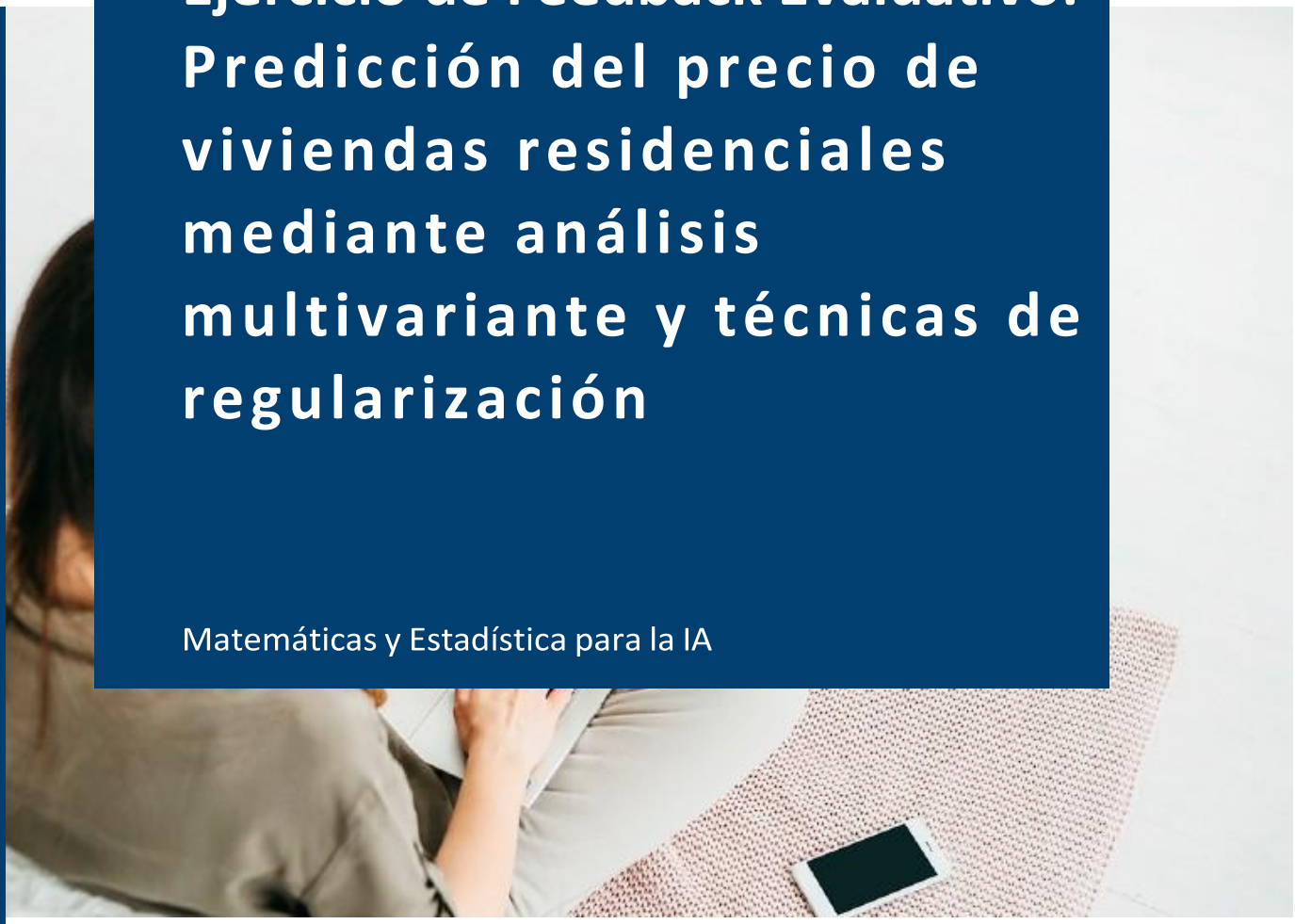
Ejercicio de Feedback Evaluativo: Predicción del precio de viviendas residenciales mediante análisis multivariante y técnicas de regularización



Matemáticas y Estadística para la IA

Para aplicar los conocimientos adquiridos hasta la Unidad 3, realizaremos un ejercicio de feedback, con la resolución de diferentes supuestos prácticos.

1. Descripción del caso

En el archivo de datos disponible, cada registro corresponde a una vivienda e incorpora información detallada sobre sus características físicas, estructurales y de contexto urbano. Entre las columnas, ***SalePrice*** representa el valor de venta de la vivienda y será la variable objetivo a predecir mediante modelos estadísticos y de machine learning. Las restantes columnas, exceptuando los campos puramente identificativos (como "*Id*"), serán utilizadas como variables predictoras.

