

Reporte de Analisis de ventas de viviendas

By Eddy silva

Introducción

El objetivo principal de este proyecto de estudio es analizar y mejorar las ventas de viviendas utilizando el conjunto de herramientas aprendidas de análisis, limpieza y transformación de datos extraídos de CSV.

El enfoque es optimizar la predicción de precios mediante el manejo de valores faltantes, detección y tratamiento de outliers, y la aplicación de técnicas de preprocesamiento como normalización y estandarización.

Análisis de los datos

Nuestro proceso de análisis de datos comenzó con la interpretación y visualización de datos , posteriormente destacamos las columnas más importantes y las imprimimos para tener mayor conocimiento de sus datos. Posteriormente analizamos los tipos de datos, porcentaje de datos faltantes entre otras anomalías de los mismos.

Manejo y limpieza de datos

Posterior al análisis y detección de anomalías en los datos comenzamos con el proceso de manejo de limpieza de datos, proceso en el cual utilizamos métodos de imputación para que nuestro modelo no tenga dificultades para el analisis y prediccion

Imputación de datos

Los métodos de imputación que ocupamos fueron diversos entre ellos la imputación por medi, la imputación por moda y imputacion por categoría, los datos que se imputan por media fue gracias a su naturaleza la cual les permite recibir la media como variables numéricas, los datos que sufrieron imputación por moda fueron datos más descriptivos que resaltan sobre el resto teniendo información más atractiva para nuestro modelo, los datos categóricos fueron rellenados con 'NA' cuando no había una categoría predominante.

Manejo de Outliers

Como manejo de outlier tenemos la opción de manejarlo por el método de eliminación o dirigirnos por el camino de la transformación para nuestro ejercicios implementamos el método de la transformación ya que permanecen los datos de una mejor forma que nuestro modelo llegaría a analizar.

Codificación de variables categóricas

Para la codificación de variables categóricas ocupamos el método One-Hot Encoding ya que nos pareció pertinente evitar introducir un orden artificial entre categorías nominales .

Normalización y Estandarización

Aplicamos método de estandarización a datos los cuales comparte conceptos como el área de esta forma se utilizó para variables con diferentes rangos o distribuciones en nuestro caso LotArea y grLivArea.

también se aplicó normalización a la variable sale price para que estuviera en el rango de 0,1 mejorando la convergencias con algunos posibles algoritmos.

Creación de nuevas variables

Se crearon nuevas variables como Quality_Index la que combina la calidad y condición de una casa esto ayudando a tener una visión integral de la calidad de vivienda.

TotalbuildingArea es otra de nuestras variables las cuales nos apoyaran en el análisis y esta nos ayuda a notar el área específica que ya está construida de cada casa que se vendió .