Estimación y Modelado del Valor de Mercado de Jugadores NBA a partir de Datos Públicos: Un Enfoque Analítico y Predictivo Multidimensional

Trabajo Fin de Máster Big Data y Business Analytics 2024/2025

Pedro Jesús Lorente Molina

18 de septiembre de 2025

Índice

1 Resumen ejecutivo

Este trabajo estima el **valor empresarial de los jugadores de la NBA** a corto plazo (t→t+1) y a **multi-horizonte** (t→t+n), integrando datos deportivos, señales de popularidad y contexto de mercado. La entrega final es un **artefacto aplicable a negocio**: una mini-aplicación Shiny que transforma puntuaciones (*scorings*) en **KPIs ejecutivos**, **rankings** y **lecturas de riesgo** listas para decidir: renovar, priorizar patrocinios, reasignar presupuesto o planificar inversiones.

La propuesta aporta tres piezas diferenciales: (1) una **predicción monetizada** por jugador; (2) una **cuantificación del riesgo** mediante intervalos *P10–P90* y métricas de *cobertura*; y (3) un **indicador de retorno** simple y accionable, ΔUSD = Valor previsto − Salario previo. El enfoque evita la fuga temporal, utiliza validación con **backtesting** por horizonte e introduce métricas útiles para negocio (eficiencia valor/coste, concentración tipo Pareto, estabilidad de rankings).

En primer lugar, definimos las magnitudes empleadas. El **índice de Gini** (entre 0 y 1) mide la desigualdad de la distribución del valor: 0 implica reparto uniforme y 1, concentración máxima; y **56 %** indica qué parte del valor total acumula el 20 % superior de jugadores. Con estas métricas observamos una **concentración de valor** elevada (Top-20 % = 56 %, Gini = 0.54), consistente con una liga de "superestrellas" y que **justifica estrategias de foco** (invertir selectivamente en la parte alta y optimizar coste en rotaciones). Estas métricas se sintetizan en el **panel de KPIs** de la Fig. 1, que actúa como cuadro de mando *general* y es **ajustable con los filtros de la aplicación** (equipo, posición, temporada y horizonte h) para estudios específicos.

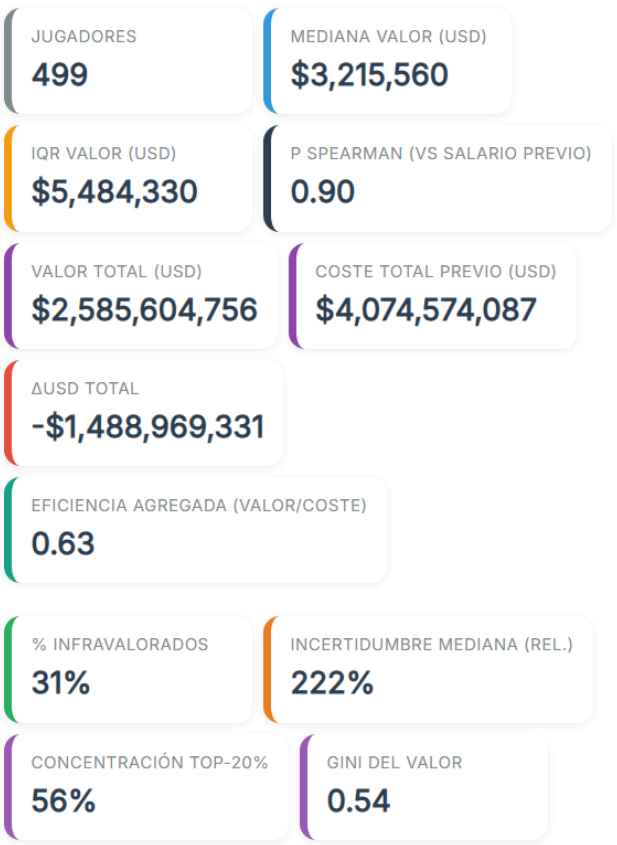


Figura 1: Panel de *KPIs* de la mini-app con filtros activos (año, equipo, posición).

Para la predicción usamos señales complementarias. **MAE** (Error Absoluto Medio, en USD) cuantifica el desvío medio en términos monetarios; **MAPE** (Error Absoluto Porcentual Medio) expresa ese desvío en porcentaje, útil para comparar entre rangos, aunque sensible a denominadores pequeños; y la **Cobertura** es la proporción de observaciones reales que caen dentro del intervalo de predicción anunciado. A horizonte $h = 1$ obtenemos MAE = 2,861,558 $, MAPE = 95 % y Cobertura = 61.85 %; a $h = 3$, MAE = 4,975,205 $, MAPE = 222.46 % y Cobertura = 54.55 %. El **aumento del error y la**

ligera caída de cobertura con el horizonte es esperable por la mayor incertidumbre temporal; por ello, para la toma de decisiones conviene priorizar **señales robustas** (p. ej., el *signo* de ΔUSD) frente a porcentajes exactos.

Definimos $\Delta\text{USD} = \text{valor estimado} - \text{coste previo}$: un **31 %** de jugadores presenta $\Delta\text{USD} > 0$, es decir, candidatos potencialmente infrapagados donde focalizar scouting, minutos o negociación.

En agregado, el **valor estimado total** (2,585,604,756 $) frente al **coste previo** (4,074,574,087 $) determina la **eficiencia media de la liga** $0{,}63 = \frac{\text{Valor}}{\text{Coste}}$. Con $0.63 < 1$, cada $1 de coste compra menos de $1 de valor estimado, lo que permite detectar **desviaciones por equipo** y orientar decisiones de tope salarial, traspasos y renovaciones hacia configuraciones con mejor relación valor/coste.

La mini-app cierra el ciclo: permite cargar nuevas predicciones y obtener, en segundos, la misma lectura estandarizada para apoyar decisiones con rigor y rapidez.

2 Introducción y objetivos

2.1 Contexto de negocio

El valor que aporta un jugador trasciende el rendimiento en pista: determina atractivo para patrocinadores, exposición mediática y capacidad de generar ingresos. Este proyecto responde a una necesidad recurrente en la industria: *cuantificar* de forma coherente y defendible el **valor económico esperado** de cada jugador y **hacerlo operativo** en decisiones cotidianas (renovaciones, activaciones de marketing, reparto de presupuesto entre equipos).

2.2 Objetivos del proyecto

- **Medir** el valor empresarial en USD por jugador, con **incertidumbre explícita**.
- **Evitar** la fuga temporal y validar con **backtesting** multi-horizonte.
- **Traducir** resultados a métricas comprensibles por negocio: ΔUSD, eficiencia, concentración, estabilidad de rankings.
- **Entregar** un artefacto productivizable: mini-app Shiny que reciba nuevos datos y devuelva predicciones e insights de inmediato.

2.3 Preguntas que la herramienta ayuda a responder

- ¿Quiénes son los **Top-N** en valor y qué **incertidumbre** rodea su estimación?
- ¿Dónde hay **oportunidades** (valor > coste) y dónde **riesgo de sobrepago**?
- ¿Qué **equipos** asignan mejor presupuesto según **eficiencia** valor/coste?
- ¿Cómo cambia el panorama a **horizontes** mayores (estabilidad de rankings, crecimiento del riesgo)?

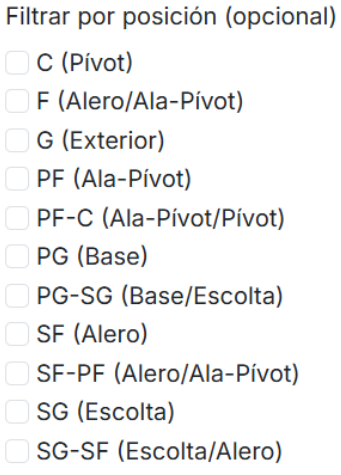
3 Datos y construcción del dataset maestro

3.1 Fuentes y scripts

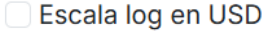
Se integran tres fuentes públicas totalmente accesibles desde Kaggle: *estadísticas generales NBA* (hasta 2023), *salarios* (2000–2025) y *market size* por equipo (2022). Se realizan dos scripts propios (Wikipedia & Google Trends, 2015–2023) configurables por fechas para poder obtener métricas de **popularidad** a partir de la extracción de datos sobre las búsquedas en Google y las visitas en Wikipedia de los diferentes jugadores. Todo se consolida en una llave jugador–año.

3.2 Resolución de identidades, equipos y posiciones

(a) Filtro por **equipo**.



(b) Filtro por **posición**.



(c) Selección de **escala log** para lectura de distribuciones.

Figura 2: Controles de filtro con normalización interna y lectura en escala log (tres vistas).

Los orígenes difieren en nombres de columnas y formatos. Se normaliza `player_id`, se genera una clave de nombre sólida y se armonizan **equipos** mediante diccionario abrev./nombre completo, devolviendo etiquetas claras (`ATL (Atlanta Hawks)`; véase Fig. 2 a). El caso sensible son las **posiciones**: la app implementa una **normalización a siglas canónicas** (`PG, SG, SF, PF, C` y combos como `SG-SF`) desde descripciones extensas y multilingües, permitiendo comparar segmentos y leer KPIs por rol (véase Fig. 2 b).

