Trabajo de Final de Grado

Datasets

Machine Learning para la predicción de eventos en la NBA

Albert Villar Ortiz

Universidad Autónoma de Barcelona

ÍNDICE GENERAL

Introducción

Este documento tiene como objetivo dar a conocer los diferentes datasets que se han generado durante todo el proyecto para comprender como esta siendo estructurada toda la información. Para cada uno de los conjuntos de datos entraremos en detalle especificando los atributos que forman parte, así como los resultados obtenidos con los diferentes métodos de feature detection utilizados.

Para este proyecto se han definido diversos datasets diferenciados básicamente por el tipo de información que contemplaban. La idea es que todos y cada uno de los datasets tengan la misma estructura, para así poder unirlos y fusionarlos con la mayor facilidad posible. Queremos entender un dataset como un módulo de datos que podemos coger y fusionar con otros o tratarlos de forma independiente, para así poder jugar con ellos con mucha facilidad.

Equipos

Este dataset corresponde al primer conjunto de datos que tratamos en el proyecto. En él vamos a poder encontrar datos que definen el enfrentamiento (día, hora, equipos…), así como información muy básica del estado de cada uno de los equipos que se enfrentan. Los atributos que podemos encontrar entonces en este dataset son los siguientes:

- **yearSeason:** Año de la temporada en el que se ha disputado el partido.
- **dateGame:** Fecha formada por día y hora en el que se produjo el partido.
- **idGame:** Identificativo único del partido.
- **localIdTeam:** Identificativo único del equipo local del partido.
- **isLocalB2B:** Valor booleano que define si el equipo local viene de B2B o no.
- **isLocalB2BFirst:** Valor booleano que define si para el equipo local este es el primer partido de un B2B.
- **isLocalB2BSecond:** Valor booleano que define si para el equipo local este es el segundo partido de un B2B.
- **localCountDaysRest:** Días que el equipo local ha tenido de descanso antes del partido.
- **localCountDaysNextGameTeam:** Días que el equipo local tiene de descanso hasta el siguiente partido.
- **awayIdTeam:** Identificativo único del equipo visitante del partido.
- **isAwayB2B:** Valor booleano que define si el equipo visitante viene de B2B o no.
- **isAwayB2BFirst:** Valor booleano que define si para el equipo visitante este es el primer partido de un B2B.
- **isAwayB2BSecond:** Valor booleano que define si para el equipo visitante este es el segundo partido de un B2B.
- **awayCountDaysRest:** Días que el equipo visitante ha tenido de descanso antes del partido.
- **awayCountDaysNextGameTeam:** Días que el equipo visitante tiene de descanso hasta el siguiente partido.

Univariate Selection

Tras la ejecución de nuestro primer modelo de detección de características/atributos hemos obtenido los siguientes resultados. Podemos observar en la columna 'Specs' el nombre de la columna que el modelo hace referencia y, justo a su derecha, con el título 'Score', podemos observar la importancia que le da al atributo respecto a la salida de nuestro dataset. Por lo tanto, hemos obtenido de forme descendente el orden de importancia de todas nuestras características en nuestro dataset de equipos.

Dicho análisis se ha realizado sobre la salida de ganador del partido:

	Specs	Score
11	awayCountDaysRest	149.400866
5	localCountDaysRest	75.637716
12	awayCountDaysNextGameTeam	51.179519
6	localCountDaysNextGameTeam	48.700238
10	isAwayB2BSecond	1.283400
8	isAwayB2B	1.227775
3	isLocalB2BFirst	0.890543
2	isLocalB2B	0.439772
9	isAwayB2BFirst	0.345390
4	isLocalB2BSecond	0.329698

Dicho análisis se ha realizado sobre la salida de anotación del equipo local:

	Specs	Score
12	awayCountDaysNextGameTeam	6423.963830
6	localCountDaysNextGameTeam	5715.609818
11	awayCountDaysRest	5657.534573
5	localCountDaysRest	4061.894062
4	isLocalB2BSecond	111.785570
2	isLocalB2B	104.412232
9	isAwayB2BFirst	86.324708
10	isAwayB2BSecond	82.304746
3	isLocalB2BFirst	64.171183
8	isAwayB2B	40.146327

