UNIVERSIDAD DE LOS ANDES
FACULTAD DE INGENIERIA
INGENIERÍA INDUSTRIAL

PROYECTO: ARRENDAMIENTO APARTAMENTOS

Presentado a
Juan F. Perez

Presentado por
Tomas Acosta Bernal - 202011237 - Análisis de Negocio - Ingeniería de Datos
Diego Alejandro Castro - 202115131 - Ciencia de Datos - Despliegue y Mantenimiento
David Felipe Pineda Verano - 202112562 - Tablero de Datos
Juan Sebastian Rojas Abril - 202013406 - Análisis de Datos

BOGOTÁ DC - COLOMBIA
29 de Enero de 2025

Índice general

Capítulo 1

Introduccion

Nuestro proyecto consiste en trabajar con una inmobiliaria que busca identificar las zonas en Estados Unidos donde se distribuyen los precios de arriendos. La finalidad es analizar las características más relevantes de un arriendo para optimizar las variables que influyen en su valor, y así diseñar estrategias que permitan mejorar la oferta y ajustar los precios de manera competitiva.

1.1. Preguntas de negocio

1.1.1. Pregunta 1.

1. ¿Cómo varía el precio de alquiler en función de la ubicación, los metros cuadrados y el número de habitaciones/baños?

Relevancia para el negocio

- Permite a la inmobiliaria identificar qué zonas (estado, ciudad, coordenadas) generan mayor rentabilidad y cómo influyen factores clave como el área o la cantidad de cuartos.
- Ayuda a decidir en qué tipo de inmuebles invertir para maximizar el retorno.

Visualizaciones

- Mapa georreferenciado (usando latitude y longitude) que muestre la distribución de precios por zona.
- Diagramas de dispersión que relacionen price vs. square_feet, diferenciados por número de habitaciones o baños.
- Gráficos de barras o boxplots para comparar el rango de precios en cada ciudad (cityname) o estado (state).

Modelo predictivo

Regresión lineal múltiple que incluya variables como bathrooms, bedrooms, square_feet y ubicación (ciudad, estado o coordenadas transformadas en variables categóricas o numéricas).

1.1.2. Pregunta 2

¿Cuales amenidades (amenities, pets_allowed, etc.) incrementan más el valor esperado de un alquiler?

Relevancia para el negocio

- Las amenidades ofrecidas (piscina, gimnasio, parqueadero, se admiten mascotas, etc.) pueden hacer la diferencia en el precio de alquiler.
- Conocer qué amenidades son más valoradas por los inquilinos ayuda a la empresa a asesorar propietarios sobre mejoras rentables.

Visualizaciones

- Gráfico de barras o heatmap que muestre la frecuencia de diferentes amenidades y su relación promedio con price.
- Boxplots agrupados por pets_allowed (sí/no), para ver si los apartamentos que admiten mascotas tienen un precio mayor o menor en promedio.

Modelo predictivo

Incorporar la variable amenities desglozada en sus categorias como one hot encoding.

Evaluar la importancia de características (feature importance) a través de modelos de árbol (Random Forest) para medir cuantitativamente qué amenidad aporta más al precio.

1.1.3. Pregunta 3.

¿Cómo podemos segmentar los apartamentos en rangos de precio alto, medio y bajo y, cuáles son las variables que más influyen en que un inmueble tenga ese nivel de precio?

Relevancia para el negocio

- Identificar claramente por qué cierto apartamento cae en la categoría "gama baja" podría orientar inversiones en remodelaciones o mejoras que eleven su precio de alquiler.
- Conocer qué características (ubicación, área, número de habitaciones, amenidades, etc.) impulsan un apartamento de un rango a otro permite sugerir mejoras rentables que aumenten el valor de arrendamiento.

Visualizaciones

Histogramas o boxplots para cada rango que muestren la distribución de variables como square_feet, bathrooms, bedrooms, etc. Gráficos de barras o treemaps que reflejen la cantidad de listados en cada categoría (alta, media, baja) según la ciudad (cityname) o estado (state).