3.3 Señales y variable objetivo

La **variable objetivo** es el valor empresarial en USD para t→t+1. No se observa de forma directa; se infiere combinando *features* deportivas (uso, eficiencia, disponibilidad), **popularidad** (pageviews, búsquedas) y **contexto** (tamaño de mercado). Se retiene el **salario previo** como aproximación del coste, y se deriva ΔUSD, una métrica puente con **ROI**.

3.4 Calidad y consistencia temporal

Se controla cobertura por temporada, duplicados y rangos válidos. Las señales de popularidad se **desfasan** para evitar fugas; la app permite leer distribuciones en **escala log** para manejar colas pesadas (véase Fig. 2 c). El maestro final se genera con un pipeline reproducible, documentado en el cuaderno de código (Anexos).

4 Exploración y transformaciones clave

4.1 Distribuciones y robustez

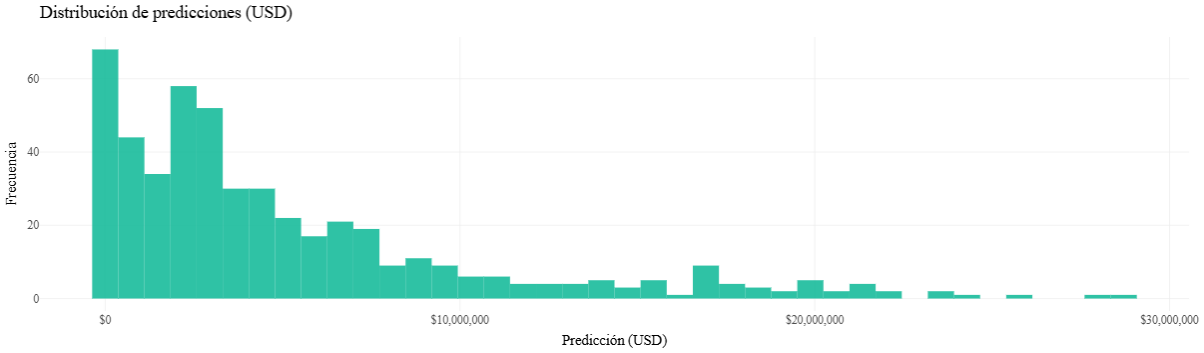


Figura 3: Distribución de *pred_usd* con opción de escala log para evitar colas pesadas.

La **Fig. 3** ofrece una lectura clara de la escala y la dispersión del *valor previsto* en USD. La asimetría con cola a la derecha sugiere concentración de valor en un subconjunto reducido de jugadores; por ello, para resumir el conjunto conviene apoyarse en la **mediana** y el **rango intercuartílico**, menos sensibles a casos extremos. Al activar la *escala log* se mejora la visibilidad de diferencias en los tramos medios y altos, evitando que unos pocos valores muy grandes distorsionen la percepción general. Aplicando filtros por **equipo**, **posición** o **año** puede observarse cómo la forma de la distribución se desplaza o ensancha, lo que ayuda a detectar segmentos con mayor potencial de retorno o, por el contrario, con mayor heterogeneidad que requiera análisis individual.

4.2 Segmentación por posición

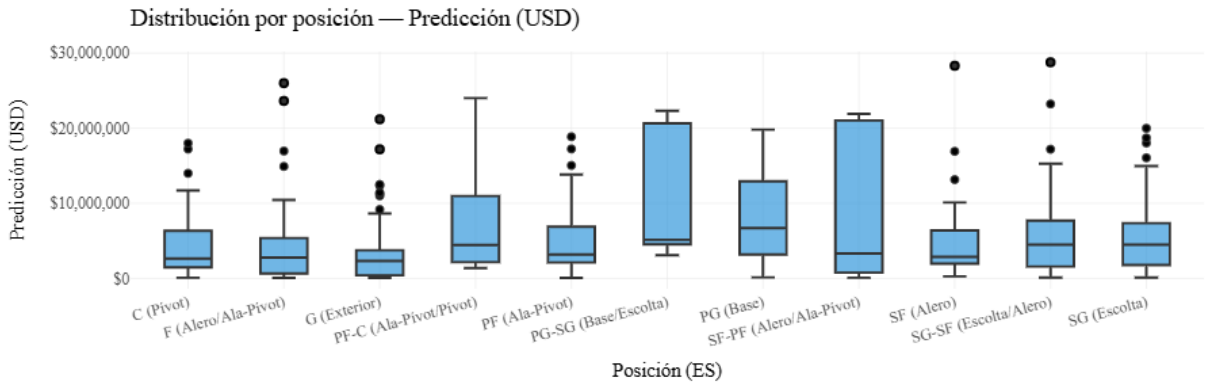


Figura 4: Distribución por **posición** (boxplots).

La **Fig. 4** compara de un vistazo los niveles de *valor previsto* entre posiciones. Cuando la mediana de un rol se sitúa claramente por encima del resto, puede hablarse de una "prima de

rol" que justifica ajustar objetivos de captación o referencias de negociación. La altura de las cajas y la longitud de los bigotes refleja la **dispersión** interna del rol: si es elevada, conviene segmentar (titulares, rotación, perfiles de uso) antes de tomar decisiones generales. Si aparecen colas pesadas, la lectura en *escala log* ayuda a distinguir mejor posiciones vecinas (por ejemplo, escoltas frente a aleros). Repetir esta comparación con filtros de **equipo** o **temporada** permite identificar contextos en los que un rol se potencia o se devalúa de forma sistemática.

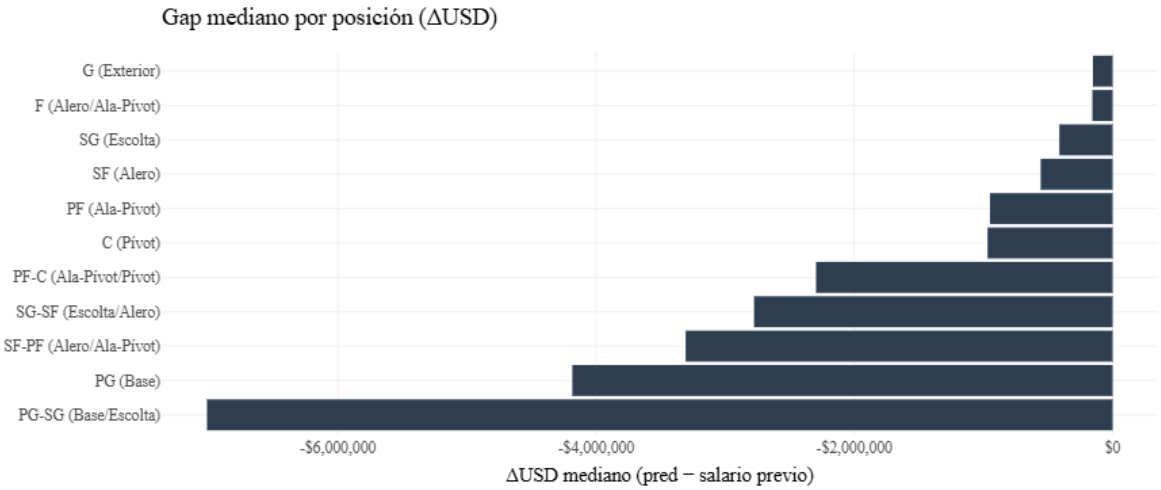


Figura 5: Gap mediano ΔUSD por **posición**.

La **Fig. 5** traduce esa comparación en un indicador accionable al mostrar el ΔUSD *mediano* por posición, esto es, la diferencia típica entre el valor estimado y el coste observado. Un signo positivo sugiere que, en promedio, el rol **aporta más de lo que cuesta** y, por tanto, es candidato a priorización en renovaciones, extensiones o fichajes; un signo negativo alerta de **riesgo de sobrepago** y aconseja negociar o reasignar presupuesto hacia posiciones más eficientes. A priori, si tomamos el conjunto completo de jugadores en activo, la lectura agregada suele reflejar **coste** $>$ **valor** (con ΔUSD global en negativo); sin embargo, esa visión esconde heterogeneidad útil. Al activar los filtros y pormenorizar, la propia gráfica revela **posiciones, equipos o jugadores** con $\Delta\text{USD} > 0$, esto es **valor** $>$ **coste**, sobre los que tiene sentido concentrar recursos. Al repetir esta lectura con filtros por **equipo** y **año**, obtenemos referencias claras para la planificación deportiva y comercial, que se traducen en listas de candidatos priorizados con un equilibrio razonable entre impacto esperado y coste.

5 Estrategia de modelado

5.1 Principios

Sin fuga temporal: las variables en t predicen $t+1$. **Horizonte explícito**: se estima cada h directamente (evita propagación de error). **Explicabilidad**: se retienen baselines interpretables como referencia y se emplean modelos robustos para capturar no linealidades sin sacrificar lectura de negocio.

