




Entrega de proyecto final

- Pablo Hünicken -



**¿Cómo podemos
predecir el valor de
venta de una casa?**

Objetivo:

- ▶ Predecir el precio de venta de las viviendas a partir de un historial de ventas en la región de Connecticut.

Contexto comercial:

- ▶ El mercado de bienes raíces en Estados Unidos es extremadamente diverso y dinámico. Abarca una amplia variedad de propiedades, desde viviendas unifamiliares hasta edificios comerciales, terrenos vacantes y propiedades de inversión.
- ▶ Para iniciar una negociación es muy útil contar con un "Precio Tentativo" que dependa de un historial de ventas casa. Asimismo, resultaría muy interesante contar con una estimación automática, dejando de lado el criterio del tasador que quizás podría ser subjetivo.

Problema comercial:

- ▶ Al no contar con el precio estimado de la vivienda, es muy difícil para el intermediario poder establecer un punto de negociación de precio entre comprador y vendedor. Por ello es que se requiere contar con un precio estimado en función a ventas históricas de casas similares para que el proceso de negociación sea más ameno para ambas partes.

Contexto analítico:

- ▶ Para poder desarrollar el modelo se cuenta con un listado de todas las ventas de bienes raíces con un precio de venta de \$2,000 o más que ocurren entre el 1 de octubre y el 30 de septiembre de cada año, haciendo un total de 997.213 registros de ventas. Para cada registro de venta, el archivo incluye: ciudad, dirección de la propiedad, fecha de venta, tipo de propiedad (residencial, departamento, comercial, industrial o terreno baldío), precio de venta y tasación de la propiedad.

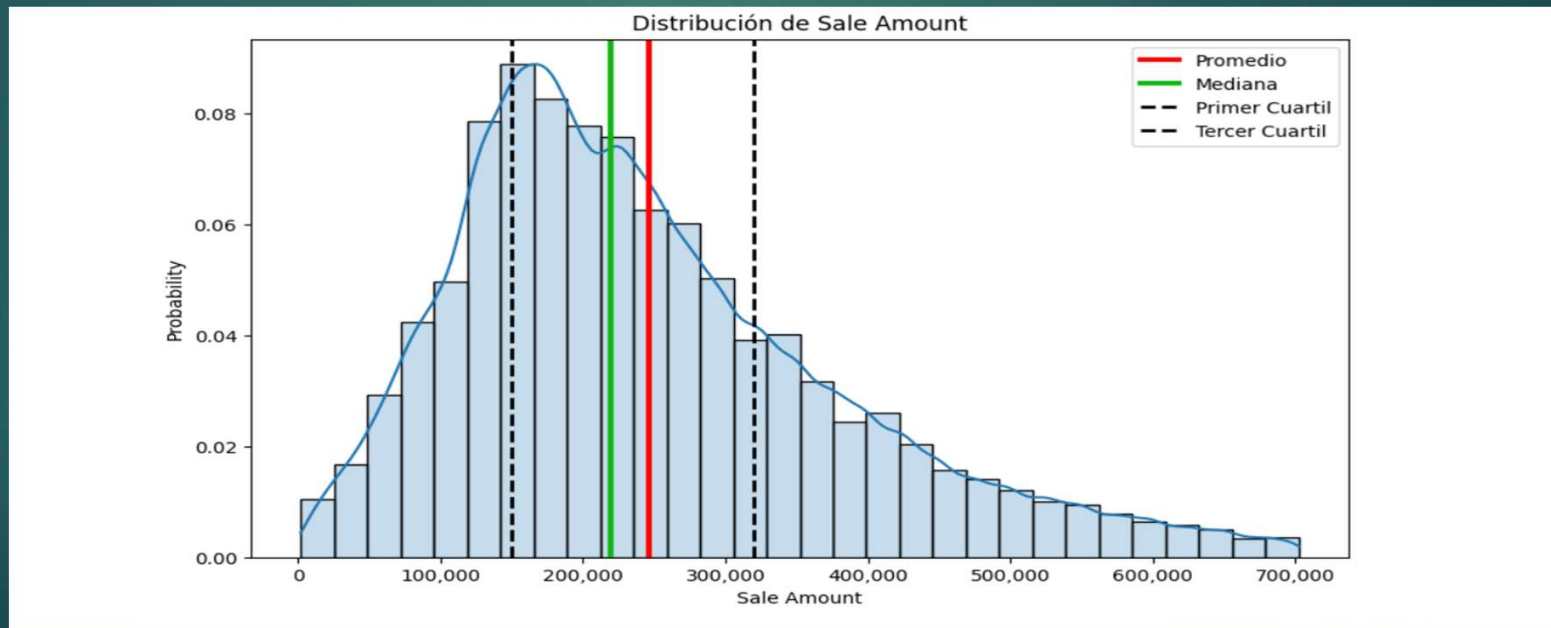
Las variables disponibles (14) son las siguientes:

1. Serial Number: número de serie
2. List Year: año de la venta. (recordemos que el año se considera desde 01/10 al 30/09 de cada año calendario)
3. Date Recorded: fecha de registro
4. Town: ciudad
5. Address: dirección
6. Assessed Value: valor tasado
7. Sale Amount: valor de venta
8. Sales Ratio: ratio de venta (valor tasado / valor de venta)
9. Property Type: tipo de propiedad
10. Residential Type: tipo residencial
11. Non Use Code: código de no uso
12. Assessor Remarks: Observaciones del asesor
13. OPM remarks: Comentario de la operación
14. Location: ubicación

Variable Target: Sale Amount

- ▶ La variable objetivo es el precio de venta de la propiedad.
- ▶ Esta variable es de tipo numérica.
- ▶ Luego de quitar valores atípicos llegamos a la siguiente conclusión de la distribución de nuestra variable target:

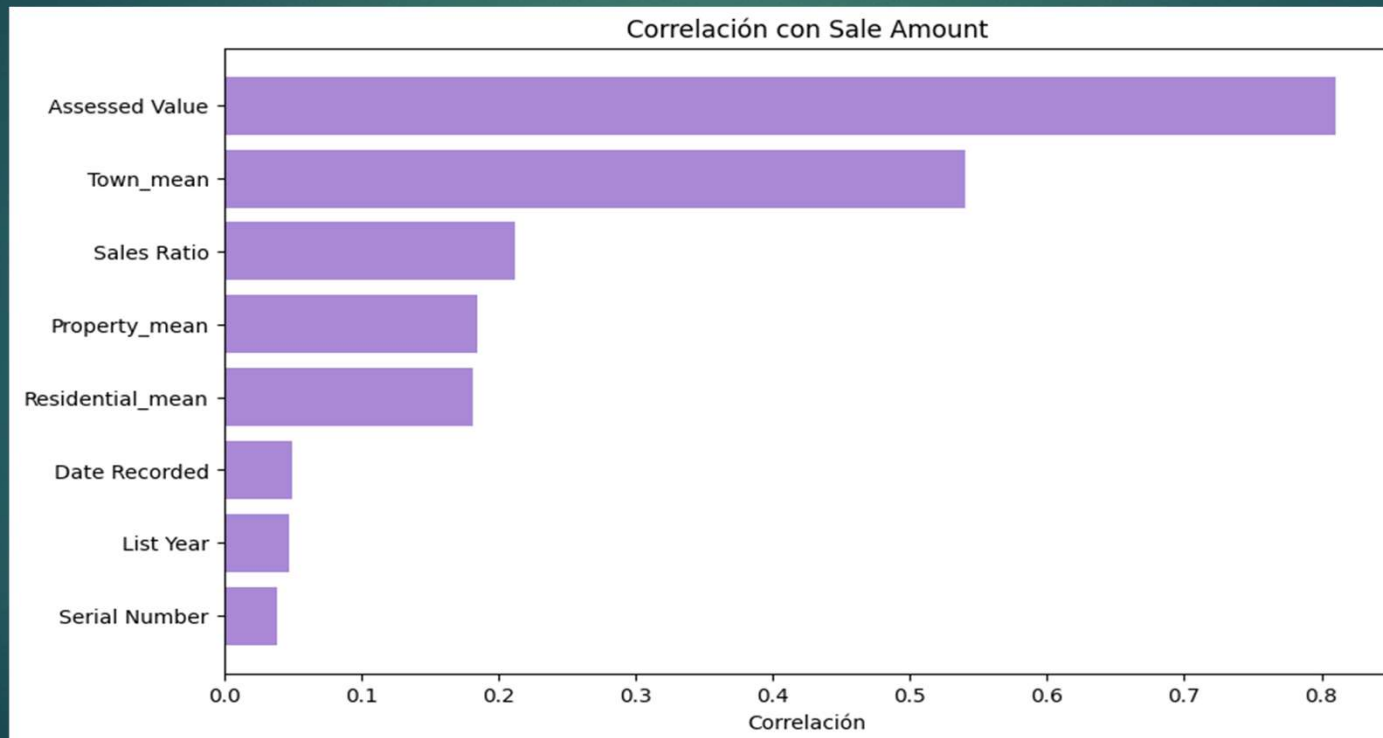
La mediana es menor que el promedio, significa que la distribución es asimétrica hacia la derecha (positiva). Esto indica que la mayoría de los valores se encuentran en la parte inferior de la distribución, lo que hace que el tercer cuartil esté más cerca de la cola derecha de la distribución.



