

TP1 - SIA: Predicción de precios de propiedades

Mediante modelos de regresión lineales y regularizados

Comisión B - Grupo 3

Integrantes:

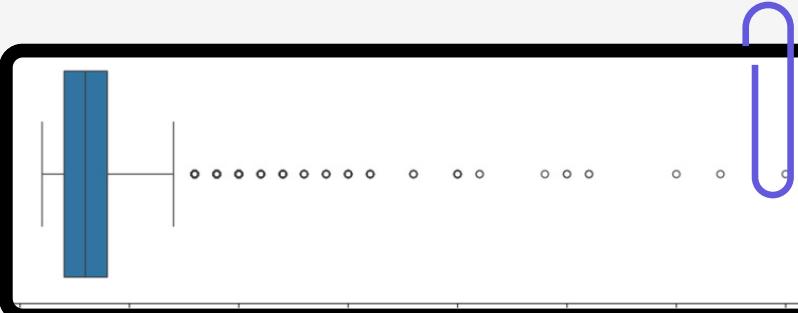
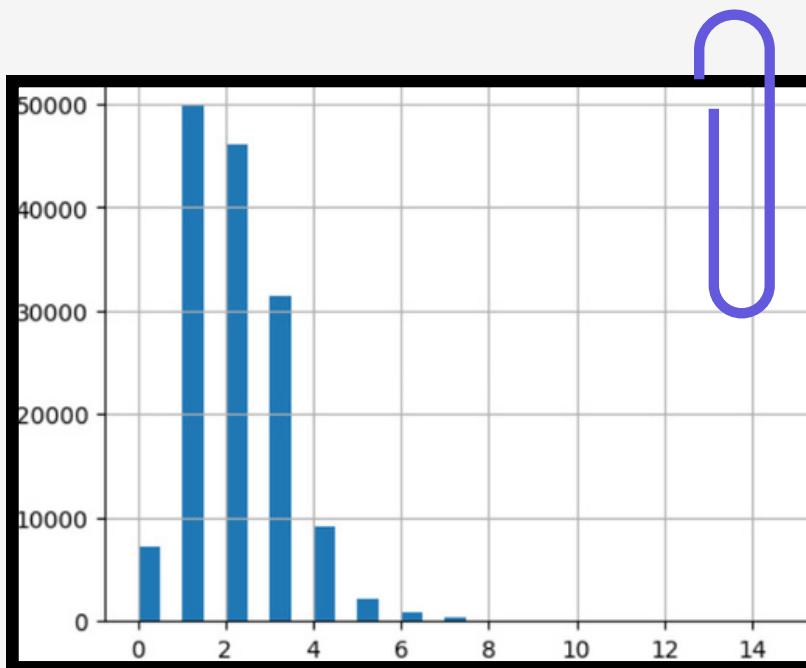
Julia Sexe (65669)

Sofia Alfie (64244)

Isidro Perasso (65595)

Tobias Tardá (65730)

Análisis exploratorio de datos



Dataset: 146.660 filas - 19 cols.

- -
- Valores faltantes** en 5 cols.
Máximo 15% aprox.
Solución: **imputación** → post sacar outliers

Histogramas: 'rooms' y 'bedrooms' 🔍 var. predictora
⚠ Afectados por faltantes

- Outliers:** 87.150 y 37.688 severos
- Solución: **eliminación** de 24.222 filas con severos
 - 'bedrooms', 'bathrooms' y 'rooms' entre 0,05 y 0,33% de severos
→ 🔍 var. predictoras

Cols. categóricas: 'I2' con 4 y 'property_type' con 10 valores únicos

- Correlaciones:** 'price' se relaciona con:
- 'surface_covered', 'surface_total', 'bathrooms' y 'bedrooms'
 - ⚠ alta colinealidad entre 'surface_covered' y 'surface_total'

Preparación de datos para modelos de ML

'surface_covered', 'l2_Bs.A.
este', 'l2_Bs.As. G.B.A.
'property_type_Casa',
'ampo', 'property_type_Co
nto', 'property_type_Dep
ercial', 'property_type_I
'property_type_Otros']



Eliminación de cols.: de las 19 nos quedamos con 5:

- 2 categóricas: 'l2' y 'property_type'
- 3 numéricas: 'rooms', 'bathrooms' y 'surface_covered'
- 1 target: 'price'

Criterios: **alta colinealidad, variables con poca utilidad o que aportarían info. repetida**

Codificación de var. categóricas: son nominales → OneHotEncoder.