5.2 Media e incertidumbre

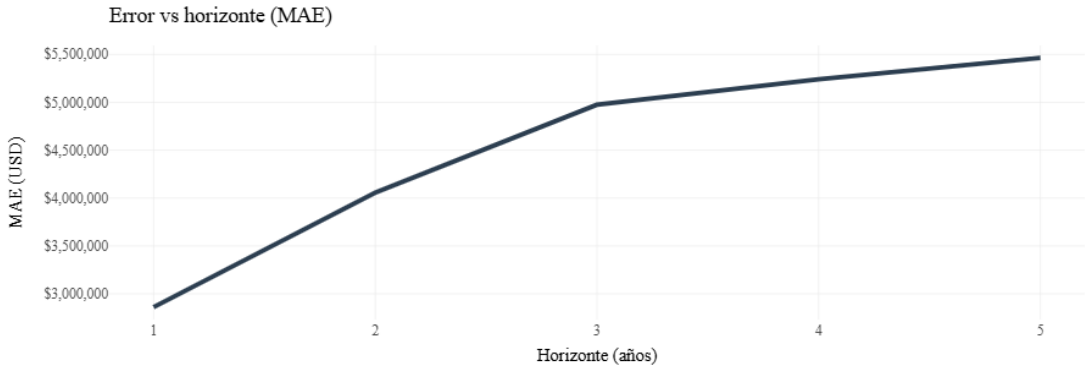
El modelo devuelve **media** (valor esperado) y un intervalo **P10–P90** por jugador. Este abanico convierte la predicción en **escenarios**: negocio discute *rango de resultados plausibles* en vez de un número único.

5.3 Variables y regularización

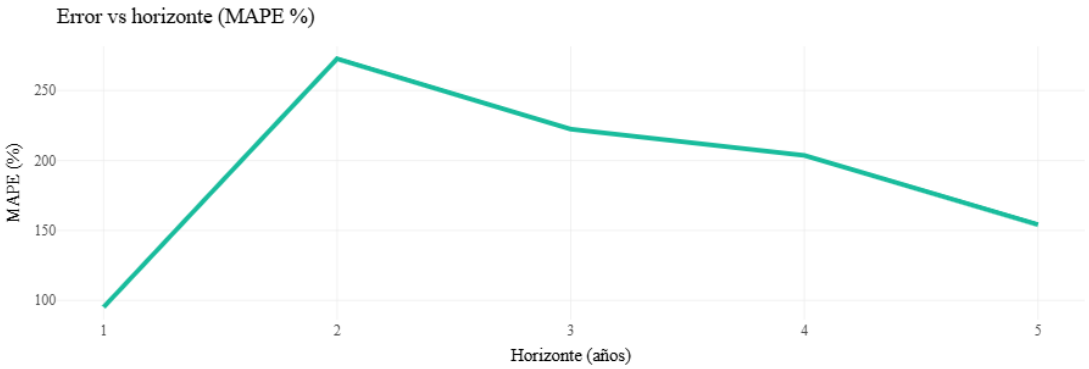
Se usan combinaciones razonadas de *features* deportivas, popularidad y contexto, con **regularización** y **validación temporal**. Se cuida el equilibrio *señal/ruido*, evitando usar variables correlacionadas con el futuro (por ejemplo, métricas generadas después del punto de corte).

6 Validación y backtesting multi-horizonte

6.1 Métricas técnicas y de negocio



(a) MAE (USD) por horizonte.



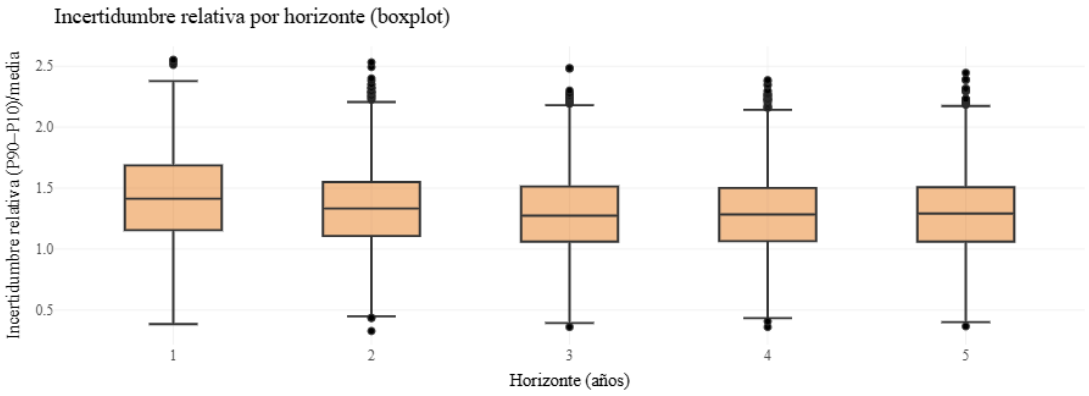
(b) MAPE (%) por horizonte.

Figura 6: Evolución del error por horizonte (dos vistas).

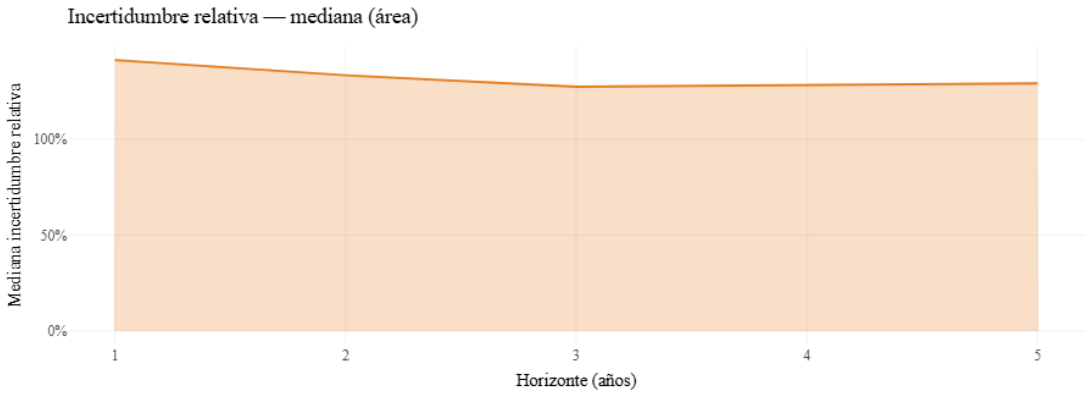
La subfigura a muestra cómo evoluciona el *MAE* en dólares a medida que aumenta el horizonte. Esta lectura da una idea directa del margen de fallo esperado en términos absolutos: lo razonable es que crezca con h, pero de forma gradual y sin saltos bruscos que indiquen zonas de inestabilidad. Con los filtros de **año**, **posición** y **equipo** puede verse si hay tramos o contextos en los que el error se dispara y conviene acotar el uso operativo a horizontes más fiables.

Por su parte, la subfigura b recoge el *MAPE* en porcentaje, lo que permite comparar segmentos con escalas distintas. Picos en horizontes concretos suelen señalar cambios de régimen o datos poco representativos; en esos casos, es preferible revisar supuestos o posponer decisiones a un horizonte donde el porcentaje de error sea más estable. Como complemento a ambas vistas, conviene seguir la **cobertura** del intervalo $[P10, P90]$ para comprobar que las bandas de incertidumbre están bien calibradas para el segmento analizado.

6.2 Cobertura e incertidumbre relativa



(a) Incertidumbre relativa por horizonte (boxplots).



(b) Mediana de la incertidumbre relativa por horizonte (área).

Figura 7: Incertidumbre relativa por h: distribución (arriba) y tendencia central (abajo).

La subfigura a muestra cómo se reparte la incertidumbre relativa en cada horizonte. Aquí entendemos incertidumbre relativa como la anchura del intervalo $[P10, P90]$ en relación con la media prevista, lo que permite comparar horizontes aunque cambie la escala. Si aparecen cajas muy altas o muchos puntos extremos en determinados h, significa que hay más casos con bandas amplias y, por tanto, más dispersión en los escenarios que contemplamos.

La subfigura b recoge la mediana de esa incertidumbre a lo largo de h y separa la tendencia general de los extremos. Si la mediana aumenta con el horizonte, también lo hace el riesgo; conviene trabajar con plazos más cortos o ser más prudente en decisiones a más años. Al aplicar

filtros por año, posición o equipo, es habitual ver que algunos roles o contextos presentan menor incertidumbre que otros; en esos casos puede ser aconsejable cerrar decisiones donde la incertidumbre es baja y diferir aquellas en las que todavía es alta. Aunque la cobertura del intervalo $[P10, P90]$ no aparece en estas vistas, su lectura conjunta es útil para comprobar que las bandas están bien calibradas en el segmento analizado.

