NBA Shot Predictor

Data Science Capstone by Durgesh Murugan

Problem

- To predict the NBA shot outcome based on all the plays from 2015-16 NBA season.
- Shot Outcome - Make or Miss (Shots only)
- Create a model that predicts target more accurately than the baseline.

Target Audience

- NBA/Basketball teams(Coaches + Players) - many also have data scientists
- Sports Analysts
- Sports Betting Companies/Gamblers??(maybe not)

Data Capstone - Outline

Acquire Data->Clean + Exploratory Data Analysis -> Feature Selection + Engineering -> Modelling + Predictions + Score Analysis ->Conclusion

The Data

- The data a is play-by-play Dataset from Basketball-Reference.com. downloaded as a .csv file.
- Each data point/row is a play.
- 600k rows + 40 columns/variables.
- Fairly clean, aside from a few nulls. No serious outliers.

0	URL	601557 non-null	object
1	GameType	601557 non-null	object
2	Location	601557 non-null	object
3	Date	601557 non-null	object
4	Time	601557 non-null	object
5	WinningTeam	601557 non-null	object
6	Quarter	601557 non-null	int64
7	SecLeft	601557 non-null	int64
8	AwayTeam	601557 non-null	object
9	AwayPlay	304900 non-null	object
10	AwayScore	601557 non-null	int64
11	HomeTeam	601557 non-null	object
12	HomePlay	296610 non-null	object
13	HomeScore	601557 non-null	int64
14	Shooter	222288 non-null	object
15	ShotType	222288 non-null	object
16	ShotOutcome	222288 non-null	object
17	ShotDist	222288 non-null	float64
18	Assister	58212 non-null	object
19	Blocker	13031 non-null	object
20	FoulType	54980 non-null	object
21	Fouler	54980 non-null	object
22	Fouled	45972 non-null	object
23	Rebounder	137001 non-null	object
24	ReboundType	137001 non-null	object
25	ViolationPlayer	2322 non-null	object
26	ViolationType	2322 non-null	object
27	TimeoutTeam	17708 non-null	object
28	FreeThrowShooter	61520 non-null	object
29	FreeThrowOutcome	61520 non-null	object
30	FreeThrowNum	61520 non-null	object
31	EnterGame	58999 non-null	object
32	LeaveGame	58999 non-null	object
33	TurnoverPlayer	37660 non-null	object
34	TurnoverType	37660 non-null	object
35	TurnoverCause	20571 non-null	object
36	TurnoverCauser	20571 non-null	object
37	JumpballAwayPlayer	2022 non-null	object
38	JumpballHomePlayer	2022 non-null	object
39	JumpballPoss	2022 non-null	object

NBA - Plays and Shots

- NBA plays are carried out by the offensive team(team who has the ball). It can end in many ways, not necessarily a shot. Stolen, out of bounds,fouled,ball fumbled, time up etc.
- Two types of plays - Shot or No Shot

All NBA shots are a part of plays.

All NBA plays do NOT have shots.

