Slide 1

Theme Song

Goooood afternoon ladies and gentleman. Today we will be diving into the world of sports analytics, and exploring some potential predictive capabilities.

Specifically, we're going to be looking at touchdown predictions in the NFL.

Slide 2

So, the dataset. The data I used was pulled from Kaggle. It was compiled by researchers at Carnegie Mellon, and it breaks down all regular season NFL plays from the years 2009 - 2017. It contains over 400,000 plays with over 100 different labeled attributes.

Slide 3

I wanted to approach this dataset by thinking about how I could build something that might be useful, or at least fun, for me or my buddies. That lead me to my first question - thinking about how I could build a predictive framework for play by play gambling. This meant the model had to be light, computationally efficient, highly accurate, using only information available from before the play occurs.

I also wanted to make sure the models I was building had some explanatory power regarding the factors that influence a touchdown, so that data and human intuition could be used to make good judgment calls on bets, and run the Vegas over/unders.

Slide 4

The first thing I had to do was narrow down the feature set. A lot of the attributes provided were labeling the play results, or the play itself. What I wanted was to be able to predict *before* the play had begun. In other words, before the ball was snapped.

Aside from general data cleaning, such as removing duplicates and no plays, I also cleared out plays that had special team, in other words non-standard, formations such as FG, Kickoff and Punt formations. This was to help simplify the analysis, and make it more relevant to the question at hand.

Finally, I made some adjustments to how the game date was being coded. I transformed the provided datetime into 2 categorical variables, Weekday and Season Week, to help provide cleaner features for analysis.

The final feature set consisted of 6 continuous variables, and 8 categorical variables

Slide 5

Next, I wanted to explore the data to help determine which models to run and to make sure my features were all relevant in helping predict touchdowns

For the continuous variables, I plotted a PairGrid to help get a quick overall visual on distributions and collinearity.

The good news is, when looking at distributions there did not appear to be any severe outliers. I confirmed this by running a quantitative summary. From this we can also spot a general lack of parametric distribution in our feature set, effectively removing Naive Bayes from our arsenal.

Finally, there were a few instances of collinearity amongst the attributes.

Slide 6

I looked further into these, and determined they were related primarily to 2 different information types:

- The length of the game - with collinearity occurring between the features representing which Drive of the game was ongoing and the Time elapsed in seconds.
- The score - with collinearity occurring between features representing Team Scores and Score Differential

I removed the Drive feature, as I felt Time elapsed in seconds more effectively communicated that information.

I also removed the Defensive Team Score, as the combination of Offensive score and Score Differential sufficiently captured that information

Slide 7

Next, I explore the categorical variables to determine whether they were in fact useful in predicting Touchdowns

I first to a look at the different temporal categories, all of which have some sort of impact.

For down

Slide 8

I then took a look at team performances by game, plotting TDs per play for all teams, for both home and away, and on the Offensive…

Slide 9

… and Defensive ends as well. The high number of outliers indicates a high variance in team performance on a game by game basis, making it an especially juicy feature for us, as we will see later on.

Slide 10

Now that I knew my features were cleaned up and good to go, I picked out and ran a few models with the results as you can see here.

I measured purely accuracy. Runtime was the length of time the prediction of the entire set took, not inclusive of fitting the model.

What was particularly interesting was how similar all the models' accuracies were. Even when toying with some parameters, I quickly realized that the impacts to performance were trivial, and that any improvement in performance could only be drawn out of additional data that could fill that 4% knowledge gap.

Ultimately, I opted for the Logistic Regression given its speed, predictive power and low computational demand.

Slide 11

With these models, I was also able to pull out coefficients and feature importance attributes to try and hone in specifically on which of the features were most useful in our predictions.

Yards from the 1st down, and total yards away from touchdown appeared to be the most influential features, with no surprise. The down and the quarter also had large impacts, likely due to the strategic differences that occur in different owns and quarters.

Surprisingly, Offensive Timeouts left also popped up fairly high in importance. Perhaps something for teams to consider before they recklessly call challenges.

And finally, we see a proliferation of different team names. The New Orleans Saints appear in all 3 models, which is a testament to their historically high scoring offense ran by Drew Brees over the last decade. More importantly, this also tells us something we don't know - what variables predict a team's success on the field.

It is likely that the the data we seek hides somewhere in that

This also opens the door up to further research questions, and gets us thinking about how to break down and analyze the black box of an NFL team. Perhaps there is were we would find our last 4% of accuracy.