PREDICCIÓN DEL PRECIO DE VIVIENDAS RESIDENCIALES MEDIANTE ANALISIS MULTIVARIANTE Y TECNICAS DE REGULARIZACION

Marcos Gómez Vega
mgomeveg@myuax.com

Índice

Introducción

1. Contextualización del Problema

El presente informe aborda el desafío de la **predicción de precios de viviendas residenciales** (SalePrice) utilizando un conjunto de datos multivariante que describe detalladamente las características físicas, estructurales y de contexto urbano de las propiedades. La valoración automática de activos inmobiliarios es una tarea con alta demanda y relevancia práctica en sectores como banca, inmobiliarias y plataformas tecnológicas. Este *dataset* proporciona un escenario ideal para aplicar técnicas avanzadas de modelado predictivo y *Machine Learning*.

2. Objetivo Principal del Estudio

El objetivo central de este trabajo es **construir y comparar modelos de regresión robustos** capaces de predecir con precisión la variable objetivo SalePrice. Para ello, se emplearán las variables restantes del conjunto de datos como variables predictoras.

3. Metodología y Modelos

La metodología se estructura en las siguientes fases clave:

- **Preprocesamiento de Datos:** Tratamiento riguroso de valores ausentes (imputación o eliminación), **transformación logarítmica** de la variable objetivo (SalePrice), codificación de variables categóricas, y estandarización de variables predictoras.
- **Análisis Exploratorio de Datos (EDA):** Identificación de patrones, tendencias y relaciones iniciales con el precio de venta.
- **Modelado y Comparación:** Se construirán y evaluarán cinco modelos predictivos basados en la regresión lineal, incluyendo:
    1. Regresión Lineal Múltiple con Reducción de Dimensionalidad (PCA).
    2. Regresión con regularización **Lasso**.
    3. Regresión con regularización **Ridge**.
    4. Regresiones Lasso y Ridge aplicadas sobre componentes principales (PCA).
- **Evaluación:** La precisión, robustez y capacidad de generalización se estimarán mediante la división del conjunto de datos en subconjuntos de entrenamiento, validación y test, y se utilizarán métricas como el **Error Cuadrático Medio de la Raíz (RMSE)** y el **Error Absoluto Medio (MAE)** para la comparación final.

Análisis Exploratorio y Justificación del Preprocesamiento

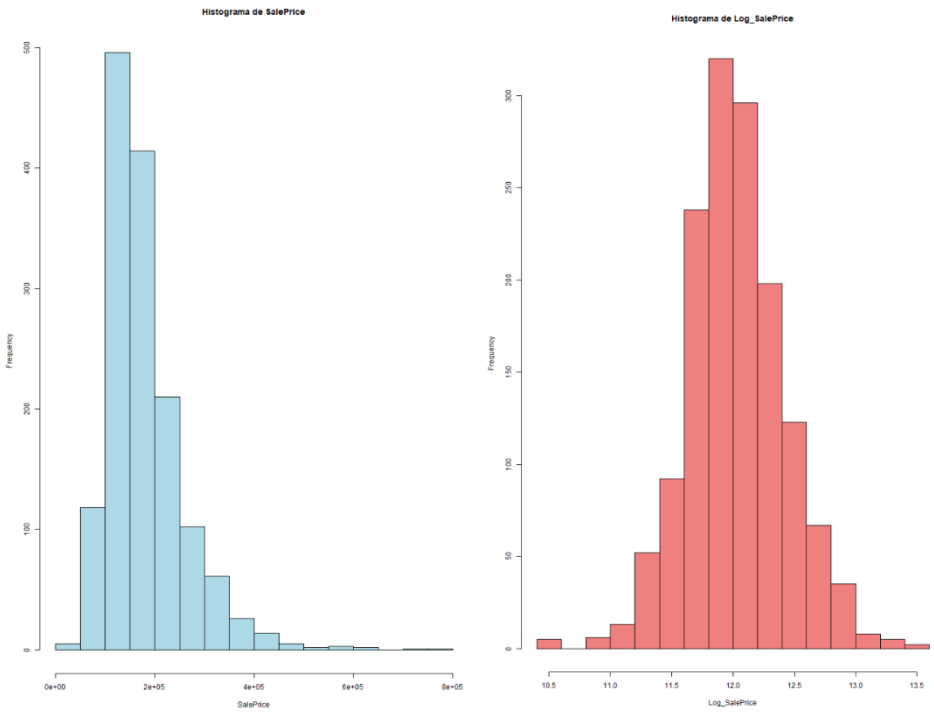
1.- Análisis de la variable objetivo

El análisis exploratorio se centró en la variable objetivo, **SalePrice**, para asegurar que se cumplieran los supuestos de distribución necesarios para los modelos de regresión lineal. Lo primero de todo fue el estudio de las estadísticas descriptivas de

