** | **-3.300e-09** |
| **Removed SKILL** | **0.067** | **-3.522e-06** |
| **Removed NUMOFF** | **0.004** | **-1.200e-09** |
| **Removed POS** | **0.007** | **-9.000e-10** |
| **Removed HEIGHT\_IN** | **0.005** | **-1.200e-09** |
| **Removed WEIGHT\_LBS** | **0.012** | **-9.000e-10** |
| **Removed COLLDIST\_MI** | **0.009** | **-1.000e-09** |
| **Removed INSTA\_LONG** | **0.023** | **6.100e-10** |
| **Removed TWIT\_LONG** | **0.016** | **6.400e-10** |
| **Removed TIK\_LONG** | **0.013** | **5.700e-11** |
| **Removed EXP\_MONTHS** | **0.014** | **-6.200e-10** |
| **Removed ClassificationCode** | **0.016** | **-2.400e-10** |
| **Removed REV\_MEN** | **0.021** | **4.000e-10** |
| **Removed EXP\_MEN** | **0.020** | **1.100e-09** |

**Appendix E: Supervised: Feature Ablation Random Forest Regressor Football**

| **Ablation MAE** | **MAE Change** | **R-squared Change** |
| --- | --- | --- |
| **Removed GRADE** | **-1.218** | **0.017** |
| **Removed AGE** | **-1.218** | **0.025** |
| **Removed SKILL** | **-1.218** | **-0.003** |
| **Removed NUMOFF** | **-1.218** | **0.023** |
| **Removed POS** | **-1.218** | **0.022** |
| **Removed HEIGHT\_IN** | **-1.218** | **0.020** |
| **Removed WEIGHT\_LBS** | **-1.218** | **0.024** |
| **Removed COLLDIST\_MI** | **-1.218** | **0.020** |
| **Removed INSTA\_LONG** | **-1.218** | **0.024** |
| **Removed TWIT\_LONG** | **-1.218** | **0.030** |
| **Removed TIK\_LONG** | **-1.218** | **0.020** |
| **Removed TOT\_FOL** | **-1.218** | **0.024** |
| **Removed RECRUIT\_YEAR** | **-1.218** | **0.024** |
| **Removed EXP\_MONTHS** | **-1.218** | **0.023** |
| **Removed ClassificationCode** | **-1.218** | **0.024** |
| **Removed REV\_MEN** | **-1.218** | **0.026** |
| **Removed EXP\_MENS** | **-1.218** | **0.022** |

**Appendix F: Supervised: Error Analysis Plots – Actual vs. Predicted NIL**

A screenshot of a graph

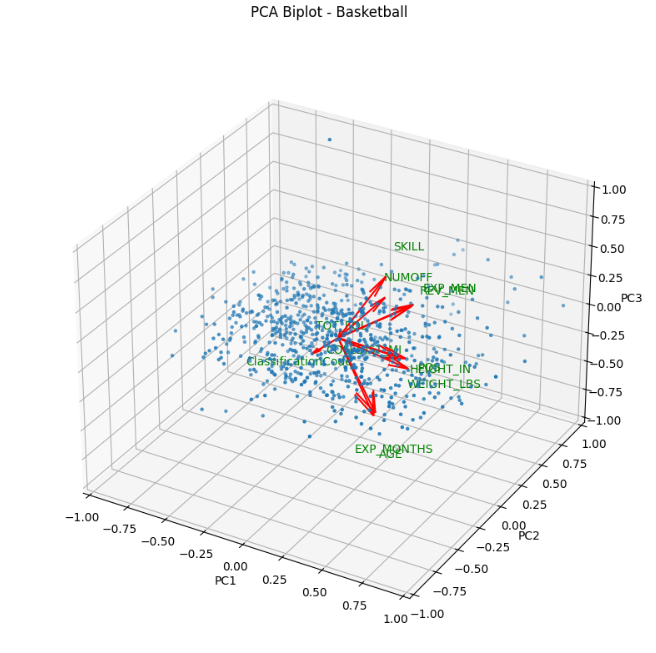
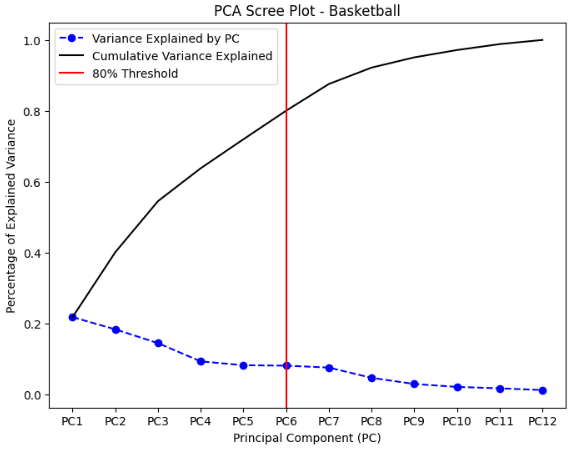
Description automatically generated

