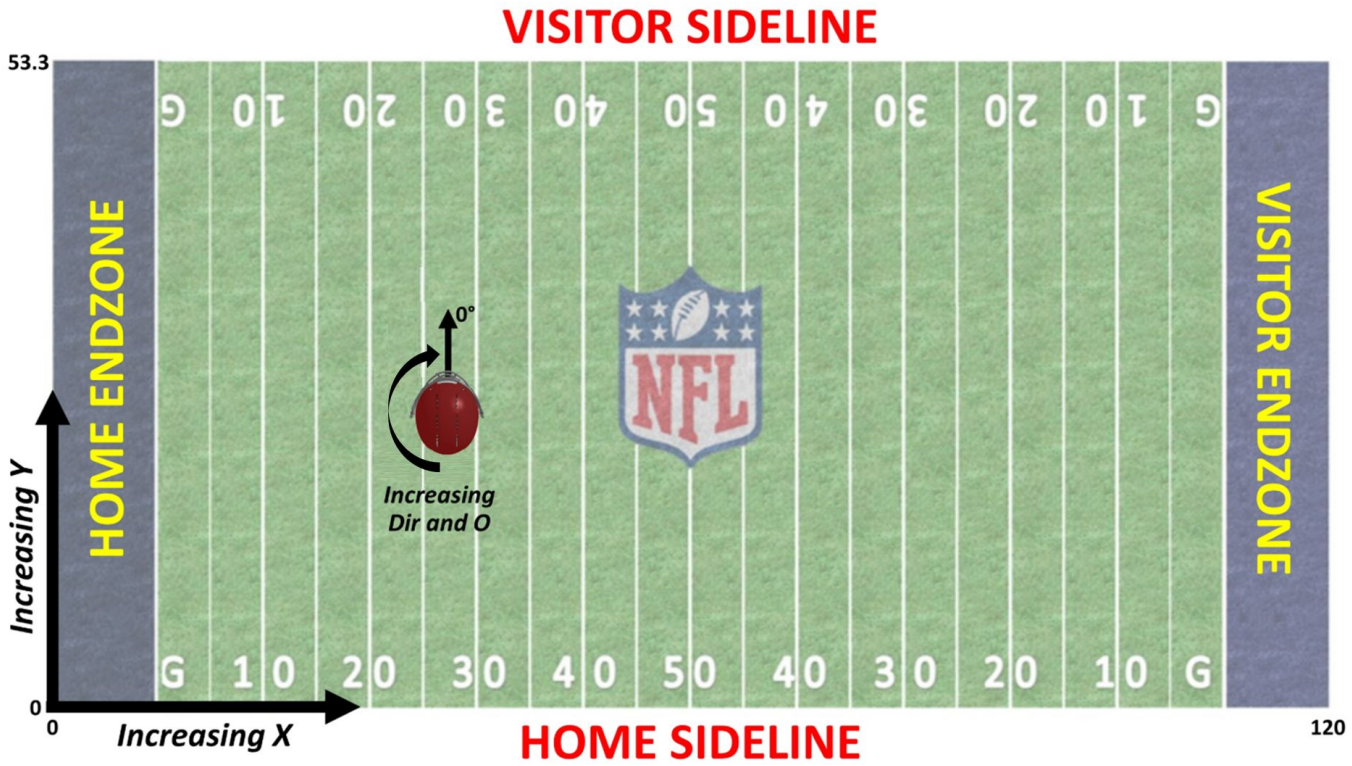
Predicting How Many Rushing Yards in a Given Run Play

Grant Nolasco

Background


VISITOR SIDELINE
53.3
HOME ENDZONE
VISITOR ENDZONE
G 10 20 30 40 50 40 30 20 10 G
G 10 20 30 40 50 40 30 20 10 G
0°
Increasing
Dir and O
Increasing Y
0
0
Increasing X
HOME SIDELINE
120

Background

Data

- The dataset comes from NFL Next Gen Stats
  - Has player tracking data/information about the offense and defense/other data about the game like weather and stadium
  - 31,007 running plays from 2017-2019 season and total of around 682,000 rows
  - 49 columns
  - Each play has 22 rows and each row has information about each player at the time of handoff

GameId	PlayId	Team	X	Y	S	A	Dis	Orientation	Dir	...	Week	Stadium	Location	StadiumType	Turf	GameWeather
2017090700	20170907000118	away	46.09	34.84	1.69	1.13	0.40	278.01	182.82	...	1	Gillette Stadium	Foxborough, MA	Outdoor	Field Turf	Clear and warm
2017090700	20170907000118	away	45.33	32.64	0.42	1.35	0.01	332.39	161.30	...	1	Gillette Stadium	Foxborough, MA	Outdoor	Field Turf	Clear and warm
2017090700	20170907000118	away	46.00	33.20	1.22	0.59	0.31	356.99	157.27	...	1	Gillette Stadium	Foxborough, MA	Outdoor	Field Turf	Clear and warm
2017090700	20170907000118	away	48.54	27.70	0.42	0.54	0.02	0.23	254.36	...	1	Gillette Stadium	Foxborough, MA	Outdoor	Field Turf	Clear and warm
2017090700	20170907000118	away	50.68	35.42	1.82	2.43	0.16	347.37	195.69	...	1	Gillette Stadium	Foxborough, MA	Outdoor	Field Turf	Clear and warm

Rules

- No external dataset
- Runtime of code cannot exceed four hours
- The goal is to minimize this function

