William Bayona 202011494

Martin Vásquez 202113314

Daniela Echavarria Yepes 202111348

Informe Proyecto 1

1.Pregunta de Negocio – William

Cliente Seleccionado: El potencial cliente es una agencia inmobiliaria que desea optimizar su portafolio, entender la dinámica del mercado y ofrecer mejores recomendaciones tanto a propietarios como a inquilinos.

Pregunta General:

"¿Cómo influyen las características del apartamento, la calidad del anuncio, las políticas de mercado en la variación de precios de alquiler del mercado inmobiliario?"

Preguntas Especificas:

1. Análisis de Precios y Tendencias

¿Cuál es la distribución de precios de alquiler por ciudad y estado?

Se propone una visualización de un mapa sobre los datos de precios, distribuidos por su latitud y longitud.

2. Factores que Influyen en el Precio

¿Cómo afectan las características físicas (habitaciones, baños, tamaño) y las fotos al precio del alquiler?

Se propone una regresión lineal múltiple que involucre las siguientes variables: Numero de Baños, Numero de Habitaciones, Tiene Foto, Se permiten mascotas, tamaño (en pies cuadrados), Estado, Origen del anuncio.

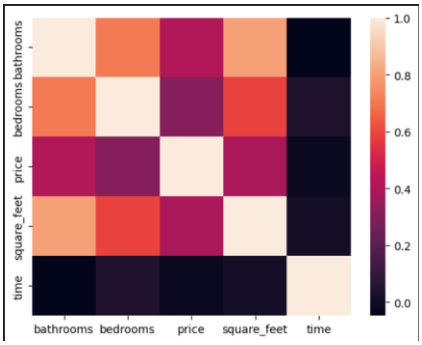
3. Calidad del Anuncio y su Impacto

¿Cómo influyen las descripciones y palabras clave en el precio?

Se propone un modelo de Random Forest basado en TF-IDF para evaluar la importancia de las palabras en títulos y descripciones, mejorando el análisis y clasificación del texto.

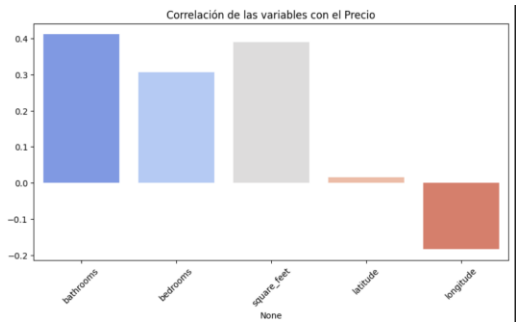
2.Exploración de Datos-Daniela

Se realiza un gráfico que muestra la correlación entre las variables independientes



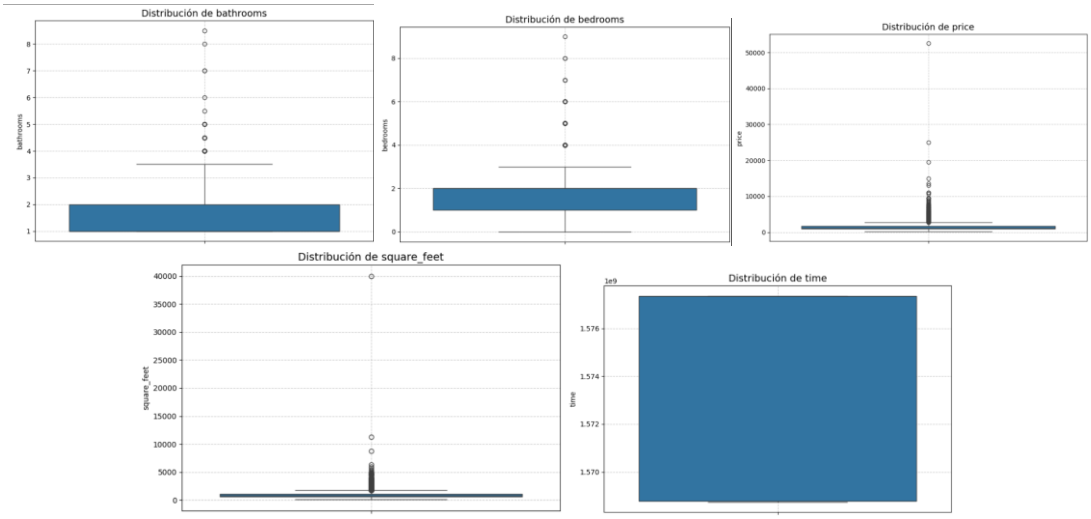
En este gráfico se puede visualizar que baños y el tamaño de un apartamento están fuertemente correlacionados con un índice de 0,9.

