Proyecto para el módulo de Álgebra

Home Value Estimator

Eugenia Barozzi

Instituto Superior Politécnico de Córdoba | Ciencia de datos e inteligencia artificial

Profesora: Sandra Aguirre

5 de Noviembre de 2024

Nombre del proyecto:

Home Value Estimator: predictor de precios de viviendas mediante machine learning, este proyecto propone el desarrollo de un modelo de machine learning específicamente de regresión lineal múltiple que implementa algoritmos para analizar datos y proporcionar un predictor de precios según características de las viviendas.

Tipo de proyecto

Tecnológico: El proyecto involucra la creación de una aplicación de análisis de datos, la implementación de modelos de machine learning y la investigación en la efectividad de estos modelos.

Espacios participantes en el módulo

Álgebra

Estos ejes temáticos se relacionan con los siguientes espacios curriculares del módulo:

1. Ciencia de datos:
    - ➢ Programación: Conocimientos de lenguajes de programación como Python.
    - ➢ Algoritmos y estructuras de datos para diseñar y optimizar algoritmos de IA.
    - ➢ Inteligencia artificial y aprendizaje automático para fundamentos y técnicas de IA Y ML.
2. Matemáticas y Estadística:
    - ➢ Cálculo: para entender los modelos matemáticos detrás de los algoritmos de IA.
    - ➢ Álgebra lineal: fundamental para el funcionamiento de algoritmos de ML.d
    - ➢ Estadística: Para el análisis de datos y la validación de modelos

Competencias y habilidades del perfil profesional

1. Pensamiento crítico y resolución de problemas:
    - Descripción: Análisis de datos complejos para tomar desiciones informadas, identificación de patrones y tendencias de los datos.
    - Competencias: La capacidad de evaluar y sintetizar información y aplicar el razonamiento lógico para resolver problemas.
2. Habilidades técnicas:
    - Descripción: Conceptos y aplicaciones del álgebra computacional, dominio de Python, herramientas como Matploit, Seaborn, Pandas, Scikit-learn.
    - Competencias: Desarrollo de habilidades en programación, análisis de datos y uso de herramientas avanzadas de visualización, machine learning y álgebra computacional.
3. Trabajo en equipo y comunicación:
    - Descripción: Colaboración con el equipo de proyecto y comunicación clara de resultados.
    - Competencias: Desarrollo de habilidades de colaboración y trabajo en equipo, esenciales para trabajar de manera multidisciplinaria.

Problemas o necesidades:

El mercado inmobiliario actual requiere herramientas precisas y eficientes para evaluar el valor de las propiedades. La falta de métodos automatizados y confiables puede llevar a estimaciones inexactas, afectando las decisiones de compra y venta. Este proyecto aborda

esta necesidad desarrollando un modelo de regresión lineal múltiple para predecir los precios de las casas basándose en sus características.

Fundamentación

Este proyecto fue elegido debido a la facilidad de acceso a datos de precios de viviendas, lo que permite probar y validar el funcionamiento de la predicción de precios de casas usando regresión múltiple.

El potencial de este proyecto es significativo, ya que puede transformar la manera en que se realizan las evaluaciones inmobiliarias, con este proyecto se pueden considerar múltiples factores que afecten el precio de la vivienda, proporcionando estimaciones más precisas y confiables. Esto beneficiaría a agentes inmobiliarios, tasadores, compradores y vendedores.

Con respecto a los profesionales de ciencia de datos este proyecto es altamente relevante, ya que nos permite abordar técnicas avanzadas de modelado, análisis de datos y desempeño. Además que proporciona experiencia práctica en el uso de herramientas necesarias para la cursada.

Visión del proyecto:

El proyecto tiene como objetivo desarrollar un modelo de regresión lineal múltiple para predecir los precios de casas basándose en sus características. Se estima que el modelo proporcionará una herramienta precisa y eficiente para evaluar el valor de las propiedades inmobiliarias.