Modelo predictivo

- Transformar price en variable categórica (baja, media, alta) y entrenar un modelo de clasificación (por ejemplo, árboles de decisión o random forest) para predecir a qué categoría pertenece un apartamento dado.
- Revisar la importancia de las variables (feature importance) para identificar cuáles factores (ubicación, número de cuartos, amenities, etc.) son más determinantes en cada categoría de precio.

1.2. Exploración de datos

1.2.1. Limpieza y Clasificación de las Variables

Inicialmente, se establecieron los ìndices de cada característica de los datos, corrigiendo errores, eliminando espacios, comas y otros caracteres innecesarios. A partir de esta limpieza y organización, se clasificaron las variables en dos tipos: **numéricas** y **categóricas**, permitiendo un análisis estructurado según su naturaleza.

1.2.2. Anàlisis de Variables Numéricas

Se comenzó con la obtención de estadísticas generales de las variables numéricas, como media, deviación, etc. . Posteriormente, se generaron histogramas para visualizar la distribución de estas variables.

Además, se elaboraron diagramas de caja para identificar la dispersión de los datos y la presencia de valores atípicos. Para complementar el análisis, se realizaron gráficos de dispersión y gráficos de violín, con el fin de explorar

la variabilidad y distribución de los datos con respecto al Precio Vs. diferentes variables.

Asimismo, se llevaron a cabo pruebas de **independencia lineal**, **Homocedasticidad**, **normalidad** y **regresión lineal**. Podemos concluir que hay una gran concentración de valores cercanos a cero y una cola larga hacia la derecha, por otro lado no sigue una normalidad estricta, dado que los puntos no se alinean adecuadamente y ademas se observa que los residuos están mayormente concentrados cerca de cero, esto sugiere la presencia de heterocedasticidad y una posible autocorrelación. Para mayor entendimiento, se pueden observar en el notebook las graficas y evaluaciones correspondientes elaboradas.

1.2.3. Análisis de Variables Categóricas

Para el caso de las variables categóricas, se calcularon sus estadísticas descriptivas, determinando la frecuencia de cada categoría. A continuación, se construyeron gráficos de barras para facilitar la interpretación de la distribución de estas variables dentro del conjunto de datos. Este análisis permitió una comprensión detallada de la estructura de los datos y sirvió como base para la aplicación de modelos predictivos en etapas posteriores.

1.3. Preparación de los datos

La preparación de datos comenzó con un análisis exhaustivo de un conjunto de 10,000 registros sobre apartamentos en renta distribuidos en 22 columnas; se identificaron y eliminaron columnas redundantes o poco relevantes (como ID, categoría, título, cuerpo del anuncio y varias relacionadas con el precio), y para garantizar la calidad de las variables seleccionadas se implementaron dos decisiones clave: primero, conservar únicamente las variables con información directamente relacionada con su respectiva columna; segundo, excluir aquellas en las que la diferencia entre categorías superaba el 90 % (es decir, si una categoría representaba más del 90 % de los datos), evitando así sesgos y preservando la utilidad del conjunto de datos para el modelado posterior.

Un aspecto crucial del preprocesamiento fue el tratamiento de la columna de amenidades. Esta información se transformó utilizando MultiLabelBinarizer para convertir las características en columnas binarias individuales. Esta transformación permitió un análisis más detallado y cuantificable de las comodidades ofrecidas en cada propiedad, facilitando la posterior evaluación de su impacto en el precio.

La limpieza de datos numéricos constituyó una parte fundamental del proceso. Se normalizaron las columnas de baños y dormitorios, creando una nueva característica que promedia ambos valores (bathrooms_bedrooms_avg). En cuanto a los valores atípicos en el precio, se implementó una estrategia de eliminación para aquellos que superaban el penúltimo valor más alto del conjunto de datos ya que se excedia por casi 40.000 de valore frente al promedio, ademas estos podían introducir sesgos significativos en el análisis.

Las variables categóricas (por ejemplo, ciudad, estado y fuente) se estandarizaron asignando valores "UNK" u "Other" a datos faltantes o para equilibrar categorías con pocos casos, reduciendo así la dimensionalidad y conservando la información esencial. Además, se creó una nueva variable categórica (price_category) a partir de los percentiles 33 y 66, clasificando los precios en "Bajo", "Medio" y "Alto" para facilitar análisis de segmentación. Finalmente, el conjunto de datos resultante, compuesto por 9,238 registros y 22 columnas procesadas, se dividió en entrenamiento (70 %) y prueba (30 %), asegurando tanto un volumen suficiente para un entrenamiento robusto como un conjunto de validación significativo para evaluar el rendimiento de los modelos predictivos.

