Primera entrega de proyecto

POR:

Cristian David Tamayo Espinosa

MATERIA:

Introducción a la inteligencia artificial

PROFESOR:

Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN 2023

1. Planteamiento del problema

El problema predictivo que se desea resolver consiste en predecir los precios de viviendas a partir de numerosas características diferentes. Estas características incluyen información sobre el tamaño del lote, la calidad de la fachada, el número de habitaciones y baños, y la ubicación geográfica, entre otras.

El objetivo es desarrollar un modelo de regresión que pueda predecir con precisión el precio de una casa en base a estas características. Para ello, se cuenta con un conjunto de datos de entrenamiento que incluye información sobre viviendas y sus precios. Además, se dispone de otro conjunto de datos de prueba que incluye información sobre viviendas, pero sin el precio.

Por lo que el objetivo final de este problema es ayudar a los compradores, vendedores y corredores de bienes raíces a tomar decisiones informadas sobre el valor de las propiedades. Además de que al predecir con precisión el precio de una casa, se puede obtener una estimación del valor real de la propiedad y evitar transacciones injustas para ambas partes.

2. Dataset

El conjunto de datos a utilizar proviene de una competición de Kaggle en la que se proporcionan datos de propiedades inmobiliarias en Ames, Iowa, Estados Unidos.

El conjunto de datos consta de cuatro archivos: train.csv, test.csv, data_description.txt y sample_submission.csv.

El archivo **train.csv** es el conjunto de datos de entrenamiento y contiene información sobre 1,460 propiedades

Mientras que el archivo **test.csv** es el conjunto de datos de prueba y contiene información sobre 1,459 propiedades.

El archivo **data_description.txt** proporciona una descripción completa de cada columna en los archivos de datos

El archivo **sample_submission.csv** es una presentación de referencia de una regresión lineal en el año y mes de venta, el tamaño del lote y el número de habitaciones.

Los campos de datos que incluyen que podremos encontrar tanto en train.csv como test.csv

SalePrice: El precio de venta de la propiedad en dólares. Esta es la variable objetivo que se intenta predecir.

MSSubClass: La clase de construcción

MSZoning: La clasificación general de zonificación

LotFrontage: Pies lineales de calle conectados a la propiedad

LotArea: Tamaño del lote en pies cuadrados

Street: Tipo de acceso a la carretera

Alley: Tipo de acceso a callejón

LotShape: Forma general de la propiedad

LandContour: Plano de la propiedad

Utilities: Tipo de servicios públicos disponibles

LotConfig: Configuración del lote

LandSlope: Pendiente de la propiedad

Entre muchos otros datos más que encontraremos allí adentro.

3. métricas

Las métricas de desempeño para evaluar las predicciones en esta competición se basan en el cálculo del error cuadrático medio (RMSE) entre el logaritmo de los valores de venta observados y el logaritmo de los valores de venta pronosticados. Tomar los logaritmos garantiza que los errores en la predicción de casas caras y baratas afecten igualmente el resultado final. En resumen, se busca minimizar el RMSE en la escala de logaritmos para obtener las predicciones más precisas posibles.

4. Desempeño

un primer criterio deseable de desempeño en producción podría ser que el modelo sea capaz de predecir con precisión los precios de venta de las viviendas con un bajo valor de RMSE, es decir, cuanto menor sea el valor de RMSE obtenido por el modelo, mejor será su desempeño en producción.

Ya que con un bajo valor de RMSE, se puede inferir que el modelo de regresión utilizado para predecir los precios de las casas es más preciso en sus predicciones.

Bibliografía:

House Prices: Advanced Regression Techniques | Kaggle. (2023). Retrieved 13 March 2023, from https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview/description