Informe Final

Integrantes

Sebastián Saldarriaga Cardona
Juan David Ríos Rodríguez
Wilmer Mario Leiva Esteban

Prof. Raul Ramos Pollan

Introducción a la Inteligencia Artificial

Universidad de Antioquia

2023-2

Introducción

El presente proyecto se centra en el desarrollo de un modelo de aprendizaje automático destinado a predecir el valor de mercado de jugadores de fútbol, utilizando como base el conjunto de datos "FIFA 19 complete player dataset" de Kaggle. La iniciativa surge de la necesidad de comprender y cuantificar los factores que influyen en la valoración de un jugador, considerando variables que abarcan desde habilidades técnicas hasta aspectos contractuales.

La esencia del enfoque radica en la consideración de factores diversos que influyen en la valoración de los jugadores, involucra un rango muy amplio de características. El propósito final es ofrecer una herramienta predictiva con aplicaciones valiosas en las negociaciones y decisiones comerciales dentro de la industria del fútbol.

Durante la ejecución del proyecto, se adoptará una métrica de desempeño estándar en el ámbito del aprendizaje automático, el Error Absoluto Medio (MAE), para evaluar la precisión del modelo. Además, se introducirá una métrica de negocio específica: la Proporción de Jugadores Valorados Correctamente. Este enfoque asegurará que el modelo no solo sea preciso desde una perspectiva técnica, sino que también cumpla con los objetivos comerciales establecidos.

Como criterio para la implementación del modelo en producción, se ha definido que al menos el 85% de las predicciones deben caer dentro de un margen de error aceptable, es decir, un rango del 10% respecto al valor real. Este criterio busca garantizar la utilidad práctica del modelo en contextos reales de negociación en la industria futbolística.

Exploración Descriptiva del Dataset

El conjunto de datos "FIFA 19 complete player dataset" de Kaggle, compuesto por 18,200 filas y 89 columnas, proporciona una visión holística de diversos aspectos relacionados con los jugadores de fútbol. La exploración general del dataset revela una amplia gama de información que abarca desde la habilidad técnica hasta la posición en el campo y aspectos contractuales.
Este panorama inicial sienta las bases para un análisis más profundo y la construcción de un modelo de aprendizaje automático capaz de predecir el valor de mercado de los jugadores de fútbol en función de estos diversos factores

Iteraciones de Desarrollo

Iteración 1:

Notebooks:

- 01 - Limpieza de datos.
- 02- Primeros modelos.

Preprocesado de Datos:

Importación del Dataset: Se procedió a obtener el token necesario de Kaggle para facilitar la importación directa del dataset a Google Colab. Una vez obtenido, se especificó el dataset deseado y se descomprimió para su posterior utilización. Esta fase permitió contar con el conjunto de datos listo para ser explorado y analizado.

Entendimiento del Dataset: Durante esta fase, se llevó a cabo un análisis exhaustivo de las columnas disponibles. Se identificaron las características más relevantes para el desarrollo del modelo, considerando el tipo de dato de cada columna. Este proceso fue esencial para realizar una primera selección de variables antes de adentrarse en procesos más sofisticados.

Limpieza del Dataset:

Eliminar columnas menos relevantes: Se realizó la eliminación de columnas que, a primera vista, no aportaban un impacto significativo al proceso de selección. Esta acción redujo la complejidad del dataset, centrándose en las variables más cruciales para el análisis.

Celdas con valores NaN: Las columnas que contenían predominantemente valores NaN fueron descartadas. Al carecer de suficientes datos, estas columnas no contribuirían de manera efectiva al modelo.

Jugadores Específicos: Se identificaron jugadores con información faltante en columnas de alta relevancia, como 'Overall' o 'Age'. Se abordó este problema mediante la incorporación manual de estos datos para evitar pérdida de información crucial.

Jugadores sin Equipo: Aquellos jugadores que no estaban asociados a ningún equipo fueron eliminados. Este paso fue significativo, ya que estos jugadores presentaban un patrón atípico y carecían de información en columnas

relacionadas con valores monetarios, generando desequilibrios en los datos.