Data Cleaning

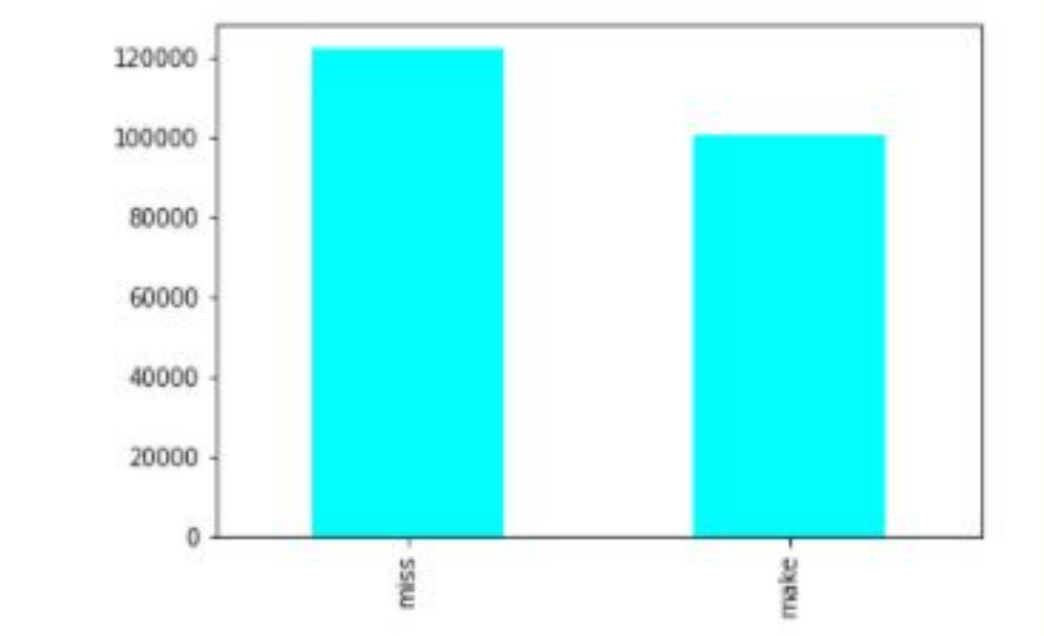
- First all the non-shooting plays were dropped
- Shot Outcomes with null values were dropped. (Make,Miss,NaN)
- Many unnecessary variables/columns were dropped(URL,location,date etc.)
- 600k columns stripped to 220k columns (approx.)
- No other significant cleaning was involved. Fairly straightforward.

EDA

Exploratory Data Analysis

- Explore Data and carry out variable selection.
- Gain major insights
- 'ShotOutcome' is the target variable - Make or Miss
- Categorical + Numeric Features (insights/summary)
- Finalise variables before modeling

