T05_ Análisis de Regresión Logística

Román Clemente Jurado

Fundamentos de la Ciencia de Datos

ÍNDICE

1. Seleccionar una variable dicotómica o recodificar una variable cuantitativa en una variable categórica de dos categorías para ajustar una regresión logística con los datos del proyecto.

Teniendo en cuenta que una variable dicotómica es aquella variable cualitativa que sólo puede tomar dos valores posibles, por ejemplo, sí/no, hombre/mujer, verdadero/falso, 1/0… Teniendo en cuenta que mi dataset no posee una variable asi para poder categorizar los datos pasaremos a la opción B y recodificaremos una clave cuantitativa para convertirla en categórica para ello:

Recodificaremos ovr para que:

- Convirtiendo el valor en 1 si el OVR ≥ 75, para jugadores sobresalientes
- 0 si el OVR < ese umbral

Por tanto, quedando asi:

- ANTES: Rodri --> ovr :91.0
- DESPUES: Rodri --> ovr: 1

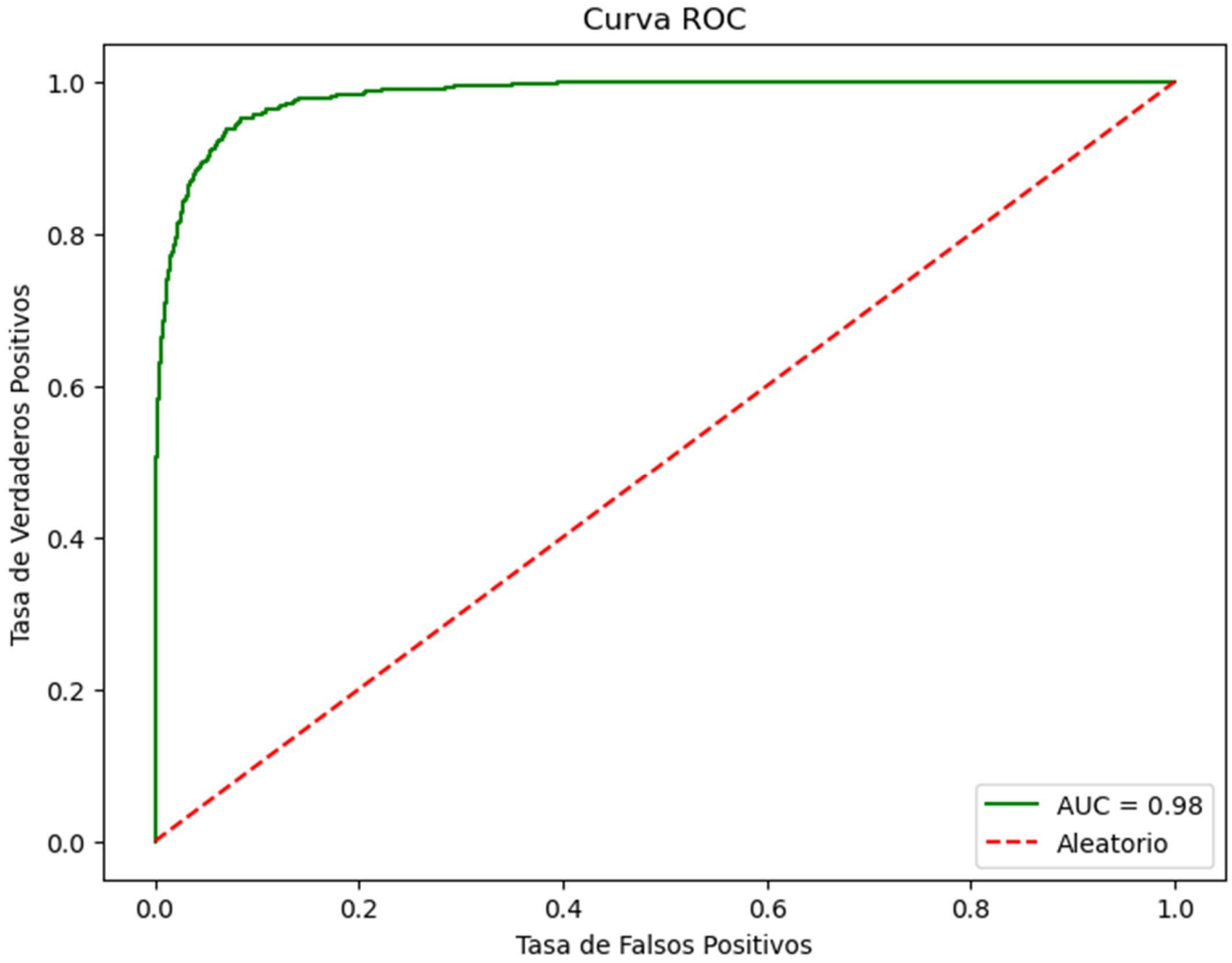
2. Se sugiere utilizar el mismo conjunto de variables explicativas preparado para la regresión lineal múltiple

Como la vez pasada usaremos las variables:

- **Reactions:** Se refieren a atributos que miden la rapidez en reaccionar ante situaciones de juego, además de habilidades de disparo
- **Composure:** Se refieren a diferentes aspectos del control del balón en variadas circunstancias
- **PHY:** Evalúan atributos como defensa y fortaleza física general, incluyendo atributos como fuerza, agresión y salto

- **DRI:** Evalúan atributos como capacidad de pase, control del balón y habilidad para regatear

3. Prestar atención a la interpretación de los resultados



Ahora mediante a la librerías de scikit-learn realizaremos un LogisticRegression para ver como se comportan los datos y que tan capaces somos de catalogar a un jugador promedio de un jugador excepcional y el resultado es la gráfica que se muestra, pero ¿Qué significa esta imagen?

