NFL Capstone - Final Report

Proposal

Drafting a quarterback in the NFL draft is a notoriously difficult thing to do. The QB who excelled in college does not always translate that success upon entering the NFL. There are some guys who were “ok” in college that go on to become all time greats. Is there a way to make this seemingly unpredictable proposition more predictable? Imagine being an NFL General Manager. Wouldn’t it be handy to be able to predict the success of any given quarterback on your draft board with high confidence? This has been on the minds of NFL fans/GM’s/coaches for as long as the draft has existed. That’s the exact problem that I am setting out to solve with my first capstone with Springboard.

Client/Problem

To put it simply, as alluded to above, the NFL overall has a predictability problem when it comes to drafting college players. Year to year, the process of selecting a great QB in the NFL draft seems to be a coin flip. If a given NFL team is in desperation mode, whiffing on the draft might cost the coach or GM their jobs. With my capstone, I want to be able to offer the reader insights as to ***who*** some of these future greats might be.

Data

For this project, I will be looking at NFL Draft and College Football Statistics data from 1970 - 2012. The NFL Draft dataset has a ton of statistics related to NFL career game performance for every player including passing yards, completions, sacks, tackles, etc (along with draft position). This will be useful for evaluating the long-term success of a player and whether or not they can be labeled a “franchise player” (A player who is often considered to be the best player on a team, or who is a younger player with the potential for a team to “build” around him). The College Football Statistics datasets gives information for every college football player and their statistics throughout their college careers. My idea is, by taking into account a response variable (Career Approximate Value) from the NFL Draft dataset along with college statistics, would there be a way for me to narrow down who the best available quarterbacks are in any upcoming draft?

Approach

My mentor, Devin, was able to share a project that a colleague of his made. In it, his colleague has a model that predicts whether a given quarterback will be a “bust” or “success”, which is very similar to what I would like to do. Ultimately, the goal for this project is to look at previous QBs who became successful and see if there might be some correlation between their success and some other variables. The simple idea is to predict who the best quarterbacks are in the upcoming draft by using the college statistics and respective career Approximate Value ratings for players that came before them. Part of my approach will also hinge on what I find during my exploratory data analysis.

Deliverables

This project could be presented in a variety of ways, but for now, I'm thinking that a Jupyter Notebook might be the best option for me. Although it would be neat to be able to present it as an interactive project - something that a reader could engage with.

Data Wrangling

For this project, I would be scraping data from Pro-Football-Reference.com using BeautifulSoup. PFR is one of the better places to collect NFL and NCAA statistics as it is revered for its seemingly unlimited potential for analysis.

I first imported the appropriate packages for wrangling the data. I then initialized the pandas dataframe I need for the NFL draft data that I would be scraping. After that, I used BeautiulSoup to scrape Pro Football Reference to get a list of players who were drafted in the NFL from 1970-2012. I then isolated only the quarterbacks in the dataframe.

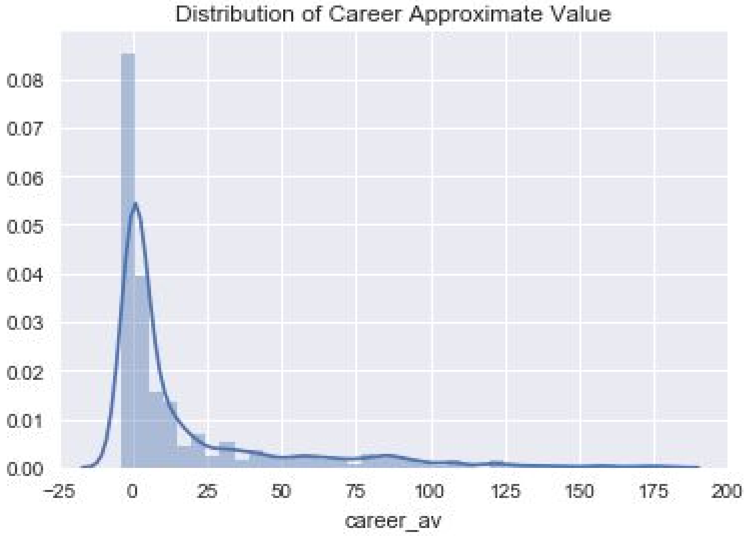
It turned out that the statistics included in the NFL draft dataframe I scraped included only the career NFL statistics for each player, not their college data (Which is what I would need for analysis later on). What I then needed to do was to get a list of names for each player in the dataframe along with a list of their respective urls for their college statistics on PFR. I had to get rid of periods in names (E.g., B.J., P.J.), apostrophes in names (E.g. O'Sullivan) and get rid of the indicator of whether or not they were eventually inducted into the NFL Hall of Fame (Denoted by ' HOF').

