Segunda 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

Informe

Hasta el momento se ha logrado realizar múltiples avances en cuestión de familiarización, preprocesado, y limpieza de datos, a su vez que la iniciación de generar algún modelo, el cual ha pesar de que se

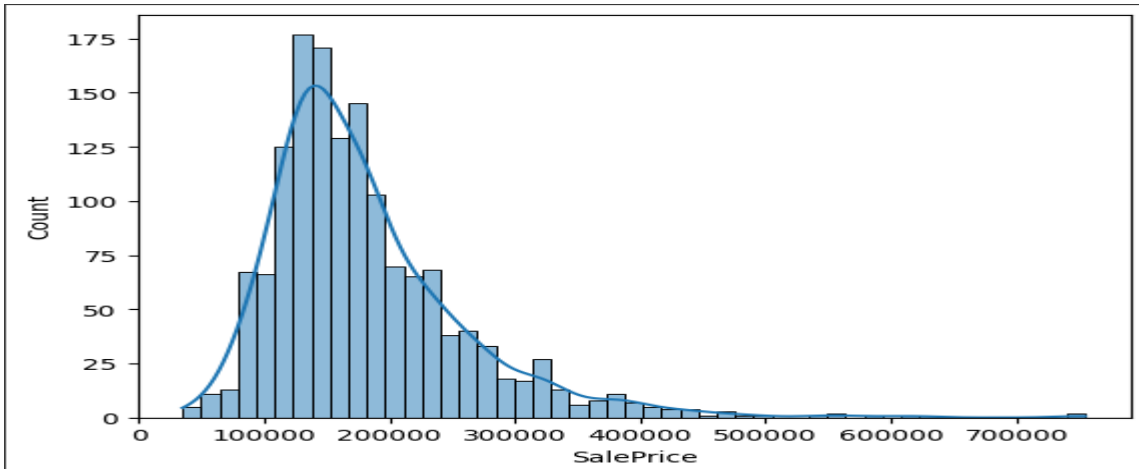
Todo con el objetivo central de generar un modelo de machine learning para predecir los precios de las casas, y de esta forma generar un modelo predictivo eficiente.

Lo primero que se realizó fue importar las bibliotecas necesarias, incluyendo numpy, pandas, matplotlib y seaborn, ya con ellas leímos los datos del conjunto de entrenamiento y del conjunto de prueba utilizando el método **pd.read_csv**. Utilice el método **data.info()** para obtener una visión general de los datos y verificar si existen valores nulos en ellos,

De igual manera, el método **data.describe()** me proporcionó estadísticas descriptivas de los datos numéricos del dataset, de esta manera pude conocer

- El número de elementos no nulos (count)
- La media (mean)
- La desviación estándar (std)
- El valor mínimo (min)
- El percentil 25 (25%)
- El percentil 50, también conocido como la mediana (50%)
- El percentil 75 (75%)
- El valor máximo (max)

Posterior a ello realice una exploración a la distribución de cómo se veía y comportaba de una manera estadística la variable objetivo que, en mi caso, es la variable del precio de venta “sale Price”. El cual gráficamente se puede observar que no se distribuye normalmente



Otra de las cosas que se realizó durante el análisis exploratorio de los datos fue observar que algunos datos categóricos realmente no aportaban mucho, por ejemplo, eliminando las características que sólo se centraban en una categoría.

Adicionalmente encontré cuales eran los datos que tenían unos valores nulos muy elevados, y se encontró que existían una categorías con valores nulos superiores al 50%, por lo tanto los decidí eliminar ya que no aportaban mucha información. También se eliminó las características que no eran útiles para la predicción de precios, como el área de la piscina o el tipo de calefacción. También eliminamos las características que tenían muchos valores perdidos y, por lo tanto, no aportaban mucha información al modelo.

Después de esa manipulación de los datos, lo siguiente fue realizar una exploración y análisis de las características categóricas y continuas por separado para comprender mejor cómo influyen en el precio de venta.

Haciendo uso de gráficos de caja y gráficos de barras se analizó los datos categóricos y los histogramas para los datos continuos, y se pudo descubrir que algunas características categóricas, como el vecindario, influyen en gran medida en el precio de venta.

Lidiando con valores atípicos

Otra cosa que se consideró muy importante fue la importancia de tratar los valores atípicos, pues estos podían de cierta manera distorsionar totalmente el modelo posterior que se buscara construir e implementar, por lo que una solución muy conveniente, útil y practica fue hacer uso del método de rango intercuartil (IQR), y esta forma poder tratar y eliminar de una manera efectiva los datos atípicos que pudiesen afectar posteriormente.

Búsqueda y construcción de un modelo

Inicialmente se ha estado buscando un modelo de regresión lineal, ya que es una técnica de aprendizaje automático **supervisado** utilizada para predecir la relación entre una variable dependiente y una o más variables independientes. Por lo que se consideraba una buena opción para observar que tan efectivo podría surgir, no obstante, se planea implementar mas tipos de modelos, de manera que al final se pueda hacer una comparación del accurasy de estos, sin embargo en el momento solo se cuenta con un solo modelo.

El modelo de regresión lineal tuvo un buen rendimiento, con una puntuación de precisión (Accuracy Score) del 0.8859647902124576. Sin embargo, hay muchos otros modelos de aprendizaje automático que se pueden utilizar para mejorar aún

más la precisión, como los modelos de árbol de decisión, los modelos de regresión logística, los modelos de redes neuronales, entre otros.

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