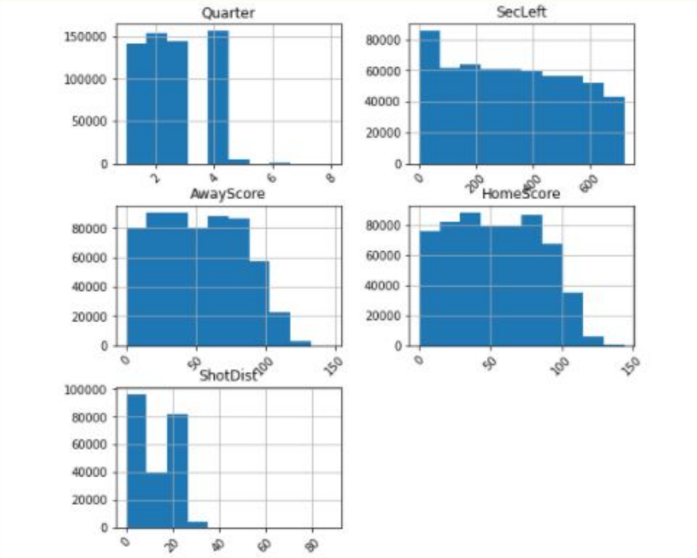
ShotOutcome

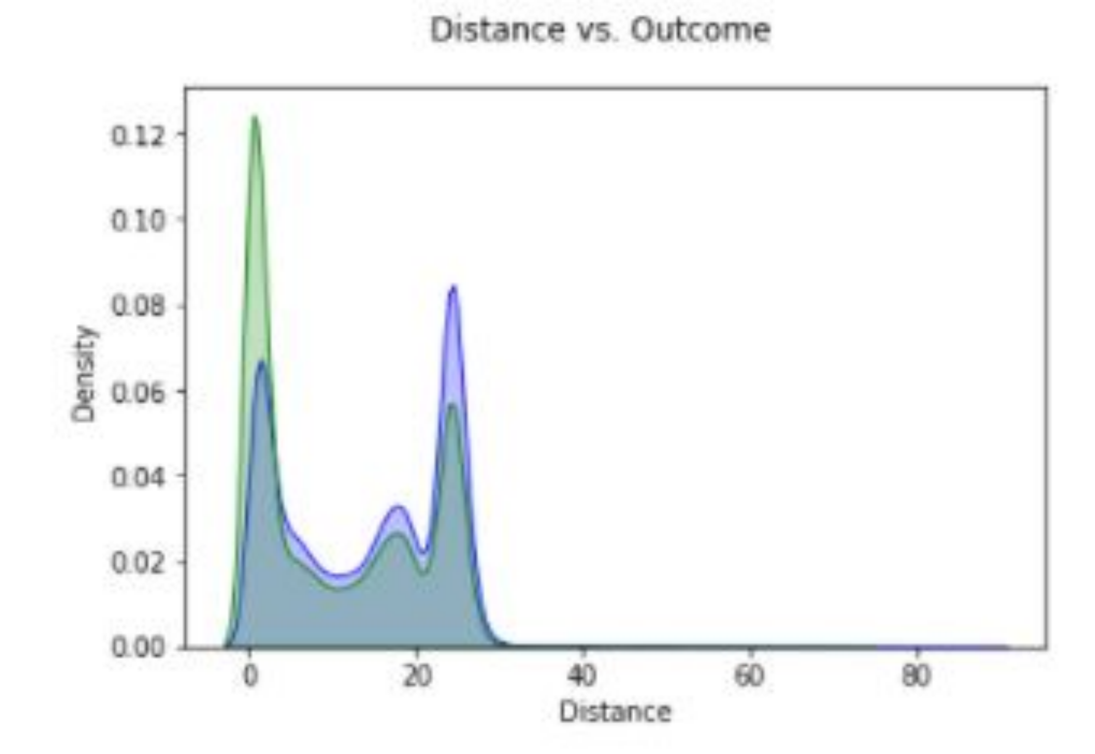



```
miss      121941
make      100347
Name: ShotOutcome, dtype: int64
```

