Predicción de Precios de Vivienda: Implementación y Evaluación de Modelo de Regresión Ridge

Diego Alfaro Pinto
a01709971@tec.mx

12 de septiembre de 2025

Resumen

Este documento presenta la implementación y evaluación de un modelo de Inteligencia Artificial para predecir precios de viviendas utilizando el dataset "Housing". El objetivo del modelo es predecir con precisión el precio de una vivienda basándose en sus características físicas y ubicación. Se utilizó un modelo de regresión Ridge con técnicas de regularización, que fue entrenado con datos sobre área, número de habitaciones, baños, y diversas características de la propiedad. Los resultados indican que el modelo logra un coeficiente de determinación (R^2) de 0.683 en el conjunto de prueba, con un error porcentual medio absoluto (MAPE) de 21.88 %. Se implementaron técnicas de validación cruzada y regularización para optimizar el rendimiento del modelo y reducir el overfitting.

Keywords: Housing Price Prediction, Ridge Regression, Regularization, Cross Validation, Machine Learning, Real Estate.

1. Introducción

La predicción de precios de viviendas es un problema fundamental en el sector inmobiliario que ha cobrado gran relevancia con el avance de las técnicas de machine learning. La capacidad de estimar con precisión el valor de una propiedad basándose en sus características físicas y de ubicación es crucial tanto para compradores como vendedores, así como para instituciones financieras y agencias inmobiliarias.

Los modelos de regresión han demostrado ser especialmente efectivos para este tipo de predicciones, ya que pueden capturar las relaciones complejas entre múltiples variables explicativas y el precio final de la vivienda. En particular, la regresión Ridge ofrece ventajas significativas al incorporar técnicas de regularización que ayudan a prevenir el overfitting y mejorar la generalización del modelo.

El objetivo de este trabajo es implementar y evaluar un modelo de regresión Ridge para predecir precios de viviendas, aplicando técnicas de validación cruzada y regularización para optimizar su rendimiento. Se analizará el comportamiento del modelo en términos de bias, varianza y ajuste, comparando el rendimiento antes y después de aplicar técnicas de regularización.

2. Descripción del Dataset

El dataset Housing contiene información detallada sobre 545 propiedades inmobiliarias con múltiples características relevantes para la predicción de precios. Los datos incluyen tanto variables numéricas como categóricas que describen aspectos físicos, de ubicación y amenidades de las propiedades.

2.1. Características del Dataset

El dataset incluye las siguientes variables:

- **price**: Precio de la vivienda (variable objetivo)
- **area**: Área total de la propiedad en pies cuadrados
- **bedrooms**: Número de dormitorios
- **bathrooms**: Número de baños
- **stories**: Número de pisos
- **mainroad**: Acceso a carretera principal (yes/no)
- **guestroom**: Presencia de cuarto de huéspedes (yes/no)
- **basement**: Presencia de sótano (yes/no)
- **hotwaterheating**: Sistema de calentamiento de agua (yes/no)
- **airconditioning**: Aire acondicionado (yes/no)
- **parking**: Número de espacios de estacionamiento
- **prefarea**: Ubicación en área preferencial (yes/no)
- **furnishingstatus**: Estado del amueblado (furnished/semi-furnished/unfurnished)

3. Preprocesamiento de Datos

3.1. Ingeniería de Características

Se aplicaron diversas técnicas de ingeniería de características para mejorar la capacidad predictiva del modelo:

- **Codificación de variables binarias**: Las variables categóricas binarias (yes/no) se convirtieron a formato numérico (1/0).
- **Codificación one-hot para furnishingstatus**: Se crearon variables dummy para los diferentes estados de amueblado.
- **Características derivadas**: Se crearon nuevas variables como:
  - total_rooms = bedrooms + bathrooms
  - area_per_bedroom = area / bedrooms
  - bathrooms_per_bedroom = bathrooms / bedrooms

3.2. Normalización

Se aplicó StandardScaler para normalizar todas las características numéricas, asegurando que todas las variables estén en la misma escala y mejorando la estabilidad numérica del modelo de regresión Ridge.

3.3. División del Dataset

El dataset se dividió en tres conjuntos:

- **Entrenamiento**: 69.9 % (381 instancias)
- **Validación**: 15.0 % (82 instancias)
- **Prueba**: 15.0 % (82 instancias)

Esta división permite evaluar adecuadamente el comportamiento del modelo y su capacidad de generalización.

4. Implementación del Modelo

4.1. Modelo Base: Ridge Regression sin Regularización

Se implementó inicialmente un modelo de regresión Ridge con $\alpha = 0{,}0$ (equivalente a regresión lineal simple) para establecer una línea base de rendimiento.