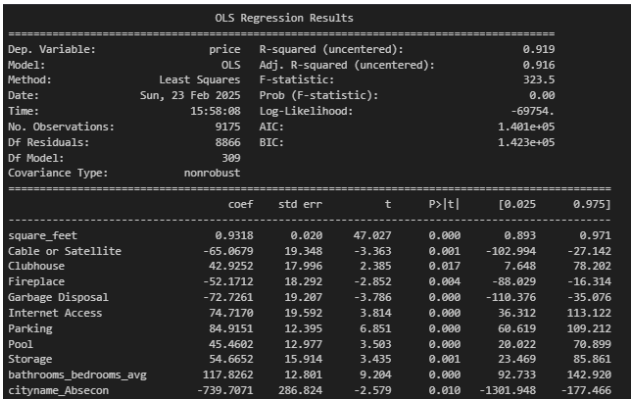
1.4. Modelamiento

1.4.1. Regresión

Para iniciar el modelamiento se procedió con la construcción de modelos de regresión, como primer paso. Se realizaron varias iteraciones con el objetivo de obtener un modelo robusto y significativo para la predicción de precios. De este modo se siguió el siguiente paso para la regresión:

1. Conversión a Variables Dummies: Para mejorar la interpretabilidad y la capacidad predictiva del modelo, todas las variables de tipo float fueron convertidas en variables dummies.

2. Regresión Inicial: Se efectuó una primera regresión considerando todas las variables disponibles en el conjunto de datos. Sin embargo, se observó que muchos parámetros no resultaban significativos.
3. Filtrado de Variables: Para mejorar la calidad del modelo, se aplicó un criterio de selección basado en la significancia estadística. Se conservaron únicamente aquellas variables con un nivel de significancia superior al 10
4. Regresión Refinada: Con el nuevo conjunto de variables filtradas, se ejecutó una regresión tomando como variable dependiente el precio. Despues de relizar esto.



OLS Regression Results

Dep. Variable:	price	R-squared (uncentered):	0.919
Model:	OLS	Adj. R-squared (uncentered):	0.916
Method:	Least Squares	F-statistic:	323.5
Date:	Sun, 23 Feb 2025	Prob (F-statistic):	0.00
Time:	15:58:08	Log-Likelihood:	-69754.
No. Observations:	9175	AIC:	1.401e+05
Df Residuals:	8866	BIC:	1.423e+05
Df Model:	309		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
square_feet	0.9318	0.020	47.027	0.000	0.893	0.971
Cable or Satellite	-65.0679	19.348	-3.363	0.001	-102.994	-27.142
Clubhouse	42.9252	17.996	2.385	0.017	7.648	78.202
Fireplace	-52.1712	18.292	-2.852	0.004	-88.029	-16.314
Garbage Disposal	-72.7261	19.207	-3.786	0.000	-110.376	-35.076
Internet Access	74.7170	19.592	3.814	0.000	36.312	113.122
Parking	84.9151	12.395	6.851	0.000	60.619	109.212
Pool	45.4602	12.977	3.503	0.000	20.022	70.899
Storage	54.6652	15.914	3.435	0.001	23.469	85.861
bathrooms_bedrooms_avg	117.8262	12.801	9.204	0.000	92.733	142.920
cityname_Absecon	-739.7071	286.824	-2.579	0.010	-1301.948	-177.466

Figura 1.1: Regresión con datos Filtrados

Como se observa en la figura se sacan todos los estadísticos y los p-values que cumplieran la regla anteriormente descrita y los diferentes estadísticos que afirman que tenemos un modelo robusto con datos significativos.

1.4.2. Random Forest

Adicionalmente, se utilizó un modelo de Random Forest Regressor para la predicción de la categoría de precios. Antes de entrenar el modelo, la variable dependiente pricecategory fue transformada a valores numéricos asignando pesos: bajo (0), medio (1) y alto (2).