- 4 cols. nuevas con categorías de 'l2'
- 10 cols. nuevas con categorías de 'property_type'

Imputación de faltantes: 22.298 nulos

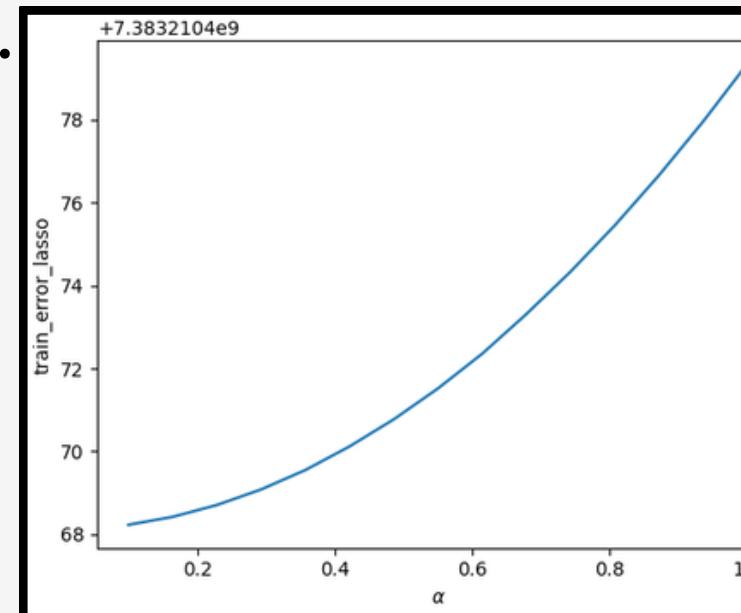
- **Estrategia:** previo agrupamiento en una categoría
- 'surface_covered': agrupación por 'rooms' y relleno con **mediana**
- 'bathrooms': agrupación por 'bedrooms' y relleno con **moda**

Separación en train y test + estandarización

Entrenamiento de modelos

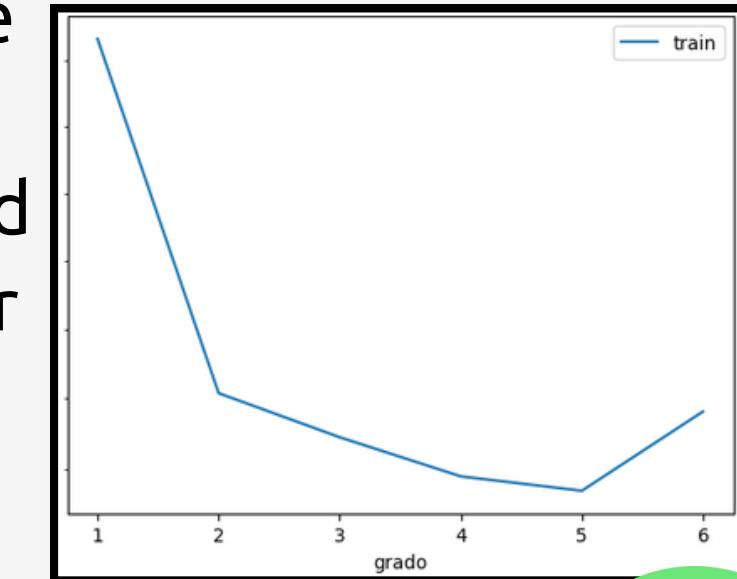
1. Regresión Lineal Múltiple

- Utilizamos las 17 variables
- Uso de Pipeline para normalizar y entrenar
- Cross validation con 5 folds
- **RMSE promedio: 78.095,2040**