6.3 Estabilidad de Top-N

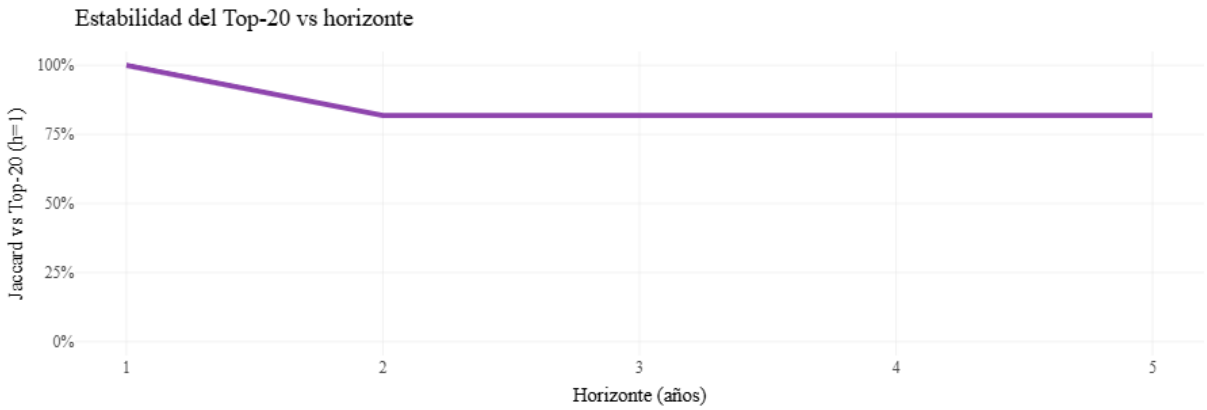


Figura 8: Estabilidad del Top-N vs h (similitud de Jaccard contra $h = 1$).

La **Fig. 8** traduce la idea de "consistencia en el tiempo" en un indicador fácil de leer: cuánto se parece el Top–N de cada horizonte al Top–N de corto plazo. Si la estabilidad se mantiene alta, los nombres clave apenas cambian y es razonable planificar a más plazo; si cae con rapidez, el **timing** importa más, porque los ganadores de hoy pueden no serlo dentro de dos o tres años. Ajustando el *Top–N* y aplicando filtros de **posición** o **equipo**, se pueden construir escenarios específicos (por ejemplo, estabilidad del Top–10 de aleros en dos franquicias candidatas) y cruzarlos con la incertidumbre de la 7 para equilibrar *potencial* y *riesgo* en la toma de decisiones.

7 Resultados: lectura para negocio

7.1 Oportunidades y riesgos individuales

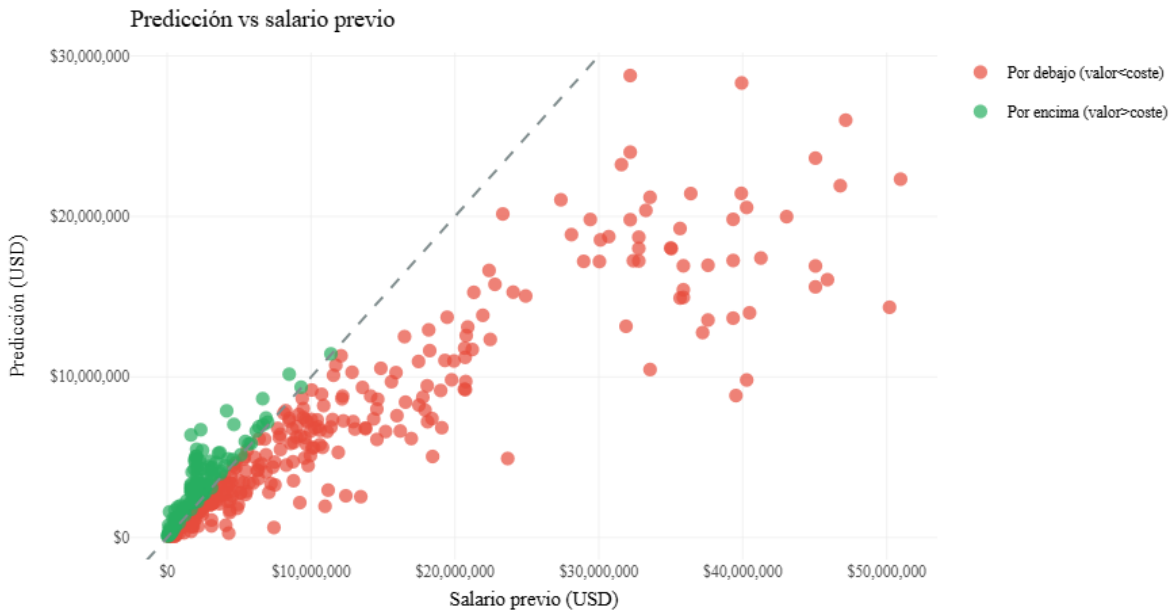
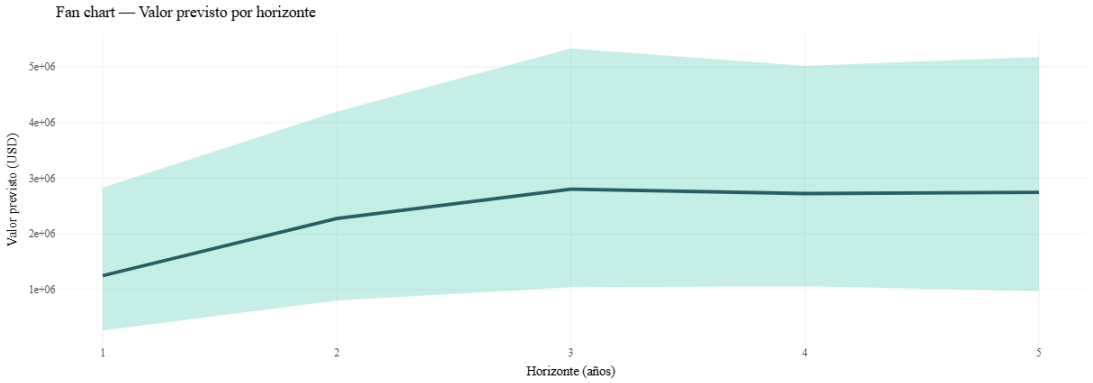


Figura 9: Predicción vs salario previo. La diagonal marca el equilibrio entre valor y coste.

La **Fig. 9** sitúa a cada jugador frente al punto de indiferencia: la diagonal representa el equilibrio entre valor estimado y coste observado. Por encima de esa línea se concentran los posibles casos de **ROI positivo**; por debajo, afloran **riesgos de sobrepago**. Los colores del gráfico ayudan a distinguir rápidamente estas situaciones y, con los filtros de **año**, **posición** y **equipo**, se observan patrones útiles: hay contextos en los que ciertos roles tienden a quedar sistemáticamente por encima o por debajo de la diagonal, lo que orienta renovaciones, extensiones o ajustes de política salarial.

Por otro lado, la Fig. 10(a) muestra la trayectoria del jugador a distintos horizontes, con la media prevista y el intervalo $[P10, P90]$. Una banda **estrecha** sugiere mayor convicción en el rango de resultados; una banda **ancha** apunta a más incertidumbre y recomienda prudencia con compromisos a más plazo. Por su parte, la Fig. 10(b) incorpora la **ficha en imagen** asociada al año actual (valor previsto, coste y ΔUSD), útil para documentar el caso y compararlo con otros candidatos. En este ejemplo se muestra un jugador concreto, pero desde el selector de la app puede visualizarse cualquier otro: tanto el *fan chart* como la ficha se actualizan automáticamente y respetan los filtros activos (año, posición, equipo).



(a) *Fan chart*: media prevista y banda $[P10, P90]$ por horizonte h.

Jugador	Equipo	Posición	Predicción USD	Salario previo USD	ΔUSD
A.J. Lawson	DAL	G (Exterior)	$414,370.34	$393,824.00	$20,546.34

(b) Ficha del jugador (imagen de la tabla asociada).

Figura 10: Ejemplo: valor previsto por horizonte y ficha asociada para un jugador (gráfico + imagen).

