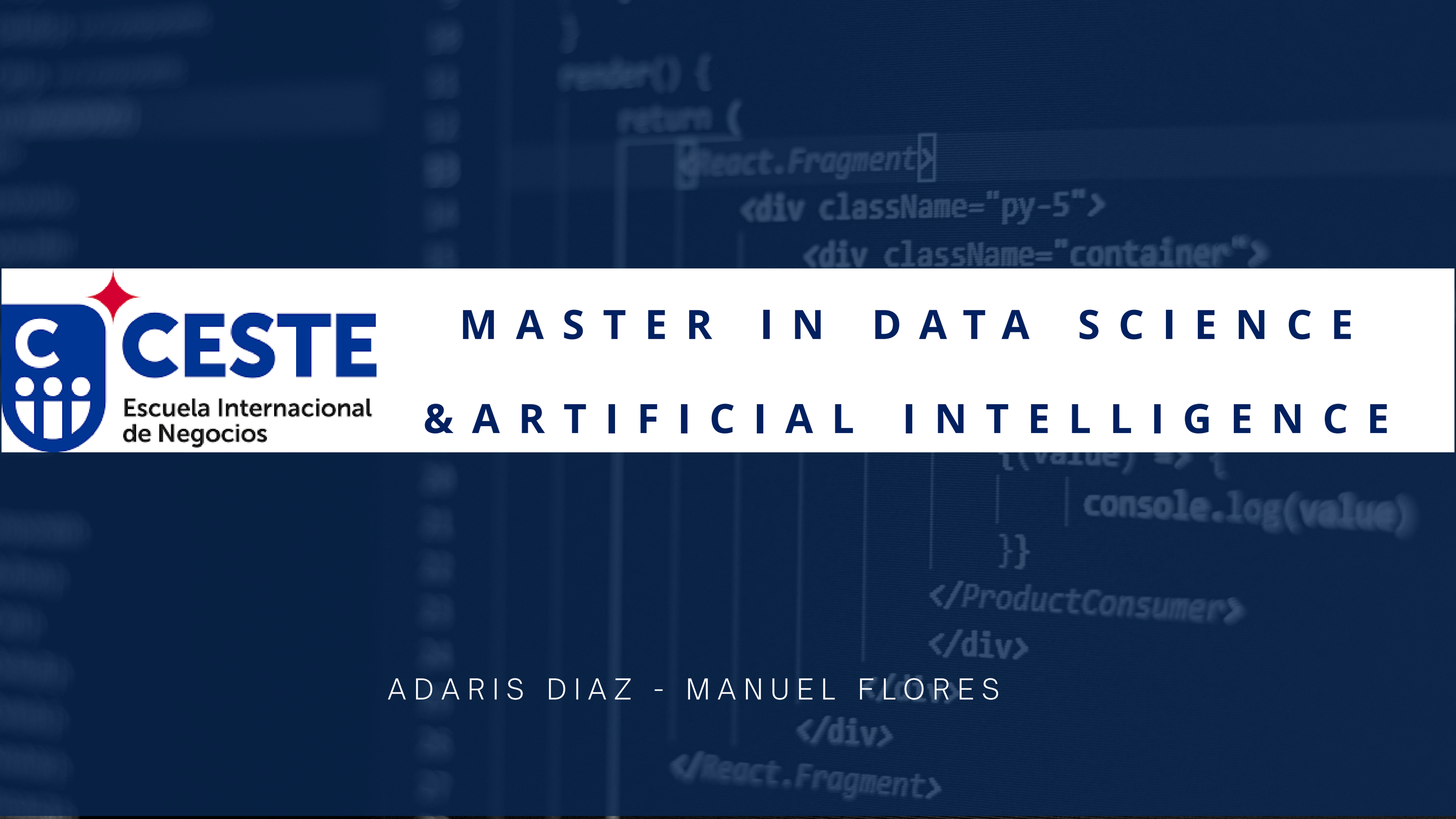


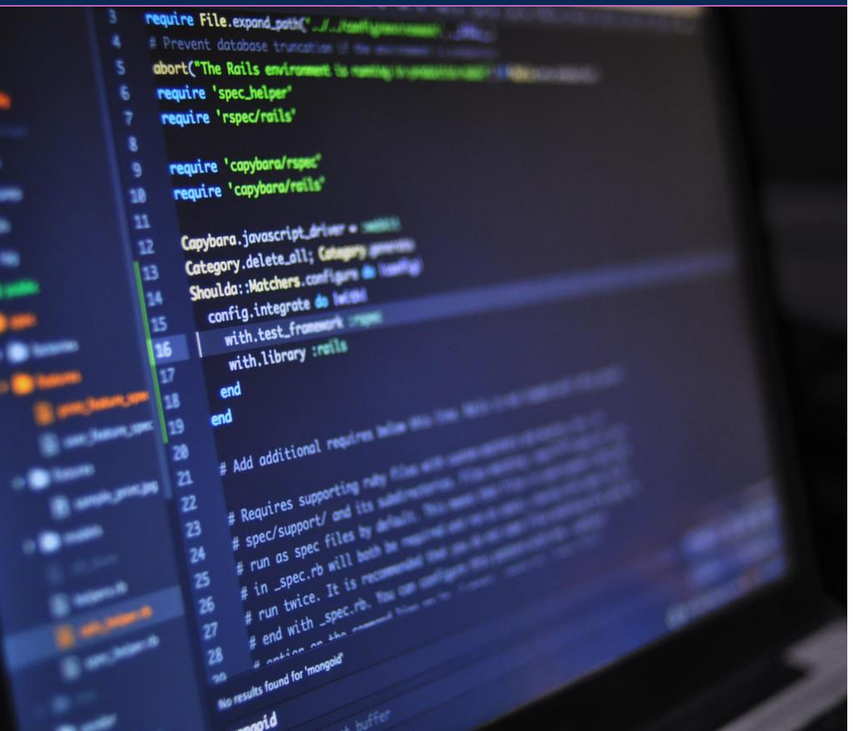
CESTE
Escuela Internacional de Negocios

MASTER IN DATA SCIENCE & ARTIFICIAL INTELLIGENCE

ADARIS DIAZ - MANUEL FLORES

HOUSE PRICES - ADVANCED REGRESSION TECHNIQUES

HOUSE PRICES - ADVANCED REGRESSION TECHNIQUES



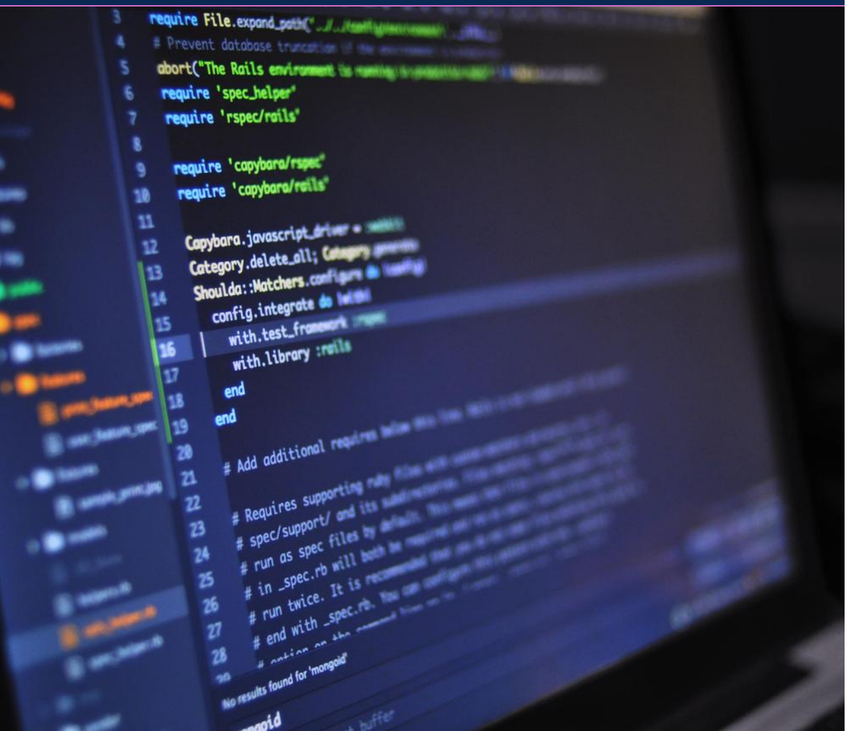
OBJETIVO: PREDECIR PRECIOS DE VIVIENDAS

DATASET: 79 VARIABLES EXPLICATIVAS (NUMÉRICAS/CATEGÓRICAS)