After getting the list of names and using that to make a list of urls for each player's college statistics page on PFR, I needed to initialize two dataframes: One for each player's passing statistics and one for their rushing statistics (Both of these categories are located in separate tables in the HTML). I then performed another scrape with BeautifulSoup to get the above statistics, but this time, while running the for loop, I had to account for bad web pages (And there were quite a few of them). Most of them were players that never really had much playing time, thus not a lot of statistics to report for them.

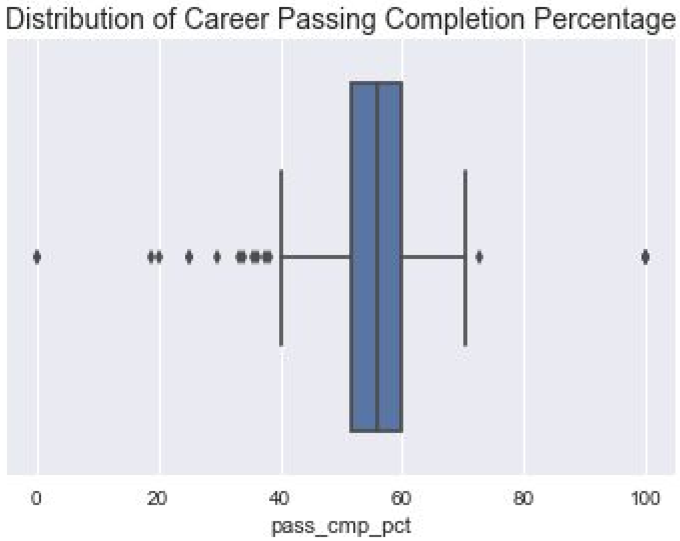
The next step in the process was merging the existing dataframes together. I performed a merge on the rushing and passing tables on the respective url links for each player. I also reordered the columns for ease of use. I noticed that some wide receivers were inadvertently included in the merged table (Which were clear due to all of the NaN values assigned to different passing categories. I rectified this issue by dropping all NaNs from the table. I had everything I needed except for the name of each player. Therefore, I had to create a function that used regular expressions to take the name of each player from their respective url links. After I used that function, I assigned a new column, ‘name’, to the merged dataframe.

The final step was cleaning the data and taking the career Approximate Value index for each player in the NFL draft dataframe and merging it with my newly merged passing/rushing college statistics dataframe. The reason for this is because I will be using the career AV as a response variable later on. After that, I dropped various columns that were deemed irrelevant for my project’s purposes (A lot of columns that were empty and/or not useful). After that, I converted all data types that needed to be to numeric instead of object for easier analysis down the road. I also filled all missing value with 0. To wrap up, I converted the dataframe to csv for quick reference later.