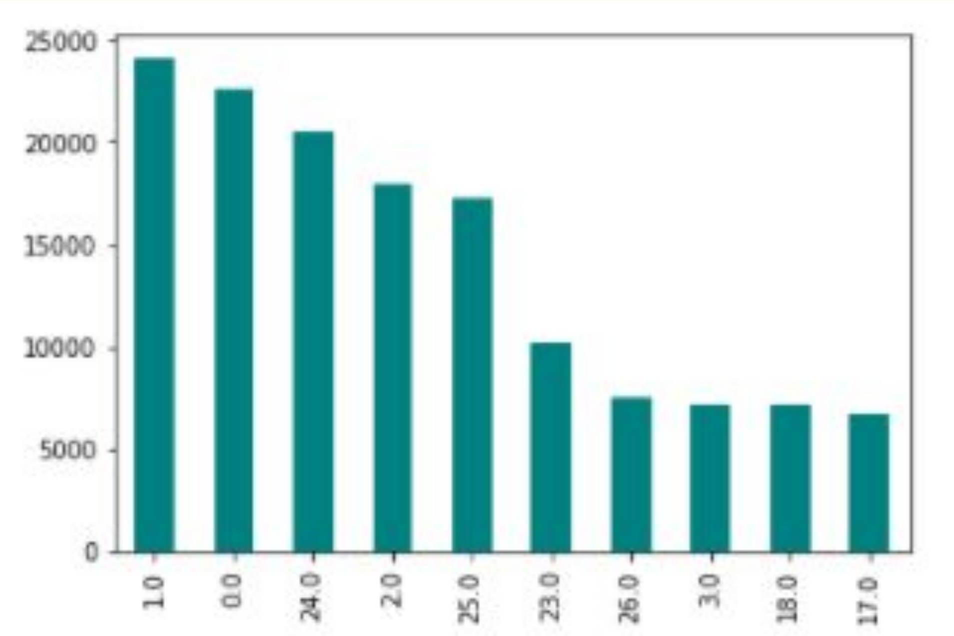
Numerical Variables

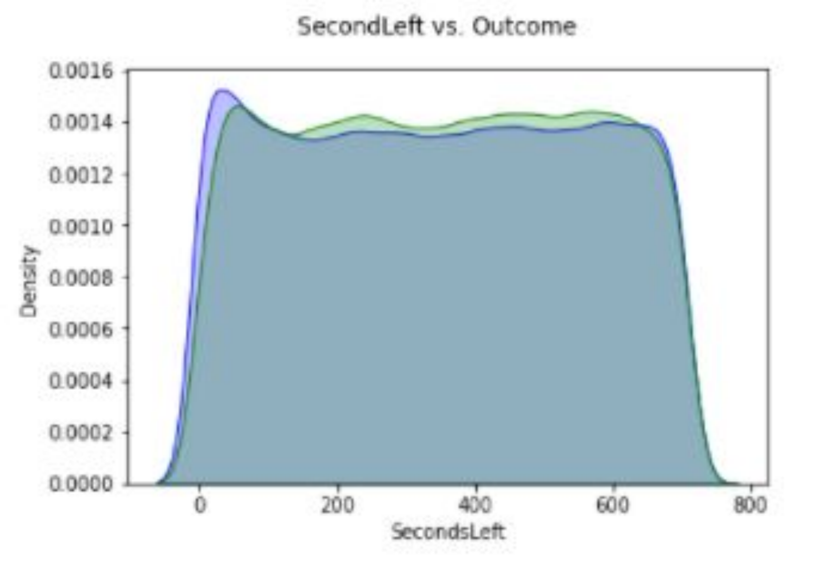
- There are 5 : Quarter, SecLeft, AwayScore, HomeScore, ShotDist
- Quarter,SecLeft,ShotDist seem the most relevant. Scores not so much.



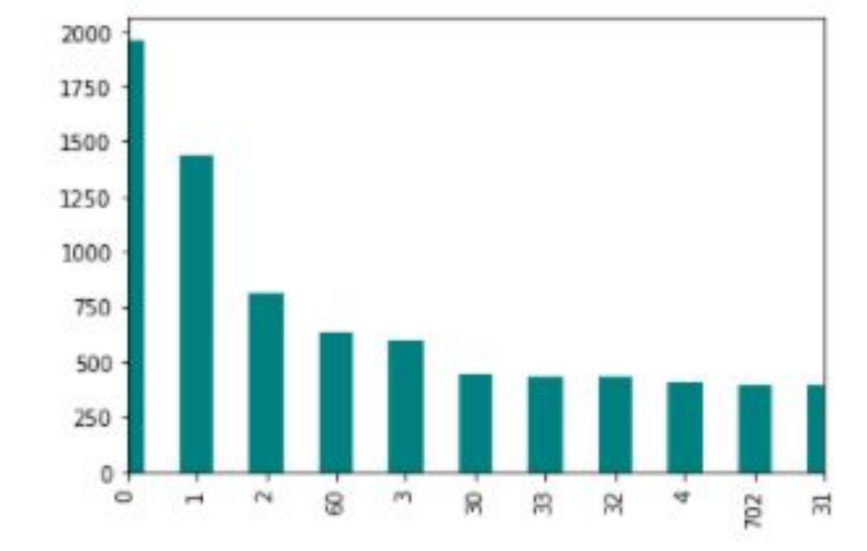

Distance vs. Outcome
Density
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
Distance