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34900  129975  163000  180921  214000  755000
```

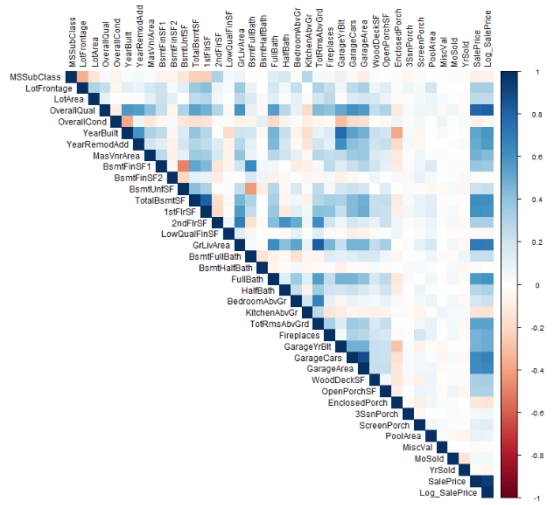
Como podemos observar, la media (**180.921€)** es significativamente mayor que la mediana (**163.000€**), lo que indica una **fuerte asimetría positiva** o sesgo hacia la derecha. Esto, junto con la gran dispersión entre el mínimo (**34.900€**) y el máximo (**755.000€**), requiere una corrección.

Para corregir esta asimetría y la heterocedasticidad asociada, se aplicó la Para ver mejor este fenómeno, mostramos los outliers de esta variable **transformación logarítmica natural** al precio de las viviendas, creando la variable **Log_SalePrice**



La comparación visual entre la distribución original y la transformada confirma el éxito de la normalización, un paso fundamental para la robustez de los modelos de regresión. El único inconveniente es que el modelo se entrena para predecir **Log_SalePrice**, por lo que los resultados finales deben deshacer este logaritmo aplicando la función exponencial para ser interpretados en la escala monetaria original

2.- Exploración de Variables Predictoras (EDA)



Se calculó la matriz de correlación de Pearson entre todas las variables numéricas para identificar las características más influyentes.

Como se observa en el mapa de calor, las variables con mayor correlación positiva con SalePrice son: *OverallQual*, GrLivArea y *GarageCars*.

La matriz también reveló instancias de **multicolinealidad**, donde algunas variables predictoras están altamente correlacionadas entre sí (ej., *GarageCars* y *GarageArea*). Esta detección justifica la posterior aplicación del **Análisis de Componentes Principales (PCA)** como una técnica de reducción de dimensionalidad para mitigar este problema.

Se generaron gráficos de dispersión (scatter plots) para las variables con mayor correlación con SalePrice para verificar la **linealidad** de la relación. En los gráficos de *GrLivArea*, *GarageArea* y *TotalBsmtSF*, se observaron **puntos atípicos** (*outliers*) extremos que podrían afectar significativamente el ajuste de los modelos de regresión, violando los supuestos básicos. Este paso de limpieza es fundamental para asegurar que las relaciones entre las variables sean lo más lineales posible, lo cual es esencial para las regresiones *Lasso* y *Ridge*.

3.- Gestión de Valores Ausentes y Duplicados

La fase de limpieza y preparación es esencial para gestionar la heterogeneidad de las variables y los datos faltantes, asegurando que el *dataset* sea apto para los modelos de regresión. Esta limpieza ha de hacerse en el conjunto de entrenamiento ya que si no podría haber el riesgo de **fuga de datos** (*data leakage*). Y más tarde se pasará esta limpieza se pasará con los datos del entrenamiento al resto de los subconjuntos de datos.

Hice una estrategia de imputación diferenciada, distinguiendo entre valores faltantes que representan una **ausencia física** de una característica y aquellos que representan un **fallo en la toma de datos**.

- **Imputación Estratégica (Ausencia = "No" o 0):** Para las variables cualitativas donde la falta de valor NA implica que la característica no existe (por ejemplo, ausencia de sótano, garaje o piscina), realicé una imputación categórica con el valor **"No"**. Esto se aplicó a variables como *Alley, BsmtQual, PoolQC, GarageType*, y *FireplaceQu*. Y un 0 en *GarageYrBlt*.