MODELO DE ENTRENAMIENTO Y RESULTADOS

Correlación

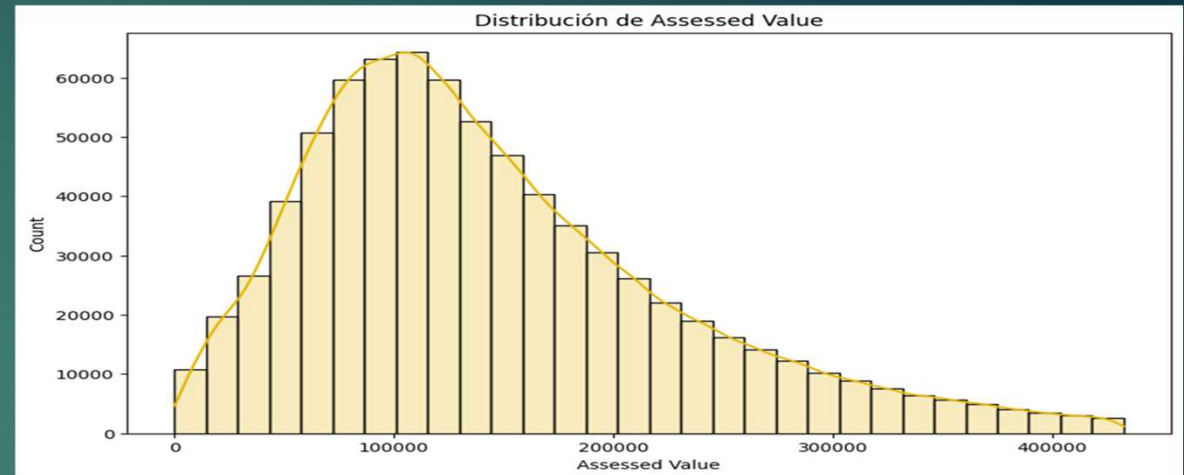
- Lo próximo que analizaremos es la correlación entre cada variable disponible, y la variable target. Debido a que no contamos con muchas variables numéricas en el DF, generaremos nuevas variables numéricas a partir de las variables categóricas, para poder la relación con la variable target. Luego procederemos a realizar un análisis más en profundidad de alguna de ellas.



