B01590499

Machine Learning Capstone: Predicting Arrests at NFL Stadiums

Background

I have been a fan of football, and the NFL more specifically, since I was a kid. The detail in which players execute sophisticated concepts play after play has always intrigued me, and been overall very fun to watch. It has only been more recently that I have become aware of the danger that fans pose to each other. Often facilitated by alcohol and the event in the game itself, fans at some games get much more rowdy than at others, and I think that being able to predict misbehavior in some form would be a very interesting project.

For my capstone project, I am proposing a project that is able to predict the number of arrests at a given NFL football game. This is a very important predictive task as with a successful model, one would be able to assess the level of risk for fans going to these NFL games. Additionally, in the hands of stadium staff, a model like this would be able to alert ushers and building security to potential threats or rule-breaking before they ever occur.

There are, however, a plethora of potential challenges to achieving this goal. Primarily, there is naturally lots of noise or unpredictability when working with a target variable like arrest counts. This is always the case when trying to predict human behavior, but even more so when predicting a very niche variable like the number of arrests at a game. Additionally, the number of arrests can prove to be a particularly hard variable to predict, as there is lots of uncertainty in the actual arrest count, as for an arrest to be made, someone has to break a law, a police officer has to observe this lawbreaking activity, and they also need to catch the violator and then make the arrest. Something like “rowdiness scale”, if it was collected in an unbiased and accurate way, may be slightly easier to forecast.

Dataset

The dataset itself is an accumulation of public records of police arrests at NFL stadiums, and is hosted publicly on Kaggle. The data was collected through requests by the Washington Post to police departments who deal with security at each individual stadium. The data is complete with all arrests made at these stadiums between the years 2011 and 2015, and has a few limitations. Only 29 out of the 31 police jurisdictions provided these data to the Washington Post, and incomplete or unreliable data were removed from the overall dataset. Some jurisdictions only provided partial data, including Buffalo, Miami, and Oakland, and additionally St. Louis provided yearly data as opposed to game level data. Finally, Detroit, Minneapolis, and Atlanta did not include parking lot arrests from their respective stadium’s data.

The first feature in the dataset is the season in which the game took place. This variable is categorical and is represented by the year in which the season began, additionally providing group structure to the dataset. The next feature is the week number within the season in which the game took place, and is an ordinal integer variable. Next we have the day of the week that the game took place, though a great majority of these games will have taken place on a Sunday. This is of course a categorical variable, and is represented as a string. We are also given the local time in which the game began, which is a continuous variable and is represented as a date object. We

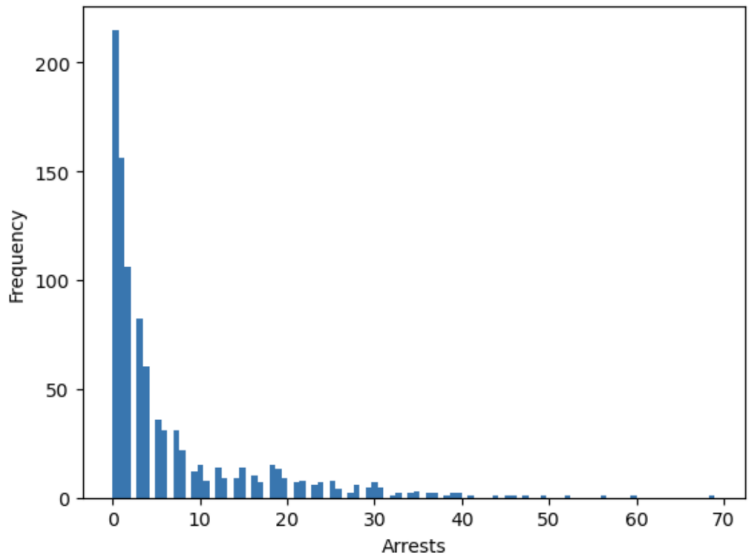
are given the home and away teams as individual categorical features, both represented as strings. Additionally, we get the home and away scores, both represented as integers and are continuous variables, along with whether the game went to overtime or not, represented as a boolean categorical variable. Finally, we are given a boolean representation of whether the game took place between two teams in the same division, followed by our target variable - a continuous representation of the number of arrests that occurred at the stadium for this game.

