
PROYECTO FIFA ML
EDA
Preparación modelo
Elección del modelo
Conclusión
Introducción
Entrenamiento
THE BRIDGE

Introducción

Propósito

Propósito

Este proyecto forma parte de uno mayor, para poder justificar una parte del mismo nos vemos en la necesidad de realizar un estudio para ver su fiabilidad

El propósito de este estudio es poder determinar porcentaje de acierto o fallo respecto a nuestra variable objetivo. Esta nos dirá la calidad de los equipos y a su vez su futura predicción dentro de la liga en la que se engloban.

Nuestra variable objetivo llamada en nuestro conjunto de datos "Overall", es una variable continua que va de 0 a 100 y nos indica la calidad del jugador.

Al ser un modelo con muchos datos nos hemos centrado en las cinco ligas más potentes y en los últimos cinco años, como ya hicimos en el EDA anterior.

Introducción

Propósito


PROYECTO FIFA ML
EDA
Preparación modelo
Elección del modelo
Conclusión
Introducción
Entrenamiento
THE BRIDGE

EDA

Validación

Correlación

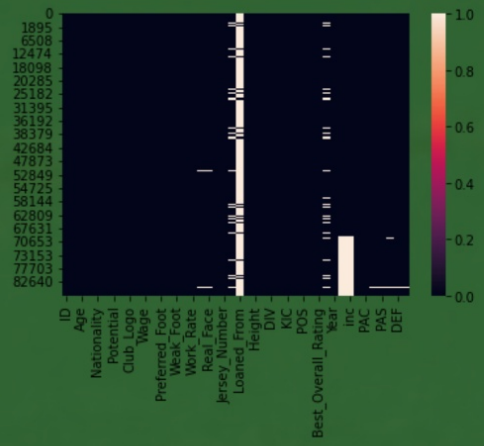
Validación de datos

Revisamos el contenido del dataframe,

Nº de filas	Nº de columnas	Variables numéricas	Variables Objeto	Variables fecha
86792	66	5 int64 / 40 float64	20	1

,
eliminaremos las variables categóricas que no aportan nada al modelo o bien aplicando la lógica de negocio, borraremos, columnas tales como, imagenes, escudos, nombres de equipo, clausulas, precio de mercado y similares. Antes de borrar estas variables tendremos que ver su correlación, lo veremos en el siguiente paso.

Eliminamos nulos sin eliminar el registro. Es necesario para poder aplicar los modelos de machine learning.



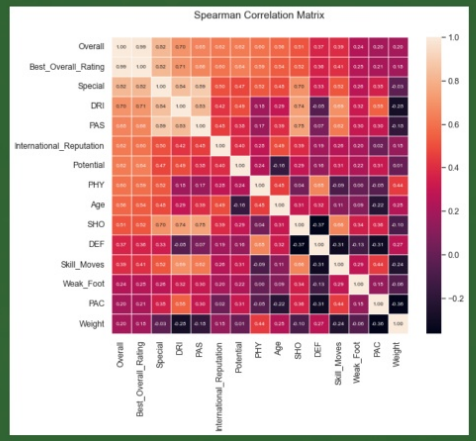
Borradas sin uso
Photo
Club_Logo
Real_Face
Position
Name
Natonality
Club
Body_type
Release_Clause
Joined
Year
Pais
Flag
Club Logo
Jersey Number
Loaned From
Contract Valid Until
Release Clause