Analizando la variable "Assessed Value"

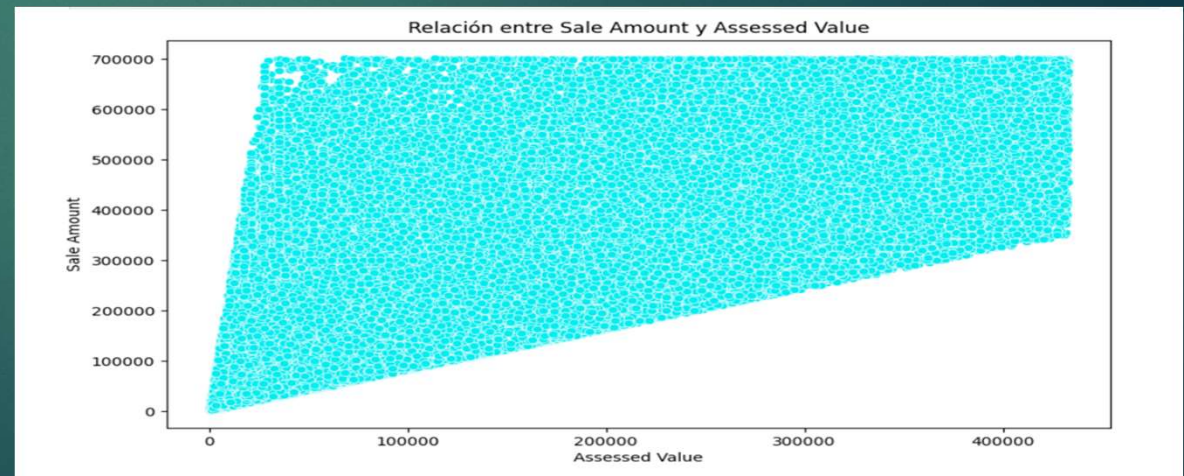
Análisis Univariado.

Podemos observar que la mayor cantidad de registros se encuentran entre los valores de tasación que van desde los 50.000 a 200.000 dólares.



Análisis Bivariado.

Se detecta una relación positiva entre las variables analizadas ya que, a medida que Assessed Value aumenta, Sale Amount también aumenta.