Intuition and Analysis

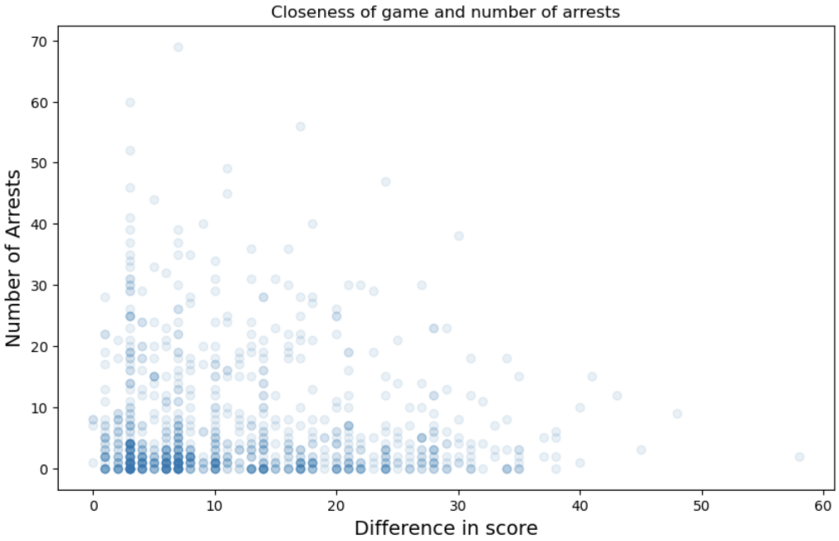
Of these raw features, there are ones that intuitively seem most likely to be useful in predicting the number of arrests. First, I would say the week number that the game happened in could have some predictive power, as intuitively the games in later weeks tend to be more important than earlier games as they have playoff implications. With more intense games, one might expect a rowdier crowd and potentially more arrests. I think the day of the week that the game is played on will have little correlation with the target variable, but the local start time of the game could be an important feature. I would expect games that are played later in the evening to potentially have more people consuming alcohol, and potentially more alcohol consumed per person. With more alcohol, I would definitely expect to see more arrests made. While I would not expect the away team to have a huge impact on arrest counts, I think the home team itself could be indicative of arrests. Fan bases are very different from team to team and city to city, so there may be some fan bases that are rowdier and get into more trouble than others. I could also potentially foresee some away teams being more hated than others, resulting in some uptick in arrests when teams are at home against the hated team. Next, I don't really think the home or away score will have much of an impact on their own towards the arrest counts. However, I think that with some feature engineering, we could create a feature such as the absolute value of the difference in scores to give us a metric for closeness of the game, and additionally, determine who won the game by the home and away scores. I think each of these engineered features, and potentially others as well, would be quite predictive towards getting an accurate estimate of the arrest count. Additionally, it makes intuitive sense to me that if a game goes into overtime, and is thus much more intense, that there would be a greater chance to see arrests as tensions are high. Finally, I foresee the divisional game boolean feature being quite predictive, as games against a true rival are always very tense.

Another very interesting use case for a model like this one is the interpretability of it. Once a model that minimizes the empirical risk has been implemented, the analysis of the model will be very interesting. For example, if it is determined that the closeness of the game is the most important factor in determining the number of arrests made, then security officers at NFL games can know to be on high alert when a game gets really close, and can maybe relax a little bit if the game is a blowout. However, if the home team is the most predictive of the number of arrests made at the stadium, then it may be a good idea to simply increase the security at these stadiums, or potentially serve less alcohol. Overall, the analysis of the model will not only be interesting in order to see what factors contribute to rowdiness or arrest counts, but they can also be used to try and reduce the number of arrests at these games.

