Resultados Analisis

Oscar Julian Rodriguez Cardenas
Introducción Machine Learning para economistas

April 29, 2022

La base de datos escogida fué tomada de Kaggle, se basa en informaciona cerca de la liga inglesa de fútbol o Premier League en la temporada 2017-2018, cuenta con 461 jugadores y 17 columnas las cuales estan compuestas de la siguiente forma:

1. **name:** Nombre del jugador.
2. **club:** Club del jugador.
3. **age:** Edad del jugador.
4. **position:** Posición en la juega dentro de la cancha.
5. **position_cat:** Categoría de la posición del jugador, donde 1 significa que es atacante, 2 significa que es mediocamposta, 3 significa que es defensor y 4 significa que es arquero.
6. **market_value**: Valor del jugador en el mercado (Millones de Euros).
7. **page_views**: Número de visitas a la página del jugador en Wkipedia desde septiembre 1 del 2016 hasta mayo 1 del 2017.
8. **fpl_value:** Valor del jugador en el juego Fantasy Premier League.
9. **fpl_sel:** Porcentaje de jugadores de Fantasy Premier League que escogen al jugador para su equipo.
10. **fpl_points:** Puntos en Fantasy Premier Leagur obtenidos por el jugador.
11. **region:** Región de la que es el jugador, donde 1 es Inglaterra, 2 es Europa, 3 es America y 4 el resto del mundo.
12. **nationality:** Nacionalidad del jugador o país de origen.
13. **new_foreign:** Indica si el jugador fué fichado de una liga diferente en está temporada de 2017/18.
14. **age_cat:** Categoiría de la edad.

15. **club_id:** Id única del club.

16. **big_club:** Indica si el jugador pertenece a uno de los seis grandes equipos o Big Six (Manchester United, Manchester City, Liverpool, Tottenham, Arsenal, Chelsea).

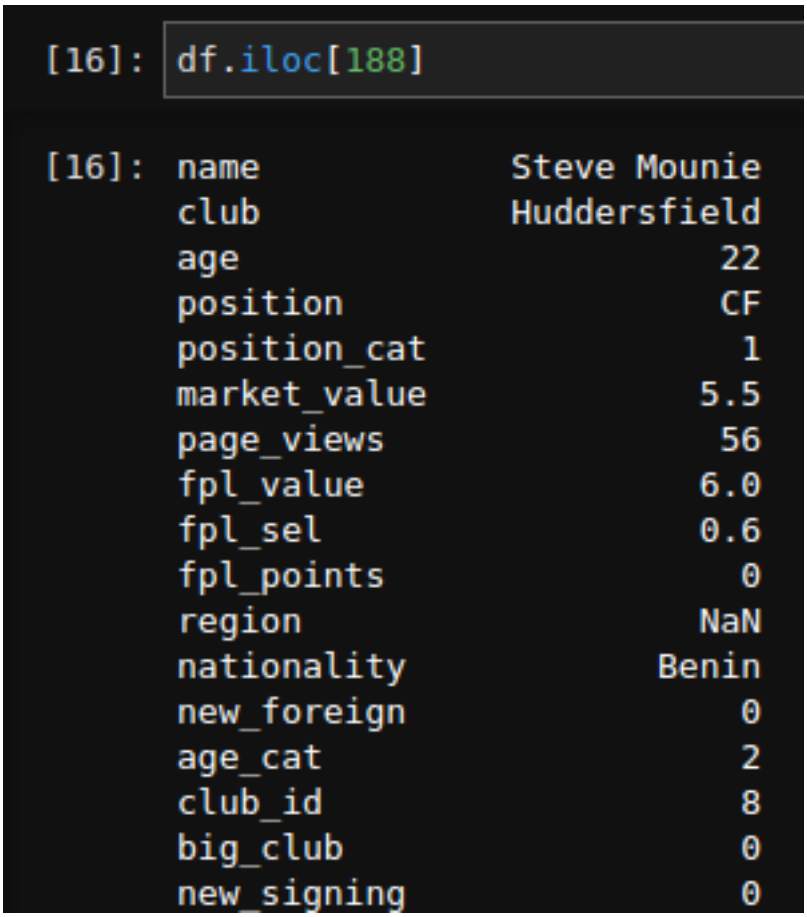
17. **new_signing:** Indica si el jugador es un nuevo fichaje en está temporada 2017/18.