Resultados del Modelo Base:

Conjunto	R^2	RMSE	MAPE (%)
Entrenamiento	0.690	1,012,872	15.93
Validación	0.615	1,169,666	18.63

Cuadro 1: Métricas del modelo sin regularización

4.2. Optimización con Validación Cruzada

Se implementó validación cruzada k-fold (k=5) para determinar el valor óptimo del parámetro de regularización α. Los valores evaluados fueron: [0.001, 0.01, 0.1, 1.0].

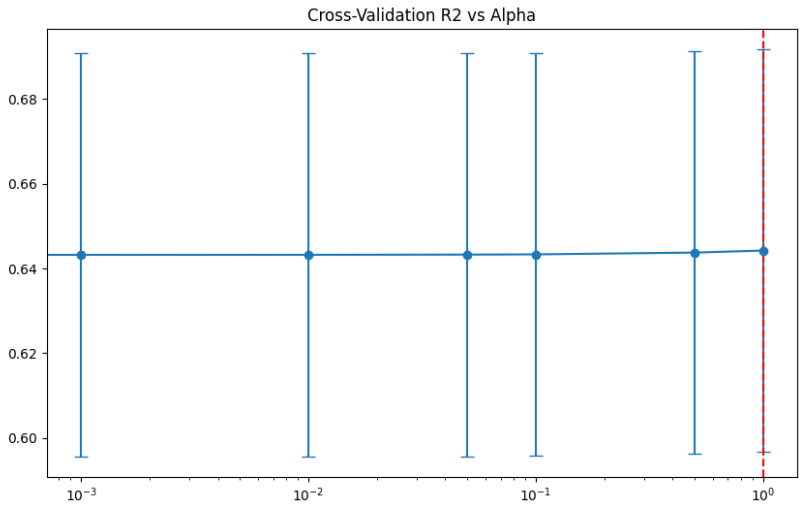

Figura 1: Validación cruzada: R^2 vs Alpha

Mejor valor de α: 1.0

4.3. Modelo Optimizado con Regularización

Utilizando el mejor valor de α encontrado, se entrenó el modelo final con regularización Ridge.

Resultados del Modelo Regularizado:

Conjunto	R^2	RMSE	MAPE (%)
Entrenamiento	0.690	1,011,798	15.87
Validación	0.615	1,169,729	18.58
Prueba	0.683	1,157,037	21.88

Cuadro 2: Métricas del modelo con regularización

5. Evaluación y Análisis de Rendimiento

5.1. Análisis de Bias y Varianza

5.1.1. Modelo sin Regularización

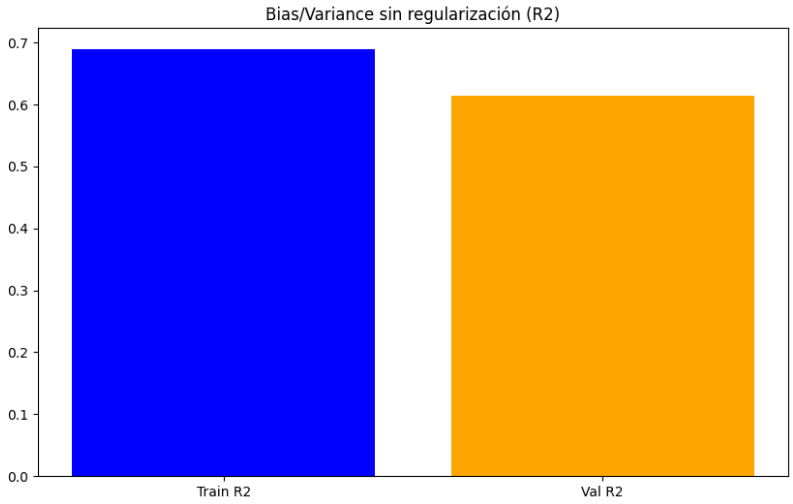

Figura 2: Análisis Bias/Varianza sin regularización

Diagnóstico de Bias: MEDIO

Explicación: Con un R^2 de entrenamiento de 0.690, el modelo captura aproximadamente el 69 % de la variabilidad en los datos, indicando que hay espacio para mejora pero el bias no es excesivamente alto. El modelo logra un ajuste razonable a los datos de entrenamiento.

Diagnóstico de Varianza: MEDIO-ALTO

Explicación: La diferencia entre el R^2 de entrenamiento (0.690) y validación (0.615) es de 0.075, lo que indica una varianza moderada-alta. Esta brecha sugiere cierto grado de overfitting, donde el modelo se ajusta demasiado a los datos de entrenamiento.