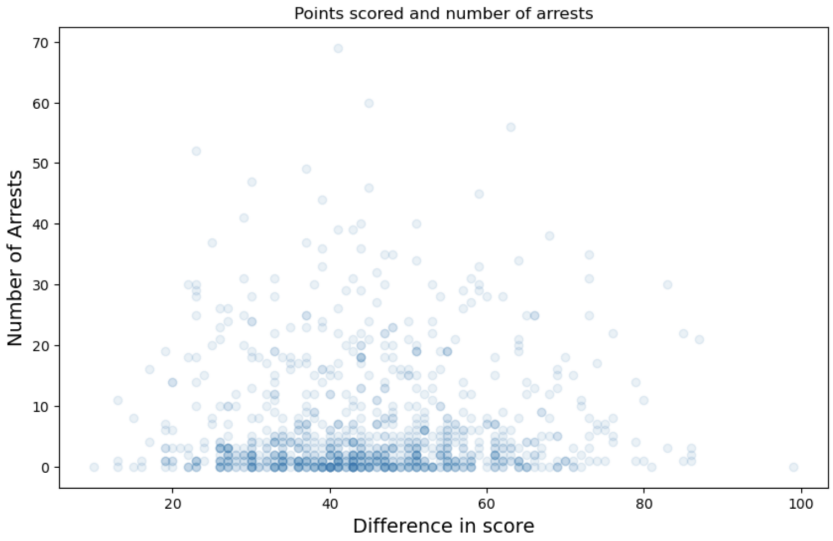
EDA



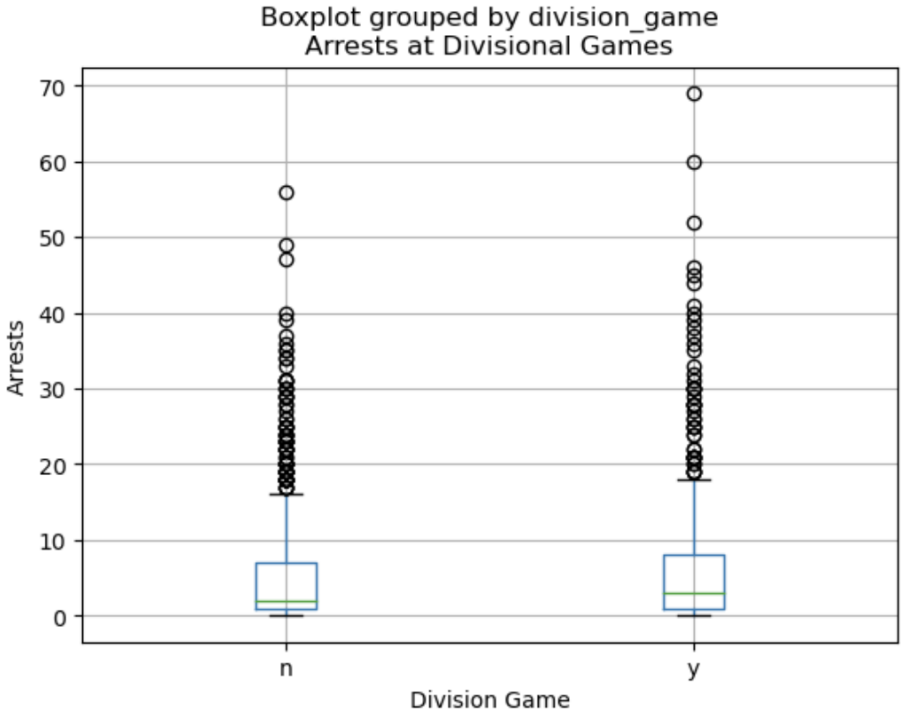
Seems to show that the target variable is a very skewed distribution, with most of the data points towards the lower end of the arrest spectrum, but some games having lots of arrests.