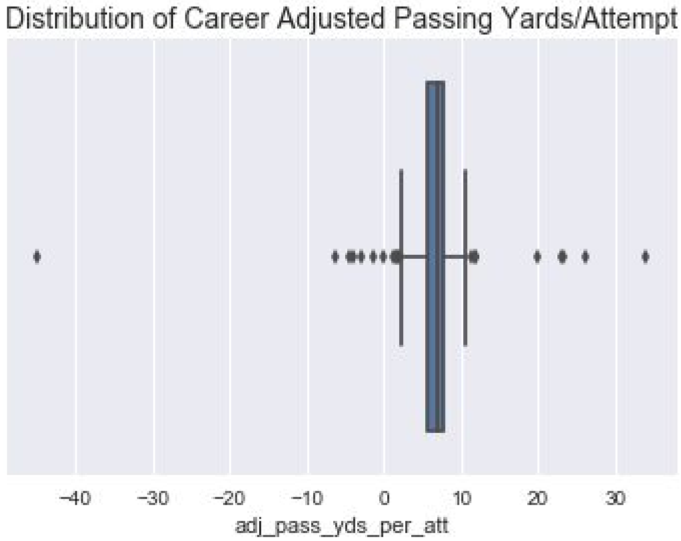
Exploratory Data Analysis



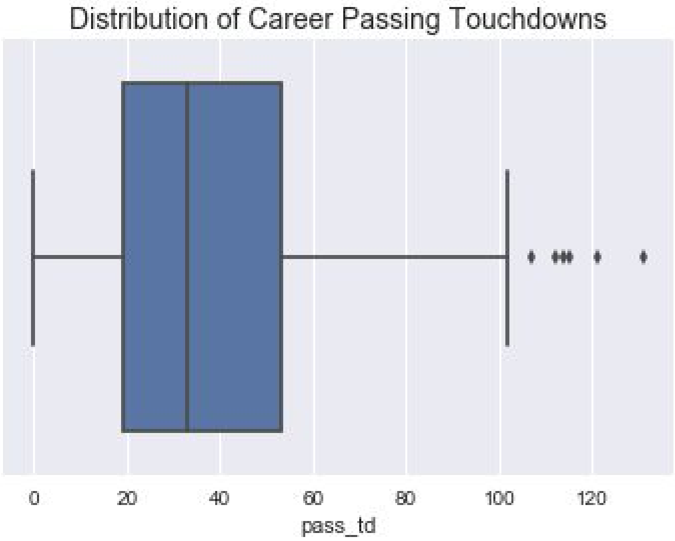
I began this section by plotting the distributions of most of the variables in the dataset. Here, I will include some of the more interesting results - such as the distribution of our target variable, career approximate value. Of all of the quarterbacks selected in every NFL draft from 1970 to 2012, very few of them have high career AV scores. The distribution centers around 0. It also has an extremely long tail - telling me that it’s very rare for a player to have a career AV score much higher than 25-30.



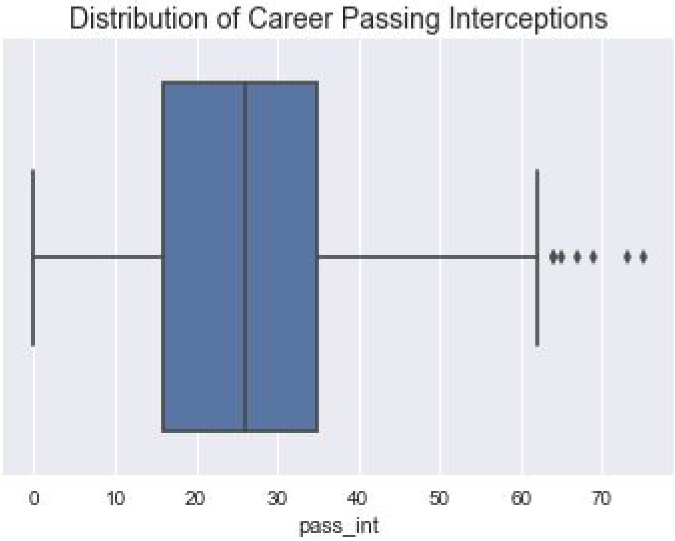
The next distribution I took a look at was career passing completion percentage. Not surprisingly, most players' percentage is around 55%. What stands out however are the huge outliers that are close to (or at) 100% career completions! After observing the data more closely, I realized that the reason these players' percentages were so high is because they only attempted a handful of passes in their college careers.



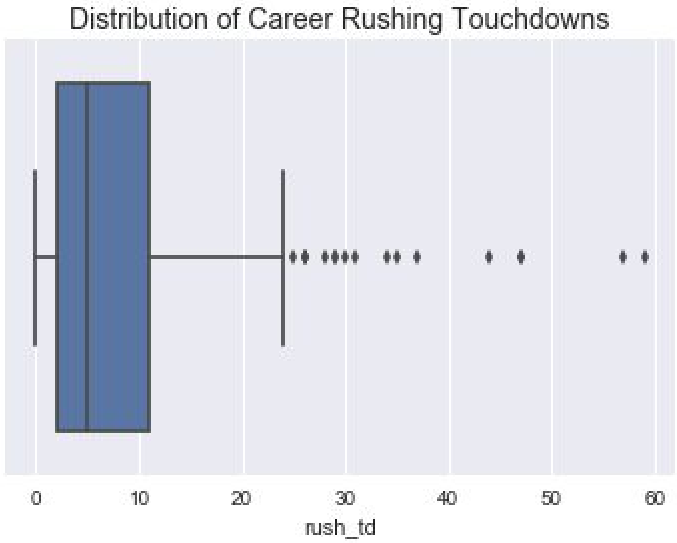
Another interesting distribution I looked at was the adjusted passing yards/attempt. Most players see their number around 7 yards or below, but there were some outliers that were much higher than that. After filtering the data frame, Most of the names that had very high adjusted passing yards/attempt went on to become household names (Sam Bradford, Russell Wilson, Cam Newton). This got me thinking that this particular statistic could be a good indicator of how good a quarterback will become in the NFL.