Asimismo, se realizó otro gráfico que compara las variables independientes con respecto a la variable dependiente (precio).



Con esta gráfica se puede determinar que la latitud y la longitud están poco correlacionadas con el precio.

Adicionalmente, se gráfico boxplots de las variables numéricas.



Se puede observar que el tamaño y el precio son las únicas variables con una gran cantidad de datos atípicos. Por otro lado, la variable tiempo presenta una gran variabilidad dentro de sus datos.

3.Preparación de Datos - Daniela

Lo primero que se realiza es eliminar duplicados del archivo, luego se eliminan datos vacíos, pero esto reduce en gran tamaño los datos, es por esta razón que se decide realizar un input a la variable pets_allowed cambiando los datos nulos a 0, no se admiten mascotas, y 1 si admitían. Asimismo, se transformó la columna has_photo a 1 si era un thumbnail o yes, de resto es un cero.

Posteriormente, se escogieron las columnas necesarias para hacer la regresión, por ejemplo, baños, habitaciones, tamaño, ya que son las que mayor correlación tienen con la variable dependiente precio. Adicionalmente, para la regresión de texto se escogieron el título, el cuerpo y el precio.

Por último, se eliminan columnas vacías y datos atípicos del precio y tamaño de los dataframes text y regresión.

4.Modelamiento – Juan Martin

Para responder a las preguntas sobre la variación de precios en el mercado inmobiliario, se desarrollaron tres modelos predictivos que permiten analizar el impacto de diferentes factores en el precio del alquiler. Para comenzar se hace una separación de los datos en 2 conjuntos, uno de entrenamiento con 80% de los datos y uno de test, con el 20% restante.

El primer modelo es una regresión lineal múltiple que utiliza variables como el número de habitaciones, baños, el tamaño del apartamento, la presencia de fotos en el anuncio, la política de mascotas, el estado y la fuente del anuncio. Estas variables se transformaron mediante un pipeline que incluye preprocesamiento de datos categóricos y numéricos, aplicando OneHotEncoder para crear las variables dummies necesarias para las variables categóricas. Esto permitió lograr un modelo que estima el precio del alquiler con base en características estructurales del inmueble con un R^2 de 0.540 teniendo un ajuste moderado. Las betas de la regresión se guardaron para utilizarse en la sección de predicción en el Dash.

Adicionalmente, se desarrollaron dos modelos de regresión basados en texto. Para ello, se empleó la concatenación del título y la descripción del anuncio, los cules se limpiaron y lematizaron (reduciéndolos a su raíz

lingüística). Posteriormente convirtiéndolos en representaciones numéricas mediante un vectorizador TF-IDF (Term Frequency - Inverse Document Frequency)una técnica de procesamiento de texto que convierte palabras en valores numéricos según su importancia en un conjunto de documentos. Se basa en la frecuencia de cada término en un documento (TF) y su rareza en la colección total de documentos (IDF), dando más peso a las palabras relevantes y reduciendo la influencia de términos comunes. Posteriormente, se entrenaron dos enfoques diferentes a partir de esta matriz: un modelo Random Forest y un modelo basado en K-Nearest Neighbors (KNN), con el objetivo de predecir cómo la redacción del anuncio influye en el precio estimado del alquiler. Para el modelo KNN, se realizó una optimización de hiperparámetros utilizando GridSearchCV llegando a 9 vecinos y distancia manhattan. Sin embargo, los resultados obtenidos no lograron un buen R^2 (0.314) ni un buen MAE (309.23) para los conjuntos de prueba, la gran discrepancia entre los R^2 (0.99) y MAE (3.194)de entrenamiento y test indica un claro problema de sobreajuste, pues es casi perfecto en para entrenamiento y muy malo para test. Por lo que este modelo no fue utilizado en el Dash. A diferencia el modelo random forest con los parámetros por defecto el cual obtuvo un R^2 (0.92) y MAE (98.34) de entrenamiento bueno pero no perfecto y en los conjunto de datos de test el R^2 (0.501)y MAE (260.65) tiene un ajuste moderado, el cual es el elegido para usarse en el dash.

Para optimizar el tiempo de ejecución en el Dash y evitar la necesidad de reentrenar los modelos en cada consulta, se exportaron los modelos entrenados en formato pkl. Esto permite que sean utilizados de manera instantánea en tiempo real dentro del Dash, mejorando la eficiencia y la experiencia del usuario.

