Sam Cohen (sjc71)
Aidan Fetterly (ajf49)
Alex Glick (alg100)
Thomas Murphy (tmm91)
Nicole Toussaint (nmt14)

CS 216 - Project Prototype

**Project Box Folder:** https://duke.box.com/s/kg0d59onvqjpa1fmp8a67ke55x9d59p3

## Part 1: Introduction and Research Questions

The Pro Bowl is the National Football League's annual all-star game. It occurs towards the end of the NFL season and top players are selected from the AFC and NFC to compete against one another. Players are voted into the Pro Bowl by coaches, players, and fans. Being selected for the Pro Bowl is an honor for players, and thus, our group is interested in gaining a deeper understanding of what makes a player more likely to be selected for the Pro Bowl. Through this project, we aim to analyze the key determinants of a player's selection into the Pro Bowl, as well as be able to predict Pro Bowl participants based on offensive skill position players' (i.e. QB, RB, TE, WR) statistics.

In this analysis, there are many factors other than statistics to consider that we might not be able to identify through the data. For instance, if some players have two positions in a given season, we need to determine how selections are made for these players. Further, statistics cannot account for game sense or teamwork. Some of the best NFL players and most frequently selected Pro Bowlers offer more than numbers for their team. We would like to identify if these non-recorded factors have a large impact. Finally, a player's team might have an effect on Pro Bowl selection. A player that shines on a bad team might have better statistics than a player that has to compete with others on their team for possession time. Alternatively, a player that is overshadowed by even more talented teammates could be a star on other teams. Meanwhile, some players reject their selection to play in the Pro Bowl because they have something even more esteemed to be preparing for: the Super Bowl. Thus, these players, who likely are the best of the best amongst NFL players, are not represented in the datasets containing the Pro Bowl players from each year. Additionally, beyond some of the qualitative data that we fail to take into account when predicting annual Pro Bowl selections, we also ignore certain quantitative statistics that can influence an NFL player's selection, such as throws made by skill positions other than quarterbacks, skill position players who are highly impactful in the return game, and the occasional blocking/tackling completed by offensive skill position players. Lastly, we do not properly account for the NFL's intentional attempt to have an even distribution of players represented in the Pro Bowl from all NFC and AFC divisions within the NFL. These are all factors that we will be considering and analyzing throughout the project.

### Research Questions:

1. How can we predict which players will be selected for the Pro Bowl based on season performance?
    a. Substantial: This question will require a detailed synthesis of many statistics based on each position, as well as a discrimination method between those who have made the Pro Bowl and those who have not.

   b. Feasible: Answering this question is feasible because we have access to specific data such as player stats, as well as historical records of which players were selected for the Pro Bowl vs. not.
   c. Relevant: This question addresses an important honor of a player's pro football career, further it is something that we, as group members, are interested in.
2. Which player statistics are key determinants for if a player will be in the Pro Bowl?
   a. Substantial: This question is substantial because it requires data wrangling from many different years in order to get a read on historical trends.
   b. Feasible: Answering this question is possible for our group in this time frame because we have an understanding of the sport, as well as annualized player statistics.
   c. Relevant: This question is important because it can provide evidence on which statistics are important for players to focus on in order to make the Pro Bowl.
3. Based on our predictive models, are there players that have been 'snubbed' in years past for Pro Bowl selections, or that were selected for the Pro Bowl that perhaps should not have been?
   a. Substantial: This question requires us to look at the accuracy of our predictive model and identify if there are players who, based on data and statistics, should have received a nomination.
   b. Feasible: This question will be possible to answer because we can simply look at what factors are predictive of a Pro Bowl nomination, and then identify if there are any players in prior years who would be predicted to be in the Pro Bowl based on their statistics, but were not selected.
   c. Relevant: This question is extremely relevant because it helps us to understand things like human error and subjectivity in professional sports. For instance, if a player had great statistics that should make them eligible for the Pro Bowl but they were not selected, it will be interesting to analyze what could have accounted for this 'human error.'