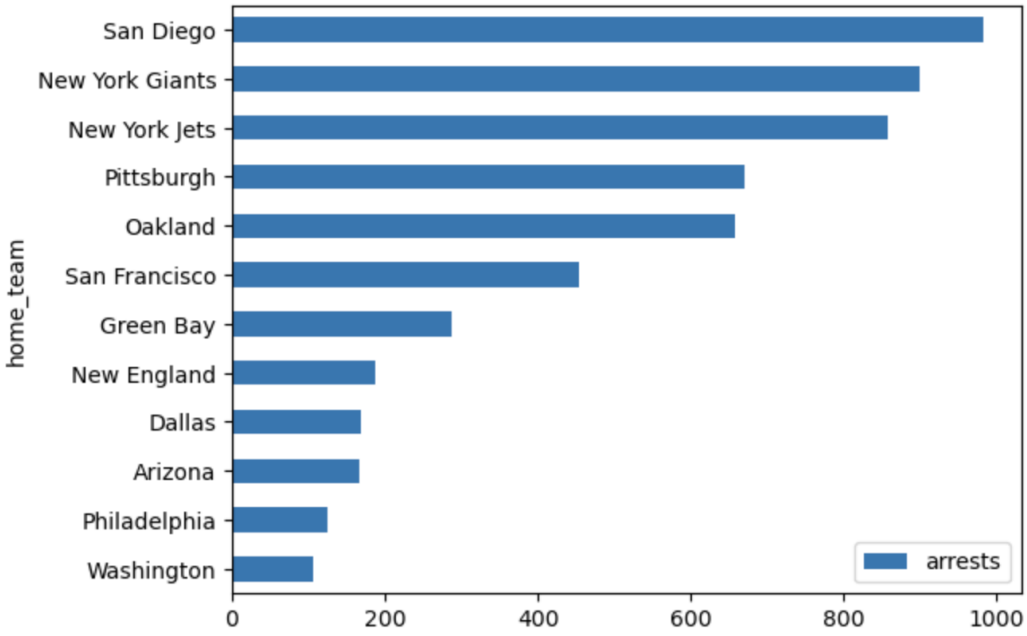
The closeness of the game seems to follow a weak pattern that as the games get less close, the number of arrests decreases.



The difference in score seems to have no impact on the number of arrests.



There does seem to be a noticeable difference in the distributions of arrests for divisional games and non-divisional games, with more arrests seemingly taking place at divisional games.



There is a noticeable difference in the total arrest count from this five year period for different stadiums, with San Diego, New York, Pittsburgh, and Oakland leading the pack here.

Models & Results

Before any modeling could take place, the engineering of features from the raw values in the dataset was a necessary step. First, the local gametime feature was converted to a categorical variable with values of afternoon, evening, and night, in order to simplify this numerical feature. Then a few features were created using the scores from both the home and away teams. A points scored feature was generated by adding these two scores together. A score difference feature was created by taking the absolute value of the difference in these two scores, meanwhile another score difference feature was created by taking the home score less the away score (with no absolute value). Finally, a home team win feature was created by comparing the score of the home team and away team, with the three outcomes being a win, a loss, or a tie (very rare - only 4 instances in 5 years).