Los tipos de datos de las columnas son de la siguiente forma:

```
[6]: df.info()

---  ------        --------------  -----
 0   name          461 non-null    object
 1   club          461 non-null    object
 2   age           461 non-null    int64
 3   position      461 non-null    object
 4   position_cat  461 non-null    int64
 5   market_value  461 non-null    float64
 6   page_views    461 non-null    int64
 7   fpl_value     461 non-null    float64
 8   fpl_sel       461 non-null    object
 9   fpl_points    461 non-null    int64
 10  region        460 non-null    float64
 11  nationality   461 non-null    object
 12  new_foreign   461 non-null    int64
 13  age_cat       461 non-null    int64
 14  club_id       461 non-null    int64
 15  big_club      461 non-null    int64
 16  new_signing   461 non-null    int64
```

A primera vista podemos observar que en la columna nacionalidad falta un dato, el cual hace referencia a la fila 188, es decir, al jugador Steve Mounie:



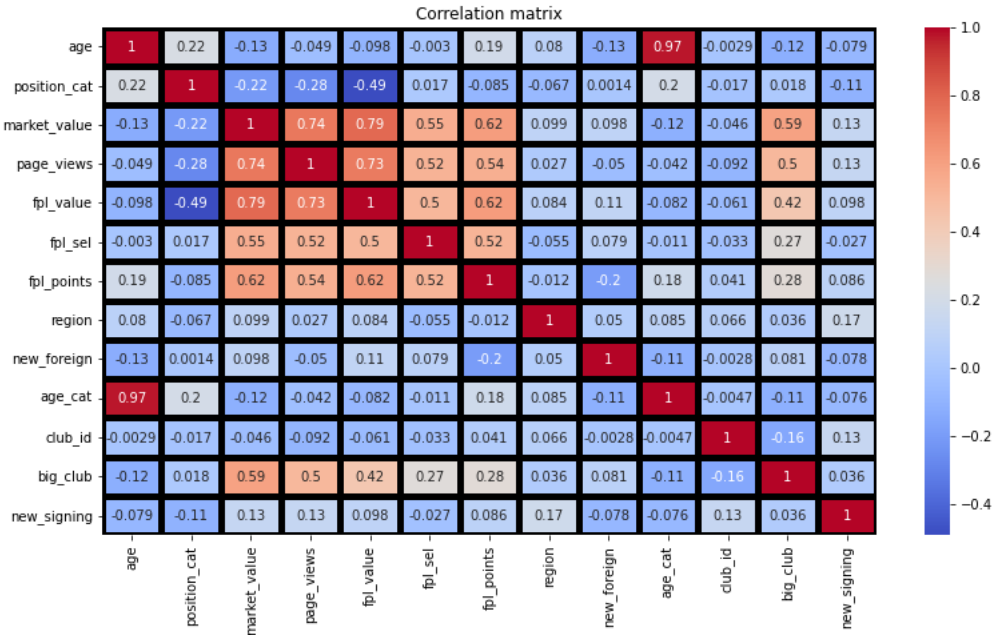
Como solo es un dato, simplemente hice una busqueda en Google y rellené el campo respectivamente.

También sabemos que la columna fpl_sel hace referencia a un porcentage, sin embargo, está de tipo objeto, es decir un string, procedemos a cambiarla a tipo float, primero removemos el signo % y posteriormente hacemos un cambio de tipo a float.