Career passing touchdowns is another statistic I decided to look at. The vast majority of players are between 20 and 50 touchdowns for their college careers, but there are some more pretty big outliers. After filtering the data frame to explore further, I saw names like Drew Brees, Peyton Manning and Phillip Rivers. This is perhaps another possible indicator for helping to predict great quarterbacks in the NFL.



I am particularly interested in the relationship between career AV and career interceptions. One would imagine that the higher the interceptions, the lower the career AV. I will keep an eye on this statistic during my analysis.

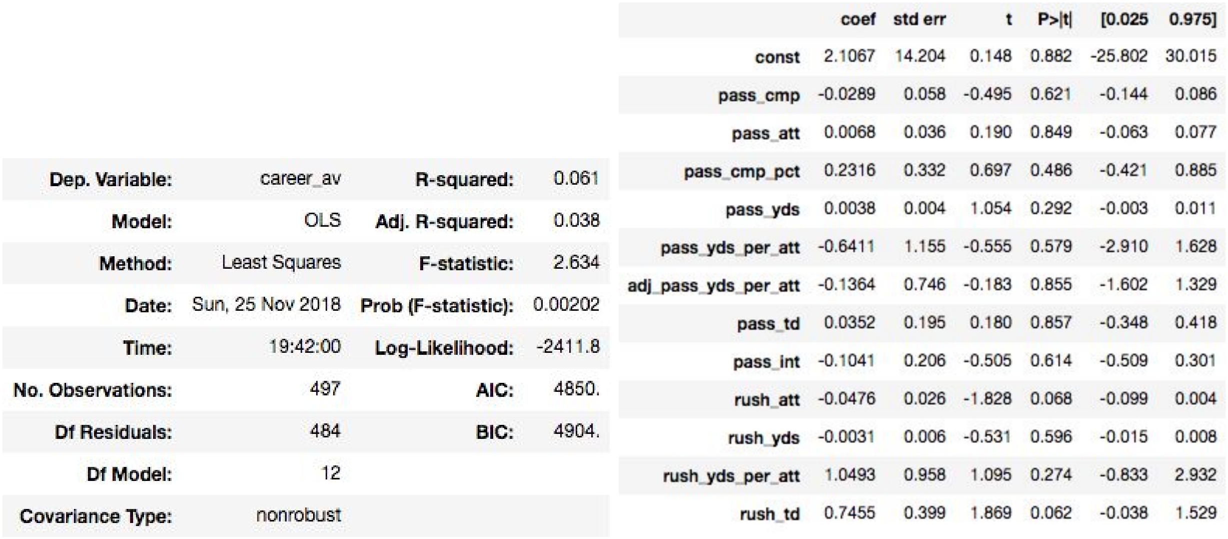


Finally I wanted to look at career rushing touchdowns. The distribution is showing quite a few outliers. Names that appear in the upper range are Vince Young, Tim Tebow and a few other notable ‘running’ quarterbacks. Historically, the NFL is a passing league and rushing quarterbacks are not usually as translatable from college to the pros as passing quartebacks. I want to see how this variable possibly relates to career AV.

Statistics

For the statistics section, I wanted to take a closer look at the relationship between career AV and the various college passing and rushing stats that I have available. I wanted to do that by using statsmodels’ multiple linear regression tool. What I did was dedicated the career AV column as my response column (y) and included all of the college passing and rushing stats columns as explanatory columns (X).