2. Objetivos

- Realizar un análisis exploratorio y descriptivo de las variables presentes en el conjunto de datos, identificando patrones, tendencias y relaciones relevantes con el precio de venta (SalePrice).
- Preprocesar adecuadamente los datos: tratamiento de valores ausentes, codificación de variables categóricas y normalización o estandarización de las variables predictoras si fuera necesario.
- Dividir el conjunto de datos en subconjuntos de entrenamiento, validación y test, utilizando técnicas apropiadas de validación cruzada para estimar el rendimiento de los modelos y evitar sobreajuste.
- Construir y comparar los siguientes modelos predictivos:
    1. Regresión lineal múltiple tras una reducción de dimensionalidad mediante análisis de componentes principales (PCA).
    2. Regresión con regularización Lasso.
    3. Regresión con regularización Ridge.
    4. Regresión Lasso y Ridge aplicadas sobre los componentes principales extraídos por PCA.
- Evaluar la precisión, robustez y capacidad de generalización de cada modelo mediante métricas adecuadas (por ejemplo, RMSE y MAE).
- Redactar un informe técnico profesional, detallando metodología, justificación estadística/matemática y conclusiones orientadas a la interpretación práctica de los resultados.

3. Dataset

El dataset disponible (*train.csv*) contiene información estructurada sobre viviendas residenciales, con cada fila representando una unidad habitacional y cada columna aportando detalles sobre distintos aspectos relevantes para la valoración inmobiliaria. Una ampliación descriptiva de las variables la podrá encontrar en el archivo adjunto *data_description.txt.*

- **Características principales del dataset**

  - **SalePrice**: Variable objetivo, representa el valor de venta final de cada vivienda (registro) en unidades monetarias.
  - Variables predictoras: Más de 70 columnas que describen atributos físicos, estructurales, de ubicación y calidad:

    1. **SalePrice** - precio de venta de la propiedad en dólares. Es la variable objetivo que quieres predecir.
    2. **MSSubClass**: clase del edificio.
    3. **MSZoning**: clasificación general de la zona.
    4. **LotFrontage**: metros lineales de calle que conectan con la propiedad.
    5. **LotArea**: tamaño de la parcela en pies cuadrados.
    6. **Street**: tipo de acceso por carretera.
    7. **Alley**: tipo de acceso por callejón.
    8. **LotShape**: forma general de la parcela.
    9. **LandContour**: nivel de planitud del terreno.
    10. **Utilities**: tipo de servicios disponibles.
    11. **LotConfig**: configuración de la parcela.
    12. **LandSlope**: pendiente del terreno.
    13. **Neighborhood**: ubicaciones físicas dentro de los límites de la ciudad de Ames.
    14. **Condition1**: proximidad a carretera principal o vía de tren.
    15. **Condition2**: proximidad a carretera principal o vía de tren (si hay una segunda).
    16. **BldgType**: tipo de vivienda.
    17. **HouseStyle**: estilo de la vivienda.
    18. **OverallQual**: calidad global de materiales y acabados.
    19. **OverallCond**: valoración global del estado de la vivienda.
    20. **YearBuilt**: año de construcción original.
    21. **YearRemodAdd**: año de reforma.
    22. **RoofStyle**: tipo de tejado.
    23. **RoofMatl**: material del tejado.
    24. **Exterior1st**: revestimiento exterior principal de la casa.
    25. **Exterior2nd**: revestimiento exterior secundario de la casa (si hay más de un material).
    26. **MasVnrType**: tipo de revestimiento de mampostería.
    27. **MasVnrArea:** superficie de revestimiento de mampostería en pies cuadrados.
    28. **ExterQual**: calidad del material exterior.
    29. **ExterCond**: estado actual del material exterior.
    30. **Foundation**: tipo de cimentación.
    31. **BsmtQual**: altura del sótano.
    32. **BsmtCond**: estado general del sótano.
    33. **BsmtExposure**: paredes de sótano con salida al exterior o nivel jardín.
    34. **BsmtFinType1**: calidad del área de sótano terminada (tipo 1).
    35. **BsmtFinSF1**: superficie terminada tipo 1 en pies cuadrados.

36. **BsmtFinType2**: calidad de la segunda zona terminada del sótano (si existe).