**Part 2: Summary of Results**

1. **How can we predict which players will be selected for the Pro Bowl based on season performance?**

   There are many components that factor into whether a player will be selected for the Pro Bowl. For our project, we aimed to predict which players would be selected for the Pro Bowl in four different offensive skill positions based on a wide variety of quantitative player statistics. In order to do so, we used two different methods of analysis: Logistic Regression and KNN models. In our logistic regression models, we split the data into training and testing data and ran one regression for each skill position. The quantitative variables that are factored into our analysis can be viewed in more detail in our code, and include things such as passing, rushing, and receiving statistics. For each regression model, we were able to predict players that would be in the Pro Bowl with 0.87-0.96 accuracy depending on the position.

   The KNN model provided a different means of predicting Pro Bowl selections, and also had high accuracy scores. The accuracy scores were between 0.88 and 0.95 depending on the position. This analysis was helpful in providing an additional model to analyze our data. Rather than just having one logistic regression model, we feel that the

KNN model helps show that the metrics/variables we are using in our model are good determinants of Pro Bowl selections and yield high accuracy scores.

2. **Which player statistics are key determinants for if a player will be in the Pro Bowl?**

While there are many quantifiable metrics that are tracked for each offensive skill position in the NFL, it is quite difficult to find a reliable/accurate data source that provides such a comprehensive list and is easily exportable. However, we were fortunate enough to find one that correctly identifies throwing statistics, rushing statistics, and receiving statistics. This allowed us to work with most of the relevant metrics for each of the four offensive skill positions simply by sorting the data by position. One caveat to this is that if there were relevant metrics that don't fall under the category of throwing, rushing, or receiving, then they were excluded despite their importance to the position, the most namely being blocking for tight ends, who serve as receivers primarily but have a critical role as blockers as well. Additionally, we cleaned the data by stripping metrics that we, through our knowledge of the sport and research conducted, determined were not relevant to the position group but still tracked.

From here, we then imported and ran a sklearn RFE (recursive feature elimination) model that identified how many metrics and which specific metrics should be considered for each position that results in the highest accuracy score when making predictions of Pro Bowl nominations. The model determines these features through recursion and iterating through all possibilities to compare against the current best predictive solution. The outcomes for each position are defined in a below section of this paper, but in summary, this RFE model informed that for the rb, wr, te, and qb positions, the optimal number of features to be considered were 6, 8, 9, and 5 respectively and results in accuracy scores of .9576, .9061, .9697, and .9322 respectively.

3. **Based on our predictive models, are there players that have been 'snubbed' in years past for Pro Bowl selections, or that were selected for the Pro Bowl that perhaps should not have been?**

Using the logistic models we developed with our data, we were able to assign Pro Bowl selection probabilities for every player in our dataset. With these probabilities, we could make a determination if a player should have made the Pro Bowl but did not. Similarly, we could make a determination if a player should not have made the Pro Bowl but did. We called these Snubs and Flakes respectively. We found that there were a considerable number of players categorized as both Snubs and Flakes in both positions; however, there were more Flakes than Snubs. Some players had a 90% probability of receiving a selection but came up empty-handed. On the other end, some players had a less than 1% chance of receiving a selection but still made the team. We recognize that Flakes are just a fancy word for False Negatives and Snubs are False Positives, so we were able to examine our confusion matrix to quantify this data as well.

**Part 3: Data Sources**

In our proposal, we cited Pro Football Reference https://www.pro-football-reference.com/years/2021/) as the website we would be using as our main data source. When we first explored the site, it appeared to have all the information we needed about players including positions, years, statistics, etc. However, upon further analysis,

we realized that a substantial amount of data was missing on the site, namely player position. Because we wanted to run analysis for Pro Bowl predictions by player type, this was a key piece of data that we were not willing to sacrifice. Ergo, we began looking for another data source that had a more complete list of players and their corresponding positions and statistics for any given season. After researching other datasets, we decided to use the Fantasy Data website (https://fantasydata.com/nfl/fantasy-football-leaders?position=5&season=2010&seasontype=1&scope=1&subscope=1&startweek=1&endweek=1&aggregatescope=1&range=1), and extracted datasets by position type from the last decade. As mentioned in our proposal, our analysis focuses on four skill positions: quarterback, wide receiver, running back, and tight end. These datasets included players statistics for receiving, rushing, and fumbling (and passing for quarterbacks). In our analysis, we aim to identify which of these statistics/which combination of statistics is most predictive of a player's nomination for the Pro Bowl at each position. We conducted separate analysis for each position, in order to achieve the most accurate results possible.