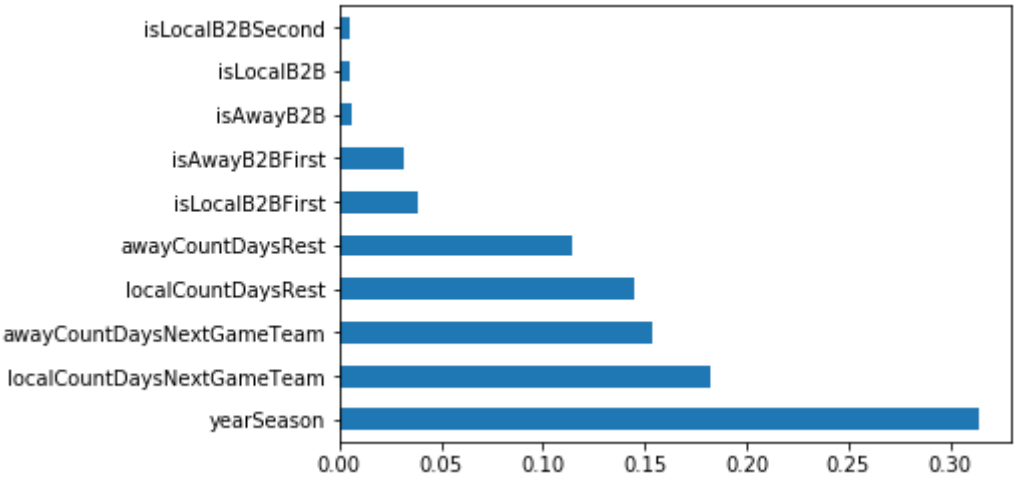
Dicho análisis se ha realizado sobre la salida de anotación del equipo visitante:

	Specs	Score
12	awayCountDaysNextGameTeam	12913.661627
6	localCountDaysNextGameTeam	11827.045409
11	awayCountDaysRest	11666.303118
5	localCountDaysRest	4699.282173
4	isLocalB2BSecond	91.632053
2	isLocalB2B	80.161808
9	isAwayB2BFirst	75.926116
10	isAwayB2BSecond	70.269264
3	isLocalB2BFirst	65.840905
8	isAwayB2B	32.916340

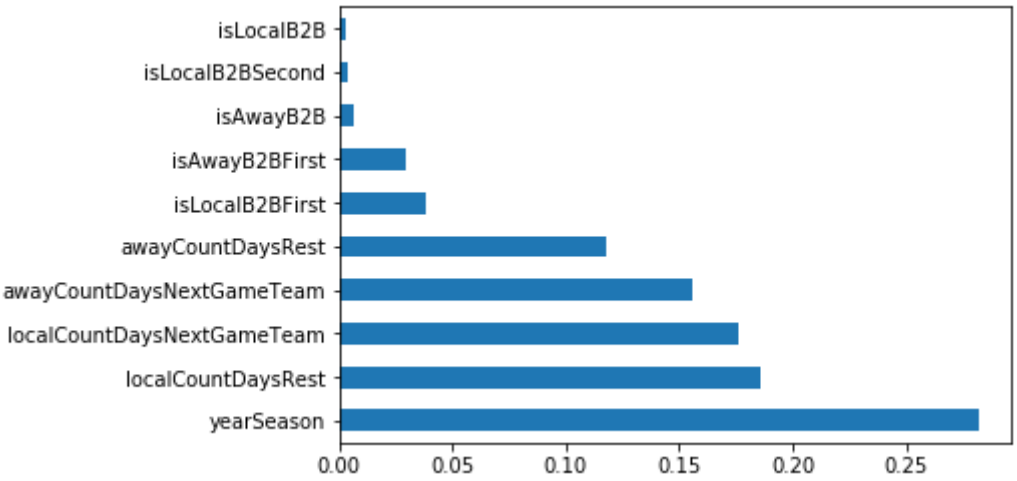
Feature Importance

Por lo que respecta al segundo y último método de selección de atributos que vamos a analizar en este proyecto, hemos realizado un procedimiento muy parecido con el objetivo de poder determinar, solamente teniendo en cuenta el dataset de equipos cuales son los datos que mas influencia tienen en nuestra. Los resultados que hemos obtenido difieren ligeramente de los obtenidos con el método anterior como podemos observar en la siguiente gráfica.

