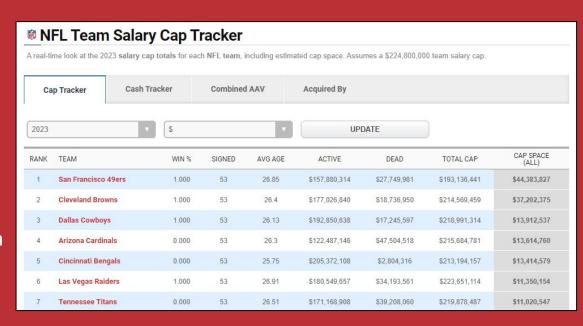


PREDICTING NFL VESTED VETERANS

Billie Kim, Casey Ng, Ethan Panal Group 28

Problem

- NFL Team General Managers (GMs) manage multi-million dollar budgets
- NFL Players become Vested Veterans after 3 years in the league
- Solution: Aid GMs with budget strategy with predictions if a rookie will make it long enough to be a vested veteran
- **Dataset:** Offensive rookie data from 1918-2019

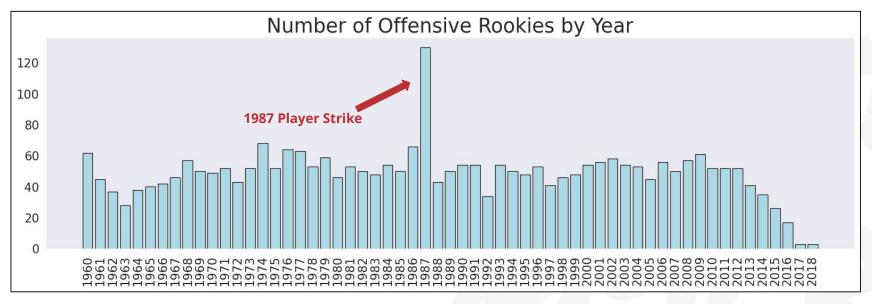


Data Cleaning/Wrangling

- Merged datasets: receiving + rushing + basic info
- Convert data types
- Standardize values
- Impute missing data
- Transformed # of years played to boolean if > 3 years

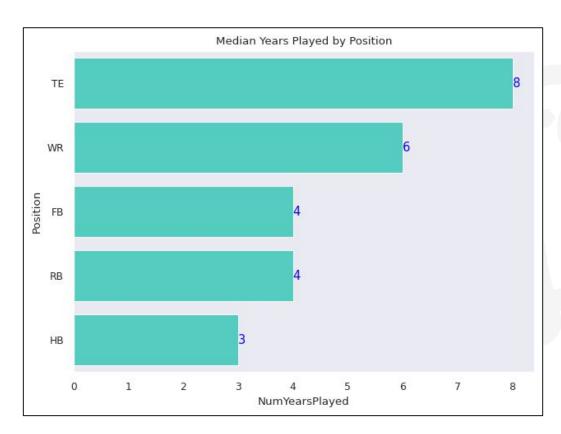
	Player_Id	Year	Team	Games_Played	Attempts	Yards_Rush	Average_Rush	Long_Rush	TDs_Rush	First_Downs_Rush		Yards_Receptions	Average_Receptions	L
0	a-b-brown	1989	NewYorkJets	16	12	63	5.3	17	0	0		10	2.5	
1	a-d- whitfield	1965	DallasCowboys	2	1	0	0	0	0	0		0	0	
2	a-j-jenkins	2012	SanFrancisco49ers	3	0	0	0	0	0	0	***	0	0	
3	aaron- bailey-2	1994	IndianapolisColts	13	0	0	0	0	0	0		30	15	
4	aaron- brooks	1999	GreenBayPackers	0	0	0	0	0	0	0		0	0	

EDA



3000+ retired offensive player stats from their rookie seasons

EDA



- Offensive players tend to last 3-8 years in league
- Running backs have the shortest careers

Correlation Heatmap

Correlation Patterns:

Rushing Stats Vs. Receiving Stats

										Fe	ature (Correl	ation H	leatma	ар												1.0
Games_Played	1.00	0.29	0.29	0.15	0.25	0.24	0.18	0.11	0.16	0.12	0.14	0.47	0.41	0.27	0.35	0.30	0.26	0.14	0.22	0.17	0.25	0.22	0.09	0.04			1.0
Attempts	0.29	1.00	0.98	0.17	0.55	0.80	0.61	0.06		0.42	0.49	0.32	0.11	-0.02	0.07	-0.02	0.02	-0.05	-0.03	-0.03	0.07	0.09	-0.07	0.12			
Yards_Rush	0.29	0.98	1.00	0.21	0.58	0.82	0.61	0.07		0.46	0.47	0.32	0.12	-0.01	0.08	-0.00	0.03	-0.04	-0.02	-0.02	0.10	0.11	-0.07	0.11			
Average_Rush	0.15	0.17	0.21	1.00	0.36	0.15	0.10	0.37	0.13	0.10	0.06	0.12	0.10	0.13	0.09	0.06	0.06	0.01	0.05	0.06	-0.02	-0.04	-0.11	-0.06			0.8
Long_Rush	0.25	0.55		0.36	1.00	0.45		0.33			0.44	0.26	0.12	-0.06	0.32	-0.00	0.16	0.13	0.10	0.07	0.07	0.07	-0.17	0.13			0.0
TDs_Rush	0.24	0.80	0.82	0.15	0.45	1.00	0.51	0.06	0.45	0.39	0.39	0.27	0.11	-0.02	0.06	0.01	0.02	-0.05	-0.02	-0.02	0.10	0.10	-0.04	0.11			
First_Downs_Rush	0.18	0.61	0.61	0.10	0.55	0.51	1.00	0.32	0.87	0.68	0.77	0.21	0.07	-0.05	0.17	-0.03	0.21	0.20	0.12	0.07	0.06	0.07	-0.11	0.16			
First_Down_Percentage_Rush	0.11	0.06	0.07	0.37	0.33	0.06	0.32	1.00	0.29	0.20	0.23	0.14	0.11	-0.01	0.23	0.03	0.34	0.39	0.28	0.22	0.01	0.02	-0.09	0.12			0.6
Rushes_Over_Twenty_Yards	0.16	0.53		0.13	0.58	0.45	0.87	0.29	1.00	0.80	0.67	0.20	0.08	-0.04	0.17	-0.01	0.22	0.18	0.14	0.10	0.07	0.08	-0.10	0.11			
Rushes_Over_Forty_Yards	0.12	0.42	0.46	0.10	0.55	0.39	0.68	0.20	0.80	1.00	0.54	0.18	0.08	-0.03	0.14	0.01	0.18	0.14	0.12	0.09	0.07	0.07	-0.08	0.08			
Fumbles	0.14	0.49	0.47	0.06	0.44	0.39	0.77	0.23	0.67	0.54	1.00	0.15	0.04	-0.05	0.13	-0.04	0.15	0.16	0.08	0.04	0.04	0.05	-0.10	0.13			
Receptions	0.47	0.32	0.32	0.12	0.26	0.27	0.21	0.14	0.20	0.18	0.15	1.00	0.93	0.30	0.58	0.70	0.62	0.25		0.40	0.31	0.23	0.13	-0.03		_	0.4
Yards_Receptions	0.41	0.11	0.12	0.10	0.12	0.11	0.07	0.11	0.08	0.08	0.04	0.93	1.00	0.43		0.82		0.24		0.48	0.33	0.24	0.16	-0.10			
Average_Receptions	0.27	-0.02	-0.01	0.13	-0.06	-0.02	-0.05	-0.01	-0.04	-0.03	-0.05	0.30	0.43	1.00	0.40	0.37	0.16	0.24	0.20	0.19	0.10	0.06	0.09	-0.23			
Long_Receptions	0.35	0.07	0.08	0.09	0.32	0.06	0.17	0.23	0.17	0.14	0.13	0.58	0.60	0.40	1.00	0.44	0.55	0.48	0.54		0.21	0.17	0.03	-0.05			
TDs_Receptions	0.30	-0.02	-0.00	0.06	-0.00	0.01	-0.03	0.03	-0.01	0.01	-0.04		0.82	0.37	0.44	1.00	0.45	0.17	0.47	0.43	0.26	0.18	0.17	-0.08		_	0.2
First_Downs_Receptions	0.26	0.02	0.03	0.06	0.16	0.02	0.21	0.34	0.22	0.18	0.15		0.59	0.16		0.45	1.00	0.62	0.92	0.69	0.19	0.15	0.08	0.05			
First_Down_Percentage_Receptions	0.14	-0.05	-0.04	0.01	0.13	-0.05	0.20	0.39	0.18	0.14	0.16	0.25	0.24	0.24	0.48	0.17	0.62	1.00	0.55	0.40	0.11	0.11	0.03	0.12			
Receptions_Over_Twenty_Yards	0.22	-0.03	-0.02	0.05	0.10	-0.02	0.12	0.28	0.14	0.12	0.08	0.55	0.58	0.20		0.47	0.92	0.55	1.00	0.78	0.18	0.13	0.09	0.03			
Receptions_Over_Forty_Yards	0.17	-0.03	-0.02	0.06	0.07	-0.02	0.07	0.22	0.10	0.09	0.04	0.40	0.48	0.19		0.43	0.69	0.40	0.78	1.00	0.13	0.09	0.07	-0.01		-	0.0
NumYearsPlayed	0.25	0.07	0.10	-0.02	0.07	0.10	0.06	0.01	0.07	0.07	0.04	0.31	0.33	0.10	0.21	0.26	0.19	0.11	0.18	0.13	1.00	0.74	0.18	0.11			
3YearsThreshold	0.22	0.09	0.11	-0.04	0.07	0.10	0.07	0.02	0.08	0.07	0.05	0.23	0.24	0.06	0.17	0.18	0.15	0.11	0.13	0.09	0.74	1.00	0.14	0.11			
Height	0.09	-0.07	-0.07	-0.11	-0.17	-0.04	-0.11	-0.09	-0.10	-0.08	-0.10	0.13	0.16	0.09	0.03	0.17	0.08	0.03	0.09	0.07	0.18	0.14	1.00	0.49			
Weight	0.04	0.12	0.11	-0.06	0.13	0.11	0.16	0.12	0.11	0.08	0.13	-0.03	-0.10	-0.23	-0.05	-0.08	0.05	0.12	0.03	-0.01	0.11	0.11	0.49	1.00		-	-0.2
	Games_Played	Attempts	'Yards_Rush	Average_Rush	Long_Rush	TDs_Rush	First_Downs_Rush	First_Down_Percentage_Rush	Rushes_Over_Twenty_Yards	Rushes_Over_Forty_Yards	Fumbles	Receptions	Yards_Receptions	Average_Receptions	Long_Receptions	TDs_Receptions	Hrst_Downs_Receptions	frst_Down_Percentage_Receptions	Receptions_Over_Twenty_Yards	Receptions_Over_Forty_Yards	NumYearsPlayed	3YearsThreshold	Height	Weight	11 T		