A screenshot of a graph

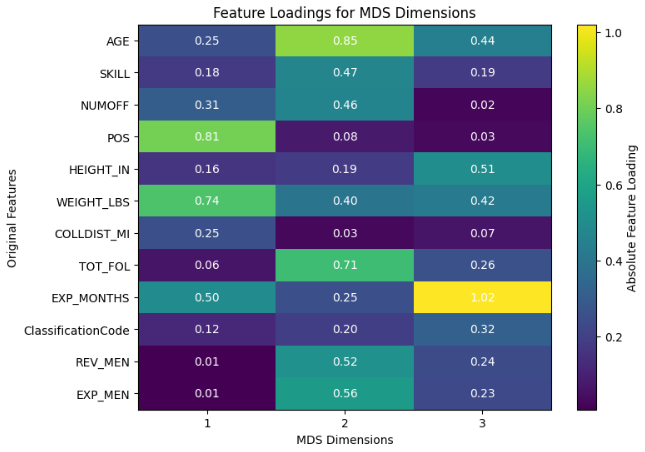
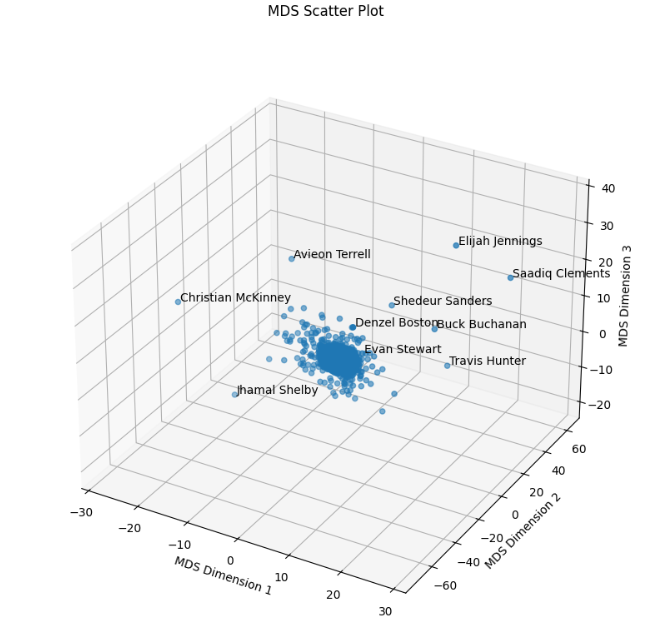
Description automatically generated