Jugadores con muchas estadísticas desconocidas: Se identificaron 48 jugadores con campos vacíos en 10 columnas. Estos jugadores fueron descartados para evitar desbalances en el entrenamiento del modelo.

Eliminación de los Porteros: Se detectó que los porteros presentaban numerosas columnas con datos NaN. Dada la falta generalizada de información relevante para este tipo de jugadores, se optó por descartarlos del dataset. Las columnas específicas relacionadas con los porteros también fueron eliminadas.

Jugadores con Release Clause en NaN: Se identificó otra columna con valores NaN, específicamente la columna 'Release Clause'. Este vacío se debía a préstamos anteriores, por lo que los jugadores afectados fueron eliminados.

Conversión de datos numéricos: Una vez asegurado que el dataset estaba libre de columnas con datos nulos, se llevó a cabo la conversión de múltiples columnas, incluyendo valores monetarios como 'Value', 'Wage', y 'Release Clause'.

Conversión de columnas categóricas: El dataset incluía cuatro columnas categóricas: 'Nationality', 'Club', 'Preferred Foot', y 'Position'. Se realizaron conversiones y reemplazos específicos, como la transformación de 'Preferred Foot' de 'right' y 'left' a 1 y 0, y la sustitución de 'Nationality' y 'Club' por la media del valor de sus jugadores respectivos. La columna 'Position' fue suprimida, considerando que la calidad específica para cada posición ya era suficiente para la categoría.

Modelos Supervisados:

Selección de Modelo de Regresión: Se optó por la Regresión Lineal como modelo inicial de regresión. Además, se dividió el dataset en conjuntos de entrenamiento y prueba, preparando el terreno para la evaluación del modelo.

Evaluación del Modelo de Regresión: Se aplicaron métricas de evaluación, destacando el Error Absoluto Medio (MAE), que mide la magnitud promedio de los errores en las predicciones del modelo. Las curvas de aprendizaje también fueron visualizadas, proporcionando una comprensión más profunda de la relación entre el tamaño del conjunto de entrenamiento y el rendimiento del modelo.

Modelos No Supervisados:

Aplicación de PCA: Se implementó PCA (Análisis de Componentes Principales) como técnica de reducción de dimensionalidad. Esto permitió visualizar la estructura del dataset y simplificar la representación de las características.

Resultados, Métricas y Curvas de Aprendizaje:

- Se evaluó el rendimiento del modelo de regresión utilizando métricas específicas, como MAE.

- Las curvas de aprendizaje se utilizaron para analizar cómo el rendimiento del modelo variaba en relación con el tamaño del conjunto de entrenamiento. Este análisis proporcionó información crucial sobre la capacidad predictiva del modelo en diferentes contextos de entrenamiento.

Iteración 2:

Notebooks:

- 03- Prueba general de características.

Preprocesado de Datos:

Descarga del Dataset: Se descargó el dataset "fifa-19-cleaned-dataset.csv" desde Google Drive utilizando el paquete gdown.

Carga y Configuración del DataFrame: El dataset se cargó en un DataFrame de Pandas, y se configuró para visualizar todas las columnas mediante pd.set_option('display.max_columns', None).

Definición de Funciones de Preprocesado: Se definieron funciones como load_data para separar características y variable objetivo, y se implementaron funciones para entrenar modelos de árbol de decisión y bosque aleatorio.

Modelos Supervisados:

Entrenamiento y Evaluación del Decision Tree: Se prepararon los datos para el modelo de Decision Tree, se entrenó y se evaluó su rendimiento. El modelo mostró un MAE de 212962.0210, MSE de 759766929467.6161 y un 77.79% de predicciones dentro del margen de error del 10%.

Entrenamiento y Evaluación del Random Forest: Se procedió con el modelo de Random Forest, y su evaluación reveló un MAE de 168736.1139, MSE de 472844833075.6188 y un destacado 86.16% de predicciones dentro del margen de error del 10%.

Resultados y Métricas:

Importancia de Características: Se visualizó la importancia de las características para el Decision Tree y para el Random Forest, destacando que la columna 'Release Clause' era la más influyente.

