Trabajo Práctico Aprendizaje Automático
Market Value Predictor

Descripción del Modelo

- ### Contexto

  En la industria del fútbol, a lo largo de su historia, la forma de evaluar a un jugador se basaba (y todavía lo hace) en el análisis cualitativo de lo observado por distintos ojeadores de un club que acudían a los estadios en busca de talentos para sus equipos, también los técnicos en base a experiencias pasadas en la línea de cal ponían a disposición su visión particular del fútbol para ver los futbolistas a contratar.
  Hoy en día con el avance de la tecnología se puso a disposición de quien lo quiera utilizar datos de los partidos encapsulados principalmente en dos tipos:
  - Event Data: Datos de eventos particulares de los partidos.
  - Tracking Data: Datos que siguen el movimiento de los jugadores a lo largo del encuentro.

  A partir de estos datos se empezaron a calcular métricas conocidas en el ambiente como e estadísticas avanzadas.
  Entonces la búsqueda de jugadores alrededor del mundo se amplió aún más con estas herramientas para elegir de una manera más analítica y con otro sustento los jugadores que un club quiere incorporar con el fin de reducir el margen de error.
  El fútbol, a su vez, no deja de ser un negocio, en el cual los clubes asignan determinados presupuesto a la incorporación de jugadores.
  En base a este contexto surge la necesidad de ,una vez seleccionados los futbolistas interesados, planificar económicamente el accionar del club en el mercado. De ahí nace nuestro interés en hacer un modelo que permita saber entre que rangos monetarios estaríamos dispuestos a pagar (basado en su rendimiento) como institución deportiva por un jugador.

- ### Objetivo

  Consiste en estimar el valor de mercado de un futbolista que un club podría pagar por él en una transferencia basado en datos de su rendimiento en la temporada anterior.

- ### Datos

  - Fbref: Estadísticas avanzadas de un jugador a lo largo de una temporada.
  - Transfermarkt: Datos de transferencias reales históricas que utilizan su precio como target para predecir el valor de mercado actual
  - Football News Articles: Datos de noticias futbolísticas de diferentes medios.

- ### Métrica

  Elegimos la métrica R2 con el objetivo de responder a la siguiente pregunta:
  ¿Por qué este ju gador vale X dinero y otro vale Y? Tratar de explicar esa variabilidad entre un jugador y otro. ¿Cuánto de la variedad de los valores de mercado observados se dan por las estadísticas avanzadas? Recordemos que buscamos asignar presupuestos para la compra de jugadores y el mercado es muy variado.

Estructura del modelo

Market Value Predictor utiliza una serie de etapas obtención de datos y de procesamiento que concluye con un modelo final de regresión que predice el valor de mercado de un jugador. A continuación se detalla la estructura completa del flujo de datos.

- ## Obtención de datos

  - A partir de la página de datos llamada "Fbref" elegimos algunas de las ligas disponibles (en nuestro caso las principales ligas de Europa) las estadísticas avanzadas de todos los jugadores de la liga a partir de la temporada 2020/2021 en adelante (2024/2025
    inclusive).
  - A partir de los nombres de los jugadores y su edad conseguimos el id correspondiente al futbolista en la página de "Transfermarkt".
  - De "Transfermarkt" obtenemos el historial de transferencias de cada jugador.
  - "Football News Articles" es obtenido de un dataset precompilado de Kaggle, el mismo junta articulos de medios como Tribuna, Skysports y Allfootball entre otros.
  - Filtramos el total de las transferencias en las temporadas para las cuales tenemos los datos estadísticos.
  - Nos quedamos con los jugadores transferidos en ese período de tiempo y a cada transferencia la unimos con las estadísticas correspondientes a la temporada previa obteniendo así el dataset con el que vamos a trabajar.
  - Además guardamos los datos de la última temporada para calcular en producción el valor actual de cada jugador.