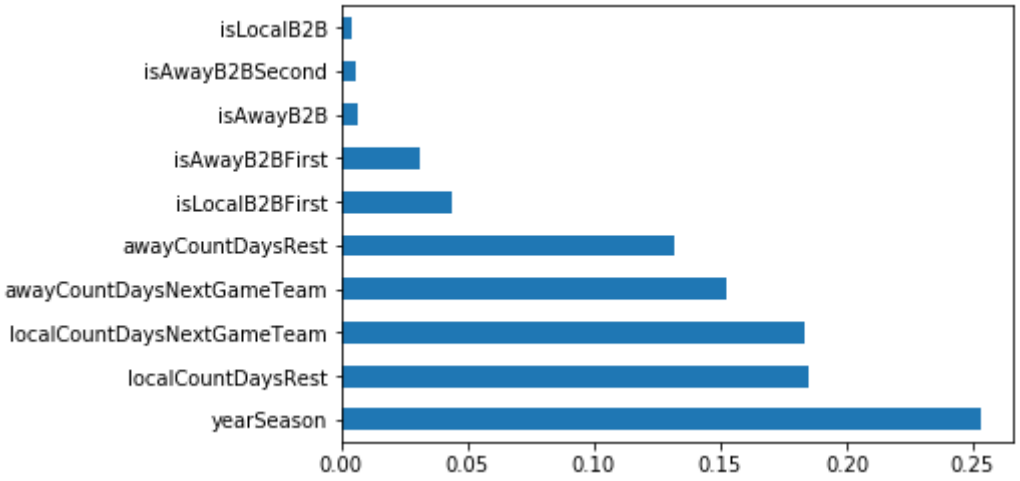
Dicho análisis se ha realizado sobre la salida de ganador del partido:



Dicho análisis se ha realizado sobre la salida de anotación del equipo local:



Dicho análisis se ha realizado sobre la salida de anotación del equipo visitante:



H2H (Head to Head)

Este dataset corresponde al segundo conjunto de datos tratados en el proyecto. En él vamos a poder encontrar datos estadísticos de los enfrentamientos previos entre los dos equipos que jugaran el partido. Los atributos que podemos encontrar entonces en este dataset son los siguientes:

- **yearSeason:** Año de la temporada en el que se ha disputado el partido.
- **dateGame:** Fecha formada por día y hora en el que se produjo el partido.
- **idGame:** Identificativo único del partido.
- **idLocal:** Identificativo único del equipo local en el partido.
- **fgmLocal:** Media del total de tiros anotados por el equipo local durante los partidos anteriores.
- **fgaLocal:** Media del total de tiros lanzados por el equipo local durante los partidos anteriores.
- **pctFGLocal:** Media del porcentaje de tiros anotados respecto al total de tiros lanzados del equipo local durante los partidos anteriores (fgmLocal/fgaLocal).
- **fg3mLocal:** Media del total de tiros de 3 puntos anotados por el equipo local durante los partidos anteriores.
- **fg3aLocal:** Media del total de tiros de 3 puntos lanzados por el equipo local durante los partidos anteriores.
- **pctFG3Local:** Media del porcentaje de tiros de 3 puntos anotados respecto al total de tiros de 3 lanzados del equipo local durante los partidos anteriores (fg3mLocal/fg3aLocal).
- **fg2mLocal:** Media del total de tiros de 2 puntos anotado por el equipo local durante los partidos anteriores.
- **fg2aLocal:** Media del total de tiros de 2 puntos lanzados por el equipo local durante los partidos anteriores.
- **pctFG2Local:** Media del porcentaje de tiros de 3 puntos anotados respecto al total de tiros de 2 puntos lanzados del equipo local durante los partidos anteriores (fg2mLocal/fg2aLocal).