Eliminación de Característica: Se eliminó la columna 'Release Clause', considerando su excesiva influencia en el modelo y así probar su rendimiento sin ella.

Reentrenamiento y Evaluación Post Eliminación: Se procedió al reentrenamiento y evaluación de ambos modelos tras la eliminación de 'Release Clause'.

- Decision Tree Post Eliminación:
  - MAE: 200252.6280, MSE: 1920095108511.3599, 85.89% dentro del margen de error del 10%.

- Random Forest Post Eliminación:
  - MAE: 159056.2394, MSE: 1018329310509.4948, 89.69% dentro del margen de error del 10%.

Pruebas de Importancia de Características Post Eliminación: Se realizó una prueba de importancia de características para ambos modelos después de la eliminación de 'Release Clause'.

Conclusión de la Iteración 2: La eliminación de la columna 'Release Clause' resultó en mejoras notables en la precisión de ambos modelos, especialmente en el caso del modelo Random Forest, que alcanzó un 89.69% de predicciones dentro del margen de error del 10%.

Iteración 3:

Notebook:

- 04 - Decision tree.

Preprocesado de Datos:

Descarga del Dataset: Se descargó el dataset "fifa-19-cleaned-dataset.csv" desde Google Drive utilizando el paquete gdown.

Carga y Configuración del DataFrame: El dataset se cargó en un DataFrame de Pandas, y se configuró para visualizar todas las columnas mediante pd.set_option('display.max_columns', None).

Definición de Funciones de Preprocesado: Se definieron funciones como load_data para separar características y la variable objetivo, y se implementaron funciones para entrenar modelos de árbol de decisión y bosque aleatorio.

Modelos Supervisados:

Entrenamiento y Evaluación del Decision Tree: Se prepararon los datos para el modelo de Decision Tree, se entrenó y se evaluó su rendimiento. El modelo original mostró un MAE de 165959.6473, MSE de 867430612071.8888 y un 84.74% de predicciones dentro del margen de error del 10%.

Análisis de Importancia de Características: Se calculó la importancia de las características para entender cuáles influyen más en las predicciones.

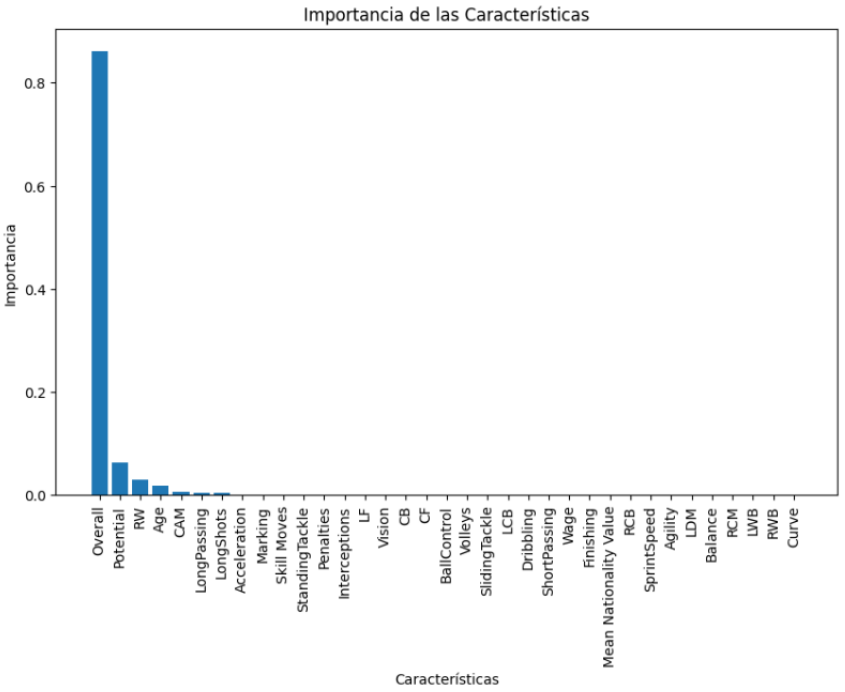
Eliminación de Características por Importancia Baja: Se identificaron y eliminaron características con importancia muy baja (casi igual a 0) para el modelo. Se estableció un umbral de importancia mínima variante para evaluar cómo cambiaba el modelo según qué columnas se eliminaban.