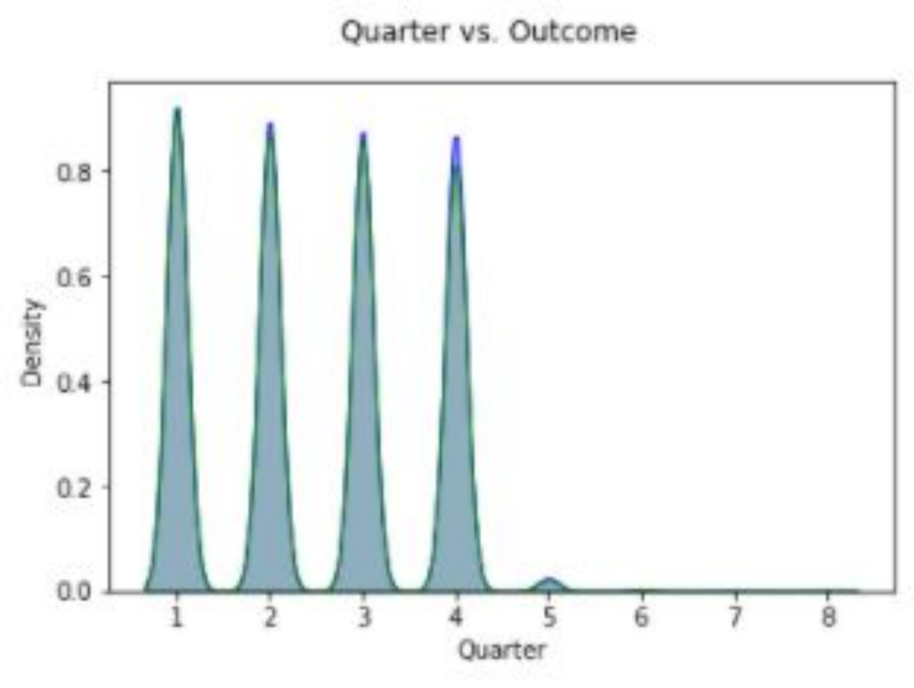
Frequency of Shot Distance in feet

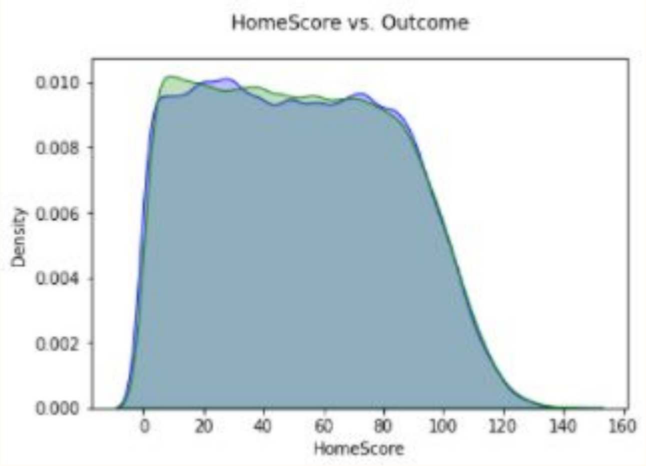


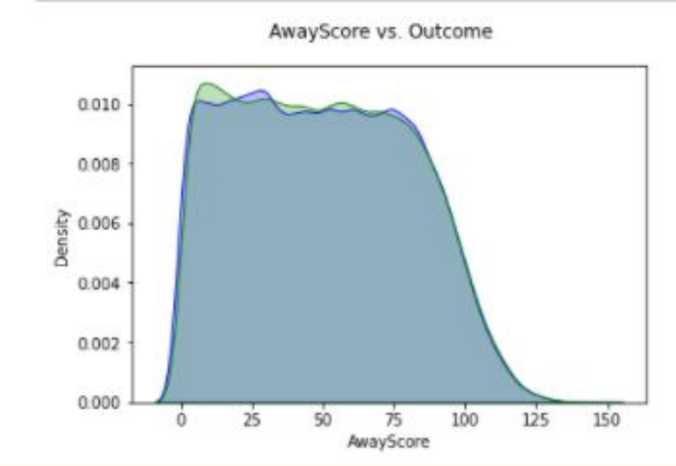

SecondLeft vs. Outcome
Density
0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
200
400
600
800
SecondsLeft