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

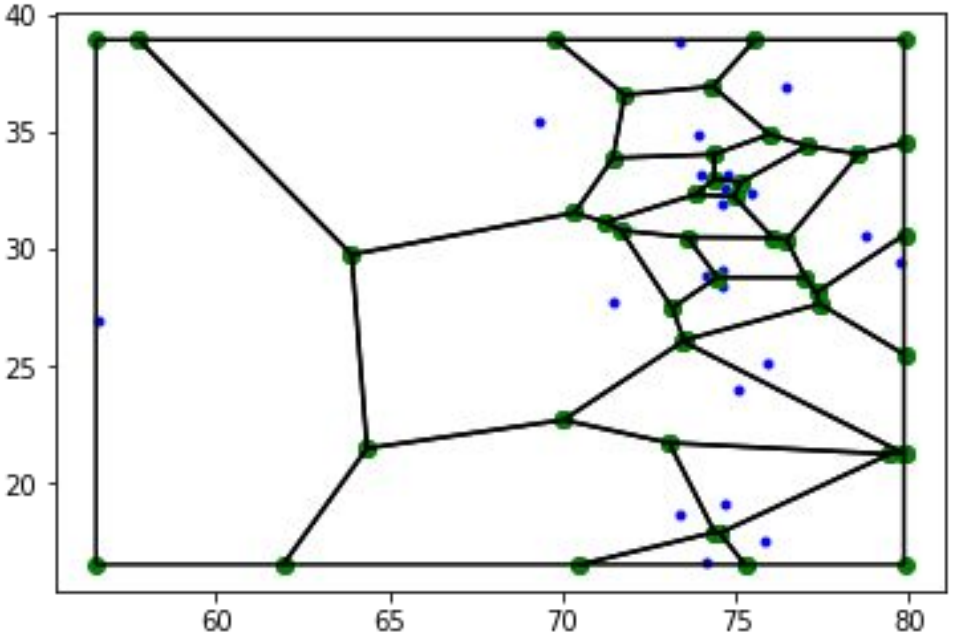
where N is the number of plays in the dataset, P is the predicted probability that the number of yards gained is less than n, Y is the actual yardage gained, and H is a heaviside step function

Challenges

- Categorical columns have multiple misspellings (e.g Sunny was written as sUNNY) and/or different interpretations of the same thing
  - Bucketed every possible string into general categories (e.g Weather column has Clear, Rainy, Overcast, Snowy, NA)
- Few of the columns had almost 10% missing values
  - Imputed the mean for numerical columns and the mode for categorical columns
- 2017 has different acceleration/directional values compared to 2018/2019
  - Had to standardize 2017 values to closely match the distribution of 2018/2019 numbers

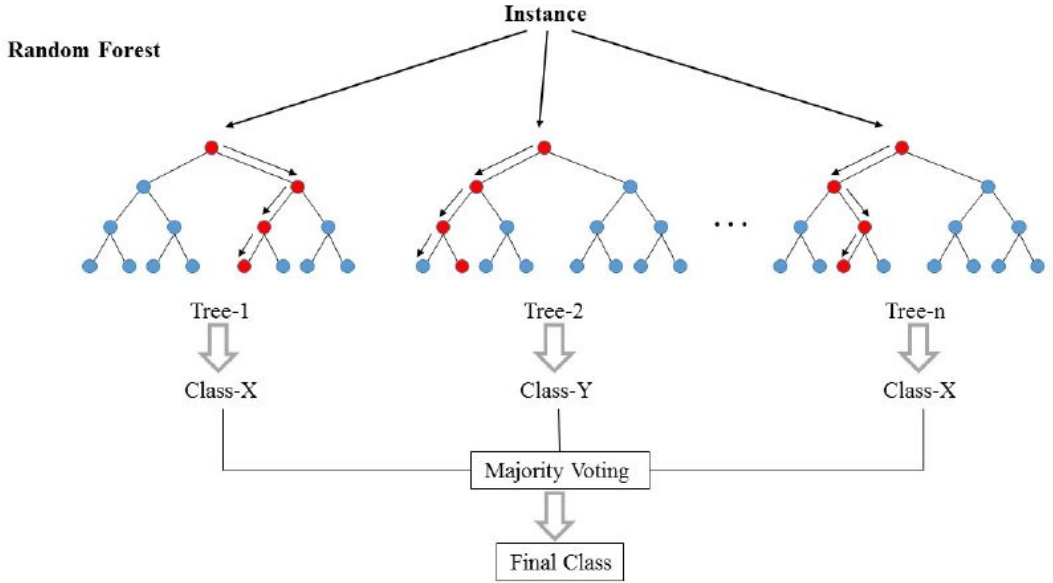
EDA

- Voronoi diagram partitions a plane into a set of regions in which each region is associated with a single point
- The region of a particular point means that that area is the closest to that particular point compared to the rest



Random Forest

- A classification/regression tool that consists of a lot of decision trees



Results

- The final dataset includes 43 columns such as the quantiles of defense's space based on the voronoi diagram and quantiles of distance from runner

Parameters	**Values Tested**
num_trees	10, 32, 55, **77**, 100
min_samples_split	**2**, 5, 10
min_samples_leaf	**1**, 2, 4
max_features	auto, **sqrt**
max_depth	**5**, 11, 17, 23, 30, None
bootstrap	True, **False**

CRPS Score of 0.013311

Additional Steps (if given time)

- Look into other algorithms to test out (e.g Neural Network)
- Potentially add in other columns that could be of importance (e.g space that the OL has)
- Do feature importance to do dimensionality reduction and potentially remove noisy features
- Once the best model has been determined, use that model and submit it to the competition!