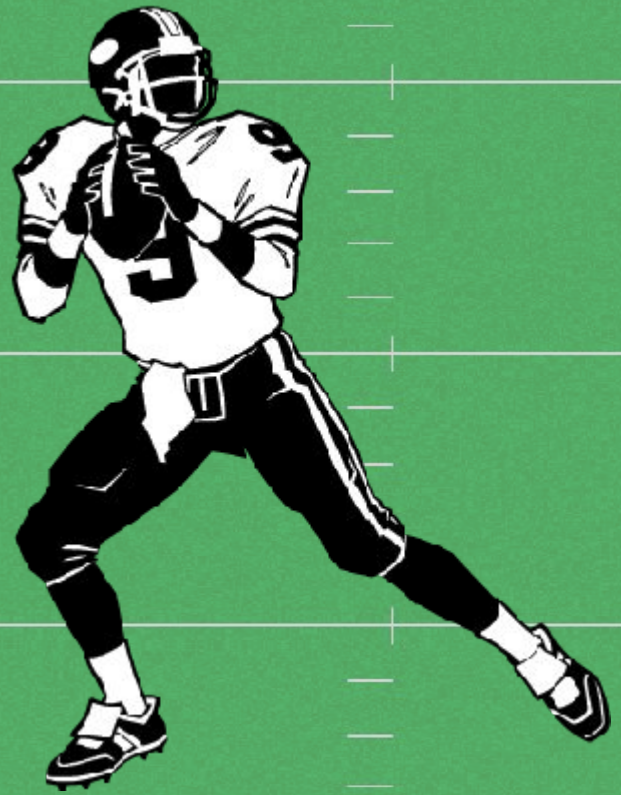


Predicting Touchdowns

Pre-Snap

Presented by:
Albert Troszczynski &
The Thinkful Team

The Data

All regular season NFL plays (2009 - 2017)

- *407,688 plays*
- *102 attributes*

Full Dataset

Research Questions

- What is the most computationally efficient model for predicting touchdowns *before the ball is snapped*?