Frequency of Seconds Left




Quarter vs. Outcome
Density
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
6
7
8
Quarter


HomeScore vs. Outcome
Density
0.000
0.002
0.004
0.006
0.008
0.010
0
20
40
60
80
100
120
140
160
HomeScore


AwayScore vs. Outcome
Density
0.000
0.002
0.004
0.006
0.008
0.010
0
25
50
75
100
125
150
AwayScore

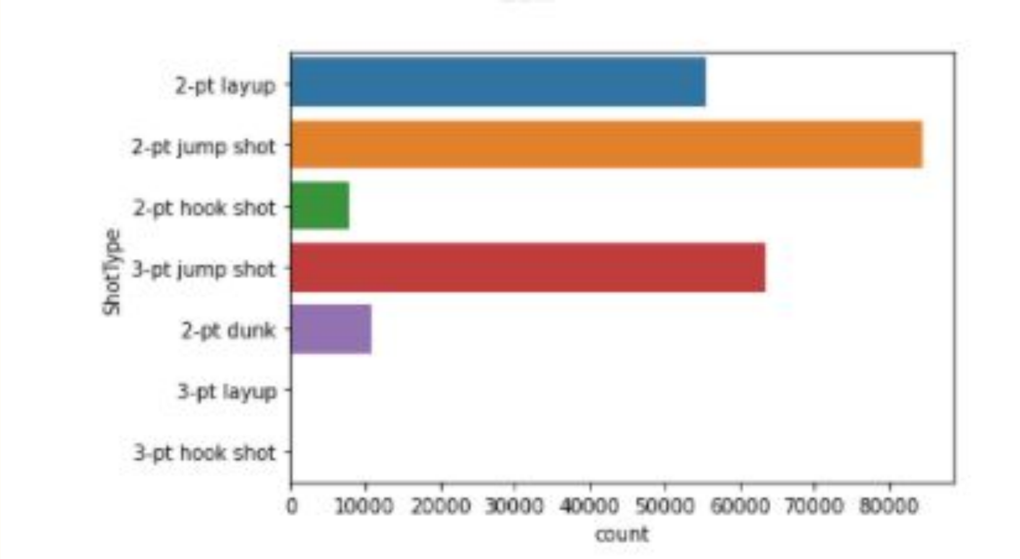
Insights + Summary (Numerical)

- Shot Distance appears to be the biggest influencer.
- Seconds Left and Quarter are smaller but not insignificant.
- Home and Away Scores seem quite insignificant. Can be ignored.

Categorical Variables

- There are 31 categorical variables. Many are not shot related. Most seem inconsequential to the target. Only 3 are interesting, rest can be rejected.
- Relevant ones: ShotType, ShotOutcome, Shooter, Assister

Shot Type



ShotOutcome	make	miss
ShotType		
2-pt dunk	9736	984
2-pt hook shot	3930	3842
2-pt jump shot	32837	51678
2-pt layup	31321	24325
3-pt hook shot	1	4
3-pt jump shot	22522	41108
3-pt layup	0	2

Shooters and Assisters

```
Shooter
S. Curry - curryst01           1933
K. Thompson - thompkl01        1838
R. Westbrook - westbru01       1837
L. James - jamesle01           1833
K. Durant - duranke01          1787
D. DeRozan - derozde01         1775
J. Harden - hardeja01          1717
D. Lillard - lillada01         1716
C. McCollum - mccolcj01        1629
P. George - georgpa01          1571
dtype: int64
```

```
Assister
R. Westbrook - westbru01    1033
R. Rondo - rondora01         838
J. Wall - walljo01           790
C. Paul - paulch01           766
D. Green - greendr01         732
L. James - jamesle01         672
R. Rubio - rubiori01         657
J. Harden - hardeja01        650
S. Curry - curryst01         617
K. Lowry - lowryky01         613
dtype: int64
```

