Universidad Francisco Marroquín
Facultad de Ciencias Económicas
Curso: Machine Learning Models
Nombre: Juan Fernando González
Fecha: 10-03-2020

Reporte

Introducción / Data

FIFA 19 Complete Player Dataset

El propósito de este reporte es poder responder a las siguiente pregunta: ¿Qué variable o variables son las que más influyen en el *Overall rating* del jugador en FIFA 19?; así como también el de analizar, procesar y obtener *insights* de los datos provenientes del *dataset* **FIFA 19 Complete Player Dataset**.

Este *dataset* fue adquirido desde la página *kaggle.com,* el cual fue publicado gracias a Karan Gadiya. Los datos que se encuentran en el archivo fueron almacenados y encontrados gracias a la herramienta de *scraping* a la base de datos de la FIFA. El *dataset* contiene datos e información sobre más de dieciocho mil jugadores de diferentes clubes, edades, y diferentes partes del mundo. Cada jugador tiene un club, su edad, rating dados ciertos parámetros, así como también, rating según habilidad, potencial, entro otras.

Algunas de las (*features* / columnas) que podremos encontrar en el *dataset* son las siguientes:

1. *Name:* El nombre del jugador.
2. *Age:* La edad del jugador
3. *Overall: Overall rating* del jugador en base a sus habilidades y cualidades (0 – 99).
4. *Ball Control:* El control del balón del jugador (0 – 99).
5. *Composure:* La composición del jugador (0 – 99).
6. *Club:* Club de fútbol en donde se encuentra el jugador.
7. *Wage:* El sueldo del jugador. (€)
8. *Value:* El valor del jugador en el mercado de FIFA. (€)
9. *Skill moves: Rating* de la habilidad con el balón de los jugadores. (0 – 99)
10. *Position:* La posición del jugador en el campo.

Por otro lado, hay 18,297 filas (*observations*), que son los valores por cada columna en el dataset. En total hay 89 *features* y 18,297 *observations*, además existen 76,948 datos

nulos, es decir, que no tiene un valor, o que no se tiene información sobre ese campo en específico. Más adelante veremos cómo tratamos con estos valores nulos.

Dada la pregunta que quiero resolver en este reporte, la variable que deseo predecir es el *Overall Rating,* por lo que será la variable dependiente. Las variables independientes, las cuales logre descubrir por medio de exploración y métodos estadísticos, fueron: *Age, Composure, Ball Control, Potential, y Reactions.* En la otra mano, no tomaremos en cuenta variables que contienen *links* o features que contengan más del ochenta por ciento de datos nulos.

Algunos de los conceptos de estadística que serán aplicados en este reporte son los siguiente: Histograma (distribución), Correlación, Regresión lineal simple, Regresión lineal múltiple, Regresión Polinomial, r^2, r^2 ajustado, entre otros.

La correlación nos indica la fuerza y dirección de una relación lineal y proporcionalidad entre dos variables. Para que dos variables sean consideradas correlacionadas, el coeficiente de correlación debe de estar cercano a 1 o a -1. Si el coeficiente de correlación es cercano a 0, esto quiere decir que no hay relación entre las 2 variables.
La regresión lineal, la cual fue el método principal usado para predecir nuestra variables dependiente, es un modelo matemático que nos permite aproximar la relación de dependencia entre una variable dependiente *y*, y una variable dependiente *x*, además de un término aleatorio, y luego, poder predecir la variable dependiente en base a la independiente.

Métodos

Luego de explorar el *dataframe*, se comenzará la limpieza de los datos. Existen varios tipos de manipulación de datos, en este trabajo se utilizaron dos, el de **imputación por valores numéricos por valores categóricos** y el de **eliminación de datos nulos**. Se eliminaron columnas que no tenían ningún dato relevante con la exploración y el análisis de los datos, por ejemplo, el ID, la duración del contrato del jugador, de donde vino el jugador, si fue cedido por otro club o no, si es especial o no, entre otras variables que no aportaban nada importante a la predicción del *overall*. Por otro lado, también realizo modificación de los valores de las columnas que son de tipo *string,* para quitarles signos (€, -, _, ,) y remplazar palabras por números, para que luego sean más fáciles de computar y comparar; así como también las abreviaciones de millones (M) y miles (k).

Una vez ya hecha la limpieza y la imputación de nuestros datos, se continuo a aplicar los métodos estadísticos y algoritmos matemáticos para visualización y análisis de los datos.