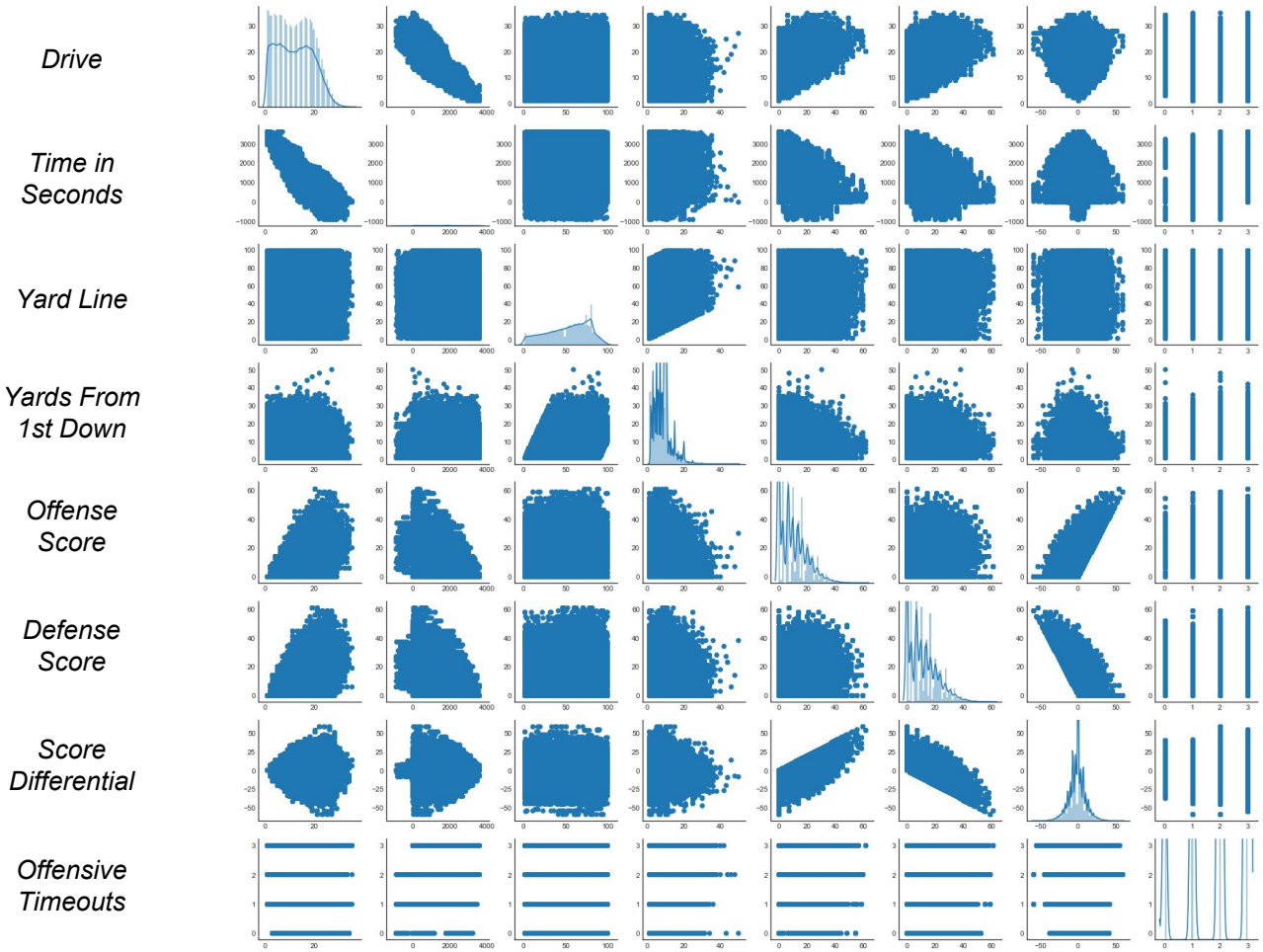
- What factors influence the likelihood of scoring a touchdown?

Feature Engineering

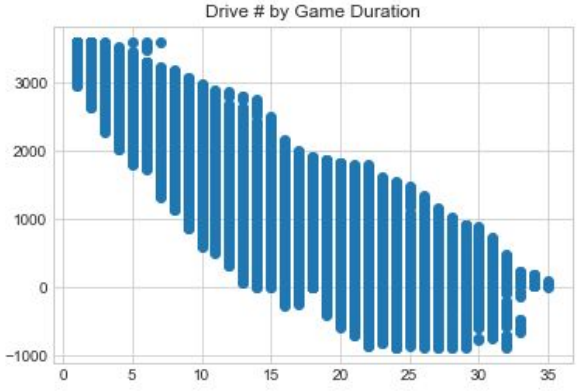
- Removed all attributes related to information *after* the ball had been snapped
- Dropped special team formations from analysis (FG, Kickoff, Punt)
- Created 2 categorical features to represent time: Weekday & Week

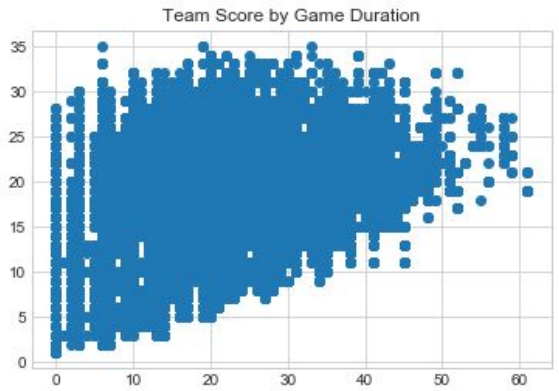
Final Feature Set: 6 Continuous / 8 Categorical (173 dummy variables)