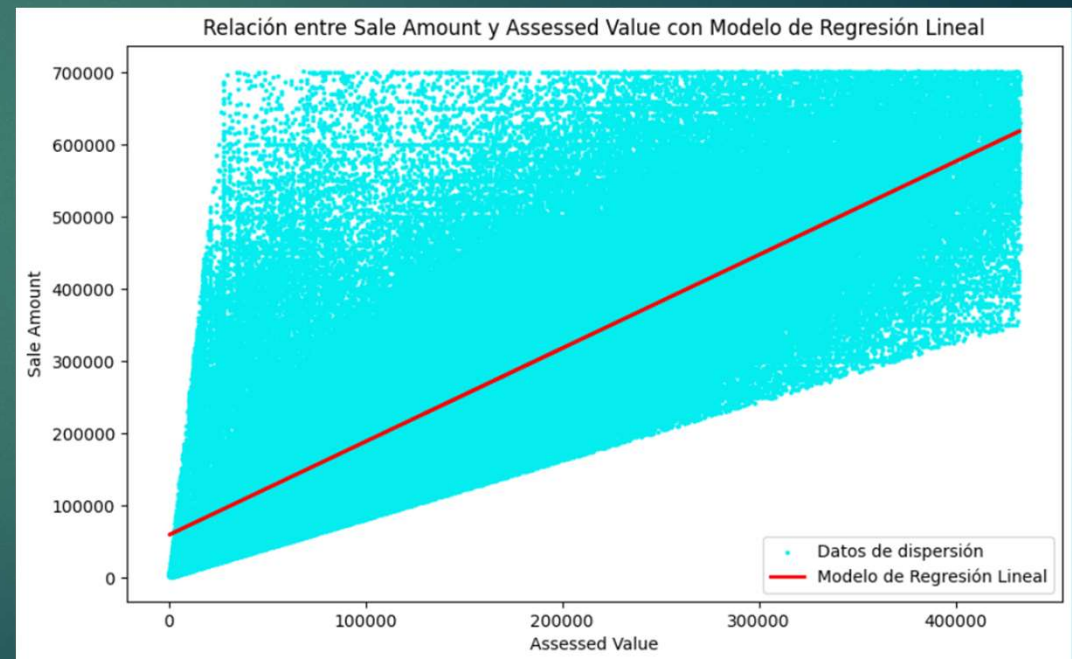


Modelo de regresión lineal inicial.

- ▶ Utilizaremos este método para poder predecir el precio de venta de las viviendas a partir de un historial de ventas en la región de Connecticut.

Relación entre Sale Amount y Assessed Value

- ▶ Intervalo de confianza solo tiene valores positivos, esto confirma la pendiente positiva.
- ▶ El valor p cero (0.000) para "Assessed Value" indica que esta variable es estadísticamente significativa en la predicción del "Sale Amount" en el modelo de regresión lineal.
- ▶ El Coeficiente de determinación (R-cuadrado) es 0.656, lo que significa que aproximadamente el 65.6% de la variabilidad en "Sale Amount" es explicada por las variables independientes incluidas en tu modelo



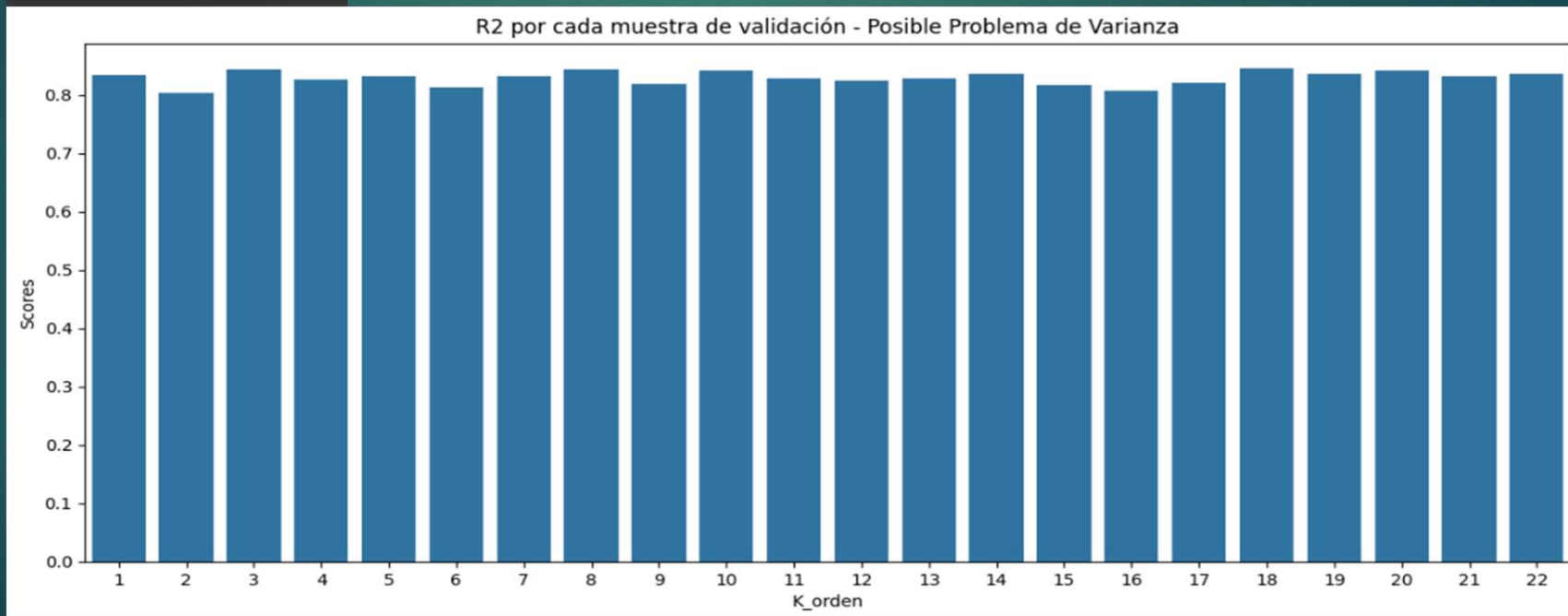
Modelo de regresión lineal final.