Feature Selection

- Domain Knowledge
- ANOVA F-Tests
- Random Forest Feature Importance

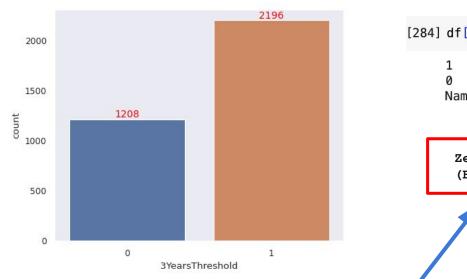
Final Features:
Weight, Games_Played,
Yards_Receptions, Yards_Rush,
Average_Rush, Attempts,
Receptions, Height,
Long_Receptions,
Average_Receptions, Long_Rush

	ANOVA	Random Forest
Weight	25.765948	0.126059
Games_Played	122.562439	0.108469
Yards_Receptions	164.037910	0.103428
Yards_Rush	29.793872	0.084811
Attempts	19.116489	0.080928
Average_Rush	1.417779	0.079855
Receptions	159.192758	0.077155
Average_Receptions	9.301793	0.076650
Height	44.161096	0.069651
Long_Receptions	79.716845	0.049955
Long_Rush	13.974476	0.045311
TDs_Rush	23.883879	0.026140
TDs_Receptions	89.453975	0.019827
First_Downs_Receptions	54.415906	0.017637
First_Downs_Rush	12.129420	0.015883
Receptions_Over_Twenty_Yards	41.218663	0.007487
Rushes_Over_Twenty_Yards	17.685941	0.006805
Fumbles	7.628763	0.003949

Baseline Accuracy







```
[284] df['3YearsThreshold'].value_counts(normalize=True)

1    0.645123
0    0.354877
Name: 3YearsThreshold, dtype: float64

Zero-Rate Classifier Accuracy
(Baseline) = 64.51 %
```