7.2 Priorización operativa

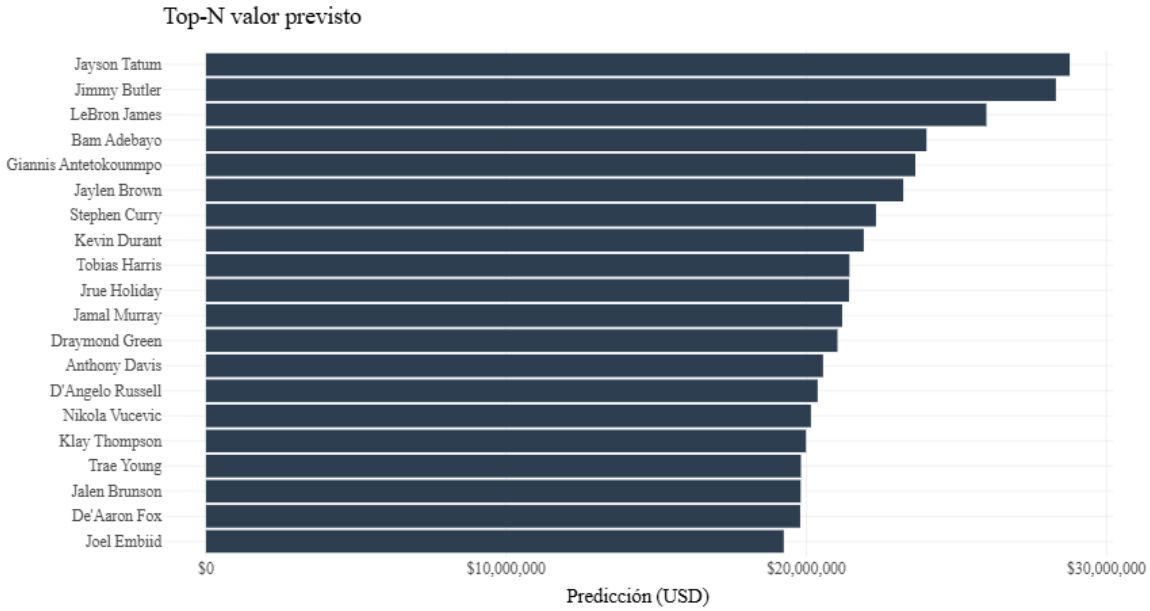


Figura 11: Top–N por valor previsto (el tamaño de N se controla en la app).

La **Fig. 11** ordena el esfuerzo: es la puerta de entrada para focalizar recursos comerciales y deportivos. Ajustar N según objetivo (por ejemplo, Top–10 para activaciones inmediatas o Top–20 para seguimiento) y aplicar filtros de **posición** o **equipo** transforma el ranking en una

lista operativa acorde al contexto. Es habitual que, tras un primer corte por Top–N, se confirme la selección con señales de incertidumbre y con el diferencial de valor–coste.

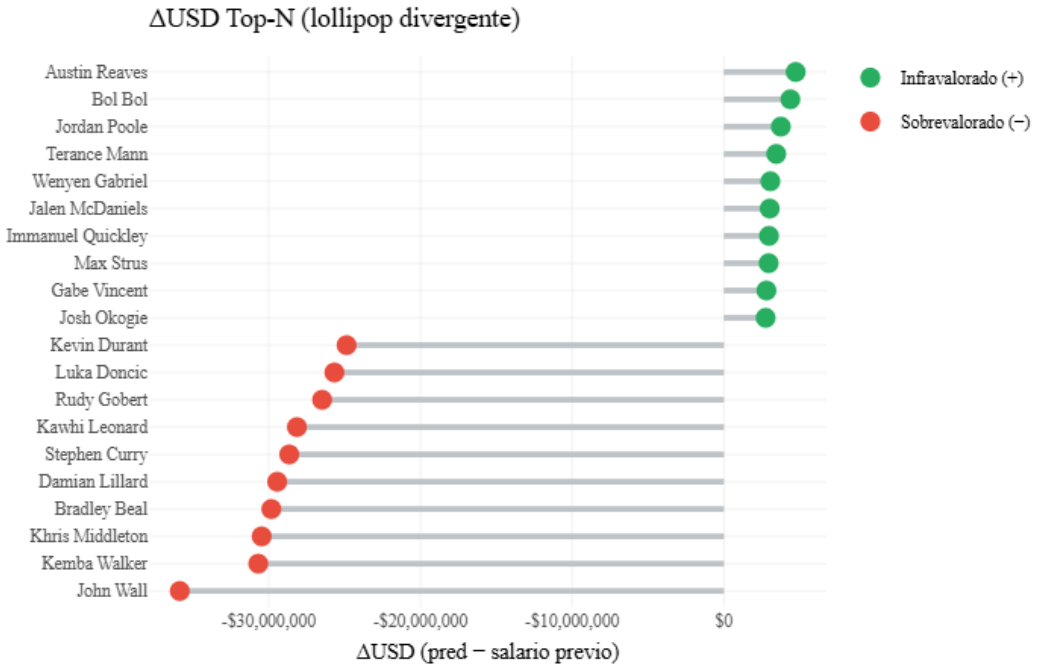


Figura 12: ΔUSD Top–N: infravalorados (derecha) vs sobrevalorados (izquierda).

La **Fig. 12** traduce el diagnóstico en acción al separar a simple vista los **candidatos a inversión** ($\Delta\text{USD} > 0$) de las **alertas de sobrepago** ($\Delta\text{USD} < 0$). Leída junto a 11, permite construir *shortlists* (listas priorizadas de candidatos) que combinan impacto esperado y eficiencia. En la práctica, el proceso es directo: partir del Top–N, filtrar por contexto, y priorizar los casos con gap positivo y riesgo razonable.

7.3 Mapa de oportunidades riesgo–retorno

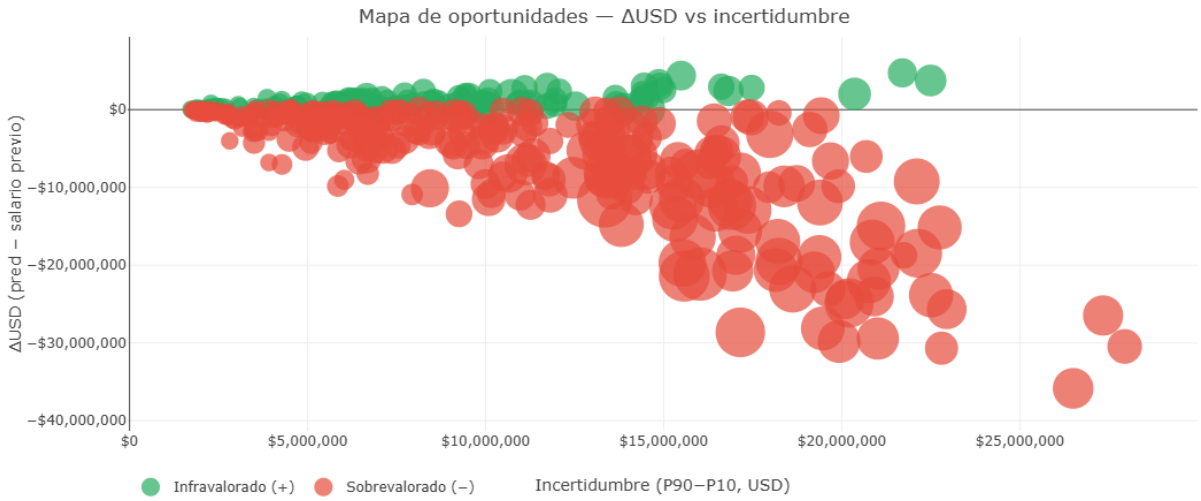


Figura 13: Mapa de ΔUSD (eje Y) frente a incertidumbre (eje X); el tamaño refleja el valor previsto.

La **Fig. 13** alinea retorno y riesgo en una sola vista. La **zona alta–izquierda** concentra el *dulce*: valor por encima del coste con incertidumbre baja; la **zona alta–derecha** sugiere potencial con mayor riesgo y pide cautela en el desembolso; la **zona baja** reúne focos de *riesgo presupuestario* al presentar ΔUSD negativo. El tamaño de la burbuja añade escala al análisis y, con filtros de **equipo**, **posición** o **año**, el mapa se convierte en un tablero de asignación de recursos adaptado a cada necesidad.

7.4 Concentración y eficiencia por equipo

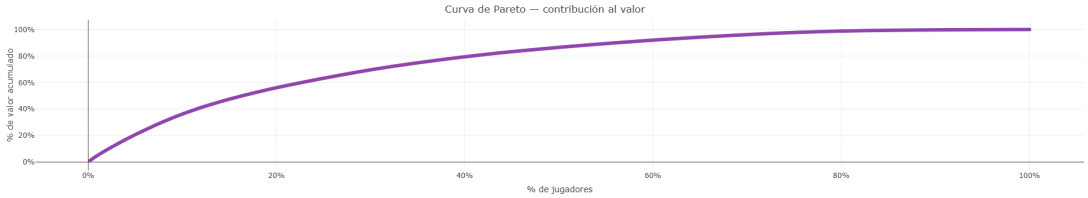


Figura 14: Curva de Pareto: porcentaje acumulado de jugadores frente a porcentaje acumulado de valor.

La **Fig. 14** muestra hasta qué punto el valor se concentra en pocos perfiles: cuanto más se aleje la curva de la diagonal, mayor concentración. Esta lectura justifica estrategias de foco y metas realistas de cobertura: captar al tramo que explica la mayor parte del valor suele rendir más que perseguir cobertura total.

