



Lead University

Gestion de tecnologia digital

Trabajo 9 - Avance del proyecto final

Carlos Aguirre

Gabriel Guzmán Mora

Andrey Mora Campos

Objetivo del proyecto

El objetivo principal de este proyecto es desarrollar un modelo de machine learning capaz de predecir el costo de una vivienda basándose en diversas características relevantes. Este modelo buscará integrar variables como la ubicación, tamaño, edad de la propiedad, y otras características específicas del inmueble para estimar su valor en el mercado actual. Al implementar técnicas avanzadas de aprendizaje automático, el proyecto no solo aspira a proporcionar estimaciones precisas que puedan ser útiles para compradores y vendedores, sino también ofrecer insights significativos a inversores y entidades financieras que operan en el sector inmobiliario.

El modelo resultante se espera que sea una herramienta robusta en la toma de decisiones, mejorando la transparencia del mercado y ayudando a nuestras partes interesadas a realizar inversiones más informadas. Además, se buscará que el modelo pueda adaptarse a cambios del mercado y a nuevas tendencias de datos, asegurando así su relevancia y utilidad a largo plazo. Este esfuerzo está alineado con nuestro compromiso de aplicar tecnología de punta para resolver problemas complejos y generar valor tanto para la empresa como para nuestros clientes.

Definición del proyecto

El uso de machine learning en la predicción del costo de las viviendas se justifica por varias razones clave. Primero, el valor de una propiedad depende de una compleja interacción de múltiples factores, incluyendo características físicas, ubicación geográfica, condiciones del mercado, y tendencias económicas, entre otros. El machine learning es capaz de manejar y analizar grandes volúmenes de datos multidimensionales y extraer patrones útiles que pueden no ser evidentes a simple vista o mediante métodos analíticos tradicionales.

Segundo, las técnicas de machine learning se adaptan continuamente a los nuevos datos, lo que permite que los modelos de predicción se actualicen y mejoren con el tiempo conforme cambian las condiciones del mercado y se disponga de nueva información. Esto es crucial en el mercado inmobiliario donde las dinámicas de precios pueden fluctuar significativamente debido a factores económicos y sociales.

Finalmente, el empleo de estas técnicas permite automatizar y escalar el proceso de valoración de propiedades, ofreciendo evaluaciones rápidas y rentables que benefician tanto a los negocios como a los consumidores. Al integrar machine learning, podemos proporcionar estimaciones de precios más precisas y personalizadas, potenciando la toma de decisiones informada y eficiente en el ámbito inmobiliario.

Tecnologías a usar:

Dado que el proyecto se centra en el desarrollo de un modelo de machine learning para predecir el costo de viviendas, las herramientas y tecnologías que podrían ser relevantes incluyen:

1. **Terraform:** útil para la gestión de la infraestructura como código, permitiendo la creación y gestión de recursos de infraestructura de manera automatizada y reproducible.
2. **Dataspell:** Creación, gestión y despliegue de modelos de machine learning. Facilitará el desarrollo de modelos predictivos y su integración con otras herramientas y sistemas.
3. **PyCharm:** Entorno de desarrollo integrado (IDE) para Python que proporciona funcionalidades avanzadas para la escritura, depuración y gestión de código Python, incluyendo soporte para la creación de modelos de machine learning.
4. **Docker:** Será utilizado para la creación de contenedores que encapsulan el entorno de ejecución del modelo de machine learning y sus dependencias, garantizando así la portabilidad y consistencia en diferentes entornos de desarrollo y producción.
5. **Postman:** Nos ayudará a probar y documentar APIs utilizadas en el proyecto, lo que incluiría APIs para el acceso a datos relevantes, como datos históricos de ventas de viviendas o información demográfica.
6. **Git y GitHub:** Git será utilizado para el control de versiones del código fuente del proyecto, permitiendo el seguimiento de cambios, la colaboración entre equipos y la gestión de ramas de desarrollo. GitHub se utilizará como plataforma para alojar el repositorio de Git, facilitando la colaboración y la integración continua.

Diseño y diagrama de la arquitectura

El diseño de la arquitectura para este proyecto podría seguir un enfoque de aprendizaje supervisado, donde se entrena un modelo de machine learning utilizando datos históricos de ventas de viviendas y características de las propiedades. Nuestra propuesta para el diseño y diagrama de la arquitectura sería:

1. **Adquisición de datos:** Se recopilarán datos históricos de ventas de viviendas, así como características relevantes de las propiedades, como ubicación, tamaño, edad, etc. Estos datos pueden provenir de fuentes públicas, bases de datos especializadas o APIs de terceros.
2. **Preprocesamiento de datos:** Los datos adquiridos se pre procesarán y limpiarán para eliminar valores atípicos, manejar valores faltantes y transformar

características si es necesario. Esto puede implicar la normalización de datos numéricos y la codificación de variables categóricas.

3. **División de datos:** Los datos se dividirán en conjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo.
4. **Selección y entrenamiento del modelo:** Se seleccionará un algoritmo de machine learning apropiado, como regresión lineal, regresión de árboles de decisión o redes neuronales, y se entrenará utilizando el conjunto de entrenamiento.
5. **Evaluación del modelo:** Se evaluará el rendimiento del modelo utilizando métricas como el error cuadrático medio (MSE) o el coeficiente de determinación (R^2) en el conjunto de prueba.
6. **Despliegue del modelo:** Una vez que el modelo ha sido entrenado y evaluado satisfactoriamente, se desplegará en un entorno de producción donde podrá ser utilizado para predecir el costo de viviendas en tiempo real.
7. **Monitorización y actualización:** Se implementará un sistema de monitorización para realizar un seguimiento del rendimiento del modelo en producción y se actualizará periódicamente con nuevos datos para garantizar su precisión y relevancia a lo largo del tiempo.

Roles

Andrey Mora Campos - Data Scientist - Data Engineer

Gabriel Guzman Mora - Machine Learning Operations Engineer

Carlos Aguirre - Data Scientist