- **ftmLocal:** Media del total de tiros libres anotados por el equipo local durante los partidos anteriores.
- **ftaLocal:** Media del total de tiros libres lanzados por el equipo local durante los partidos anteriores.
- **pctFTLocal:** Media del porcentaje de tiros libres anotados respecto al total de tiros libres lanzados del equipo local durante los partidos anteriores (ftmLocal/ftaLocal).
- **orebLocal:** Media del total de rebotes ofensivos que ha realizado el equipo local durante los partidos anteriores.
- **drebLocal:** Media del total de rebotes defensivos que ha realizado el equipo local durante los partidos anteriores.
- trebLocal:
- **astLocal:** Media del total de asistencias que ha realizado el equipo local durante los partidos anteriores.
- **stlLocal:** Media del total de robos de balón que ha realizado el equipo local durante los partidos anteriores.
- **blkLocal:** Media del total de tapones que ha realizado el equipo local durante los partidos anteriores.
- **tovLocal:** Media del total de balones perdidos que ha realizado el equipo local durante los partidos anteriores.
- **pfLocal:** Media de las faltas personales realizadas por el equipo local durante los partidos anteriores.
- **ptsLocal:** Media del total de puntos realizados por el equipo local durante los partidos anteriores.
- **plsmnsLocal:** Media de la valoración total del equipo local durante los partidos anteriores.
- **idAway:** Identificativo único del equipo visitante en el partido.
- **fgmAway:** Media del total de tiros anotados por el equipo visitante durante los partidos anteriores.
- **fgaAway:** Media del total de tiros lanzados por el equipo visitante durante los partidos anteriores.
- **pctFGAway:** Media del porcentaje de tiros anotados respecto al total de tiros lanzados del equipo visitante durante los partidos anteriores (fgmLocal/fgaLocal).
- **fg3mAway:** Media del total de tiros de 3 puntos anotados por el equipo visitante durante los partidos anteriores.
- **fg3aAway:** Media del total de tiros de 3 puntos lanzados por el equipo visitante durante los partidos anteriores.
- **pctFG3Away:** Media del porcentaje de tiros de 3 puntos anotados respecto al total de tiros de 3 lanzados del equipo visitante durante los partidos anteriores (fg3mLocal/fg3aLocal).
- **fg2mAway:** Media del total de tiros de 2 puntos anotado por el equipo visitante durante los partidos anteriores.
- **fg2aAway:** Media del total de tiros de 2 puntos lanzados por el equipo visitante durante los partidos anteriores.

- **pctFG2Away:** Media del porcentaje de tiros de 3 puntos anotados respecto al total de tiros de 2 puntos lanzados del equipo visitante durante los partidos anteriores (fg2mLocal/fg2aLocal).
- **ftmAway:** Media del total de tiros libres anotados por el equipo visitante durante los partidos anteriores.
- **ftaAway:** Media del total de tiros libres lanzados por el equipo visitante durante los partidos anteriores.
- **pctFTAway:** Media del porcentaje de tiros libres anotados respecto al total de tiros libres lanzados del equipo visitante durante los partidos anteriores (ftmLocal/ftaLocal).
- **orebAway:** Media del total de rebotes ofensivos que ha realizado el equipo visitante durante los partidos anteriores.
- **drebAway:** Media del total de rebotes defensivos que ha realizado el equipo visitante durante los partidos anteriores.
- treb**Away**:
- **astAway:** Media del total de asistencias que ha realizado el equipo visitante durante los partidos anteriores.
- **stlAway:** Media del total de robos de balón que ha realizado el equipo visitante durante los partidos anteriores.
- **blkAway:** Media del total de tapones que ha realizado el equipo visitante durante los partidos anteriores.
- **tovAway:** Media del total de balones perdidos que ha realizado el equipo visitante durante los partidos anteriores.
- **pfAway:** Media de las faltas personales realizadas por el equipo visitante durante los partidos anteriores.
- **ptsAway:** Media del total de puntos realizados por el equipo visitante durante los partidos anteriores.
- **plsmnsAway:** Media de la valoración total del equipo visitante durante los partidos anteriores.