Métrica	**Valor**
Accuracy en entrenamiento (R^2)	0.90
Accuracy en test (R^2)	0.33
MAE (Error Absoluto Medio)	0.49
RMSE (Raíz del Error Cuadrático Medio)	0.66

Tabla 1.1: Resultados del modelo Random Forest

El modelo muestra un alto desempeño en el conjunto de entrenamiento, pero una menor capacidad de generalización en el conjunto de prueba, lo que sugiere posible sobreajuste. Se recomienda evaluar ajustes en los hiperparámetros y considerar técnicas como reducción de dimensionalidad o regularización para mejorar el rendimiento del modelo en datos no vistos.

Por otro lado se desarrolló otro modelo de Random Forest capaz de clasficiar estas variables obteniendo los siguientes resultados.

El modelo obtuvo una precisión del 71.27 porciento y un F1 Score de 71.23 porciento, lo que indica un desempeño razonable en la clasificación de las categorías de precios. Se observa que el modelo mantiene un buen equilibrio entre precisión y recall, lo que sugiere que las clases están bien diferenciadas. Sin embargo, se podrían probar técnicas de balanceo de datos o ajuste de hiperparámetros para mejorar la predicción en categorías con menor representación.

Despues sacando el modelo se sacaron las variables más significativas de este modelo para tenerlas en cuenta en las preguntas del análisis financiero.

Métrica	Valor
F1 Macro	0.7142668506109578
F1 Micro	0.7149863760217984
Precisión	0.7126939248948818
Accuracy	0.7149863760217984
F1 Score	0.7123068874782054
Recall	0.7149863760217984

Tabla 1.2: Resultados del modelo Random Forest - Clasificación

1.4.3. Red neuronal

Se implementó una Red Neuronal utilizando las mismas variables para la predicción de la categoría de precios. Los datos fueron escalados antes del entrenamiento, asegurando que los valores estuvieran en un rango adecuado para la red neuronal.

Métrica	Valor
Accuracy (training set)	0.91
Accuracy (test set)	0.71
Mín. y máx. de `X_train_scaled`	0.0 / 1.0000000000000002

Tabla 1.3: Resultados del clasificador NN

El modelo de red neuronal mostró un mejor desempeño en el conjunto de prueba en comparación con el modelo de Random Forest, lo que indica que la red neuronal logró capturar mejor las relaciones no lineales en los datos.

1.5. Diseño del Tablero

El tablero desarrollado en Dash permite ingresar y visualizar información clave sobre precios de alquiler y sus determinantes. A continuación, se describen los principales elementos considerados:

- **Valores de entrada:** Se permite al usuario filtrar por ubicación (ciudad, estado), metros cuadrados, número de habitaciones y baños, así como por amenidades.
- **Resultados generados:** Se muestran gráficos interactivos que permiten analizar tendencias y patrones en los precios de alquiler.
- **Visualizaciones:**
  - Mapa georreferenciado que muestra la distribución de precios por zona.
  - Diagramas de dispersión para relacionar precio vs. área, diferenciados por número de habitaciones o baños.
  - Gráficos de barras y boxplots para comparar el rango de precios en distintas ciudades y estados.
  - Gráficos que muestran la influencia de las amenidades en el precio de alquiler.
- **Instrucciones para el usuario:** Se presentan en una sección visible del tablero, explicando cómo interactuar con los filtros y las visualizaciones.
- **Distribución de elementos:** Se organizan de manera intuitiva para facilitar la navegación y el análisis de datos.

1.5.1. Respuestas a las Preguntas de Interés

Pregunta 1: Variación del Precio en Función de la Ubicación y Características

Relevancia para el negocio: Permite identificar qué zonas generan mayor rentabilidad y cómo influyen factores clave como el área o la cantidad de cuartos. Esto ayuda a decidir en qué tipo de inmuebles invertir.

Visualizaciones:

- Mapa georreferenciado con distribución de precios.
- Diagramas de dispersión precio vs. metros cuadrados, diferenciados por número de habitaciones.
- Boxplots y gráficos de barras comparando precios en distintas ciudades y estados.