First Attempt



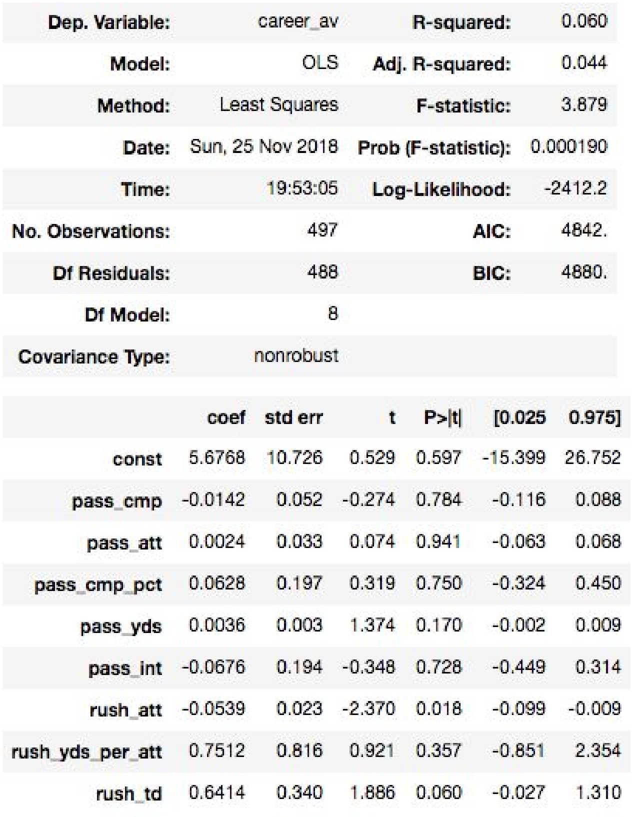
Dep. Variable:	career_av	**R-squared:**	0.061
Model:	OLS	**Adj. R-squared:**	0.038
Method:	Least Squares	**F-statistic:**	2.634
Date:	Sun, 25 Nov 2018	**Prob (F-statistic):**	0.00202
Time:	19:42:00	**Log-Likelihood:**	-2411.8
No. Observations:	497	**AIC:**	4850.
Df Residuals:	484	**BIC:**	4904.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.1067	14.204	0.148	0.882	-25.802	30.015
pass_cmp	-0.0289	0.058	-0.495	0.621	-0.144	0.086
pass_att	0.0068	0.036	0.190	0.849	-0.063	0.077
pass_cmp_pct	0.2316	0.332	0.697	0.486	-0.421	0.885
pass_yds	0.0038	0.004	1.054	0.292	-0.003	0.011
pass_yds_per_att	-0.6411	1.155	-0.555	0.579	-2.910	1.628
adj_pass_yds_per_att	-0.1364	0.746	-0.183	0.855	-1.602	1.329
pass_td	0.0352	0.195	0.180	0.857	-0.348	0.418
pass_int	-0.1041	0.206	-0.505	0.614	-0.509	0.301
rush_att	-0.0476	0.026	-1.828	0.068	-0.099	0.004
rush_yds	-0.0031	0.006	-0.531	0.596	-0.015	0.008
rush_yds_per_att	1.0493	0.958	1.095	0.274	-0.833	2.932
rush_td	0.7455	0.399	1.869	0.062	-0.038	1.529

The summary of the model told me that the R-squared value was 0.061, which right off the bat was not very optimistic. What this means is that all of these different college stats only explain around 6.1% of the variance in career AV, which is not exactly robust.

The p-values for the individual X values were mostly very high as well. What this tells us is that the higher the p-value, the lower the statistical significance to the model of said X value the specific p-value is associated with. The X values with the highest p-values were: passing attempts, adjusted passing yards/attempt and passing touchdowns. This is somewhat curious because one would think that there would be a clearly positive relationship between career AV and these stats (Which would lead to lower p-values), but that is not the case here. I decided to then rerun the model and take out some of the X values with higher p-values.