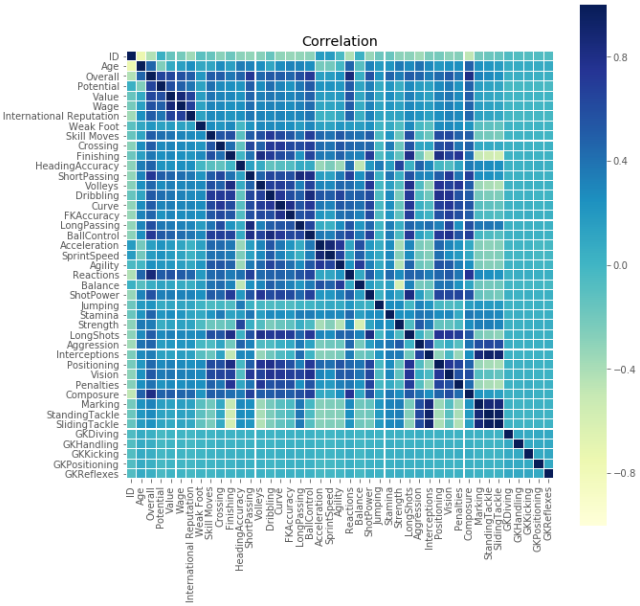
Como primer método, se empleó el **diagrama de correlación**, con el objetivo de poder visualizar y analizar la relación que tienen las variables independientes con la variable dependiente. Luego, se procedió a realizar la **regresión lineal simple** por cada variable independiente, y así poder ver la predicción, que tan bien se relacionaban y explicaban el modelo. Esto lo sabíamos por medio de la métrica de $r^2, r^2 ajustado\ y\ RMSE\ (root\ mean\ square\ error)$. Por otro lado, también se trató de aplicar **funciones de penalización** a ciertos modelos. Estas penalizaciones podrían ayudar al modelo a poder decidir qué variables son significativas para el modelo y cuáles no, así como también, normalizar nuestra predicción y reducir el RMSE. Los métodos de penalización utilizados fueron **Ridge** y **Lasso**. Debido a que una *feature* independiente no se lograba explicar con totalidad en una regresión lineal simple, se utilizó la **regresión polinomial** para poder tener un mejor resultado y porcentaje del r^2 aumentando el grado de complejidad del modelo. Por último, se realizó una **regresión lineal múltiple** agrupando las variables independientes que fueron escogidas, y así tratar de predecir el Overall de los jugadores a partir de ellas.

Resultados

A continuación, se presentarán los resultados obtenidos de los métodos aplicados al *dataset,* así como también el análisis y deducciones del mismo.

Los primeros resultados que obtuve fueron sacados de la información que me proporcionaban el diagrama de correlación:

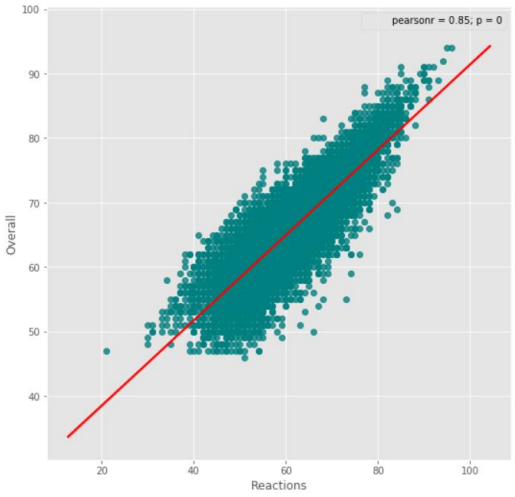
Correlación de la variables del dataset:



Como podemos observar, las variables que están fuertemente correlacionadas con la variable dependiente son: *Composure, Reactions, Age, Potential, Value, Ball Controll y Short passing.* De estas 7 variables, solo escogeremos las mejores o las que tiene mayor fuerza en la relación con el Overall, para así poder tener un mejor modelo y ver cuál de ellas puede determinar mejor por si sola nuestra variable dependiente.

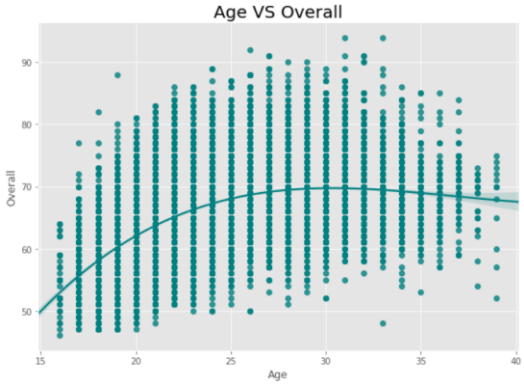
Luego, se procedió a realizar la regresión lineal simple.

Regresión lineal simple: *Overall vs Reaction*



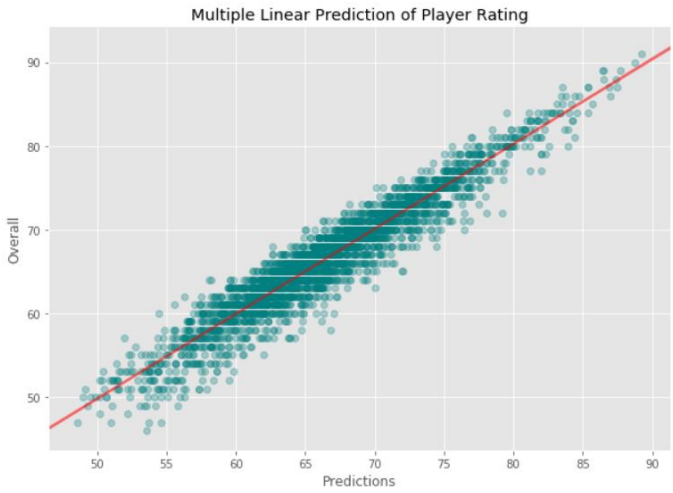
Como se puede observar, existe una relación positiva y fuerte entre la variable independiente *Reaction*, con la variable a predecir, *Overall*. Esto quiere decir que a medida que incrementa el rating de la reacción del jugador, mejor Overall tendrá. Por otro lado, la predicción dio los siguientes valores: $r^2 = 0.72$ y $RMSE$ 3.62; lo que nos indica que la variable si logra explicar de buena manera la variabilidad en los datos, y que la desviación entre los puntos y la recta de predicción es baja.

