**Our ETL journey**

List of sites and sources:

1.

2.

3…

We started with 10 years of NFL financial data from 2011 through 2021 in 10 different datasets.

We primarily used Pandas to combine and clean and transform the datasets into one.

* Merge\_csv.ipynb

Once all of the years of financial data were combined we explored the data and eliminated some columns that were purely categorical and not necessary or needed to define our original hypothesis.

* NFL Financials.ipynb

We then collected the Team statistics and transformed the category\columns into a more standard format, renaming, capitalizations, teams that changed names\locations from one year to another, also dropping and cropping the dataset to conform to the 10 year period so our 2 data sets match chronologically.

* NFL\_Standings\_fixed.ipynb

Our next step was to merge the 2 datasets. We performed a left merge on “Year” and “Team” and came up with our final working dataset.

* Dataset\_merge.ipynb

Our final working dataset has 22001 rows and 24 columns

We loaded our data into a .csv called “NFL Merged\_df.csv”

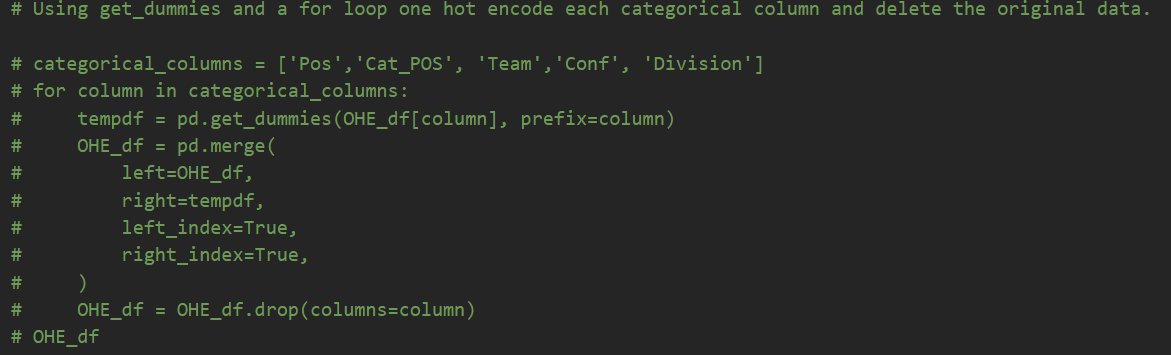
**Our Machine learning journey:**

* OHE.ipynb

We separated out 3 targets or outcomes:

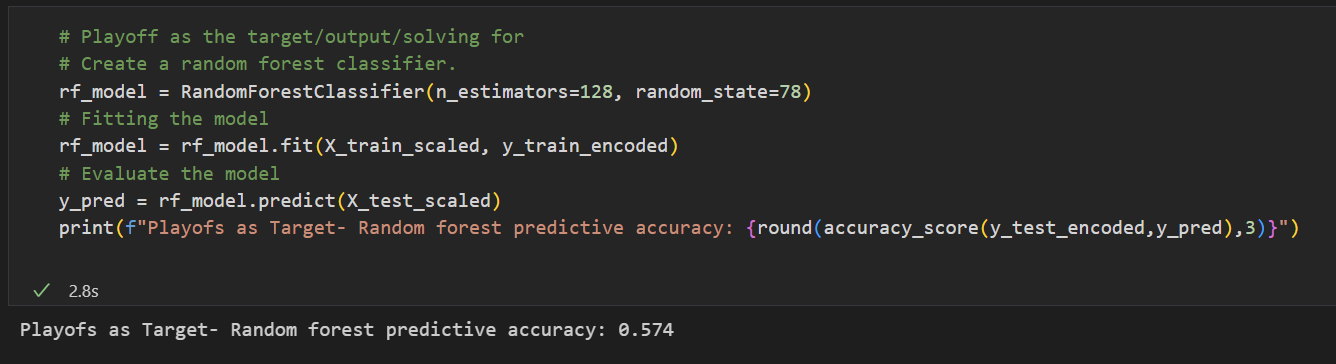
* Playoffs
* Won\_Superbowl
* Win%

We initially OHE the data using get\_dumies to keep as much information as we could by converting all of the categorical column data to numerical format so they could be included into the ML model.

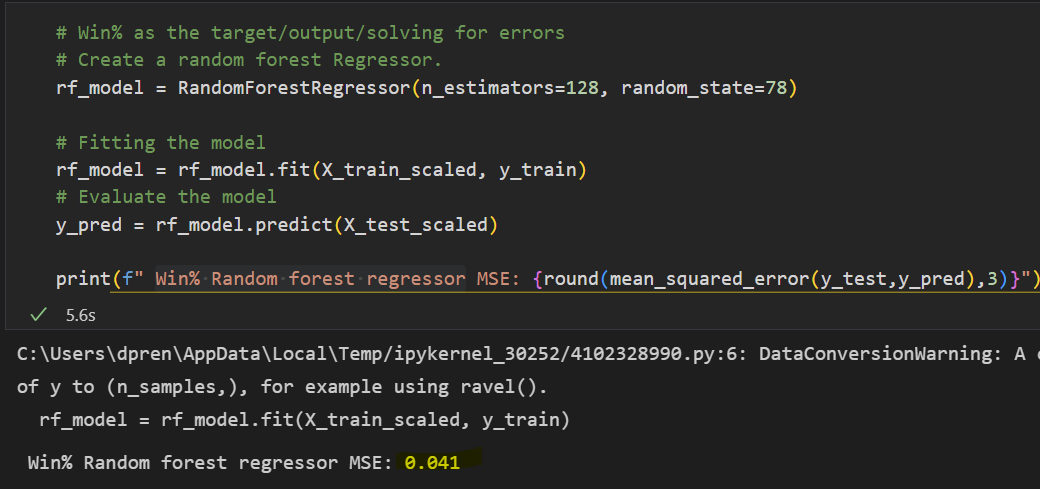


But this did not give us any advantage, and actually skewed our initial ML models.

Ultimately we dropped the entire categorical column data and did not use OHE. We also settled on using Playoffs in a Random Forest Predictive model as our main focus for the machine learning. As you can see it gave us a 57.4% chance of predicting which team would make it to the playoffs. Only slightly better than flipping a coin. This is still a significant finding if not satisfying.



Wrapping up the ML portion of this project we wanted to see if we could see how accurately the Win% data could be ratified.

We did validate our data with using Win% in a Random Forest Regressor using the MSE: mean\_squared\_error. 

This was showing a 4% error rate, which we thought was acceptable given the data we collected.