VARIABLE OBJETIVO: PRECIO DE VENTAS (SALESPRICE)

MÉTRICA: Root-Mean-Squared-Error (RMSE)

FLUJO DEL PROYECTO



ANÁLISIS EXPLORATORIO DE DATOS (EDA)

PREPROCESAMIENTO Y LIMPIEZA

INGENIERÍA DE CARACTERÍSTICAS

MODELADO Y EVALUACIÓN

OPTIMIZACIÓN Y SUMISIÓN

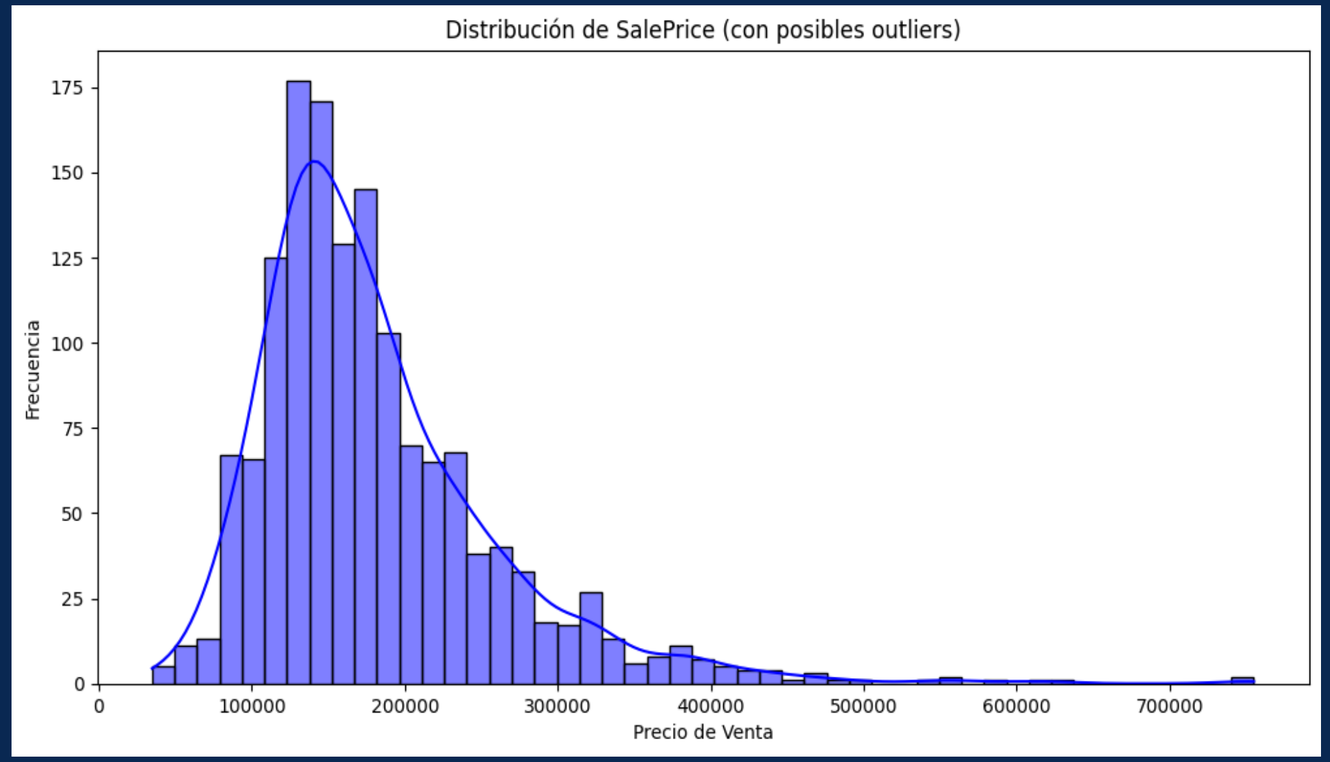
ANÁLISIS EXPLORATORIO (EDA)

DISTRIBUCIÓN DE SALEPRICE (SESGADA → NECESIDAD DE TRANSFORMACIÓN LOGARÍTMICA)