Re-Entrenamiento y Evaluación Post Eliminación:

Decision Tree Post Eliminación:

- MAE: 160461.1733, MSE: 820198134961.0038, 85.11% dentro del margen de error del 10% al eliminar las características que aportan menos de un 0.01%.

- **Pruebas de Importancia de Características Post Eliminación:** Se realizó una nueva prueba de importancia de características para el modelo Decision Tree después de la eliminación de características.

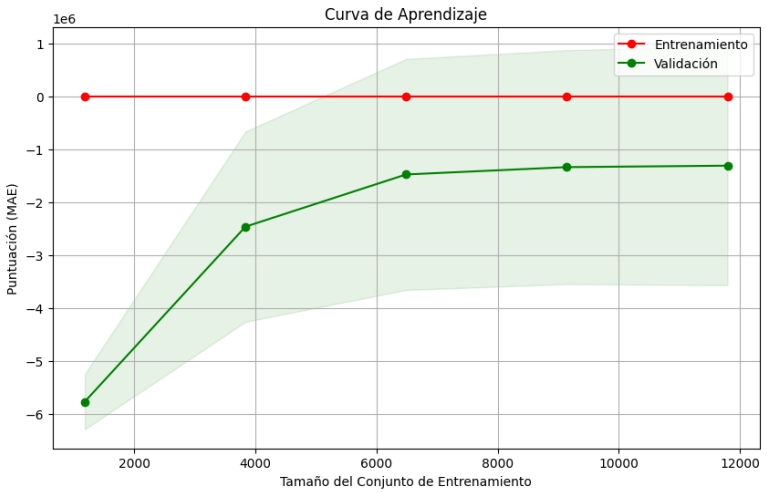


Resultados, Métricas y Curvas de Aprendizaje:

Conclusión de la Iteración 3:

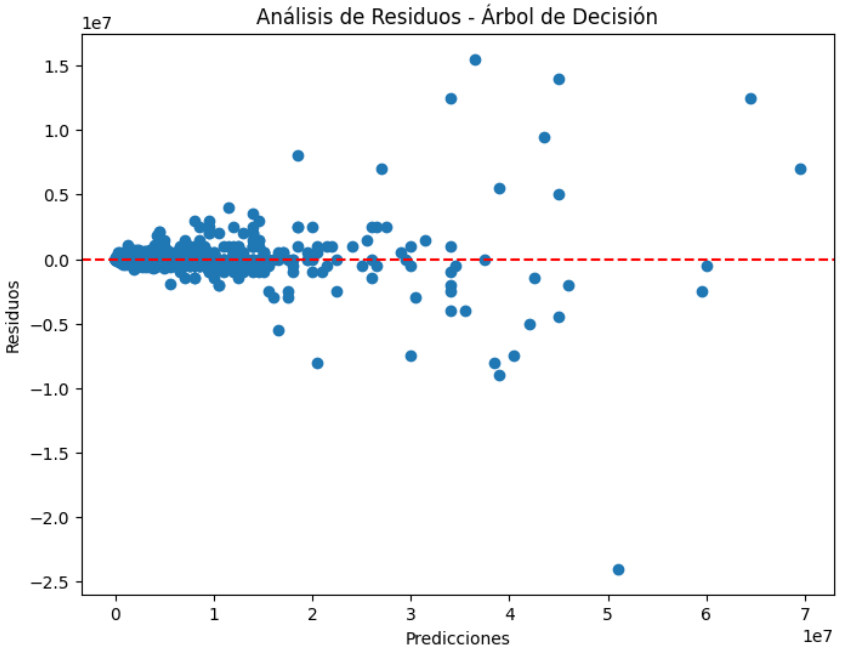
La eliminación de características con importancia muy baja resultó en una mejora en la precisión del modelo Decision Tree, con un MAE de 160461.1733 y un 85.11% de predicciones dentro del margen de error del 10%. Este enfoque más simple podría preferirse ya que aumenta la eficiencia del modelo a la vez que reduce su complejidad (pasó de 69 columnas a 35 tras la eliminación).