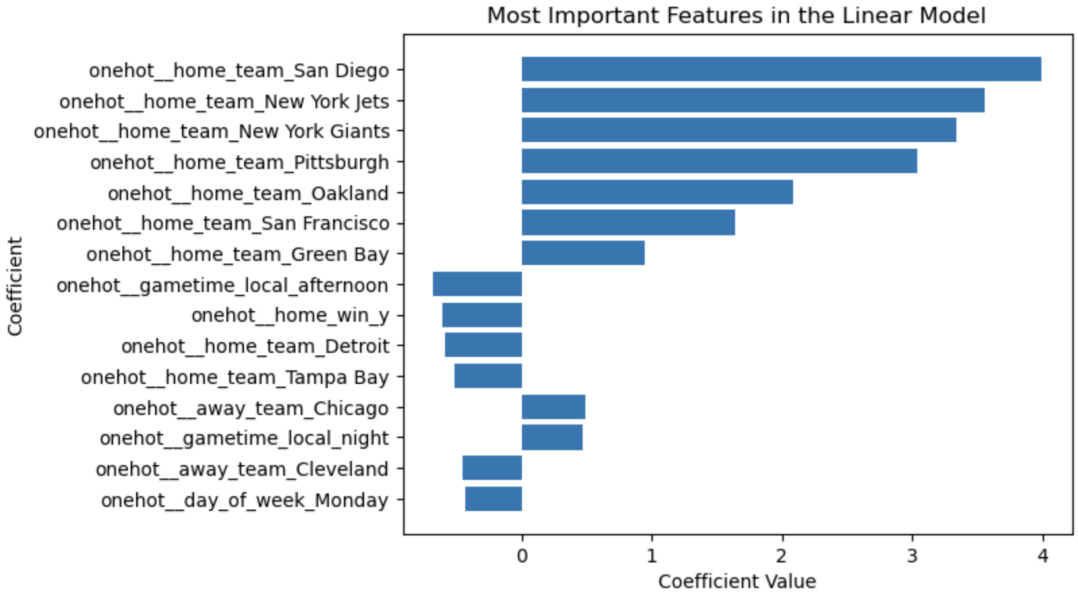
Once the dataset was completed, there was a bit more preprocessing that needed to be done. All of the categorical features needed to be one-hot encoded into their own features, the ordinal features needed to have numerical values assigned to each category, and the continuous features needed to be standardized. After this preprocessing, it is best practice for all remaining features to be standardized as well, such that when interpreting coefficients to a linear model, all features have a mean of zero and a standard deviation of one. This allows us to be able to compare the coefficients directly for determining feature importance. The dataset was split into training, validation, and testing, with a random split of 60%, 20%, and 20% of the data points into each section, respectively.

After preprocessing was complete, four different machine learning algorithms were utilized in order to predict the number of arrests that take place at an NFL game. These algorithms are: simple linear regression (OLS), linear regression with L1 regularization, linear regression with L2 regularization, and support vector regression. Linear regression, the algorithm with the least complexity, had no hyperparameters to tune, so we can simply fit the regression model on the training data and then test for correctness on the validation set. This model scored particularly poorly on the mean squared error (MSE) metric for the validation set (MSE of 3e24).

Linear regression with L1 regularization has the regularization coefficient to tune, of course. 15 different alphas were tried, ranging from 1e-2 to 1e2 spaced evenly in logspace. The optimal alpha, in terms of MSE on the validation set was the smallest alpha of 1e-2, and this resulted in a much better MSE of 38.1. Linear regression with L2 regularization also has the regularization coefficient to tune. Again, 15 different alphas were attempted, also ranging from 1e-2 to 1e2 spaced evenly in logspace. The optimal alpha, in terms of MSE on the validation set was again 1e-2, and this resulted in an MSE of 38.2.

Finally, Support Vector regression (SVR), a method that has two hyperparameters to tune - gamma and C - was used. 15 different gammas were tested, ranging from 1e-3 to 1e5 in logspace, and 15 different values of C were also tried, ranging from 1e-2 to 1e2 again in logspace. All 225 of the above combinations of C and gamma were used to train a SVR model, before testing with the MSE metric on the validation set. The combination of C and gamma that yielded the best result was a C of 100, and a gamma of 0.0037. This gave a MSE on the validation set of 41, thus this more complex model was outperformed by the linear regression with L1 regularization.

Overall, the linear regression model with L1 regularization yielded the best results on the validation set of data. Thus, this model was used to predict the number of arrests at NFL stadiums on the test set. This model scored a MSE of 28 on this test set. Additionally, some basic interpretation of this model was conducted. Based on the graph below, it is clear that the stadium itself is the most important predictor of the arrests that will be made at the stadium. This is a clear conclusion due to the fact that a great majority of the highest 15 regression coefficients are coefficients for the one hot encoding of a home team. However, aside from the stadium itself, the other important predictors of arrests are: a win by the home team, the local time of the game, as well as the game taking place on a Monday.