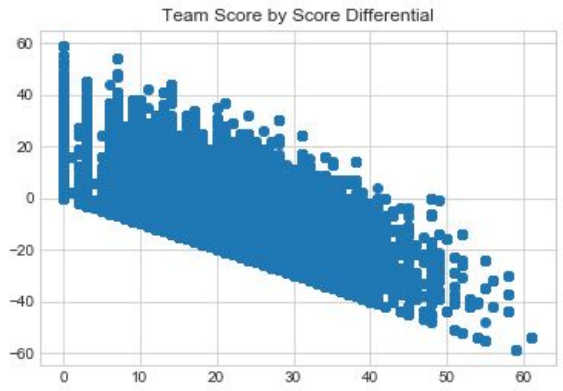
Collinearity and Distribution Analysis of Continuous Variables



	Drive	TimeSecs	yrdline100	ydsfrom1stdown	OffTeamScore	DefTeamScore	ScoreDiff	OffTimeouts	Touchdown
Drive	1.000000	-0.943548	-0.013830	0.022242	0.666066	0.667269	-0.037186	-0.286669	-0.003928
TimeSecs	-0.943548	1.000000	0.066493	-0.010125	-0.676789	-0.683320	0.042716	0.312748	-0.012735
yrdline100	-0.013830	0.066493	1.000000	0.231691	-0.025704	0.003238	-0.025488	0.060667	-0.276037
ydsfrom1stdown	0.022242	-0.010125	0.231691	1.000000	0.006267	0.017767	-0.011020	-0.006588	-0.132329
OffTeamScore	0.666066	-0.676789	-0.025704	0.006267	1.000000	0.386280	0.515851	-0.202339	-0.000894
DefTeamScore	0.667269	-0.683320	0.003238	0.017767	0.386280	1.000000	-0.590922	-0.215733	0.001032
ScoreDiff	-0.037186	0.042716	-0.025488	-0.011020	0.515851	-0.590922	1.000000	0.023397	-0.001740
OffTimeouts	-0.286669	0.312748	0.060667	-0.006588	-0.202339	-0.215733	0.023397	1.000000	-0.019108
Touchdown	-0.003928	-0.012735	-0.276037	-0.132329	-0.000894	0.001032	-0.001740	-0.019108	1.000000