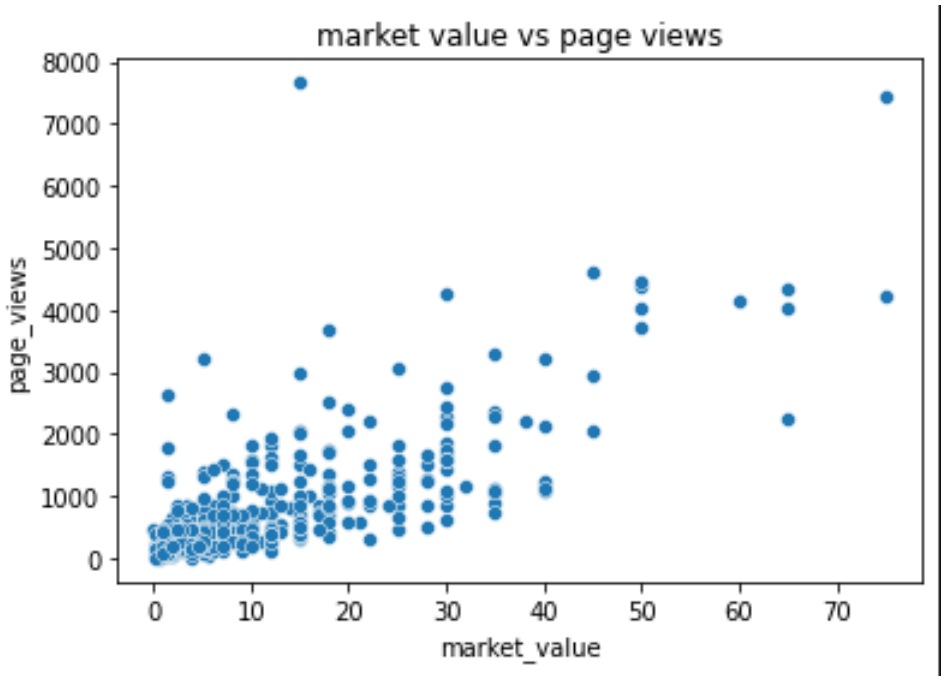
Habiendo realizado esto, ya si verificamos nuevamente la base no tenemos más datos nulos, en esta base de datos no es necesario realizar mucho preprocesamiento de datos.

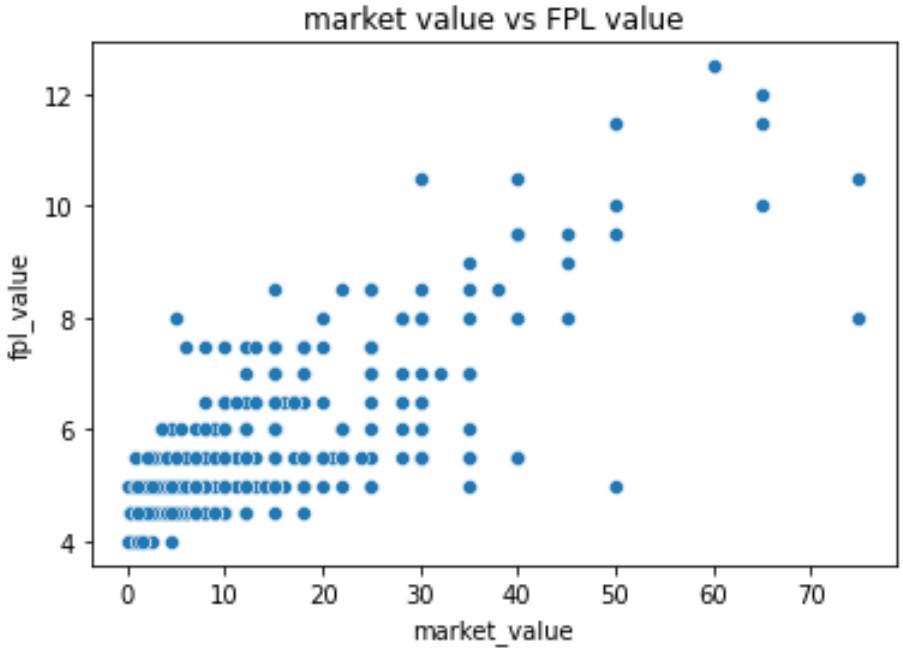
El modelo de machine learning que busco realizar es una regresión, posiblemente lineal, la variable objetivo es el valor del jugador en el mercado , para esto buscaremos correlaciones entre esta variable continua y otras variables, primero