Univariate Selection

Tras la ejecución de nuestro primer modelo de detección de características/atributos hemos obtenido los siguientes resultados. Podemos observar en la columna ‘Specs’ el nombre de la columna que el modelo hace referencia y, justo a su derecha, con el título ‘Score’, podemos observar la importancia que le da al atributo respecto a la salida de nuestro dataset. Por lo tanto, hemos obtenido de forme descendente el orden de importancia de todas nuestras características en nuestro dataset de equipos.

Dicho análisis se ha realizado sobre la salida de ganador del partido:

	Specs	Score
46	winsAway	223.847759
45	winsLocal	178.581321
28	fg3aAway	16.742112
19	blkLocal	14.172979
22	ptsLocal	13.906216
44	ptsAway	11.874628
37	drebAway	10.975616
27	fg3mAway	10.832497
6	fg3aLocal	9.169565
9	fg2aLocal	8.954497

Dicho análisis se ha realizado sobre la salida de anotación del equipo local:

	Specs	Score
45	winsLocal	453.165370
46	winsAway	365.035656
6	fg3aLocal	167.224598
22	ptsLocal	133.369535
28	fg3aAway	83.404150
5	fg3mLocal	81.723403
44	ptsAway	68.525552
25	fgaAway	55.032493
2	fgmLocal	49.857049
17	astLocal	47.034502

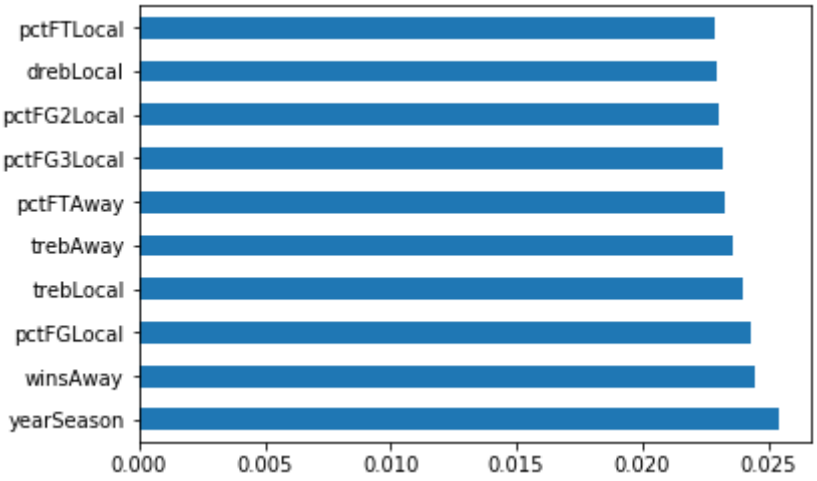
Dicho análisis se ha realizado sobre la salida de anotación del equipo visitante:

	Specs	Score
46	winsAway	473.389745
45	winsLocal	323.914501
28	fg3aAway	154.390782
44	ptsAway	124.206302
6	fg3aLocal	91.663317
27	fg3mAway	78.058365
3	fgaLocal	60.031132
22	ptsLocal	59.830381
39	astAway	52.357202
24	fgmAway	52.035247

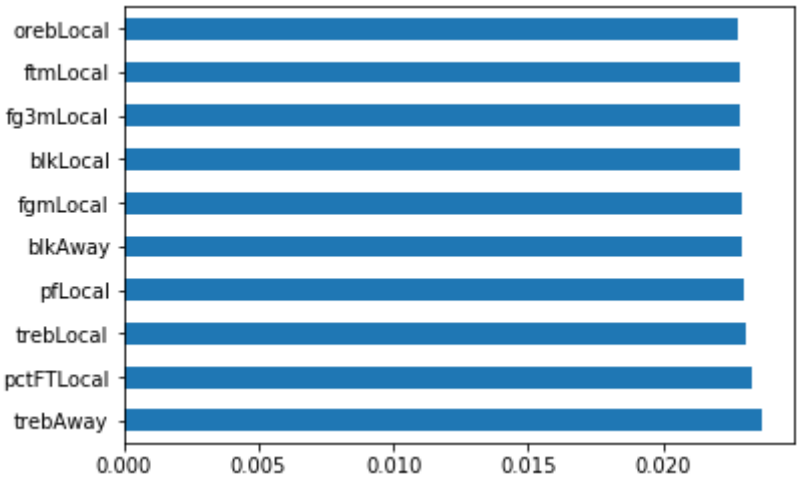
Feature Importance

Por lo que respecta al segundo y último método de selección de atributos que vamos a analizar en este proyecto, hemos realizado un procedimiento muy parecido con el objetivo de poder determinar, solamente teniendo en cuenta el dataset de enfrentamientos cuales son los datos que más influencia tienen en nuestra salida.