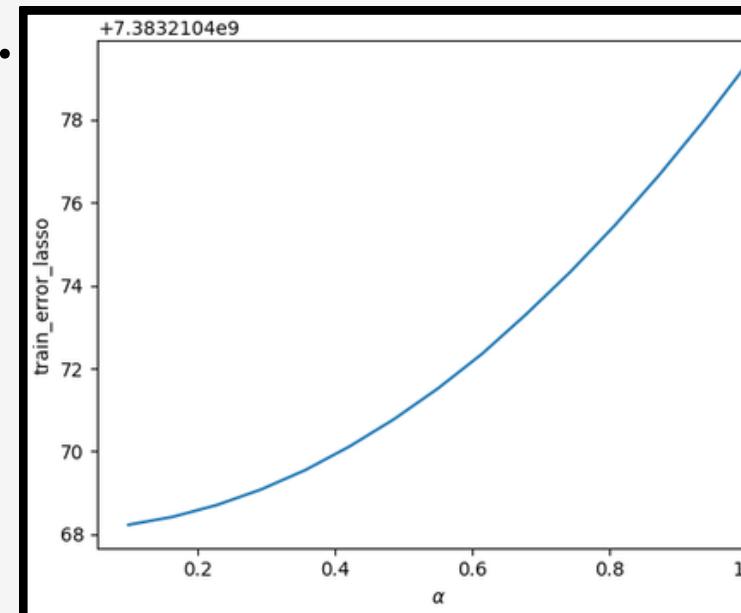
2. Regresión Polinomial

- Sacamos las variables que representan a 'property_type' → lentitud
- Exploración: mínimo error en test con grado 5
- **RMSE promedio: 74.691,4752**



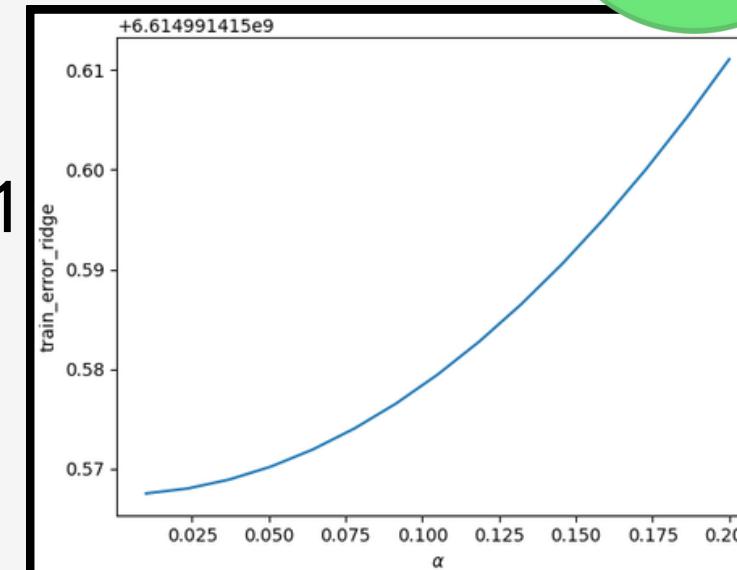
3. Lasso Regression

- Utilizamos las 3 var. num.
- Exploración de alpha: ¿cómo varía el error? \rightarrow elección alpha = 0,1
- Polinomio grado 4
- **RMSE promedio: 82.156,3417**



