DS502 Project Proposal

Mia Barger, Quincy Hershey, Alex Moore, Ethan Prihar

Problem Description

https://www.kaggle.com/c/nfl-big-data-bowl-2020

The national football league (NFL) has compiled data on plays where an individual runs the ball. During a play like this, an individual carrying the football must run as far as they can down the field with their team's support, while the other team attempts to stop them. The NFL wishes to predict, in real time, how many yards the player with the ball will run. This prediction will be used by the NFL's analysts during and after plays throughout the football season to better understand what contributes to more successful runs.

Data Description

The following excerpt from the NFL Big Data Bowl Kaggle page describes the data available:

Each row in the file corresponds to a single player's involvement in a single play. The dataset was intentionally joined (i.e. denormalized) to make the API simple. All the columns are contained in one large dataframe which is grouped and provided by PlayId.

GameId - a unique game identifier
PlayId - a unique play identifier
Team - home or away
X - player position along the long axis of the field. See figure below.
Y - player position along the short axis of the field. See figure below.
S - speed in yards/second
A - acceleration in yards/second^2
Dis - distance traveled from prior time point, in yards
Orientation - orientation of player (deg)
Dir - angle of player motion (deg)
NflId - a unique identifier of the player
DisplayName - player's name
JerseyNumber - jersey number
Season - year of the season
YardLine - the yard line of the line of scrimmage
Quarter - game quarter (1-5, 5 == overtime)
GameClock - time on the game clock

PossessionTeam - team with possession
Down - the down (1-4)
Distance - yards needed for a first down
FieldPosition - which side of the field the play is happening on
HomeScoreBeforePlay - home team score before play started
VisitorScoreBeforePlay - visitor team score before play started
NflIdRusher - the NflId of the rushing player
OffenseFormation - offense formation
OffensePersonnel - offensive team positional grouping
DefendersInTheBox - number of defenders lined up near the line of scrimmage, spanning the width of the offensive line
DefensePersonnel - defensive team positional grouping
PlayDirection - direction the play is headed
TimeHandoff - UTC time of the handoff
TimeSnap - UTC time of the snap
Yards - the yardage gained on the play (we are predicting this)
PlayerHeight - player height (ft-in)
PlayerWeight - player weight (lbs)
PlayerBirthDate - birth date (mm/dd/yyyy)
PlayerCollegeName - where the player attended college
Position - the player's position (the specific role on the field that they typically play)
HomeTeamAbbr - home team abbreviation
VisitorTeamAbbr - visitor team abbreviation
Week - week into the season
Stadium - stadium where the game is being played
Location - city where the game is being player
StadiumType - description of the stadium environment
Turf - description of the field surface
GameWeather - description of the game weather
Temperature - temperature (deg F)
Humidity - humidity
WindSpeed - wind speed in miles/hour
WindDirection - wind direction

Prediction type

The problem could be approached as either regression or classification. We will approach the problem as a regression problem, for a couple of reasons. First, there would be a large number of classes with little data in each class if we approached it as a classification problem. Second, a classification problem would predict only integer values for a single play's yardage, while a regression can incorporate continuous values. While we will need to submit a file which predicts yards as discrete values, using a regression model will still allow for

more precise measurements for our predictions because it allows for the fact that yardage is not discrete in real life.

Methods

Before we begin constructing the model we'll need to attempt to convert the categorical attributes to ordinal attributes. For example, **HomeTeamAbbr** could be converted to information on the offensive and defensive power of the team. Other categorical attributes can be one hot encoded. There is a possibility of developing a simulation of the play which doesn't use machine learning. The advantage of this is that it's easier for a machine learning model to learn the difference between what a theory-based model predicted and the truth than it is for a neural network to learn the theory itself. The disadvantage of this theory-based model creation is that we might not have enough time to work on the machine learning if we use it all making a theory-based model. After we have finished processing the data we plan on using a neural network. The first neural network we'll attempt to use is a regular feed forward network. Although unlikely, it is possible that at any given point in a play, we have enough information to know how many yards the ball carrier will travel. It is likely though that the previous states of the play are relevant when predicting the total yards run, therefore the second model we will construct is an RNN with LSTM nodes. LSTM nodes are better than normal perceptrons because they can learn for themselves how long past information stays relevant. An alternative to the RNN is a 1-D CNN. The 1-D CNN will look for trends in a stream of play data, which may lead to other insight useful to predicting the total yards run that an RNN might not have been able to see. It is likely that multiple models will perform well, and if so, an ensembling process will be used to make the final prediction.

Error Metrics and Algorithms for assessing them

The Kaggle competition stipulates use of the Continuous Ranked Probability Score (CRPS) based on the cumulative distribution of yards gained. The formula is as follows:

$$C = \frac{1}{199N} \sum_{m=1}^{N} \sum_{n=-99}^{99} (P(Y \leq n) - H(n - Y_m))^2$$

where P is the predicted distribution, N is the number of plays in the test set, Y is the actual yardage and $H(x)$ is the Heaviside step function. The Kaggle competition will not score the submission if the CDF is a decreasing function.

Comments and Concerns

The NFL data has an interesting structure which gives us unique insights and challenges. Rather than each row being an observation, collections of rows unified by a 'Play I.D.' represent one play during a football game, with a response yards gained. There are many interesting preprocessing techniques to transform this data into something we can model. There are countless exploration topics in the data on topics such as matchups, player performance, and team organization that can motivate our models and insights about this unique data.

The high variability in the data is one concern for our regression. An identical play (players, stadium, latent factors) could lead to a 0 yard gain or a 100-yard touchdown. For this reason we will have to negotiate with extreme noise in the development of our model, and in some cases a lack of predictor correlation to the response.