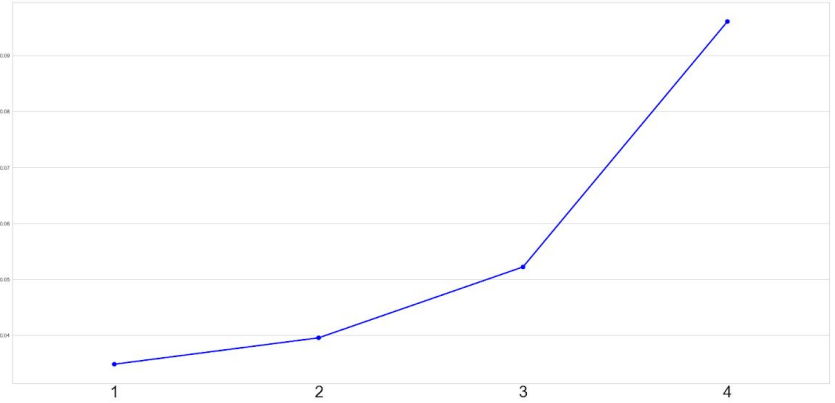




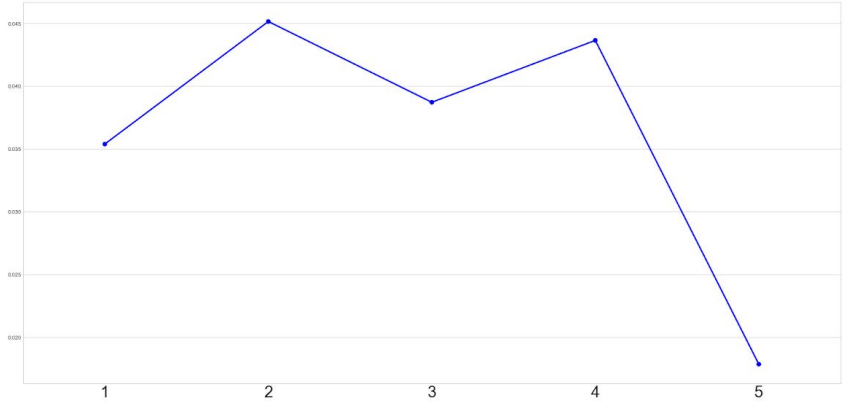


Touchdowns per Play

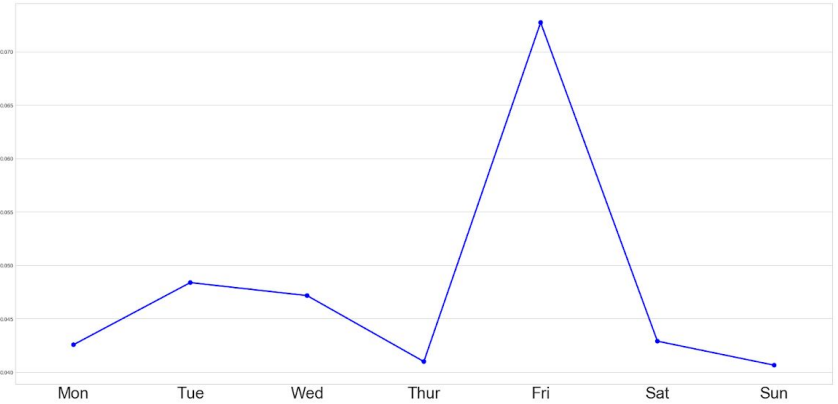
Down



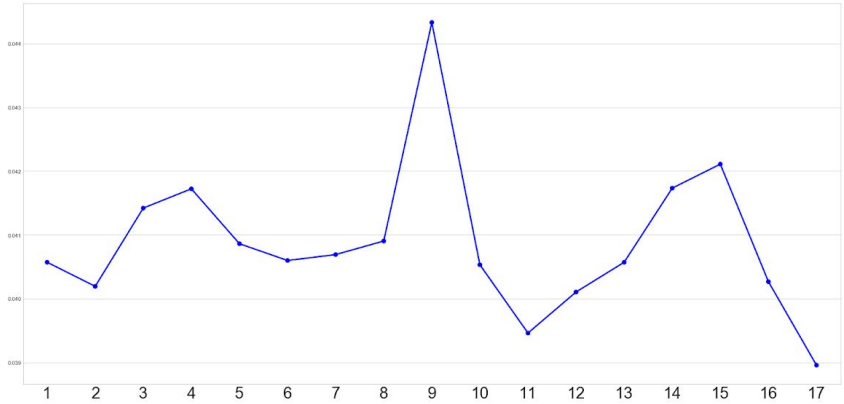
Quarter



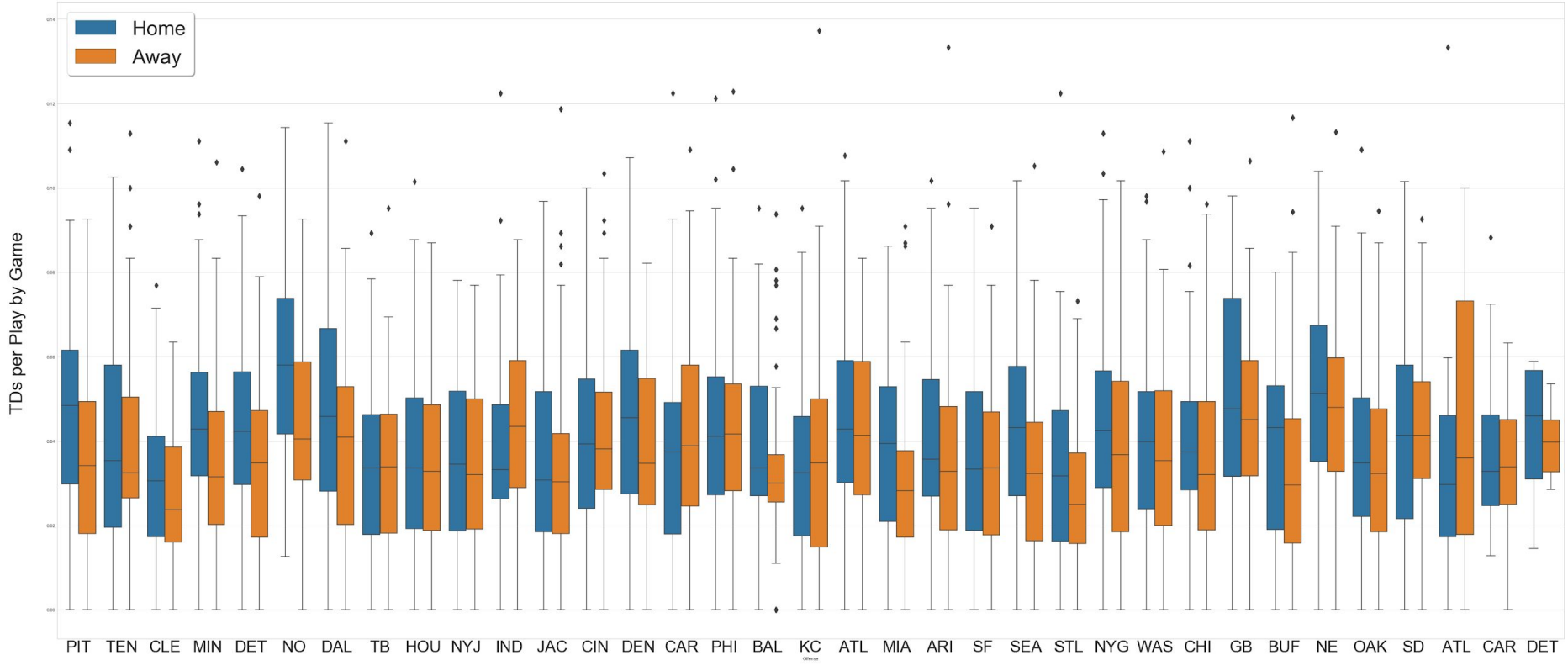
Weekday



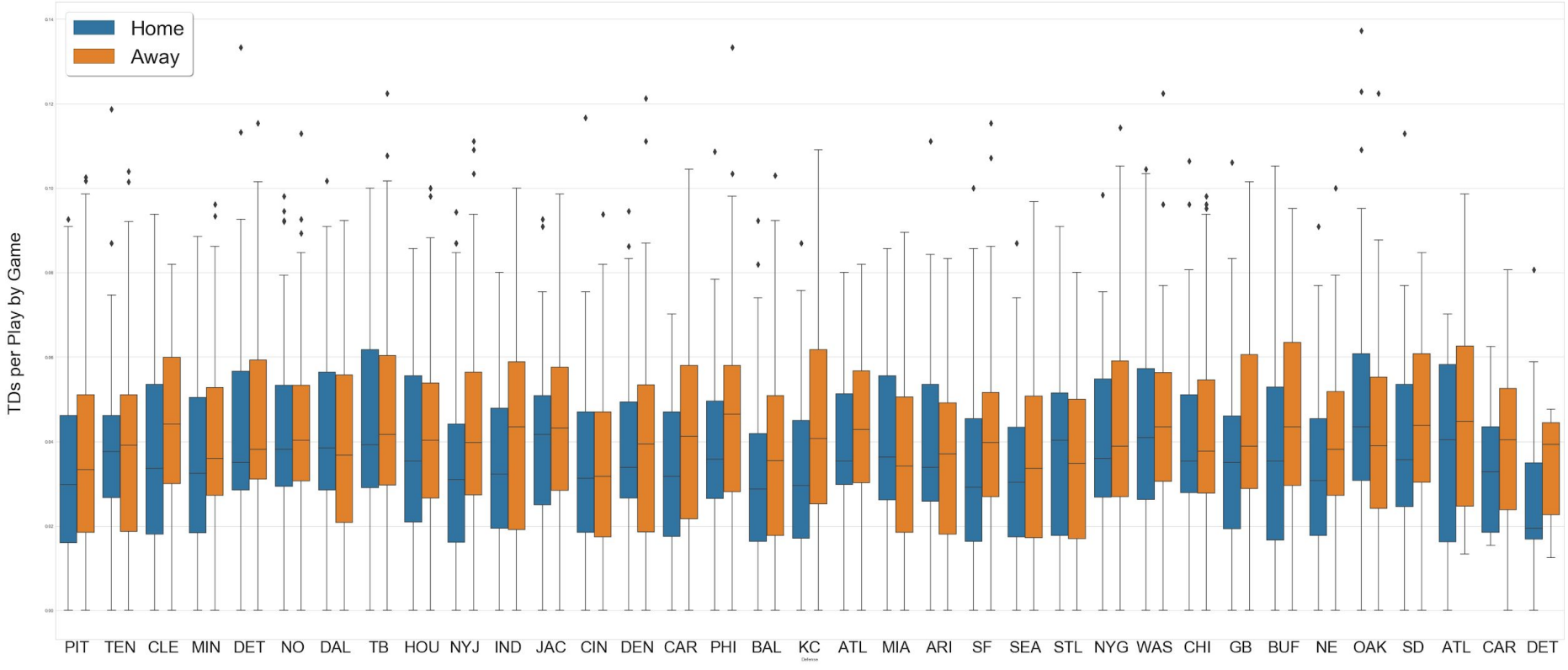
Week



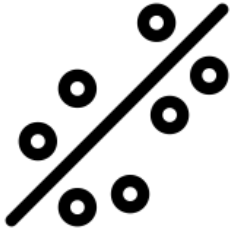
Offensive Performance



Defensive Performance



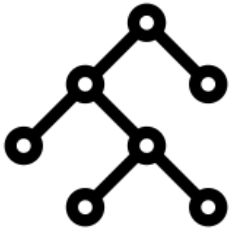
Logistic Regression



