INFORME #1

PROYECTO:

Mercado inmobiliario ruso de Sberbank

INTEGRANTES:

Jhon Jader Caro Sánchez

CC. 1001137636
Ingeniería de Sistemas

Leider Steven Caro Mejía

CC. 1017260248
Ingeniería Civil

José Manuel Ladino Villa

CC. 1010075481
Ingeniería Civil

Descripción del problema

En este trabajo el objetivo **es predecir el precio de las casas en diferentes zonas de Rusia,** esto es especialmente importante en el contexto de una economía rusa volátil y un mercado inmobiliario con características complejas que hacen que las predicciones de precios sean un desafío único. Dicha base de datos fue proporcionada por uno de los bancos más grandes y antiguos de Rusia con el fin de poder hallar modelos que fueran más precisos a la hora de determinar el precio de inmobiliarios y ofrecer dicho servicio predictivo a sus clientes.

Descripción de la base de datos

Usaremos la base de datos de Kaggle de la siguiente competición (https://www.kaggle.com/competitions/sberbank-russian-housing-market) que tiene 30472 número de muestras (casas) solo en la base de datos de entrenamiento y 422 columnas

Las variables se dividen en dos bases de datos:
1) **Variables relacionadas directamente con el inmueble (Algunas de las más relevantes):**

1. precio de venta (esta es la variable objetivo)
2. identificación de la transacción o venta del inmueble
3. fecha de la transacción
4. área total en metros cuadrados, incluidas logias, balcones y otras áreas no residenciales
5. superficie habitable en metros cuadrados, excluyendo logias, balcones y otras áreas no residenciales
6. piso: para apartamentos, piso del edificio
7. número de pisos del edificio
8. año de construcción
9. número de salas de estar
10. área de cocina
11. condición del apartamento
12. nombre del distrito
13. población total en el área
14. población de área por género
15. población menor de edad
16. población en edad de trabajar
17. población en edad de jubilación
18. cantidad de edificios en el área
19. número de cafeterías dentro del área
20. número de centros comerciales en el área
21. número de zonas industriales en el área
22. número de zonas verdes
23. Distancia en KM hasta el metro
24. Distancia en KM hasta Carretera circular de Moscú

25. Distancia en KM hasta centro de la ciudad

2) **Variables relacionadas con la economía al momento de la venta del inmueble(Algunas de las más relevantes):**

1. Fecha de la transacción
2. Precio del Petróleo crudo (Rublo)
3.Crecimiento del PIB real
3. Indicador de inflación
4. Crecimiento del índice de precios al consumidor
5. Crecimiento del índice de precios al productor
6. Deflactor del PIB
7. Balanza comercial (como porcentaje del año anterior)
8. Tipo de cambio rublo/USD
9. Tipo de cambio rublo/EUR
10. Importación/exportación neta de capital
11. Provisión por pedidos en Rusia (para el desarrollador)
12. Provisión por pedidos en Moscú (para el desarrollador)
13. Índice RTS
14. Índice MICEX
15.índice MICEX para bonos gubernamentales
16. Volumen de depósitos de los hogares
17. Crecimiento del volumen de los depósitos de la población
18. Tasa de interés promedio sobre depósitos
19. Volumen de préstamos hipotecarios
20. Crecimiento de los préstamos hipotecarios
21. Tasa media ponderada de los préstamos hipotecarios
22. Crecimiento del producto regional bruto de la entidad constituyente de la Federación de Rusia donde se encuentra el Apartamento
23. Ingreso promedio per cápita
24.Crecimiento de la renta real disponible de la población
25. salario: Salario mensual promedio
26. Crecimiento de los salarios nominales
27. Costo de una canasta fija de bienes y servicios de consumo para comparaciones
28. Volumen de trabajos de construcción realizados (millones de rublos)
29.El índice del volumen físico de inversión en activos fijos (en precios comparables en% al mes correspondiente del año anterior)
30. Tasa de aumento/disminución natural de la población (1.000 personas)
31. Aumento (disminución) de la migración de la población
32. Crecimiento poblacional total
33. Indices de parto
34. Indices de mortalidad

Métricas de evaluación

Como métrica de machine learning, se utilizará el RMSLE (Root Mean Squared Logarithmic Error), que es la métrica definida por la competencia en cuestión. El RMSLE mide el error promedio de los logaritmos de las predicciones y los valores reales. Esto tiende a penalizar de manera más significativa las grandes discrepancias en valores más grandes. Supongamos que tenemos un conjunto de datos con dos observaciones: Valor real de la casa (en dólares): 100,000 y la predicción del modelo (en dólares): 120,000 segunda observación: Valor real de la casa (en dólares): 40,000, predicción del modelo (en dólares): 20,000.

Primero, calculamos el error logarítmico para cada observación:

Error logarítmico = Ln(120,000) - Ln(100,000) = 0.1823
Error logarítmico = Ln(20,000) - Ln(40,000) = -0.6931

Luego, calculamos el error cuadrático de estos errores logarítmicos:

Error cuadrático = (0.1823)^2 = 0.0333
Error cuadrático = (-0.6931)^2 = 0.4800

Ahora, tomamos el promedio de estos errores cuadráticos:

Promedio de errores cuadráticos = (0.0333 + 0.4800) / 2 = 0.2567

Finalmente, calculamos la raíz cuadrada de este promedio para obtener el RMSLE:

RMSLE = √(0.2567) ≈ 0.5067

En este ejemplo, el valor del RMSLE es aproximadamente 0.5067. Esto significa que, en una escala logarítmica, el modelo tiene un error promedio del 50.67%. Cuanto más cercano a cero sea el RMSLE, mejor será el rendimiento del modelo en términos de predicción de precios de viviendas.

Se puede realizar un análisis de impacto en las ventas con esta métrica. Por ejemplo, se podría calcular cuántos ingresos adicionales se generan como resultado de las predicciones más precisas del modelo en comparación con un enfoque menos preciso.

Criterio deseable de desempeño en producción

1) En general, se espera que el modelo tenga un error de predicción bajo, inicialmente esperamos que el RMSLE este por debajo del 15%. Esto significa que las predicciones del modelo deben estar lo más cerca posible de los valores reales de los precios de las casas.

2) Evaluar si el modelo está contribuyendo a aumentar las ventas, la retención de clientes y la satisfacción del cliente.