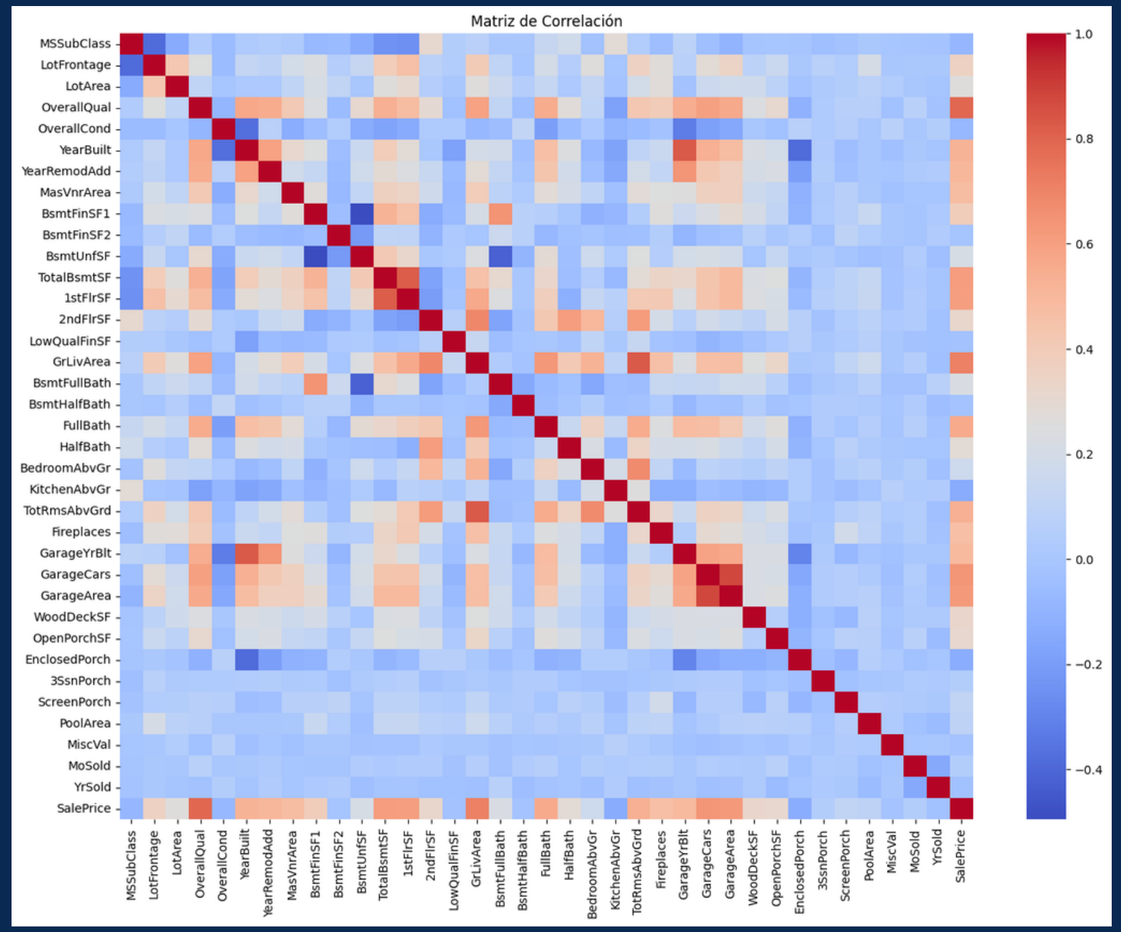
CORRELACIÓN CON VARIABLES COMO GRLIVAREA, OVERALLQUAL

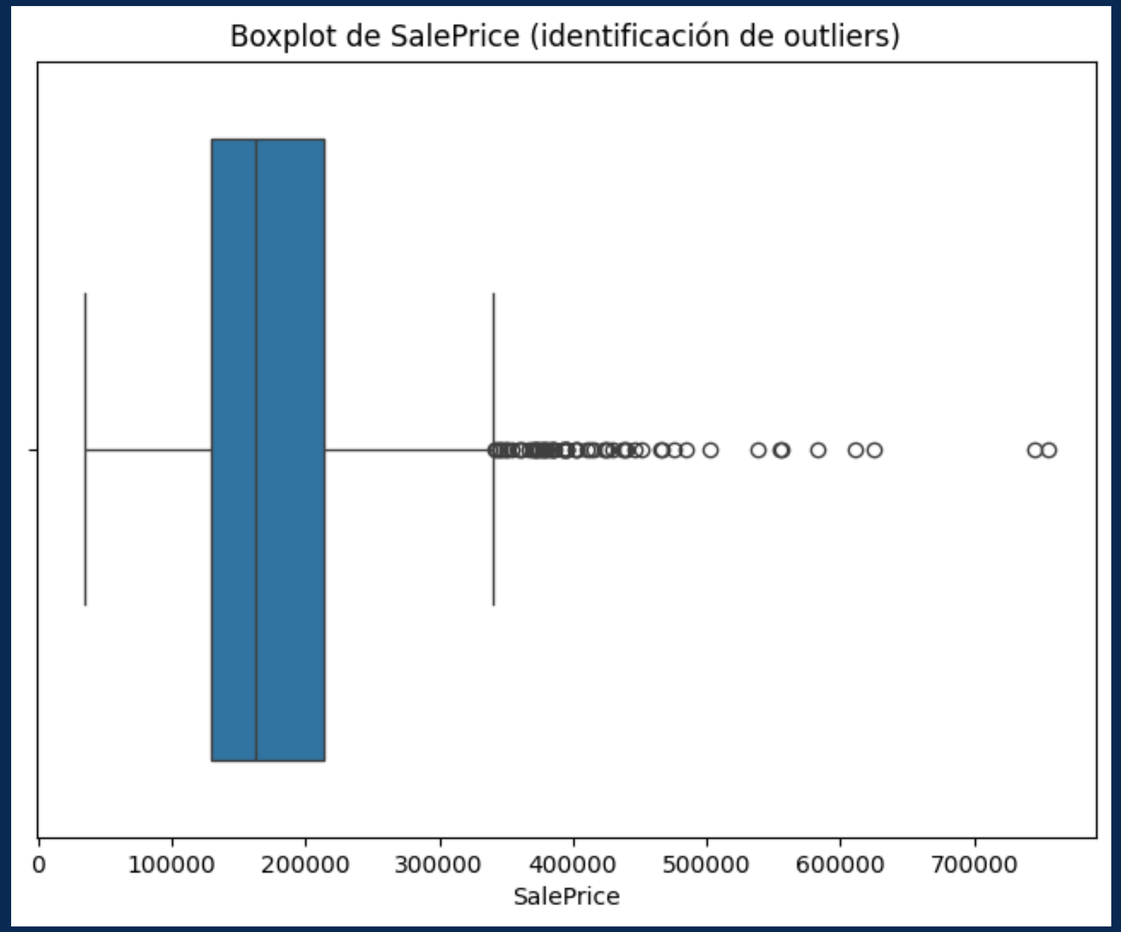
RELACIÓN CON CATEGÓRICAS: NEIGHBORHOOD, HOUSESTYLE (IMPACTO EN PRECIO MEDIANO).