- Objetivo general: Desarrollar un modelo de regresión lineal múltiple que prediga con precisión los precios basándose en las características específicas relevantes según el análisis.
- Objetivos específicos:
    1. Recolectar y preparar datos: Obtener un conjunto de datos de precios de viviendas y sus características, y realizar la limpieza y normalización de los datos.
    2. Desarrollar el modelo: Implementar el modelo de regresión lineal múltiple utilizando técnicas de aprendizaje automático.
    3. Evaluar el modelo: Validar el modelo utilizando métricas de rendimiento como el R^2, el error cuadrático medio (MSE) y la precisión.
    4. Predicción: Aplicar el modelo para predecir precios de viviendas.
    5. Integración de Conceptos de Álgebra y Matemática: Utilizar herramientas matemáticas para el análisis de datos.
- Metas: Calidad: Alcanzar una precisión del modelo de al menos 80% en la predicción de precios de viviendas, Cantidad: Evaluar el modelo utilizando al menos 1000 registros de datos de viviendas.

Selección de acciones

Objetivo específico	Acciones	Habilidades a lograr
Recolectar datos y limpieza	Buscar fuentes de datos confiables, limpiar y normalizar los datos.	Análisis de datos, manipulación de datos, limpieza de datos.

Desarrollar el modelo	implementar el modelo de regresión lineal múltiple utilizando Python y bibliotecas como scikit-learn.	Programación en Python, modelado predictivo, uso de bibliotecas de aprendizaje automático.
Predicción	Utilizar el modelo seleccionado para predecir precios en nuevos datos,	Habilidad para aplicar e interpretar modelos predictivos a datos reales.
Evaluar el modelo	Validar el modelo utilizando un conjunto de datos de prueba y calcular métricas de rendimiento.	Evaluación de modelos, interpretación de métricas de rendimiento.
Integración de Conceptos de Álgebra	Utilizar conceptos estadísticos y matemáticos para la interpretación de resultados y validación del modelo	Dominio de álgebra lineal y estadística aplicada. habilidad para integrar estos conceptos matemáticos.

Resumen y Producto final:

Diseño del Modelo

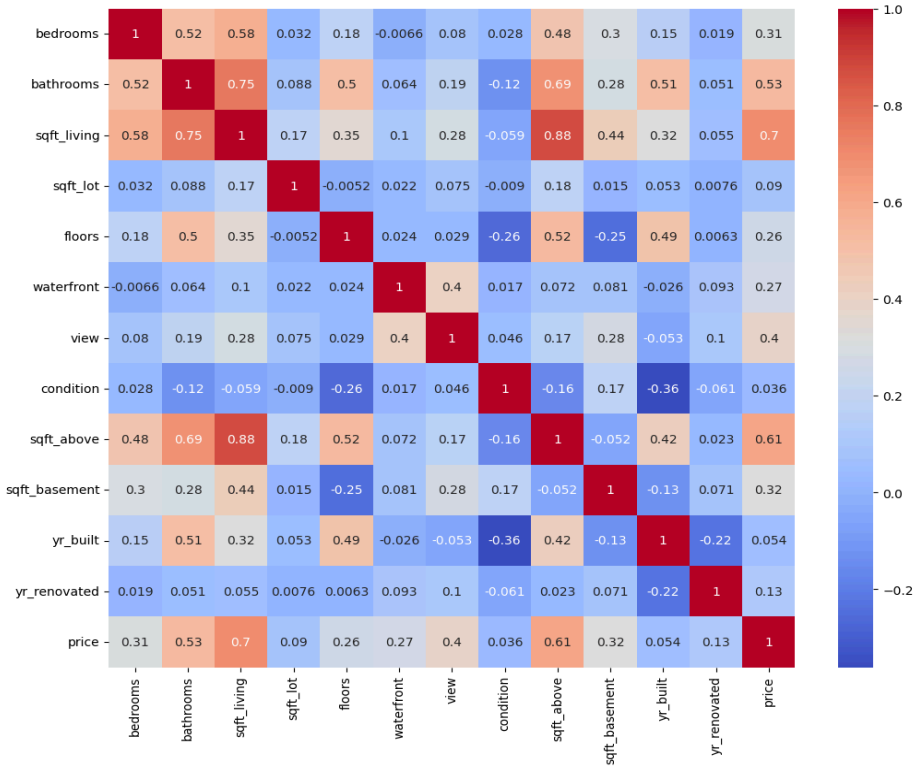
El modelo de regresión lineal múltiple se diseñó para predecir los precios de las casas utilizando un conjunto de datos, se escogieron las variables independientes las cuales son: el número de habitaciones, baños, tamaño del sótano, espacio habitable y otras características relevantes obtenidas según los análisis de correlación que fue útil para identificar qué variables tienen relaciones significativas