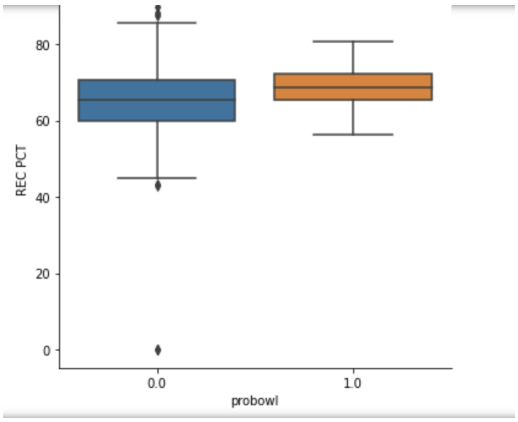
Another important component of data wrangling for this project was collecting data on Pro Bowl selections. In order to make predictions about the Pro Bowl, we plan to utilize past Pro Bowl nomination data as a binary variable in our logistic regression model, to use as a reference to identify how accurate our predictions truly are. Thus, when looking for data sources, we decided to use a separate dataset of Pro Bowl football nomination rosters throughout the past decade from the Football Reference website.

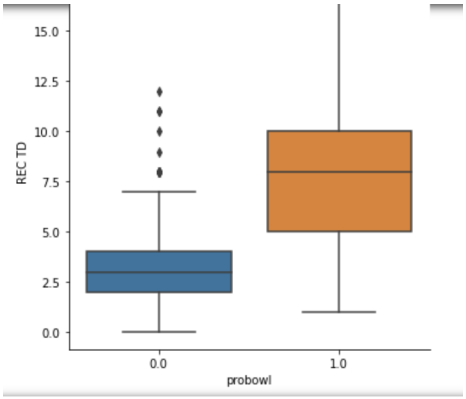
Once all the data had been wrangled for each position and each year, we had 40 excel/csv files that we imported to Jupyter Notebook of player data (10 years x 4 positions = 40 total csvs), and an additional 40 excel/csv files of Pro Bowl nomination data. In order to combine the Pro Bowl data with our player statistic data, we merged on the "NAME" column, providing us with a dataset for each year from 2010 to 2020 that also included whether or not the player had been nominated for the Pro Bowl. Finally, in order to further simplify the data and get it into a more readable format, we combined the data from each year to create one large dataframe with data from the past decade for each skill position (quarterback, wide receiver, running back, and tight end). After completing the wrangling and cleaning stage of this project, we had four dataframes to work with and run analysis on.

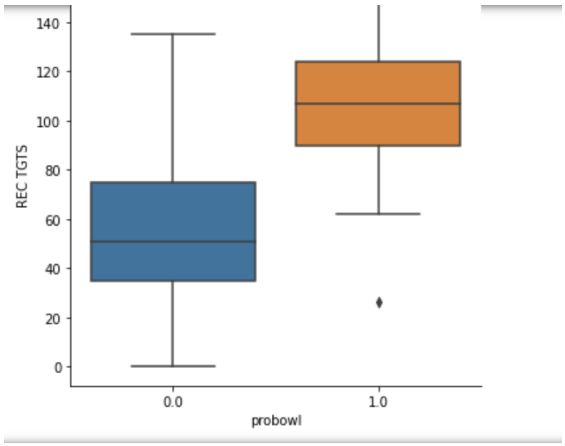
**Part 4: Results and Methods**

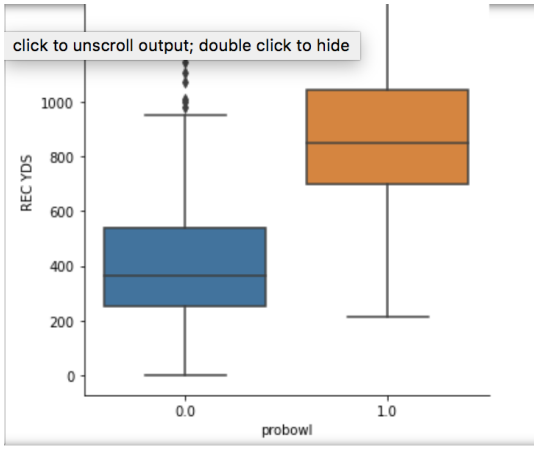
As mentioned in the data analysis section, before conducting any data analysis, the first step was to wrangle and clean our data. The result was that the data was converted into a more readable and concise format, and allowed for analysis to be completed more efficiently. When assessing how we wanted to address our research questions, we decided that an integral first step would be to plot the data and visualize some of the statistics. In order to do so, we utilized box plots. In each plot, the x axis was Pro Bowl nomination (1 for Pro Bowl nomination, 0 for no Pro Bowl nomination), and the y axis was a player statistic. There are two boxes in each histogram that we created, with the left one (blue) representing the data for non-Pro Bowl players and the right one (orange) representing the data for Pro Bowl players. In our Jupyter Notebook, we created a separate cell to plot histograms for each player type. In many of the graphs, we see a common trend where the average value is much lower for non-Pro Bowl nominated players as opposed to Pro Bowl nominated players. This makes sense in the context of the data because Pro Bowl nominated players should, on average, be expected to have higher averages in key statistics