Modelo predictivo: Se implementó una regresión lineal múltiple que incluye variables como número de baños, habitaciones, área y ubicación.

Pregunta 2: Influencia de las Amenidades en el Precio del Alquiler

Relevancia para el negocio: Permite conocer qué amenidades son más valoradas y asesorar a propietarios sobre mejoras rentables.

Visualizaciones:

- Gráfico de barras o heatmap mostrando la relación de cada amenidad con el precio promedio.
- Boxplots para comparar precios entre apartamentos que admiten o no mascotas.

Modelo predictivo: Se empleó un modelo de árboles de decisión (Random Forest) para evaluar la importancia de las diferentes amenidades.

Pregunta 3: Segmentación de Apartamentos en Rango de Precios

Relevancia para el negocio: Identificar por qué un apartamento pertenece a una categoría de precio específica ayuda a tomar decisiones estratégicas para mejorar su rentabilidad.

Visualizaciones:

- Histogramas y boxplots mostrando la distribución de variables como metros cuadrados y número de habitaciones en cada categoría de precio.
- Gráficos de barras o treemaps reflejando la cantidad de listados por categoría y ciudad.

Modelo predictivo: Se transformó la variable precio en categorías (baja, media, alta) y se entrenó un modelo de clasificación con árboles de decisión para predecir la categoría de un apartamento.

1.5.2. Conclusión

El tablero permite explorar el impacto de diferentes variables en los precios de alquiler mediante visualizaciones interactivas y modelos predictivos. Facilita la toma de decisiones informadas para la inversión inmobiliaria y la mejora de propiedades en alquiler.

1.6. Evaluacion

1.6.1. Pregunta Negocio 1

Para abordar la primera pregunta de negocio se aplicó un modelo de regresión lineal, que reveló que las variables determinantes son la ubicación, el tamaño (en metros cuadrados) y la cantidad de habitaciones y baños. A continuación, se detalla cómo cada una de estas variables influye en el precio de alquiler:

- La variable square_feet presenta un coeficiente positivo (0.93). Esto indica que, a mayor área del inmueble, mayor es el precio de alquiler, lo que respalda la importancia del tamaño como factor de valoración.

- La variable bathrooms_bedrooms_avg tiene un coeficiente muy elevado (117.83), lo que sugiere que la combinación o equilibrio entre el número de habitaciones y baños influye de manera significativa en el precio. Es decir, un mayor confort o funcionalidad del inmueble se traduce en un valor de alquiler más alto.

- Las variables relacionadas con la ubicación (identificadas como cityname_ y state_) muestran coeficientes de gran magnitud, tanto positivos como negativos. Esto indica que la zona donde se encuentra el inmueble es probablemente el factor más determinante en el precio de alquiler. Por ejemplo, ciudades como Hollywood, Brookline o Palo Alto presentan coeficientes muy altos, sugiriendo que en esas áreas los precios tienden a ser considerablemente superiores. En contraste, otras ciudades muestran coeficientes negativos, lo que indica precios relativamente más bajos en comparación con la categoría de referencia.

1.6.2. Pregunta 2

Al examinar los coeficientes del modelo podemos concluir lo siguiente respecto a las amenidades que afectan el valor de alquiler:

- Internet Access (coef. +74.72): La conectividad es altamente valorada en el mercado actual, lo que indica que ofrecer acceso a internet robusto puede justificar un precio de alquiler superior.
- Parking (coef. +84.92): El estacionamiento es crucial, especialmente en zonas urbanas con alta densidad de tráfico o en áreas donde el espacio es limitado.
- Storage (coef. +54.67): Disponer de espacio adicional para almacenamiento es percibido como un valor agregado importante, aumentando el atractivo del inmueble.
- Clubhouse y Pool (coef. +42.93 y +45.46, respectivamente): Estas amenidades recreativas generan valor, probablemente por el estilo de vida que ofrecen y la diferenciación en el mercado.
- La variable bathrooms_bedrooms_avg (coef. +117.83) no es una amenidad per se, pero refleja que una mejor configuración interna (una buena proporción entre baños y dormitorios) tiene un impacto significativo en el precio, lo cual es crucial al definir el perfil del inmueble.