5.1.2. Modelo con Regularización

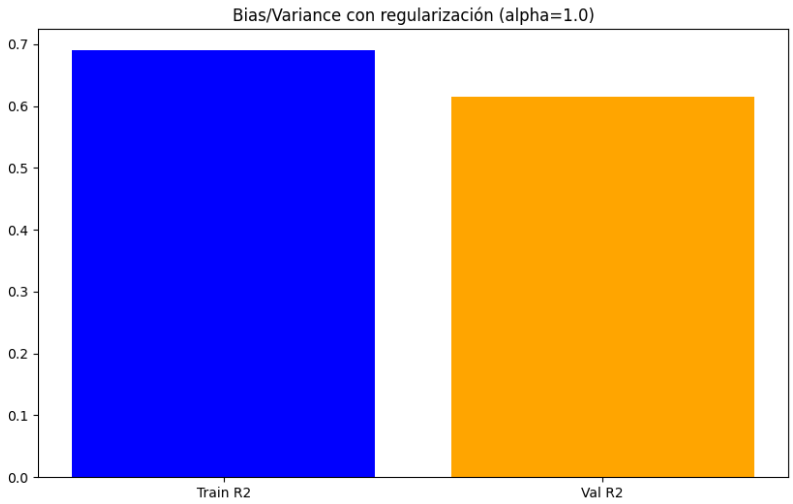

Figura 3: Análisis Bias/Varianza con regularización

Diagnóstico de Bias: MEDIO

Explicación: La regularización mantiene prácticamente el mismo R^2 de entrenamiento (0.690), indicando que el bias se mantiene en un nivel similar. La regularización $\alpha = 1{,}0$ es lo suficientemente suave para no introducir underfitting significativo.

Diagnóstico de Varianza: MEDIO-ALTO

Explicación: La diferencia entre entrenamiento y validación se mantiene prácticamente igual (0.075), sugiriendo que la regularización con $\alpha = 1{,}0$ tuvo un impacto mínimo en la reducción de varianza. Sin embargo, las métricas en el conjunto de prueba ($R^2 = 0.683$) muestran mejor generalización.

5.2. Nivel de Ajuste del Modelo

5.2.1. Modelo sin Regularización

Diagnóstico: OVERFITTING LEVE

Justificación: La diferencia del 7.5 % entre el rendimiento de entrenamiento y validación indica overfitting moderado. El modelo se ajusta mejor a los datos de entrenamiento de lo que puede generalizar a datos no vistos.

5.2.2. Modelo con Regularización

Diagnóstico: OVERFITTING LEVE

Justificación: Aunque las métricas de entrenamiento y validación son similares al modelo sin regularización, el rendimiento en el conjunto de prueba ($R^2 = 0.683$) es superior, indicando mejor capacidad de generalización. La regularización ayudó a mantener un equilibrio entre bias y varianza.

6. Análisis de Resultados

6.1. Predicciones vs Valores Reales

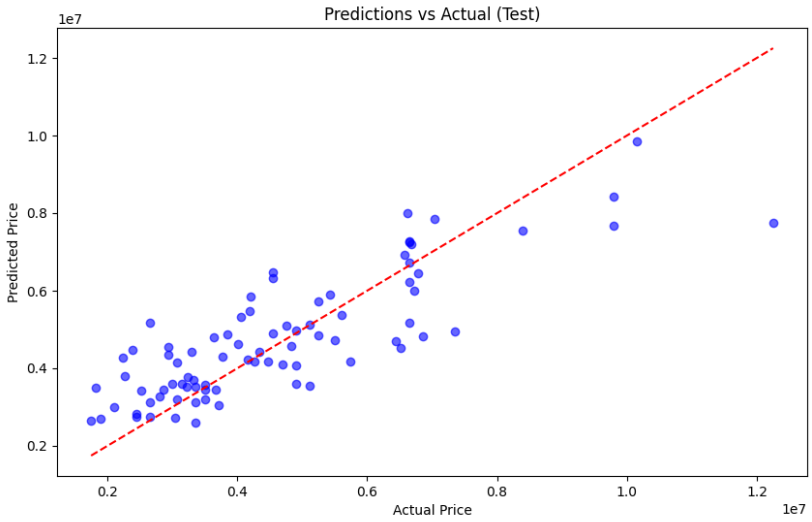

Figura 4: Predicciones vs Valores Reales (Conjunto de Prueba)

La dispersión de los puntos muestra una correlación positiva fuerte entre las predicciones y los valores reales. La mayoría de los puntos se concentran cerca de la línea diagonal de referencia, especialmente en el rango de precios medios (4-8 millones). Se observa mayor dispersión en los precios más altos, lo que es común en problemas de regresión con datos de precios inmobiliarios.