**Appendix F: Unsupervised: Basketball PCA**

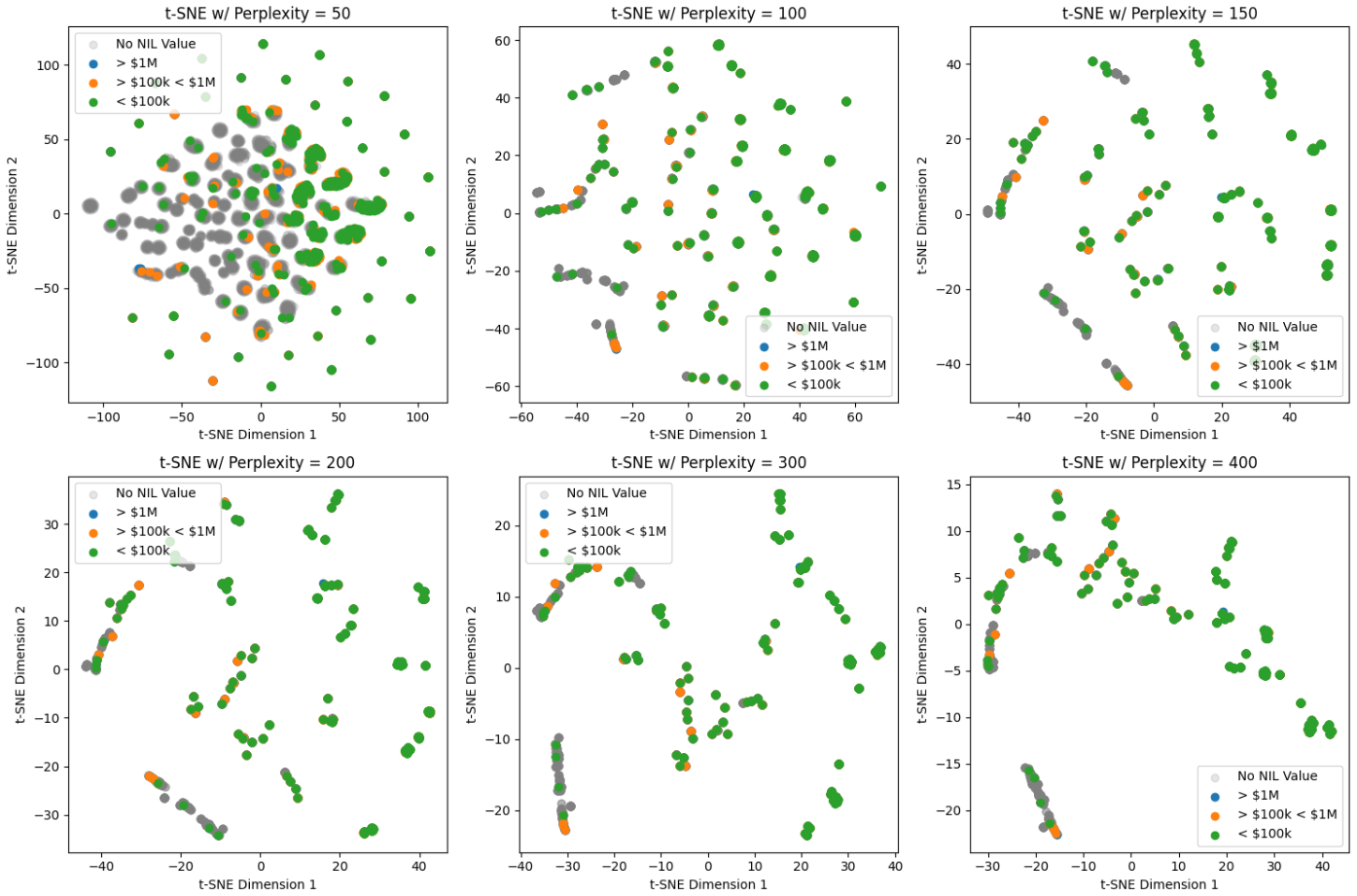
1: Unsupervised Basketball PCA Analysis