37. **BsmtFinSF2**: superficie terminada tipo 2 en pies cuadrados.
38. **BsmtUnfSF**: superficie sin terminar del sótano en pies cuadrados.
39. **TotalBsmtSF**: superficie total del sótano en pies cuadrados.
40. **Heating**: tipo de calefacción.
41. **HeatingQC**: calidad y estado de la calefacción.
42. **CentralAir**: aire acondicionado central.
43. **Electrical**: sistema eléctrico.
44. **1stFlrSF**: superficie de la primera planta en pies cuadrados.
45. **2ndFlrSF**: superficie de la segunda planta en pies cuadrados.
46. **LowQualFinSF**: superficie terminada de baja calidad (todas las plantas) en pies cuadrados.
47. **GrLivArea**: superficie habitable sobre rasante (a nivel de suelo o superior) en pies cuadrados.
48. **BsmtFullBath**: número de baños completos en el sótano.
49. **BsmtHalfBath**: número de aseos (medio baño) en el sótano.
50. **FullBath**: número de baños completos sobre rasante.
51. **HalfBath**: número de aseos (medio baño) sobre rasante.
52. **Bedroom**: número de dormitorios por encima del nivel del sótano.
53. **Kitchen**: número de cocinas.
54. **KitchenQual**: calidad de la cocina.
55. **TotRmsAbvGrd**: número total de habitaciones sobre rasante (no incluye baños).
56. **Functional**: valoración de la funcionalidad de la vivienda.
57. **Fireplaces**: número de chimeneas.
58. **FireplaceQu**: calidad de la chimenea.
59. **GarageType**: ubicación del garaje.
60. **GarageYrBlt**: año en que se construyó el garaje.
61. **GarageFinish**: acabado interior del garaje.
62. **GarageCars:** capacidad del garaje en número de coches.
63. **GarageArea**: superficie del garaje en pies cuadrados.
64. **GarageQual**: calidad del garaje.
65. **GarageCond**: estado del garaje.
66. **PavedDrive**: entrada de vehículos pavimentada.
67. **WoodDeckSF**: superficie de la terraza de madera en pies cuadrados.
68. **OpenPorchSF**: superficie del porche abierto en pies cuadrados.
69. **EnclosedPorch**: superficie del porche cerrado en pies cuadrados.
70. **3SsnPorch**: superficie del porche de tres estaciones en pies cuadrados.
71. **ScreenPorch**: superficie del porche con mosquitera en pies cuadrados.
72. **PoolArea**: superficie de la piscina en pies cuadrados.
73. **PoolQC**: calidad de la piscina.
74. **Fence**: calidad de la valla.
75. **MiscFeature**: característica adicional no incluida en otras categorías.
76. **MiscVal**: valor en dólares de la característica adicional.
77. **MoSold**: mes de venta.
78. **YrSold:** año de venta.
79. **SaleType**: tipo de venta.
80. **SaleCondition**: condición de la venta.

- **Tipos de variables**
  - **Numéricas continuas**: LotArea, GrLivArea, SalePrice, YearBuilt, GarageArea, TotalBsmtSF, entre otras.
  - **Numéricas discretas**: OverallQual, OverallCond, Fireplaces, BedroomAbvGr, GarageCars.
  - **Categóricas**: MSZoning, Street, Neighborhood, HouseStyle, GarageType, SaleCondition.
  - **Identificadores**: "Id" (No debe usarse como predictor).

- **Observaciones**
  - El dataset presenta valores faltantes en algunas variables (por ejemplo, PoolQC, Alley) que requieren tratamiento específico durante el preprocesamiento.
  - Las variables cubren atributos estructurales, contexto urbano/social y calidad, lo que permite realizar análisis multivariante y explorar relaciones complejas para la predicción de precios.
  - La riqueza y variedad de las columnas lo hacen ideal para aplicar análisis exploratorio, reducción de dimensionalidad y técnicas avanzadas de regularización en regresión.
  - Este conjunto refleja la complejidad real de un problema de predicción de precios inmobiliarios y facilita el desarrollo de modelos robustos y comparativos.