* Theoretical Baseline for Classification *

ML Algorithm Comparison

LogisticRegression()

Accuracy Score: 0.678015608759227

DecisionTreeClassifier()

Accuracy Score: 0.6372769233479556

RandomForestClassifier()

Accuracy Score: 0.7132718496203513

KNeighborsClassifier()

Accuracy Score: 0.68262424466642

GradientBoostingClassifier()

Accuracy Score: 0.7191418705847119

GaussianNB()

Accuracy Score: 0.5789167239222733

SVC()

Accuracy Score: 0.7073921392456354

Method: Looping thru Mean Cross-Validation Accuracy Scores







Base Models

- 1. Random Forest
- 2. Gradient Boosting
- 3. Support Vector Machine

TRAINING ACCURACY, TESTING ACCURACY:

(0.9987405541561712, 0.7064579256360078)

(0.781696053736356, 0.7211350293542075)

(0.7246011754827876, 0.7045009784735812)

Hyperparameter Tuning

SVC Best Hyperparameters:

- 'C': 10
- 'gamma': 'scale'
- 'kernel': 'rbf'

Best CV Score: **0.7150**

Accuracy on Test Set: **0.6967**

Baseline Accuracy: 0.7074

```
svc_param_grid = {
    'C': [0.1, 1, 10, 100, 1000],
    'gamma': ['scale', 'auto'],
    'kernel': ['rbf', 'poly', 'sigmoid']
}
```

Hyperparameter Tuning

Gradient Boost Best Hyperparameters:

- subsample: 0.5
- n_estimators: 2000
- min_samples_split: 5
- min_samples_leaf: 8
- max_features: 'log2'
- max_depth: None
- loss: 'log_loss'
- learning_rate: 0.001

Best CV Score: 0.7254

Accuracy on Test Set: 0.7221

Baseline Accuracy: 0.7191

```
gb param grid = {
    'loss':['log loss', 'exponential'],
    'learning rate':[0.001, 0.01, 0.1, 0.2],
    'subsample':[0.5, 0.8, 1.0],
    'max depth': [20, 40, 60, 80, 100, None],
    'max features': ['sqrt', 'log2', None],
    'min samples leaf': [1, 2, 4, 8],
    'min samples split': [2, 5, 10],
    'n estimators': [100, 400, 1000, 2000]
```

Hyperparameter Tuning

Random Forest Best Hyperparameters:

- n_estimators: 400
- min_samples_split: 10
- min_samples_leaf: 8
- max_features: 'sqrt'
- max_depth: 60
- criterion: 'log_loss'
- bootstrap: True

Best CV Score: 0.7313

Accuracy on Test Set: **0.7290**

Baseline Accuracy: 0.7132

```
rf_param_grid = {
    'criterion': ['gini', 'entropy', 'log_loss'],
    'bootstrap': [True, False],
    'max_depth': [20, 40, 60, 80, 100, None],
    'max_features': ['auto', 'sqrt', 'log2', None],
    'min_samples_leaf': [1, 2, 4, 8],
    'min_samples_split': [2, 5, 10],
    'n_estimators': [100, 200, 400, 800, 1200, 2000]
}
```

Performance Metrics Evaluations

	Accuracy	Recall	Precision	F1 Score
Random Forest	0.728963	0.869231	0.746367	0.803127
Gradient Boosting	0.722114	0.873846	0.737662	0.800000
Support Vector Machine	0.696673	0.836923	0.727273	0.778255

MAKKAKKAKA

CONCLUSION/NEXT STEPS



SINGLE MODEL?

- LESS EXPENSIVE
- SIMPLICITY
- GENERALLY, LOWER
 PERFORMANCE
 - **NEED LESS DATA**



SPECIALIZED MODELS?

- MORE EXPENSIVE
- MORE COMPLEX
- GENERALLY, HIGHER
 PERFORMANCE
 - **NEED MORE DATA**

Final Deliverable - 2020 Rookies

	Player_Id	Prediction
1	a-j-dillon	1
7	adrian-killins-jr	1
28	anthony-mcfarland-2	0
33	antonio-gandy-golden	1
34	antonio-gibson-2	1
35	antonio-williams	1
53	brandon-aiyuk	1
67	cam-akers	1
74	ceedee-lamb	1
78	chase-claypool	1
90	clyde-edwards-helaire	1
93	cole-kmet	1
103	d-andre-swift	1
122	darnell-mooney	1