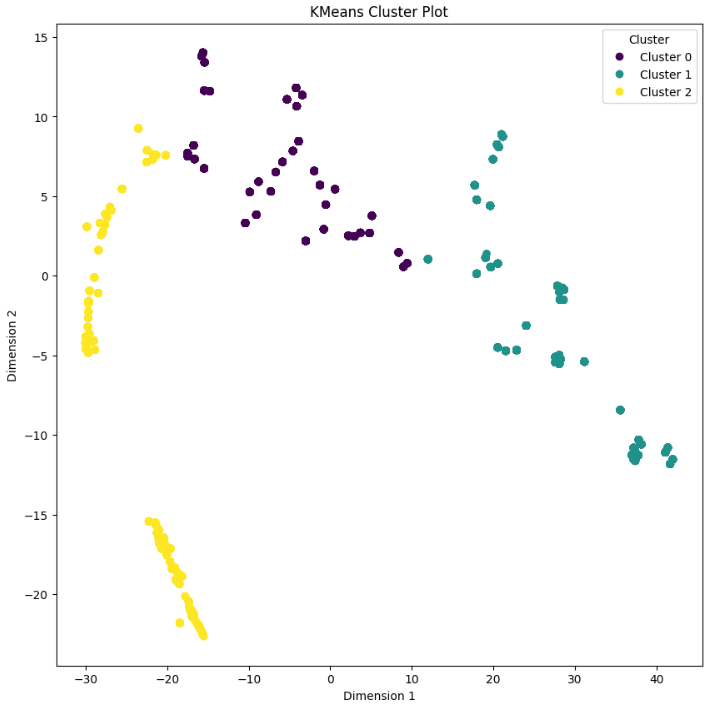
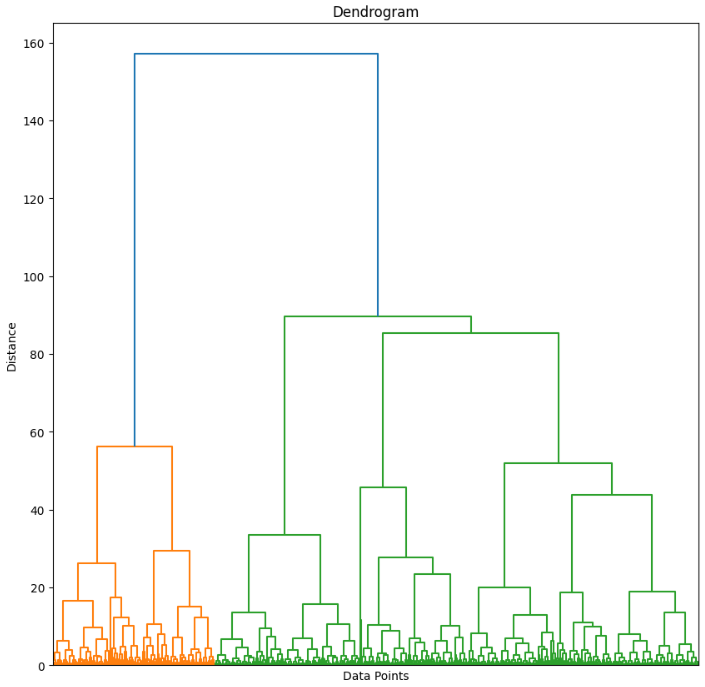
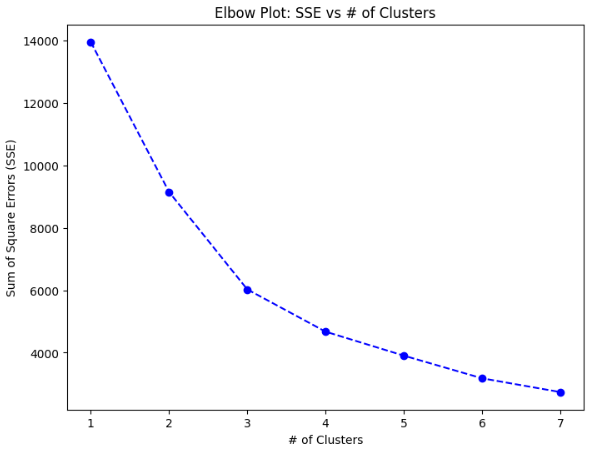
2. Unsupervised Football MDS Analysis