than their non-Pro Bowl nominated counterparts. Below are some of the plotted histograms for tight ends (we chose to display our findings for tight ends at random among the four options with the intention that it is representative of the histograms for the other offensive skill positions). It is clear from these graphs that certain statistics have more variation than others in terms of the difference in averages between Pro Bowl vs non-Pro Bowl players. For example, the difference in the averages of Pro Bowl vs non-Pro Bowl players is much greater for the REC YDS (receiving yards) statistic, as opposed to REC PCT (team winning percentage). Therefore we can conclude that REC YDS may have greater predictive power when it comes to assessing if a player is likely to make the Pro Bowl. We conducted this same analysis with a variety of different statistics for all four skill positions. Through running this preliminary analysis, we were able to visualize how each of the different statistics affects Pro Bowl nominations for each position, which gave us a deeper understanding of the variables we were working with.




After looking at these initial visuals, the next step of our analysis involved running logistic regressions for each of the four skill positions. In order to do so, we imported LogisticRegression from sklearn.linear_model, as well as accuracy_score and plot_confusion_matrix. Utilizing scikit-learn made it easy to run multiple logistic regressions. In each logistic regression, our target variable was the binary "probowl" variable in our dataframe defined as 1 for Pro Bowl nomination and 0 for no Pro Bowl nomination. The data being used in the regression was every numerical statistic in each of the skill position datasets. For example,

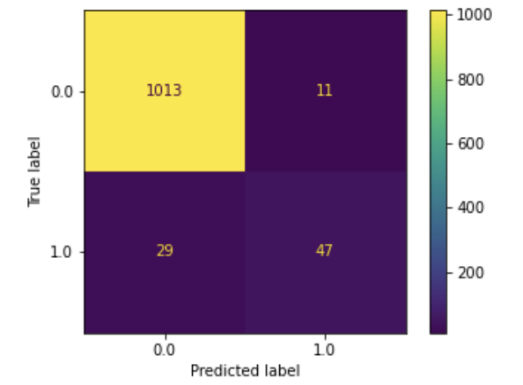
for the running back position this included the variables: "GMS", "REC TGTS", "REC REC", "REC YDS", "RUSH ATT", "RUSH YDS", "RUSH AVG", "RUSH TD", "FUM", "LST." For each skill position, we also ran two regressions; one was a simple regression in which we did not divide the data into training vs test data while the other regression involved splitting the data into training vs testing groups and then running a regression to predict the test data. Both of these types of regressions are outlined in detail for each skill position below.

Simple Logistic Regression Results

When running logistic regression on each dataset as a whole, we get the below accuracy scores for each skill position.

| Skill Position | Accuracy Score |
|---|---|
| Quarterback | 0.9273885 |
| Running Back | 0.9636364 |
| Tight End | 0.9345455 |
| Wide Receiver | 0.9390909 |

For each skill position we also plotted a confusion matrix, like the one pictured below. This provides insights into the breakdown of the predictions from the logistic regression. For example, in this confusion matrix we can see that of the 1100 total running backs, our logistic regression accurately predicted 47/76 Pro Bowl nominations and 1013 non Pro Bowl nominations. This left a total of 11 false positives (players who were predicted to be nominated for the Pro Bowl but were not) as well as 29 false negatives (players who were not predicted to be nominated for the Pro Bowl but were). This same analysis was run for each of the other three skill positions, and can be viewed in our Jupyter Notebook file linked at the end of this section. Although this data provides us with some insights into how to predict Pro Bowl nominations and which statistics are the best indicators of Pro Bowl nominations, it is also a bit skewed due to the nature of the regression. This regression was run on the entire data set and did not account for training vs testing data. Thus, we can think of these regressions as “cheating” in a way, by returning accuracy scores that may be more precise than when data is separated into training and testing data. These were the first regressions we ran, and after running them, our group realized that an important subsequent step in our analysis would be to account for training and testing data. Thus, while we thought this data was still important to include and analyze, the section below outlines logistic regression results using training data, which provide more accurate accuracy scores.