- ▶ Luego de modificar algunas variables para convertirlas en numéricas, incorporar y analizar todas las variables que observamos con mayor correlación, nos quedamos con las siguientes.
 - ▶ 1) Assessed Value: valor tasado
 - ▶ 2) Town_mean: ciudad. Variable categórica (Town) transformada por medio de función MEAN para poder utilizar dicha variable en el modelo.
 - ▶ 3) Variables Dummy creadas por medio del método One Hot Encoding.
 - ▶ 4) Nuevos campos creados a partir de campos categóricos:
 - Address: dirección (si el registro tiene dirección = 1, si el registro no tiene dirección = 0)
 - Location: ubicación (si el registro tiene ubicación = 1, si el registro no tiene ubicación = 0)
 - Assessor Remarks: Observaciones del asesor (si el registro tiene observaciones = 1, si el registro no tiene observaciones = 0)
 - OPM remarks: Comentario de la operación (si el registro tiene comentarios = 1, si el registro no tiene comentarios = 0)
- ▶ El Coeficiente de determinación (R-cuadrado) es 0.741, lo que significa que aproximadamente el 74.1% de la variabilidad en "Sale Amount" es explicada por las variables independientes incluidas en tu modelo.

Buscando posibles problemas de Varianza

- ▶ Un posible problema en la varianza del modelo significa que las métricas de validación en distintas submuestras es muy diferente, por lo que podríamos tener problemas de **Generalización**.
- ▶ Como podemos observar en la gráfica, no se detecta problema de varianza.

```
R2 Promedio: 0.8291  
R2 Desvio: 0.0119  
R2 CV: 0.0143
```



Conclusiones generales.

► Podemos obtener las siguientes observaciones:

- Nuestro R-cuadrado mejoró de 0.675 a 0.741, desde nuestro primer modelo al último que aplicamos.
- R-squared (R-cuadrado): El valor de R-cuadrado es 0.741, lo que significa que aproximadamente el 74.1% de la variabilidad en "Sale Amount" es explicada por las variables independientes incluidas en el modelo. Esto nos da un indicio de que el modelo de regresión lineal parece tener un buen ajuste.
- Quitamos todas las variables cuyos P valores > 0.1 y nuestro R-cuadrado continúa siendo de 0.741. Es por esto que utilizaremos este modelo, con P valores < 0.1 que fue el umbral definido para el análisis.
- No se detectaron problemas de Varianza. El desvío y el coeficiente de variación presenten números aceptables.
- Al contar con resultados muy similares tanto en DF_TRAIN como en DF_TEST podríamos decir que no estamos en presencia de sobreajuste o sub ajuste.
- Al aplicar un análisis de hiperparámetros detectamos que el mejores parámetros encontrados es polinomio de grado 1. Originalmente, estábamos utilizando un polinomio de grado 2 en nuestro modelo de regresión lineal. Al reemplazar este polinomio de grado 2 por uno de grado 1, observamos que el coeficiente de determinación (R^2) es inferior al obtenido con el modelo original. Dado este resultado, hemos decidido continuar utilizando el modelo de regresión lineal original, con el polinomio de grado 2.