El modelo reducido presenta la siguiente curva aprendizaje:



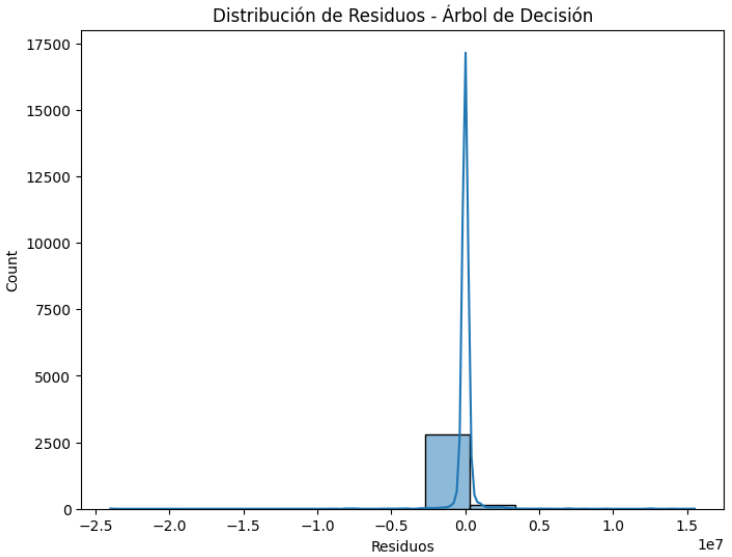
Esta curva revela que la inclusión de nuevos datos al dataset no presentaría una mejora sustancial en su eficiencia, por lo que no parece ser necesario realizar nuevas mediciones.

Adicionalmente se realizó el análisis de residuos para evaluar la calidad de ajuste de nuestro modelo de regresión. Este análisis nos ayudó a verificar si se cumplen los supuestos clave, como homocedasticidad y normalidad de los residuos, proporcionándonos información crucial sobre la validez y eficacia de nuestro modelo.



Con base en el análisis de residuos vs. predicciones, se observa que la gran mayoría de los datos se concentran cerca de cero en el eje y, indicando una adecuada alineación con el supuesto de homocedasticidad. Sin embargo, la forma de embudo a medida que avanza a la derecha sugiere una variabilidad ligeramente mayor en los residuos para ciertos rangos de predicciones. Aunque hay indicios de heterocedasticidad en esta pequeña proporción de datos, la homocedasticidad general del modelo puede considerarse aceptable.

Ahora, para revisar el supuesto de normalidad se hizo el siguiente gráfico:



Mean Absolute Error - Árbol de Decisión: 160461.17327907766

R-squared - Árbol de Decisión: 0.9739734368229317

La forma de la gráfica sugiere que el modelo en su mayoría está haciendo predicciones precisas, presentando una distribución alrededor de cero que corresponde con la distribución normal. Sin embargo puede tener valor investigar más a fondo esos valores atípicos que contribuyen a ese pico tan alto en la gráfica

En general los resultados parecen mostrar que el modelo tiene un rendimiento bastante bueno, con un bajo MAE y un alto R2.

Iteración 4:

Notebook:

- 05 - Random Forest.

Preprocesado de Datos:

Preparación del Entorno: Instalación de librerías, Se instaló la librería gdown para la gestión de descargas.

Carga y Exploración del Dataset:

- **Descarga y Carga del Dataset:** Se descargó el dataset "fifa-19-cleaned-dataset.csv" desde Google Drive utilizando el paquete gdown. El conjunto de datos fue cargado en un DataFrame de Pandas.

- **Visualización del DataFrame:** Se configuró el DataFrame para visualizar todas las columnas y se mostraron las primeras filas del DataFrame.

Exploración de Características: Se presentó un listado de todas las columnas presentes en el DataFrame.