Regresión Polinomial: *Overall vs Age*



En este caso, primero se aplicó la regresión lineal simple, pero dado a su resultado se aplicó una regresión múltiple para poder mejorar la predicción entre la edad y el overall rating del jugador. Se utilizo el grado 3 para poder explicar o mejorar el modelo, y se obtuvo una mejora casi insignificativa del 0.08 por ciento en r^2.

Regresión lineal múltiple: *Overall vs Features independientes*



Por último, el modelo de regresión lineal múltiple que hace la predicción y la relación entre las 5 variables independientes ya antes mencionadas, y la variable independiente, siendo esta el Overall del jugador. Y tal y como se mira en la gráfica, el modelo múltiple fue el mejor por mucho para poder predecir el Overall rating. Obtuvo el resultado de 0.91 en r^2, 0.91 en $r^2 ajustado$ y 2.0 en *RMSE.*

Conclusiones

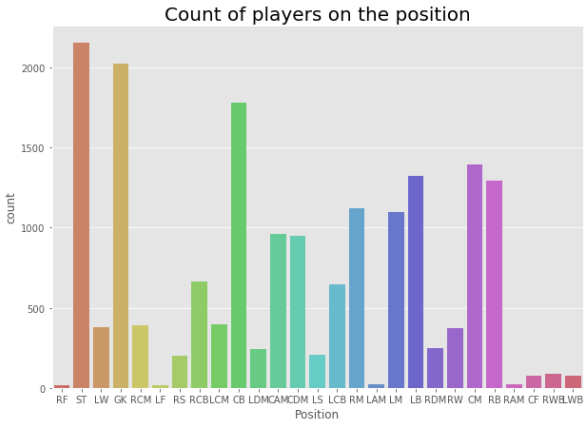
Luego del proceso de exploración y análisis de los datos ya antes mencionados, puedo concluir y responder con certeza a mi pregunta del inicio. Dados los resultados anteriormente mencionados, se puede concluir que el modelo de regresión lineal múltiple es el mejor modelo para poder predecir el Overall rating de un jugador. La edad, la compostura, el control de balón y el potencial, fueron las variables que se utilizaron a lo largo de la exploración, y se pudo demostrar que por si solas, no son suficientes para poder predecir con exactitud y eficacia el Overall rating de un jugador, sin embargo, al unir o concatenarse, pueden expresar y predecir el rating de un jugador con un 91% de efectividad, y con una deviación en los residuos del modelo de 2 puntos, lo cual es bajo. Por otro lado, como recomendación puedo decir que las funciones de penalización son útiles solo en casos especiales, donde existan varias variables que pueden ser útiles o

no para un modelo, como también, el hecho de ver que tanto mejora o empeora nuestro *RMSE.* Y por último, la variable qué mas peso tiene o mas influenza tiene al hacer la predicción del *Overall*, es el potencial del jugador, por lo que si un futbolista quiere tener un buen rating de *Overall*, debe asegurarse de tener un alto *rating* en Potencial.

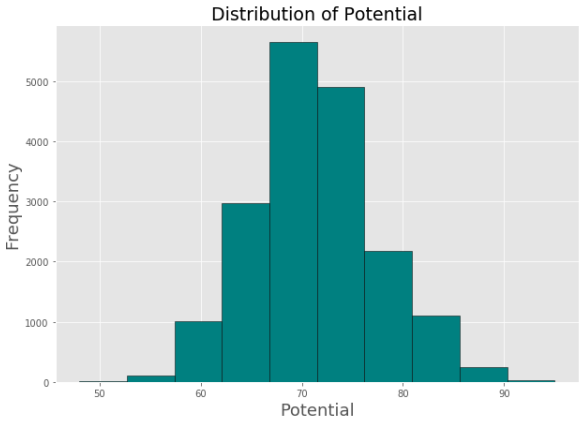
(Seaborn, 2012)
(Statinfer, 2018)
(Stephanie, 2016)

Apéndice

Graficas de distribución del *dataset*:



Most of the players on the dataset are STs (Strikers).



Referencias

Creech, S. (2019). *Statistically SIgnificant* . From https://www.statisticallysignificantconsulting.com/RegressionAnalysis.htm

Seaborn. (2012). *Seaborn*. From https://seaborn.pydata.org/generated/seaborn.lmplot.html

Statinfer. (2018, 01 24). *Statinfer*. From https://statinfer.com/204-1-7-adjusted-r-squared-in-python/

Stephanie. (2016, October 25). *Datasciencecentral*. From https://www.statisticshowto.datasciencecentral.com/rmse/