OUTLIERS: DETECTADOS EN PRECIOS ALTOS

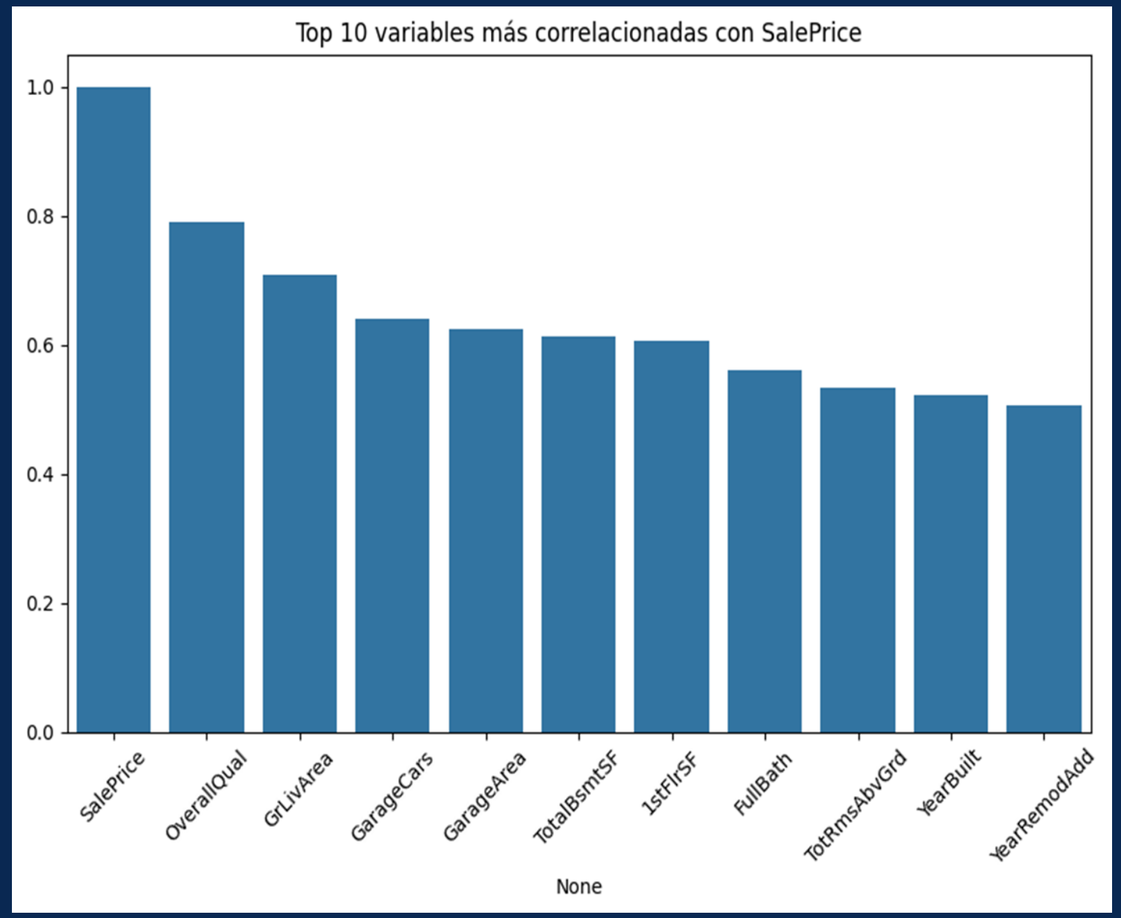


ANÁLISIS EXPLORATORIO (EDA)





PREPROCESAMIENTO DE DATOS



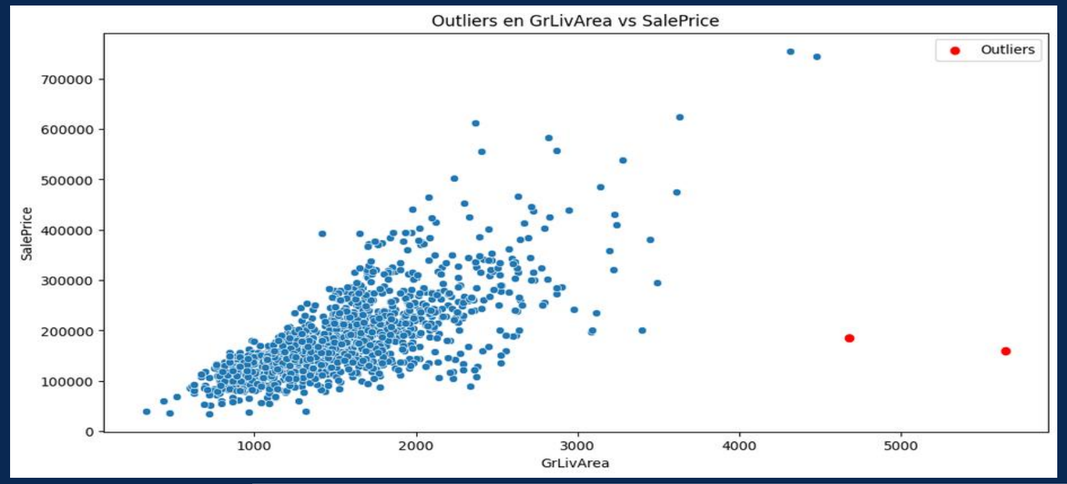
Transformaciones Aplicadas

- Log-transform en SalePrice (para cumplir con métrica RMSE logarítmica).
- Eliminación de la columna Id.

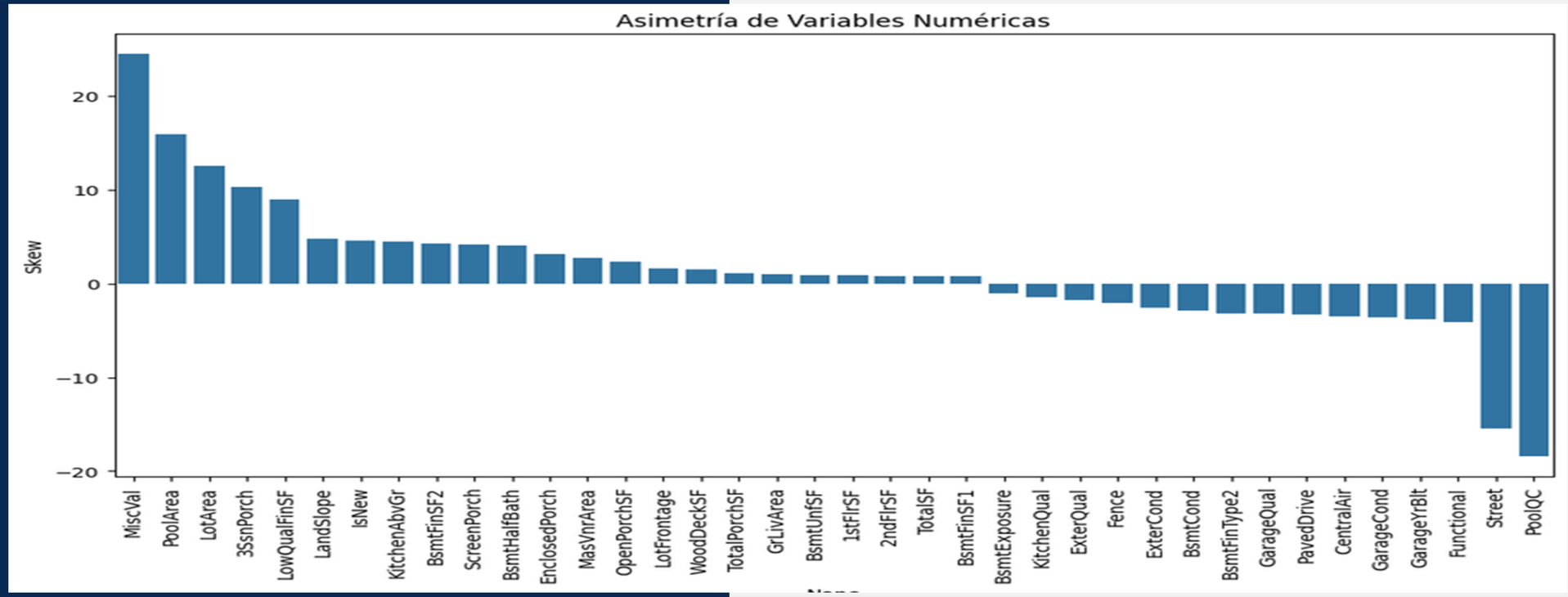
Tratamiento de datos faltantes

- LotFrontage: Rellenado con mediana.