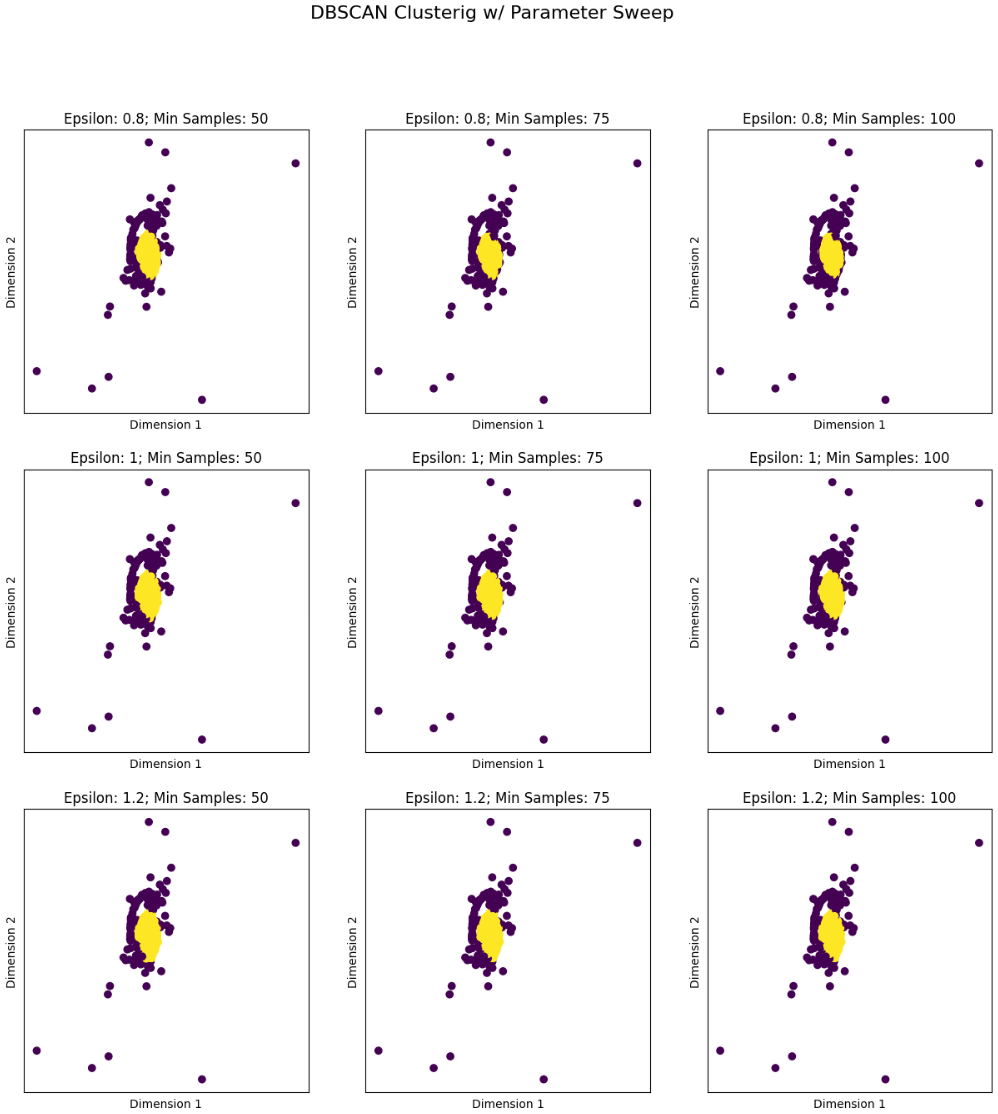
3. : t-SNE Parameter Sweep of the Football Dataset



**4.** Football Elbow Plot, Dendrogram, KMeans w/ t-SNE

****

5. DBSCAN of the Football Dataset



6. KMeans PCA and KMeans t-SNE for Football