Ahora mismo estamos viendo una gráfica ROC (Receiver Operating Characteristic), la gráfica ROC es una herramienta estadística que se utiliza para evaluar la capacidad discriminativa de una prueba diagnóstica dicotómica, o, en otras palabras, nos sirve para definir que tan buena es nuestra capacidad deductiva para discernir y catalogar

Los Ejes X e Y representan la Tasa de Falsos Positivos o FPR, que mide el porcentaje de observaciones negativas que el modelo clasificó incorrectamente como positivas mientras que la otra

representa la Tasa de Verdaderos Positivos o TPR o sensibilidad, que mide el porcentaje de observaciones positivas correctamente clasificadas como positivas respectivamente

Mientras la curva verde muestra cómo varían las tasas de TPR y FPR a medida que se ajusta el umbral de decisión del modelo, una curva que se acerca al vértice superior izquierdo indica un modelo con buen desempeño, ya que tiene una alta sensibilidad y una baja tasa de falsos positivos, mientras que, la línea diagonal roja representa un modelo aleatorio, que no tiene capacidad predictiva. Esto ocurre si el modelo asigna probabilidades sin información útil, un modelo bien entrenado se alejará de esta recta con una curva hacia arriba

El área debajo de la curva es una métrica que mide el área bajo la curva ROC. Este valor va de 0 a 1:

- 1.0: Modelo perfecto
- 0.5: Modelo sin capacidad predictiva

En este caso, el AUC = 0.98 indica un modelo bueno al acercarse al 1.0

4. Prestar al tratamiento de variables para los modelos

Como ya dijimos usaríamos ROC para interpretar el entrenamiento, pero no sería la única herramienta, para ello también usaremos una matriz de confusión y un Reporte de Clasificación

- Una matriz de confusión es una herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real o sea en términos prácticos nos permite ver qué tipos de aciertos y errores está teniendo nuestro modelo a la hora de pasar por el proceso de aprendizaje
    - Matriz de Confusión:
        - [[4573 82]
        - [124 426]]
    - ¿Qué significa estos datos?
        - Fila 1 (Clase 0.0 - No sobresaliente):
            - 4573 corresponde a jugadores correctamente clasificados como no sobresalientes
            - 82: corresponde a jugadores incorrectamente clasificados como sobresalientes (falsos positivos)
        - Fila 2 (Clase 1.0 - Sobresaliente):

- 124 corresponde a jugadores sobresalientes que fueron clasificados como no sobresalientes (falsos negativos).
- 426 corresponde a jugadores correctamente clasificados como sobresalientes.

- Pero, nuevamente hago la pregunta, ¿Cómo interpreto yo estos datos?
  - Podemos decir que el modelo clasifica muy bien a los jugadores no sobresalientes al ser mayoría, pero tiene algo de dificultad para identificar a los sobresalientes

- El reporte de clasificación evalúa el rendimiento de un algoritmo de clasificación

Métrica	Clase "No sobresaliente"	Clase "Sobresaliente"	Promedios generales
Precisión	0.97	0.84	0.91
Recall	0.98	0.77	0.88
F1-Score	0.98	0.81	0.89
Accuracy	-	-	0.96

  - Precisión:
    - Clase "No sobresaliente" indica que todos los jugadores clasificados como no sobresalientes, el 97% realmente lo son
    - Clase "Sobresaliente" nos dice que todos los jugadores clasificados como sobresalientes, el 84% realmente lo son
  - Recall
    - Clase "No sobresaliente" El modelo identifica correctamente el 98% de los jugadores no sobresalientes
    - Clase "Sobresaliente" El modelo identifica correctamente el 77% de los jugadores sobresalientes
  - F1-Score:
    - Mide el balance entre precisión y recall. Es más bajo para la clase sobresaliente porque la identificación de esta clase no es tan precisa, probablemente se deba a los pocos jugadores excepcionales que realmente hay y que además también afecta los outliers y como los hemos trabajado para regularlo
  - Accuracy:
    - El modelo clasifica correctamente al 96% de los jugadores. Este es un buen indicador del rendimiento general

Variable	Coeficiente	Odds Ratio
reactions	0.396398	1.486461
composure	0.030975	1.031460
phy	0.100509	1.105733

dri	0.172147	1.187852

Los coeficientes del modelo indican el impacto de cada variable explicativa en la probabilidad de que un jugador sea sobresaliente, por tanto, si el valor de Odds Ratio es mayor a 1, el aumento en esa variable incrementa la probabilidad de que el jugador sea sobresaliente

Como ya analizamos en su momento las variables más influyentes en la probabilidad de que un jugador sea sobresaliente son:

1. Reactions, 48.6% por unidad
2. Dri, 18.8% por unidad

Composure tiene un impacto menor, pero aún positivo.

Lenguaje de Negocio

1. Desempeño del Modelo:
    - El modelo tiene un buen desempeño general con un 96% de precisión. Sin embargo, podría mejorar en la identificación de jugadores sobresalientes, ya que actualmente alcanza un recall del 77% para esta clase
2. Factores Clave:
    - Las habilidades técnicas, reactions, dri y phy, tienen un impacto significativo en la probabilidad de que un jugador sea sobresaliente. Por ejemplo, mejorar en reacciones puede aumentar considerablemente la probabilidad de que un jugador sea sobresaliente

Fomentar el desarrollo de las habilidades más influyentes en los jugadores para mejorar su rendimiento general y aumentar sus probabilidades de destacar ademas analizar técnicas de balance de datos, como sobremuestreo de la clase minoritaria, para mejorar la identificación de jugadores sobresalientes

UAX Universidad Alfonso X el Sabio

GRACIAS

UAX.COM