- bedrooms: Habitaciones
- bathrooms: Baños
- sqft_basement: Tamaño del sótano en pies cuadrados
- sqft_above: Espacio interior de la vivienda sobre el nivel del suelo en pies cuadrados
- sqft_living: Espacio habitable en pies cuadrados

Elección de variables

Es una tabla que muestra los coeficientes de correlación de Pearson entre todas las variables numéricas seleccionadas. La matriz de correlación que se genera es una representación algebraica de estas relaciones.

```
numerical_features = ['bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors','waterfront', 'view', 'condition',
corr_matrix = data[numerical_features].corr()
```



Luego, separamos las variables seleccionadas donde **X**: Contiene las variables independientes (características) que se utilizarán para predecir el precio de las casas, **y**: Contiene la variable dependiente (precio de las casas).

```
] # Separar las características (X) y la variable objetivo (y)

  X = data[['bedrooms', 'bathrooms','sqft_basement', 'sqft_above','sqft_living']]
  y = data['price']
```

Proceso de Predicción en Regresión Lineal Múltiple

Se dividen los datos de entrenamiento y de prueba siendo usualmente 80% de los datos se utiliza para entrenar el modelo y el 20% para probarlo.

Luego se crea y entrena el Modelo con: LinearRegression() que crea una instancia del modelo de regresión lineal.

Ajuste del Modelo:

Cuando se ejecuta `model.fit(X_train, y_train)`, el modelo calcula los coeficientes (β) y la intersección (β_0) que mejor se ajustan a los datos de entrenamiento.

- Fórmula: $\beta = (X^T X)^{-1} X^T y$

model.fit(X_train, y_train): Entrena el modelo utilizando los datos de entrenamiento. Durante este proceso, el modelo ajusta los coeficientes para minimizar la suma de los errores al cuadrado entre los valores predichos y los valores reales.

```
# dividir los datos ne entranmiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
imputer = SimpleImputer(strategy='mean')

# crear modelo, y entrenarlo con los datos de entrenamiento.
model = LinearRegression() #Esto genera una ecuación lineal que intenta predecir

# y (el precio) en función de las variables en X (Rooms, Bathroom, BuildingArea).
model.fit(X_train, y_train)
```

Predicción:

Al ejecutar `model.predict(X_test)`, el modelo aplica la ecuación de regresión lineal múltiple a los datos de prueba (X_test) utilizando los coeficientes e intersección calculados.

- Fórmula: $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \ + \beta_n x_n$

- **Y** es la variable dependiente que queremos predecir
- **b** es la intersección o término constante.
- **b_1, b_2..,** son los coeficientes de las variables independientes.
- **X_1, X_2,** son las variables independientes (por ejemplo, número de habitaciones, tamaño del sótano, etc.).

```
# Realizar predicciones en los datos de prueba
print(X_test)
y_pred = model.predict(X_test)
```

Cuando se ejecuta `model.predict(X_test)`, el modelo utiliza esta ecuación para calcular las predicciones (y^) para cada fila de X_test

En resumen, `model.predict` aplica la ecuación de regresión lineal múltiple utilizando los coeficientes e intersección calculados durante el ajuste del modelo para generar las predicciones.