Second Attempt

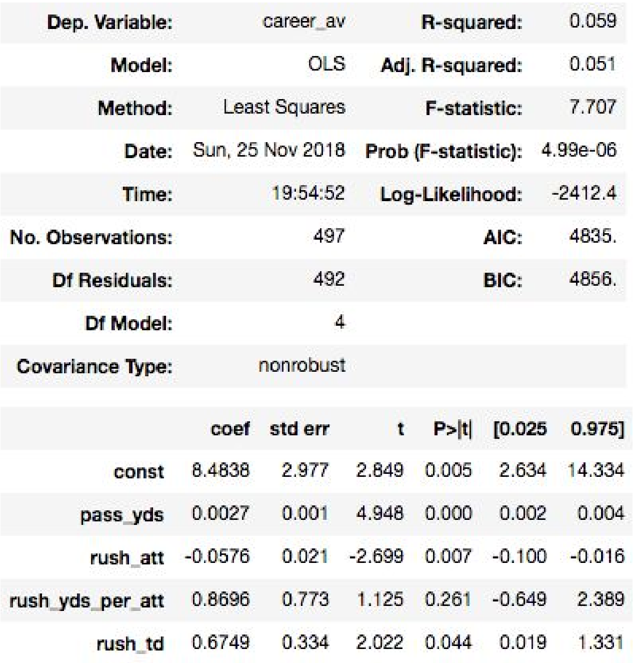


Dep. Variable:	career_av	**R-squared:**	0.060
Model:	OLS	**Adj. R-squared:**	0.044
Method:	Least Squares	**F-statistic:**	3.879
Date:	Sun, 25 Nov 2018	**Prob (F-statistic):**	0.000190
Time:	19:53:05	**Log-Likelihood:**	-2412.2
No. Observations:	497	**AIC:**	4842.
Df Residuals:	488	**BIC:**	4880.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	5.6768	10.726	0.529	0.597	-15.399	26.752
pass_cmp	-0.0142	0.052	-0.274	0.784	-0.116	0.088
pass_att	0.0024	0.033	0.074	0.941	-0.063	0.068
pass_cmp_pct	0.0628	0.197	0.319	0.750	-0.324	0.450
pass_yds	0.0036	0.003	1.374	0.170	-0.002	0.009
pass_int	-0.0676	0.194	-0.348	0.728	-0.449	0.314
rush_att	-0.0539	0.023	-2.370	0.018	-0.099	-0.009
rush_yds_per_att	0.7512	0.816	0.921	0.357	-0.851	2.354
rush_td	0.6414	0.340	1.886	0.060	-0.027	1.310

This time, I ran the model without including passing yards/attempt, adjusted passing yards/attempt, passing touchdowns and rushing yards. The R-squared for this model returned 0.060, which was slightly lower than our last model. I decided to leave the passing attempts variable in there to see what would happen. It turns out that the p-value jumped from .84 to .94! Some of the other notably high p-values for this iteration of the model were pass completions, passing completion percentage, passing yards and interceptions.

Third Attempt

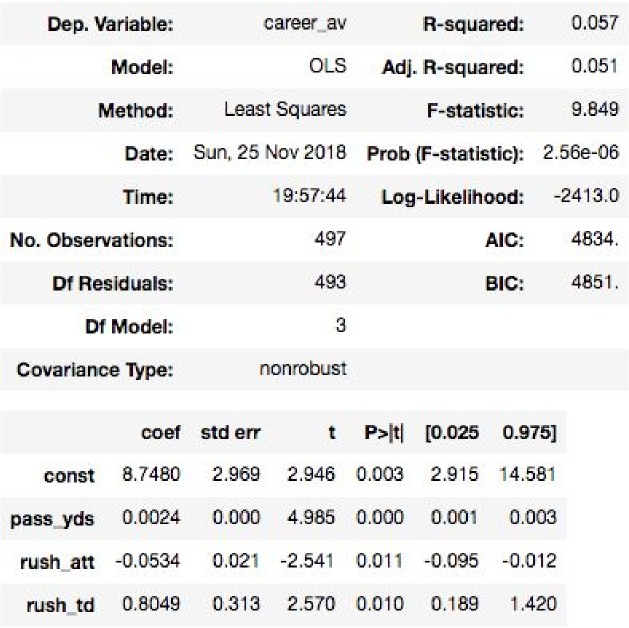


Dep. Variable:	career_av	**R-squared:**	0.059
Model:	OLS	**Adj. R-squared:**	0.051
Method:	Least Squares	**F-statistic:**	7.707
Date:	Sun, 25 Nov 2018	**Prob (F-statistic):**	4.99e-06
Time:	19:54:52	**Log-Likelihood:**	-2412.4
No. Observations:	497	**AIC:**	4835.
Df Residuals:	492	**BIC:**	4856.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	8.4838	2.977	2.849	0.005	2.634	14.334
pass_yds	0.0027	0.001	4.948	0.000	0.002	0.004
rush_att	-0.0576	0.021	-2.699	0.007	-0.100	-0.016
rush_yds_per_att	0.8696	0.773	1.125	0.261	-0.649	2.389
rush_td	0.6749	0.334	2.022	0.044	0.019	1.331

I then repeated the model once again, this time taking out all of the variables except pass yards, rushing attempts, rushing yards/attempt and rushing touchdowns. This time, the R-squared value was even lower (Around .059) and I also started seeing the p-values drop fairly significantly. The only one that was out of hand this time was the rushing yards/attempt p-value.