- **Imputación por Tendencia Central:** Para las variables numéricas con pocos valores faltantes que no indicaban ausencia, como *LotFrontage*, imputé el valor faltante con la **media** de la columna. Es importante señalar que, para mantener la **rigurosidad metodológica** y prevenir el riesgo de **fuga de datos** (*data leakage*), el valor de la media se calculó **únicamente** en el conjunto de **Entrenamiento**.

Posteriormente, ese valor fue aplicado para imputar los NA en los conjuntos de Validación y Test.

- **Eliminación de Filas:** Tras las imputaciones estratégicas, eliminé las filas restantes que contenían NA (que representaban fallos en la recolección o eran casos muy aislados) para garantizar la integridad del conjunto de datos. También se revisó y confirmó la inexistencia de filas duplicadas en el *dataset*.

Para corregir esta no linealidad y mejorar la robustez del modelo, procedí a eliminar los *outliers* de las variables predictoras más influyentes utilizando viendo las graficas los puntos que eran extremos y así poder eliminarlos poniendo unos límites.

4.- Codificación de variables Categóricas y Estandarización

Esta es la continuación lógica de la preparación de datos, asegurando que todas las variables predictoras sean aptas para los modelos lineales y **PCA**.

- **Codificación One-Hot y Ordinal:** Las variables nominales (sin orden) se transformaron mediante **One-Hot Encoding**, mientras que las variables ordinales (con jerarquía, ej., calidades) se mapearon a valores numéricos para preservar su orden.
- **Estandarización Z-score:** Apliqué la **estandarización Z-score** a todas las variables predictoras (numéricas y codificadas) para asegurar que ninguna característica dominara el proceso de modelado, siendo este un paso imprescindible antes de aplicar el **Análisis de Componentes Principales (PCA)**.

Implementación de Técnicas Multivariantes

1.- División de datos

Para garantizar la robustez del modelo y una estimación no sesgada del rendimiento, dividí el *dataset* preprocesado en tres subconjuntos disjuntos:

- **Entrenamiento (60%):** Utilizado para ajustar los parámetros de los modelos.
- **Validación (20):** Utilizado para el ajuste de hiperparámetros en los modelos *Lasso* y *Ridge* mediante validación cruzada (cv.glmnet).
- **Test ($20%):** Reservado para la evaluación final y robustez del modelo.

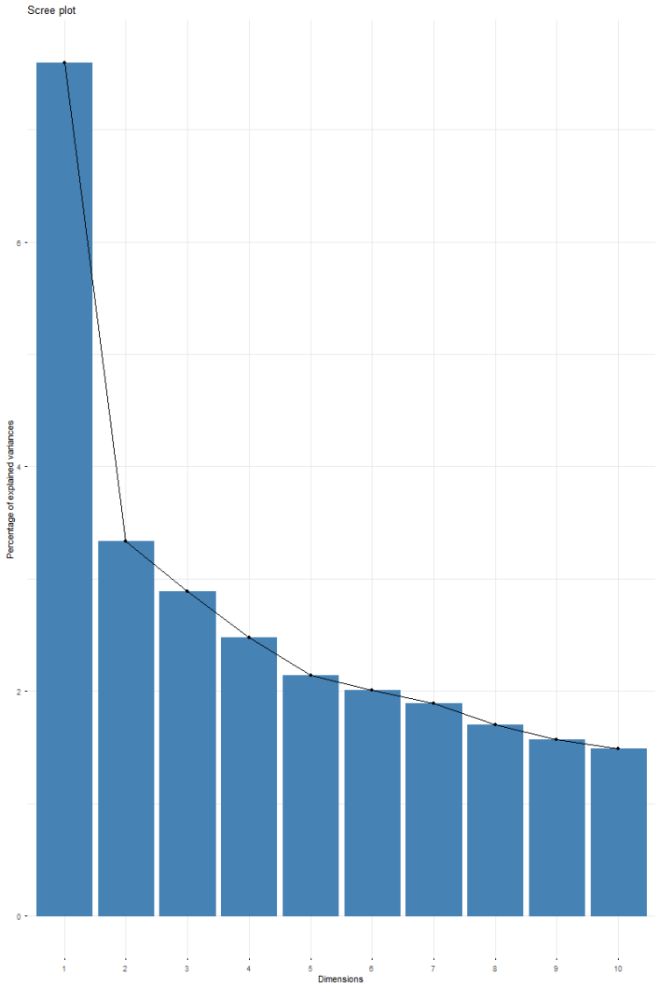
Antes de aplicar la limpieza de datos y el PCA, y dado que si no podría haber el riesgo de **fuga de datos** (*data leakage*), también realicé un **filtrado de variables con varianza cero** (variables que solo tenían un valor único), eliminando estas columnas de los tres conjuntos para evitar errores matemáticos.

2.- Análisis de Componentes Principales

El PCA es una técnica multivariante clave utilizada para reducir la dimensionalidad de nuestro conjunto de variables predictoras y mitigar la **multicolinealidad** previamente detectada en el EDA.

Aplicación: Ejecuté el PCA sobre el conjunto de entrenamiento estandarizado (Z-score), utilizando la función *prcomp*. Esto me generó **Componentes Principales (PC)**, que son

combinaciones lineales de las variables predictoras originales. Obtenemos tantos PC como variables predictoras tenía en el *dataset* original.



Selección de Componentes: a selección del número óptimo de componentes principales se basó en el criterio de la **varianza acumulada explicada**.

Para determinar cuántos Componentes Principales conservar, se suele examinar el *Scree Plot* (Gráfico de codo), que muestra la varianza explicada por cada dimensión:

Salto del Codo: es un método visual para elegir el número óptimo de Componentes Principales. Se observa el punto en la gráfica donde la pendiente de la varianza explicada cambia drásticamente, pasando de una caída empinada a una curva más plana. Observando el gráfico, se detecta un salto o "codo" significativo tras el **Primer Componente Principal** (PC1), y la curva se suaviza considerablemente.

Decisión Metodológica: La regla empírica del codo sugeriría retener solo PC1 o PC2. Sin embargo, la varianza individual de los primeros componentes es relativamente baja (PC1 explica menos del 8% de la varianza total). Si aplicáramos la regla del codo, retendríamos un porcentaje insuficiente de información, lo que llevaría a un modelo de regresión con baja capacidad explicativa.

Por esta razón, en lugar de la regla del codo, elegí el criterio de la **Varianza Acumulada**, reteniendo el número mínimo de componentes necesarios para explicar al menos el 85% de la información total del *dataset*.

- **Resultado:** Este criterio condujo a la selección de **91 Componentes Principales**, logrando una reducción de dimensionalidad robusta al pasar de unas 200 a 91 variables, y preservando la capacidad explicativa del modelo.

Transformación: Una vez que obtuve los PC del conjunto de entrenamiento, apliqué las rotaciones (las "fórmulas" del PCA) a los conjuntos de **Validación** y **Test** para transformarlos a la misma base dimensional, generando los conjuntos finales (*df_pca_train, df_pca_val, df_pca_test*) para el modelado.

Descripción y Creación de Modelos

1.- Fundamentos de la Regularización

Para abordar la complejidad inherente a un modelo con un gran número de variables predictoras (después de la codificación y PCA), utilicé técnicas de **regularización**. El objetivo principal de la regularización es **prevenir el sobreajuste** (*overfitting*) y **mejorar la capacidad de generalización** del modelo en datos que nunca ha visto.

La regularización funciona introduciendo un "castigo" o **penalización** al modelo cada vez que intenta hacer los coeficientes (pesos de las variables) demasiado grandes. Si un coeficiente es muy grande, significa que el modelo está demasiado apegado a esa variable en particular, lo que generalmente conduce al sobreajuste.

Este castigo se controla mediante un hiperparámetro llamado *lambda*. Una lambda más grande implica una penalización más fuerte y coeficientes más pequeños. Esta lambda fue optimizada mediante **validación cruzada** en el conjunto de validación.

Regresión Lasso (L1)

La Regresión Lasso aplica una penalización que se basa en el **valor absoluto de los coeficientes**.

- **¿Qué hace?** La penalización Lasso tiene una propiedad única: si la influencia de una variable es muy baja, puede forzar su coeficiente a ser **exactamente cero**.
- **Efecto:** Lasso realiza **selección automática de variables** (*feature selection*). Esto simplifica el modelo final, ya que elimina las características menos importantes, lo que facilita la interpretabilidad.

Regresión Ridge (L2)

La Regresión Ridge aplica una penalización que se basa en el **cuadrado de los valores de los coeficientes**.

- **¿Qué hace?** Ridge es excelente para manejar la **multicolinealidad** (cuando dos variables están altamente correlacionadas, como GarageCars y GarageArea). Su penalización es suave y distribuye el peso de las variables correlacionadas de manera más uniforme.
- **Efecto:** Reduce el tamaño de todos los coeficientes, acercándolos a cero, pero **sin llevarlos nunca exactamente a cero**. Esto significa que todas las variables predictoras (o Componentes Principales) permanecen en el modelo, aunque con una influencia reducida.