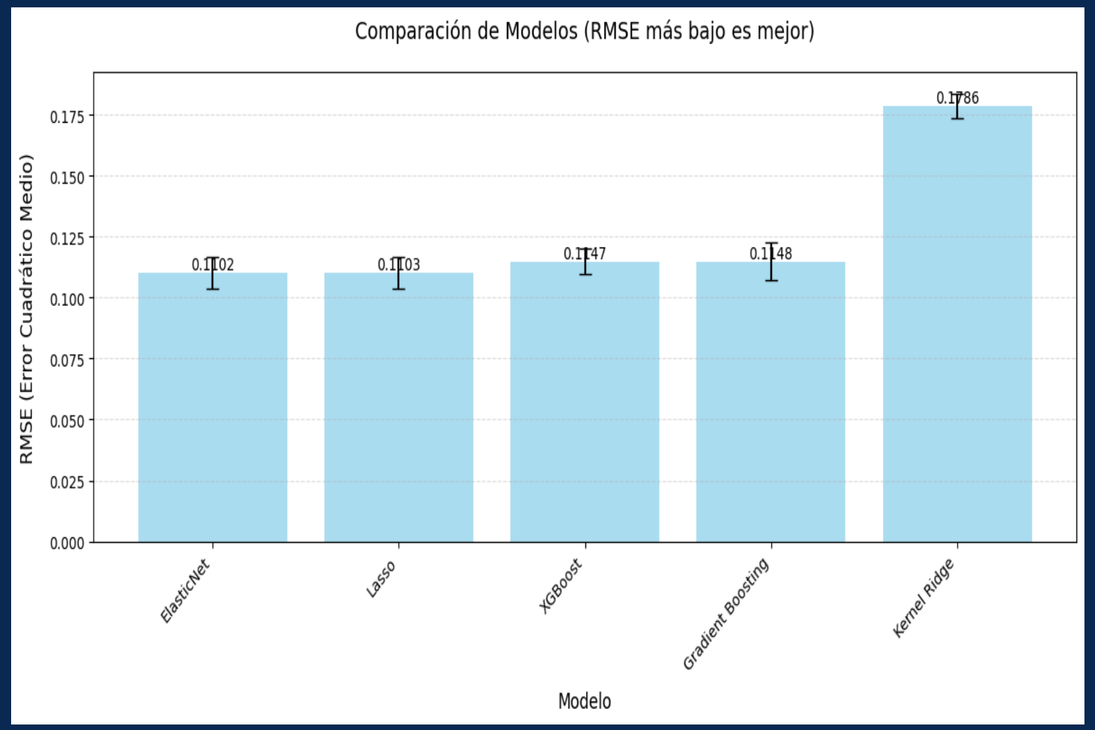
INGENIERÍA DE CARACTERÍSTICAS



- ✓ Creación de nuevas variables: TotalSF, TotalBath, TotalPorchSF, HouseAge, IsNew, Remodeled
- ✓ Agrupamiento de vecindarios por precio mediano (NeighborhoodBinned)
- ✓ Conversión de variables numéricas a categóricas (como MSSubClass, OverallCond)
- ✓ Uso de LabelEncoder para variables ordinales
- ✓ One-hot encoding para categóricas restantes



MODELADO



LASSO REGRESSION (REGULARIZACIÓN L1)

ELASTICNET (COMBINACIÓN L1 Y L2)

RANDOM FOREST

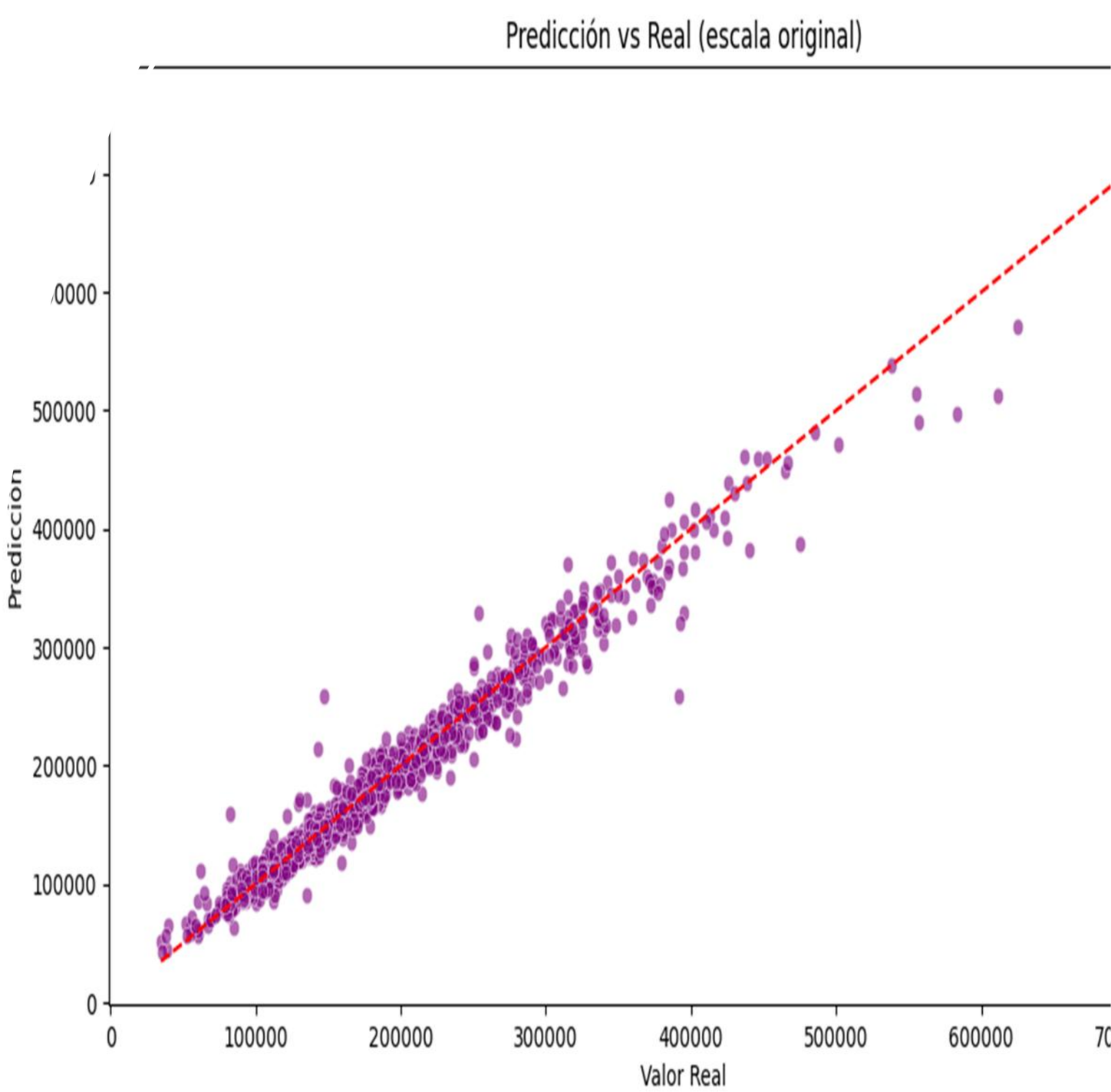
GRADIENT BOOSTING (XGBOOST).