4. Ridge Regression

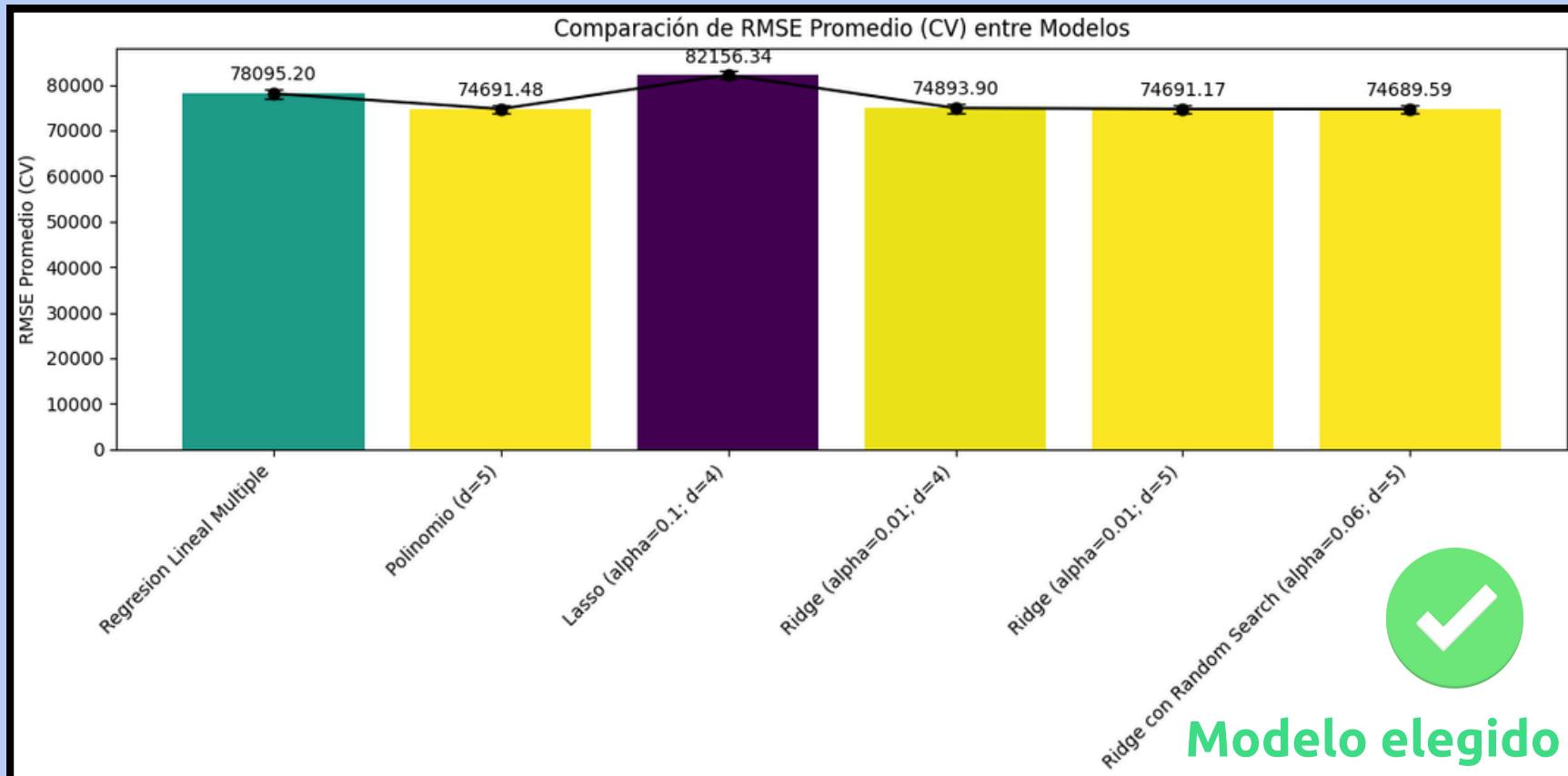
- Mismas variables que polinómica
- Exploración \rightarrow alpha = 0,01
- Polin. grado 4 RMSE promedio: 74.893,9008
- Grado 5 RMSE promedio: 74.691,1695



Ajuste fino

Modelo Ridge

- RandomizedSearch → 20 combinaciones, 5 folds
- Mejor combinación de hiperparámetros: **grado = 5 y alpha = 0,0634**
- Mejor **RMSE (CV)**: **74.689,5872**



Testeo

Elección: Ridge de grado = 5 y alpha ≈ 0,06

- Entrenamiento con todo el set de train
- Predicciones en test → **RMSE: 74.241,14**
- Variable + influyente: '**surface_covered**', en potencias y combinación con 'rooms' y 'bathrooms'