2.- Creación de los 5 modelos de regresión

Construí y evalué cinco modelos de regresión para comparar la efectividad de la reducción de dimensionalidad (PCA) frente a la regularización (Lasso/Ridge) y sus combinaciones:

Modelo	Variable Predictoras	Técnica Principal	Objetivos Primario
Regresión Lineal Múltiple	91 componentes Principales (PC)	Regresión sin regularización	Establecer el *baseline* del rendimiento post-PCA
Regresión Lasso (L1)	Variables Originales	Regularización	Selección de variables y reducción del *overfitting*.
Regresión Ridge (L2)	Variables Originales	Regularización	Mitigar la multicolinealidad y reducir el tamaño de coeficientes.
Regresión Lasso con PCA	91 componentes Principales (PC)	Regularización + PCA	Evaluar si la regularización mejora el rendimiento tras la reducción de dimensionalidad.
Regresión Ridge con PCA	91 componentes Principales (PC)	Regularización + PCA	Evaluar el efecto conjunto de ambas técnicas en la precisión final.

Los modelos de regularización fueron ajustados utilizando la librería glmnet con **validación cruzada** (cv.glmnet) sobre el conjunto de validación para determinar el hiperparámetro óptimo de penalización lambda

Modelo	Penalización	Hiperparámetro Óptimo	Df
Regresión Lasso	alpha=1	0.00527	Df=72
Regresión Ridge	alpha=0	0.1539	Df=200
Regresión Lasso + PCA	alpha=1	0.003963	Df=61
Regresión Ridge + PCA	alpha=0	0.03447	Df=91

Análisis de Lasso (L1): Es importante destacar que el valor **Df** (Degrees of Freedom) para los modelos Lasso indica cuántas variables predictoras mantienen un coeficiente distinto de cero. El modelo Lasso original logró la mayor selección de variables, reduciendo las 200 variables aproximadas originales a solo **72 coeficientes activos**, lo que demuestra su eficacia para simplificar la estructura del modelo.

Evaluación de Resultados

1.- Métrica de Rendimiento y Comparativa Final

La evaluación de la precisión y robustez de los modelos se realizó en el conjunto de **Test** mediante el **RMSE** (Error Cuadrático Medio de la Raíz) y el **MAE** (Error Absoluto Medio). Para la interpretación práctica, las predicciones fueron transformadas de vuelta a la escala original (SalePrice) aplicando la función exponencial (exp(x)).

Modelo	MAE_E	MAE_T	RMSE_E	RMSE_T	R^2_T	Robustez
Lineal con PCA	179.392	188.828	197.794	207.181	0,8997	Overfitting Moderado
Lasso (L1)	177.253	179.734	190.378	194.911	**0,9371**	Buena Generalización
Ridge (L2)	**176.945**	176.471	189.599	190.163	0,9300	Buena Generalización
Lasso con PCA	177.023	**172.397**	189.617	185.543	0,9098	Buena Generalización
Ridge con PCA	175.824	173.612	**187.134**	**185.601**	0,9130	Buena Generalización

2.- Análisis de Robustez y Generalización

Analicé la diferencia entre el error de Entrenamiento (E) y el error de Test (T) para evaluar la capacidad de generalización de cada modelo:

- **Regresión Ridge (L2)**: Es el modelo más **robusto** en términos de error absoluto, ya que presenta el **menor error en el conjunto de Test** (**MAE_T** de 176,471€), y una excelente capacidad explicativa (**R^2_T** de 0.9300). La brecha entre **MAE_E** (176,945€) y **MAE_T** (176,471€) es mínima, lo que indica una **Buena Generalización** (ausencia de *overfitting*).
- **Modelos de Regularización (Lasso y Ridge sin PCA)**: Estos modelos demuestran una buena robustez, ya que la brecha entre MAE_E y MAE_T es pequeña. La penalización lambda cumplió su función de mantener los coeficientes bajo control.