Discussion

One of the most interesting conclusions from the modeling portion of this project is that the stadium itself is the best predictor for the number of arrests at that stadium. There are quite a few reasons why this would be the case. The trivial reason for this is that often with predictive tasks, the best predictor for a particular event is the most similar event that we have access to in a data set. In this data set, there are dozens of examples of arrests at a given stadium to look at in the training set when trying to make a prediction for the same stadium in the test set. Another possible reason for this conclusion is the set of laws that the stadium police have to follow for the particular City or State, or even the leniency of the officers at the given stadium. Oftentimes different jurisdictions abide by different sets of laws, so the behavior of fans who are arrested at one stadium may not be worthy of an arrest at a different stadium. In terms of the leniency of the officers, the group of police at a given stadium and their level of activity can be assumed to be relatively constant, thus if a particular group of police arrest around four fans per game, it might be a good prediction that they would arrest four fans at the next game. This frequency of arrest may also simply be a function of the capacity of the officers at stadiums to make arrests.

The second key takeaway from the analysis of feature importance is that a win by the home team, and the time of the game are the most important non-stadium features. Intuitively, both of these features make sense for this model. A victory from the home team will no doubt give the home fans the result they were rooting hard for, and thus make them less likely to participate in any illegal activities. Whereas a loss may give a boost of adrenaline to all of the fans who paid good money to go watch their beloved team win, and in turn cause them to make mistakes they may regret. Additionally, the games taking place in the afternoon generally have less arrests, and those taking place at night tend to have more arrests. This also makes intuitive sense as drinking habits and questionable behavior tend to be a bit amplified as the sun goes down.

However, even though the interpretation of the model makes a fair bit of intuitive sense to me as the data scientist, this method of assessing an NFL game on its propensity for illegal activity is definitely not without risk. There are lots of biases in a model like this, particularly when taking a set of data from the past and using it for predictive policing in the future. For example, the football games at the stadium in San Diego are likely to have more arrests than a game in any other stadium according to the best model from this analysis. Yet, just because San Diego may have made the most arrests in the past does not mean that the fans at the San Diego stadium were deserving of the arrest that were made. This could in fact be a major red flag that is telling us something about the biases of the San Diego Police. If, for example, there are many unjust arrests made at the San Diego Stadium, and our model was picking up on this, then a risk assessment of the San Diego stadium using this model would only be perpetuating the

pre-existing biases in our society. No matter how you frame it, whenever you are trying to use machine learning to execute some form of predictive policing, there will always be questions of equity and accountability. This model and analysis is no different, as the biases from the different Cities and stadiums have clearly been picked up by this model.

The biggest limitations of this project are twofold. The aforementioned biases for each stadium are glaring, and the predictive power of the model is very limited when the stadiums themselves are the best predictors for arrests. If this project were to be continued and improved, it should consider building models for individual stadiums one at a time in order to get more personalized and hopefully more accurate results. It would even be beneficial to add in some alternative sources of data not included in this data set, such as number of alcoholic beverages sold, or even the number of fans in attendance for the game. These additional features could help remove some of the confounding variables that led to the disparity in the arrests data. I believe that individualizing the models to the stadiums on top of adding in relevant data would help reduce the bias in the model and increase the predictive power. An alternative way to lower the bias of the model would be to remove the home and away teams from the data set altogether. This way, the model would no longer be able to learn anything specific about the stadium itself, but could just use other non-identifying features to predict arrest outcomes. While this method should in theory remove the bias from the model, it is still possible that bias will remain, as features like the score of the home team can still be correlated with the home team itself (maybe some teams were higher scoring than others from 2011-2015).

Conclusion

Overall, given a dataset of arrests at NFL stadiums from 2011 through 2015, this project aimed to answer the question, can you predict the number of arrests at NFL games? Through exploratory data analysis, it was found that games are much more likely to have few to no arrests than to have several, and that intuitive metrics like closeness of the game or whether it was a divisional game seem to qualitatively have little impact on the arrest counts. However, it was observed that there is great disparity between home teams and the amount of arrests they have at their stadiums. Four models were used, and linear regression with L1 regularization performed best on the validation set, and generalized well to the test set. Upon further examination of the best model, it was found that the stadiums themselves were indeed the best predictors of arrests. While this was helpful for prediction in this task, it may instead be perpetuating instances of bias at these stadiums, if this model will be used for predictive policing. So, it is important to recognize the limitations of this project, and to correct them to the best of our abilities if we are to continue in this line of research.