Modelos Supervisados:

Entrenamiento de Random Forest: Se definieron funciones para cargar datos, entrenar el modelo de Random Forest y evaluar su rendimiento.

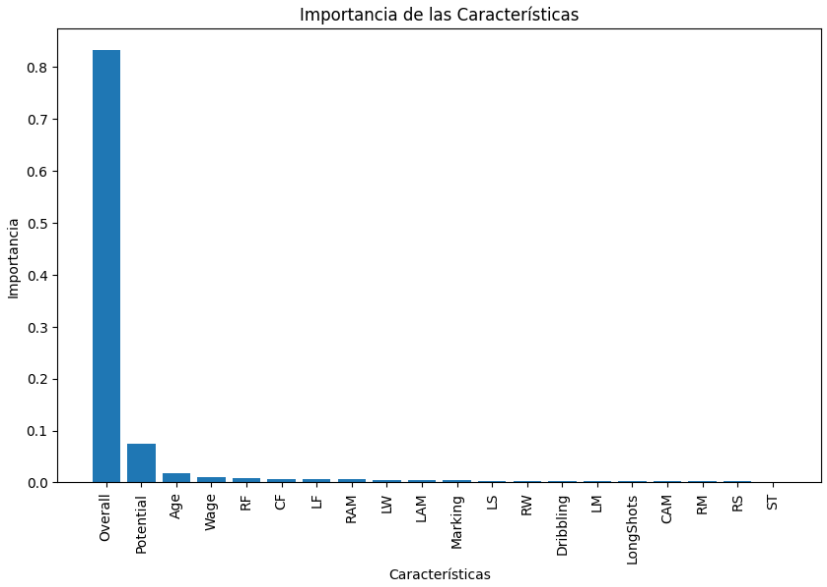
Análisis de Importancia de Características:

- **Importancia de Características:**
  Se realizó un análisis para entender cuáles características influyen más en las predicciones del modelo de Random Forest.
- **Visualización de Importancia:**
  Se visualizó la importancia de las características en un gráfico, mostrando las más relevantes.

Pruebas de Eliminación de Características:

- **Prueba de Remoción:** Se realizaron pruebas eliminando características con impacto muy bajo en el modelo. Se evaluó el rendimiento del Random Forest después de la eliminación.

- **Resultados tras la eliminación de características de bajo impacto:** Se estableció un umbral de importancia (0.1%) y se identificaron características por debajo de este umbral para su eliminación.

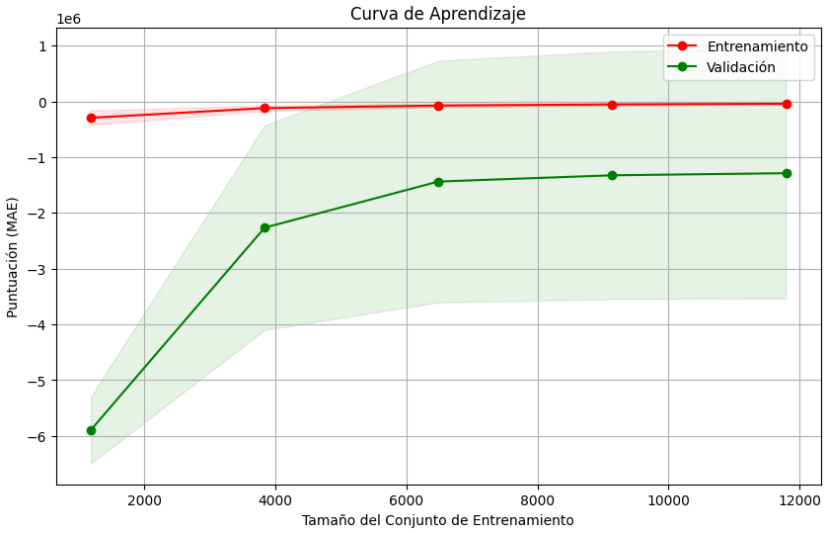


Resultados, Métricas y Curvas de Aprendizaje:

Conclusión de la Iteración 4:

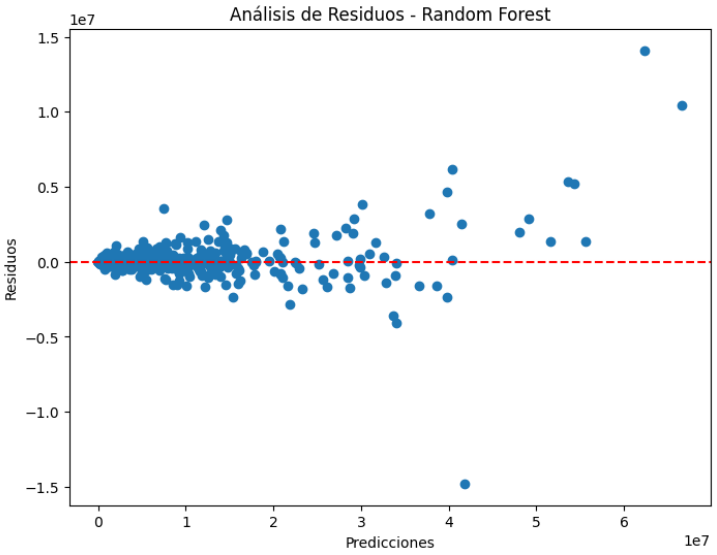
Se entrenó el modelo de Random Forest con un MAE de 115063.9030, MSE de 316000962559.3422 y un 89.25%% de predicciones dentro del margen de error del 10%. La eliminación de características poco relevantes no afectó significativamente el rendimiento del modelo (se redujo en aproximadamente 1%) mientras que sí redujo significativamente su complejidad (pasó de 69 a 20 columnas).

De igual manera que para el Decision Tree, se generó la curva de aprendizaje para el Random Forest:

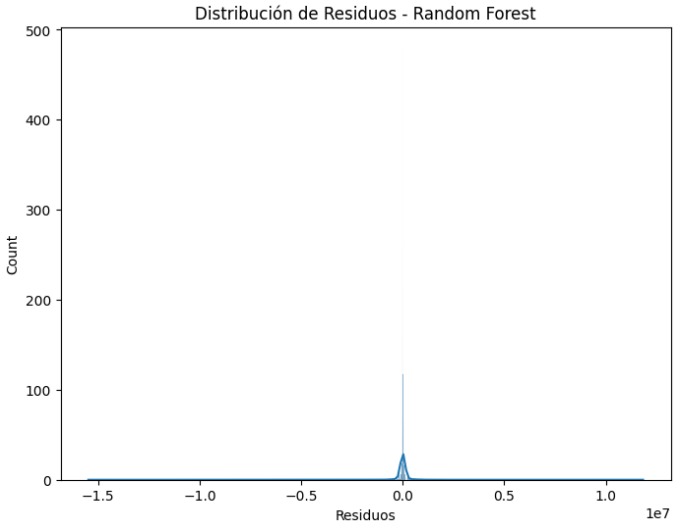


En esta, la línea de entrenamiento permanece relativamente estable, indicando que el modelo se ajusta bien a los datos de entrenamiento. La línea de validación aumenta inicialmente, alcanzando un punto de estabilización alrededor de 6000 o 7000, sugiriendo que agregar más datos ya no mejora significativamente la capacidad de generalización del modelo, de igual forma que pasa con la del decision tree.

Y nuevamente se realizó el análisis de residuos para evaluar la calidad de ajuste de nuestro modelo de regresión, ayudándonos a comprobar los supuestos ya mencionados, proporcionándonos información crucial sobre la validez y eficacia de nuestro modelo.



Tras examinar la relación entre los residuos y las predicciones en el análisis, se nota que la gran mayoría de los datos se agrupa alrededor del valor cero en el eje y, lo cual sugiere una concordancia adecuada con el supuesto de homocedasticidad. No obstante, la presencia de una forma de embudo a medida que avanzamos hacia la derecha indica una variabilidad ligeramente mayor en los residuos para ciertos intervalos de predicciones. Aunque hay señales de heterocedasticidad en esta porción reducida de datos, la homocedasticidad general del modelo parece ser aceptable. Este patrón es consistente con lo observado en el análisis de residuos del modelo de árbol de decisión, demostrando una tendencia similar en ambos casos.