Final Attempt



Dep. Variable:	career_av	**R-squared:**	0.057
Model:	OLS	**Adj. R-squared:**	0.051
Method:	Least Squares	**F-statistic:**	9.849
Date:	Sun, 25 Nov 2018	**Prob (F-statistic):**	2.56e-06
Time:	19:57:44	**Log-Likelihood:**	-2413.0
No. Observations:	497	**AIC:**	4834.
Df Residuals:	493	**BIC:**	4851.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	8.7480	2.969	2.946	0.003	2.915	14.581
pass_yds	0.0024	0.000	4.985	0.000	0.001	0.003
rush_att	-0.0534	0.021	-2.541	0.011	-0.095	-0.012
rush_td	0.8049	0.313	2.570	0.010	0.189	1.420

I ran the model one last time and took out rushing yards/attempt. What I had left was passing yards, rushing attempts and rushing touchdowns. Passing yards returned a p-value of

0.000, rushing attempts returned 0.011 and rushing touchdowns returned 0.010 - All of which were within the threshold of statistical significance established at the beginning. The R-squared this time also dropped even more to 0.057.

Our model suggests that passing yards, rushing attempts and rushing touchdowns are the most statistically significant variables to include in the regression. However, the variables only explain around 5.7% of the total variance in career AV. What this means is that there is a lot more variance that can only be explained by other variables that we do not have access to here. I'm also fairly surprised at the variables determined to be statistically significant. I'm left wondering why are total college passing yards more significant to this model than interceptions or passing touchdowns for example? These results are a little confounding and I'm hoping to get to understand them a bit better moving forward.

Machine Learning

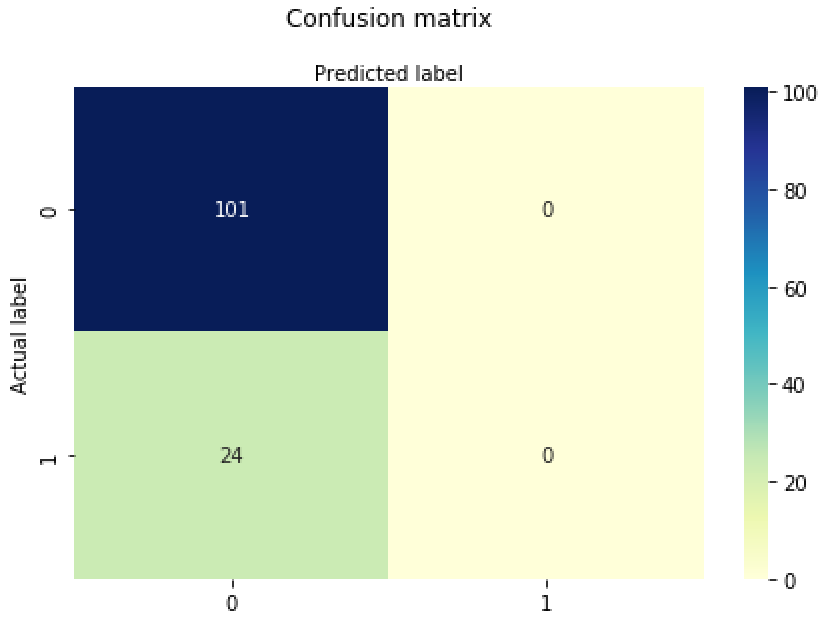
This section of my capstone project consisted of trying out a few different machine learning models in order to identify which one would be best suited for my problem - Whether or not we can identify great NFL quarterbacks based solely on their college statistics.

Scaled Multiple Linear Regression

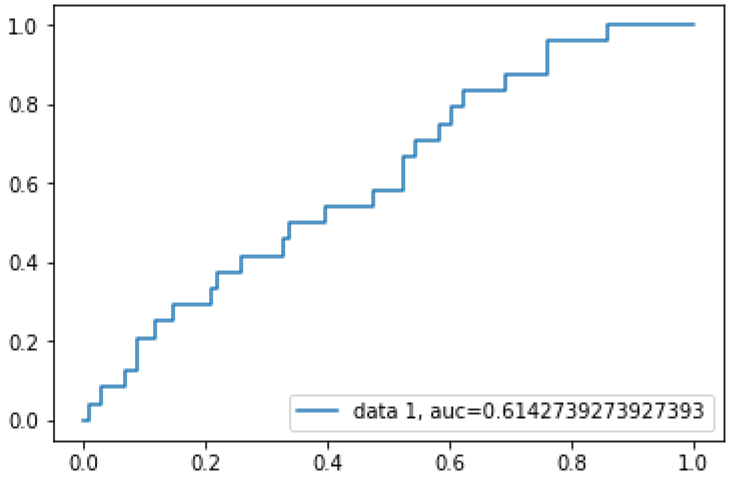
To begin, I scaled all of the numeric data from 0.0 to 1.0. By scaling the data, I hoped that it would make the multiple linear regression model that I used in the data storytelling/statistics notebook more robust. After the scaling process, I ran the model just as I had done last time, only to find that the results were nearly identical. The R-squared and adjusted R-squared values were the same. I tried to take away variables whose p-values were noticeably high, and continued to take away and rerun the model a few more times until I was left with a few variables that made the model as optimal as possible from an explained variance standpoint. However, the adjusted R-squared at the end of this process was still only 0.051, which points to a fairly bad model. Basically, what it is telling me is that the model explains only 5% of the variability of the response data around its mean. So 95% of that variance is explained by data that I have not use and/or do not have access to.