EDA

Validación

Correlación

Correlación



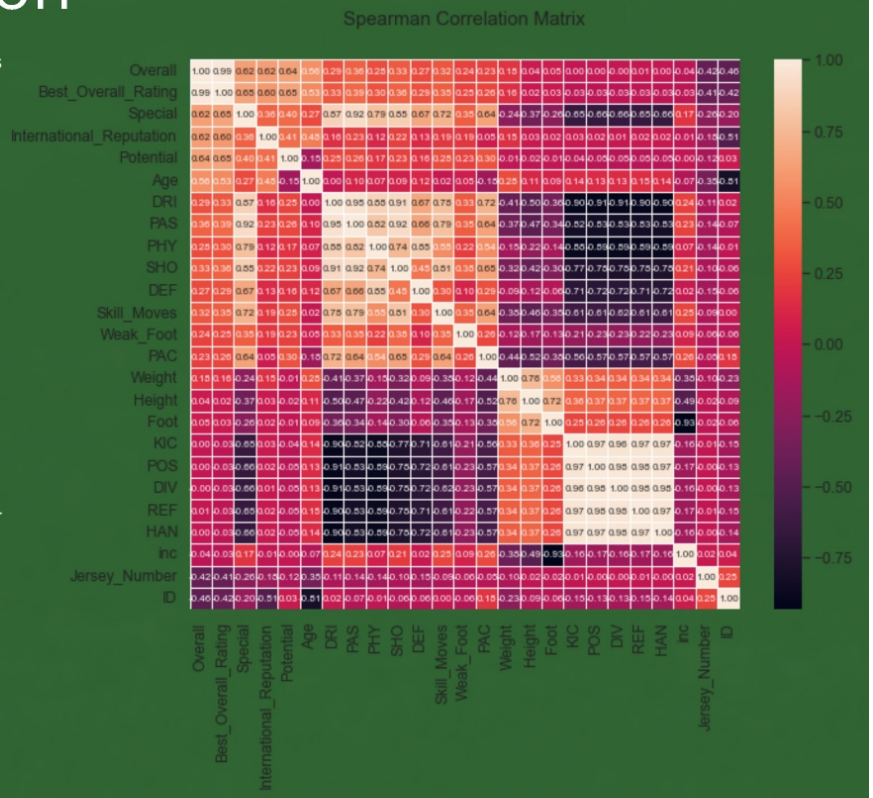
Con el dataframe ya limpio de nulos, nos centramos en revisar las variables que nos quedan y como se relaciona estas entre sí.

Nos vemos en la necesidad de separar los jugadores de campo de los porteros evitando eliminar variables que tengan una correlación alta con la variable objeto. Para los jugadores de campo que es lo que nos ocupa.

Borradas tipo
KIC
POS
DIV
REF
HAN

Eliminamos también del modelo aquellas con una correlación muy baja al igual que las objeto que no aportaban información su contenido.

Borradas
Foot
inc
Height
inc
Jersey_number
ID



EDA

Validación

Correlación


PROYECTO FIFA ML
EDA
Preparación modelo
Elección del modelo
Conclusión
Introducción
Entrenamiento
THE BRIDGE

Preparacion del modelo

Categorización

Escalado

Categorización

Una vez limpio el dataframe, solo nos quedan analizar ciertas variables categóricas que pueden ser necesarias incluir o no en el modelo .

Categóricas
Value
Wage
Prefered_Foot
Work_rate
Loaned_From
Best_Position

Revisando una a una, podemos determinar que el salario , el valor de mercado, el ritmo de trabajo o si es una cesión, no influyen en la variable target. La única de ellas que puede tener influencia es "Prefered_Foot", la cual convertiremos a numérica y borraremos el resto de variables del dataframe.

Preparacion del modelo

Categorización

Escalado

Escalado

Antes de poder entrenar el modelo necesitamos que todas las variables númericas del dataframe tengan la misma escala, para ello escalamos todas salvo la variable target

Head of dataset is:

Age	Potential	Special	International_ Reputation	Weak_Foot	Skill_Moves	Best_Overall_ Rating	DefensiveAw areness	PAC	SHO	PAS	DRI	DEF	PHY	Preferred_Foot_Right	Overall
0.46	0.84	1.00	1.00	0.75	0.50	0.88	0.56	0.64	0.97	0.86	0.80	0.40	0.84	1.0	87
0.35	0.92	0.99	0.75	1.00	0.75	0.96	0.69	0.73	0.94	0.99	0.85	0.48	0.77	1.0	91

Ya tendriamos el modelo listo para ser entrenado.

Preparacion del modelo

Categorización

Escalado


PROYECTO FIFA ML
EDA
Preparación modelo
Elección del modelo
Conclusión
Introducción
Entrenamiento
THE BRIDGE

Entrenamiento

Entrenamiento

Entrenamiento

Entrenamos y dividimos el dataset en conjunto de entrenamiento y de testing utilizando un 80% entrenamiento y un 20% para la prueba.