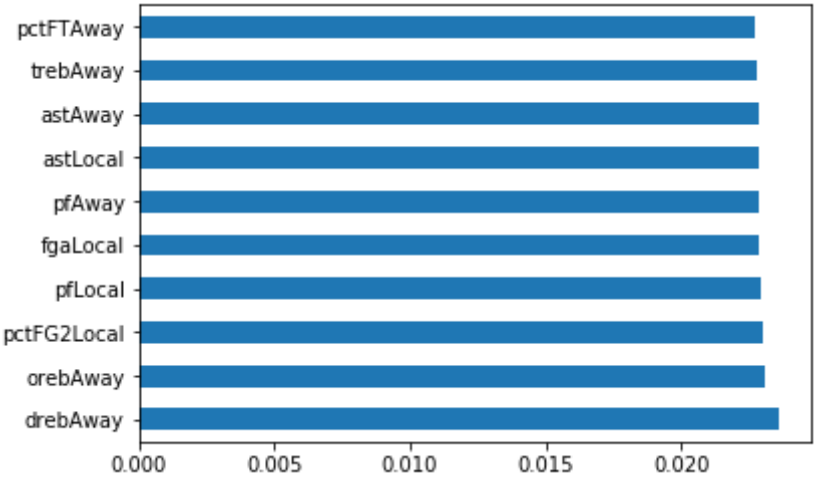
Dicho análisis se ha realizado sobre la salida de ganador del partido:



Dicho análisis se ha realizado sobre la salida de la anotación del equipo local:



Dicho análisis se ha realizado sobre la salida de la anotación del equipo visitante:



Jugadores/Participantes

Este dataset corresponde al tercer conjunto de datos tratados en el proyecto. En él vamos a poder encontrar datos sobre quién va a participar en él y quien por lo contrario se va a mantener ausente. Los atributos que podemos encontrar entonces en este dataset son los siguientes:

- **yearSeason:** Año de la temporada en el que se ha disputado el partido.
- **dateGame:** Fecha formada por día y hora en el que se produjo el partido.
- **idGame:** Identificativo único del partido.
- **idLocal:** Identificativo único del equipo local en el partido.
- **activePlayersL:** Listado con los identificativos de los jugadores del equipo local que han participado en el partido.
- **inactivePlayersL:** Listado con los identificativos de los jugadores del equipo local que no han participado en el partido.
- **idAway:** Identificativo único del equipo visitante en el partido.
- **activePlayersA:** Listado con los identificativos de los jugadores del equipo visitante que han participado en el partido.
- **inactivePlayersA:** Listado con los identificativos de los jugadores del equipo visitante que no han participado en el partido.

Cuotas/Probabilidades

Este dataset corresponde al cuarto conjunto de datos tratados en el proyecto. En él vamos a poder encontrar datos sobre cuáles son las probabilidades que les ha asignado las casas de apuestas a la victoria a cada uno de los equipos. Los atributos que podemos encontrar entonces en este dataset son los siguientes:

- **yearSeason:** Año de la temporada en el que se ha disputado el partido.
- **dateGame:** Fecha formada por día y hora en el que se produjo el partido.
- **idGame:** Identificativo único del partido.
- **idLocal:** Identificativo único del equipo local del partido.
- **oddLocalWin:** Cuota establecida por las casas de apuestas a la victoria del equipo local.
- **idAway:** Identificativo único del equipo visitante del partido.
- **oddAwayWin:** Cuota establecida por las casas de apuestas a la victoria del equipo visitante.