Logistic Regression

I then tried to perform a logistic regression model. Before I began, I established a response column to help the model divide the players into two categories: If a player had a career Approximate Value of greater than or equal to 0.15, he was considered a 'good' quarterback. If his career AV was less than 0.15, he was considered 'bad'. I ran the model and visualized the confusion matrix in the form of a heatmap.

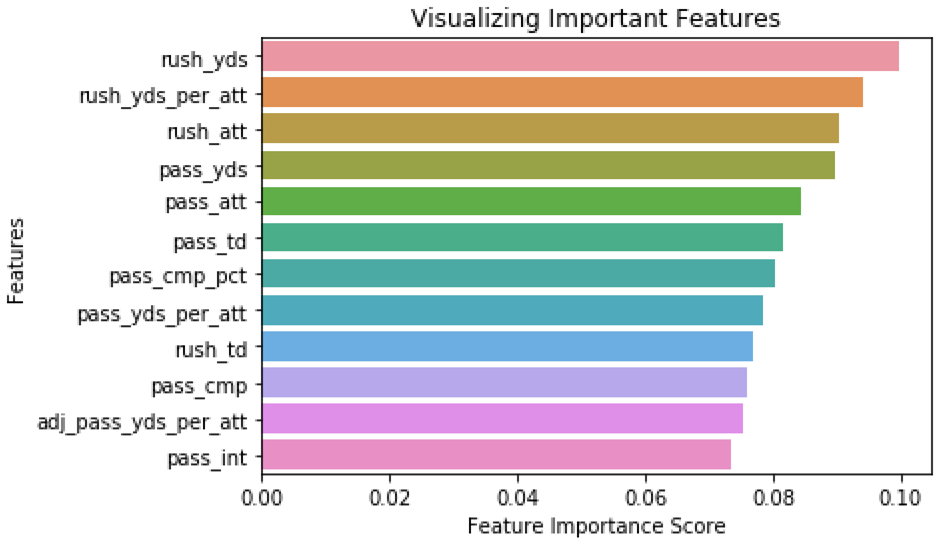


It told us that the diagonal values (101, 0) were actual predictions and the reverse-diagonal values (24, 0) were incorrect predictions. Our model's accuracy score was around 0.80, which is pretty good, but the precision and recall scores were at 0, which is not very good at all. What this says is that our model, while fairly accurate, is not precise at all. This equates to a very weak model, and AUC score (Which was at 0.6) helped us confirm that as seen in the plot below.



Random Forest

By far, the best model that I used was the random forest model. After running the model, I received an accuracy score of 0.744.



I was even able to visualize the feature importance score for all of the individual features and take out some of the less important ones. This helped me adjust my model and end up with an accuracy score at 0.776!

Conclusion

So why were the models overall fairly weak? The most likely explanation is that our data only explains a small percentage of the variance of the response data around its mean because there is more valuable data out there than can help explain more of that variance. The NFL combine, for example, is an annual scouting event where NFL scouts measure physical attributes of players (Speed, vertical, strength, height, weight, etc.). Perhaps this information is more important in helping to predict which quarterbacks will become NFL greats. One important thing about my capstone is the exploratory data analysis I was able to perform on the data set. I got to see some very interesting trends and feel that I understand a bit better why these statistics matter and how rare it actually is to see an ‘elite’ quarteback in college.

If I had more time with this project, I would explore combine data along with other data sets and use them in conjunction with college statistics to see if that could help improve the accuracy and precision of the models I used and to see if that could help to better predict which quarterbacks will become great ones in the NFL.

Unfortunately, I was not able to accurately predict which college quarterbacks would become great with this project, but what I did find is that college data alone is not enough to accurately predict something as complex as NFL greatness.