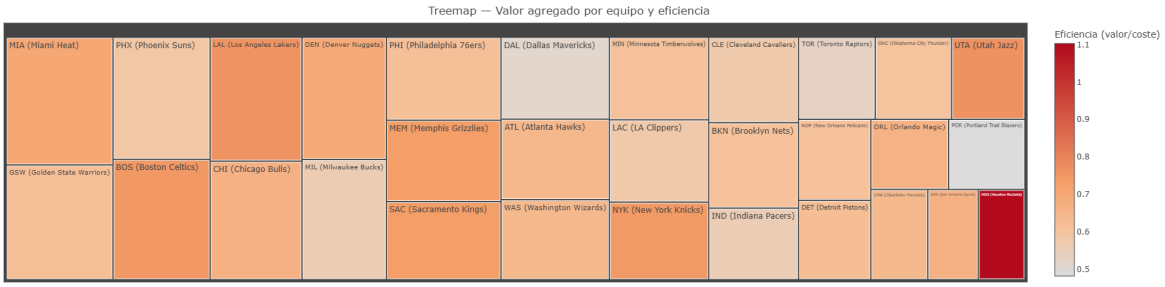


Figura 15: Treemap por equipo: área $\propto$ valor agregado; color $\propto$ eficiencia valor/coste.

La **Fig. 15** aterriza esa concentración en clave de **asignación presupuestaria**. El área refleja el valor agregado por equipo y el color la **eficiencia** (valor/coste): tonos favorables sugieren buen encaje entre inversión y retorno; tonos desfavorables invitan a revisar salarios, roles o captación. Combinado con 14, el treemap permite decidir *dónde* aumentar inversión, *dónde* sostener y *dónde* recortar para mejorar el rendimiento global.

8 De modelo a producto: app Shiny

8.1 Qué resuelve para negocio

La app reduce la **latencia** entre un nuevo *scoring* y su lectura ejecutiva: en unos clics se obtienen KPIs, rankings, vistas de riesgo y fichas por jugador con filtros coherentes de **año**, **posición** y **equipo**. El lenguaje se **uniformiza** (misma definición de ΔUSD, misma *eficiencia* valor/coste), lo que recorta fricciones y acelera la alineación entre áreas deportivas y comerciales. En la práctica, esto significa pasar de "mirar hojas de cálculo" a "tomar decisiones comparables" sobre una base común.

8.2 Robustez de ingesta y normalización

En entornos reales, las fuentes y los formatos **cambian con el tiempo**. Por eso la app reconoce **nombres alternativos de columnas**, **mapea posiciones** desde descripciones largas a siglas canónicas y **armoniza equipos** (abreviaturas y nombres completos) para que los filtros funcionen de forma consistente. Cuando falta información clave, la interfaz lo señala y ofrece métricas de **cobertura de mapeo** para que el usuario sepa cuántos registros quedan plenamente operativos. El objetivo es que *fallar sea difícil* y que, si ocurre, el problema sea visible y trazable.

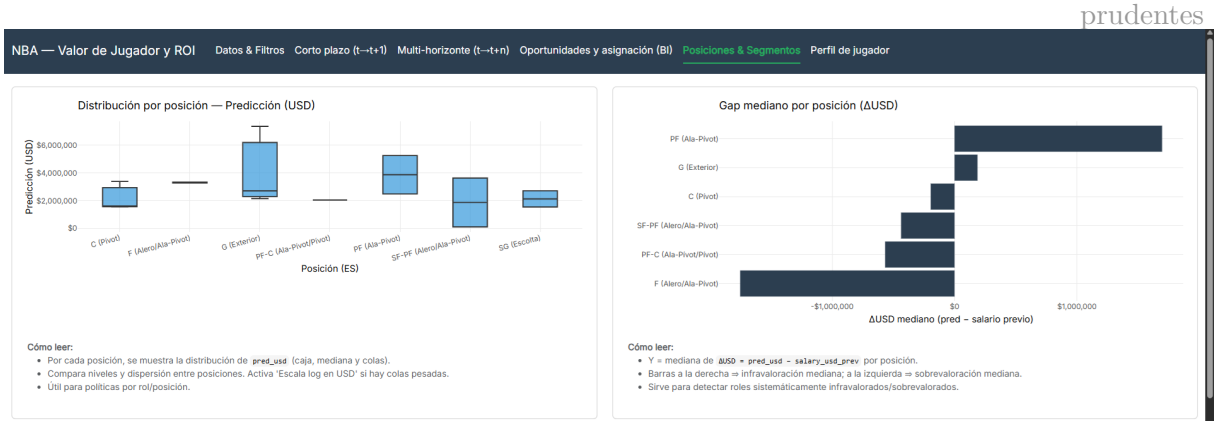


Figura 16: Estructura de la app: pestañas y navegación por flujos (corto plazo, multi-horizonte, BI y perfil de jugador).

La **Fig. 16** muestra la navegación por pestañas que guía el análisis: del corto plazo a los horizontes futuros, de lo agregado al detalle, y de ahí a la ficha individual. Cada vista comparte los mismos filtros, de modo que las conclusiones sean comparables entre pantallas.

8.3 Escalabilidad y extensiones

El artefacto ya cubre el ciclo completo (ingesta en cuaderno, salidas en `reports/` y lectura en la app), y puede crecer sin rehacer piezas. La ampliación natural es sustituir las entradas en fichero por conectores a **APIs** o **bases de datos** con actualizaciones programadas (por temporada o corte diario), manteniendo la misma parametrización de año, horizontes y filtros. En paralelo, el despliegue de la app puede pasar a un entorno **multiusuario** con **autenticación**, **roles** y **trazabilidad** explícita (qué versión de *scoring* y qué filtros sustentan cada exportación).

Además, el núcleo puede exponerse como **servicio reutilizable** (endpoints para *scoring* y bandas P10–P90 por jugador/horizonte), integrable en sistemas corporativos y con **escenarios** "what–if" desde la propia app. Operativamente, encaja en un *MLOps* ligero: ejecuciones periódicas, **versionado** de modelos/pipelines, **monitorización** (MAE, MAPE, cobertura, estabilidad del Top–N) y **alertas** cuando se crucen umbrales de interés (p. ej., ΔUSD > 0 con incertidumbre baja). Con ello, el sistema pasa de entrega analítica puntual a **capacidad operativa** sostenida.

9 Discusión: por qué funciona y cuándo ser prudentes

9.1 Lo que explica el modelo

El valor no es solo rendimiento deportivo: también **atención** y **contexto**. Al incorporar señales de popularidad y tamaño de mercado, el modelo recoge **palancas económicas** que las métricas puramente deportivas no capturan. La validación temporal y la **estabilidad** de los rankings en varios horizontes sugieren que la señal es persistente y útil para priorizar, especialmente cuando se combina con lectura de **incertidumbre** y de **cobertura**.

9.2 Zonas grises y sesgos

La **atención** no es lo mismo que la *intención de compra* y, en determinados momentos, puede inflarse por la coyuntura mediática. También hay factores fuera del modelo como **condiciones**

contractuales, dinámicas de vestuario o decisiones institucionales que alteran la relación entre valor y salario. Por eso se muestran **bandas de incertidumbre** y métricas de **calibración**: ayudan a decidir con mejor contexto, pero no sustituyen el juicio deportivo o comercial. La recomendación es usar el modelo como un **asistente de decisión** y contrastar sus señales con información cualitativa y conocimiento del terreno.

10 Gobernanza y operación del artefacto

10.1 Ciclo de vida sugerido

El artefacto se opera en tres frentes coordinados. En **datos**, conviene versionar entradas, controlar **esquemas** y vigilar **cobertura** por temporada y entidad. En **modelo**, es imprescindible un **checklist anti–fuga temporal**, validación por horizonte h y **registro** continuo de métricas (MAE, MAPE, cobertura y estabilidad tipo Jaccard). En **app**, ayuda fijar **dependencias** (bloqueo de librerías), automatizar **pruebas de ingesta** (nombres alternativos, posiciones) y estandarizar **exportables** para informes.

10.2 Seguridad y cumplimiento

El acceso a las fuentes debe ser de **solo lectura** y el tratamiento de datos personales **mínimo** (no aplica PII en este caso). La auditoría debe registrar **versiones de scoring**, filtros aplicados y **artefactos visuales** compartidos, idealmente bajo **SSO** y control de roles. Con esto, el trazado de decisiones queda documentado.

11 Conclusiones

Este proyecto pone a disposición del usuario una **medición operativa** del valor económico del jugador que es, a la vez, interpretable, accionable y extensible. Partiendo de fuentes heterogéneas y de calidad desigual, se ha construido un *pipeline* que normaliza identidades, equipos y posiciones; integra señales de rendimiento, popularidad y contexto de mercado; y transforma todo ello en una estimación monetizada del valor para el próximo ciclo, con **incertidumbre explícita** y lectura consistente de ΔUSD (valor menos coste). El resultado no es solo un conjunto de números: es un **lenguaje común** para decidir, que reduce la latencia entre la generación de un nuevo *scoring* y su uso en reuniones deportivas y comerciales.