- ## Preprocesamiento

  - **Selección de columnas útiles:** En base a nuestro entendimiento del juego elegimos las estadísticas que mejor se adecúan a la hora de evaluar el rendimiento, eliminando las redundantes.
  - **Eliminación de valores nulos:** Se descartan las filas incompletas con alto grado de faltantes en las estadísticas ya que se considera que no hay suficiente información del jugador.
  - **Reducción de dimensiones:** Se agrupan estadísticas altamente correlacionadas (correlación > 0.95). A cada grupo se le aplica el algoritmo de PCA para reducir las dimensiones.
  - **Codificación de variables categóricas:** Utilizamos distintos tipos de codificación como MultiLabelBinarizer para los jugadores que comparten varias posiciones o OneHotEncoder para las nacionalidades y las ligas donde juegan.
  - **Limpieza de Texto:** Normalizamos el texto para evitar problemas por mayúsculas, eliminamos stopwords.
  - **Tokenización y Lematización:** Dividimos el texto del contenido de las noticias en palabras o subpalabras para su análisis, las reducimos a su forma base.
  - **Extracción de Entidades:** Utilizamos NER para identificar jugadores de football mencionados en las noticias.
  - **Detección de contexto y análisis de sentimientos:** Analizamos los temas importantes mencionados como transferencias, lesiones, desempeño y analizamos llos sentimientos de las noticias para saber si afectaban positiva o negativamente al jugador.

  Todas estas transformaciones son persistidas para su reutilización en producción

- ## Modelo de Machine Learning Automatizado

  Utilizamos Pycaret para seleccionar automáticamente el mejor modelo que se ajuste a la

métrica elegida haciendo validación cruzada con 5 folds (k-fold)
Se guarda al final el modelo elegido para utilizarlo en producción.

- **Salida**

  El valor estimado (fee) en euros.

Gestión de Riesgos

- ## Obtención de datos

  El hecho de asociar los datos de distintas páginas genera riesgos ya que lo hicimos con el nombre y la edad del futbolista. Al haber (pocos casos en comparación al promedio) jugadores con el mismo nombre y edad hace que se pierda la información total de un jugador. Si este jugador resultó transferido durante el período en análisis se pierde una transferencia que sumaría al dataset.

  También existe la posibilidad de que existan jugadores distintos con el mismo id, ya que la búsqueda en Transfermarkt de un jugador lo asocie a otro. En ese caso, nos priva de utilizarlo en las transferencias, pero si nos permite calcular su valor de mercado en base a sus estadísticas.

  Existen casos en el mundo del fútbol en el que se dan transferencias a mitad de temporada por lo que las estadísticas que se reflejan en Fbref son las de la primera parte de la misma. Entonces, ahí nos estamos perdiendo información valiosa del resto de su temporada por cuestiones de como se obtienen los datos de la página.

  El modelo se basa en observar las estadísticas de la última temporada por lo que existen casos de jugadores que normalmente tienen un valor de mercado alto pero pasaron la mayor parte de la misma lesionado. Este es un modelo que no es tolerante a las lesiones de larga duración. Podríamos gestionarlo atrasando estadísticas a dos temporadas atrás pero dado que el rango de tiempo disponible en Fbref es muy chico tendríamos muy pocas transferencias

- ## Procesamiento de datos

  La variable categórica del equipo en el que juega el jugador hace que se amplíe mucho las dimensiones dado a la gran cantidad de equipos involucrados. Esto se gestionó utilizando la competición donde radica el equipo.

  Además la importancia del equipo en el modelo es demasiado alta en comparación al resto de las estadísticas y lo que estamos valuando es como varía el precio de mercado en base a su rendimiento.

Trabajo Futuro

El historial de lesiones es fundamental para evaluar el rendimiento en una temporada por lo que añadirlo sería un aliciente que mejore el modelo.

Actualmente estamos trabajando con jugadores de campo, ya que el arquero tiene otras estadísticas avanzadas que influyen en su rendimiento. Añadir al modelo a los arqueros profundizaría mas en los jugadores a acceder.

Extender a mas temporadas previas las estadísticas para obtener un mejor contexto.

Hacer otro modelo que se explique en otro tipo de variables distintas a las estadísticas, datos mas cualitativos y de contexto, que expliquen la otra parte del valor de mercado y cuanto influye.

Extender el modelo de NLP para identificar también personas que no sean jugadores pero también pueden llegar a afectar el precio del mismo, como puede llegar a ser el entrenador de un equipo en algunos casos.

Integrantes

- Luciano Costa, 102104 luccosta@fi.uba.ar
- Weng Xu Marcos Tomás, 109153 mweng@fi.uba.ar
- Theo Jorge, 107862 tjorge@fi.uba.ar
- Nicolas Tonizzo 107820 ntonizzo@fi.uba.ar