4. Recomendaciones para el desarrollo

- **Variable objetivo:** Deja muy claro que deberás predecir la columna SalePrice utilizando las restantes variables, salvo identificadores.
- **Análisis exploratorio:** Emplea visualizaciones (diagramas de dispersión, histogramas, boxplots, mapas de calor de correlaciones) para examinar la distribución de SalePrice y su relación con las variables más importantes.
- **Tratamiento de datos:** Identifica y gestiona cualquier dato ausente (.NA) mediante imputación o eliminación justificada. Codifica variables categóricas de forma adecuada (por ejemplo, one-hot encoding).
- **Normalización/Estandarización:** Considera aplicar escalado a las variables predictoras, especialmente relevante antes de usar PCA, Lasso o Ridge.
- **Reducción de dimensionalidad:** Aplica PCA y selecciona el número óptimo de componentes principales en función de la varianza explicada y la interpretación del modelo.
- **Regularización:** Ajusta los hiperparámetros de regularización para Lasso y Ridge usando la validación cruzada del conjunto de entrenamiento/validación. Compara cómo cada técnica afecta la selección de variables y el rendimiento predictivo.
- **División de datos:** Separa correctamente tus datos en entrenamiento, validación y test (ejemplo típico: 60%-20%-20%). Usa validación cruzada para obtener estimaciones más robustas.
- **Evaluación:** Compara los modelos con métricas cuantitativas (RMSE, MAE, MSE) y visualiza las predicciones frente a los valores reales de SalePrice.
- **Documentación y reflexión:** Redacta un informe que explique puntualmente los pasos seguidos, las decisiones metodológicas tomadas y el valor interpretativo de los resultados. Incluye recomendaciones para la aplicación práctica del modelo en valoración de viviendas.
- **Presentación:** Usa visualizaciones claras y tablas de resumen para la interpretación de resultados. El informe debe tener una presentación técnica y profesional acorde al trabajo solicitado.

5. ¿Por qué es interesante este Dataset?

Este dataset es especialmente interesante y valioso porque refleja la complejidad real del mercado inmobiliario y representa una oportunidad para aplicar técnicas de análisis predictivo y machine learning sobre datos multivariantes que incluyen información física, estructural, urbana y de calidad de las viviendas.

- **Razones por las que es interesante**

  - **Aplicabilidad profesional:** Permite entrenar modelos de valoración automática de viviendas, una tarea con enorme demanda y relevancia práctica en sectores como banca, inmobiliarias, urbanismo y plataformas tecnológicas.
  - **Riqueza y variedad de variables:** Posee 80 características categóricas y numéricas, lo que facilita la exploración de relaciones complejas y multimodales entre atributos y el precio final de las propiedades.
  - **Escenario ideal para aprendizaje:** Es perfecto para experimentar con todas las fases del análisis de datos moderno: EDA, técnicas de limpieza y transformación, reducción de dimensionalidad, y comparación de modelos de regresión convencional y regularizada.
  - **Impacto social y económico:** Modelar correctamente el precio de las viviendas ayuda a familias, empresas e instituciones en la toma de decisiones informadas de compra, venta e inversión.
  - **Complejidad realista:** El dataset presenta datos reales, con valores faltantes, variables mixtas y escenarios de multicolinealidad, simulando los desafíos profesionales que se enfrentan en proyectos de automáticos de clasificación inmobiliaria.
  - **Comparabilidad y benchmarking**: Facilita la comparación objetiva de diferentes modelos y técnicas estadísticas, permitiendo evaluar interprerabilidad, capacidad predictiva y robustez frente a las características particulares de los datos.

Analizar este dataset posibilita construir modelos predictivos robustos y transferibles, aprender prácticas avanzadas de preprocesamiento y valorar el impacto de la inteligencia artificial en mercados dinámicos como el inmobiliario.

6. Entregables

- Script reproducible y bien comentado, que permita seguir paso a paso el análisis, el preprocesamiento, el desarrollo de modelos y la evaluación de resultados (*50% de la nota*).

- Informe profesional en PDF (máximo 10 páginas), con alto énfasis en formulaciones matemáticas y estadísticas, incluyendo (*50% de la nota*):
  - Introducción y contextualización del problema (*10%*).
  - Fases de análisis exploratorio y justificación del preprocesamiento (*20%*).
  - Descripción, implementación y comparación de los modelos requeridos (*20%*).
  - Evaluación de resultados y discusión fundamentada (*25%*).
  - Conclusiones técnicas y posibles líneas de mejora (*25%*).

7. Recursos complementarios de apoyo

- **Valores faltantes**
  - https://epirhandbook.com/es/new_pages/missing_data.es.html
  - https://youtu.be/xZophFn6tkA?si=0c1xvh-fqYZbP7t_
- **Codificación de variables categóricas**
  - https://librovivodecienciadedatos.ai/preparacion-de-datos.html
  - https://rpubs.com/jboscomendoza/vairables_dummy_con_r
  - https://keepcoding.io/blog/codificacion-de-variables-categoricas/
  - https://www.youtube.com/watch?v=byMkc2swr8s

8. Referencias

- https://github.com/Praveen76/House-Prices--Advanced-Regression-Techniques
- https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

G R A C I A S

UAX.COM