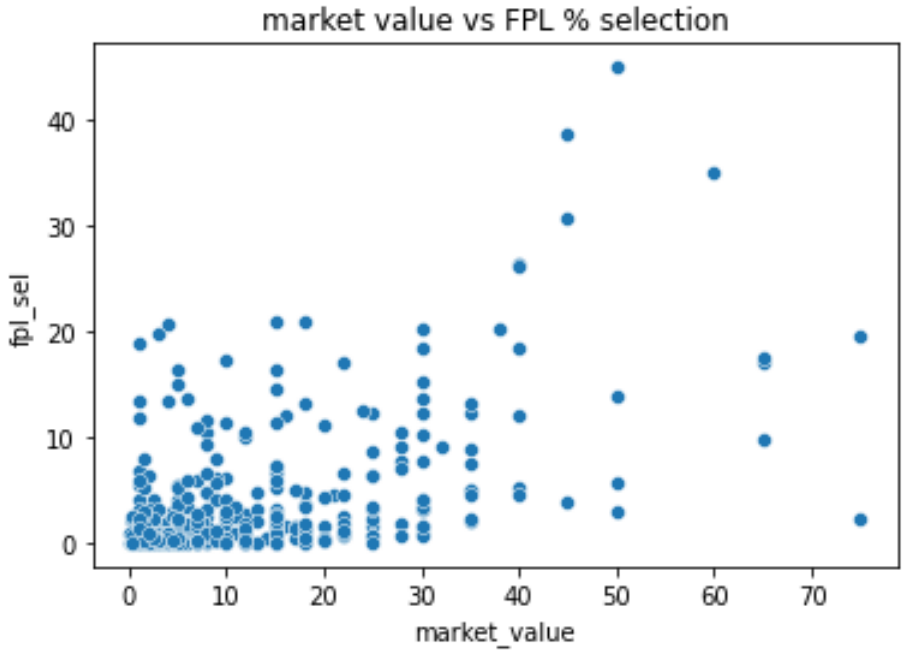
veamos la matriz de correlaciones:

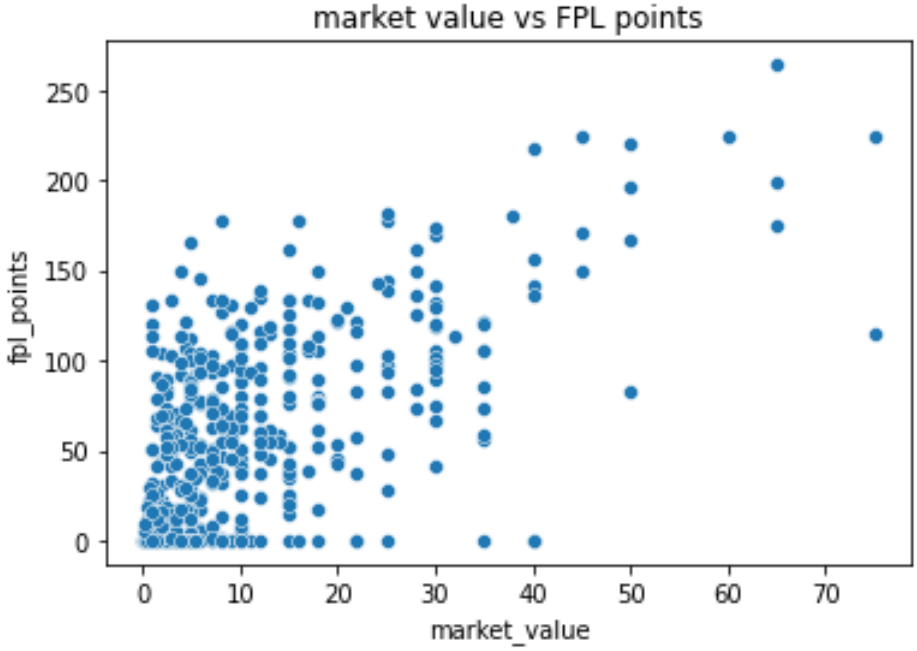


Si nos enfocamos en nuestra variable objetivo market_value podemos ver que tiene fuertes correlaciones con page_views, fpl_value, fpl_sel,fpl_points y big_club, veamos una a una:

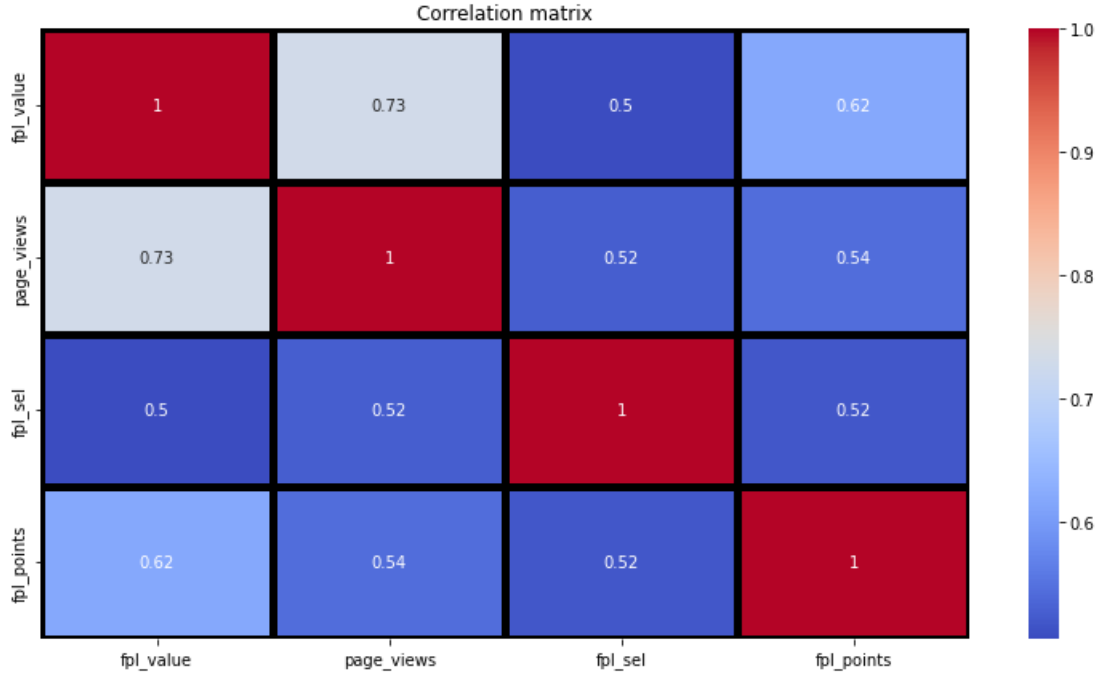

market value vs page views
page_views
market_value
0
1000
2000
3000
4000
5000
6000
7000
8000
0
10
20
30
40
50
60
70


market value vs FPL value
fpl_value
market_value
4
6
8
10
12
0
10
20
30
40
50
60
70


market value vs FPL % selection
fpl_sel
market_value
0
10
20
30
40
50
60
70


market value vs FPL points
fpl_points
market_value
0
50
100
150
200
250
10
20
30
40
50
60
70

Como podemos observar evidentemente hay fuertes corerlaciones entre estas variables y la variable objetivo, sin embargo, también existen correlaciones entre estas variables, veamos esto:



Por lo tanto dejaremos la variable con más correlación y desecharemos las demás, es decir, dejaremos fpl_value con una correlación del 0.79.También dejaremos la variable big_club como ruido, pues yo creo que el ser perteneciente a un equipo del big six está relacionado con el precio en el mercado, finalmente tenemos dos variables como predictoras y una variable a predecir, creo que con una regresión lineal pues hay buena correlación y así obtener buenos resultados y un modelo con muy bajo sobreajuste, sin embargo, si no es así, podemos probar otras técnicas de regresión como decisión tree regresor, regresión polinómica y mirar si las otras características estan relacionadas no linealmente con la variable de salida, también podriamos tratar con support vector machines pero para regresión.