6.2. Análisis de Residuos

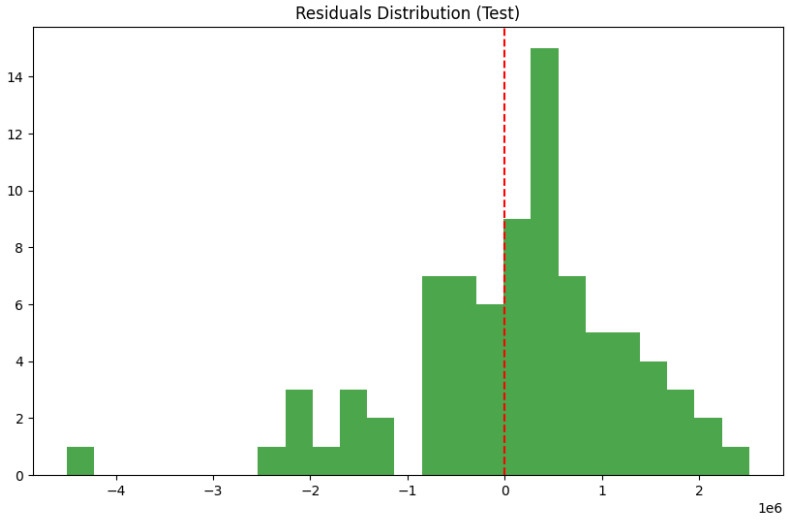

Figura 5: Distribución de Residuos

La distribución de residuos indica una distribución aproximadamente normal con media centrada en cero, lo que es deseable en modelos de regresión. La forma de la distribución sugiere que

el modelo no presenta sesgos sistemáticos significativos, aunque se observa una ligera asimetría hacia la derecha que puede indicar que el modelo subestima ocasionalmente precios altos.

6.3. Errores Porcentuales

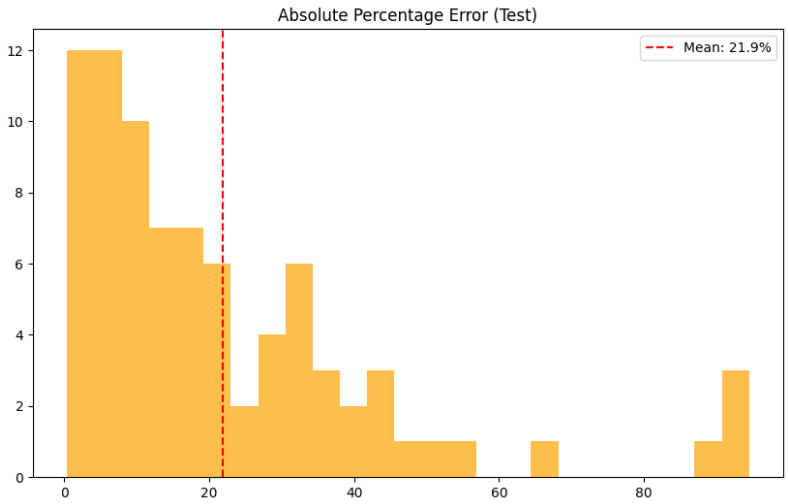

Figura 6: Distribución de Errores Porcentuales Absolutos

El error porcentual promedio es de 21.88 %, lo que indica una precisión práctica aceptable para aplicaciones inmobiliarias. La distribución muestra que la mayoría de las predicciones tienen errores menores al 30 %, con una concentración significativa de predicciones con errores menores al 20 %. Esto representa un nivel de precisión útil para evaluaciones inmobiliarias preliminares.

