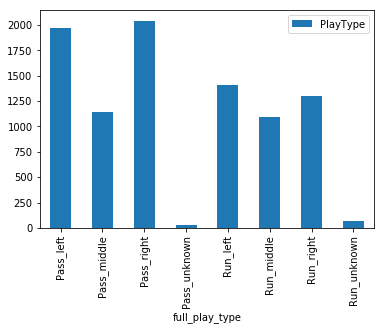
**Predicting NFL Plays for Optimum Success**

In the NFL, a single play-call can be the difference between a win or a loss. A designed run play may work well in a certain situation, while it may work terribly in a different situation. Therefore, it is vital to make use of analytics to know *when* and *where* these calls need to be made. The NFL is fairly new to analytics, so you often see new innovations and projects geared towards making insightful decisions. For example, Kaggle hosts an annual NFL data scientist competition called the “Big Data Bowl”. (National Football League, n.d.)It is a football analytics competition affording college students and professionals the opportunity to utilize historical data sets of the same player tracking data used by teams and suggest innovations about how football is played and coached. The winners are invited to Scouting combine to present their project to coaches. From this competition you can find many different viewpoints of analyst around the world and some of those projects gave me the inspiration for my own project. Some of my inspiration also came from social media as well. There’s a large community of sports analysts who interact and encourage new ideas. Ron Yurko (Twitter, n.d.)(who’s nflscrapr package has been used in many varieties) and Micheal Lopez (Director of Analytics for the NFL) are prime examples.

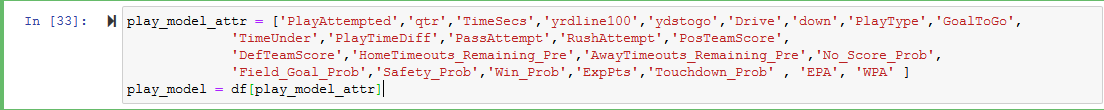
**Transformation, Data Cleaning and Exploratory Data Analysis**

Before using my dataset, I had to transform and clean it. It had many unnecessary columns, relating to specific play and player personnel. So, I selected the ones I would need for my models to predict the next play type. There were missing values in those valuable columns as well, so the next step was removing rows where this was the case. I then performed Exploratory Data Analysis on the dataset. I looked at a league wide view of the plays ran each quarter. This can be easily visualized on a bar plot by running a count and grouping by Pass attempts and Rush attempts. Then I visualized the plays ran on each down around the league in similar fashion. The Redskins have 13 distinct opponents in the 2020 season, so one wishing to build upon this model could create 13 separate datasets and run models on. To narrow my exploratory data analysis on the Redskins divisional rivals, I did the same as mentioned before on 3 distinct datasets (filtered on when the possession team was the Cowboys, Eagles or Giants)

Next, I needed to get a better understanding of the plays being ran. In the dataset, it was too vague. For example, in the “PlayType” column there were only values such as Pass, Run, Kickoff, Punt, etc. I needed to know where these passes and runs where being directed. What’s the point of telling an NFL coach that they should expect a pass for the next play, when you could possibly tell them where that pass could be thrown. This would better prepare a defense for whatever an offense may bring. I created a function that created a new column, “Full\_play\_type” that describes this. I also created visualizations using bar plots that illustrate the variance.

 See Figure for plot of the Philadelphia Eagles plays.

**Model Construction**

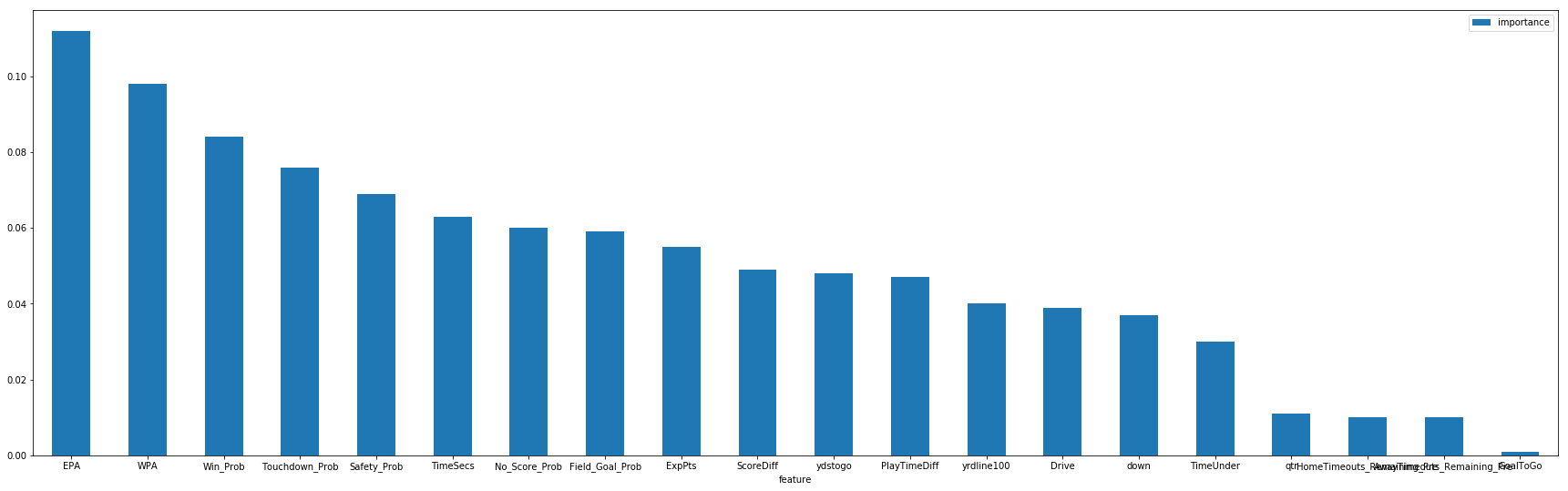
Now onto the good part. To begin my model construction, I had to carefully choose my features. Now, a lot of the columns in this data where not very useful and could potentially lower the performance of any model drastically. Out of 102 columns, I selected 25 that I believed would provide valuable information to help our model make predictions for the next play. 

I then filtered out records that I wouldn’t be using to predict a run or pass. For example, this included removing rows where the playtype was a kickoff, punt, no play, field goal, etc. I added a new column called ScoreDiff which would replace two other columns: PosTeamScore & DefTeamScore. (Used to signify the scores of the current Offense and Defense teams. Then I proceeded to fill out missing values with 0 for the model to work. Possibly in future attempts, I could fill out these values with a quarterly average value instead. Finally, I separated the datasets into two; my features table and my target values (PlayType).

After preparing the data for the models, I decided on 4 algorithms that I would be comparing accuracies and area under the curves. I choose Decision Tree Classifier, Random Forest Classifier, Logistic Regression and KNeighbors Classifier. They performed fairly well for a first run, with Random Forest being the best of the bunch. A screenshot of a social media post

Description automatically generated

I then examined the importance of each of my features to see which ones are playing the vital roles in making predictions. Unsurprisingly, EPA and WPA were the top two. (They are statistically calculated values already).



**NEXT STEPS**

For my next delivery, I will try to tune my model to get an even better accuracy. I will try out different features (even some I may have discarded). I will be aiming for around 85% accuracy across each algorithm.

# References

National Football League. (n.d.). *NFL Operations*. Retrieved from Big Data Bowl: https://operations.nfl.com/the-game/big-data-bowl/

Twitter. (n.d.). Retrieved from Ron Yurko: https://twitter.com/Stat\_Ron/