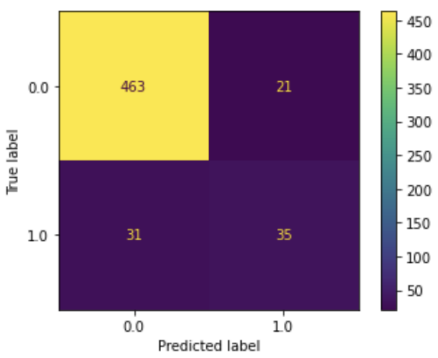


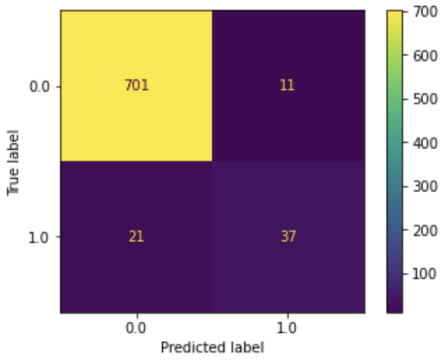

Logistic Regression with Training vs Testing Data Results

In order to split our data into training and testing data, we utilized the train_test_split function from sklearn.model_selection. We held 30% of the data for testing, and left the other 70% of the data for training the model. Below are the results for accuracy scores based on logistic regression that utilized training vs test data.

| Skill Position | Accuracy Score on Test Data |
|---|---|
| Quarterback | 0.90545 |
| Running Back | 0.95844 |
| Tight End | 0.90909 |
| Wide Receiver | 0.93506 |

As can be seen from these results, after fitting the data using the training data, we can predict whether players from the test data will make the Pro Bowl or not with accuracy ranging from 90.545% to 95.844% based on skill position. These accuracy scores are fairly precise, depending on how we define “precise,” in which we mean depending on what we want to use our analysis for. For example, if we are using this model to predict future nomination and place bets, we may want to look for more variables to include in the model or find other ways to increase the accuracy percentage to be in the high 90s for each of the skill positions. However, if we simply want to look at historical data and try to predict Pro Bowl nominations without anything at stake financially, it appears that our accuracy scores are fairly precise and can provide solid predictions for future test data based on the training data we fed the model.




Quarterback Matrix

Running Back Matrix

KNN Model

Along with a logistic regression model to predict Pro Bowl selections, we created a K-nearest neighbors model to predict Pro Bowl selection for each skill position. We wanted another model that took a different approach, and we wanted to compare the accuracy difference between this model and our logistic regression. We approached the model in a standard manner, splitting the data into testing and training, with a 30% and 70% split respectively. We used random seed 2162021 to randomize the split. We then performed a gridsearch to identify the best value for the K hyperparameter at each position. Finally, we found the accuracy of our model on our training and testing data. These results are summarized below:

| **Position** | **n_neighbors** | **Training Accuracy** | **Testing Accuracy** |
|---|---|---|---|
| Quarterback | 6 | 0.91074 | 0.88135 |
| Tight End | 28 | 0.91428 | 0.91515 |
| Wide Receiver | 7 | 0.94285 | 0.93939 |
| Running Back | 12 | 0.96493 | 0.95757 |

In general, we found our KNN model to have great accuracy, especially in the Running Back position. However, the model did not outperform our logistic regression model, and in fact many of the accuracy percentages are extremely close to the results yielded from the logistic regression model. The wide receiver and tight end position accuracy scores of the KNN model outperformed the logistic regression model, however, the opposite is true for quarterback and running back. One thing we found interesting was the n_neighbors hyper-parameter for Tight End. This value is much higher than the hyper-parameters for the other positions, we are unsure if this is a quirk of the position or just randomly occurred.