Mean Absolute Error - Random Forest: 113483.38419803324

R-squared - Random Forest: 0.9908513204296178

Una vez más, al examinar la segunda gráfica, se obtienen resultados similares al decision tree, aunque con una notable diferencia: el pico en la gráfica es menos pronunciado en comparación con el modelo de árbol de decisión. La forma general de la gráfica indica que el modelo en su mayoría realiza predicciones precisas pero puede ser importante realizar más investigaciones sobre los valores atípicos que contribuyen a ese pico menos acentuado en la gráfica. Esta, de igual forma parece presentar una distribución normal.

Retos y Consideraciones de Despliegue

Retos Principales

1. Escalabilidad: Uno de los desafíos principales es asegurar que el modelo pueda manejar un volumen significativo de predicciones en tiempo real, especialmente en entornos donde la cantidad de datos de jugadores de fútbol puede ser considerable. La infraestructura y la arquitectura del sistema deben diseñarse para escalar de manera eficiente y garantizar un rendimiento consistente incluso bajo cargas de trabajo elevadas.

2. Seguridad de Datos: El modelo debe implementarse en un entorno seguro que proteja la integridad y la confidencialidad de los datos de los jugadores, evitando posibles brechas de seguridad o accesos no autorizados.

3. Interpretabilidad del Modelo: La capacidad de entender y explicar las decisiones del modelo es crucial, especialmente en industrias como la del fútbol, donde las decisiones basadas en datos pueden tener un impacto significativo. Garantizar la interpretabilidad del modelo permitirá a los usuarios, como agentes deportivos o directores técnicos, confiar en las predicciones y comprender los factores que influyen en ellas.

4. Actualización Continua: Los datos en la industria del fútbol están en constante cambio debido a transferencias, nuevas temporadas y cambios en el rendimiento de los jugadores. Es esencial implementar un mecanismo eficiente para actualizar el modelo de manera regular, ya sea mediante la reentrenamiento periódico o la adopción de técnicas de aprendizaje incremental.

5. Monitorización del Rendimiento: Establecer un sistema de monitorización continua es esencial para detectar posibles degradaciones en el rendimiento del modelo. La monitorización permitirá identificar y abordar rápidamente problemas relacionados con cambios en la distribución de datos o en la calidad de las predicciones.

Consideraciones Adicionales

1. Documentación Clara: Proporcionar documentación detallada sobre la implementación del modelo, sus requisitos y limitaciones es crucial para aquellos que interactúan con el sistema.

2. Capacitación del Usuario: Garantizar que los usuarios clave, como analistas deportivos y directores técnicos, estén capacitados para comprender y utilizar eficazmente las predicciones del modelo es esencial. Esto contribuirá a una adopción exitosa y a la obtención de beneficios significativos en el ámbito profesional.

Conclusiones

- La capacidad del modelo para mantener un rendimiento sólido, incluso después de la optimización y eliminación de características menos relevantes, destaca la robustez del enfoque adoptado, lo que sugiere que el modelo es capaz de generalizar bien y no depende excesivamente de características específicas.

- El modelo final, especialmente en el caso del Random Forest, exhibe un rendimiento sobresaliente con un MAE bajo, un MSE significativamente reducido y un alto porcentaje de predicciones dentro del margen de error del 10%. Este éxito destaca la eficacia de la estrategia adoptada para el preprocesamiento y entrenamiento del modelo.

- los resultados evidencian una mejora sustancial al optar por el modelo de Random Forest en comparación con el árbol de decisión. La significativa reducción en el Error Absoluto Medio (MAE) y el aumento del coeficiente de determinación (R2) indican una mayor precisión en las predicciones del modelo de Random Forest. A pesar del tiempo de ajuste de hiper parámetros más prolongado, este modelo demuestra ser más robusto y preciso. Aunque persisten las limitaciones sobre la generalización, estos hallazgos sugieren que el Random Forest es una elección superior, destacando su potencial para aplicaciones prácticas.