Accuracy: 96.07%

Run Time: 0.348 s

Random Forest



Accuracy: 96.07%

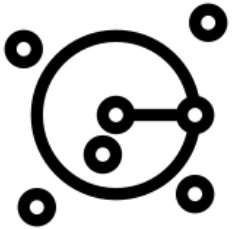
Run Time: 2.098 s

Linear SVM

Accuracy: 95.9%

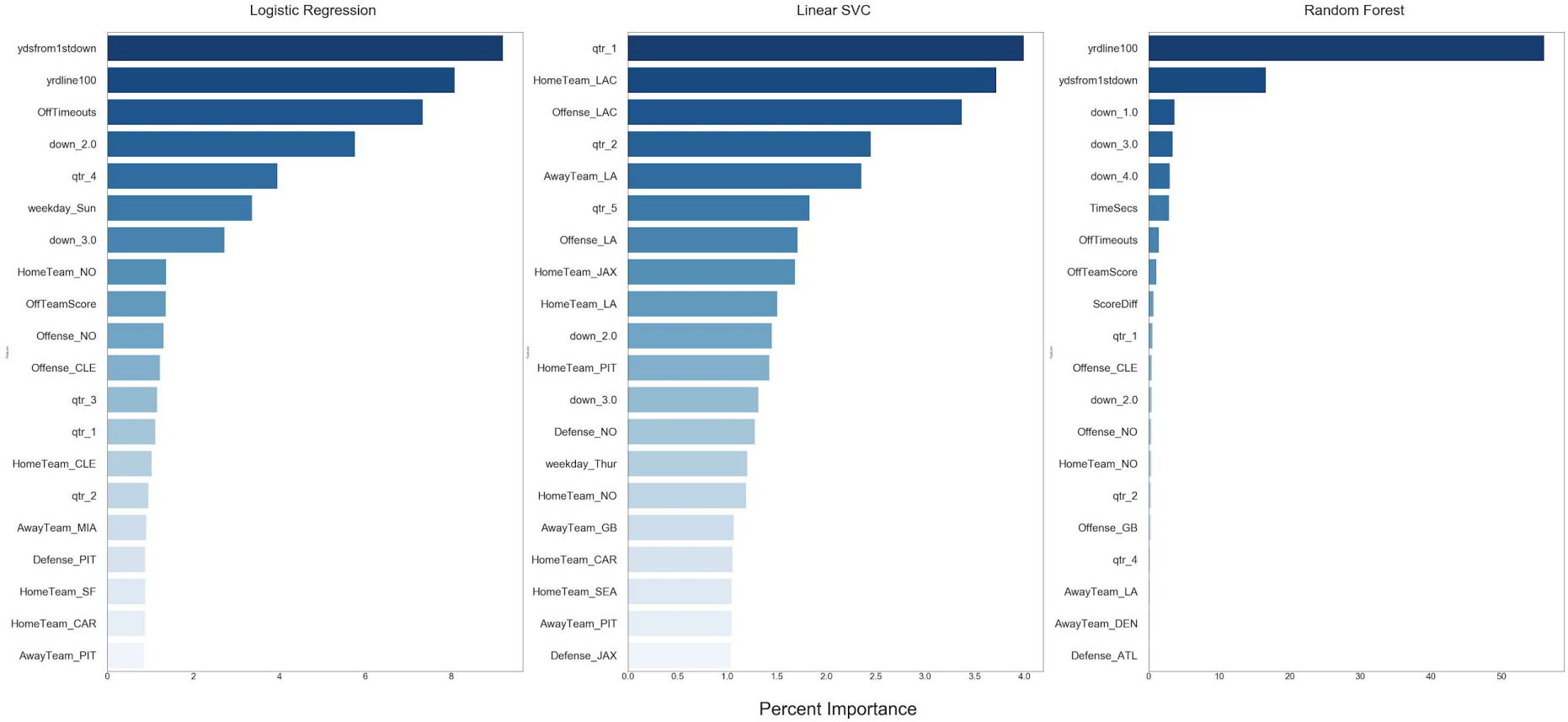
Run Time: 0.336 s

K-Nearest Neighbor



Accuracy: 96.1%

Run Time: 336 s


Highest Importance Features by Model
Logistic Regression
ydsfrom1stdown
yrdline100
OffTimeouts
down_2.0
qtr_4
weekday_Sun
down_3.0
HomeTeam_NO
OffTeamScore
Offense_NO
Offense_CLE
qtr_3
qtr_1