Evaluación del modelo

Para evaluar al modelo, usaremos el r2(coeficiente de determinación), MSE(error cuadrático medio), y el MAE(error absoluto medio).

```
r2 = r2_score(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
```

1. Coeficiente de Determinación $R2$, Valor: 0.51

Aproximadamente el 51% de la variabilidad en los precios de las casas puede ser explicada por las variables independientes en tu modelo. Esto indica que el modelo tiene una capacidad explicativa moderada, pero aún hay un 49% de la variabilidad que no está siendo capturada, lo que sugiere la presencia de otros factores influyentes no considerados en el modelo.
La interpretación de la métrica es que mide la proporción de la variabilidad total de la variable dependiente que es explicada por el modelo.

2. Error Cuadrático Medio (MSE), Valor: 74,198,772,469.47

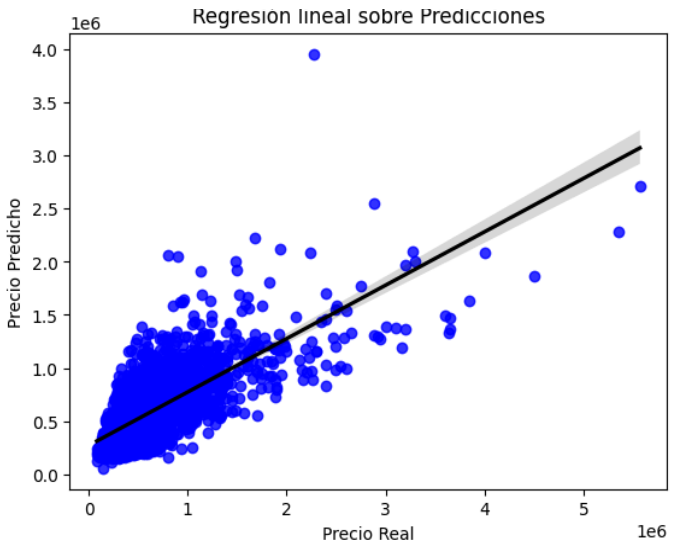
Interpretación: El MSE indica que, en promedio, las predicciones del modelo se desvían de los valores reales en una cantidad significativa. Un MSE alto sugiere que las predicciones del modelo no son muy precisas y que hay margen para mejorar la precisión del modelo.
La interpretación de la métrica es que mide el promedio de los errores al cuadrado entre valores predichos y los valores reales. Un MSE más bajo indica un mejor ajuste.

3. Error Absoluto Medio (MAE), Valor: 174,405.68

Interpretación: El MAE indica que, en promedio, las predicciones del modelo se desvían de los valores reales en aproximadamente 174,405.68 unidades monetarias (por ejemplo, dólares). Esto proporciona una medida de la magnitud promedio del error en las predicciones sin considerar la dirección del error.
La interpretación de la métrica es que mide el promedio de los errores absolutos entre los valores predichos y los valores reales, un MAE más bajo indica un mejor ajuste del modelo.

Conclusión

El modelo tiene una moderada capacidad explicativa: El $R2$ de 0.51 sugiere que el modelo explica una proporción moderada de la variabilidad en los datos, es decir un 50%.
Precisión Limitada: Tanto el MSE como el MAE indican que hay una desviación significativa entre las predicciones y los valores reales, lo que sugiere que el modelo podría beneficiarse de mejoras adicionales.



Viabilidad del Modelo

El modelo puede ser útil como una primera aproximación para predecir los precios de las casas, pero hay margen para mejorar su precisión y capacidad explicativa. Para hacerlo más viable se podría considerar:

1. Incluir más variables relevantes: Agregar características adicionales que puedan influir en el precio de las casas.
2. Utilizar técnicas de regularización: Como Ridge o Lasso, para reducir el sobreajuste.
3. Probar con modelos más complejos: Como árboles de decisión, bosques aleatorios o redes neuronales.

En resumen, el modelo actual es un buen punto de partida, pero se beneficiaría de mejoras adicionales para ser más preciso y confiable en sus predicciones.

- Para ver el sólo el código del modelo desarrollado: AlgebraRgresion
- Para ver los datos y el proyecto desarrollado, entrar a github: Regresiónmúltiple