Biggest Snubs and Flakes from Pro Bowl Nomination

With the logistic model, we can make some interesting observations between what our model predicted would happen and reality. We wanted to examine which players, in the four positions between 2010-2020, were not likely to make the Pro Bowl but did and were likely to make the Pro Bowl but didn't. Those that made the Pro Bowl and, according to our model, should not have, are called Flakes. Those that did not make it to the Pro Bowl and, according to our model should have, are called Snubs.

To find the probability that each player in our dataset would make the Pro Bowl, we used the predict_proba function to append a row of probability. Predict_proba returns the probability of a 0 output in the logistic model, so we had to subtract by one to find the predicted probability of a player making the Pro Bowl. We then selected players that made the Pro Bowl and sorted by probability ascending to find the flakes, and selected players that did not make the Pro Bowl and sorted by probability descending to find the snubs. We did this for each of our four skill positions. We also plotted the confusion matrix for each logistic regression. The confusion matrix gives us an easy way to identify how many snubs and flakes there are by simply looking at the values in the first and third quadrant. Here is a summary of the results:

**Flakes:**

| **Position** | **Biggest Flake** | **Pro Bowl Prob** | **Number of Flakes** |
|---|---|---|---|
| QB | Andy Dalton ‘14 | 0.048145 | 37 |
| TE | Zach Miller ‘10 | 0.001114 | 25 |
| WR | Mecole Hardman ‘19 | 0.000347 | 45 |
| RB | Darren Sproles ‘16 | 0.003511 | 29 |

**Snubs:**

| **Position** | **Biggest Snub** | **Pro Bowl Prob** | **Number of Snubs** |
|---|---|---|---|
| QB | Matt Ryan ‘18 | 0.900013 | 20 |
| TE | Jordan Reed ‘15 | 0.953479 | 11 |
| WR | Pierre Garcon ‘13 | 0.903081 | 17 |
| RB | Alfred Morris ‘12 | 0.940765 | 11 |

SKL Feature Selection Technique RFE

To attempt to improve our model, we ran a feature selection using a SKL RFE function learned from outside of class, which determined the most important factors in the dataset for

determining a Pro Bowl player. This process is essentially feature ranking with recursive feature elimination. Given an external estimator that assigns weights to features, the goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through any specific attribute. Then, the least important features are pruned from the current set of features. This procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached. After importing sklearn datasets, feature_selection, and SVR, we arbitrarily chose five n_features_to_select. We started with the running back position and imported RFE and SVR from the sklearn package. We then set the predictor and target values, ran the logistic model fit, and then applied the RFE function. The printed result displays which columns were selected as the five best predictors of Pro Bowl nominations. However, we then wanted to find the optimum number of features to consider that will result in the highest accuracy possible given our dataset. To achieve this, we essentially ran a loop that iterated through all potential numbers of optimal features to be used as prediction to test each possibility and observe their accuracy. A limitation of this is if the amount of columns is extremely large, the code will take a very long time to run. After determining the optimal number of features, I then printed the accuracy and the column names that were selected. For rb, the optimal number of features was 6, the accuracy was .9576, and the best factors were GMS, REC TGTS, REC REC, RUSH AVG, RUSH TD, and FUM. We then repeated this process for the other three positions. The optimum number of features for wr is 8 with a score of .9061 with best factors of GMS, REC TGTS, REC PCT, RUSH ATT, RUSH YDS, RUSH TD, FUM, and LST. The optimum number of features for te is 9 with a score of .9697 with best factors of GMS, REC REC, REC TD, RUSH ATT, RUSH YDS, RUSH AVG, RUSH TD, FUM, and LST. Lastly, the optimum number of features for qb is 5 with a score of .9322 with best factors of GMS, CMP, ATT, PCT, and AVG. We can see that the highest score achieved is for te, which has a score of .9697 and 9 factors. The lowest score is for wr, which has a score of .9061 and has 8 factors. These observations are interesting because we are able to see consistently over four large datasets, which variables are best for predicting Pro Bowl nominations for different offensive skill positions. However, one hesitation with reading into this too much is that there are non-statistical factors at play here that we fail to account for, as already mentioned. Additionally, while some statistical measurements might be representative of typical Pro Bowl nominated players, they are not valued as much for certain skill positions. This model does not take that into consideration, placing as much value in rushing yards for tight ends as receiving yards for tight ends.