Features/Variables Selected

- Shot Distance
- Seconds Left in the quarter
- Quarter
- Shot Type
- Assister
- Shooter

- Shot Outcome(Target) -make or miss

Modelling

This is a classification problem. The following models were mainly used :

- Logistic Regression
- Random Forests
- Decision Trees

Logistic Regression

Baseline Accuracy: 0.54

Model Score: 0.81, accuracy score: 0.81

Training set score: 0.8168, Test set score: 0.8152 (check overfitting/underfitting)

ROC-AUC Score: 0.8862

```
Confusion Matrix
[[13107  7069]
 [ 1147 23135]]
--------------------------------------------------------------------------------
Classification Report
              precision    recall  f1-score   support

        make       0.92      0.65      0.76     20176
        miss       0.77      0.95      0.85     24282

    accuracy                           0.82     44458
   macro avg       0.84      0.80      0.81     44458
weighted avg       0.84      0.82      0.81     44458

--------------------------------------------------------------------------------
Accuracy 81.52 %
```

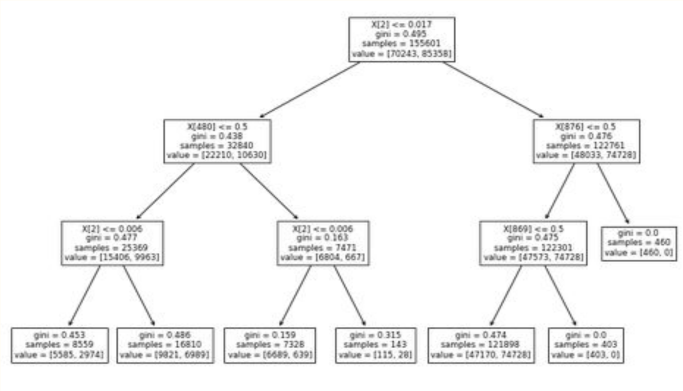
```
Confusion matrix

 [[13107  7069]
 [ 1147 23135]]

True Positives(TP) =  13107

True Negatives(TN) =  23135

False Positives(FP) =  7069

False Negatives(FN) =  1147
```

Decision Tree

Model Score : 0.63

Model accuracy score with criterion gini index: 0.6235

Training set score: 0.6285, Test set score: 0.6231 (check overfitting/underfitting)




```
Confusion matrix

[[ 6434 13742]
 [ 2997 21285]]
```

	precision	recall	f1-score	support
make	0.68	0.32	0.43	20176
miss	0.61	0.88	0.72	24282
accuracy			0.62	44458
macro avg	0.64	0.60	0.58	44458
weighted avg	0.64	0.62	0.59	44458