Entrenamiento

Entrenamiento


PROYECTO FIFA ML
EDA
Preparación modelo
Elección del modelo
Conclusión
Introducción
Entrenamiento
THE BRIDGE

Elección del modelo

Metricas

Modelos

Métricas

Queremos medir el error del modelo respecto a los datos reales, para ello nos hemos basado en el error absoluto medio MAE y nos hemos apoyado como complemento en el error cuadrático medio MSE y en la raíz error cuadrático medio RMSE ,consiguiendo una mayor certeza en los errores que comete el modelo. Para elegir una o varias métricas como criterio veremos sus valores y como se comportan en los distintos modelos.

Sin cross validation

Modelo	MAE	MSE	RMSE
xgboost	0,593	0,718	0,847
rf	0,601	0,804	0,897
catboost	0,605	0,741	0,861
et	0,611	0,814	0,902
lightgbm	0,648	0,821	0,906
dt	0,716	1,507	1,228
gbr	0,814	1,131	1,064
huber	0,899	1,262	1,123
br	0,899	1,250	1,118
lr	0,899	1,250	1,118
ard	0,900	1,252	1,119
omp	1,024	1,527	1,236
par	1,216	2,209	1,486
knn	1,218	2,491	1,578
gnb	1,280	3,223	1,795
ada	1,284	2,404	1,550
KRR	2,183	9,605	3,099
elastic	5,787	50,428	7,101
lasso	6,096	56,031	7,485
dummy	6,167	57,336	7,572
lassolars	6,167	57,336	7,572

Con cross validation

Modelo	MAE	MSE	RMSE
catboost	0,784	0,913	0,834
xgboost	0,786	0,926	0,857
rf	0,792	0,952	0,906
et	0,795	0,948	0,906
lightgbm	0,811	0,943	0,890
dt	0,862	1,118	1,254
gbr	0,907	1,023	1,046
huber	0,949	1,061	1,126
br	0,950	1,060	1,124
lr	0,950	1,060	1,124
ard	0,950	1,060	1,124
omp	1,014	1,114	1,240
knn	1,122	1,273	1,621
par	1,123	1,297	1,417
ada	1,126	1,235	1,523
gnb	1,129	1,333	1,777
KRR	1,493	1,701	2,894
elastic	2,368	2,634	6,937
lasso	2,432	2,704	7,313
dummy	2,446	2,720	7,397
lassolars	2,446	2,720	7,397

Elección del modelo

Metricas

Modelos

Modelos

De cara a elegir el modelo en los dos últimos pasos con parámetros e hyperparametros , hemos realizado una criba y elegido solo los modelos cuyo MAE están por debajo de 1 para aplicar parámetros y por debajo de 0'9 para los hyperparametros.

Modelo	MAE	MSE	RMSE
rf	0,785	0,940	0,883
xgboost	0,786	0,926	0,857
et	0,795	0,949	0,902
lightgbm	0,811	0,943	0,890
catboost	0,867	0,985	0,971
dt	0,867	1,122	1,258
huber	0,949	1,061	1,126
ard	0,950	1,060	1,124
br	0,950	1,060	1,124
lr	0,950	1,060	1,124
gbr	2,445	2,718	7,390

Modelo	MAE	MSE	RMSE
Light Gradient Boosting Machine	0,588	0,734	0,857
Extreme Gradient Boosting	0,632	0,870	0,933
Extra Trees Regressor	0,692	0,968	0,984
Decision Tree Regressor	0,720	1,486	1,219
CatBoost Regressor	0,835	1,256	1,121
Random Forest Regressor	0,835	1,256	1,121

Vemos que el modelo que tiene mejores valores en las tres métricas es Light gradient boosting machine, siendo el modelo elegido para su uso.

Elección del modelo

Metricas

Modelos

PROYECTO FIFA ML

EDA

Preparación modelo

Elección del modelo

Conclusión

Introducción

Entrenamiento

THE BRIDGE

Conclusión

Predicción

Conclusión

Una vez hemos elegido el modelo para la predicción utilizamos el 20 % del dataframe que habiamos guardado al principio y obtenemos el porcentaje de error en valor abosluto MAPE, siendo este porcentaje

0'57 %

Lo cual es un valor más que aceptable para poder validar nuestra parte del proyecto.

¡¡¡¡GRACIAS!!!!

Conclusión

Predicción


PROYECTO FIFA ML
EDA
Preparación modelo
Elección del modelo
Conclusión
Introducción
Entrenamiento
THE BRIDGE