Desde el punto de vista metodológico, la combinación de predicción puntual y bandas $[P10, P90]$ permite separar **tendencia** y **riesgo**. Las métricas de validación (error absoluto medio (MAE) en dólares, error porcentual (MAPE) y cobertura del intervalo) muestran comportamientos coherentes con el aumento del horizonte: el error crece de forma gradual, y la cobertura sirve para comprobar que las bandas están bien calibradas. Además, la **estabilidad de rankings** por horizontes (medida como similitud del Top–N frente al corto plazo) aporta una señal adicional: cuando el *ranking* se mantiene, hay más base para comprometer recursos a futuro; cuando cambia con rapidez, el **timing** pasa a ser central.

En la parte de **extracción de valor**, los datos revelan patrones útiles para priorizar. A nivel agregado, la concentración del valor (lectura tipo Pareto y Gini) justifica estrategias de foco: gran parte del impacto reside en un tramo acotado de jugadores. A nivel individual, la

proyección frente al salario previo sitúa de un vistazo **oportunidades** (valor > coste) y **riesgos** (valor < coste); el *fan chart* añade contexto de convicción al mostrar cómo se abre o cierra la banda por horizonte. Para pasar de diagnóstico a acción, el Top–N por valor y el gráfico divergente de ΔUSD facilitan **listas priorizadas** que equilibran impacto y eficiencia, mientras que el *bubble* riesgo–retorno (incertidumbre en el eje X, ΔUSD en el eje Y y tamaño por escala) ayuda a ordenar inversiones según tolerancia al riesgo. Por su parte, la lectura por **posición** y por **equipo** traduce estas señales a políticas de rol y a **asignación presupuestaria**: el treemap muestra dónde el valor agregado es mayor y cómo de bien se convierte en eficiencia (valor/coste).

El método seguido aporta, además, **trazabilidad y reproducibilidad**. Cada vista comparte filtros coherentes (año, posición, equipo), de modo que las conclusiones sean comparables entre pantallas y sesiones. La app Shiny encapsula el flujo completo (ingesta robusta, normalización, KPIs, figuras y fichas) y convierte la analítica en **producto**: del dato a la decisión sin depender de iteraciones técnicas largas. Esto tiene impacto organizativo inmediato: equipos deportivos y comerciales trabajan sobre la misma definición de ΔUSD, la misma noción de eficiencia y el mismo criterio de riesgo, lo que reduce fricciones y acelera la alineación.

Conviene, no obstante, **ser prudentes** en la interpretación. La atención no equivale siempre a intención de compra, y coyunturas mediáticas o institucionales pueden sesgar temporalmente las señales; lesiones, traspasos y contextos tácticos también alteran la relación entre valor y salario. El sistema reconoce estas zonas grises exponiendo bandas de incertidumbre y métricas de calibración: su función es **asistir** la decisión, no sustituir el criterio experto. La recomendación práctica es modular el horizonte de decisión cuando la incertidumbre crece, revisar supuestos ante picos de error en segmentos concretos y contrastar resultados con evidencia cualitativa.

En conjunto, el trabajo demuestra que es posible **medir, comparar y priorizar** con una base cuantitativa común: sabemos *dónde* se concentra el valor, *qué* perfiles lo generan con mejor relación valor/coste y *cuándo* la incertidumbre aconseja acelerar o esperar. La estandarización de conceptos (valor monetizado, ΔUSD, eficiencia), la validación multi–horizonte y la interfaz de lectura rápida habilitan decisiones más informadas en renovación de contratos, captación, asignación de presupuesto y activación comercial. Esta es la principal conclusión: **del dato se extrae acción**, y de la acción, retorno medible, siempre que se gobierne el ciclo de vida del artefacto y se mantenga la disciplina de lectura conjunta de valor y riesgo.

12 Líneas futuras

El siguiente paso es la ingesta **en tiempo (casi) real** (lesiones, traspasos, noticias) y la evaluación de **escenarios** (*what–if*) con supuestos editables. La **explicabilidad local** por jugador ayudará a entender *por qué* sube o baja su valoración. En el plano operativo, el despliegue corporativo con **SSO**, **logging** y **alertas** permitirá activar informes y avisos automáticos. A medio plazo, la integración con **presupuestación** y con **ventas de patrocinio** cerrará el bucle entre predicción, planificación y retorno.

13 Referencias bibliográficas

[1] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3.ª ed.). OTexts. https://otexts.com/fpp3/

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2.ª ed.). Springer. https://hastie.su.domains/ElemStatLearn/

[3] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2.ª ed.). Springer. https://www.statlearning.com/

[4] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[5] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[6] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[7] Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101. https://doi.org/10.2307/1412159

[8] Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association, 9*(70), 209–219. https://doi.org/10.2307/2276207

[9] Gini, C. (1912). Variabilità e mutabilità. *Studi Economico-Giuridici della R. Università de Cagliari, 3*, 1–158. (Edición clásica). Recurso introductorio moderno: https://doi.org/10.1007/978-3-030-53953-5_5

[10] Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles, 37*, 547–579. https://doi.org/10.5169/seals-266450

[11] Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2023). *shiny: Web Application Framework for R* (v1.7.x) [Paquete R]. Posit. https://shiny.posit.co/

[12] Google. (s. f.). *Google Trends Help: About Google Trends data.* (Consulta de 2025). https://support.google.com/trends/answer/4365533

[13] Wikimedia Foundation. (s. f.). *Pageviews API (Analytics Query Service).* (Consulta de 2025). https://wikitech.wikimedia.org/wiki/Analytics/AQS/Pageviews

[14] Kaggle (usuario: wyattowalsh). (s. f.). *Basketball* [Conjunto de datos]. Kaggle. https://www.kaggle.com/datasets/wyattowalsh/basketball/data

[15] Kaggle (usuario: ratin21). (s. f.). *NBA Player Salaries 2000–2025* [Conjunto de datos]. Kaggle. https://www.kaggle.com/datasets/ratin21/nba-player-salaries-2000-2025

[16] Kaggle (usuario: ratin21). (s. f.). *2022 NBA Team Market Size* [Conjunto de datos]. Kaggle. https://www.kaggle.com/datasets/ratin21/2022-nba-team-market-size

A Anexos

A.1 Alcance y contenido

Este anexo documenta el cuaderno único que contiene de principio a fin el flujo técnico del proyecto, disponible en: *Cuaderno de Colab (TFM)*. El cuaderno reúne en un único flujo todo el trabajo: la ingesta y normalización de las fuentes, la construcción del *dataset maestro* jugador–año, el *scoring* $t \rightarrow t{+}1$, la proyección multi–horizonte h, el cálculo de métricas de validación (MAE, MAPE, cobertura y estabilidad de rankings) y la generación de las tablas y figuras que alimentan el informe y la app de Shiny. Su propósito es garantizar consistencia metodológica, trazabilidad y reproducibilidad entre ejecuciones.

A.2 Ejecución

El cuaderno está preparado para ejecutarse de forma secuencial. Tras abrir el enlace, se recomienda crear una copia en Drive y ejecutar las celdas en orden. En síntesis, el cuaderno:

1. Instala dependencias.
2. Monta o descarga las fuentes de datos configuradas.
3. Expone controles de ejecución (año objetivo, horizontes h, rutas de entrada/salida, filtros básicos por posición/equipo).
4. Emite resúmenes de cobertura y validación durante el proceso.
5. Persiste ficheros y figuras en subcarpetas estandarizadas.

La parametrización reproduce la lógica de filtros utilizada en la app, garantizando coherencia entre las vistas analíticas y los resultados del informe.

A.3 Entradas, salidas y correspondencia con el informe

El cuaderno consume como entradas los ficheros consolidados de estadísticas NBA, salarios y *market size*, junto con señales de popularidad (Wikipedia y Google Trends) y, opcionalmente, un *mapping* jugador→equipo/posición. Las salidas reproducen exactamente los insumos del documento y de la app: `scoring` $t \rightarrow t{+}1$, predicciones multi–horizonte, métricas de backtesting y un conjunto de imágenes en PNG.

La Tabla 1 recoge los **archivos de entrada** del proyecto (datos brutos) y, al final, los **cuatro ficheros** que la app Shiny espera en su propia carpeta.