OPTIMIZACIÓN Y RESULTADOS

Técnicas empleadas:

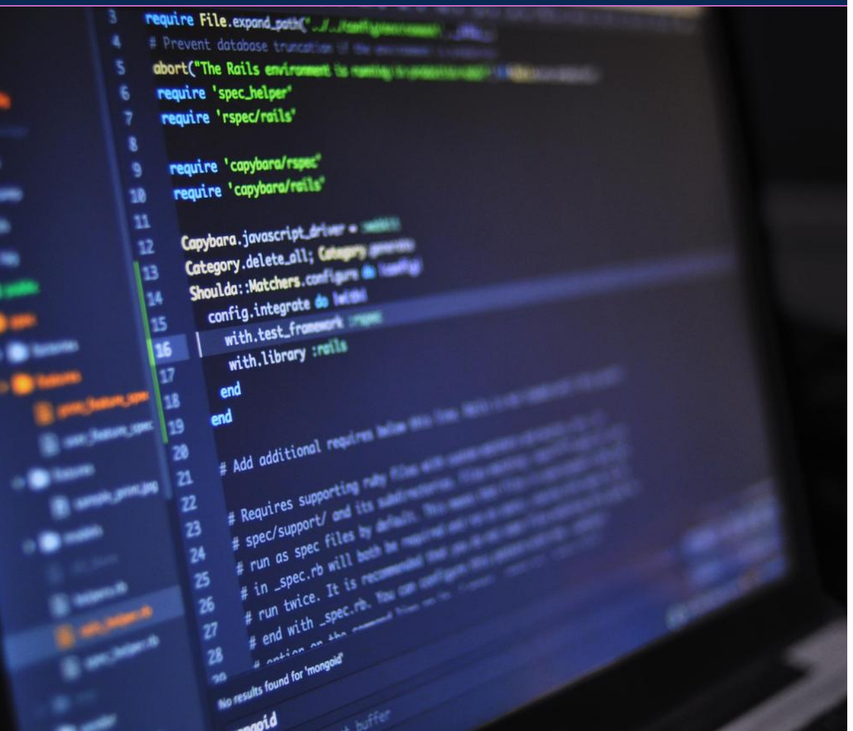
- GridSearchCV para hiperparámetros
- Validación cruzada (KFold)
- Resultado: Score 0.12143 (391 de 4736 participantes)

391	ADARIS DIAZ		0.12143	2	14d



STORE SALES - TIME SERIES FORECASTING

HOUSE PRICES - ADVANCED REGRESSION TECHNIQUES



OBJETIVO: PRONOSTICAR VENTAS DE: “LA CORPORACIÓN FAVORITA”

DATASET: DATOS DE VENTAS SEMANALES DE MÚLTIPLES TIENDAS (TRAIN.CSV, TEST.CSV)

VARIABLE OBJETIVO:
VENTAS (SALES)

VARIABLES CLAVES:

FECHA (DATE)

FAMILIA DE PRODUCTOS (FAMILY)

VENTAS (SALES)

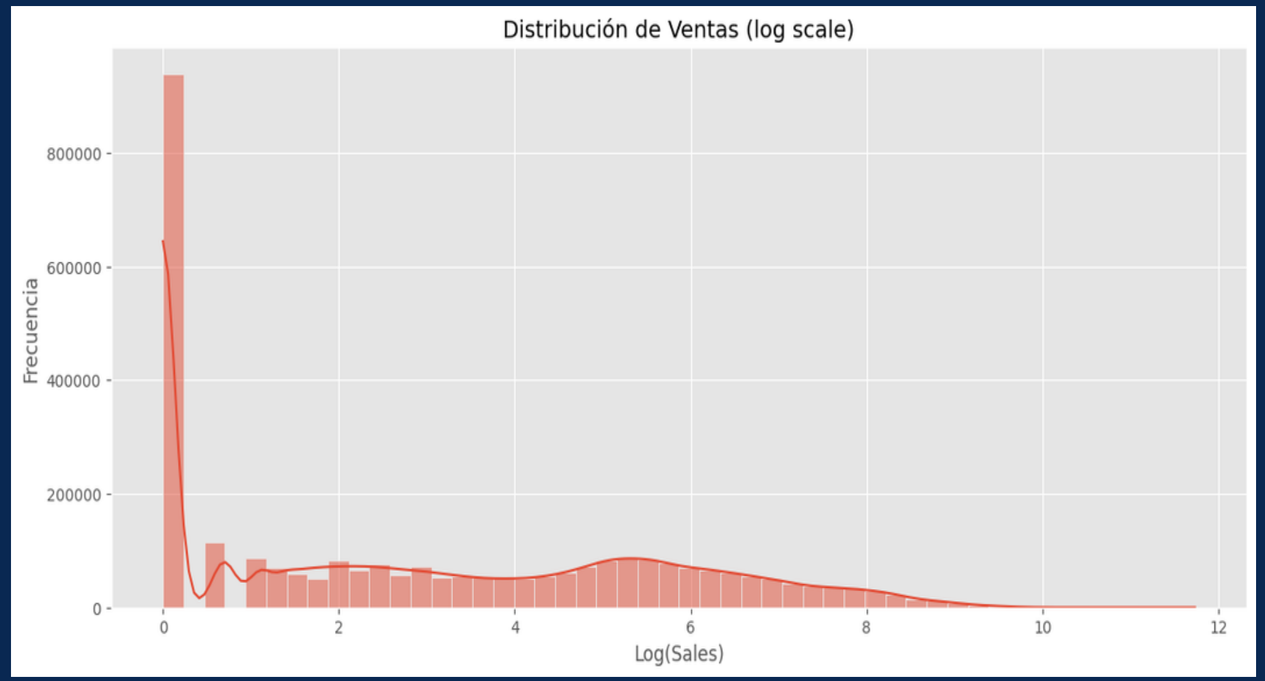
PROMOCIONES (ONPROMOTION)

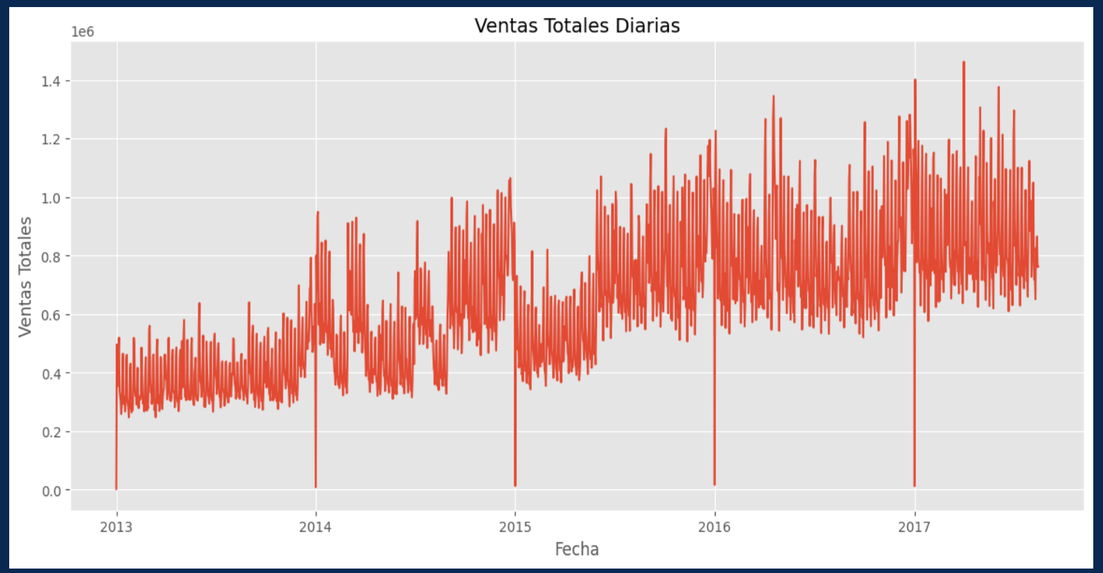
DATOS EXTERNOS COMO PRECIOS DEL PETRÓLEO (OIL.CSV) Y DÍAS FESTIVOS (HOLIDAYS_EVENTS.CSV)