5.Diseño y desarrollo del tablero - William

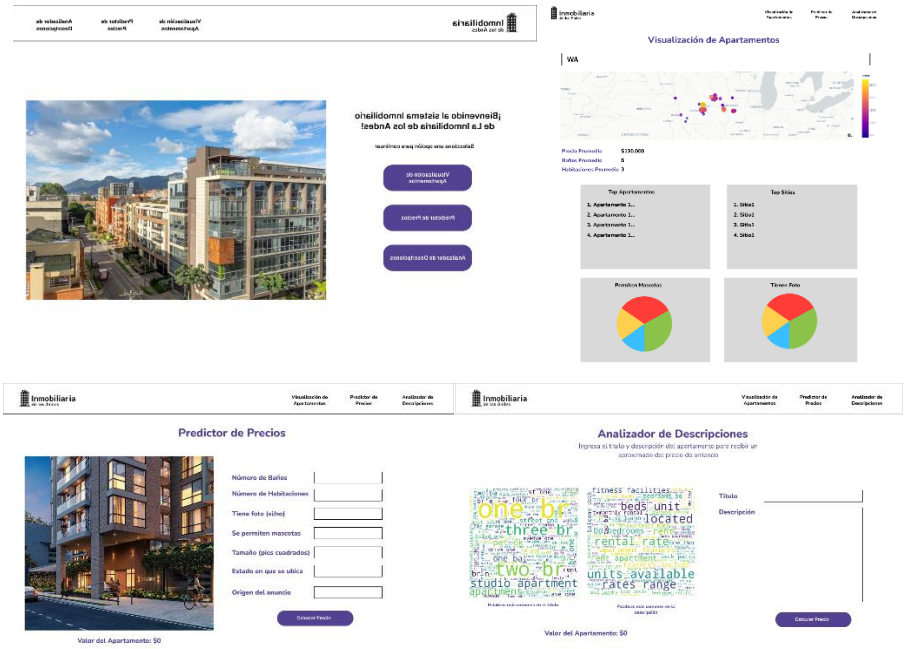
El tablero ha sido diseñado con tres pantallas principales para facilitar la exploración y predicción de precios en el mercado inmobiliario. Se han creado wireframes previos para definir la estructura y disposición de los elementos, asegurando una experiencia de usuario clara e intuitiva.

En la primera pantalla, los usuarios pueden visualizar información relevante del arrendamiento de filtrada por estado, lo que les permite explorar tendencias y comparar precios en diferentes estados.

En la segunda pantalla, un formulario permite ingresar características clave del inmueble, como metros cuadrados y número de habitaciones, para realizar una regresión lineal y estimar su precio basado en datos históricos.

Finalmente, en la tercera pantalla, los usuarios pueden ingresar un título y una descripción del inmueble para obtener un precio estimado utilizando procesamiento de lenguaje natural.

Se presentan a continuación los wireframes realizados:



6.Evaluación – William

Los resultados obtenidos permiten evaluar cómo las características del apartamento, la calidad del anuncio, las políticas de mercado y las tendencias temporales influyen en la variación de los precios de alquiler del mercado inmobiliario.

En la primera pantalla, a través del análisis de precios y tendencias, se identificó la distribución de precios por ciudad y estado mediante visualizaciones interactivas en un mapa con filtros. Esta visualización responde de manera adecuada a la primera pregunta de negocio, ya que permite analizar la variabilidad de los precios en diferentes ubicaciones y facilita la identificación de patrones espaciales.

En cuanto a los factores que influyen en el precio, la regresión lineal múltiple demostró que variables como el número de habitaciones, baños, la presencia de fotos, la política de mascotas, el tamaño del inmueble y el origen del anuncio tienen un impacto significativo en la estimación del precio. Con un R^2 de 0.54 y un MAE de 257, el modelo ofrece una aproximación significativa a la relación entre las características del inmueble y su precio de alquiler, proporcionando una respuesta satisfactoria a la segunda pregunta de negocio. Sin embargo, la precisión del modelo podría mejorarse mediante técnicas de ajuste y la inclusión de variables adicionales.

Respecto al impacto de la calidad del anuncio en el precio, el análisis mediante un modelo de Random Forest basado en una matriz TF-IDF permitió evaluar la relevancia de las palabras en títulos y descripciones, destacando en la fijación del precio. Aunque el modelo mostró un R^2 de prueba de 0.5041 con un MAE de 259.36, su desempeño es suficiente para identificar tendencias y factores clave en la redacción de anuncios efectivos.

En manera conjunta estos hallazgos permiten, en efecto, definir de que forma las características del apartamento, la calidad del anuncio, las políticas de mercado influyen en la variación de precios de alquiler del mercado inmobiliario, por lo que se concluye afirmativamente que ha sido resuelta la pregunta principal de negocio.

7.Despliegue y mantenimiento – Juan Martin

Para poder tener una mejor visualización de datos como se menciona en el punto anterior se implementó la herramienta previamente diseñada y esta misma fue subida a una maquina virtual de AWS. El tablero se puede acceder desde e siguiente enlace:

http://54.175.101.168:8050/home