Target/Resultados

Este dataset corresponde al último conjunto de datos tratados en el proyecto. En él vamos a poder encontrar los resultados de cada uno de nuestros enfrentamientos. Es decir, este dataset será considerado el target de todos los nuestros modelos matemáticos. Los atributos que podemos encontrar entonces en este dataset son los siguientes:

- **yearSeason:** Año de la temporada en el que se ha disputado el partido.
- **dateGame:** Fecha formada por día y hora en el que se produjo el partido.
- **idGame:** Identificativo único del partido.
- **localPts:** Puntos anotados por el equipo local en ese partido.
- **awayPts:** Puntos anotados por el equipo visitante en ese partido.
- **winner:** Te mostrará un 0 si el ganador es el equipo local y un 1 si el ganador es el equipo visitante.

Dataset total

En este dataset vamos a contar con todas las características que hemos comentado en los puntos anteriores. Una vez unidos todos los atributos procedemos también a realizar la obtención de características.

Univariate Selection

Tras la ejecución de nuestro primer modelo de detección de características/atributos hemos obtenido los siguientes resultados. Podemos observar en la columna 'Specs' el nombre de la columna que el modelo hace referencia y, justo a su derecha, con el título 'Score', podemos observar la importancia que le da al atributo respecto a la salida de nuestro dataset. Por lo tanto, hemos obtenido de forme descendente el orden de importancia de todas nuestras características en nuestro dataset total.

Dicho análisis se ha realizado sobre la salida de ganador del partido:

	Specs	Score
91	inactPlayerAway11	3.192117e+07
92	inactPlayerAway12	1.605234e+07
30	actPlayerLocal3	1.293235e+07
47	inactPlayerLocal8	1.253536e+07
29	actPlayerLocal2	1.091816e+07
31	actPlayerLocal4	1.065150e+07
39	actPlayerLocal12	9.554232e+06
46	inactPlayerLocal7	9.296589e+06
44	inactPlayerLocal5	9.032648e+06
45	inactPlayerLocal6	8.974005e+06

Dicho análisis se ha realizado sobre la salida de anotación del equipo local:

	Specs	Score
39	actPlayerLocal12	2.054806e+08
92	inactPlayerAway12	1.644440e+08
91	inactPlayerAway11	8.216955e+07
100	inactPlayerAway8	7.298931e+07
38	actPlayerLocal11	6.762586e+07
49	inactPlayerLocal10	6.207822e+07
102	inactPlayerAway10	5.174262e+07
44	inactPlayerLocal5	4.383186e+07
45	inactPlayerLocal6	4.226589e+07
99	inactPlayerAway7	4.211061e+07

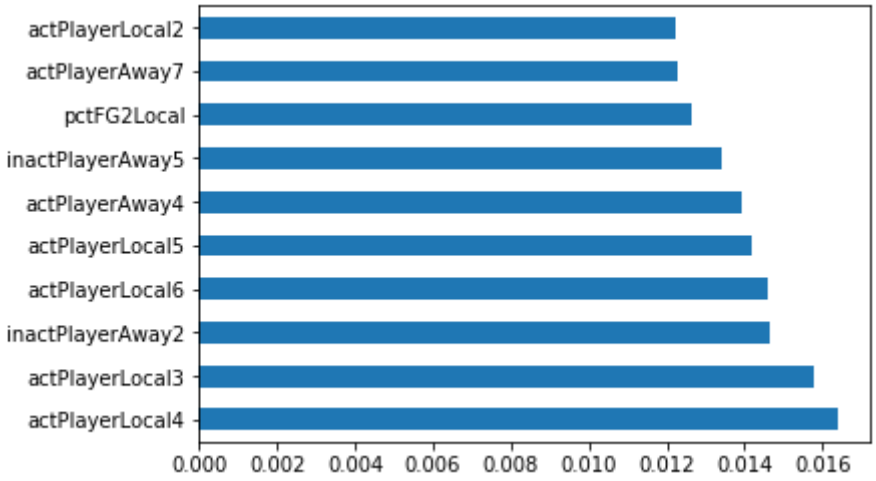
Dicho análisis se ha realizado sobre la salida de anotación del equipo visitante:

	Specs	Score
100	inactPlayerAway8	1.384967e+08
39	actPlayerLocal12	1.283368e+08
92	inactPlayerAway12	9.396128e+07
45	inactPlayerLocal6	8.694924e+07
44	inactPlayerLocal5	8.412871e+07
49	inactPlayerLocal10	8.237642e+07
91	inactPlayerAway11	6.555767e+07
38	actPlayerLocal11	6.554571e+07
47	inactPlayerLocal8	6.173889e+07
97	inactPlayerAway5	5.772619e+07

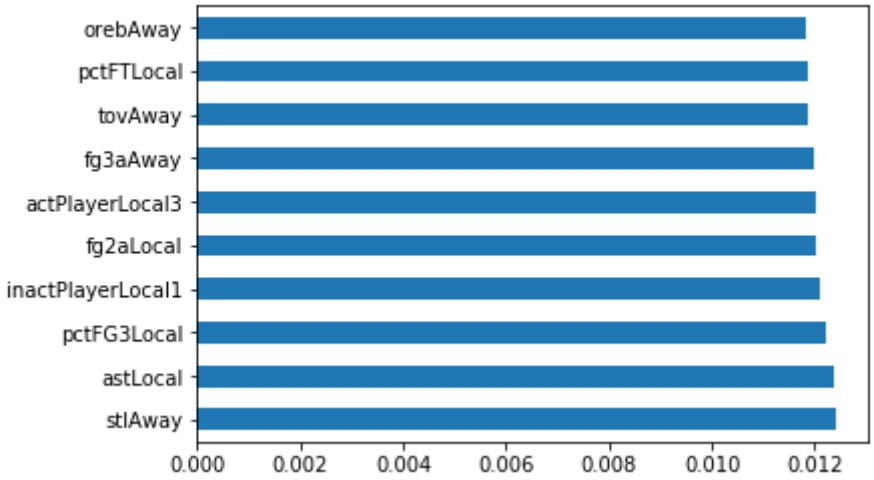
Feature Importance

Por lo que respecta al segundo y último método de selección de atributos que vamos a analizar en este proyecto, hemos realizado un procedimiento muy parecido con el objetivo de poder determinar, solamente teniendo en cuenta todo el dataset cuales son los datos que más influencia tienen en nuestra salida.

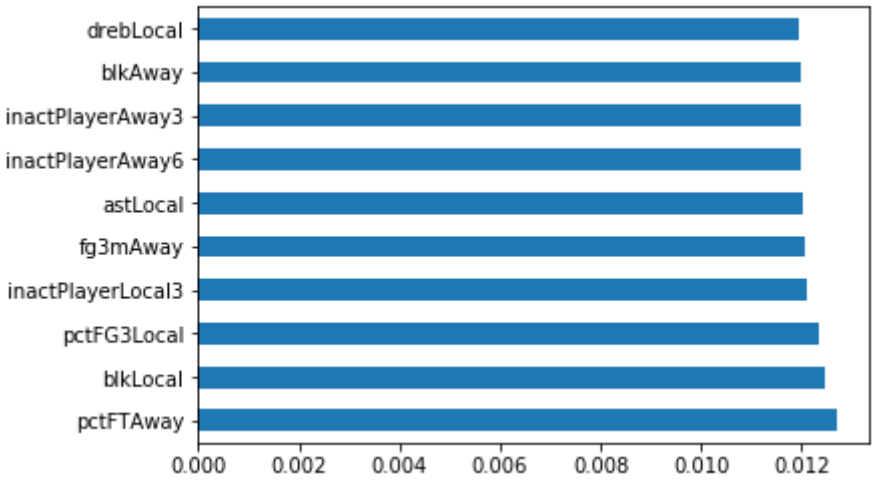
Dicho análisis se ha realizado sobre la salida de ganador del partido:



Dicho análisis se ha realizado sobre la salida de anotación del equipo local:



Dicho análisis se ha realizado sobre la salida de anotación del equipo visitante:


drebLocal
blkAway
inactPlayerAway3
inactPlayerAway6
astLocal
fg3mAway
inactPlayerLocal3
pctFG3Local
blkLocal
pctFTAway
0.000
0.002
0.004
0.006
0.008
0.010
0.012