MÉTRICA: RMSLE (ROOT MEAN SQUARED LOGARITHMIC ERROR)

ANÁLISIS EXPLORATORIO (EDA)

- DISTRIBUCIÓN DE VENTAS: SESGO POSITIVO QUE REQUIRIÓ TRANSFORMACIÓN LOGARÍTMICA.
- TENDENCIA TEMPORAL: CLARA ESTACIONALIDAD ANUAL Y SEMANAL.
- CORRELACIÓN ENTRE VARIABLES
- IDENTIFICACIÓN DE VALORES FALTANTES

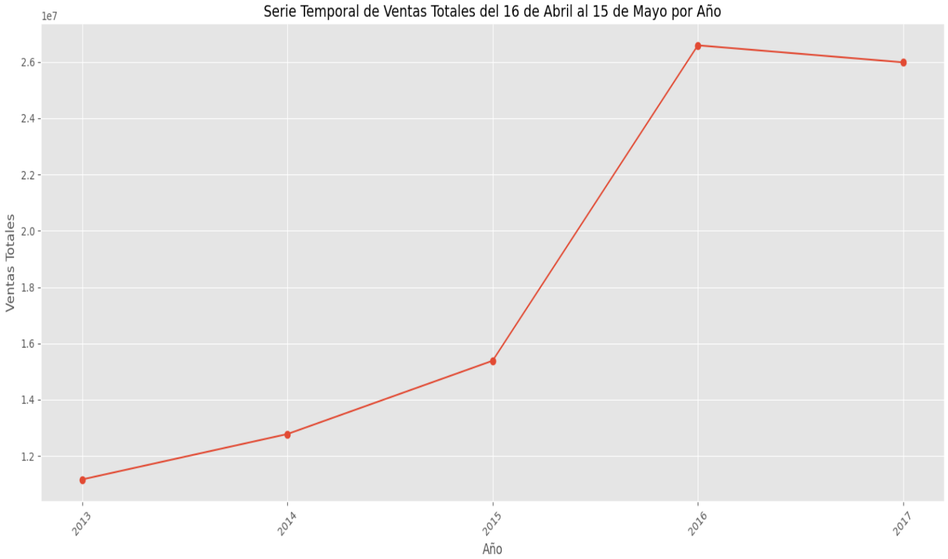


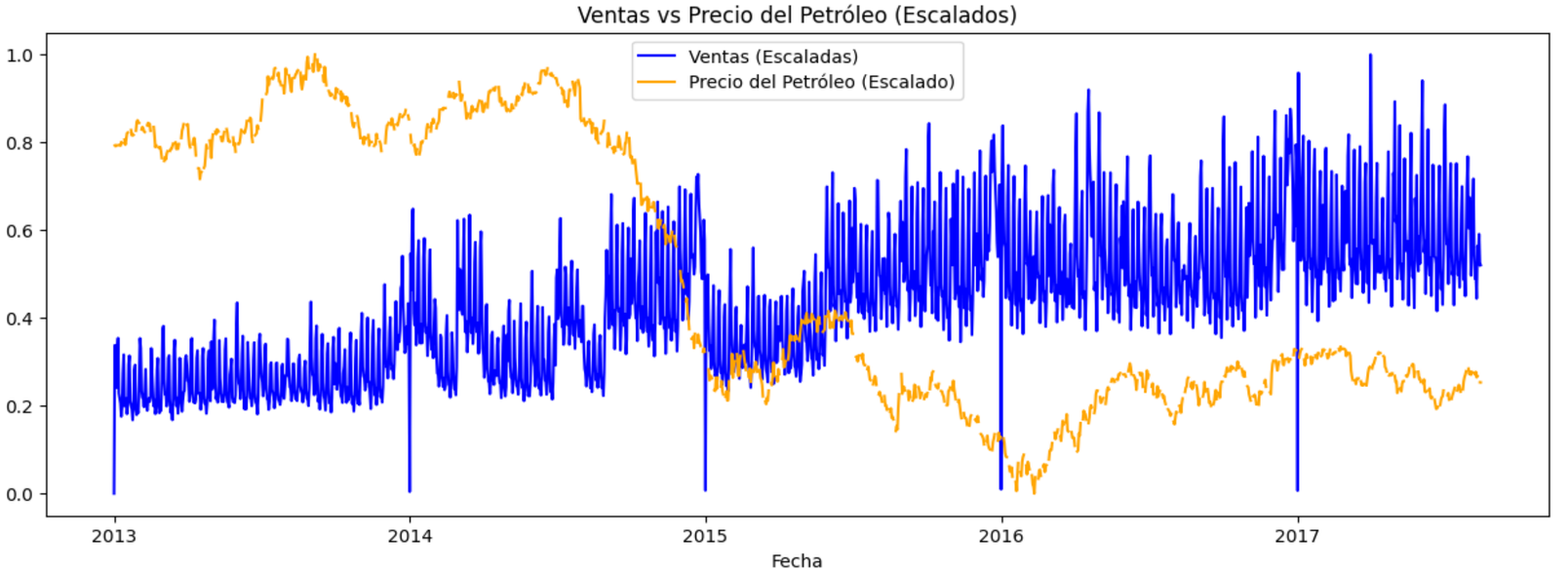


ANÁLISIS EXPLORATORIO

(EFECTO DE VARIABLES EXTERNAS)

- CORRELACIÓN NEGATIVA ENTRE PRECIO DEL PETRÓLEO Y VENTAS.
- IMPACTO POSITIVO DE LAS PROMOCIONES EN LAS VENTAS PROMEDIO.
- DÍAS FESTIVOS CLASIFICADOS COMO NACIONAL, REGIONAL Y LOCAL.
- EFECTO DEL TERREMOTO.





PREPROCESAMIENTO Y LIMPIEZA DE DATOS

TÉCNICAS APLICADAS

IMPUTACIÓN DEL PRECIO DEL PETRÓLEO CON INTERPOLACIÓN LINEAL Y FORWARD-FILL PARA MANTENER CONTINUIDAD.