Carpeta	Archivo / Patrón	Tipo	Descripción
data/raw/csv/	game_info.csv	Entrada	Info básica del partido (fecha, asistencia, duración).
data/raw/csv/	game_summary.csv	Entrada	Resumen del partido, equipos y estado.
data/raw/csv/	inactive_players.csv	Entrada	Jugadores inactivos por partido.
data/raw/csv/	line_score.csv	Entrada	Puntuaciones por cuarto y prórrogas.
data/raw/csv/	officials.csv	Entrada	Árbitros asignados.
data/raw/csv/	other_stats.csv	Entrada	Estadísticas adicionales por partido/equipo.
data/raw/csv/	common_player_info.csv	Entrada	Metadatos de jugador (id, nombre, equipo, posición).
data/raw/csv/	draft_combine_stats.csv	Entrada	Métricas del draft combine.
data/raw/csv/	draft_history.csv	Entrada	Historial de picks del draft.
data/raw/csv/	game.csv	Entrada	Registro maestro de partidos por temporada.
data/raw/csv/	team.csv	Entrada	Metadatos de franquicias (id, nombre, ciudad).
data/raw/csv/	team_details.csv	Entrada	Arena, propietarios y dirección de la franquicia.
data/raw/csv/	team_history.csv	Entrada	Fundaciones/traslados y cambios de nombre.
data/raw/csv/	team_info_common.csv	Entrada	Estadísticas/atributos agregados del equipo.
data/raw/csv/	player.csv	Entrada	Maestro de jugadores (ids y nombres normalizados).
data/raw/csv/	play_by_play.csv	Entrada	Secuencia detallada de jugadas.
data/raw/csv/	NBA Player Salaries_ 2000-2025.csv	Entrada	Rendimiento y salarios de jugadores (2000–2025).
data/raw/csv/	2022 NBA Team Market Size.csv	Entrada	Tamaño de mercado por equipo (2022).
data/raw/csv/	pageviews_2015_2023.csv	Entrada	Visitas a Wikipedia por jugador (2015–2023).
data/raw/csv/	googletrends_interest_yearly 2015_2023_batched.csv	Entrada	Interés anual en Google Trends (2015–2023).
scripts/R_Shiny_app/	app.R	app	Aplicación de R Shiny
scripts/R_Shiny_app/	scoring_results_full_ .csv	Entrada (app)	Scoring $t \rightarrow t{+}1$: pred_usd, P10/P90 y salario previo.
scripts/R_Shiny_app/	scoring_multi_forecast_ .csv	Entrada (app)	Predicciones multi–horizonte: media, P10 y P90 por jugador.
scripts/R_Shiny_app/	mh_metrics.csv	Entrada (app)	Métricas de backtesting por h: MAE, MAPE, cobertura, estabilidad.
scripts/R_Shiny_app/	common_player_info.csv	Entrada (app)	Mapping jugador→equipo/posición para etiquetas.

Cuadro 1: Inventario de **entradas**: datos brutos y ficheros requeridos.

La Tabla 2 recoge los **archivos de salida** generados por el cuaderno y los artefactos entrenados.

Carpeta	Archivo / Patrón	Tipo	Descripción
data/processed/	player_master_year.csv	Salida	*Dataset maestro* jugador–año tras normalización e integración.
data/processed/	scoring_inputs.csv	Salida (op.)	Snapshot de variables de entrada normalizadas usadas en el scoring $t \rightarrow t{+}1$; facilita auditoría y réplica del experimento.
data/processed/	team_mapping.csv	Salida (op.)	Diccionario de abreviaturas/nombres estandarizados de equipos.
reports/	scoring_results_full_y2023_.csv	Salida	Scoring $t \rightarrow t{+}1$ (corte 2023).
reports/	scoring_results_full_.csv	Salida	Scoring $t \rightarrow t{+}1$: pred_usd, P10/P90, salario previo.
reports/	scoring_batch_latest_y2023.csv	Salida (op.)	Lote resumido de scoring para compatibilidad.
reports/	scoring_multi_forecast_y2023_h5.csv	Salida	Predicciones multi–horizonte h=1..5 (media, P10, P90).
reports/	scoring_multi_forecast_.csv	Salida	Predicciones multi–horizonte por jugador.
reports/	mh_predictions.csv	Salida (op.)	Predicciones multi–horizonte (formato alternativo).
reports/	metrics_backtest_multihorizon.csv	Salida	MAE, MAPE, cobertura y estabilidad (Jaccard) por h.
reports/	mh_metrics.csv	Salida	Métricas de backtesting por h (equivalente/alternativa).
reports/	run_manifest.json	Salida (op.)	Manifiesto de ejecución (parámetros, versiones, *checksums*).
models/	model_gbr.joblib	Salida	Modelo Gradient Boosting (punto).
models/	model_rf.joblib	Salida	Modelo Random Forest (punto).
models/	model_ridge.joblib	Salida	Modelo Ridge (baseline lineal).
models/	pipeline_point_gbr.joblib	Salida	*Pipeline* de inferencia para predicción puntual.
models/	pipeline_q10_gbr.joblib	Salida	*Pipeline* de inferencia para P10.
models/	pipeline_q90_gbr.joblib	Salida	*Pipeline* de inferencia para P90.

Cuadro 2: Inventario de **salidas**: datasets procesados, resultados y artefactos de modelos.

Nota: el inventario de archivos y el árbol de directorios son aproximados y pueden presentar pequeñas variaciones respecto a la estructura real.

Estructura de carpetas (1/2). La siguiente rama resume la estructura de directorios y la ubicación de los principales archivos de **entrada** y **salida**.

```
project-root/
 data/
   raw/
     csv/
        2022 NBA Team Market Size.csv
        NBA Player Salaries_2000-2025.csv
        common_player_info.csv
        draft_combine_stats.csv
        draft_history.csv
        game.csv
        game_info.csv
        game_summary.csv
        googletrends_interest_yearly_2015_2023_batched.csv
        inactive_players.csv
        line_score.csv
        officials.csv
        other_stats.csv
        pageviews_2015_2023.csv
        play_by_play.csv
        player.csv
        team.csv
        team_details.csv
        team_info_common.csv
   processed/
         player_master_year.csv              (salida)
         scoring_inputs.csv                  (salida, opcional)
         team_mapping.csv                    (salida, opcional)
 reports/
   scoring_results_full_y2023_*.csv          (salida)
   scoring_results_full_*.csv                (salida)
   scoring_batch_latest_y2023.csv            (salida, opcional)
   scoring_multi_forecast_y2023_h5.csv       (salida)
   scoring_multi_forecast_*.csv              (salida)
   mh_predictions.csv                        (salida, opcional)
   metrics_backtest_multihorizon.csv         (salida)
   mh_metrics.csv                            (salida)
   run_manifest.json                         (salida, opcional)
```

Estructura de carpetas (2/2, continuación).

```
project-root/
 models/
   model_gbr.joblib
   model_rf.joblib
   model_ridge.joblib
   pipeline_point_gbr.joblib
   pipeline_q10_gbr.joblib
   pipeline_q90_gbr.joblib
 scripts/
   popularity_metrics/
        Fetch_Wikipedia_Pageviews_NBA.ipynb
        GoogleTrends_Batched_Resilient.ipynb
   R_Shiny_app/
         app.R
         scoring_results_full_*.csv          (entrada para app)
         scoring_multi_forecast_*.csv        (entrada para app)
         mh_metrics.csv                      (entrada para app)
         common_player_info.csv              (entrada para app)
 notebooks/
       TFM_Cuaderno_Código_Lorente_Molina_Pedro_Jesús
```

A.4 Reproducibilidad y trazabilidad

Para reproducir el TFM hay tres pasos. Primero, descarga la carpeta del proyecto desde Carpeta del proyecto y descomprímela manteniendo la estructura (`data/`, `reports/`, `models/`, `scripts/`, ...). Segundo, abre el cuaderno en Colab desde Cuaderno de Colab (TFM), crea una copia en Drive y ejecútalo de arriba a abajo; el propio cuaderno instala dependencias, prepara los datos y genera las salidas en `data/processed/` y `reports/`. Tercero, utiliza la app Shiny: **opción A (online)** en App Shiny (TFM); **opción B (local)** ejecutando `app.R` desde `scripts/R_Shiny_app/` con los cuatro archivos de entrada en esa misma carpeta (`scoring_results_full_*.csv`, `scoring_multi_forecast_*.csv`, `mh_metrics.csv`, `common_player_info.csv`). De esta forma, con las mismas entradas y parámetros, las salidas son reproducibles.

A.5 Integración con la app Shiny

Las salidas siguen convenciones de nombres detectadas por la app. En ausencia de carga manual, la aplicación busca por defecto ficheros como `scoring_results_full_*.csv`, `scoring_multi_forecast_*.csv` y `mh_metrics.csv` en las rutas del proyecto. Este acoplamiento permite un ciclo de trabajo corto: generación en el cuaderno, lectura inmediata en la interfaz y consistencia entre informe, app y artefactos persistidos (ver también Sección 10).

A.6 Licencias y créditos

Este trabajo utiliza datos públicos descargados de Kaggle y se acoge a los *Data Terms* de Kaggle y a las condiciones específicas de cada conjunto, manteniendo la atribución y el uso permitido. En particular: el dataset general de baloncesto *Basketball* (wyattowalsh), el histórico de salarios *NBA Player Salaries 2000–2025* (ratin21) y el tamaño de mercado por franquicia *2022 NBA Team Market Size* (ratin21). El uso que se hace aquí es académico/no comercial; no se redistribuyen los ficheros originales fuera del ámbito del TFM y se conserva la autoría de sus

creadores y de la plataforma. Las señales de popularidad (Wikipedia y Google Trends) se han obtenido a través de sus interfaces públicas respetando sus términos de servicio; el proyecto no trata información personal identificable (PII). Para facilitar la auditoría y la reproducibilidad, guarda junto al PDF final el paquete de artefactos generado en la ejecución y mantén accesible el enlace al Cuaderno de Colab (TFM).