Box Folder Code

Our Jupyter Notebook file as well as all the individual csvs used to conduct our analysis for this project (including csvs for each year and each player position type) can be found in this Box folder: https://duke.box.com/s/kg0d59onvqjpa1fmp8a67ke55x9d59p3

(Please note that our goal as a group was to utilize the skills learned in this class to wrangle and clean the data in Jupyter Notebook using numpy and pandas functions, so much of the data found in the Box folder is in individual CSVs. There are many files and the data can be viewed in a more concise format in the Jupyter Notebook file itself.)

**Part 5: Limitations and Future Work**

While our group has worked tremendously hard to ensure the integrity of our project, there are certain limitations we faced and certain parts of our analysis that could be expanded on by future analysts going forward. A notable limitation of the dataset we used as the basis for our analysis was that the data we employed did not contain information on players' divisions within their respective conferences. While the Pro Bowl as an event is tailored to include all the best players at each position across the two conferences, there is potential for consideration to be given to try and create relatively even representation across the league's eight geographic divisions. Therefore, colleagues who may look to expand on our research in the future may find it valuable to employ this information.

Additionally, the Fantasy Data dataset which we utilized for our analysis did not contain some secondary statistics which help to fully demonstrate an individual player's talent and their merit with regards to making the Pro Bowl. For example, while they may not necessarily be the most prolific pass-catchers or runners in the league, certain running backs and wide receivers thrive in their roles as return specialists, using tremendous speed and instincts to affect a game through Special Teams plays. Further, an incredibly important part of playing the position of Tight End is the ability to serve as a pass and/or run-blocker for one's team. While we did not have access to statistics in regards to these areas of the game from the dataset we chose, their merit as additions for researchers is undeniable. As we saw in our analysis of the biggest "flakes" and "snubs" from recent Pro Bowls, two players - Mecole Hardman and Darren Sproles - stood out as extreme outliers as individuals who we failed to predict to make the Pro Bowl. Knowledgeable NFL fans would note that these two players have served as two of the most dynamic kick returners in recent years, so expanding on our available data to include such statistics would absolutely be beneficial to extend our results with greater accuracy.

Further, there are many observable aspects of the NFL game that serve as determinant factors of a player's performance and prestige that are near impossible to quantify into accessible data, and the absence of these metrics serves as an additional limitation to our project. For example, intangible traits such as leadership and work ethic are important axioms of the game of football which drive the outcomes of games but do not necessarily show up in individual statistics, thereby contributing to an upper bound on the predictive power of our model. In addition to these unobserved factors, there are non-football dynamics which also contribute to an individual's selection to a Pro Bowl. Given that Pro Bowl selection consists of one-third fan voting and two-thirds voting amongst players and coaches, factors such as player/team popularity, inter-team rivalries, and/or team performance (and therefore media coverage) may play a role in selection decisions. It is unquestionable that players on the Dallas Cowboys will tend to have an easier path to making the Pro Bowl as compared to players on the Jacksonville Jaguars given the differences in the sizes and scopes of the teams' fan bases. Thus, if future researchers looked to include information and data on yearly team performance and rankings of fan basis, one could expect to see further benefit to the predictive power of our model.

Some future work that could be done in this area is expanding the range of positions to every position. Of course, these four positions don't even cover the entirety of the offense, let alone the defense. We chose these positions because their success can easily be measured by

their statistics, whereas a lineman or defensive player does not have the same quantifications. Perhaps the model would have to consider more advanced categorical variables associated with a player to make an accurate assessment on these positions. This leads into the next point of improvement in future works: more advanced variables. For the most part (outside QB rating), the variables we looked at for each player were basic counting statistics like reception yards, completions, and rush yards. There exists more advanced unique statistics that are made from these base statistics and other metrics. An inclusion of these stats could lead to a more accurate model that accounts for further relationships in those base statistics. These advanced statistics could also help a weighting limitation that we observed in our modeling. We took a lot of statistics for each position, and some of those statistics or variables are much more important in valuing a player than others. For instance, we examined Wide Receivers receiving yards and rushing yards. While a WR might be considered more valuable with good rushing yards, it is the receiving yards that really matter to a WR. Advanced statistics could account for this and improve the model, or future works could account for this discrepancy in their model. Finally, a weighting system for both the current and future advanced metrics in the regression analysis may improve the power of the project overall.