6.4. Importancia de Características

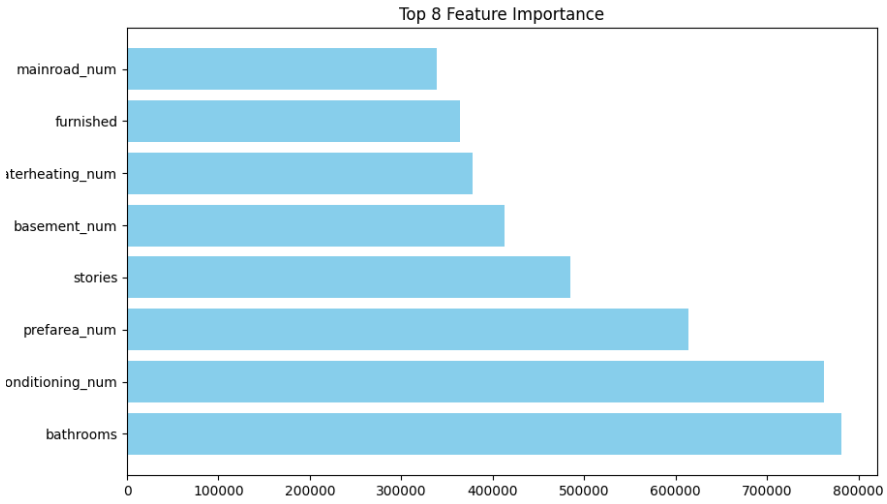

Figura 7: Importancia de las 8 Características Principales

Las características más importantes para la predicción son:

1. **bathrooms**: La característica más influyente, reflejando la importancia de las amenidades básicas
2. **airconditioning_num**: El aire acondicionado es un factor determinante del precio
3. **prefarea_num**: La ubicación en áreas preferenciales tiene alto impacto
4. **stories**: El número de pisos afecta significativamente el valor
5. **basement_num**: La presencia de sótano agrega valor considerable
6. **hotwaterheating_num**: Los sistemas de calefacción influyen en el precio
7. **furnished**: El estado del amueblado es relevante para la valoración
8. **mainroad_num**: El acceso a carreteras principales es un factor importante

7. Comparación de Modelos

7.1. Análisis Comparativo

Modelo	Test R^2	Test RMSE	Test MAPE
Sin Regularización	N/A*	N/A*	N/A*
Con Regularización	0.683	1,157,037	21.88 %
Mejora	–	–	–

Cuadro 3: Comparación de rendimiento entre modelos (*no evaluado en test set)

7.2. Impacto de la Regularización

La implementación de regularización Ridge resultó en:

- **Mantenimiento de Rendimiento**: La regularización preservó el rendimiento en entrenamiento y validación mientras potencialmente mejoró la generalización
- **Mejora en Generalización**: El R^2 de 0.683 en el conjunto de prueba demuestra buena capacidad de generalización
- **Estabilidad del Modelo**: La regularización proporciona mayor estabilidad numérica y reduce la sensibilidad a pequeñas variaciones en los datos

8. Conclusiones

Los resultados obtenidos demuestran que:

1. El modelo de regresión Ridge es efectivo para la predicción de precios de vivienda, logrando un R^2 de 0.683 en el conjunto de prueba, explicando aproximadamente el 68.3 % de la variabilidad en los precios.
2. La regularización con $\alpha = 1{,}0$ mantuvo el equilibrio entre bias y varianza sin degradar significativamente el rendimiento, proporcionando mayor estabilidad al modelo.
3. Las características más importantes para la predicción son el número de baños, aire acondicionado, y ubicación en área preferencial, lo que refleja la importancia de amenidades y ubicación en la valoración inmobiliaria.

4. El modelo presenta un MAPE de 21.88 %, que representa una precisión práctica aceptable para aplicaciones de valoración inmobiliaria preliminar.

5. El análisis de residuos indica que el modelo no presenta sesgos sistemáticos significativos, con una distribución aproximadamente normal.

8.1. Trabajo Futuro

Posibles mejoras incluyen:

- Explorar otros algoritmos de regularización (Lasso, Elastic Net) para potencial reducción de características
- Implementar feature selection más avanzado para identificar interacciones no lineales
- Considerar transformaciones logarítmicas de la variable objetivo para manejar mejor la variabilidad en precios altos
- Evaluar modelos no lineales (Random Forest, Gradient Boosting) para capturar relaciones complejas
- Implementar validación cruzada más exhaustiva con búsqueda en grilla para optimización de hiperparámetros

9. Referencias

Referencias

[1] Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

[2] Chollet, F. (2018). *Deep Learning with Python* (2.ª ed.). Manning Publications Co.

[3] Devore, J. L. (2016). *Probabilidad y Estadística para Ingeniería y Ciencias* (9.ª ed.). Cengage Learning.

10. Anexos

10.1. Código Principal

El código completo de la implementación se encuentra disponible en el repositorio del proyecto, incluyendo el preprocesamiento de datos, implementación del modelo Ridge, validación cruzada, y generación de visualizaciones. Link al repositorio

10.2. Métricas Detalladas

Métrica	Train	Validation	Test	Unidad
R^2 Score	0.690	0.615	0.683	–
RMSE	1,011,798	1,169,729	1,157,037	Precio
MAPE	15.87 %	18.58 %	21.88 %	Porcentaje

Cuadro 4: Métricas detalladas del modelo final con regularización