Algunas características como Cable or Satellite, Fireplace y Garbage Disposal presentan coeficientes negativos (aproximadamente -65.07, -52.17 y -72.73, respectivamente). Esto podría indicar que estas amenidades, tal vez en combinación con otras variables, se encuentran en inmuebles de menor segmento o en contextos donde no aportan tanto valor percibido. Otra posibilidad es que, en el contexto actual, estas características sean menos determinantes para los inquilinos o incluso puedan ser sustituidas por alternativas más modernas.

Las amenidades que generan un impacto positivo, especialmente aquellas relacionadas con conectividad, estacionamiento y espacios adicionales, son áreas estratégicas en las que enfocar las inversiones. Recomendar a propietarios invertir en mejoras como el acceso a internet de alta calidad, ampliar opciones de estacionamiento y proporcionar almacenamiento extra puede aumentar notablemente el valor de alquiler de sus propiedades.

Dado que los inquilinos valoran significativamente amenidades como las áreas recreativas (clubhouse, pool) y la configuración interna óptima (balance entre baños y dormitorios), estos aspectos pueden convertirse en un factor diferenciador en mercados competitivos.

1.6.3. Pregunta 3

Para abordar esta pregunta, mediante Random Forest para clasificacion se agruparon los apartamentos en tres categorías: alto, medio y bajo precio. Una vez definidos los grupos,fue posible analizar qué variables tienen mayor

peso en la clasificación de cada inmueble. A partir de los resultados obtenidos, podemos destacar lo siguiente:

Variables con mayor impacto global:

- square_feet (0.198): El área del inmueble es la variable más influyente. Un mayor tamaño se asocia, en general, a un precio de alquiler superior.
- longitude (0.086) y latitude (0.065): Estas variables capturan la ubicación geográfica. La posición exacta del apartamento influye notablemente en el precio, reflejando la importancia de la zona o barrio.
- bathrooms_bedrooms_avg (0.074): La proporción o promedio entre baños y dormitorios es clave para determinar la calidad interna del inmueble y, por ende, su precio.
- state_CA (0.021): Indica que, en el caso de California, los apartamentos tienden a tener precios diferentes (probablemente más altos) en comparación con otras regiones, reflejando dinámicas de mercado locales.
- pets_allowed (0.019): Permitir mascotas añade un valor percibido, aunque su impacto es menor en comparación con las variables anteriores. Parking (0.016), Pool (0.014), Patio/Deck (0.014) y Dishwasher (0.013): Estas amenidades, si bien tienen coeficientes menores, en conjunto pueden influir en la decisión de arrendamiento, particularmente en mercados donde los servicios y la comodidad son diferenciadores clave.

Identificar que el área del inmueble (square_feet) y la ubicación (capturada a través de longitude y latitude) son los principales impulsores del precio permite focalizar las inversiones en propiedades que, por su tamaño y localización, tienen mayor potencial de rentabilidad. Además, mejorar la proporción entre habitaciones y baños (bathrooms_bedrooms_avg) puede ser una estrategia de remodelación para pasar de una categoría de precio bajo a una de precio medio o alto. Saber que características específicas (como la presencia de estacionamiento, piscina o áreas al aire libre) aportan valor adicional ayuda a asesorar a propietarios para que realicen mejoras enfocadas. Por ejemplo, en mercados competitivos o en inmuebles catalogados como gama baja, incluir amenidades como Parking o Pool puede ser un factor diferenciador que incremente el precio de alquiler.

La segmentación por rangos de precio, combinada con el análisis de las variables que más influyen (tamaño, ubicación, configuración interna y ciertas amenidades), permite a la inmobiliaria diseñar estrategias de inversión y marketing más precisas y orientadas a maximizar la rentabilidad de su portafolio de apartamentos.

1.7. Despliegue y mantenimiento

Para el despliegue, primero creamos una instancia en EC2 y transferimos el archivo Tablero1.py desde la terminal de la máquina local hacia la máquina virtual. Luego instalamos las dependencias necesarias y subimos los archivos CSV solicitados por el script. Una vez configurada la máquina virtual y lista para un entorno interactivo, la detuvimos con el fin de no consumir créditos innecesariamente. Los soportes se pueden visualizar en el repositorio.