Cloudy With a Chance of Football (Cohort 23): Capstone Project Data Architecture (Expanded)

Data Ingestion

Data Limits
Years: 2019 and 2020
Positions: offense and whole-team defense
Platform: ESPN

NFL Savant
Individual player data
nflsavant_ingestor.py

Fantasy Football Data Pros
Weekly fantasy stats
2019 ESPN projection data
datapros_ingestor.py

Kaggle Data
2019 stats and 2020 projections
kaggle_ingestor.py

Carnegie Mellon
Temperature and precipitation data
cmu_ingestor.py

Munging and Wrangling

Fuzzy Wuzzy
seaborn
NumPy
matplotlib
Pandas

nflsavant_cleaner_eda.ipynb
datapros_cleaner_eda.ipynb
kaggle_cleaner_eda.ipynb
cmu_cleaner_eda.ipynb

nflsavant.csv
datapros.csv
kaggle.csv
cmu.csv

SQLite

players

player_id	unique_id	player_name
0	tombrady	Tom Brady
1	joshallen	Josh Allen

week_table

year	week	week_start	week_id
2019	1	20190907	2019_1
2019	2	20190914	2019_2

positions

player_id	position
0	qb
1	qb

player_age

player_id	week_id	age
0	2019_1	42.1
1	2019_1	25.52

player_dob

player_id	dob
0	19770803
1	19960521

team_home_time_zone

team_name	week_id	time_zone
NE	2019_1	-5
BUF	2019_1	-5

player_team

player_id	week_id	team_name
0	2019_1	NE
1	2019_1	BUF

game_location

week_id	team_name	location
2019_1	NE	1
2019_1	BUF	0

injury_status

player_id	week_id	injury_status
0	2019_1	0
1	2019_1	0

days_since_last_game

team_name	week_id	days_since_last_game
NE	2019_2	6
BUF	2019_2	5

time_displacement_table

week_id	team_name	hours_displaced
2019_1	NE	0
2019_2	BUF	0

performance

player_id	week	predicted	actual	performance
0	2019_1	18.20	21.90	1
1	2019_1	29.35	27.10	0
0	2019_2	25.87	18.50	1
1	2019_2	19.93	28.98	1

weather

week_id	team_name	temperature_min	temperature_max	precipitation
2019_1	NE	58	75	.1
2019_1	BUF	35	48	3

opponents table

week_id	team_name	game_date	game_time	week_day	ARI	ATL	BAL	BUF	CAR	...	NO	NYG	NYJ	PHI	PIT	SF	SEA	TB	TEN	WAS
2019_1	NE	20190907	19	6	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
2019_2	BUF	20190906	12	6	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0

stats_table

player_id	week_id	PassingYds	PassingTD	Int	PassingAtt	Cmp	RushingAtt	RushingYds	RushingTD	Rec	Tgt	ReceivingYds	ReceivingTD	FL
0	2019_1	324	5	0	20	17	3	6	0	0	0	0	0	0
1	2019_1	405	4	0	32	25	4	12	0	0	0	0	0	0

Columns

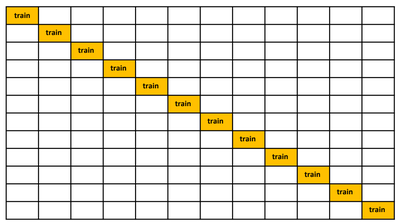
Column Name	Data Type	Unique Variables or Description
player_id	integer	
unique_id	text	firstnamelastname
player_name	text	Firstname Lastname
year	integer	2019 2020
week	integer	
week_start	datetime	YYYY-MM-DD
week_id	text	YYYY_WW
position	text	qb, rb, te, wr, k
age	real	
dob	datetime	YYYY-MM-DD
time_zone	integer	Hours away from Greenwich mean time of home time zone
team_name	text	Official team abbreviation codes
injury_status	integer	0 : no injury 1 : questionable 2 : out
days_since_last_game	integer	
hours_displaced	integer	Number of hours away from home time zone
temperature_min	integer	
temperature_max	integer	
precipitation	real	
PassingYds	real	
PassingTD	integer	
Int	integer	
PassingAtt	integer	
Cmp	integer	
RushingAtt	integer	
RushingYds	real	
RushingTD	integer	
Rec	integer	
Tgt	integer	
ReceivingYds	real	
ReceivingTD	integer	
FL	integer	
location	integer	0 : away 1 : home
predicted	real	ESPN prediction for player fantasy points
actual	real	Actual number of ESPN fantasy points achieved
performance	integer	0 : predicted < actual 1 : actual < predicted
game_time	integer	hours of day from 1 – 24 in home team time zone
week_day	integer	day of week during which game takes place, Monday = 0, Sunday = 6

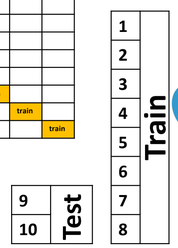
Machine Learning

Model 3: Running Backs
Model 2: Quarterbacks
Model 1: All players
Model 7: Defense
Model 6: Kickers
Model 5: Wide Receivers
Model 4: Tight Ends

model3.ipynb
model2.ipynb
model1.ipynb
model7.ipynb
model6.ipynb
model5.ipynb
model4.ipynb

12-fold time aware cross validation



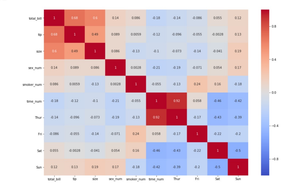


80/20 train/test split

scikit-learn

Computation and Analysis

- Using a correlation matrix, determine which features are important and would be useful in classifying over or under-performers
- Use Pandas Shift to make previous week performance features useful in predicting time aware target variable or over or under-performing
- Normalize all data from 0 to 1
- Convert text data to integer data
- Check on class imbalance and resolve class imbalance if detected
- Conduct automated testing of scripts



Reporting and Visualization

Operationalize machine learning models by exporting the models to python scripts, building an app with D3.js and Flask to accept basic user input and provide predictions on whether a player will over or under-perform in the 2021 NFL season