- **Lineal con PCA**: Muestra la **mayor diferencia** entre **MAE_E** (179,392€) y **MAE_T** (188,828€), lo que indica un **sobreajuste moderado**. A pesar de esto, su **R^2** en entrenamiento (0.9575) es alto, pero decae en test (0.8997), confirmando la dificultad para generalizar.
- **Lasso/Ridge con PCA**: Estos modelos ofrecieron el **menor error de Test (MAE_T de 172,397€)**. Esto sugiere que la combinación de reducción de dimensionalidad y regularización débil puede ser la estrategia más efectiva para predecir precios en la escala original.

3.- Selección y Discusión del Modelo Óptimo

Modelo Óptimo: Ridge con PCA

El modelo óptimo seleccionado es la Regresión Ridge con PCA debido a que presenta el menor Error Absoluto Medio en el conjunto de Test (MAE_T), que es la métrica más relevante para el rendimiento de negocio en la escala original.

- **Precisión Absoluta (MAE_T):** Con un **MAE_T de 173,612€,** este modelo (junto con Lasso+PCA) tiene el menor error promedio absoluto en la predicción del precio de venta, lo que es crucial para la valoración inmobiliaria.
- **Capacidad Explicativa (R^2_T):** El modelo explica el **91.30%** de la varianza de los precios, manteniendo una alta validez estadística en datos no vistos.
- **Robustez:** Mantiene una **Buena Generalización** (MAE_E: 175,824€ vs. MAE_T: 173,612€), lo que lo hace confiable para su uso en producción.

Conclusiones técnicas y posibles líneas de mejora

1.- Conclusiones Técnicas y Valoración Final

El objetivo principal del estudio era construir y comparar modelos robustos para la predicción de SalePrice.

- **Modelo Óptimo Seleccionado**: El modelo que ofreció el mejor compromiso entre precisión y robustez en la escala original fue la **Regresión Ridge aplicada sobre los 91 Componentes Principales**.
- **Rendimiento y Precisión**:
  - **Capacidad Explicativa**: El modelo explica el 91.30% de la varianza del precio (R^2_T), demostrando una alta validez estadística en datos no vistos.
  - **Precisión de Negocio (MAE)**: Con un MAE_T de **173,612€**, el error es aceptable para la valoración inmobiliaria, siendo el más bajo de todos.
- **Análisis Metodológico Central (PCA vs. Regularización)**: El paso más crítico para el rendimiento fue la aplicación del **PCA**, que logró simplificar la estructura del problema. La aplicación de la regularización Ridge tras el PCA mejoró marginalmente el MAE de Test, lo que indica que **eliminar la multicolinealidad vía PCA y luego aplicar una ligera penalización (Ridge) es la mejor estrategia**.

2.- Posibles Líneas de Mejora

Para incrementar la precisión y la robustez del modelo, se podrían explorar las siguientes líneas de trabajo:

- **Modelos No Lineales**: Si bien la regresión lineal funcionó bien (gracias a la transformación logarítmica), se podría probar el rendimiento de modelos no lineales más complejos como **Gradient Boosting** o **Random Forest**. Estos modelos no dependen de los supuestos de linealidad y podrían capturar relaciones que las regresiones Ridge pasaron por alto.

- **Ingeniería de Características (*Feature Engineering*)**: Crear nuevas variables que combinen información relevante, como el cálculo del precio por metro cuadrado habitable (SalePrice / GrLivArea) o una puntuación de calidad total, podría mejorar la capacidad predictiva.

- **Investigación de Outliers Amplificados**: Dado que el **RMSE_T (185,601€)** sigue siendo significativamente más alto que el **MAE_T (173,612€)**, el modelo aún comete algunos errores muy grandes debido a la transformación exponencial. Se podrían investigar los casos específicos donde el modelo falla drásticamente para identificar patrones atípicos y tratarlos de forma específica.