HomeTeam_CLE
qtr_2
AwayTeam_MIA
Defense_PIT
HomeTeam_SF
HomeTeam_CAR
AwayTeam_PIT
0
2
4
6
8
Linear SVC
qtr_1
HomeTeam_LAC
Offense_LAC
qtr_2
AwayTeam_LA
qtr_5
Offense_LA
HomeTeam_JAX
HomeTeam_LA
down_2.0
HomeTeam_PIT
down_3.0
Defense_NO
weekday_Thur
HomeTeam_NO
AwayTeam_GB
HomeTeam_CAR
HomeTeam_SEA
AwayTeam_PIT
Defense_JAX
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Random Forest
yrdline100
ydsfrom1stdown
down_1.0
down_3.0
down_4.0
TimeSecs
OffTimeouts
OffTeamScore
ScoreDiff
qtr_1
Offense_CLE
down_2.0
Offense_NO
HomeTeam_NO
qtr_2
Offense_GB
qtr_4
AwayTeam_LA
AwayTeam_DEN
Defense_ATL
0
10
20
30
40
50
Percent Importance

Conclusions

- We can predict touchdowns with incredible speed and accuracy
- Teams make strategic, in game decisions that impact touchdown likelihood
- There is a black box of missing data, most likely related to the details of how teams operate and their personnel, that opens the door to further analysis