CONVERSIÓN DE DÍAS FESTIVOS EN VARIABLES BINARIAS (0/1) POR TIPO: NACIONAL, REGIONAL Y LOCAL, ENLAZADOS POR FECHA Y UBICACIÓN.

TRATAMIENTO DE EVENTOS Y OUTLIERS

DETECTAMOS UN EVENTO ATÍPICO EN ABRIL-MAYO 2016 DEBIDO A UN TERREMOTO.

SE AJUSTARON MANUALMENTE LAS VENTAS LOGARÍTMICAS DE ESE PERIODO CON BASE EN PROMEDIOS HISTÓRICOS MULTIANUALES.

ESTE TRATAMIENTO FUE CRUCIAL PARA EVITAR SESGOS EN EL ENTRENAMIENTO DEL MODELO.

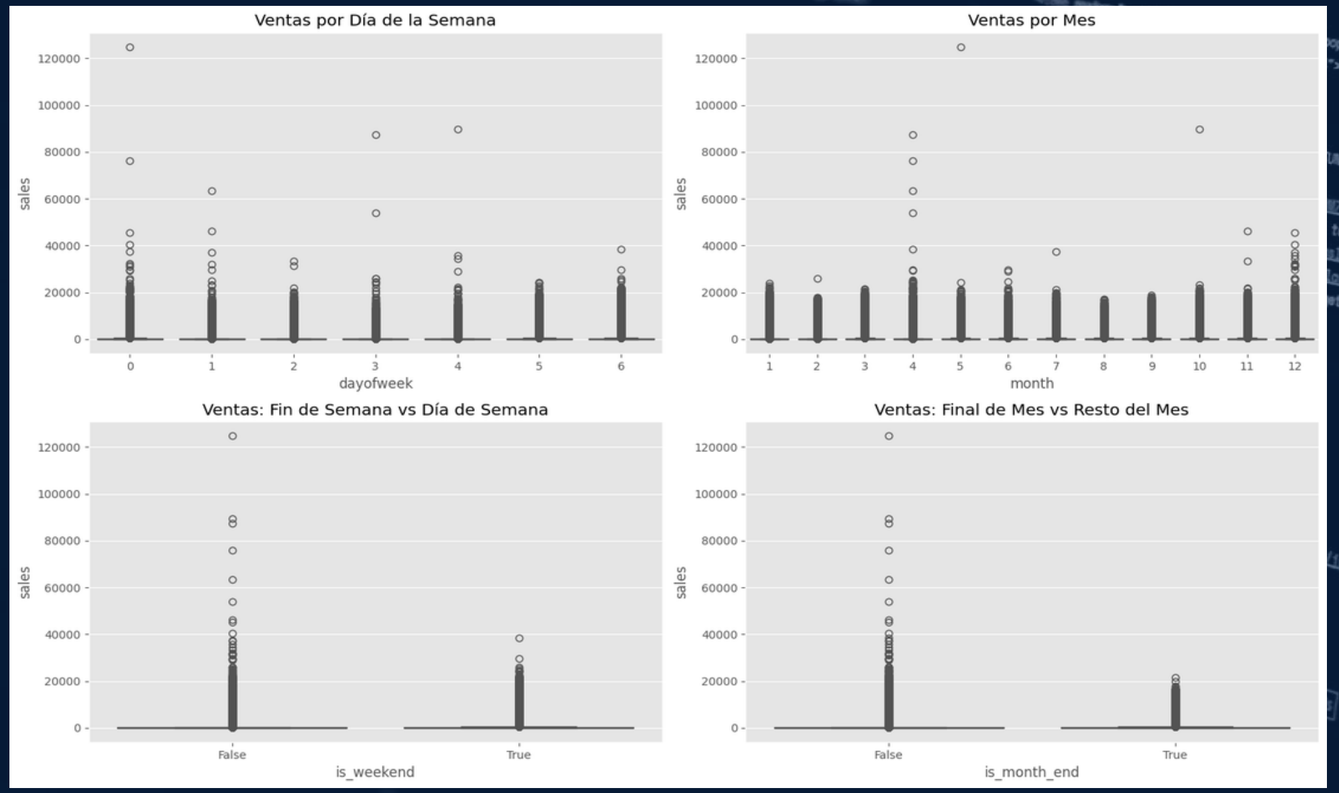
INGENIERÍA DE CARACTERÍSTICAS

VARIABLES TEMPORALES CREADAS:

- DÍA DE LA SEMANA, FIN DE MES, DÍA DE PAGO
- ESTACIONALIDAD ANUAL: MES, SEMANA DEL AÑO

OTRAS VARIABLES CLAVE:

- PROMEDIOS MÓVILES POR TIENDA Y FAMILIA (7, 14 Y 30 DÍAS)
- PROPORCIÓN DE PRODUCTOS EN PROMOCIÓN
- CODIFICACIÓN DE UBICACIÓN Y TIPO DE TIENDA



MODELADO Y EVALUACIÓN

MODELO: LightGBM

VALIDACIÓN: TIMESERIESSPLIT PARA EVITAR FUGA TEMPORAL

MÉTRICA: RMSLE PERSONALIZADA PARA PENALIZAR ERRORES RELATIVOS

SCORE FINAL: 0.45907 (115 de 858 participantes)

PRINCIPALES DESAFÍOS

INTEGRACIÓN DE FUENTES DE DATOS DISPARES (TIENDAS, PETRÓLEO, CALENDARIO, TRANSACCIONES)

MODELADO ROBUSTO ANTE EVENTOS NO RECURRENTES

NO SE PUEDA USAR CUALQUIER MODELO. INTENTAMOS CON OTROS QUE DABAN VALORES NO ACEPTABLE, O NO TERMINABAN DE EJECUTAR

REFLEXIONES FINALES

HOUSE PRICES - ADVANCED REGRESSION TECHNIQUES

CONCLUSIONES

- Importancia del preprocesamiento
- Buen desempeño al combinar varios modelos

LECCIONES APRENDIDAS

- El desempeño del modelo estuvo fuertemente influenciado por la calidad del preprocesamiento y la ingeniería de características (Feature engineering), lo que demuestra que estos pasos son tan críticos como la arquitectura del modelo en sí.
- La transformación logarítmica fue crucial para la métrica
- La regularización ayuda a prevenir overfitting
- El EDA inicial ahorró problemas posteriores

RECOMENDACIONES

- Probar modelos más complejos (redes neuronales)
- Más ingeniería de características

STORE SALES - TIME SERIES FORECASTING

CONCLUSIONES

- Importancia del preprocesamiento y el uso de todos los datos disponibles
- Se construyó un modelo preciso, eficiente y explicable.

LECCIONES APRENDIDAS

- La transformación logarítmica fue crucial para la métrica
- El desempeño del modelo estuvo fuertemente influenciado por la calidad del preprocesamiento y la ingeniería de características, lo que demuestra que estos pasos son tan críticos como la arquitectura del modelo en sí.
- Evaluación adecuada con splits temporales evita sobreajuste.

RECOMENDACIONES

- Mejorar la ingeniería de características **(feature engineering)**
- Explorar otros modelos: híbridos y redes neuronales para series temporales

MUCHAS GRACIAS!