Random Forests

Model Score : 0.64

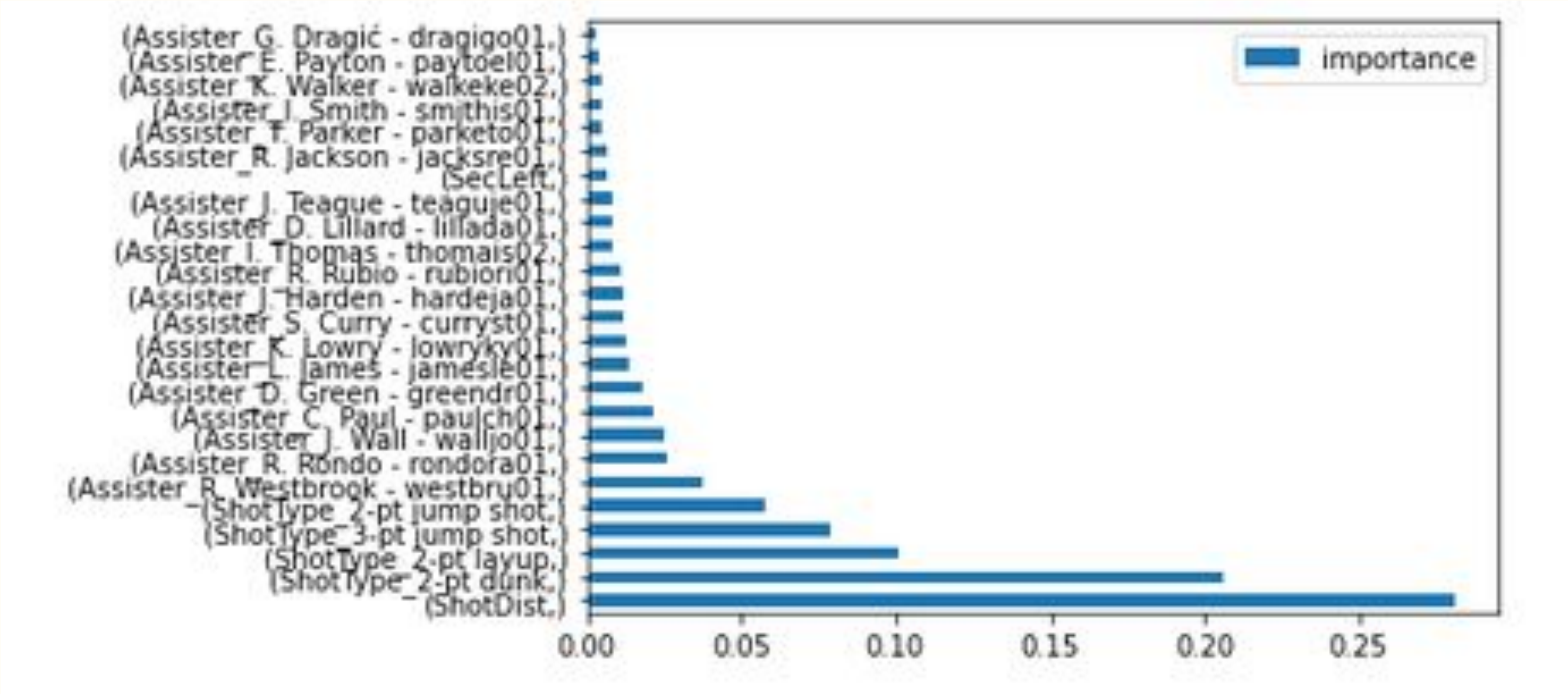
Model accuracy score : 0.6411

Training set score: 0.6414, Test set score: 0.6411 (check overfitting/underfitting)

```
Confusion matrix

[[ 7190 12986]
 [ 2968 21314]]
```

	precision	recall	f1-score	support
make	0.71	0.36	0.47	20176
miss	0.62	0.88	0.73	24282
accuracy			0.64	44458
macro avg	0.66	0.62	0.60	44458
weighted avg	0.66	0.64	0.61	44458

Feature Importance (Random Forests)



Conclusion

- The Logistic Regression model performs the best at predicting the Outcome with an accuracy score of 81%. A strong improvement over the baseline(54%).
- The other two models, Random forests and Decision Trees are not as good but still better than baseline at approximately 61/62%.
- Other models like SVM,XGBoost were tried but not so great.
- All the models show no signs of overfitting/underfitting.

- Shot Distance and Shot Type are the major factors influencing shot outcomes.
- Followed by Assisters.

Moving Forward

- Using a shot log dataset over a play by play-by-play dataset. May have more variables that may help in predicting target(ShotOutcome) better.
- Adding more seasons and increasing the size of dataset. More representative of players.
- Working on improving scores in Random Forests/Decision Trees.
- Applying other models as well, Neural Nets,k-NN.
- Spend less time doing EDA, more time modelling.

- Creating a recommendation system based on players. Recommend shots type + distance. - may need a shot log data set. Host System on website?

Thank You